Landwirtschaftliche Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität zu Bonn

Institut für Photogrammetrie

Der EM-Algorithmus

bei Schätz- und Klassifikationsproblemen

Diplomarbeit

zur
Erlangung des Grades
Diplom-Ingenieur
(Dipl.-Ing.)
der
Fachrichtung
Vermessungswesen

Vorgelegt am 13. November 2000 von
Marc Luxen
aus Aremberg

Landwirtschaftliche Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität zu Bonn

Institut für Photogrammetrie

Der EM-Algorithmus

bei Schätz- und Klassifikationsproblemen

Diplomarbeit

zur
Erlangung des Grades
Diplom-Ingenieur
(Dipl.-Ing.)
der
Fachrichtung
Vermessungswesen

Vorgelegt am 13. November 2000 von
Marc Luxen
aus Aremberg

Betreuer:	Prof. Dr.-Ing. Wolfgang Förstner Dipl.-Ing. Ansgar Brunn
Ausgegeben am:	20. Oktober 2000
Vorgelegt am:	13. November 2000

Landwirtschaftliche Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität zu Bonn

Institut für Photogrammetrie

Diplomaufgabe

für Herrn cand. geod. Marc Luxen

Der EM-Algorithmus bei Schätz- und Klassifikationsproblemen

Eine zentrale Aufgabe der Photogrammetrie besteht heute in der Entwicklung von Verfahren zur automatischen Wiedererkennung von Objekten in digitalen Bildern. Im Zusammenhang hiermit treten häufig Schätz- und Klassifikationsprobleme auf, die mit traditionellen geodätischen Parameterschätzverfahren nicht oder nur mit verhältnismäßig großem Aufwand gelöst werden können. Beziehen sich beispielsweise die Beobachtungen, die zur Bestimmung von objektrelevanten Parametern aus einem Bild abgeleitet werden, auf unterschiedliche Objekte und geht aus dem Beobachtungsmaterial nicht hervor, welche Beobachtung zu welchem Objekt gehört, so muß zur Schätzung der Objektparameter eine Zuordnung der Beobachtungen zu den einzelnen Objekten erfolgen. Mit klassischen Methoden der Parameterschätzung ist es nicht ohne weiteres möglich, dieses Zuordnungsproblem *gemeinsam* mit der Parameterschätzung zu lösen.

Im Jahre 1977 wurde in einem von der Royal Statistical Society veröffentlichten Aufsatz der amerikanischen Professoren A.P. Dempster, N.M. Laird und D.B. Rubin[1] der sogenannte *Expectation Maximization (EM) Algorithmus* vorgestellt. Hierbei handelt es sich um ein Verfahren zur Schätzung von Parametern aus unvollständigen Beobachtungsdaten, das u.a. die Lösung des Zuordnungsproblems gemeinsam mit der Parameterschätzung erlaubt und daher für die automatische Bildinterpretation von großem Interesse ist.

Aufgabe des Diplomanden ist es, die theoretischen Grundlagen des EM-Algorithmus aufzuarbeiten und anhand eines frei wählbaren Beispiels aufzuzeigen, wie der EM-Algorithmus eingesetzt werden kann, wenn im Zusammenhang mit einer Parameterschätzung gleichzeitig Zuordnungsprobleme zu lösen sind.

(Prof. Dr.-Ing. W. Förstner)

Betreuer: W. Förstner, A. Brunn
Ausgegeben am: 20. Oktober 2000

Abgabetermin: 20. Januar 2001
Abgegeben am: 13. November 2000

[1] A.P. Dempster, N.M. Laird, D.B. Rubin (1977): *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society, Series B (Methodological), Bd. 39, Nr.1, 1977, S. 1-38

E r k l ä r u n g :

Hiermit versichere ich, die vorliegende Arbeit selbständig und nur unter Nutzung der im Literaturverzeichnis aufgeführten Quellen angefertigt zu haben.

Aremberg, 13. November 2000

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die vorliegende Diplomarbeit beschäftigt sich mit dem sogenannten *Expectation Maximization (EM)-Algorithmus*, einem Verfahren zur Schätzung unbekannter Parameter aus unvollständigen Beobachtungsdaten. Die Arbeit stellt eine Zusammenstellung der Ergebnisse verschiedener Veröffentlichungen zu diesem Thema dar und zeigt darüber hinaus anhand eines einfachen Beispiels auf, wie der EM-Algorithmus zur Lösung von Schätz- und Klassifikationsproblemen eingesetzt werden kann.

Angesichts der Tatsache, daß Schätzverfahren primär im Rahmen des Fachbereiches Ausgleichungsrechnung und Statistik behandelt werden, stellt sich die Frage, warum die Auseinandersetzung mit einem Verfahren zur Parameterschätzung aus unvollständigen Beobachtungsdaten in einer Diplomarbeit aus dem Bereich der *Photogrammetrie* erfolgt. In dieser Einleitung wird daher die Motivation zur vorliegenden Arbeit aus der Sicht der Photogrammetrie erläutert. Hierzu wird zunächst in Abschnitt 11 in allgemeiner Weise die enge Verknüpfung der Photogrammetrie mit der Ausgleichungsrechnung und Statistik beschrieben. Anschließend werden kurz einige für diese Arbeit wesentliche Aspekte der klassischen Methoden der geodätischen Ausgleichungsrechnung erläutert und es werden anhand eines Beispiels Grenzen dieser klassischen Methoden aufgezeigt. Abschließend erfolgt eine Abgrenzung der Parameterschätzung aus unvollständigen Daten von den klassischen Verfahren und es wird anhand einer konkreten Aufgabe skizziert, inwiefern Verfahren zur Parameterschätzung aus unvollständigen Beobachtungsdaten im Rahmen der Photogrammetrie eingesetzt werden können.

11 Statistische Methoden in der Photogrammetrie

Im Zusammenhang mit photogrammetrischen Auswertungen spielen statistische Methoden und Parameterschätzverfahren eine wesentliche Rolle. Dies betrifft sowohl die klassische Photogrammetrie als auch moderne Verfahren der digitalen Bildverarbeitung. Im Hinblick auf das Ziel, bei photogrammetrischen Auswertungen Resultate zu erzielen, die von zufälligen Meßunsicherheiten und groben Fehlern möglichst wenig beeinträchtigt werden, werden die Zielgrößen einer Auswertung in der Regel mittels mehr oder weniger robusten[1] Schätzverfahren aus z.T. hochredundanten Beobachtungen bestimmt.

Im Bereich der *klassischen Photogrammetrie* sind in diesem Zusammenhang beispielhaft

[1] Unter der Robustheit eines Schätzverfahrens wird hier die Unanfälligkeit der Schätzung gegenüber Ausreißern in den Beobachtungen verstanden

die bekannten Orientierungs- und Kalibrierungsverfahren[2] zu nennen, also

- der Räumlicher Rückwärtsschnitt (RRS),
- der Räumlicher Vorwärtsschnitt (RVS),
- die Relative Orientierung,
- die Absolute Orientierung,
- die Bündelblockausgleichung im Rahmen einer Aerotriangulation und
- die Testfeldkalibrierung.

Bei jedem dieser Verfahren werden Parameterschätzungen durchgeführt, um jeweils eine Menge $\mathcal{P}$ unbekannter Parameter[3] mittels einer Menge $\mathcal{V}$ gemessener Größen unter Berücksichtigung der Meßunsicherheiten zu bestimmen. Zwar existieren zu einigen der Verfahren auch direkte Lösungen oder zumindest direkte Näherungslösungen, hierbei handelt es sich jedoch meist um Spezialfälle hinsichtlich der Anzahl der verarbeiteten Bilder oder Beobachtungen. Daher werden die direkten Lösungen in der überwiegenden Zahl der Anwendungen allenfalls zur Bestimmung von Näherungswerten für die unbekannten Parameter verwendet, um die endgültigen Schätzwerte in einer Ausgleichung zu ermitteln.

Im Rahmen der *digitalen Bildverarbeitung* auf Grauwertbildern finden statistische Methoden ebenfalls umfangreiche Anwendung. Hervorzuheben sind in diesem Zusammenhang insbesondere Verfahren zur *automatischen Bildinterpretation*, also Verfahren zur automatischen (Wieder-) Erkennung von Objekten in Grauwertbildern. Bei diesen Verfahren wird vor allem auf statistische Klassifikationsverfahren zurückgegriffen, die i.d.R. entsprechend dem sogenannten *dreistufigen Konzept der Bildinterpretation* auf unterschiedlichen Abstraktionsebenen zur Anwendung kommen (vgl.[Fuchs 1998]). Nach einer Vorverarbeitung des Bildes auf der *unteren Ebene* der Bildinterpretation[4] wird bei der sogenannten *Merkmalsextraktion* auf der *mittleren Ebene der Bildinterpretation* jedes Bildpixel mittels statistischer Klassifikationsverfahren nach seiner Zugehörigkeit zu einer Punktstruktur, einer linienhaften Struktur oder einer homogenen Fläche im Bild klassifiziert. Im Hinblick auf diese Klassifikation werden die Eigenschaften der einzelnen Strukturelemente statistisch modelliert, um für jedes Pixel mittels Hypothesentests eine Entscheidung bezüglich der Zugehörigkeit zur Merkmalsklasse *„Punkt"*, *„Linie"* oder „Fläche" treffen zu können (vgl. [Fuchs 1998]). Im Anschluß daran wird auf der *oberen Ebene der Bildinterpretation* den einzelnen Strukturelementen, bzw. Gruppen von Strukturelementen jeweils die Bezeichnung einer Objektklasse zugeordnet, wobei wieder statistische Klassifikationsverfaren zum E¡insatz kommen. Aus einer Menge von im Objektraum möglicherweise vorkommender Objekte wird also das Objekt bestimmt, das im Bild mit größter Wahrscheinlichkeit vorkommt. Im Ergebnis werden geschätzte Bezeichnungen der im Bild dargestellten Objekte erhalten(vgl.[Hornegger 1996]).

Zusammenfassend kann man sagen, daß es aufgrund der hier angedeuteten Vielzahl von Anwendungen statistischer Methoden in der Photogrammetrie durchaus gerechtfertigt ist, sich im Rahmen der Photogrammetrie mit den theoretischen Grundlagen der Ausgleichungsrechnung und Statistik zu beschäftigen, zumal die klassischen Methoden der geodätischen Ausgleichungsrechnung bei bestimmten Aufgaben aus der Photogrammetrie

[2] Zur detaillierten Erläuterung der Verfahren wird auf die Standardliteratur der Photogrammetrie (z.B. [Kraus 1994]) verwiesen

[3] Die Menge $\mathcal{P}$ der unbekannten Parameter umfasst je nach Aufgabe Parameter der Inneren und Äußeren Orientierung der verwendeten Kameras sowie die Koordinaten von Neupunkten.

[4] Die Vorverarbeitung kann u.a. aus einer geometrischen Transformation (etwa einer Entzerrung), einer radiometrischen Korrektur (etwa einer Kontrastanpassung) und-/oder einer Glättung bestehen.

nicht direkt angewendet weren können, so daß sie an die konkrete Aufgabe angepasst, ergänzt oder erweitert werden müssen.

Im folgenden Abschnitt werden kurz die herkömmlichen Methoden der klassischen geodätischen Ausgleichungsrechnung beschrieben und es wird anhand eines Beispiels gezeigt, daß die klassische Ausgleichungsrechnung im Zusammenhang mit bestimmten Schätz- und Klassifikationsproblemen an ihre Grenzen stößt. Da insbesondere die Beschreibung der Schätzverfahren sehr allgemein gehalten ist, sei an dieser Stelle im Hinblick auf eine detaillierte Beschreibung auf [KOCH 1998] verwiesen.

12 Klassische Schätzverfahren und deren Grenzen

121 Prinzip der klassischen geodätischen Ausgleichung

Den klassischen geodätischen Parameterschätzverfahren ist gemeinsam, daß eine Menge $\mathcal{P}$ unbekannter Parameter aufgrund einer Menge $\mathcal{V}$ gemessener Größen (Beobachtungen) bestimmt wird, die mit den unbekannten Parametern in einem bestimmten Zusammenhang stehen. Der Zusammenhang zwischen den Beobachtungen und den unbekannten Parametern geht dabei aus einem mathematischen Modell hervor, von dem vorausgesetzt wird, das es aufgrund gewisser Gesetzmäßigkeiten, denen die Beobachtungen in Abhängigkeit von den unbekannten Parametern genügen, a priori bekannt ist. Es wird also für jede Beobachtung vorausgesetzt, daß von vorneherein bekannt, in welchem mathematischen Zusammenhang sie mit den unbekannten Parametern steht. Die Gleichungen, die den Zusammenhang zwischen den Beobachtungen und den unbekannten Parametern beschreiben, werden *Beobachtungsgleichungen* genannt. Die Beobachtungen werden aufgrund der Meßunsicherheiten als stochastische Variable aufgefasst und im Rahmen der Ausgleichung gemäß ihrer Präzision gewichtet.

Im Hinblick auf die Abgrenzung der üblichen Schätzverfahren von den Schätzverfahren aus unvollständigen Daten kommt folgender Aussage eine wesentliche Bedeutung zu:

Bei den Standardschätzverfahren aus vollständigen Beobachtungsdaten handelt es sich bei allen im Modell eingeführten Observablen um tatsächlich zu messende *bzw.* tatsächlich gemessene *Größen.*

Üblicherweise wird ein linearer Zusammenhang zwischen den Beobachtungen und den Parametern hergestellt, was auf Parameterschätzungen in linearen Modellen führt[5]. Das bekannte *Gauß-Markoff-Modell* stellt ein solches lineares Modell dar, in dem die unbekannten Parameter mittels unterschiedlicher Schätzmethoden ermittelt werden können. Meist wird hier die beste lineare erwartungstreue Schätzung der unbekannten Parameter bestimmt. Im Gauß-Markoff-Modell stimmt diese Schätzung mit der Schätzung nach der Methode der kleinsten Quadrate und der Maximum-Likelihood-Methode im Falle normalverteilter Beobachtungen überein (vgl. [KOCH 1998]).

[5] ggf. folgt der lineare Zusammenhang aus einem nichtlinearen Zusammenhang durch Taylor-Entwicklung mit Näherungswerten für die unbekannten Parameter und Abbruch nach dem linearen Term

122 Grenzen der klassischen geodätischen Ausgleichungsrechnung

Die klassische geodätische Ausgleichungsrechnung stößt u.a dann an ihre Grenzen, wenn der Zusammenhang zwischen den Beobachtungen und den unbekannten Parametern nicht eindeutig angegeben werden kann, d.h. wenn das der Ausgleichung zugrundeliegende Beobachtungsmodell a priori nicht eindeutig festliegt.

Um dies zu verdeutlichen, sei das *Beispiel* betrachtet, bei dem die Abszissen u_i und Ordinaten y_i von Punkten zweier in einem ebenen uy-Koordinatensystem dargestellten Geraden $\mathcal{K}_1$ und $\mathcal{K}_2$ gemessen worden seien, um hieraus die Parameter *Steigung* und *Nullablage* der beiden Geraden zu bestimmen (vgl. Abbildung 122). Dabei sei es versäumt worden, festzuhalten, welche Beobachtungen zu welcher der beiden Geraden gehören, so daß diese Informationen für die Ausgleichung fehlen. Offensichtlich ist es nicht möglich, aus den verbleibenden Informationen die

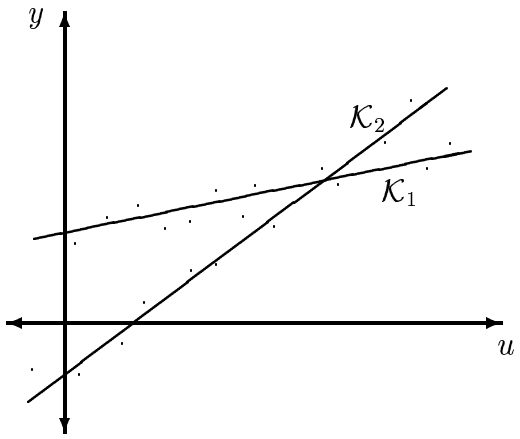


Abbildung 1:

Parameter beider Geraden mittels herkömmlicher geodätischer Schätzmethoden zu bestimmen, da (zumindest ohne Visualisierung der Daten) keine der Beobachtungen eindeutig einer der beiden Geraden zugeordnet werden kann. Es ist zwar bekannt, daß jede Beobachtung entweder zur Gerade $\mathcal{K}_1$ oder zur Gerade $\mathcal{K}_2$ gehört, aber nicht zu welcher von beiden. Somit ist der Zusammenhang zwischen den unbekannten Parametern der Ausgleichung und den Beobachtungen a priori nicht eindeutig angebbar und damit eine klassische Ausgleichungsrechnung nicht durchführbar.

Zur Lösung des Problems bedarf es der Anwendung eines Verfahrens zur sogenannten *Parameterschätzung aus unvollständigen Beobachtungsdaten*, wie sie im nächsten Abschnitt beschrieben ist. Der *Expectation-Maximization Algorithmus* stellt ein solches Verfahren dar; mit ihm ist es möglich, zum einen die einzelenen Beobachtungen hinsichtlich ihrer Zugehörigkeit zu einer der beiden Geraden zu klassifizieren und dabei gleichzeitig Schätzwerte für die Geradenparameter zu bestimmen (vgl. hierzu Kapitel 4)

Dem gegebenen Beispiel scheint zunächst der Bezug zur Photogrammetrie zu fehlen; es lässt sich allerdings auf Aufgaben der digitalen Bildverarbeitung verallgemeinern, bei denen bestimmte Kenngrößen (d.h. unbekannte Parameter) mehrerer im Bild erscheinender Objekte vollautomatisch bestimmt werden sollen, wobei die Zusammenhänge zwischen den hierzu herangezogenen Beobachtungen (z.B. den Ergebnissen einer Merkmalsextraktion) und den gesuchten Parametern bis auf die Zuordnung der Beobachtungen zu den in der Bildszene dargestellen Objekten bekannt ist. Da sich -wie beschrieben- solche Aufgaben mit den herkömmlichen Methoden der geodätischen Ausgleichungsrechnung nicht

ohne weiteres lösen lassen, braucht man Verfahren zur Parameterschätzung aus unvollständigen Beobachtungsdaten. Das Prinzip der Parameterschätzung aus unvollständigen Daten soll daher im nächsten Abschnitt näher beschrieben werden.

13 Parameterschätzung aus unvollständigen Daten

Im Folgenden wird der Unterschied zwischen Parameterschätzungen aus unvollständigen Beobachtungsdaten und den herkömmlichen Schätzverfahren erläutert.

Abgrenzung der Schätzung aus unvollständigen Beobachtungsdaten

Wie in Abschnitt 12 erläutert, wird bei den klassischen geodätischen Parameterschätzverfahren vorausgesetzt, daß als Beobachtungen nur *tatsächlich gemessene* Werte verwendet werden. Es kann u.U. allerdings auch von Vorteil sein, in die Ausgleichung sogenannte *„fehlende Beobachtungen"* einzuführen. Fehlende Beobachtungen sind Zufallsparameter[6] , die wie *Beobachtungen* in eine Ausgleichung eingeführt werden, die aber tatsächlich *nicht* beobachtet worden sind. Hierbei kann es sich um Größen handeln, die entweder prinzipiell durch Messungen nicht zugänglich sind oder um meßbare Größen, die im konkreten Fall aus bestimmten Gründen nicht gemessen worden sind. Bei fehlenden Beobachtung handelt es sich also um Variablen einer Parameterschätzung, die aufgrund einer *Überparametrisierung des Modells* entstehen, also durch Einführung zusätzlicher Parameter in eine Ausgleichung, die durch bestehende Gesetzmäßigkeiten zwischen den tatsächlichen Beobachtungen und den eigentlichen Modellparametern nicht erklärbar sind.

Eine solche Überparametrisierung erscheint zunächst nicht plausibel, da es als wenig sinnvoll erscheint, nicht beobachtete Größen trotzdem formal als Beobachtungen in die Ausgleichung einzubringen. Bei näherer Betrachtung sieht man jedoch ein, daß die Einführung fehlender Beobachtungen insbesondere dann Sinn macht, wenn bestimmte Größen des mathematischen Modells einer Ausgleichung zwar nicht direkt meßbar sind, aber bekannt ist, daß zwischen ihnen und den durch Messungen zugänglichen Beobachtungen Korrelationen bestehen und wie groß diese sind. Würden diese unzugänglichen Beobachtungen nicht mit in die Ausgleichung einbezogen, so bliebe auch das Wissen um diese Wechselbeziehungen ungenutzt. Werden im Gegensatz dazu durch eine Überparametrisierung des Schätzproblems Zufallsparameter für die nich zugänglichen Beobachtungen eingeführt, so kann die Kenntnis um die Korrelationen zwischen den versteckten und tatsächlichen Beobachtungen in der Ausgleichung verwertet werden.

Ein anschauliches, aus dem Bereich der Photogrammetrie stammendes Beispiel für Parameterschätzungen mit fehlenden Beobachtungen stellt die Klassifikation dreidimensionaler Objekte aus (nur) zweidimensionalen Bildern im Rahmen einer Einzelbildauswertung dar (vgl. [Hornegger 1996]): Ein dreidimensionales Objekt des Objektraumes wird im Bild nur zweidimensional dargestellt, die Tiefeninformation geht also bei der Abbildung verloren. Trotz dieses Informationsverlustes bei der Abbildung soll aber das räumliche Objekt aus dem ebenen Bild erkannt werden. Es liegt also nahe, im Rahmen einer Überparametrisierung die fehlenden Tiefeninformationen als fehlende Beobachtungen mit in die Parameterschätzung für die Bestimmung der Objektbezeichnung einzuführen.

[6] Unter Zufallsparametern werden hier unbekannte Parameter einer Ausgleichung verstanden, die nicht als feste Größen, sondern als Zufallsvariablen aufgefasst werden.

Es stellt sich also die Frage nach Parameterschätzverfahren, die fehlende Beobachtungen zulassen, also auf unvollständigen Beobachtungsdaten arbeiten. Der in dieser Arbeit untersuchte EM-Algorithmus stellt ein solches Schätzverfahren dar. Neben diesem Algorithmus existieren weitere Verfahren, wie z.B. den *DA (Data Augmentation Algorithms)* oder den *PMDA (Poor Man's Data Augmentation Algorithm)*. In Bezug auf diese Verfahren sei hier auf die Literatur [TANNER 1993] verwiesen.

14 Zusammenfassung / Überblick über die Arbeit

In dieser Einleitung wurde anhand von einiger Beispiele aufgezeigt, daß in der Photogrammetrie Methoden der Ausgleichungsrechnung und Statistik eine sehr starke Rolle spielen, so daß es nahe liegt, sich auch im Rahmen der Photogrammetrie mit diesen Methoden auseinanderzusetzen. Außerdem wurden die herkömmlichen, auf vollständigen Beobachtungsdaten arbeitenden Verfahren der klassischen geodätischen Ausgleichungsrechnung kurz skizziert und anhand eines Beispiels deren Grenzen aufgezeigt. Es wurde angedeutet, das in bestimmten Fällen, in denen eine klassische Ausgleichungsrechnung nicht möglich ist, eine Parameterschätzung aus unvollständigen Beobachtungsdaten zum Ziel führen kann. Die in den Abschnitten 12) und 13 gegebenen Beispiele zeigen, daß die Parameterschätzung aus unvollständigen Daten für die Photogrammetrie durchaus relevant ist.

Im Folgenden wird der EM-Algorithmus als ein iteratives Verfahren zur Schätzung von Parametern aus unvollständigen Daten vorgestellt. Nach einer allgemeinen Definition dieses Algorithmus erfolgt eine ausführliche Auseinandersetzung mit seinen Eigenschaften. Hier werden insbesondere die Bedingungen betrachtet, unter denen die EM-Iterationen konvergieren, und die Geschwindigkeit des Algorithmus. Anschließend wird der EM-Algorithmus zur Lösung einer konkreten Aufgabe in einem Programm realisiert. Den Abschluss der Arbeit bildet eine Zusammenfassung.

Kapitel 2

Der EM-Algorithmus

Der in dieser Arbeit diskutierte Ansatz zur Parameterschätzung aus unvollständigen Daten ist der EM-Algorithmus. Er soll in diesem Abschnitt in seiner allgemeinen Form definiert werden. Im Hinblick auf die Notation sind hierzu einige Vorbemerkungen notwendig.

21 Begriffe und Notation

In der Einleitung wurde im Zusammenhang mit Parameterschätzverfahren abstrakt von einer Menge $\mathcal{P}$ unbekannter Parameter gesprochen, die mit Hilfe einer Menge $\mathcal{V}$ von Beobachtungen bestimmt werden. Im Folgenden werden nun die u unbekannten Parameter $\beta_1, \beta_2, \ldots, \beta_u$ einer Parameterschätzung in dem $u \times 1$ Vektor

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1, & \beta_2, & \ldots & \beta_u \end{bmatrix}^T \tag{210.1}$$

zusammengefasst. Die unbekannten Parameter werden zunächst als fest vorausgesetzt, sie stellen also keine Zufallsvariablen dar. Der *Parameterraum* $\mathcal{B}$ ist eine Untermenge des u-dimensionalen euklidschen Vektorraumes $\mathbb{R}^u$ und umfasst alle im Rahmen des Schätzproblems theoretisch möglichen numerischen Realisierungen des Vektors $\boldsymbol{\beta}$. In dem $n_y \times 1$ Zufallsvektor

$$\boldsymbol{y} = \begin{bmatrix} y_1, & y_2, & \ldots & y_{n_y} \end{bmatrix}^T \tag{210.2}$$

sind die n_y *tatsächlichen* Beobachtungen, also die durch Messungen zugänglichen Beobachtungen $y_1, y_2, \ldots y_n$ enthalten. Die n_x durch Messungen nicht zugänglichen Beobachtungen $x_1, x_2, \ldots x_{n_x}$ werden in dem $n_x \times 1$ Zufallsvektor

$$\boldsymbol{x} = \begin{bmatrix} x_1, & x_2, & \ldots & x_{n_x} \end{bmatrix}^T \tag{210.3}$$

zusammengefasst. Die Mengen aller numerischen Realisierungen von $\boldsymbol{x}$ bzw. $\boldsymbol{y}$ bilden die *Stichprobenräume* $\mathcal{X}$ bzw. $\mathcal{Y}$.

Der $n_z \times 1$ Zufallsvektor

$$\boldsymbol{z} = \begin{bmatrix} z_1, & z_2, & \ldots & z_{n_z} \end{bmatrix}^T \tag{210.4}$$

ist der sogenannte *vollständige* Beobachtungsvektor. Hierbei handelt es sich *nicht* zwangsläufig um die einfache Zusammenfassung der Vektoren $\boldsymbol{x}$ und $\boldsymbol{y}$ in einem gemeinsamen

Vektor. Der Vektor $\boldsymbol{z}$ umfaßt zwar stets alle fehlenden Beobachtungen aus $\boldsymbol{x}$, muß aber im allgemeinen nicht unbedingt auch tatsächliche Beobachtungen aus $\boldsymbol{y}$ enthalten. Anhand des folgenden einfachen Beispiels soll die Struktur des Vektors $\boldsymbol{z}$ verdeutlicht werden:

Betrachtet werde eine Parameterschätzung mit den (gemessenen) Beobachtungen $y_i, i \in \{1 \dots n_y\}$.

Fall 1: Es soll angenommen werden, daß die Beobachtung y_1 *der Summe zweier Beobachtungen* x_1 *und* x_2 *entspricht, die aber direkt nicht meßbar sind, also* $y_1 = x_1 + x_2$. *Der Vektor* $\boldsymbol{y}$ *der tatsächlichen Beobachtungen, der Vektor* $\boldsymbol{x}$ *der fehlenden Beobachtungen und der Vektor* $\boldsymbol{z}$ *der vollständigen Beobachtungen haben dann die Form*

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_{n_y} \end{bmatrix} \quad , \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \textit{und} \quad \boldsymbol{z} = \begin{bmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \\ \dots \\ y_{n_y} \end{bmatrix}$$

Der vollständige Beobachtungsvektor enthält also neben den fehlenden Beobachtungen alle tatsächlichen Beobachtungen.

Fall 2: Werden sämtliche tatsächlichen Beobachtungen als Summe zweier versteckter Beobachtungen betrachtet, also $y_i = x_{2i} + x_{2i-1}$, *so ergibt sich*

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_{n_y} \end{bmatrix} \quad , \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_{2n_y} \end{bmatrix} \quad \textit{und} \quad \boldsymbol{z} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_{2n_y} \end{bmatrix} = \boldsymbol{x},$$

der vollständige Beobachtungsvektor enthält also außer den versteckten Beobachtungen keine weiteren Elemente.

Die Gesamtheit aller numerischer Realisierungen des Vektors $\boldsymbol{z}$ bildet den Stichprobenraum $\mathcal{Z}$ der vollständigen Beobachtungen.

Allgemein kann davon ausgegangen werden,daß der tatsächliche Beobachtungsvektor $\boldsymbol{y}$ und der vollständige Beobachtungsvektor $\boldsymbol{z}$ in einem mathematischen Zusammenhang stehen, da beide funktional von den selben unbekannten Parametern abhängen. Es wird daher vorausgesetzt, daß der Stichprobenraum $\mathcal{Y}$ durch eine Abbildung

$$B : \mathcal{Z} \to \mathcal{Y}$$

aus dem Stichprobenraum $\mathcal{Z}$ hervorgeht, so daß sich $\boldsymbol{y}$ durch Abbildung

$$B : \boldsymbol{z} \mapsto \boldsymbol{y} = B(\boldsymbol{z}) \tag{210.5}$$

aus $\boldsymbol{z}$ ergibt[1].

Im Fall 1 des oben gegebenen Beispiels ist die Abbildung B *beispielsweise mit*

$$B : \boldsymbol{z} \mapsto \boldsymbol{y} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I} \end{bmatrix} \cdot \boldsymbol{z}$$

[1] Man kann sich vorstellen, daß sich die unbekannten Parameter mit Hilfe der Elementen des Vektors $\boldsymbol{z}$ bestimmen ließen, wenn $\boldsymbol{z}$ vollständig beobachtet werden könnte. Die so ermittelten Parameter bestimmen unmittelbar die Elemente des tatsächlichen Beobachtungsvektors $\boldsymbol{y}$. Somit kann $\boldsymbol{y}$ als Abbild von $\boldsymbol{z}$ verstanden werden.

gegeben, worin $\mathbf{0}$ *eine Nullmatrix der Dimension* $2 \times n_y$ *und* $\boldsymbol{I}$ *eine* $n_y \times n_y$ *Einheitsmatrix bedeuten. Im Fall 2 dieses Beispiels gilt*

$$B : \boldsymbol{z} \mapsto \boldsymbol{y} = \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \end{bmatrix} \cdot \boldsymbol{z}.$$

Die Abbildung B ist i.a. nicht eindeutig, d.h. die Gleichung $\boldsymbol{y} = B(\boldsymbol{z})$ ist bei gegebenem $\boldsymbol{z}$ in der Regel für mehrere Vektoren $\boldsymbol{z}$ erfüllt. Daher wird die Menge

$$\mathcal{Z}(\boldsymbol{y}) = \{\boldsymbol{z} \mid \boldsymbol{y} = B(\boldsymbol{z})\} \tag{210.6}$$

eingeführt. Sie besteht aus allen Vektoren $\boldsymbol{z}$, die für ein gegebenes $\boldsymbol{y}$ die Gleichung $\boldsymbol{y} = B(\boldsymbol{z})$ erfüllen.

In Fall 1 des Beispiels besteht $\mathcal{Z}(\boldsymbol{y})$ *beispielsweise aus allen Vektoren* $\boldsymbol{z}$*, die die Bedingung* $z_2 = y_1 - z_1$ *erfüllen.*

Nachdem nun die in der Diplomarbeit verwendeten Bezeichnungen und die Notation erläutert wurden, wird kurz die Maximum-Likelihood-Methode beschrieben. Einerseits kann die Maximum-Likelihood-Methode für sich allein bereits zur Schätzung von Parametern aus unvollständigen Beobachtungen verwendet werden, andererseits stellt sie den Ausgangspunkt für den EM-Algorithmus dar.

22 Maximum-Likelihood-Methode

Im Zusammenhang mit dem Expectation-Maximization-Algorithmus spielt der Maximum-Likelihood-Schätzer eine wesentliche Rolle. So wird z.B. innerhalb der einzelnen Iterationsschritte der Erwartungswert der logarithmierten Likelihoodfunktion berechnet und maximiert (vgl. Abschnitt 232). Dies führt im Optimalfall auf eine Maximum-Likelihood-Schätzung der unbekannten Parameter. Wie in Abschnitt 222 gezeigt wird, eignet sich außerdem die Maximum-Likelihood-Methode als solche bereits zur Parameterschätzung aus unvollständigen Daten, wenn als Dichtefunktion eine von den tatsächlichen Beobachtungen $\boldsymbol{y}$ abhängige Randdichte verwendet wird. Daher wird die Maximum-Likelihood-Schätzung in diesem Abschnitt kurz erläutert. Im nächsten Kapitel wird dann der EM-Algorithmus definiert.

221 Definition

Die Maximum-Likelihood-Methode zählt zu den klassischen Methoden der Parameterschätzung. Im Unterschied zu anderen herkömmlichen Verfahren, wie z.B. der besten linearen erwartungstreuen Schätzung oder der Methode der kleinsten Quadrate kann eine Parameterschätzung nach der Maximum-Likelihood-Methode nur durchgeführt werden, wenn die parametrische Dichtefunktion der Beobachtungen bekannt ist. Liegen normalverteilte[2] Beobachtungen vor, so führt die Maximum-Likelihood-Schätzung im Gauß-

[2]Aufgrund der Gültigkeit des Zentralen Grenzwertsatzes der mathematischen Statistik kann in einer Vielzahl von Anwendungen von normalverteilten Beobachtungen ausgegangen werden.

Markoff-Modell auf die gleichen Schätzwerte wie die Methode der kleinsten Quadrate und die beste lineare erwartungstreue Schätzung.(vgl. [KOCH 1998])

Die Definition der Maximum-Likelihood-Schätzung vorbereitend wird zunächst die sogenannte *Likelihoodfunktion* definiert[3].

Definition: [LIKELIHOODFUNKTION]
Der Zufallsvektor $\boldsymbol{z}$ der Beobachtungen besitze die von den unbekannten, festen Parametern abhängige Dichte $f(\boldsymbol{\beta})$, dann ist die Likelihoodfunktion $l(\boldsymbol{z} \mid \boldsymbol{\beta})$ *definiert durch*

$$l(\boldsymbol{z} \mid \boldsymbol{\beta}) = f(\boldsymbol{\beta}) \tag{221.1}$$

Die Bestimmung der Schätzwerte für die unbekannten Parameter erfolgt durch Maximierung der Likelihoodfunktion. D.h. es wird der Parametervektor $\boldsymbol{\beta}$ bestimmt, für den die Dichtefunktion $l(\boldsymbol{z} \mid \boldsymbol{\beta})$ der Beobachtungen $\boldsymbol{z}$ maximal wird. Dies folgt aus der Definition

Definition: [MAXIMUM-LIKELIHOOD-METHODE]
Der Zufallsvektor $\boldsymbol{z}$ der Beobachtungen besitze die Likelihoodfunktion $l(\boldsymbol{z} \mid \boldsymbol{\beta})$, dann bezeichnet man als Maximum-Likelihood-Methode *die Schätzung $\bar{\boldsymbol{\beta}}$ der festen Parameter $\boldsymbol{\beta}$, die maximale Werte für $l(\boldsymbol{z} \mid \boldsymbol{\beta})$ liefert, also*

$$\bar{\boldsymbol{\beta}} = \arg\sup_{\boldsymbol{\beta}} l(\boldsymbol{z} \mid \boldsymbol{\beta}) \tag{221.2}$$

Durch Variation von $\boldsymbol{\beta}$ ergibt sich also als Schätzung $\bar{\boldsymbol{\beta}}$ der Vektor, der zusammen mit dem Beobachtungsvektor $\boldsymbol{z}$ die maximale Dichte erzielt. Umgangssprachlich kann man sagen, es werden als Schätzwerte die Parameter gewählt, die gemäß der Dichtefunktion am besten zu den Beobachtungen passen.

Anstatt die Likelihoodfunktion $l(\boldsymbol{z} \mid \boldsymbol{\beta})$ selbst zu maximieren wird meist die logarithmierte Likelihoodfunktion

$$L(\boldsymbol{\beta}) := L(\boldsymbol{z} \mid \boldsymbol{\beta}) := \log l(\boldsymbol{z} \mid \boldsymbol{\beta}) \tag{221.3}$$

maximiert. Dies vereinfacht im allgemeinen die Berechnungen und führt im Zusammenhang mit der EDV auch auf numerisch sicherere Ergebnisse. Aufgrund der strengen Monotonie und damit der Umkehrbarkeit der Logarithmusfunktion ist die Maximierung der logarithmierten Likelihoodfunktion gleichbedeutend mit der Maximierung der Likelihoodfunktion selbst. In diesem Zusammenhang ist auch der aus der Informationstheorie stammende Begriff der *Information* von Interesse. Unter der Information $I(\boldsymbol{z})$, die mit der Realisierung $\boldsymbol{z}$ eines Zufallsvektors verbunden ist, wird der ,,Grad der Überraschung" verstanden, die das Auftreten des Vektors $\boldsymbol{z}$ auslöst (vgl. [FÖRSTNER 1991]). Mathematisch ist die Information $I(\boldsymbol{z})$ eines Zufallsvektors $\boldsymbol{z}$ definiert als negativer Logarithmus seiner Dichte (vgl. [FÖRSTNER 1991]):

$$I(\boldsymbol{z}) = \log \frac{1}{f(\boldsymbol{z})} = -\log f(\boldsymbol{z}). \tag{221.4}$$

Entsprechend gilt

$$I(\boldsymbol{z} \mid \boldsymbol{\beta}) = -\log l(\boldsymbol{z} \mid \boldsymbol{\beta}) = -L(\boldsymbol{y} \mid \boldsymbol{\beta}) \tag{221.5}$$

Somit entspricht die Maximierung der logarithmierten Likelihoodfunktion der Minimierung der zum Vektor $\boldsymbol{z}$ gehörenden Information. Für die geschätzten Parameter wird also der Grad der Überraschung im Hinblick auf die Beobachtungen $\boldsymbol{z}$ minimal.

[3]Die Definitionen dieses Abschnittes entsprechen denen in in [KOCH 1998]

222 Schätzung aus unvollständigen Beobachtungsdaten mittels Maximum-Likelihood-Methode

Wie zu Beginn des Abschnitts 22 angedeutet, kann die Maximum-Likelihood-Methode für sich bereits zur Schätzung unbekannter Parameter aus unvollständigen Beobachtungsdaten verwendet werden. Die Vorgehensweise hierzu soll im folgenden erläutert werden:

Gemäß der Definition (221.2) ist zur Bestimmung der unbekannten Parameter $\boldsymbol{\beta}$ mittels der Maximum-Likelihood-Methode die Likelihoodfunktion $L(\boldsymbol{z} \mid \boldsymbol{\beta})$ zu maximieren. Im Falle eines Schätzproblems aus unvollständigen Daten steht man nun vor dem Problem, daß $\boldsymbol{z}$ fehlende, also tatsächlich nicht gemessene Beobachtungen enthält. Somit kann, trotzdem die Dichtefunktion $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ der vollständigen Beobachtungen[4] in Abhängigkeit von den unbekannten Parametern als bekannt gilt, die Likelihoodfunktion $L(\boldsymbol{z} \mid \boldsymbol{\beta})$ nicht unmittelbar ausgewertet und maximiert werden.

Der Weg zur Berechnung der Schätzwerte führt über die tatsächlichen Beobachtungen $\boldsymbol{y}$, die aus den vollständigen Beobachtungen durch die Abbildung $\boldsymbol{y} = B(\boldsymbol{z})$ hervorgehen: Die von den unbekannten Parametern abhängige Dichtefunktion $g(\boldsymbol{y} \mid \boldsymbol{\beta})$ der tatsächlichen Beobachtungen $\boldsymbol{y}$ ergibt sich aus der (bekannten) Dichtefunktion $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ der vollständigen Beobachtungen als verallgemeinerte[5] Randdichte (vgl. [DEMPSTER ET AL. 1968]) zu

$$g(\boldsymbol{y} \mid \boldsymbol{\beta}) = \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int f(\boldsymbol{z} \mid \boldsymbol{\beta}) \, d\boldsymbol{z}. \tag{222.1}$$

Der Integrationsbereich ist hierbei die Menge $\mathcal{Z}(\boldsymbol{y})$, es wird also also über sämtliche Vektoren $\boldsymbol{z}$ integriert, die durch die Abbildung (210.5) auf $\boldsymbol{y}$ abgebildet werden.

Die Randdichte $g(\boldsymbol{y} \mid \boldsymbol{\beta})$ aus (222.1) kann nun als Likelihoodfunktion verwendet werden, so daß die unbekannten Parameter aufgrund der tatsächlichen Beobachtungen $\boldsymbol{y}$ nach der Maximum-Likelihood-Methode bestimmt werden können.

Auf diese Weise kann prinzipiell mittels Maximum-Likelihood-Schätzung ein Schätzproblem mit unvollständigen Beobachtungsdaten gelöst werden. Allerdings werden in der Regel andere Verfahren zur Parameterschätzung aus unvollständigen Beobachtungsdaten vorgezogen, was im allgemeinen 2 Gründe hat: Zum einen ist die Bestimmung der Randdichte $g(\boldsymbol{y} \mid \boldsymbol{\beta})$ mittels numerischer Integration sehr *rechenintensiv*. Zum anderen bleibt bei der beschriebenen Methode das *Wissen um die Spezialität der Dichtefunktion* $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ *ungenutzt*. Im allgemeinen wird eine ganze Menge $\mathcal{F}$ von Dichtefunktionen $f^g(\boldsymbol{z} \mid \boldsymbol{\beta})$ existieren, die durch Integration über $\mathcal{Z}(\boldsymbol{y})$ zu der gleichen Randdichte führen wie $f(\boldsymbol{z} \mid \boldsymbol{\beta})$. Das vorliegende Wissen darüber, um welche dieser Dichtefunktionen es sich im konkreten Fall handelt, bleibt also bei der vorgestellten Methode unberücksichtigt. Es liegt hingegen die Vermutung nahe, daß sich dieses Wissen in der Parameterschätzung vorteilhaft einbringen lässt. Dies leistet der EM-Algorithmus. Er wird im folgenden nach einer Ableitung seiner Schlüsselgleichung allgemein definiert.

---

[4] Unter den vollständigen Beobachtungen werden im Rahmen dieser Arbeit die Elemente des vollständigen Beobachtungsvektors verstanden.

[5] Die in [KOCH 1998] auf Seite 98 angegebne Randverteilung stellt einen Spezialfall der Gleichung (222.1) dar. Sie ergibt sich aus (222.1), falls $\boldsymbol{z} = (x_1, \quad x_2, \quad \ldots \quad , x_n)$ und $\boldsymbol{y} = (x_{i+1}, \quad \ldots \quad , x_n)$ gilt. Daher wird (222.1) hier als verallgemeinerte Randdichte bezeichnet.

23 Definition des (G)EM-Algorithmus

In diesem Abschnitt wird der *Expectation-Maximization-Algorithmus* (EM-Algorithmus) bzw. der allgemeinere *Generalized Expectation-Maximization-Algorithmus* (GEM-Algorithmus) allgemein definiert. Zunächst wird jedoch eine Gleichung abgeleitet, die als Schlüsselgleichung des (G)EM-Algorithmus[6] angesehen werden kann, da sie maßgeblich zu dessen Verständnis beiträgt.

231 Schlüsselgleichung des (G)EM-Algorithmus

Zur Ableitung der Schlüsselgleichung wird zunächst die bedingte Dichte $f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta})$ des Vektors $\boldsymbol{z}$ der vollständigen Beobachtungen eingeführt für den Fall, daß $\boldsymbol{y}$ und $\boldsymbol{\beta}$ gegeben sind. Nach [Dempster et al. 1968], Gl. (2.5) gilt

$$f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}) = \frac{f(\boldsymbol{z} \mid \boldsymbol{\beta})}{g(\boldsymbol{y} \mid \boldsymbol{\beta})}. \tag{231.1}$$

Somit folgt für die in Abschnitt 222 benutzte Likelihoodfunktion

$$g(\boldsymbol{y} \mid \boldsymbol{\beta}) = \frac{f(\boldsymbol{z} \mid \boldsymbol{\beta})}{f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta})}$$

und für die hierzu gehörende logarithmierte Likelihoodfunktion

$$L(\boldsymbol{\beta}) := L(\boldsymbol{y} \mid \boldsymbol{\beta}) = \log g(\boldsymbol{y} \mid \boldsymbol{\beta}) = \log f(\boldsymbol{z} \mid \boldsymbol{\beta}) - \log f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}). \tag{231.2}$$

Ziel ist es nun, die logarithmierte Likelihoodfunktion (231.2) durch Variation von $\boldsymbol{\beta}$ zu maximieren, um eine Maximum-Likelihood-Schätzung der unbekannten Parameter $\boldsymbol{\beta}$ zu erhalten. Die Dichtefunktion $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ des vollständigen Beobachtungsvektors $\boldsymbol{z}$ wird dabei bis auf die unbekannten Parameter als bekannt vorausgesetzt. Da die logarithmierte Likelihoodfunktion unter Ausnutzung dieser Kenntnis erfolgen soll, wird auf die Berechnung der Likelihoodfunktion nach (222.1) verzichtet und stattdessen die rechten Seite von (231.2) maximiert. Eine Hürde stellt in diesem Zusammenhang die Tatsache dar, daß der vollständige Beobachtungsvektor $\boldsymbol{z}$ fehlende, also numerisch unbekannte Beobachtungen enthält, so daß weder $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ noch $f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta})$ in Abhängigkeit von $\boldsymbol{\beta}$ direkt ausgewertet werden können. Die Grundidee zur Überwindung dieser Hürde besteht darin, die fehlenden Beobachtungen aus $\boldsymbol{z}$ durch deren Erwartungswerte[7] zu ersetzen, um dann die rechte Seite von (231.2) durch Variation von $\boldsymbol{\beta}$ zu maximieren. Dies ist das Grundprinzip des EM-Algorithmus, der als iteratives Verfahren definiert ist (vgl. Abschnitt 232).

Zunächst wird vorausgesetzt, daß mit $\hat{\boldsymbol{\beta}}^{(i)}$ vorläufige Schätzwerte für die unbekannten Parameter vorliegen. Mit Hilfe der Gleichung

$$L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) = \log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) = \log f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) - \log f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)}) \tag{231.3}$$

sollen nun neue Schätzwerte $\hat{\boldsymbol{\beta}}^{(i+1)}$ für $\boldsymbol{\beta}$ berechnet werden, wozu die logarithmierte Likelihoodfunktion $L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)})$ als Zufallsvariable aufgefasst wird. Es kann der Erwartungswert dieser Zufallsvariable über $\mathcal{Z}(\boldsymbol{y})$ berechnet werden, wobei zur Berechnung der

[6] Soweit in dieser Arbeit der Ausdruck *(G)EM-Algorithmus* verwendet wird, beziehen sich die entsprechenden Aussagen sowohl auf den EM-Algorithmus als auch auf den GEM-Algorithmus

[7] Die Berechnung dieser Erwartungswerte erfolgt iterativ und stützt sich jeweils auf vorher ermittelte Schätzwerte für die unbekannten Parameter. Das Verfahren wird noch ausführlich erläutert.

benötigten Dichtefunktionen $f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta})$ die vorläufigen Schätzwerte $\hat{\boldsymbol{\beta}}^{(i)}$ herangezogen werden. Aufgrund der Linearität des Erwartungswertoperators erhält man zunächst

$$E[L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] = E[\log f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] - E[\log f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}]. \quad (231.4)$$

Hierin gilt für die linke Seite

$$\begin{aligned} E[L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] &= E[\log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] \\ &= \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int \log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})\, d\boldsymbol{z} \\ &= \log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})\, d\boldsymbol{z} \qquad (231.5) \\ &= \log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) \\ &= L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}). \end{aligned}$$

Durch Einsetzen von (231.5) in (231.4) erhält man die

Schlüsselgleichung des EM-Algorithmus:

$$L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) = E[\log f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] - E[\log f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] \quad (231.6)$$

bzw.

$$L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) = Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}). \quad (231.7)$$

Hierin bedeuten $Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ und $H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ Abkürzungen für die beiden Erwartungswerte in (231.6). Die Größe $Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ wird in diesem Zusammenhang als *Kullback-Leibler-Statistik* und $-H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ als *Entropie*[8] bezeichnet.(Vgl. hierzu u.a. [Hornegger 1996],[Förstner 1991])

Explizit gilt für die **Kullback-Leibler-Statistik**

$$\begin{aligned} Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) &:= E[\log f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] \\ &= \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) d\boldsymbol{z} \qquad (231.8) \end{aligned}$$

und für die **Entropie**

$$\begin{aligned} H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) &:= E[\log f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] \\ &= \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int \log f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)}) f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) d\boldsymbol{z}, \qquad (231.9) \end{aligned}$$

[8] Der Begriff der Entropie stammt aus der Informationstheorie. Unter der Entropie versteht man den Mittleren Wert der zu übermittelnden Information. (vgl. [Förstner 1991])

wobei $f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)})$ mit (231.1) berechnet werden kann[9].

Bei gegebenem Vektor $\hat{\boldsymbol{\beta}}^{(i+1)}$ lassen sich die beiden Erwartungswerte der rechten Seite von (231.6) bzw. (231.7) mittels der Integrale (231.8) und (231.9) und damit die logarithmierte Likelihoodfunkion $L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)})$ (ggf. durch numerische Integration) berechnen.

Gleichung (231.6) bzw. (231.7) kann als *Schlüsselgleichung* des EM - Algorithmus angesehen werden, da sie die Grundlage für das nachfolgend vorgestellte iterative Verfahren zur Bestimmung der unbekannten Parameter bereitstellt. Aus dem beschriebenen Verfahren folgt unmittelbar der EM-Algorithmus.

Iteratives Verfahren:

1.) Es wird ein Startvektor $\hat{\boldsymbol{\beta}}^{(0)}$ als vorläufige Schätzung für die unbekannten Parameter gewählt. Die Wahl erfolgt mehr oder weniger willkürlich.

2.) Ausgehend von der vorläufigen Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der unbekannten Parameter wird (231.6) in Verbindung mit (231.8) und (231.9) dazu verwendet, die logarithmierte Likelihoodfunktion $L(\boldsymbol{y} \mid \boldsymbol{\beta}^{(i+1)})$ für alternative Parameter $\boldsymbol{\beta}^{(i+1)}$ zu berechnen.[10] Es werden diejenigen Parameter $\hat{\boldsymbol{\beta}}^{(i+1)}$ bestimmt, für die $L(\boldsymbol{y} \mid \boldsymbol{\beta}^{(i+1)})$ maximal wird. Diese Parameter stellen neue (verbesserte) Schätzwerte für die unbekannten Parameter dar.

3.) Falls die Abweichung $\|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\|$ der neuen Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ von der vorherigen Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ einen Schwellwert ϵ überschreitet, wird das Verfahren mit den verbesserten Schätzwerten bei 1.) fortgesetzt.

4.) Die endgültigen Schätzwerte $\hat{\boldsymbol{\beta}}^{*}$ werden ausgegeben.

Das iterative Vorgehen ist notwendig, da in jeder Iteration die erhaltenen vorläufigen Schätzwerte von den Schätzwerten des vorherigen Schrittes abhängen.

Übergang auf den EM-Algorithmus:

Beim oben beschriebenen iterativen Verfahren wird in jeder Iteration das Maximum der logarithmierten Likelihoodfunktion entweder analytisch oder numerisch bestimmt. Wegen

$$L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) = Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}).$$

müssen dazu die Kullback-Leibler-Statistik $Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ und die negative Entropie $H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ in Abhängigkeit von $\hat{\boldsymbol{\beta}}^{(i+1)}$ analytisch oder numerisch berechnet werden. Wie aus (231.1) und den Erläuterungen zu (222.1) und (231.9) hervorgeht, führt dabei insbesondere die numerische Berechnung der Entropie $-H$ zu erheblichem Rechenaufwand.

Zur Verringerung des Rechenaufwandes macht man sich im Hinblick auf den EM-Algorithmus folgende Eigenschaft der Entropie zunutze: Es kann gezeigt werden, daß im $(i+1)$-ten Iterationsschritt die negative Entropie $H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ für jeden beliebigen Vektor $\boldsymbol{\beta}'$

[9] Wie weiter unten erläutert, wird im Rahmen des EM-Algorithmus bei der Auswertung von (231.6) bzw. (231.7) auf die Berechnung der Entropie verzichtet. Somit braucht die aufwendige Berechnung von $f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)})$ und die zugehörige Integration tatsächlich nicht zu erfolgen

[10] Die Berechnung der logarithmierten Likelihoodfunktion und die anschließende Maximierung kann in vielen Fällen analytisch erfolgen. Ist dies nicht möglich, so muß mit numerischen Verfahren gearbeitet werden, bei denen die logarithmierte Likelihoodfunkion (231.6) numerisch berechnet.

mindestens so groß wird wie die sich mit den „alten" Parametern $\hat{\boldsymbol{\beta}}^{(i)}$ ergebende Größe $H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$. Die entsprechende Aussage liefert der

Satz: *Es seien $\hat{\boldsymbol{\beta}}^{(i)}$ und $\boldsymbol{\beta}'$ zwei Vektoren des Parameterraumes $\mathcal{B}$, dann gilt*

$$H(\boldsymbol{\beta}' \mid \hat{\boldsymbol{\beta}}^{(i)}) \leq H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}). \tag{231.10}$$

Identität herrscht genau dann, wenn überall in $\boldsymbol{z}$ gilt

$$f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) = f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}') \tag{231.11}$$

Beweis: Betrachte die Differenz

$$\begin{aligned}
&H(\boldsymbol{\beta}' \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \\
&\quad = E[\log f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}') \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] - E[\log f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] \\
&\quad = E[\log f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}') - \log f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] \\
&\quad = E[\log \left(\frac{f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}')}{f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})} \right) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] \\
&\quad = \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int \log \left\{ \frac{f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}')}{f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})} \right\} f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) d\boldsymbol{z}
\end{aligned} \tag{231.12}$$

Durch Taylor-Entwicklung der Logarithmusfunktion $\log x$ an der Stelle $x_0 = 1$ ergibt sich bei Verwendung des Lagrange-Restgliedes vom Grad 2 (vgl. [Helfrich II 1996])

$$\log(x) = (x-1) - \underbrace{\frac{1}{2\xi^2} \cdot (x-1)^2}_{\geq 0},$$

mit ξ zwischen 1 und x, also $\xi \in [x;1]$ falls $x < 1$ bzw. $\xi \in [1;x]$ falls $x \geq 1$. Daraus folgt $\log(x) \leq x - 1$, wobei Gleichheit genau dann gilt, wenn $x = 1$. Hiermit ergibt sich

$$\begin{aligned}
&H(\boldsymbol{\beta}' \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \\
&\quad \leq \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int \left\{ \frac{f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}')}{f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})} - 1 \right\} f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) d\boldsymbol{z} \\
&\quad = \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int \{ f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}') - f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) \} d\boldsymbol{z} \\
&\quad = \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}') d\boldsymbol{z} - \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) \} d\boldsymbol{z} = 1 - 1 = 0,
\end{aligned}$$

da nach (222.1) und (231.1) gilt

$$\int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(\phi)}) d\boldsymbol{z} = \frac{1}{g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(\phi)})} \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(\phi)}) d\boldsymbol{z} = 1 \quad \text{für } \phi = (i, i+1).$$

Hieraus folgt schließlich die erste Aussage $H(\boldsymbol{\beta}' \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \leq 0$.

Entsprechend der im Zusammenhang mit der Taylorentwicklung der Logarithmusfunktion getroffenen Aussagen gilt das Gleichheitszeichen genau dann, wenn in (231.12) und damit überall in $\boldsymbol{z}$ gilt $f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\beta}') = f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})$. Somit sind beide Aussagen des Satzes bewiesen. □

Aus der Eigenschaft $H(\boldsymbol{\beta}' \mid \hat{\boldsymbol{\beta}}^{(i)}) \leq H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ resultiert im Hinblick auf den EM-Algorithmus nun folgende Überlegung: Wird bei dem oben beschriebenen iterativen Verfahren nur die Kullback-Leibler-Statistik Q betrachtet und werden dabei im $(i+1)$-ten Iterationsschritt Schätzwerte $\hat{\boldsymbol{\beta}}^{(i+1)}$ gefunden, für die $Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ gilt, so ist damit wegen $H \geq 0$ nach (231.9) und $H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \leq H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ gleichzeitig eine Vergrößerung der logarithmierten Likelihoodfunktion

$$L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) = Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i)})$$

im Vergleich zu $L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i)})$ verbunden. Aus diesem Grund verzichtet man auf die (rechentintensive) Auswertung der negativen Entropie H und beschränkt sich bei den Iterationen des EM-Algorithmus auf die Maximierung der Kullback-Leibler-Statistik Q[11].

Im folgenden wird der EM-Algorithmus formal definiert.

232 Definition des EM-Algorithmus

Wie im letzten Abschnitt bereits erläutert, besteht der EM-Algorithmus in der iterativen Berechnung und Maximierung der Kullback-Leibler-Statistik $Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$. Die Definition des EM-Algorithmus in seiner allgemeinen Form lautet (vgl. [Dempster et al. 1968]):

Definition: [EM-Algorithmus]
Für den Zufallsvektor $\boldsymbol{z}$ der vollständigen Beobachtungen sei die von den unbekannten, festen Parametern $\boldsymbol{\beta}$ abhängige Wahrscheinlichkeitsdichte $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ bekannt. In $\boldsymbol{z}$ seien fehlende, also durch Messungen nicht direkt zugängliche Beobachtungen wie ggf. auch tatsächliche Beobachtungen enthalten. Die tatsächlichen Beobachtungen seien im Vektor $\boldsymbol{y}$ und die unzugänglichen Beobachtungen im Vektor $\boldsymbol{x}$ zusammengefaßt. Dann besteht der *Expectation Maximization - Algorithmus (EM-Algorithmus)* aus der iterativen Anwendung des *Expectation*-Schrittes (*E-Schritt*) und des *Maximization*-Schrittes (*M-Schritt*):

E-Schritt: Berechne die Kullback-Leibler-Statistik

$$Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = E[\log f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i+1)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}] \qquad \text{nach (231.8)} \tag{232.1}$$

M-Schritt: Maximiere die Kullback-Leibler-Statistik, um die neue Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ zu erhalten.

$$\hat{\boldsymbol{\beta}}^{(i+1)} := \arg \max_{\hat{\boldsymbol{\beta}}^{(i+1)} \in \mathcal{B}} Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \tag{232.2}$$

Hierin bezeichnet $\mathcal{B}$ den Parameterraum der unbekannten Parameter.

Beim EM-Algorithmus wird im M-Schritt die Maximierung der Kullback-Leibler-Statistik gefordert. Es muß also gelten

$$Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq Q(\boldsymbol{\beta}' \mid \hat{\boldsymbol{\beta}}^{(i)}) \quad \text{für alle Paare} \quad (\boldsymbol{\beta}', \hat{\boldsymbol{\beta}}^{(i)}) \in \mathcal{B} \times \mathcal{B}. \tag{232.3}$$

[11] Die Beschränkung auf die Maximierung der Kullback-Leibler-Statistik führt dazu, daß die Werte $\hat{\boldsymbol{\beta}}^{(i+1)}$, für die Q maximal wird, nicht unbedingt auch die logarithmierte Likelihoodfunktion maximieren. Daher sind für die Kullback-Leibler-Statistik und die logarithmierte Likelihoodfunktion differenzierte Konvergenzbetrachtungen erforderlich(vgl. Kap 3.

Fordert man hingegen nur, daß in der Iteration der neue Schätzwert $\hat{\boldsymbol{\beta}}^{(i+1)}$eine *mindestens so große* Kullback-Leibler-Statistik liefert wie der vorherige Schätzwert $\hat{\boldsymbol{\beta}}^{(i)}$, so erhält man den sogenannten *Generalized EM -Algorithmus (GEM).*

Definition:[GEM-ALGORITHMUS]
Wird im M-Schritt lediglich gefordert, daß die im $(i+1)$-ten Iterationsdurchgang mit der neuen Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ berechnete Kullback-Leibler-Statistik mindestens so groß ist wie die mit der alten Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$berechnete Kullback-Leibler-Statistik, also

$$Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \tag{232.4}$$

so bezeichnet man den zugehörigen Algorithmus als *Generalized Expectation-Maximization-Algorithmus (GEM-Algorithmus).*

Diese Definitionen stellen die Grundlage für die nachfolgenden Untersuchungen der Eigenschaften des EM-Algorithmus bzw. des GEM-Algorithmus dar. Der Ablauf der Parameterschätzung aus unvollständigen Beobachtungsdaten mittels des EM-Algorithmus ist in Abbildung 1 schematisch dargestellt.

233 Zusammenfassung

In diesem Kapitel wurde die Maximum-Likelihood-Methode kurz vorgestellt und erläutert, wie sich damit Parameterschätzungen aus unvollständigen Beobachtungsdaten durchführen lassen. Im Anschluß daran wurde die Schlüsselgleichung des EM-Algorithmus abgeleitet und das sich daraus ergebende iterative Verfahren zur Parameterschätzung aus unvollständigen Daten beschrieben. Es wurde aufgezeigt, wie durch aus diesem iterativen Verfahren der EM-Algorithmus entsteht. Abschließend wurden der EM-Algorithmus und der GEM-Algorithmus formal definiert und deren Unterschied wurde erläutert.

Im nächsten Kapitel werden die Eigenschaften des (G)EM-Algorithmus beschrieben. Für die konkrete Anwendung dieses Algorithmus ist dabei insbesondere die Frage zu stellen, unter welchen Bedingungen der Algorithmus wie schnell zu welchem Ergebnis führt.

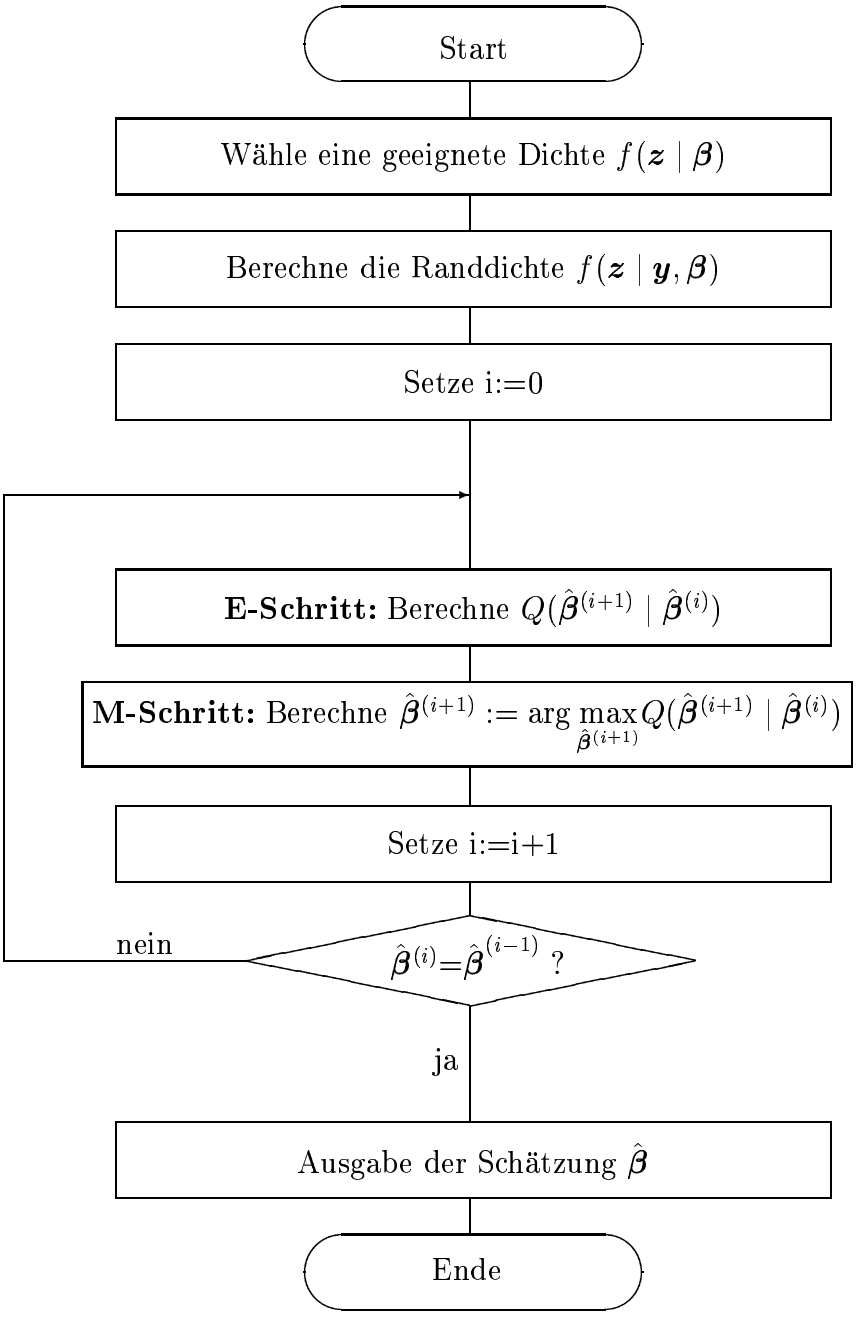


Abbildung 1: Ablaufschema des EM-Algorithmus

Kapitel 3

Eigenschaften des EM-Algorithmus

In diesem Kapitel werden die Eigenschaften des EM-Algorithmus beschrieben. Insbesondere ist hierbei von Interesse, unter welchen Bedingungen der Algorithmus konvergiert und ob er tatsächlich wie gewünscht die logarithmierte Likelihoodfunktion maximiert.

Wie sich aus den Definitionen (232.2) bzw. (232.4) ergibt, stellt der EM-Algorithmus einen Spezialfall des allgemeineren GEM-Algorithmus dar. Daher werden im Weiteren vor allem die Eigenschaften des GEM-Algorithmus wiedergegeben. Für den EM-Algorithmus gelten diese Eigenschaften dann entsprechend.

31 Vorbemerkungen

Im Rahmen des EM-Algorithmus werden in jeder Iteration neue (verbesserte) Schätzwerte $\hat{\boldsymbol{\beta}}^{(i+1)}$ für die unbekannten Parameter unter Zuhilfenahme der in der vorherigen Iteration erhaltenen Schätzwerte $\hat{\boldsymbol{\beta}}^{(i)}$ bestimmt. Es entsteht eine Folge von Parametervektoren

$$\left\{\hat{\boldsymbol{\beta}}^{(i)}\right\}_{i\geq 0} = \hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \dots \,, \tag{310.1}$$

deren einzelne Folgenglieder jeweils vom vorherigen Folgenglied abhängen. Aufrund dieser Abhängigkeit wird davon ausgegangen, daß der Parametervektor $\hat{\boldsymbol{\beta}}^{(i+1)}$ der $(i+1)$-ten Iteration formal aus dem Parametervektor $\hat{\boldsymbol{\beta}}^{(i)}$ der i-ten Iteration durch eine Abbildung $A : \mathcal{B} \to \mathcal{B}$ hervorgeht; der Maximization-Schritt wird also beschrieben durch

$$A : \hat{\boldsymbol{\beta}}^{(i)} \mapsto \hat{\boldsymbol{\beta}}^{(i+1)} = A(\hat{\boldsymbol{\beta}}^{(i)}). \tag{310.2}$$

Im Allgemeinen ergeben sich für verschiedene Startwerte $\hat{\boldsymbol{\beta}}^{(0)}$ in der i-ten Iteration unterschiedliche Schätzwerte $\hat{\boldsymbol{\beta}}^{(i)}$. Die zu den Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ gehörenden Werte $L(\hat{\boldsymbol{\beta}}^{(i)}) = L(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i)})$ der logarithmierten Likelihoodfunktion bilden die Folge

$$\left\{L(\hat{\boldsymbol{\beta}}^{(i)})\right\}_{i\geq 0} = L(\hat{\boldsymbol{\beta}}^{(0)}), L(\hat{\boldsymbol{\beta}}^{(1)}), L(\hat{\boldsymbol{\beta}}^{(2)}), \dots \,. \tag{310.3}$$

Entsprechend der Definition (232.2) kommen beim EM-Algorithmus in jeder Iteration $(i+1)$ als neue Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ alle Vektoren $\hat{\boldsymbol{\beta}}$ in Frage, die die Kullback-Leibler-Statistik $Q(\hat{\boldsymbol{\beta}} \mid \hat{\boldsymbol{\beta}}^{(i)})$ in dieser Iteration maximieren; sie bilden die Lösungsmege $N^{(i+1)}$ des M-Schrittes der $(i+1)$-ten Iteration. Wegen der Abhängigkeit der Kullback-Leibler-Statistik von der Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ hängt die Lösungsmenge der $(i+1)$-ten Iteration von der

Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ aus der i-ten Iteration ab. Diesem Umstand Rechnung tragend, wird die Abbildung

$$N : \mathcal{B} \to \mathcal{U} \subset \mathcal{B} \quad \text{bzw.} \tag{310.4}$$

$$N : \hat{\boldsymbol{\beta}}^{(i)} \mapsto N(\hat{\boldsymbol{\beta}}^{(i)}) = \{\boldsymbol{\beta} : Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq Q(\bar{\boldsymbol{\beta}} \mid \hat{\boldsymbol{\beta}}^{(i)}) \; \forall \bar{\boldsymbol{\beta}} \in \mathcal{B}\} \tag{310.5}$$

des Parameterraumes $\mathcal{B}$ auf die Menge $\mathcal{U}$ aller Teilmengen des Parameterraums eingeführt. Durch sie wird jeder Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der i-ten Iteration die sich hieraus ergebende Lösungsmenge $N^{(i+1)} = N(\hat{\boldsymbol{\beta}}^{(i)})$ des M-Schrittes der $(i+1)$-ten Iteration zugeordnet[1]. Analog wird entsprechend der Definition (232.4) für den GEM-Algorithmus die Abbildung

$$M : \mathcal{B} \to \mathcal{U} \subset \mathcal{B} \quad \text{bzw.} \tag{310.6}$$

$$M : \hat{\boldsymbol{\beta}}^{(i)} \mapsto M(\hat{\boldsymbol{\beta}}^{(i)}) = \{\boldsymbol{\beta} : Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})\} \tag{310.7}$$

des Schätzwertes $\hat{\boldsymbol{\beta}}^{(i)}$ auf die Teilmenge $M(\hat{\boldsymbol{\beta}}^{(i)})$ des Parameterraumes eingeführt. Mit ihr ergibt sich die Lösungsmenge $M^{(i+1)}$ des M-Schrittes in der $(i+1)$-ten EM-Iteration zu $M^{(i+1)} = M(\hat{\boldsymbol{\beta}}^{(i)})$[1].

Bei den Abbildungen N und M wird jeweils ein Punkt $\hat{\boldsymbol{\beta}}^{(i)}$ auf eine Menge $N(\hat{\boldsymbol{\beta}}^{(i)})$ bzw. $M(\hat{\boldsymbol{\beta}}^{(i)})$ abgebildet. Eine solche Abbildung bezeichnet man als eine *Punkt-Menge-Abbildung (PMA)*. Im Zusammenhang mit Punkt-Menge-Abbildungen wird häufig der Begriff der *Abgeschlossenheit* verwendet.

Definition: [Abgeschlossenheit einer Punkt-Menge-Abbildung]
Man bezeichnet eine auf der Menge $\mathcal{B}$ erklärte Punkt-Menge-Abbildung $M : \mathcal{B} \to \mathcal{U}$ von Punkten $\boldsymbol{\beta} \in \mathcal{B}$ auf Mengen $M(\boldsymbol{\beta}) \in \mathcal{U}$ als abgeschlossen an der Stelle $\boldsymbol{\beta}^*$, *falls sie folgende Eigenschaft besitzt:*
Ist $\{\boldsymbol{\beta}^{(k)}\}_{k\geq 0}$ mit $\boldsymbol{\beta}^{(k)} \in \mathcal{B}$ eine im Urbild $\mathcal{B}$ definierte Folge mit dem Grenzwert $\boldsymbol{\beta}^*$, d.h. $\boldsymbol{\beta}^{(k)} \to \boldsymbol{\beta}^*$ und $\{L^{(k)}\}_{k\geq 0}$ eine auf der Menge $\mathcal{U}$ definierte Folge mit dem Grenzwert L^*, d.h. $L^{(k)} \to L^*$, deren Folgenglieder $L^{(k)}$ jeweils in der Menge $M(\boldsymbol{\beta}^{(k)})$ liegen, also $L^{(k)} \in M(\boldsymbol{\beta}^{(k)})$, dann folgt $L^* \in M(\boldsymbol{\beta}^*)$. (310.8)
Die Abgeschlossenheitsbedingung für Punkt-Menge-Abbildungen ist vergleichbar mit der Stetigkeitsbedingung für Abbildungen, bei denen Punkte auf Punkte abgebildet werden (Vgl. [Kerner et al. 1995], S. 277).

In den folgenden Abschnitten wird bei der Beschreibung der Eigenschaften des (G)EM-Algorithmus auf die Vorbemerkungen dieses Abschnittes zurückgegriffen.

32 Eigenschaften des (G)EM-Algorithmus

Bereits aus den Erläuterungen zur Schlüsselgleichung des EM-Algorithmus in Abschnitt 231 und der Definition (232.4) des GEM-Algorithmus gehen zwei Eigenschaften des (G)EM-Algorithmus hervor:

1.) In der $(i+1)$-ten Iteration liefert die neue Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ eine mindestens gleich große Kullback-Leibler-Statistik wie die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der vorherigen Iteration i, $Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$. Dies ergibt sich unmittelbar aus der Definition (232.4) des GEM-Algorithmus.

[1] Im folgenden ist also zu berücksichtigen, daß das Symbol N entweder im Zusammenhang mit der Lösungsmenge $N^{(i+1)}$ eines M-Schrittes oder mit der Abbildung $N(\hat{\boldsymbol{\beta}}^{(i)})$ auftritt. Entsprechendes gilt für M.

2.) Die negative Entropie H wird in der $(i+1)$-ten Iteration mit der neuen Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ höchstens so groß wie mit der Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der vorherigen Iteration i, $H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \leq H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$. (vgl. (231.11))

Über das aus diesen Eigenschaften resultierende Verhalten der logarithmierten Likelihoodfunktion $L(\boldsymbol{\beta})$ während der EM-Iterationen wurde bisher noch keine explizite Aussage getroffen. Das Verhalten von $L(\boldsymbol{\beta})$ ist aber besonders von Interesse, da das Ziel des EM-Algorithmus ja darin besteht, zu einer Maximum-Likelihood-Schätzung zu gelangen. Im nächsten Abschnitt wird daher das Verhalten der logarithmierten Likelihoodfunktion im Rahmen des GEM-Algorithmus beschrieben und es werden Bedingungen genannt, unter denen die logarithmierte Likelihoodfunktion zu einem lokalen Maximum bzw. zu einem stationären Punkt konvergiert. Im darauf folgenden Abschnitt erfolgen dann Betrachtungen zur Konvergenz der Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$.

321 Zum Verhalten der Likelihoodfunktion während der (G)EM-Iterationen

Eine erste Aussage bezüglich des Verhaltens der logarithmierten Likelihoodfunktion in den (G)EM-Iterationen liefert der

Satz: *Im Rahmen des GEM-Algorithmus wächst die Folge $L(\hat{\boldsymbol{\beta}}^{(0)}), L(\hat{\boldsymbol{\beta}}^{(1)}), \ldots$ der in den Iterationen erhaltenen logarithmierten Likelihoodfunktionen $L(\boldsymbol{\beta}) = \log g(\boldsymbol{y} \mid \boldsymbol{\beta})$ monoton, d.h. es gilt für jede Iteration $(i+1)$*

$$L(\hat{\boldsymbol{\beta}}^{(i+1)}) \geq L(\hat{\boldsymbol{\beta}}^{(i)}), \tag{321.1}$$

wobei das Gleichheitszeichen genau dann gültig ist, wenn sowohl

$$Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \tag{321.2}$$

als auch

$$f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)}) = f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) \quad \textit{für alle } \boldsymbol{z} \in \mathcal{Z} \tag{321.3}$$

gilt.

BEWEIS: Nach (231.7) gilt für die Differenz der logarithmierten Likelihoodfunktion zweier aufeinanderfolgender Iterationen i und $(i+1)$

$$\begin{aligned} L(\hat{\boldsymbol{\beta}}^{(i+1)}) - L(\hat{\boldsymbol{\beta}}^{(i)}) = {} & Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \\ & - [Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \boldsymbol{\beta}^{(i-1)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \boldsymbol{\beta}^{(i-1)})] \end{aligned} \tag{321.4}$$

Zum Beweis des Satzes muß gezeigt werden, daß dieser Ausdruck nicht negativ ist und daß Gleichheit genau unter den genannten Bedingungen gilt. Hierzu wird zunächst die rechte Seite von (321.4) umgeformt. Es gilt

$$\begin{aligned} & Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i-1)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i-1)}) \\ & \quad = E[\log f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i-1)}] - E[\log f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i-1)}] \\ & \quad = E[\log f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i)}) - \log f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i-1)}] \end{aligned}$$

$$= E[\log \frac{f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i)})}{f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i-1)}], \qquad \text{mit } \frac{f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(i)})}{f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})} \overset{(231.1)}{=} g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i)})$$

$$= E[\log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i)}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i-1)}] = \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int \log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i)}) f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i-1)}) d\boldsymbol{z}$$

$$= \log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i)}) \int_{\mathcal{Z}(\boldsymbol{y})} \cdots \int f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i-1)}) d\boldsymbol{z}$$

$$= \log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i)}).$$

Nach (231.3) und (231.7) gilt aber auch $\log g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(i)}) = Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$, so daß für die rechte Seite von (321.4) folgt

$$L(\hat{\boldsymbol{\beta}}^{(i)}) = Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i-1)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i-1)}) = Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}). \quad (321.5)$$

Zum Beweis des Satzes ist somit die Beziehung

$$Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - [Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})] \geq 0$$

zu verifizieren. Aus der Definition (232.4) des GEM-Algorithmus und der Beziehung $H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ aus (231.11) folgt

$$Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) + H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$$
$$= \underbrace{[Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})]}_{=:R \geq 0} + \underbrace{[H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})]}_{=:S \geq 0} \geq 0.$$

Hieraus folgt die erste Aussage.
Gleichheit besteht genau dann, wenn R=S=0 gilt. Nach (231.11) ist dies genau dann der Fall, wenn (321.2) und (321.3) erfüllt sind. Damit ist der Satz vollständig bewiesen. □

Das monotone Wachstum der logarithmierten Likelihoodfunktion während der (G)EM-Iterationen stellt eine wichtige Eigenschaft des (G)EM-Algorithmus dar: Zusammen mit der unten geforderten Kompaktheit des Parameterraumes $\mathcal{B}$ (vgl. (322.2)) gewährleistet sie die Konvergenz der logarithmierten Likelihoodfunktion im Rahmen des (G)EM-Algorithmus (vgl. die Betrachtungen zur Konvergenz der logarithmierten Likelihoodfunktion in Abschnitt 322).

Als erste, direkte Konsequenz der Monotonie der Folge $\left\{L(\hat{\boldsymbol{\beta}}^{(i+1)})\right\}_{i \geq 0}$ ergibt sich der

Satz: *Das Maximum $L^{(max)}$ der logarithmierten Likelihoodfunktion stellt einen Fixpunkt des GEM-Algorithmus dar.* (321.6)

Denn falls in der i-ten Iteration des (G)EM-Algorithmus mit $\hat{\boldsymbol{\beta}}^{(i)}$ vorläufige Schätzwerte für die unbekannten Parameter gefunden werden, für die die logarithmierte Likelihoodfunktion $L(\boldsymbol{\beta})$ maximal wird, also $L(\hat{\boldsymbol{\beta}}^{(i)}) = L^{(max)} \geq L(\boldsymbol{\beta})$ für alle $\boldsymbol{\beta} \in \mathcal{B}$, so folgt aus (321.1) für die darauf folgende Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)} \in \mathcal{B}$ in der $(i+1)$-ten Iteration

1.) $L(\hat{\boldsymbol{\beta}}^{(i+1)}) = L(\hat{\boldsymbol{\beta}}^{(i)}) = L^{(max)}$, die logarithmierte Likelihoodfunktion nimmt also für die neuen Schätzwerte $\hat{\boldsymbol{\beta}}^{(i+1)}$ wieder den gleichen Wert an wie im i-ten Schritt,

und somit aus (321.2) und (321.3)

2.) $Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$, die Kullback-Leibler-Statistik nimmt in der $(i+1)$-ten Iteration den gleichen Wert an wie in der i-ten Iteration,

3.) $f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)}) = f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})$ für alle $\boldsymbol{z}$, die Dichtefunktion $f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i+1)})$ ist überall in $\boldsymbol{z}$ identisch mit der Dichtefunktion $f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(i)})$, die mit Hilfe der Schätzwerte der i-ten Iteration berechnet wird.

Obwohl sich der Wert der logarithmierten Likelihoodfunktion in nachfolgenden Iterationen nicht mehr ändert, wenn sie in einer Iteration i einmal ihr Maximum erreicht hat, bedeutet dies jedoch *nicht*, daß damit auch ihr Argument, also die Schätzung $\hat{\boldsymbol{\beta}}$ für die unbekannten Parameter festliegt. Denn nimmt die logarithmierte Likelihoodfunktion ihren Maximalwert für mehrere, voneinander verschiedene $\boldsymbol{\beta} \in \mathcal{B}$ an, so besteht die Möglichkeit, daß die Schätzungen der nachfolgenden Iterationen zwischen diesen Vektoren alterniert, ohne daß die logarithmierte Likelihoodfunktion ihren Wert ändert. Setzt man jedoch voraus, daß die logarithmierte Likelihoodfunktion ihr Maximum nur in einem Punkt $\boldsymbol{\beta}^*$ annimmt und daß dieses Maximum im i-ten EM-Iterationsschritt ereicht wird, also $L(\hat{\boldsymbol{\beta}}^{(i)}) = L(\boldsymbol{\beta}^*) > L(\boldsymbol{\beta})$ für alle $\boldsymbol{\beta} \in \mathcal{B} \setminus \{\hat{\boldsymbol{\beta}}^{(i+1)}\}$, so folgt für die nächste Iteration des GEM-Algorithmus

$$\hat{\boldsymbol{\beta}}^{(i+1)} = \hat{\boldsymbol{\beta}}^{(i)} = \boldsymbol{\beta}^*.$$

(Dies läßt sich durch Widerspruch beweisen. Nimmt man an, daß $\hat{\boldsymbol{\beta}}^{(i+1)} \neq \hat{\boldsymbol{\beta}}^{(i)}$, so gilt $L(\hat{\boldsymbol{\beta}}^{(i+1)}) > L^{(max)} > L(\hat{\boldsymbol{\beta}}^{(i+1)}) \in \mathcal{B}$, was einen Widerspruch zur Annahme darstellt.)

Zusammenfassend kann also der folgende Satz festgehalten werden:

Satz: *Hat die logarithmierte Likelihoodfunktion $L(\boldsymbol{\beta})$ innerhalb der (G)EM-Iterationen einmal ihren Maximalwert $L^{(max)}$ erreicht, so ändert sich ihr Wert in den nachfolgenden Iterationen nicht mehr. Besitzt die logarithmierte Likelihoodfunktion nur an einer Stelle $\boldsymbol{\beta}^* \in \mathcal{B}$ dieses Maximum, so liegt dann auch der endgültige Schätzwert $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$ fest, sonst liegt eine mehrdeutige Lösung des Optimierungsproblems vor.* (321.7)

322 Betrachtungen zur Konvergenz der logarithmierten Likelihoodfunktion

Die Tatsache, daß die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)}\}_{i \geq 0}$ der logarithmierten Likelihoodfunktion während der (G)EM-Iterationen monoton wächst und daß das globale Maximum L^* einen Fixpunkt des Algorithmus darstellt, lässt noch keine Aussage darüber zu, ob und ggf. unter welchen Bedingungen $\{L(\hat{\boldsymbol{\beta}}^{(i)}\}_{i \geq 0}$ überhaupt konvergiert. Wird in einer Iteration das globale Maximum der logarithmierten Likelihoodfunktion erhalten, so folgt zwar aus den bisherigen Überlegungen, daß $\{L(\hat{\boldsymbol{\beta}})\}_{i \geq 0}$ konvergiert. Es ist bisher jedoch nicht geklärt, ob das Maximum L^* in den (G)EM-Iterationen überhaupt erreicht wird.

Für die Betrachtungen zur Konvergenz der logarithmierten Likelihoodfunktion wird von folgenden Annahmen ausgegangen:

Annahme 1: *Die Menge $\mathcal{B}_{\boldsymbol{\beta}_0} = \{\boldsymbol{\beta} \in \mathcal{B} : L(\boldsymbol{\beta}) \geq L(\boldsymbol{\beta}_0)\}$ ist für jedes $L(\boldsymbol{\beta}_0) > -\infty$ kompakt.* (322.1)

Annahme 2: *Die Funktion $L(\boldsymbol{\beta})$ ist stetig in $\mathcal{B}$ und differenzierbar im Inneren von $\mathcal{B}$.* (322.2)

Annahme 3: *Die logarithmierte Likelihoodfunktion nimmt ihren Maximalwert L^* im Innern des Parameterraums $\mathcal{B}$ an, d.h. das globale Maximum ist gleichzeitig auch ein lokales Maximum der logarithmierten Likelihoodfunktion $L(\boldsymbol{\beta})$.* (322.3)

Bemerkungen:

Entsprechend der Ausführungen in Abschnitt 21 handelt es sich bei dem Parameterraum $\mathcal{B}$ um eine Teilmenge des u-dimensionalen euklidschen Vektorraumes, $\mathcal{B} \subset \mathbb{R}^u$. Entsprechend den Ausführungen in [KERNER ET AL. 1995] ist eine solche Teilmenge des $\mathbb{R}^u$ genau dann kompakt, wenn sie abgeschlossen[3] und beschränkt[4] ist. Mit Annahme 1 wird also davon ausgegangen, daß es sich bei der Lösungsmenge des Maximization-Schrittes in jeder Iteration des GEM-Algorithmus jeweils um eine abgeschlossene und beschränkte Menge handelt. Dies ist eine sehr schwache Forderung, so daß Annahme 1 nahezu immer richtig ist[5]. Auch die mit den Annahmen 2 und 3 verbundenen Forderungen nach Stetigkeit und Differenzierbarkeit der logarithmierten Likelihoodfunktion bzw. nach dem Auftreten des globalen Maximums der logarithmierten Likelihoodfunktion im Inneren des Parameterraumes, sind für viele Dichtefunktionen kontinuierlicher Zufallsvariable - insbesondere für die Normalverteilung und die Exponentialverteilung (vgl. [DEMPSTER ET AL. 1968]) - erfüllt. Somit sind die nachfolgend unter den Annahmen 1, 2 und 3 getroffenen Aussagen praktisch von sehr allgemeiner Gültigkeit.

Die beiden Annahmen (322.1) und (322.2) lassen nachstehende *Folgerung* zu:
Nach (322.1) entstammt jedes Glied der Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ der beschränkten Menge $\mathcal{B}_{\hat{\boldsymbol{\beta}}^{(0)}}$. Somit ist $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ eine beschränkte Folge. Wegen der angenommenen Stetigkeit der logarithmierten Likelihoodfunktion $L(\boldsymbol{\beta})$ nach (322.2) folgt hieraus, daß auch die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ beschränkt ist. Insgesamt handelt es sich also bei $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ um eine monoton wachsende und unter den (praktisch sehr häufig zutreffenden) Annahmen (322.1) und (322.2) um eine nach oben beschränkte Folge. Nach dem Satz von BOLZANO-WEIYERSTRASS (vgl. [KERNER ET AL. 1995], S. 35) ist jede solche monotone und beschränkte Folge konvergent, womit die Konvergenz von $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ folgt. Dies führt auf den

Satz: *Mit Ausnahme der Fälle, in denen die Annahmen (322.1) und (322.2) nicht zutreffen, ist beim (G)EM-Algorithmus die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ der Werte der logarithmierten Likelihoodfunktion, die sich mit den in den (G)EM-Iterationen erhaltenen Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ für die unbekannten Parameter ergeben, konvergent. Der Grenzwert wird mit L^* bezeichnet.* (322.4)

Nachdem die Konvergenz der logarithmierten Likelihoodfunktion im Rahmen des (G)EM-Algorithmus nachgewiesen ist, stellt sich die Frage, ob es sich bei dem Grenzwert L^* um das gesuchte globale Maximum der logarithmierten Likelihoodfunktion handelt; die bisherigen Überlegungen lassen diesbezüglich noch keine Aussage zu.

Treffen die Annahmen 1, 2 und 3 zu, so handelt es sich bei dem Grenzwert L^* entweder

- um einen Wert, der zu einem lokalen Maximum der logarithmierten Likelihoodfunktion gehört, das mit ihrem globalen Maximum übereinstimmt, *oder*
- um einen Wert, der zu einem sonstigen lokalen Maximum der logarithmierten Likelihoodfunktion gehört *oder*
- um einen Wert, der zu einem stationären Punkt[6] gehört.

[3] Eine Teilmenge $\mathcal{E}$ des $\mathbb{R}^u$ heißt abgeschlossen, wenn ihr Komplement $\bar{\mathcal{E}} = \mathbb{R}^u \backslash \mathcal{E}$ offen ist, d.h. wenn zu jedem Punkt $\bar{\mathbf{e}} \in \bar{\mathcal{E}}$ eine Kugel mit Mittelpunkt in $\bar{\mathbf{e}}$ existiert, die vollständig in $\bar{\mathcal{E}}$ liegt.

[4] Eine Teilmenge des R^u heißt beschränkt, wenn es eine Kugel mit endlichem Radius gibt,die die Menge volltsändig einschließt.

[5] Zumindest dem Verfasser ist kein Fall bekannt, in dem Annahme 1 nicht erfüllt wäre.

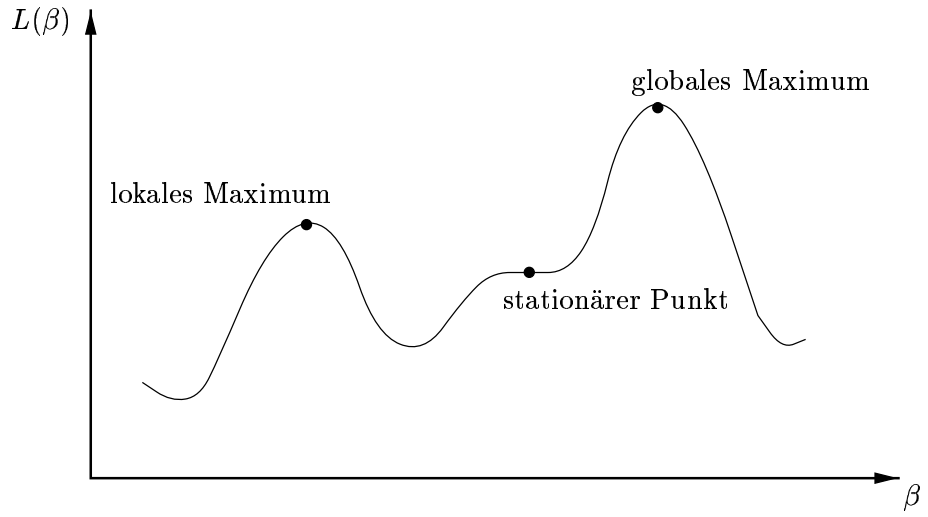


Abbildung 1: Theoretisch mögliche Konvergenzpunkte des EM-Algorithmus

Für den Fall, daß bei einer Parameterschätzung nur ein unbekannter Parameter β zu bestimmen ist, sind die möglichen Fälle in Abbildung 1 dargestellt.

Selbst wenn im M-Schritt des (G)EM-Algorithmus globale Optimierungsverfahren zur Maximierung der Kullback-Leibler-Statistik eingesetzt werden, ist es i.a. nicht sicher, daß der (G)EM-Algorithmus zum gesuchten globalen Maximum der logarithmierten Likelihoodfunktion führt. Das liegt daran, daß beim (G)EM-Algorithmus die negative Entropie H unberücksichtigt bleibt: Nach der Schlüsselgleichung des (G)EM-Algorithmus 231.7 gilt mit der Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der i-ten (G)EM-Iteration die Beziehung $L(\boldsymbol{\beta}) = Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})$. Im Rahmen des (G)EM-Algorithmus wird die hierin enthaltene negative Entropie H nicht weiter berücksichtigt, sondern im M-Schritt wird lediglich die Kullback-Leibler-Statistik maximiert. Diese Vernachlässigung der Größe H führt dazu, daß der (G)EM-Algorithmus auf eine Schätzung $\hat{\boldsymbol{\beta}}$ führt, die i.a. *nicht* zugleich die Größe $L = Q + H$ global maximiert. So kommt es zur Konvergenz der logarithmierten Likelihoodfunktion gegen einen Wert, der zu einem (vom globalen Maximum verschiedenen) lokalen Maximum oder zu einem stationären Punkt gehört.

Im folgenden soll nun erklärt werden, unter welchen Bedingungen im Rahmen des (G)EM-Algorithmus mit der Konvergenz der logarithmierten Likelihoodwerte zu einem globalen Maximum, einem lokalen Maximum oder zu einem stationären Punkt (z.B. Sattelpunkt) zu rechnen ist.

Grundlegend für die hier angestellten Konvergenzbetrachtungen ist das folgende

Theorem über globale Konvergenz

Satz: *Es sei* $\{\boldsymbol{a}_i\}_{i=0}^{\infty}$, $\boldsymbol{a}_i \in \mathcal{A}$ *eine Folge von Punkten mit* $\boldsymbol{a}_{i+1} \in F(\boldsymbol{a}_i)$, *wobei* F *eine auf der Menge* $\mathcal{A}$ *definierte Punkt-Menge-Abbildung bezeichnet. Weiter sei mit* $\Gamma \subset \mathcal{A}$

[6] Ein stationärer Punkt der logarithmierten Likelihoodfunktion ist ein Punkt, in dem der Gradient der logarithmierten Likelihoodfunktion zwar gleich dem Nullvektor ist (d.h. die Notwendige Bedingung für ein lokales Extremum ist in diesem Punkt erfüllt), bei dem es sich jedoch um kein lokales Extremum der logarithmierten Likelihoodfunktion handelt.

eine Lösungsmenge[7] *gegeben und es gelte:*

1.) Alle Punkte $\boldsymbol{a}_i$ *sind Elemente einer kompakten Menge* $\mathcal{S} \subset \mathcal{A}$

2.) Die Punkt-Menge-Abbildung F *ist abgeschlossen in* $\mathcal{A} \backslash \Gamma$

3.) Es ist eine stetige Funktion $\gamma(\boldsymbol{a})$ *auf* $\mathcal{A}$ *definiert, für die gilt*

a) gehört $\boldsymbol{a}$ *nicht zur Lösungsmenge , also* $\boldsymbol{a} \notin \Gamma$*, dann gilt für alle Punkte* $\boldsymbol{b}$ *des Bildes* $F(\boldsymbol{a})$ *die Beziehung* $\gamma(\boldsymbol{b}) > \gamma(\boldsymbol{a})$

b) gehört $\boldsymbol{a}$ *zur Lösungsmenge, also* $\boldsymbol{a} \in \Gamma$*, dann gilt für alle Punkte* $\boldsymbol{b}$ *des Bildes* $F(\boldsymbol{a})$ *die Beziehung* $\gamma(\boldsymbol{b}) \geq \gamma(\boldsymbol{a})$

Dann liegen alle Konvergenzpunkte der Folge $\{\boldsymbol{a}_i\}_0^\infty$ *in der Lösungsmenge* Γ *und* $\{\gamma(\boldsymbol{a}_i)\}_{i\geq 0}$ *konvergiert monoton mit* $\lim_{i\to\infty} \gamma(\boldsymbol{a}_i) = \gamma(\boldsymbol{a})$ *für ein* $\boldsymbol{a} \in \Gamma$ (322.3)

Auf den Beweis dieses Satzes wird hier verzichtet, da er den Rahmen dieser Arbeit sprengen würde. Er kann in ([Zangwill 1969]) gefunden werden.

Nach dem Theorem über globale Konvergenz konvergiert also eine Folge $\{\gamma(\boldsymbol{a}_i)\}_{i\geq 0}$, deren Folgenglieder $\gamma(\boldsymbol{a}_i)$ sich als Funktion der Glieder $\boldsymbol{a}_i$ einer Folge $\{\boldsymbol{a}_i\}_{i\geq 0}$ ergeben, wenn für $\{\boldsymbol{a}_i\}$ eine Lösungsmenge Γ vorgegeben ist und die Bedingungen 1.), 2.), 3.a) und 3.b) erfüllt sind. Dies gilt für allgemeine Folgen $\{\boldsymbol{a}_i\}$, Punkt-Menge-Abbildungen F, Lösungsmengen Γ und Funktionen γ. Somit gilt dieses Theorem auch in dem Spezialfall des (G)EM-Algorithmus, in dem

- es sich bei $\{\boldsymbol{a}_i\}_{i\geq 0}$ um die Folge der beim (G)EM-Algorithmus jeweils im M-Schritt erhaltenen Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ für die Unbekannten Parameter handelt, also $\boldsymbol{a}_i = \hat{\boldsymbol{\beta}}^{(i)}$
- $\{\gamma(\boldsymbol{a}_i)\}_{i\geq 0}$ die entsprechende Folge der logarithmierten Likelihoodwerte darstellt, also $\gamma(\boldsymbol{a}_i) = L(\hat{\boldsymbol{\beta}}^{(i)})$
- F die Punkt-Menge-Abbildung ist, die die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der i-ten (G)EM-Iteration auf die Lösungsmenge $M^{(i+1)}$ bzw. $N^{(i+1)}$ der $(i+1)$-ten Iteration abbildet, also $F(\boldsymbol{a}_i) = M(\hat{\boldsymbol{\beta}}^{(i)})$ beim GEM-Algorithmus und $F(\boldsymbol{a}_i) = M(\hat{\boldsymbol{\beta}}^{(i)})$ beim EM-Algorithmus.
- die Lösungsmenge entweder die Menge Γ^L Lokaler Maxima im Inneren von $\mathcal{B}$, die Menge Γ^S stationärer Punkte oder die Menge Γ^G globaler Maxima der logarithmierten Likelihoodfunktion ist.

Die Menge $\mathcal{A}$ entspricht dann dem Parameterraum $\mathcal{B}$. Es können mit dem Theorem über globale Konvergenz also die Bedingungen formuliert werden, unter denen im Rahmen des (G)EM-Algorithmus die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 1}$ gegen ein globales oder lokales Maximum oder gegen einen stationären Punkt konvergiert.

Zunächst soll angegeben werden, unter welchen Bedingungen die aus den (G)EM-Iterationen resultierende Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ gegen einen stationären Punkt der logarithmierten Likelihoodfunktion konvergiert. Als Lösungsmenge Γ wird hierzu die Menge Γ^S der stationären Punkte von $L(\boldsymbol{\beta})$ im Inneren von $\mathcal{B}$ gewählt. Wegen Annahme (322.1) ist die Bedingung 1 des Theorems über globale Konvergenz für den (G)EM-Algorithmus erfüllt. Genau so ist wegen des monotonen Wachstums der logarithmierten Likelihoodfunktion nach (321.1) für den (G)EM-Algorithmus die Bedingung (3b) erfüllt. Als Konsequenz aus dem Theorem über globale Konvergenz ergibt sich also unmittelbar der

Satz: *Sei* $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ *eine Folge von Schätzungen der unbekannten Parameter aus den Iterationen des GEM-Algorithmus, so daß gilt* $\hat{\boldsymbol{\beta}}^{(i+1)} \in M(\hat{\boldsymbol{\beta}}^{(i)})$*. Weiter sei*

[7] Die Lösungsmenge enthält diejenigen Punkte, mit denen alle bis auf endlich viele Folgenglieder der Folge $\{\boldsymbol{a}_i\}_{i=0}^\infty$ zusammenfallen sollen. Sie wird aufgrund gewisser Vorgaben festgelegt.

1. *die Punkt-Menge-Abbildung M, die die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der unbekannten Parameter der i-ten Iteration auf die Lösungsmenge $M^{(i+1)} = M(\hat{\boldsymbol{\beta}}^{(i)})$ des M-Schrittes der $(i+1)$-ten Iteration abbildet, für $\hat{\boldsymbol{\beta}}^{(i)} \in \mathcal{B}\backslash\Gamma^S$ abgeschlossen, wobei Γ^S die Menge aller stationären Punkte der logarithmierten Likelihoodfunktion im Innern von $\mathcal{B}$ bezeichnet.*
2. *Es gelte $L(\hat{\boldsymbol{\beta}}^{(i+1)}) > L(\hat{\boldsymbol{\beta}}^{(i)})$ für alle $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^S$*

Dann sind alle Konvergenzpunkte von $\{\hat{\boldsymbol{\beta}}^{(i)}\}$ stationäre Punkte von L und $L(\hat{\boldsymbol{\beta}}^{(i)})$ konvergiert monoton mit dem Grenzwert $\lim_{i\to\infty} L(\hat{\boldsymbol{\beta}}^{(i)}) = L^* = L(\boldsymbol{\beta}^*)$ für ein $\boldsymbol{\beta}^* \in \Gamma^S$ (322.6)

Eine entsprechende Aussage in Bezug auf die Konvergenz zu lokalen Maxima erhält man, in dem man mit der Lösungsmenge Γ die Menge Γ^L der lokalen Maxima identifiziert. Man erhält den

Satz: *Sei $\{\hat{\boldsymbol{\beta}}^{(i)}\}$ eine Folge von Schätzungen der unbekannten Parameter aus den Iterationen des GEM-Algorithmus, mit $\hat{\boldsymbol{\beta}}^{(i+1)} \in M(\hat{\boldsymbol{\beta}}^{(i)})$. Weiter sei*

1. *die Punkt-Menge-Abbildung M, die die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der unbekannten Parameter der i-ten Iteration auf die Lösungsmenge $M^{(i+1)} = M(\hat{\boldsymbol{\beta}}^{(i)})$ des M-Schrittes der $(i+1)$-ten Iteration abbildet, für $\hat{\boldsymbol{\beta}}^{(i)} \in \mathcal{B}\backslash\Gamma^L$ abgeschlossen, wobei Γ^L die Menge aller lokalen Maxima der logarithmierten Likelihoodfunktion im Innern von $\mathcal{B}$ bezeichnet.*
2. *$L(\hat{\boldsymbol{\beta}}^{(i+1)}) > L(\hat{\boldsymbol{\beta}}^{(i)})$ für alle $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^L$.*

Dann sind alle Konvergenzpunkte von $\{\hat{\boldsymbol{\beta}}^{(i)}\}$ lokale Maxima von L und $L(\hat{\boldsymbol{\beta}}^{(i)})$ konvergiert monoton mit dem Grenzwert $\lim_{i\to\infty} L(\hat{\boldsymbol{\beta}}^{(i)}) = L^* = L(\boldsymbol{\beta}^*)$ für ein $\boldsymbol{\beta}^* \in \Gamma^L$. (322.7)

Mit den Sätzen (322.6) bzw. (322.7) sind die Voraussetzungen für die Konvergenz der logarithmierten Likelihoodfunktion gegen lokale Maxima bzw. gegen stationäre Punkte in allgemeiner Weise für den GEM-Algorithmus formuliert. Die Voraussetzung 1. der Abgeschlossenheit der Abbildung M in $\mathcal{B}\backslash\Gamma$ ist hierbei allerdings wenig anschaulich und kann für den GEM-Algorithmus in der konkreten Anwendung nicht oder nur schwer nachgeprüft werden. Im Gegensatz dazu ergibt sich für den EM-Algorithmus als Spezialfall des GEM-Algorithmus eine anschauliche und verifizierbare Bedingung für die Abgeschlossenheit der Punkt-Menge-Abbildung M: Nach [Wu 1982], Gl. (10) kann gezeigt werden, daß die Abgeschlossenheit der zum M-Schritt des EM-Algorithmus gehörenden Punkt-Menge-Abbildung M bereits folgt, wenn die Kullback-Leibler-Statistik $Q(\psi \mid \phi)$ sowohl in ψ als auch in ϕ stetig ist. Da die meisten in den Anwendungen verwendeten Dichtefunktionen diese Bedingungen erfüllen, ist mit der Forderung nach Abgeschlossenheit von M eine nur geringe Einschränkung der Allgemeinheit verbunden.

Wie sich aus nachfolgendem Satz ergibt, folgt aus der Stetigkeit von $Q(\psi \mid \phi)$ in ψ und ϕ nicht nur die Abgeschlossenheit der Abbildung M sondern auch die Konvergenz der logarithmierten Likelihoodfuntion gegen einen stationären Punkt.

Satz: *Die Kullback-Leibler-Statistik $Q(\psi \mid \phi)$ sei stetig sowohl in ψ als auch in ϕ. Dann sind alle Konvergenzpunkte jeder Realisierung $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ eines EM-Algorithmus stationäre Punkte der logarithmierten Likelihoodfunktion $L(\hat{\boldsymbol{\beta}}^{(i)})$ und $L(\hat{\boldsymbol{\beta}}^{(i)})$ konvergiert monoton gegen $L^* = L(\boldsymbol{\beta}^*)$ für einen stationären Punkt $\boldsymbol{\beta}^*$.* (322.8)

Beweis: Es muß nachgewiesen werden, daß die Voraussetzungen des Satzes (322.6) erfüllt sind. Da die Stetigkeitsbedingung nach Voraussetzung erfüllt ist, ist die Forderung nach Abgeschlossenheit der Punkt-Menge-Abbildung M erfüllt. Es muß also noch gezeigt werden, daß $L(\hat{\boldsymbol{\beta}}^{(i+1)}) > L(\hat{\boldsymbol{\beta}}^{(i)})$ für alle $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^S$ gilt.
Betrachtet werde ein Punkt $\hat{\boldsymbol{\beta}}^{(i)} \in \mathcal{B}$, der nicht zur Menge Γ^S der stationären Punkte gehört, also $\hat{\boldsymbol{\beta}}^{(i)} \in \mathcal{B}\backslash\Gamma^S$. Nach (231.11) gilt $H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq H(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})$ für alle $\boldsymbol{\beta} \in$

$\mathcal{B}\backslash\{\hat{\boldsymbol{\beta}}^{(i)}\}$, so daß $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(i)}$ die Größe $H(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})$ maximiert. Als notwendige Bedingung hierfür gilt[8]

$$\frac{\partial H(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})}{\partial \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}^{(i)}} =: \boldsymbol{D}^{10} H(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})\Big|_{\hat{\boldsymbol{\beta}}^{(i)}} = \boldsymbol{D}^{10} H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = 0$$

Dieses Ergebnis setzt man in

$$\boldsymbol{D}L(\hat{\boldsymbol{\beta}}^{(i)}) := \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}^{(i)}} = \underbrace{\frac{\partial Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})}{\partial \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}^{(i)}}}_{=:\boldsymbol{D}^{10} Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})} - \underbrace{\frac{\partial H(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})}{\partial \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}^{(i)}}}_{=\boldsymbol{D}^{10} H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})=0} = \boldsymbol{D}^{10} Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$$

ein. Da der Punkt $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^S$ kein stationärer Punkt der logarithmierten Likelihoodfunktion ist, muß der Gradient $\boldsymbol{D}L(\hat{\boldsymbol{\beta}}^{(i)})$ und damit $\boldsymbol{D}^{10} Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ in diesem Punkt vom Nullvektor verschieden sein:

$$\boldsymbol{D}L(\hat{\boldsymbol{\beta}}^{(i)}) = \boldsymbol{D}^{10} Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \neq \mathbf{0} \quad \text{für jedes } \hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^S.$$

Nach der Definition (232.2) des EM-Algorithmus nimmt in der i-ten Iteration die Kullback-Leibler-Statistik $Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})$ für die neue Schätzung $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(i+1)}$ der unbekannten Parameter ihr globales Maximum an. An der Stelle $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(i)}$ kann Q diesen Maximalwert nicht annehmen, da wegen $\boldsymbol{D}^{10} Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \neq \mathbf{0}$ die notwendige Bedingung für ein lokales Maximum nicht erfüllt ist, so daß es mindestens einen Vektor $\boldsymbol{\beta}$ mit $Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)}) > Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ gibt. Es muß daher für die neue Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ gelten

$$Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) > Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}),$$

womit unter Berücksichtigung von $H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ aus (231.11) und mit (321.5) folgt

$$\begin{aligned} L(\hat{\boldsymbol{\beta}}^{(i+1)}) =& Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \\ >& Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \\ \overset{(321.5)}{=}& Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i-1)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i-1)}) = L(\hat{\boldsymbol{\beta}}^{(i)}). \end{aligned}$$

Hieraus folgt die Aussage. □

Ein dem Satz (322.8) entsprechender Satz für die konvergenz des EM-Algorithmus gegen ein lokales Maximum gilt *nicht.*

BEGRÜNDUNG: Anhand eines Gegenbeispiels kann gezeigt werden, daß aus $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^L$ nicht zwangsläufig $L(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ folgt. Betrachtet sei der Fall, in dem es sich bei $\hat{\boldsymbol{\beta}}^{(i)}$ zwar um kein lokales Maximum, aber um einen stationären Punkt der logarithmierten Likelihoodfunktion handelt, also $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^L$ aber $\hat{\boldsymbol{\beta}}^{(i)} \in \Gamma^S$. Wegen der Stationarität von L an der Stelle $\hat{\boldsymbol{\beta}}^{(i)}$ gilt dann

$$DL(\hat{\boldsymbol{\beta}}^{(i)}) = D^{10} Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = 0,$$

und somit ist $L(\hat{\boldsymbol{\beta}}^{(i+1)}) > L(\hat{\boldsymbol{\beta}}^{(i)})$ in stationären Punken nicht gegeben. Ausgehend von einem stationären Punkt erfolgt also keine weitere Konvergenz gegen ein lokales Maximum. Damit ist Bedingung 2. von Satz (322.7) nicht erfüllt und ein dem Satz (322.8) entsprechender Satz für die Konvergenz der logarithmierten Likelihoodfunktion gegen ein lokales Maximum gilt nicht.

[8] Im Rahmen dieser Arbeit bezeichnet $\boldsymbol{D}^{lm} H(\phi \mid \psi)$ bzw. $\boldsymbol{D}^{lm} Q(\phi \mid \psi)$ die gemischt-partielle Ableitung

$$\boldsymbol{D}^{lm} H(\phi \mid \psi) = \frac{\partial^{(l+m)} H(\phi \mid \psi)}{\partial \phi^l \partial \psi^m} \qquad \text{bzw.} \qquad \boldsymbol{D}^{lm} Q(\phi \mid \psi) = \frac{\partial^{(l+m)} Q(\phi \mid \psi)}{\partial \phi^l \partial \psi^m}.$$

Um sicherzustellen, daß der EM-Algorithmus gegen ein lokales Maxima konvergiert, muss eine weitere Bedingung eingeführt werden. Falls innerhalb der EM-Iterationen der Fall eintritt, daß die Schätzwerte $\hat{\boldsymbol{\beta}}^{(i)}$ einen stationären Punkt der Likelihoodfunktion darstellen, so muss im nächsten Iterationsschritt ein $\hat{\boldsymbol{\beta}}^{(i+1)}$ gefunden werden, der eine größere Kullback-Leibler-Statistik liefert. Eine Bedingung hierfür ist, daß für alle $\hat{\boldsymbol{\beta}}^{(i)} \in \Gamma^S \backslash \Gamma^L$ gilt $\sup_{\boldsymbol{\beta} \in \mathcal{B}} Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)}) > Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$. Mit dieser Bedingung ergibt sich der

Satz: *Die Kullback-Leibler-Statistik $Q(\psi \mid \phi)$ sei stetig sowohl in ψ als auch in ϕ und es gelte*

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)}) > Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \qquad \textit{für jedes } \hat{\boldsymbol{\beta}}^{(i)} \in \Gamma^S \backslash \Gamma^L. \tag{322.9}$$

Dann sind alle Konvergenzpunkte jeder Realisierung $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i \geq 0}$ eines EM-Algorithmus lokale Maxima der logarithmierten Likelihoodfunktion $L(\hat{\boldsymbol{\beta}}^{(i)})$ und $L(\hat{\boldsymbol{\beta}}^{(i)})$ konvergiert monoton (322.10) *gegen $L^* = L(\boldsymbol{\beta}^*)$ für ein lokales Maximum $\boldsymbol{\beta}^* \in \Gamma^L$.*

Die Voraussetzung (322.9) lässt sich praktisch schwer nachprüfen und ist daher überwiegend von theoretischem Interesse. Alternative Bedingungen für die Konvergenz der logarithmierten Likelihoodfunktion gegen ein lokales Maximum werden in der Literatur jedoch nicht gegeben.

Zusammenfassend lassen sich folgende Eigenschaften des EM-Algorithmus in Bezug auf die Konvergenz der logarithmierten Likelihoodfunktion festhalten:

Zur Konvergenz der logarithmierten Likelihoodfunktion im Rahmen der EM-Iterationen:

1. Ist die Kullback-Leibler-Statistik $Q(\psi \mid \phi)$ stetig in beiden Argumenten, so konvergiert die Likelihoodfunktion im Rahmen des EM-Algorithmus gegen einen stationären Punkt.
2. Gilt zusätzlich $\sup_{\boldsymbol{\beta} \in \mathcal{B}} Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)}) > Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ für jedes $\hat{\boldsymbol{\beta}}^{(i)} \in \mathcal{L} \backslash \mathcal{T}$, so konvergiert die Likelihoodfunktion gegen ein lokales Maximum.

323 Betrachtungen zur Konvergenz der Folge geschätzter Parameter

Die bisherigen Konvergenzbetrachtungen beziehen sich ausschließlich auf die Konvergenz der logarithmierten Likelihoodfunktion im Rahmen des EM-Algorithmus. Es wurde gezeigt, daß die Folge der logarithmierten Likelihoodwerte stets konvergiert und daß unter bestimmten Bedingungen Konvergenz gegen einen stationären Punkt oder gegen ein lokales Maximum der logartithmierten Likelihoodfunktion eintritt. Die Konvergenz der Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i \geq 0}$ stellt jedoch keinesfalls sicher, daß auch die Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i \geq 0}$ der Argumente der logarithmierten Likelihoodfunktion, also der Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ für die unbekannten Parameter $\boldsymbol{\beta}$ konvergiert[9]. Da das Ziel des (G) EM-Algorithmus letztlich darin besteht, optimale Schätzwerte zu bestimmen, ist die Konvergenz der Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i \geq 0}$ ebenfalls von

[9] Nach [Wu 1982] ist die umgekehrte Aussage jedoch richtig, falls $D^{10} H(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})$ stetig ist, d.h. aus der Konvergenz der Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i \geq 0}$ folgt dann die Konvergenz der Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i \geq 0}$.

Interesse. Im folgenden wird daher erläutert, unter welchen Bedingungen die Folge der Schätzungen im Rahmen des (G)EM-Algorithmus konvergiert und um welche Art Grenzwert es sich im Falle der Konvergenz handelt. Hierzu werden zunächst alle stationären Punkte bzw. lokalen Maxima der logarithmierten Likelihoodfunktion, zu denen der selbe Likelihoodwert a gehört, in einer Menge $\Gamma^S(a)$ bzw. $\Gamma^L(a)$ zusammengefasst:

$$\Gamma^S(a) := \{\boldsymbol{\beta} \in \Gamma^S : L(\boldsymbol{\beta}) = a\} \tag{323.1}$$

$$\Gamma^L(a) := \{\boldsymbol{\beta} \in \ \Gamma^L : L(\boldsymbol{\beta}) = a\} \tag{323.2}$$

Nach den Sätzen (322.6) bzw. (322.7) konvergiert $L(\hat{\boldsymbol{\beta}}^{(i)})$ im Rahmen der GEM-Iterationen unter bestimmten Bedingungen, $L(\hat{\boldsymbol{\beta}}^{(i)}) \to L(\boldsymbol{\beta}^*) = L^*$, wobei jeder Konvergenzpunkt $\hat{\boldsymbol{\beta}}^*$ in der Menge Γ^S der stationären Punkte bzw. in der Menge Γ^L der lokalen Maxima liegt. Besteht die Menge $\Gamma^S(L^*)$ bzw. die Menge $\Gamma^L(L^*)$ nur aus einem einzigen Punkt $\boldsymbol{\beta}^*$, so konvergiert die Folge unter diesen Bedingungen offensichtlich gegen diesen Punkt. Es gilt somit für die Konvergenz der Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ der aus Satz (322.6) resultierende

Satz: *Es sei $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ eine Folge von Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ für die unbekannten Parameter $\boldsymbol{\beta}$ beim GEM-Algorithmus und es sei*

1. *die Abbildung M, die die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der unbekannten Parameter der i-ten Iteration auf die Lösungsmenge $M^{(i+1)} = M(\hat{\boldsymbol{\beta}}^{(i)})$ des M-Schrittes der $(i+1)$-ten Iteration abbildet, M für $\hat{\boldsymbol{\beta}}^{(i)} \in \mathcal{B}\backslash\Gamma^S$ abgeschlossen, wobei Γ^S die Menge aller stationären Punkte der logarithmierten Likelihoodfunktion im Innern des Parameterraumes $\mathcal{B}$ bezeichnet;*
2. *$L(\hat{\boldsymbol{\beta}}^{(i+1)}) > L(\hat{\boldsymbol{\beta}}^{(i)})$ für alle $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^L$,*

so daß gemäß (322.6) die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ der logarithmierten Likelihoodwerte $L(\hat{\boldsymbol{\beta}}^{(i)})$ für $i \to \infty$ gegen den Grenzwert $L^* = L(\boldsymbol{\beta}^*)$ mit $\boldsymbol{\beta}^* \in \Gamma^S$ konvergiert.
Besteht ferner die Menge $\Gamma^S(L^*) = \{\boldsymbol{\beta} \in \Gamma^S : L(\boldsymbol{\beta}) = L^*\}$ der stationären Punkte von $L(\boldsymbol{\beta})$, deren logarithmierter Likelihoodwert gleich dem Grenzwert L^* ist, nur aus einem einzigen Element, also $\Gamma^S(L^*) = \{\boldsymbol{\beta}^*\}$, so konvergiert $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ für $i \to \infty$ gegen diesen stationären Punkt $\boldsymbol{\beta}^*$. (323.3)

Entsprechend gilt, resultierend aus Satz (322.7) für die Konvergenz der Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$, falls $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ gegen ein lokales Maximum der logarithmierten Likelihoodfunktion konvergiert

Satz: *Es sei $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ eine Folge von Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ für die unbekannten Parameter $\boldsymbol{\beta}$ beim GEM-Algorithmus und es sei*

1. *die Abbildung M, die die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der unbekannten Parameter der i-ten Iteration auf die Lösungsmenge $M^{(i+1)} = M(\hat{\boldsymbol{\beta}}^{(i)})$ des M-Schrittes der $(i+1)$-ten Iteration abbildet, für $\hat{\boldsymbol{\beta}}^{(i)} \in \mathcal{B}\backslash\Gamma^L$ abgeschlossen, wobei Γ^L die Menge aller lokalen Maxima der logarithmierten Likelihoodfunktion im Innern des Parameterraumes $\mathcal{B}$ bezeichnet;*
2. *$L(\hat{\boldsymbol{\beta}}^{(i+1)}) > L(\hat{\boldsymbol{\beta}}^{(i)})$ für alle $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^L$,*

so daß gemäß (322.7) die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ der logarithmierten Likelihoodwerte $L(\hat{\boldsymbol{\beta}}^{(i)})$ für $i \to \infty$ gegen den Grenzwert $L^* = L(\boldsymbol{\beta}^*)$ mit $\boldsymbol{\beta}^* \in \Gamma^L$ konvergiert.
Besteht weiter die Menge $\Gamma^L(L^*) = \{\boldsymbol{\beta} \in \Gamma^L : L(\boldsymbol{\beta}) = L^*\}$ der lokalen Maxima von $L(\boldsymbol{\beta})$, deren logarithmierter Likelihoodwert gleich dem Grenzwert L^* ist, nur aus einem einzigen Element, also $\Gamma^L(L^*) = \{\boldsymbol{\beta}^*\}$, so konvergiert $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ für $i \to \infty$ gegen dieses lokale Maximum $\boldsymbol{\beta}^*$.

(323.4)

Die Bedingungen, daß die Menge $\Gamma^S(L^*)$ der stationären Punkte bzw. die Menge $\Gamma^L(L^*)$ der lokalen Maxima von $L(\boldsymbol{\beta})$ nur aus einem Element bestehen, sind relativ scharf und für viele multimodale[10] Dichtefunktionen nicht erfüllt. Es können jedoch auch weniger scharfe Bedingungen formuliert werden, unter denen die Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ des (G)EM-Algorithmus konvergiert. Beispielsweise kann gezeigt werden, daß – falls Γ eine diskrete[11] Menge ist – die Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ bereits konvergiert, wenn $\lim_{i\to\infty} \|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| = 0$ gilt (vgl.[Wu 1982]). Eine entsprechende Aussage liefert der

Satz: *Es sei $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ eine Folge von Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ für die unbekannten Parameter $\boldsymbol{\beta}$ beim GEM-Algorithmus und es sei*

1. *die Abbildung M, die die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der unbekannten Parameter der i-ten Iteration auf die Lösungsmenge $M^{(i+1)} = M(\hat{\boldsymbol{\beta}}^{(i)})$ des M-Schrittes der $(i+1)$-ten Iteration abbildet, für $\hat{\boldsymbol{\beta}}^{(i)} \in \mathcal{B}\backslash\Gamma^S$ abgeschlossen, wobei Γ^S die Menge aller stationären Punkte der logarithmierten Likelihoodfunktion im Innern des Parameterraumes $\mathcal{B}$ bezeichnet;*
2. *$L(\hat{\boldsymbol{\beta}}^{(i+1)}) > L(\hat{\boldsymbol{\beta}}^{(i)})$ für alle $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^S$,*

so daß gemäß (322.6) die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ der logarithmierten Likelihoodwerte $L(\hat{\boldsymbol{\beta}}^{(i)})$ für $i \to \infty$ gegen den Grenzwert $L^* = L(\boldsymbol{\beta}^*)$ mit $\boldsymbol{\beta}^* \in \Gamma^S$ konvergiert.
Gilt weiter $\|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| \to \mathbf{0}$ für $i \to \infty$, dann liegen alle Konvergenzpunkte $\boldsymbol{\beta}^*$ von $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ in einer zusammenhängenden und kompakten Teilmenge von $\Gamma^S(L^*)$. Handelt es sich bei $\Gamma^S(L^*)$ um eine diskrete Menge, dann konvergiert $\hat{\boldsymbol{\beta}}^{(i)}$ gegen ein $\boldsymbol{\beta}^*$ in $\Gamma^S(L^*)$. (323.5)

Beweis: Hinsichtlich dieses Beweises sei auf [Wu 1982] verwiesen. □

Wie bereits aus den Vorbemerkungen zu Satz (322.8) hervorgeht ist die Abgeschlossenheitsbedingung 1.) speziell für den EM-Algorithmus erfüllt, wenn die Kullback-Leibler-Statistik $Q(\psi \mid \phi)$ sowohl in ψ als auch in ϕ stetig ist. Weiter geht aus dem Beweis zu Satz (322.8) hervor, daß beim EM-Algorithmus auch Bedingung 2.) von (323.5) erfüllt ist. Damit ergibt sich der

Satz: *Beim EM-Algorithmus sei die Kullback-Leibler-Statistik $Q(\psi \mid \phi)$ stetig sowohl in ψ als auch in ϕ. Gilt weiter $\lim_{i\to\infty} \|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| = \mathbf{0}$, dann liegen alle Konvergenzpunkte $\boldsymbol{\beta}^*$ von $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ in einer zusammenhängenden und kompakten Teilmenge von $\Gamma^S(L^*)$. Handelt es sich bei $\Gamma^S(L^*)$ um eine diskrete Menge, dann konvergiert $\hat{\boldsymbol{\beta}}^{(i)}$ gegen ein $\boldsymbol{\beta}^*$ in $\Gamma^S(L^*)$.* (323.6)

Ein dem Satz (323.5) entsprechender Satz ergibt sich wieder für die Konvergenz gegen lokale Maxima der logarithmierten Likelihoodfunktion:

Satz: *Es sei $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ eine Folge von Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ für die unbekannten Parameter $\boldsymbol{\beta}$ beim GEM-Algorithmus und es sei*

1. *die Abbildung M, die die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der unbekannten Parameter der i-ten Iteration auf die Lösungsmenge $M^{(i+1)} = M(\hat{\boldsymbol{\beta}}^{(i)})$ des M-Schrittes der $(i+1)$-ten Iteration*

[10] Eine multimodale Dichtefunktion ist eine Dichtefunktion, die durch Zusammensetzung (beispielsweise als Linearkombination) mehrerer Basisdichtefunktionen entsteht.

[11] Eine Menge wird als *diskrete* Menge bezeichnet, falls ihre Zusammenhangskomponenten jeweils nur aus genau einem Element bestehen.

abbildet, für $\hat{\boldsymbol{\beta}}^{(i)} \in \mathcal{B}\backslash\Gamma^L$ *abgeschlossen, wobei* Γ^L *die Menge aller lokalen Maxima der logarithmierten Likelihoodfunktion im Innern des Parameterraumes* $\mathcal{B}$ *bezeichnet;*

2. $L(\hat{\boldsymbol{\beta}}^{(i+1)}) > L(\hat{\boldsymbol{\beta}}^{(i)})$ *für alle* $\hat{\boldsymbol{\beta}}^{(i)} \notin \Gamma^L$,

so daß gemäß (322.6) die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ *der logarithmierten Likelihoodwerte* $L(\hat{\boldsymbol{\beta}}^{(i)})$ *für* $i \to \infty$ *gegen den Grenzwert* $L^* = L(\boldsymbol{\beta}^*)$ *mit* $\boldsymbol{\beta}^* \in \Gamma^L$ *konvergiert.*
Gilt weiter $\|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| \to \mathbf{0}$ *für* $i \to \infty$, *dann liegen alle Konvergenzpunkte* $\boldsymbol{\beta}^*$ *von* $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ *in einer zusammenhängenden und kompakten Teilmenge von* $\Gamma^L(L^*)$. *Handelt es sich bei* $\Gamma^L(L^*)$ *um eine diskrete Menge, dann konvergiert* $\hat{\boldsymbol{\beta}}^{(i)}$ *gegen ein* $\boldsymbol{\beta}^*$ *in* $\Gamma^L(L^*)$. (323.7)

An dieser Stelle soll ein weiterer Satz angeführt werden, der Bedingungen für die Konvergenz der Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ im Rahmen eines (G)EM-Algorithmus nennt. Zunächst wird dazu die Abkürzung

$$\mathcal{L}(L) = \{\boldsymbol{\beta} \in \mathcal{B} : L(\boldsymbol{\beta}) = L\} \tag{323.8}$$

eingefürt für die Menge aller Parametervektoren $\boldsymbol{\beta}$, für die die logarithmierte Likelihoodfunktion $L(\boldsymbol{\beta})$ den Wert L annimmt. Hiermit gilt dann der

Satz: *Es sei* $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ *eine mittels des GEM-Algorithmus erhaltene Folge von Schätzungen* $\hat{\boldsymbol{\beta}}^{(i)}$ *für die unbekannten Parameter* $\boldsymbol{\beta}$ *und es gelte* $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$. *Weiter sei* $\boldsymbol{D}^{10}Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ *sowohl in* $\boldsymbol{\psi}$ *als auch in* $\boldsymbol{\phi}$ *stetig.*
Dann konvergiert $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ *gegen einen stationären Punkt* $\boldsymbol{\beta}^*$ *der logarithmierten Likelihoodfunktion* $L(\boldsymbol{\beta})$ *mit* $L(\boldsymbol{\beta}^*) = \lim_{i\to\infty} L(\hat{\boldsymbol{\beta}}^{(i)}) = L^*$, *wenn* entweder

(a) *die Menge* $\mathcal{L}(L^*)$ *aus einem einzigen Element besteht, also* $\mathcal{L}(L^*) = \{\boldsymbol{\beta}^*\}$ oder (323.9)

(b) $\lim_{i\to\infty} \|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| = \mathbf{0}$ *gilt und* $\mathcal{L}(L^*)$ *diskret ist.*

BEWEIS: Zunächst muß nachgewiesen werden, daß die Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ für $i \to \infty$ konvergiert, falls entweder Bedingung (a) oder Bedingung (b) erfüllt ist.

Wegen der vorausgesetzten Stetigkeit von $\boldsymbol{D}^{10}Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ sowohl in $\boldsymbol{\psi}$ als auch in $\boldsymbol{\phi}$ folgt nach [HELFRICH I 1995] zunächst die Differenzierbarkeit von $Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ und schließlich die Stetigkeit von $Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ in $\boldsymbol{\psi}$ und $\boldsymbol{\phi}$. Wie aus dem Beweis von Satz (322.8) hervorgeht, folgt aus der Stetigkeit von Q und der Voraussetzung $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = 0$ die Konvergenz der Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ gegen einen Grenzwert L^*, so daß die Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ für $i \to \infty$ der Menge $\mathcal{L}(L^*)$ angehören müssen.[12] Besteht gemäß Bedingung (a) die Menge $\mathcal{L}(L^*)$ nur aus einem einzigen Punkt, so folgt hieraus offensichtlich die Konvergenz der Parameterfolge, $\lim_{i\to\infty} \hat{\boldsymbol{\beta}}^{(i)} = \boldsymbol{\beta}^* \in \mathcal{L}(L^*)$ gegen diesen Punkt $\boldsymbol{\beta}^*$. Sofern Bedingung (b) gilt, folgt die Konvergenz der Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ gegen (genau) ein $\boldsymbol{\beta}^* \in \mathcal{L}(L^*)$ wegen Satz (323.5).

Nachdem die Konvergenz der Parameterfolge bewiesen ist, muß nachgewiesen werden, daß es sich bei dem Grenzwert $\boldsymbol{\beta}^*$ in beiden Fällen (a) und (b) um einen stationären Punkt der logarithmierten Likelihoodfunktion handelt. Hierzu ist zu zeigen, daß $\boldsymbol{D}L(\boldsymbol{\beta}^*)$ identisch mit dem Nullvektor ist. Es gilt zunächst

$$\begin{aligned} \boldsymbol{D}L(\boldsymbol{\beta}^*) &= \lim_{i\to\infty} \boldsymbol{D}L(\hat{\boldsymbol{\beta}}^{(i+1)}) \\ &= \lim_{i\to\infty} \boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - \lim_{i\to\infty} \boldsymbol{D}^{10}H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}). \end{aligned} \tag{323.10}$$

[12] An dieser Stelle sei angemerkt, daß hierdurch keineswegs die Konvergenz der Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ nachgewiesen ist, da die Möglichkeit besteht, daß $\hat{\boldsymbol{\beta}}^{(i)}$ für $i \to \infty$ zwischen mehreren Vektoren der Menge $\mathcal{L}(L^*)$ alterniert.

Nach Voraussetzung gilt $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$, so daß wegen der vorausgesetzten Stetigkeit von $\boldsymbol{D}^{10}Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ folgt

$$\boldsymbol{D}^{10}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) = \lim_{i\to\infty} \boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}. \tag{323.11}$$

Da nach (231.11) für alle $\boldsymbol{\beta} \in \mathcal{B}$ gilt $H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \geq H(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})$, nimmt $H(\boldsymbol{\psi} \mid \hat{\boldsymbol{\beta}}^{(i)})$ für $\boldsymbol{\psi} = \hat{\boldsymbol{\beta}}^{(i)}$ ein lokales Maximum an, es gilt also $\boldsymbol{D}^{10}H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$. Damit folgt

$$\lim_{i\to\infty} \boldsymbol{D}^{10}H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \boldsymbol{D}^{10}H(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) = \mathbf{0}. \tag{323.12}$$

Einsetzen von (323.12) und (323.11) in (323.10) liefert

$$\boldsymbol{D}L(\boldsymbol{\beta}^*) = \mathbf{0},$$

womit folgt, daß $\boldsymbol{\beta}^*$ ein stationärer Punkt der logarithmierten Likelihoodfunktion $L(\boldsymbol{\beta})$ ist. Damit ist der Satz bewiesen. □

Der Vorteil von Satz (323.9) gegenüber den vergleichbaren Sätzen (323.3) und (323.5) besteht darin, daß er ohne die (in der konkreten Anwendung schwer verifizierbaren) Bedingungen 1. und 2. dieser Sätze auskommt. An die Stelle dieser Bedingungen treten die Forderung nach Stetigkeit von $\boldsymbol{D}^{10}Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ und die Bedingung $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$. Die Stetigkeitsbedingung kann in diesem Zusammenhang für eine große Anzahl von Dichtefunktionen $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ als erfüllt gelten. Insbesondere für den EM-Algorithmus ist auch die Voraussetzung $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$ erfüllt, da die Kullback-Leibler-Statistik $Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)}$ gemäß der Definition des M-Schrittes beim EM-Algorithmus für $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(i+1)}$ ein lokales Maximum der logarithmierten Likelihoodfunktion darstellt.

Offensichtlich handelt es sich bei der Menge $\Gamma^S(L^*)$ um eine Teilmenge der Menge $\mathcal{L}(L^*)$, es gilt also $\Gamma^S(L^*) \subset \mathcal{L}(L^*)$. Daher stellt die Bedingung (a), daß $\mathcal{L}(L^*)$ genau ein Element $\boldsymbol{\beta}^*$ enthält, eine schärfere Bedingung dar als die Voraussetzung des Satzes (323.3), daß $\Gamma^S(L^*)$ aus genau einem Element besteht. Ebenso ist die Forderung (b) nach Diskretheit der Menge $\mathcal{L}(L^*)$ schärfer als in (323.5) die Forderung nach Diskretheit der Menge $\Gamma^S(L^*)$. Im Vergleich zu den Sätzen (323.3) und (323.5) stellt Satz (323.9) also restriktivere Anforderungen an die Funktion $L(\boldsymbol{\beta})$.

Die Bedingung $\lim_{i\to\infty} \|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| = \mathbf{0}$ kann nur in Einzelfällen und dann auch nur mit verhältnismäßig großem Aufwand nachgewiesen werden. Daher kann (b) nur in Ausnahmefällen als Kriterium für die Konvergenz der Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ dienen. Es verbleibt in der überwiegenden Zahl der Anwendungen das Kriterium (a).

Ein wichtiger Spezialfall des Satzes (323.9) liegt vor, wenn die logarithmierte Likelihoodfunktion eine unimodale Funktion ist, d.h. wenn $L(\boldsymbol{\beta})$ nur einen einzigen stationären Punkt besitzt, der zugleich ein lokales Maximum von L ist. Dies ist beispielsweise der Fall, wenn es sich bei der Dichtefunktion $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ der vollständigen Beobachtungen um die Normalverteilung handelt. Wird in einem solchen Fall mit dem EM-Algorithmus gearbeitet, so sind die Bedingungen (a) und $Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$ von Satz (323.9) erfüllt und die Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ konvergiert gegen den stationären Punkt, also das lokale Maximum der logarithmierten Likelihoodfunktion. Somit gilt der

Satz: *Die logarithmierte Likelihoodfunktion $L(\boldsymbol{\beta})$ sei im Innern des Parameterraums $\mathcal{B}$ eine unimodale Funktion und $\boldsymbol{\beta}^*$ bezeichne den einzigen stationären Punkt. Weiter sei $\boldsymbol{D}^{10}Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ stetig in $\boldsymbol{\psi}$ und in $\boldsymbol{\phi}$. Dann konvergiert beim EM-Algorithmus die Folge*

$\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ der in den einzelnen Iterationen bestimmten Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ für die unbekannten Parameter $\boldsymbol{\beta}$ gegen den stationären Punkt $\hat{\boldsymbol{\beta}}^{*}$, dem einzigen lokalen Maximum der logarithmierten Likelihoodfunktion $L(\boldsymbol{\beta})$. (323.13)

Dieser Satz wird in der Praxis am häufigsten verwandt, um die Konvergenz der Parameterfolge gegen ein lokales Maximum der logarithmierten Likelihoodfunktion nachzuweisen.

Die bis hierhin abgeleiteten Eigenschaften des EM-Algorithmus in Bezug auf die konvergenz der Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ sind in Übersicht 3.1 zusammengefaßt.

Hierin sind vor allem die Bedingungen von Interesse, unter denen die Parameterfolge gegen ein lokales Maximum der logarithmierten Likelihoodfunktion konvergiert, da ja bei der Parameterschätzung das lokale Maximum der logarithmierten Likelihoodfunktion gesucht wird, das diese Funktion auch global maximiert. Im Allgemeinen werden mit dem EM-Algorithmus für unterschiedliche Startvektoren $\hat{\boldsymbol{\beta}}^{(0)}$ unterschiedliche Schätzungen $\boldsymbol{\beta}^{*}$ erhalten. Daher sollte der EM-Algorithmus zur Lösung eines Parameterschätzproblems stets mehrfach unter Verwendung verschiedener Startwerte durchgeführt werden, um die erhaltenen Ergebnisse einer Auswertung durch weitere Auswertungen mit unterschiedlichen Startwerten zu stützen.

324 Betrachtungen zur Konvergenzgeschwindigkeit

Wie bereits in Abschnitt 231 erläutert, besteht das Ziel des (G)EM-Algorithmus darin, Schätzwerte für die unbekannten Parameter zu erhalten, die die logarithmierte Likelihoodfunktion der unvollständigen Beobachtungsdaten maximieren. Nach Möglichkeit sollte die Schätzung eindeutig sein. In den vorangegangenen Abschnitten wurde diskutiert, unter welchen Bedingungen der EM-Algorithmus bzw. der GEM-Algorithmus zu diesem angestrebten Ziel führt: Konvergiert die Folge der logarithmierten Likelihoodwerte gegen ein lokales Maximum, so werden mittels des (G)EM-Algorithmus Schätzwerte für die unbekannten Parameter erhalten. Konvergiert darüber hinaus auch die Parameterfolge gegen ein lokales Maximum, so sind diese Schätzwerte auch eindeutig.

Neben der zunächst im Vordergrund stehenden Frage, wann der (G)EM-Algorithmus konvergiert, ist hier wie bei grundsätzlich allen iterativ ablaufenden Algorithmen von besonderem Interesse, wie rasch sich die iterativ berechneten Größen im Verlauf der Iterationen ihrem Grenzwert annähern. In der Regel entscheidet diese sogenannte *Konvergenzgeschwindigkeit* darüber, welcher Algorithmus aus einer Reihe geeigneter Algorithmen für eine bestimmte Aufgabe ausgewählt wird. In diesem Abschnitt wird daher zunächst ein Maß für die Konvergenzgeschwindigkeit des EM-Algorithmus entwickelt und schließlich werden allgemeine Aussagen über die Konvergenzgeschwindigkeit des (G)EM-Algorithmus getroffen. Die Ausführungen entsprechen denen in [Dempster et al. 1968].

Die Angabe eines Maßes für die Konvergenzgeschwindigkeit beschränkt sich auf den EM-Algorithmus, was folgenden Grund hat: Beim EM-Algorithmus müssen die neuen Schätzwerte $\hat{\boldsymbol{\beta}}^{(i+1)}$ jeder Iteration i die Kullback-Leibler-Statistik $Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ maximieren. Da dies eine sehr scharfe Anforderung an die Neuschätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ darstellt, ist in jeder Iteration die Menge der als neue Schätzung in Frage kommenden Vektoren $\boldsymbol{\beta}$ relativ klein. Somit ist ausgehend vom Startwert $\boldsymbol{\beta}^{(0)}$ die Variationsbreite möglicher Realisierungen der Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ relativ eng und in erster Näherung wird die Konvergenzgeschwindigkeit aller dieser Realisierungen die gleiche sein. Im Gegensatz dazu wird beim GEM-Algorithmus in der i-ten Iteration von der neuen Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ lediglich verlangt, daß sie keine geringere Kullback-Leibler-Statistik liefert als die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$

der vorherigen Iteration. Wegen der Schwachheit dieser Forderung kommt i.d.R. in jeder GEM-Iteration als neue Schätzung eine größere Menge von unterschiedlicher Vektoren in Frage, was dazu führt, daß die möglichen Realisierungen der Parameterfolge sich stark voneinander unterscheiden können. Für einige dieser möglicher Realisierungen werden sich höhere Konvergenzgeschwindigkeiten ergeben als für andere, so daß ein einheitliches Maß für die Konvergenzgeschwindigkeit beim GEM-Algorithmus nicht angegeben werden kann.

Das hier verwendete Maß für die Konvergenzgeschwindigkeit ist wie folgt definiert:

Definition: [KONVERGENZGESCHWINDIGKEITSMASS]
Bei einem Iterationsverfahren werde in jedem Iterationsschritt i der Vektor $\hat{\boldsymbol{\beta}}^{(i+1)}$ auf Grundlage des Ergebnisses $\hat{\boldsymbol{\beta}}^{(i)}$ einer vorherigen Iteration $(i-1)$ berechnet und die so entstehende Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ konvergiere gegen den Punkt $\boldsymbol{\beta}^*$. Dann gilt als Maß für die Konvergenzgeschwindigkeit *des Iterationsverfahrens die Größe*

$$v = \frac{\|\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\|}{\|\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\|} \tag{324.1}$$

Zur Ableitung dieses Maßes für die Konvergenzgeschwindigkeit beim EM-Algorithmus wird wieder die Abbildung A aus (310.2) betrachtet, die den Zusammenhang zwischen der neuen Schätzung $\hat{\boldsymbol{\beta}}^{(i+1)}$ der i-ten Iteration und der Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ der vorherigen Iteration beschreibt. Es gilt

$$A : \hat{\boldsymbol{\beta}}^{(i+1)} = \boldsymbol{A}(\hat{\boldsymbol{\beta}}^{(i)}) \qquad \text{mit } \boldsymbol{A}(\hat{\boldsymbol{\beta}}^{(i)}) = \left[a_1(\hat{\boldsymbol{\beta}}^{(i)}), a_2(\hat{\boldsymbol{\beta}}^{(i)}), \ldots, a_u(\hat{\boldsymbol{\beta}}^{(i)})\right]^T. \tag{324.2}$$

Es wird vorausgesetzt, daß die Folge $\{\hat{\boldsymbol{\beta}}^{(i)}\}$ der Schätzungen für die unbekannten Parameter $\boldsymbol{\beta}$ im Rahmen der EM-Iterationen gegen den Punkt $\boldsymbol{\beta}^*$ konvergiert. Durch Taylor-Entwicklung der Abbildung $\boldsymbol{A}$ im Konvergenzpunkt $\boldsymbol{\beta}^*$ erhält man

$$\boldsymbol{A}(\hat{\boldsymbol{\beta}}^{(i)}) = \boldsymbol{A}(\boldsymbol{\beta}^*) + \boldsymbol{DA}(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*) + \mathcal{O}(\|\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\|^2), \tag{324.3}$$

wobei

$$\boldsymbol{DA}(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \begin{bmatrix} \frac{\partial a_1(\boldsymbol{\beta})}{\partial \beta_1}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} & \cdots & \frac{\partial a_1(\boldsymbol{\beta})}{\partial \beta_u}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \frac{\partial a_u(\boldsymbol{\beta})}{\partial \beta_1}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} & \cdots & \frac{\partial a_u(\boldsymbol{\beta})}{\partial \beta_u}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \end{bmatrix} \tag{324.4}$$

die *Jacobimatrix* der Abbildung A im Punkt $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ bedeutet. Falls die Zahl der Iterationen gegen unendlich strebt, nähert sich wegen der Konvergenz der Parameterfolge der Punkt $\hat{\boldsymbol{\beta}}^{(i)}$ beliebig nahe dem Grenzwert $\boldsymbol{\beta}^*$, also $\hat{\boldsymbol{\beta}}^{(i)} \to \boldsymbol{\beta}^*$ für $i \to \infty$. Daher kann der letzte Term in (324.3) für genügend große i vernachlässigt werden:

$$\boldsymbol{A}(\hat{\boldsymbol{\beta}}^{(i)}) = \boldsymbol{A}(\boldsymbol{\beta}^*) + \boldsymbol{DA}(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*) \qquad \text{für } i \text{ genügend groß,}$$

so daß für genügend große i gilt

$$\underbrace{\boldsymbol{A}(\hat{\boldsymbol{\beta}}^{(i)})}_{=\hat{\boldsymbol{\beta}}^{(i+1)}} - \underbrace{\boldsymbol{A}(\boldsymbol{\beta}^*)}_{=\boldsymbol{\beta}^*} = \hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^* = \boldsymbol{DA}(\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*) \tag{324.5}$$

Offensichtlich wird die Geschwindigkeit, mit der sich die Schätzungen $\hat{\boldsymbol{\beta}}^{(i)}$ dem Grenzwert $\boldsymbol{\beta}^*$ annähern, für genügend große i maßgeblich von der Jacobi-Matrix $\boldsymbol{DA}(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \boldsymbol{DA}(\boldsymbol{\beta}^*)$ bestimmt. Für die Norm der Differenz $\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*$ ergibt sich zunächst

$$\|\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\| = \left\| \boldsymbol{DA}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*) \right\|. \tag{324.6}$$

Aus dem Ausdruck der rechten Seite dieser Gleichung soll nun die Norm $\|\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\|$ separiert werden. Hierzu muß die zur euklidschen Vektornorm zugehörige *Matrixnorm* eingeführt werden. Nach [Kerner et al. 1995] ist eine *Matrixnorm* eine auf einer Matrix definierte Norm und jede Vektornorm $\|\boldsymbol{b}_{o\times 1}\|$ besitzt eine *zugehörige Matrixnorm* $\|\boldsymbol{B}\|_M$ für quadratische Matrizen $\boldsymbol{B}$ der Dimension o,

$$\|\boldsymbol{B}\|_M = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{Bx}\|}{\|\boldsymbol{x}\|}. \tag{324.7}$$

Die der euklidschen Vektornorm zugehörige Matrixnorm wird als *Spektralnorm* bezeichnet. Für sie gilt (vgl. [Kerner et al. 1995])

$$\|\boldsymbol{B}\|_{M,euklid.} = \sqrt{\lambda_{max}(\boldsymbol{B}^T\boldsymbol{B})} =: \sqrt{\lambda_{max}^{(\boldsymbol{B}^T\boldsymbol{B})}}, \tag{324.8}$$

wobei $\lambda_{max}^{(\boldsymbol{B}^T\boldsymbol{B})}$ den größten Eigenwert der Matrix $\boldsymbol{B}^T\boldsymbol{B}$ bedeutet. Im folgenden wird bei der Matrixnorm auf die Indizes M und *euklid.* verzichtet, da im konkreten Fall zwischen der Matrix- und der Vektornorm anhand der Darstellung von Matrizen mittels Großbuchstaben und von Vektoren mittels Kleinbuchstaben unterschieden werden kann und weil hier ausschließlich die Spektralnorm verwendet wird. Wegen der Definition des *Supremums* als kleinste obere Schranke gilt

$$\|\boldsymbol{B}\| = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{Bx}\|}{\|\boldsymbol{x}\|} \geq \frac{\|\boldsymbol{Bx}\|}{\|\boldsymbol{x}\|} \quad \text{und damit} \quad \|\boldsymbol{B}\|\cdot\|\boldsymbol{x}\| = \sqrt{\lambda_{max}^{(\boldsymbol{B}^T\boldsymbol{B})}}\cdot\|\boldsymbol{x}\| \geq \|\boldsymbol{Bx}\|. \tag{324.9}$$

Überträgt man dieses Ergebnis auf Gl. (324.6), so ergibt sich

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\| &= \left\| \boldsymbol{DA}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*) \right\| \\ &\leq \|\boldsymbol{DA}(\boldsymbol{\beta}^*)\| \cdot \|\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\| = \sqrt{\lambda_{max}^{([\boldsymbol{DA}(\boldsymbol{\beta}^*)]^T\boldsymbol{DA}(\boldsymbol{\beta}^*))}} \cdot \|\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\|\end{aligned} \tag{324.10}$$

und als Maß für die Konvergenzgeschwindigkeit des EM-Algorithmus erhält man die Quadratwurzel des größten Eigenvektors der Matrix $\boldsymbol{DA}(\boldsymbol{\beta}^*)$,

$$v = \frac{\|\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\|}{\|\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\|} = \sqrt{\lambda_{max}^{([\boldsymbol{DA}(\boldsymbol{\beta}^*)]^T\boldsymbol{DA}(\boldsymbol{\beta}^*))}}. \tag{324.11}$$

(Vergleiche hierzu [Dempster et al. 1968] und [Hornegger 1996].)[13]
Es es gilt also mit $\sqrt{\lambda_{max}^{([\boldsymbol{DA}(\boldsymbol{\beta}^*)]^T\boldsymbol{DA}(\boldsymbol{\beta}^*))}} = \lambda_{max}^{\boldsymbol{DA}(\boldsymbol{\beta}^*)}$ die Beziehung

$$v = \lambda_{max}^{\boldsymbol{DA}(\boldsymbol{\beta}^*)}. \tag{324.12}$$

[13]Im Rahmen der Literaturrecherche zur vorliegenden Arbeit wurde weder der Nachweis dieser Aussage noch ein Gegenbeispiel hierzu gefunden. Auf die Richtigkeit der in [Dempster et al. 1968] und [Hornegger 1996] getroffenen Aussage, daß als Maß für die Konvergenzgeschwindigkeit der größte Eigenwert der Matrix $\boldsymbol{DA}(\boldsymbol{\beta}^*)$ wird dennoch vertraut.

Ist ein Eigenwert der Matrix $\boldsymbol{DA}(\boldsymbol{\beta}^*)$ größer als eins, so kann es sich nach den Ausführungen in [DEMPSTER ET AL. 1968] bei $\boldsymbol{\beta}^*$ um einen Sattelpunkt der logarithmierten Likelihoodfunktion handeln: Wird im Rahmen der EM-Iterationen für einen Startwert $\hat{\boldsymbol{\beta}}^{(0)}$ als Schätzung ein Sattelpunkt $\hat{\boldsymbol{\beta}}^*$ der logarithmierten Likelihoodfunktion erhalten so bleibt die Schätzung $\hat{\boldsymbol{\beta}}$ in nachfolgenden Iterationen in diesem Sattelpunkt. Wird jedoch eine Schätzung $\boldsymbol{\beta}'$ erhalten, die auch nur minimal von $\boldsymbol{\beta}^*$ abweicht, so bewegen sich nachfolgende Schätzungen wegen (324.11) vom Sattelpunkt fort.
Hat $\boldsymbol{DA}(\boldsymbol{\beta}^*)$ einen mit eins identischen Eigenwert, so weist dies darauf hin, daß $\boldsymbol{\beta}^*$ auf einer Kurve im $\mathbb{R}^u$ liegt, deren Punkte allesamt Maxima der logarithmierten Likelihoodfunktion darstellen (vgl. [DEMPSTER ET AL. 1968]).

Der nachfolgende Satz gibt Aufschluß darüber, wie die Matrix $\boldsymbol{DA}(\boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ praktisch berechnet werden kann:

Satz: *Es sei* $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ *die Parameterfolge eines (G)EM-Algorithmus mit den Eigenschaften*

1.) $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ *konvergiert gegen einen Punkt und* $\boldsymbol{\beta}^*$ *im Innern des Parameterraums* $\mathcal{B}$

2.) $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$

Dann gilt

$$\boldsymbol{D}L(\boldsymbol{\beta}^*) = \mathbf{0}, \tag{324.13}$$

und

$$\boldsymbol{DA}(\hat{\boldsymbol{\beta}}^*) = \left[\boldsymbol{D}^{20}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\right]^{-1} \boldsymbol{D}^{20}H(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*). \tag{324.14}$$

BEWEIS: Nach (231.7) gilt $L(\hat{\boldsymbol{\beta}}^{(i+1)}) = Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$. Somit gilt für den Gradienten $\boldsymbol{D}L(\hat{\boldsymbol{\beta}}^{(i+1)})$ der logarithmierten Likelihoodfunktion in Abhängigkeit von $\hat{\boldsymbol{\beta}}^{(i+1)}$

$$\begin{aligned}\boldsymbol{D}L(\hat{\boldsymbol{\beta}}^{(i+1)}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}^{(i+1)}} &= \frac{\partial Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)}}{\partial \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}^{(i+1)}} - \frac{\partial H(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(i)})}{\partial \boldsymbol{\beta}} \\ &= \underbrace{\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})}_{=0 \text{ nach Vorauss.2.)}} - \boldsymbol{D}^{10}H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \\ &= -\boldsymbol{D}^{10}H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \end{aligned} \tag{324.15}$$

und nach Übergang auf den Grenzwert für $i \to \infty$, also für unendlich viele Iterationen folgt

$$\boldsymbol{D}L(\beta^*) = \lim_{i\to\infty} \boldsymbol{D}L(\hat{\boldsymbol{\beta}}^{(i+1)}) = -\lim_{i\to\infty} \boldsymbol{D}^{10}H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}, \tag{324.16}$$

da nach (323.12) $\lim_{i\to\infty} \boldsymbol{D}^{10}H(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$ gilt. Somit folgt die erste Aussage $\boldsymbol{D}L(\beta^*) = 0$.
Zum Beweis der zweiten Aussage wird die von den beiden Vektoren $\hat{\boldsymbol{\beta}}^{(i)}$ und $\hat{\boldsymbol{\beta}}^{(i+1)}$ abhängige Größe $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ in einer Taylorreihe um $\boldsymbol{\beta}^*$ entwickelt. Bekanntlich gilt für eine vom Vektor $\boldsymbol{u}$ abhängige skalarwertige Funktion $b(\boldsymbol{u})$ (vgl. [HELFRICH II 1996])

$$b(\boldsymbol{u}_0 + \boldsymbol{h}) = b(\boldsymbol{u}_0) + \boldsymbol{h}^T \nabla b(\boldsymbol{u})|_{\boldsymbol{u}_0} + \frac{1}{2}\boldsymbol{h}^T H(\boldsymbol{u})|_{\boldsymbol{u}_0+\theta\boldsymbol{h}}\boldsymbol{h} \qquad \text{mit } 0 \leq \theta \leq 1, \tag{324.17}$$

wobei $\nabla b(\boldsymbol{u})|_{\boldsymbol{u}_0}$ den Gradienten der Funktion b an der Stelle $\boldsymbol{u}_0$ und $H(\boldsymbol{u})|_{\boldsymbol{u}_0+\theta\boldsymbol{h}}$ die *Hesse-Matrix* der Funktion b für einen Punkt $\boldsymbol{u}_0 + \theta\boldsymbol{h}$ auf der geradlinigen Verbindung zwischen

$\boldsymbol{u}_0$ und $\boldsymbol{u}_0 + \boldsymbol{h}$ bedeuten.
Setzt sich der Vektor $\boldsymbol{u}$ aus den beiden Vektoren $\boldsymbol{u}_1$ und $\boldsymbol{u}_2$ zusammen, also $\boldsymbol{u} = \left(\boldsymbol{u}_1^T, \boldsymbol{u}_2^T\right)^T$ und entsprechend $\boldsymbol{u}_0 = \left(\boldsymbol{u}_{10}^T, \boldsymbol{u}_{20}^T\right)^T$ und $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2)^T$, so folgt hieraus

$$b(\begin{bmatrix}\boldsymbol{u}_{10}\\ \boldsymbol{u}_{20}\end{bmatrix} + \begin{bmatrix}\boldsymbol{h}_1\\ \boldsymbol{h}_2\end{bmatrix}) = b(\begin{bmatrix}\boldsymbol{u}_{10}\\ \boldsymbol{u}_{20}\end{bmatrix}) + \begin{bmatrix}\boldsymbol{h}_1\\ \boldsymbol{h}_2\end{bmatrix}^T \nabla b(\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{bmatrix})\Big|_{\left[\begin{smallmatrix}\boldsymbol{u}_{10}\\ \boldsymbol{u}_{20}\end{smallmatrix}\right]} + \frac{1}{2}\begin{bmatrix}\boldsymbol{h}_1\\ \boldsymbol{h}_2\end{bmatrix}^T H(\begin{bmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{bmatrix})\Big|_{\left[\begin{smallmatrix}\boldsymbol{u}_{10}\\ \boldsymbol{u}_{20}\end{smallmatrix}\right]+\theta\left[\begin{smallmatrix}\boldsymbol{h}_1\\ \boldsymbol{h}_2\end{smallmatrix}\right]} \begin{bmatrix}\boldsymbol{h}_1\\ \boldsymbol{h}_2\end{bmatrix}. \quad (324.18)$$

Angewendet auf die k-te Komponente $q_k^{10}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ des Vektors $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \left(q_1^{10}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}), q_2^{10}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}), \ldots, q_u^{10}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})\right)^T$ liefert diese Gleichung

$$q_k^{10}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = q_k^{10}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) + \begin{bmatrix}\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\\ \hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\end{bmatrix}^T \begin{bmatrix}\frac{\partial q_k^{10}(\hat{\boldsymbol{\beta}}^{(i+1)}|\hat{\boldsymbol{\beta}}^{(i)})}{\partial \hat{\boldsymbol{\beta}}^{(i+1)}}\Big|_{(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)}\\ \frac{\partial q_k^{10}(\hat{\boldsymbol{\beta}}^{(i+1)}|\hat{\boldsymbol{\beta}}^{(i)})}{\partial \hat{\boldsymbol{\beta}}^{(i)}}\Big|_{(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)}\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\\ \hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\end{bmatrix}^T \boldsymbol{H}_k \begin{bmatrix}\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\\ \hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\end{bmatrix} \quad (324.19)$$

wobei die konkrete Gestalt der Hessematrix $\boldsymbol{H}_k$ für die weiteren Überlegungen irrelevant ist. Nach Voraussetzung 1.) verschwindet in dieser Gleichung die linke Seite und wegen $\boldsymbol{D}^{10}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) = \lim_{i\to\infty} \boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \boldsymbol{0}$ auch der erste Term der rechten Seite. Mit den Abkürzungen

$$\frac{\partial q_k^{10}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})}{\partial \hat{\boldsymbol{\beta}}^{(i+1)}} =: \boldsymbol{q}_k^{20}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) \text{ und } \frac{\partial q_k^{10}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})}{\partial \hat{\boldsymbol{\beta}}^{(i)}} =: \boldsymbol{q}_k^{11}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}),$$

folgt

$$0 = \begin{bmatrix}\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\\ \hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\end{bmatrix}^T \begin{bmatrix}\boldsymbol{q}_k^{20}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\\ \boldsymbol{q}_k^{11}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\\ \hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\end{bmatrix}^T \boldsymbol{H}_k \begin{bmatrix}\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\\ \hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\end{bmatrix} \quad (324.20)$$

Wegen $\hat{\boldsymbol{\beta}}^{(i+1)} = \boldsymbol{A}(\hat{\boldsymbol{\beta}}^{(i)})$ und $\boldsymbol{\beta}^* = \boldsymbol{A}(\boldsymbol{\beta}^*)$ gilt

$$\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^* = \boldsymbol{A}(\hat{\boldsymbol{\beta}}^{(i)}) - \boldsymbol{A}(\boldsymbol{\beta}^*)$$

und mit der Taylor-Entwicklung der Funktion $\boldsymbol{A}$

$$\boldsymbol{A}(\hat{\boldsymbol{\beta}}^{(i)}) = \boldsymbol{A}(\boldsymbol{\beta}^*) + \underbrace{\frac{\partial \boldsymbol{A}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}^*}}_{=\boldsymbol{D}\boldsymbol{A}(\boldsymbol{\beta}^*)} (\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*)$$

unter Vernachlässigung von Termen der Ordnung 2 und größer gilt

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*\\ \hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*\end{bmatrix} = \begin{bmatrix}\boldsymbol{D}\boldsymbol{A}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*)\\ (\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*)\end{bmatrix} = \begin{bmatrix}\boldsymbol{D}\boldsymbol{A}(\boldsymbol{\beta}^*)\\ \boldsymbol{I}\end{bmatrix} (\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*).$$

Einsetzen in (324.20) liefert

$$0 = (\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*)^T \begin{bmatrix}\boldsymbol{D}\boldsymbol{A}(\boldsymbol{\beta}^*)\\ \boldsymbol{I}\end{bmatrix}^T \begin{bmatrix}\boldsymbol{q}_k^{20}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\\ \boldsymbol{q}_k^{11}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\end{bmatrix} + \frac{1}{2}(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*)^T \underbrace{\begin{bmatrix}\boldsymbol{D}\boldsymbol{A}(\boldsymbol{\beta}^*)\\ \boldsymbol{I}\end{bmatrix}^T \boldsymbol{H}_k \begin{bmatrix}\boldsymbol{D}\boldsymbol{A}(\boldsymbol{\beta}^*)\\ \boldsymbol{I}\end{bmatrix}}_{:=\boldsymbol{C}} (\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*) \quad (324.21)$$

und nach Ausklammern von $(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*)^T$ gilt

$$0 = (\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*)^T \left\{ \begin{bmatrix} \boldsymbol{DA}(\boldsymbol{\beta}^*) \\ \boldsymbol{I} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{q}_k^{20}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) \\ \boldsymbol{q}_k^{11}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) \end{bmatrix} + \frac{1}{2}\boldsymbol{C}(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*) \right\} \tag{324.22}$$

Da $\hat{\boldsymbol{\beta}}^{(i+1)} - \boldsymbol{\beta}^*$ i.a. nicht der Nullvektor ist, muß der Ausdruck innerhalb der Klammer {} der Nullvektor sein. Es folgt also

$$\mathbf{0} = \begin{bmatrix} \boldsymbol{DA}(\boldsymbol{\beta}^*) \\ \boldsymbol{I} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{q}_k^{20}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) \\ \boldsymbol{q}_k^{11}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) \end{bmatrix} + \frac{1}{2}\boldsymbol{C}(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*) \tag{324.23}$$

und durch Übergang auf den Grenzwert für $i \to \infty$, bei dem der zweite Term wegen $\lim_{i\to\infty}(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^*) = 0$ verschwindet

$$\mathbf{0} = \begin{bmatrix} \boldsymbol{DA}(\boldsymbol{\beta}^*) \\ \boldsymbol{I} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{q}_k^{20}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) \\ \boldsymbol{q}_k^{11}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) \end{bmatrix} = \boldsymbol{DA}(\boldsymbol{\beta}^*)^T \boldsymbol{q}_k^{20}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) + \boldsymbol{q}_k^{11}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*). \tag{324.24}$$

Hieraus ergibt sich

$$\mathbf{0} = \boldsymbol{DA}(\boldsymbol{\beta}^*)^T \left[\boldsymbol{q}_1^{20}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*), \ldots, \boldsymbol{q}_u^{20}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\right] + \left[\boldsymbol{q}_1^{11}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*), \ldots, \boldsymbol{q}_u^{11}(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\right],$$

wobei $\mathbf{0}$ die Nullmatrix bedeutet. Mit den Jacobi-Matrizen

$$\boldsymbol{D}^{20}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \begin{bmatrix} \boldsymbol{q}_1^{20}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})^T \\ \cdots \\ \boldsymbol{q}_u^{20}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})^T \end{bmatrix} \tag{324.25}$$

bzw.

$$\boldsymbol{D}^{11}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \begin{bmatrix} \boldsymbol{q}_1^{11}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})^T \\ \cdots \\ \boldsymbol{q}_u^{11}(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})^T \end{bmatrix}, \tag{324.26}$$

die die 1. Ableitungen des Gradienten $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)})$ der Kullback-Leibler-Statistik nach den Elementen des ersten Argumentes $\hat{\boldsymbol{\beta}}^{(i+1)}$ bzw. nach den Elementen des 2. Argumentes $\hat{\boldsymbol{\beta}}^{(i)}$ entalten, ergibt sich

$$\mathbf{0} = \boldsymbol{DA}(\boldsymbol{\beta}^*)^T \boldsymbol{D}^{20}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)^T + \boldsymbol{D}^{11}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)^T \tag{324.27}$$

und daraus

$$\mathbf{0} = \boldsymbol{D}^{20}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\boldsymbol{DA}(\boldsymbol{\beta}^*) + \boldsymbol{D}^{11}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*). \tag{324.28}$$

Schließlich folgt

$$\boldsymbol{DA}(\boldsymbol{\beta}^*) = -\left[\boldsymbol{D}^{20}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\right]^{-1} \boldsymbol{D}^{11}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) \tag{324.29}$$

Nach (321.5) gilt $L(\hat{\boldsymbol{\beta}}^{(i)}) = Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = Q(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i-1)}) - H(\hat{\boldsymbol{\beta}}^{(i)} \mid \hat{\boldsymbol{\beta}}^{(i-1)})$, so daß in der i-ten Iteration $L(\hat{\boldsymbol{\beta}}^{(i)})$ unabhängig von der vorherigen Schätzung $\boldsymbol{\beta}^{(i-1)}$ ist. Es gilt also $\boldsymbol{D}^{11}L(\hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$ und aus $L(\boldsymbol{\beta}^*) = Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) - H(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)$ folgt

$$\mathbf{0} = \boldsymbol{D}^{11}L(\boldsymbol{\beta}^*) = \boldsymbol{D}^{11}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) - \boldsymbol{D}^{11}H(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*), \tag{324.30}$$

so daß gilt

$$\boldsymbol{D}^{11}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) = \boldsymbol{D}^{11}H(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*. \tag{324.31}$$

Für das Element $\left[\boldsymbol{D}^{11}H(\boldsymbol{\psi} \mid \boldsymbol{\phi})\right]_{kl}$ der k-ten Zeile und der l-ten Spalte der Matrix $\boldsymbol{D}^{11}H(\boldsymbol{\psi} \mid \boldsymbol{\phi}$ gilt mit $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_u)^T$ und $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_u)^T$

$$\begin{aligned}
\left[\boldsymbol{D}^{11}H(\boldsymbol{\psi} \mid \boldsymbol{\phi})\right]_{kl} &= \frac{\partial}{\partial\psi_k}\left(\frac{\partial}{\partial\phi_l}\int_{\mathcal{Z}(\boldsymbol{y})}\cdots\int \log f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi}) f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\phi}) d\boldsymbol{z}\right) \\
&= \frac{\partial}{\partial\psi_k}\int_{\mathcal{Z}(\boldsymbol{y})}\cdots\int \log f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi}) \frac{\partial f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\phi})}{\partial\phi_l} d\boldsymbol{z} \\
&= \int_{\mathcal{Z}(\boldsymbol{y})}\cdots\int \frac{1}{f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})} \frac{\partial f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})}{\partial\psi_l} \frac{\partial f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\phi})}{\partial\phi_k} d\boldsymbol{z}. \qquad (324.32)
\end{aligned}$$

Entsprechend gilt für das Element $\left[\boldsymbol{D}^{20}H(\boldsymbol{\psi} \mid \boldsymbol{\phi})\right]_{kl}$ der k-ten Zeile und der l-ten Spalte der Matrix $\boldsymbol{D}^{20}H(\boldsymbol{\psi} \mid \boldsymbol{\phi})$

$$\begin{aligned}
\left[\boldsymbol{D}^{20}H(\boldsymbol{\psi} \mid \boldsymbol{\phi})\right]_{kl} &= \frac{\partial}{\partial\psi_k}\left(\frac{\partial}{\partial\psi_l}\int_{\mathcal{Z}(\boldsymbol{y})}\cdots\int \log f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi}) f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\phi}) d\boldsymbol{z}\right) \\
&= \frac{\partial}{\partial\psi_k}\int_{\mathcal{Z}(\boldsymbol{y})}\cdots\int \frac{1}{f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})} \frac{\partial f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})}{\partial\psi_l} f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\phi}) d\boldsymbol{z} \\
&= \int_{\mathcal{Z}(\boldsymbol{y})}\cdots\int -\frac{1}{f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})^2} \frac{\partial f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})}{\partial\psi_k} \frac{\partial f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})}{\partial\psi_k} f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\phi}) d\boldsymbol{z} \\
&\qquad + \underbrace{\int_{\mathcal{Z}(\boldsymbol{y})}\cdots\int \frac{1}{f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})} \frac{\partial^2 f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})}{\partial\psi_k\partial\psi_l} f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\phi}) d\boldsymbol{z}}_{= \frac{\partial^2}{\partial\psi_k\partial\psi_l}\int\cdots\int_{\mathcal{Z}(\boldsymbol{y})} f(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\psi}) d\boldsymbol{z} = \frac{\partial^2}{\partial\psi_k\partial\psi_l} 1 = 0} \\
&= -\int_{\mathcal{Z}(\boldsymbol{y})}\cdots\int \frac{1}{f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})} \frac{\partial f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})}{\partial\psi_k} \frac{\partial f(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi})}{\partial\psi_k} d\boldsymbol{z}. \qquad (324.33)
\end{aligned}$$

Aus (324.32) und (324.33) folgt

$$\boldsymbol{D}^{11}H(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) = -\boldsymbol{D}^{20}H(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*) \qquad (324.34)$$

Durch Einsetzen von (324.31) und (324.34) in (324.29) erhält man schließlich

$$\boldsymbol{DA}(\boldsymbol{\beta}^*) = \left[\boldsymbol{D}^{20}Q(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)\right]^{-1} \boldsymbol{D}^{20}H(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^*)$$

und damit die zweite Aussage. □

In der Literatur zum (G)EM-Algorithmus wird dem EM-Algorithmus eine relativ langsame Konvergenz zugeschrieben (vgl. [Dempster et al. 1968], [Hornegger 1996], [Redner 1984]. Wie in [Dempster et al. 1968] angedeutet, kann gezeigt werden, das der EM-Algorithmus um so schneller konvergiert, je weniger fehlende Beobachtungen in die Ausgleichung eingeführt werden. Der GEM-Alorithmus konvergiert wegen der weniger strengen Anforderungen an die Schätzung $\hat{\boldsymbol{\beta}}^{(i)}$ im M-Schritt noch langsamer als der EM-Algorithmus. Besonders dann, wenn sich die Kullback-Leibler-Statistik eines Parameterschätzproblems nicht analytisch angeben lässt, so das sie im Rahmen der EM-Iterationen mittels numerischer Verfahren berechnet werden muß, ergibt sich ein enormer Rechenaufwand. Insbesondere in einem solchen Fall lohnt es sich vor dem Einsatz des

EM-Algorithmus sorgfältig zu prüfen, ob das gewünschte Ergebnis nicht schneller mit einem anderen Algorithmus erhalten werden kann.
Oft kann die Kullback-Leibler-Statistik einer Parameterschätzung aus unvollständigen Daten jedoch analytisch angegeben werden, so daß sich der im Zusammenhang mit dem EM-Algorithmus anfallende Rechenaufwand stark reduziert. Es bleibt allerdings dann immer noch die Tatsache, daß der EM-Algorithmus i.d.R. aufgrund seiner langsamen Konvergenz sehr viele EM-Iterationen erforderlich macht.

33 Zusammenfassung

In diesem Kapitel wurden die allgemeinen Eigenschaften des EM-Algorithmus diskutiert. Ausgehend von der Betrachtung des Verhaltens der logarithmierten Likelihoodfunktion im Rahmen der EM-Iterationen und dem Nachweis ihrer Konvergenz wurde erläutert, unter welchen Bedingungen die logarithmierte Likelihoodfunktion im Rahmen der (G)EM-Iterationen wie gewünscht maximiert wird bzw. unter welchen Bedingungen Konvergenz gegen einen stationären Punkt erfolgt. Im Anschluß daran wurden Bedingungen genannt, unter denen die Parameterfolge gegen einen stationären Punkt bzw. gegen ein lokales Maximum der logarithmierten Likelihoodfunktion konvergiert, sofern die Folge der Likelihoodwerte konvergiert. Zum Abschluß des Kapitels erfolgten einige Bemerkungen zur Konvergenzgeschwindigkeit des (G)EM-Algorithmus, wobei insbesondere auf die Langsamkeit des Verfahrens hingewiesen wurde.

Im folgenden Kapitel wird der EM-Algorithmus nun im Zusammenhang mit einer konkreten Aufgabe angewandt, um zu zeigen, wie der Algorithmus zur Parameterschätzung und Klassifikation eingesetzt werden kann.

Übersicht 3.1: **Zur Konvergenz der Parameterfolge:**

1. Konvergiert die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ beim (G)EM-Algorithmus gemäß Satz (322.6) für $i \to \infty$ gegen L^*, den Funktionswert $L^* = L(\boldsymbol{\beta})$ eines stationären Punktes $\boldsymbol{\beta} \in \Gamma^S$ von $L(\boldsymbol{\beta})$, dann gilt für das Konvergenzverhalten der Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ für $i \to \infty$:

   (a) Besteht die Menge $\Gamma^S(L^*) = \{\boldsymbol{\beta} \in \Gamma^S : L(\boldsymbol{\beta}) = L^*\}$ der stationären Punkte $\boldsymbol{\beta}$ der logarithmierten Likelihoodfunktion mit $L(\boldsymbol{\beta}) = L^*$ aus genau einem Punkt $\boldsymbol{\beta}^*$, dann konvergiert $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ gegen diesen Punkt, $\lim_{i\to\infty} \hat{\boldsymbol{\beta}}^{(i)} = \hat{\boldsymbol{\beta}}^*$.

   (b) Falls $\|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| \to \mathbf{0}$ für $i \to \infty$ gilt und $\Gamma^S(L^*)$ eine diskrete Menge ist, so konvergiert $\hat{\boldsymbol{\beta}}^{(i)}$ gegen ein $\boldsymbol{\beta}^*$ in $\Gamma^S(L^*)$. Ist $\Gamma^S(L^*)$ nicht diskret, dann liegen zumindest alle Konvergenzpunkte $\boldsymbol{\beta}^*$ von $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ in einer zusammenhängenden und kompakten Teilmenge von $\Gamma^S(L^*)$.

2. Konvergiert die Folge $\{L(\hat{\boldsymbol{\beta}}^{(i)})\}_{i\geq 0}$ beim (G)EM-Algorithmus gemäß Satz (322.7) für $i \to \infty$ gegen L^*, den Funktionswert $L^* = L(\boldsymbol{\beta})$ eines lokalen Maximums $\boldsymbol{\beta} \in \Gamma^L$ von $L(\boldsymbol{\beta})$, dann konvergiert die Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ für $i \to \infty$ gegen ein lokales Maximum $\boldsymbol{\beta}^*$ von $L(\boldsymbol{\beta})$, falls eine der folgenden Bedingungen erfüllt ist:

   (a) Die Menge $\Gamma^L(L^*) = \{\boldsymbol{\beta} \in \Gamma^S : L(\boldsymbol{\beta}) = L^*\}$ der lokalen Maxima $\boldsymbol{\beta}$ mit $L(\boldsymbol{\beta}) = L^*$ besteht nur aus einem einzigen Punkt $\boldsymbol{\beta}^*$. $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ konvergiert dann gegen diesen Punkt, $\lim_{i\to\infty} \hat{\boldsymbol{\beta}}^{(i)} = \hat{\boldsymbol{\beta}}^*$.

   (b) Falls $\|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| \to \mathbf{0}$ für $i \to \infty$ gilt und $\Gamma^L(L^*)$ eine diskrete Menge ist, so konvergiert $\hat{\boldsymbol{\beta}}^{(i)}$ gegen ein $\boldsymbol{\beta}^*$ in $\Gamma^L(L^*)$. Ist $\Gamma^L(L^*)$ nicht diskret, dann liegen zumindest alle Konvergenzpunkte $\boldsymbol{\beta}^*$ von $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ in einer zusammenhängenden und kompakten Teilmenge von $\Gamma^L(L^*)$.

3. Die Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ des (G)EM-Algorithmus konvergiert gegen einen stationären Punkt $\boldsymbol{\beta}^*$ der logarithmierten Likelihoodfunktion, wenn in jeder Iteration $\boldsymbol{D}^{10}Q(\hat{\boldsymbol{\beta}}^{(i+1)} \mid \hat{\boldsymbol{\beta}}^{(i)}) = \mathbf{0}$ gilt und $\boldsymbol{D}^{10}Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ in $\boldsymbol{\psi}$ und $\boldsymbol{\phi}$ stetig ist und zugleich entweder

   (a) die Menge $\mathcal{L}(L^*)$ aus einem einzigen Element besteht, also $\mathcal{L}(L^*) = \{\boldsymbol{\beta}^*\}$ oder

   (b) $\lim_{i\to\infty} \|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| = \mathbf{0}$ gilt und $\mathcal{L}(L^*)$ diskret ist.

Speziell für den EM-Algorithmus gilt:

4. Die Parameterfolge $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ des EM-Algorithmus konvergiert gegen einen stationären Punkt $\boldsymbol{\beta}^*$, falls die Kullback-Leibler-Statistik $Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ in $\boldsymbol{\psi}$ und $\boldsymbol{\phi}$ stetig ist, $\lim_{i\to\infty} \|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\| = \mathbf{0}$ gilt und $\Gamma^S(L^*)$ eine diskrete Menge ist. Unabhängig von der Diskretheit von $\Gamma^S(L^*)$ liegen liegen alle Konvergenzpunkte $\boldsymbol{\beta}^*$ von $\{\hat{\boldsymbol{\beta}}^{(i)}\}_{i\geq 0}$ in einer zusammenhängenden und kompakten Teilmenge von $\Gamma^S(L^*)$.

5. Ist $L(\boldsymbol{\beta})$ im Innern des Parameterraums $\mathcal{B}$ eine unimodale Funktion mit dem lokalen Maximum $\boldsymbol{\beta}^*$ und ist $\boldsymbol{D}^{10}Q(\boldsymbol{\psi} \mid \boldsymbol{\phi})$ stetig in $\boldsymbol{\psi}$ und $\boldsymbol{\phi}$, so konvergiert die Parameterfolge gegen das lokale Maximum.

Kapitel 4

Beispiel: Parameterschätzung und Klassifikation

Im folgenden soll anhand des Bereits in der Einleitung skizzierten, sehr einfachen Beispiels aufgezeigt werden, wie der EM-Algorithumus zur Parameterschätzung bei gleichzeitiger Lösung eines Zuordnungsproblems eingesetzt werden kann.

41 Aufgabe: Bestimmung der Parameter zweier ausgleichender Kurven in einem Bild

Gegeben sei ein CCD-Bild (d.i. ein Bild im Rasterformat), in dem zwei linienhafte Objekte $\mathcal{O}_1$ und $\mathcal{O}_2$ – hier eine Hochspannungsleitung und eine Straße – abgebildet sind (vgl. Abb. 41a). Zur Überführung des Bildes in ein Vektorformat seien die Bildkoordinaten u_i und y_i einzelner Punkte auf den Mittellinien[1] der beiden Objekte bestimmt und in der Datei `Linien.dat` abgespeichert worden. Als Beobachtungsmaterial liegt also eine Meßreihe der Form

u_i	u_1	u_2	$\dots$	u_N
y_i	y_1	y_2	$\cdots$	y_N

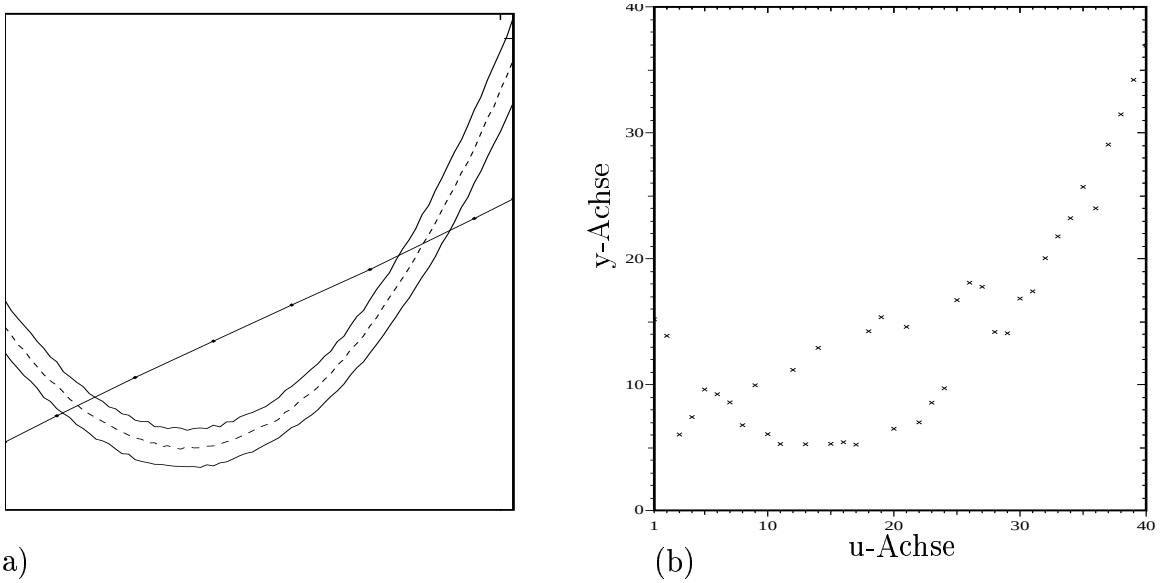
(410.1)

vor. In Abb. 41b sind die einzelnen Objektpunkte[2] in einem Koordinatensystem dargestellt.

Bei den Koordinatenmessungen seien die N Objektpunkte in willkürlicher Reihenfolge angemessen worden, so daß in der Datei `Linien.dat` die zu Objekt $\mathcal{O}_1$ und $\mathcal{O}_2$ gehörenden Koordinatenpaare zufällig aufeinander folgen. (Das bedeutet, daß es nicht möglich ist, aus der Reihenfolge der einzelnen Koordinatenpaare in der Koordinatenliste auf deren Zugehörigkeit zu einem der beiden Objekte zu schließen.) Bei den Messungen sei es weiter versäumt worden, festzuhalten, welches gemessene Koordinatenpaar zu welchem Objekt gehört. Für die weitere Auswertung ist demnach zwar bekannt, daß jedes gemessene Koordinatenpaar zu einem der beiden Objekte $\mathcal{O}_1$ (Hochspannungsleitung) oder $\mathcal{O}_2$ (Straße) gehört, es geht aus dem Datenmaterial jedoch nicht hervor, zu welchem von beiden (vgl. Anhang A2).

[1] Im Falle der Straße sind dies Punkte auf dem Mittelstreifen und im Falle der Hochspannungsleitung Punkte auf dem mittleren Leitungskabel.

[2] Wenn im folgenden von Objektpunkten oder Punkten von Objekten die Rede ist, so sind damit jeweils die Punkte auf den Mittellinien der beiden Objekte gemeint.



(a) (b)

Abbildung 1: (a) Bild mit zwei linienhaften Obekten (Straße und Hochspannungsleitung). (b) Gemessene Punkte auf den Mittellinien der beiden Objekte. Zur Visualisierung der Messunsicherheiten wurde die Messwertstreuung in der Abbildung extrem groß gewählt, was insbesondere aus den zu der Hochspannungsleitung gehörenden Punkten ersichtlich ist.

Die *Aufgabe* besteht nun darin, eine Software zu entwickeln, die nach Vorgabe zweier Beobachtungsmodelle,[3] mit denen die Mittellinien der beiden Objekte $\mathcal{O}_1$ und $\mathcal{O}_2$ im Bild parametrisiert werden können, dazu in der Lage ist, nur aufgrund der in der Datei `Linien.dat` abgelegten Daten automatisch diejenigen Prameter der beiden Modelle zu schätzen, mit denen jeweils die optimale Approximation der Mittellinie der Hochspannungsleitung bzw. der Straße erreicht wird. Hierbei muß das Problem der Zuordnung der einzelnen Koordinatenpaare zu einem der beiden Objekte gelöst werden.

Um die nachfolgenden Betrachtungen möglichst einfach zu halten, wird hier davon ausgegangen, daß die Mittellinie der Hochspannungsleitung (Objekt $\mathcal{O}_1$) durch eine Gerade und die Mittellinie des der Straße (Objekt $\mathcal{O}_2$) durch ein Polynom 2ten Grades approximiert werden können. (Prinzipiell können hier beliebige Funktionen gewählt werden, die zur Parametrisierung der beiden Mittellinien im Bild geeignet sind.) Weiter werden der Einfachheit halber die Abszissenwerte u_i der beobachteten Punkte als fest vorgegeben angesehen und nur die Ordinaten y_i als mit Meßunsicherheiten behaftete Beobachtungen betrachtet.

Zur Eignung herkömmlicher Schätzverfahren für die Aufgabe

An dieser Stelle sei darauf hingewiesen, daß die oben formulierte Aufgabe mit herkömmlichen[4] Methoden der Parameterschätzung im Gauß Markoff Modell *nicht* ohne weiteres gelöst werden kann. Mit herkömmlichen Schätzmethoden können aus einem Satz von Be-

[3] Gemeint ist hier eine Modellierung der Beobachtungen im Sinne eines Gauß Markoff Modells bzw. eines Gauß Helmert Modells. (vgl. [Koch 1998])

[4] Unter den herkömmlichen Schätzverfahren wird hier die Parameterschätzung im Gauß Markoff Modell nach der Methode der kleinsten Quadrate, der Maximum - Likelihood -Methode und nach der Methode der besten linearen erwartungstreuen Schätzung verstanden (vgl. [Koch 1998])

obachtungsdaten jeweils nur Schätzwerte für die unbekannten Parameter eines einzigen Beobachtungsmodells abgeleitet werden; es ist nicht möglich, aus einem Beobachtungssatz gleichzeitig die Parameter mehrerer Modelle zu schätzen, von denen für jede Beobachtung jeweils nur eines zutrifft[5]. Übertragen auf die vorliegende Aufgabe bedeutet dies, daß mit herkömmlichen Schätzmethoden aus den Beobachtungsdaten lediglich *entweder* Schätzwerte für die unbekannten Parameter der Gerade *oder* Schätzwerte für die unbekannten Parameter des Polynoms 2ten Grades bestimmt werden können. Sollen die Parameter *beider* Modelle geschätzt werden, so werden zwei getrennte Ausgleichungen erforderlich. Diese Ausgleichungen führen allerdings nur dann zu optimalen Approximationen der Mittellinien von $\mathcal{O}_1$ und $\mathcal{O}_2$, wenn vorab bekannt ist, welche Beobachtungen in welche der beiden Ausgleichungen einzuführen sind, d.h. welche Koordinatenpaare zu welchem Objekt gehören; denn wird beispielsweise in die Schätzung der Geradenparameter (d.h. der Parameter zur Beschreibung der Mittellinie der Hochspannungsleitung $\mathcal{O}_1$) eine Beobachtung (d.h. ein Koordinatenpaar) eingeführt, die tatsächlich zu der Straße (Objekt $\mathcal{O}_2$) gehört, so stellt diese Beobachtung in der Ausgleichung zur Schätzung der Geradenparameter einen Ausreißer dar, der das Ergebnis der Ausgleichung verfälscht.

Es ist demnach wegen der Unkenntnis um die Zuordnung der gemessenen Koordinaten zu den beiden Objekten $\mathcal{O}_1$ und $\mathcal{O}_2$ nicht ohne weiteres möglich, die oben gestellte Aufgabe mit herkömmlichen Schätzmethoden zu lösen, da das Zuordnungsproblem mit herkömmlichen Schätzverfahren zumindest nicht gleichzeitig mit der Parameterschätzung gelöst werden kann.

42 Lösung mittels des EM-Algorithmus:

Im folgenden wird aufgezeigt, wie die oben gestellte Aufgabe durch Anwendung des EM-Algorithmus gelöst werden kann. Dazu wird die Aufgabe zunächst als Schätzproblem mit unvollständigen Beobachtungsdaten formuliert. Anschließend wird auf dieses Schätzproblem dann der EM-Algorithmus angewandt.

421 Formulierung der Aufgabe als Schätzproblem aus unvollständigen Beobachtungsdaten

Zur Formulierung der Aufgabe als Schätzprolem aus unvollständigen Beobachtungsdaten werden zunächst im Vektor $\boldsymbol{y}$ der zugänglichen Beobachtungen die y-Werte der Meßreihe (410.1) zusammengefasst:

$$\boldsymbol{y} = [y_1 \ , \ y_2 \ , \ \cdots \ , \ y_N]^T \tag{421.1}$$

Um die folgenden Betrachtungen möglichst einfach zu halten, wird angenommen, daß die Beobachtungen y_i paarweise statistisch unabhängig voneinander sind, d.h. für die Kovarianz $C(y_i, y_j)$ zweier Beobachtungen y_i und y_j gilt

$$C(y_i, y_j) = 0 \text{ für } i \neq j.$$

Weiter wird angenommen, daß die zur Hochspannungsleitung gehörenden Beobachtungen untereinander gleich genau sind und die Varianz σ_1^2 besitzen. Ebenso werden die zur

[5]Man beachte, daß es sich bei der multivariaten Parameterschätzung (vgl. [Koch 1998]) im Sinne einer Deformationsanalyse um eine mehrfache Bestimmung *derselben* Parameter handelt und nicht um die Bestimmung von Parametern mehrerer Modelle

Straße gehörenden Beobachtungen als untereinander gleich genau angenommenen, ihre Varianz beträgt σ_2^2. [6] Somit ergibt sich die Kovarianzmatrix $\boldsymbol{D}(\boldsymbol{y})$ der Beobachtungen als Diagonalmatrix

$$\begin{aligned}\boldsymbol{D}(\boldsymbol{y}) =& diag\,[V(y_1), V(y_2), \ldots, V(y_N)] \\ \text{mit } V(y_i) =& \begin{cases}\sigma_1^2 & \text{falls } y_i \text{ zu } \mathcal{O}_1 \text{ (Hochspannungsleitung) gehört,} \\ \sigma_2^2 & \text{falls } y_i \text{ zu } \mathcal{O}_2 \text{ (Straße) gehört.}\end{cases}\end{aligned} \tag{421.2}$$

Die Beobachtungen y_i gehören jeweils entweder zum Modell „Gerade" oder zum Modell „Polynom 2. Grades". Zur Charakterisierung der Beobachtungen kommen also folgende Modellansätze im Sinne eines Gauß-Markoff-Modells in Frage:

Modell 1 (Objekt $\mathcal{O}_1$, „Hochspannungsleitung"), Linearen Approximation:

$$y_i + r_{i1} = a_0 + a_1 u_i = \underbrace{[1\ ,\ u_i]}_{:=\overset{1}{\boldsymbol{m}}_i^T} \underbrace{\begin{bmatrix}a_0\\ a_1\end{bmatrix}}_{:=\boldsymbol{\beta}_1} = \overset{1}{\boldsymbol{m}}_i^T \boldsymbol{\beta}_1 \qquad \text{mit } V(y_i) = \sigma_1^2 \text{ und } C(y_i, y_j) = 0$$

$$\text{für } i, j \in \{1, \ldots, N\}\ j \neq i \tag{421.3}$$

Modell 2 (Objekt $\mathcal{O}_2$, „Straße"), Approximation durch ein Polynom 2ten Grades:

$$y_i + r_{i2} = b_0 + b_1 u_i + b_2 u_i^2 = \underbrace{[1, u_i, u_i^2]}_{:=\overset{2}{\boldsymbol{m}}_i^T} \underbrace{\begin{bmatrix}b_0\\ b_1\\ b_2\end{bmatrix}}_{:=\boldsymbol{\beta}_2} = \overset{2}{\boldsymbol{m}}_i^T \boldsymbol{\beta}_2 \text{ mit } V(y_i) = \sigma_2^2 \text{ und } C(y_i, y_j) = 0$$

$$\text{für } i, j \in \{1, \ldots, N\}\ , j \neq i \tag{421.4}$$

Hierin stellen die Elemente von $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$ sowie σ_1 und σ_2 die unbekannten Parameter des Schätzproblems dar. Sie werden im Vektor

$$\boldsymbol{\beta} = \left[\boldsymbol{\beta}_1^T\ ,\ \sigma_1\ ,\ \boldsymbol{\beta}_2^T\ ,\ \sigma_2\right]^T \tag{421.5}$$

zusammengefasst. Mit r_{i1} und r_{i2} sind die Beobachtungsresiduen bezeichnet, d.h. die Abweichungen der Beobachtungen y_i von ihren Erwartungswerten $\hat{y}_{i1}$ bzw. $\hat{y}_i 2$ in den Modellen 1 und 2.

Wie in Abschnitt 41 angedeutet, fehlen im Beobachtungsmaterial Informationen über die Zugehörigkeit der einzelnen Beobachtungen y_i zu Modell 1 oder Modell 2, so daß diese Informationen im folgenden als unzugängliche Beobachtungen aufgefaßt werden. Formal wird für jede tatsächliche Beobachtung y_i ein Vektor

$$\boldsymbol{x}_i = \begin{bmatrix}x_{i1}\\ x_{i2}\end{bmatrix} \in \{\boldsymbol{e}_1, \boldsymbol{e}_2\} \qquad \text{mit } \boldsymbol{e}_1 = \begin{bmatrix}1\\ 0\end{bmatrix}\ ,\ \boldsymbol{e}_2 = \begin{bmatrix}0\\ 1\end{bmatrix} \tag{421.6}$$

nicht zugänglicher Beobachtungen x_{i1} und x_{i2} eingeführt. In Abhängigkeit davon, ob die Beobachtung y_i zur Hochspannungsleitung (Modell 1) oder zur Straße (Modell 2) gehört, gilt $\boldsymbol{x}_i = \boldsymbol{e}_1$ oder $\boldsymbol{x}_i = \boldsymbol{e}_2$. Gehört die Beobachtung y_i zu Modell 1, so gilt $\boldsymbol{x}_i = \boldsymbol{e}_1$; im

[6] U.U. sind die Mittellinien der beiden Objekte im Bild nicht gleich gut lokalisierbar. (Beispielsweise ist die Mittellinie einer Hochspannungsleitung durch das mittlere Kabel schärfer festgelegt als die Mittellinie einer Straße durch den doch (relativ breiten) Mittelstreifen.) Daher ist zwischen der Varianz der zu Objekt 1 und Objekt 2 gehörenden Beobachtungen zu unterscheiden.

anderen Fall, in dem die Beobachtung y_i zu Modell 2 gehört, gilt $\boldsymbol{x}_i = \boldsymbol{e}_2$. Daher wird $\boldsymbol{x}_i$ auch als *Indikatorvektor* mit den *Indikatoren* x_{i1} und x_{i2} bezeichnet.

Die Vektoren unzugänglicher Beobachtungen $\boldsymbol{x}_i,\ i \in \{1, \ldots N\}$ werden in dem Vektor $\boldsymbol{x}$

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \ldots \\ \boldsymbol{x}_N \end{bmatrix} = [x_{11}\ ,\ x_{12}\ ,\ \ldots\ , x_{N1}\ ,\ x_{N2}]^T \tag{421.7}$$

zusammengefasst. Dies ist der unzugängliche Beobachtungsvektor des zu lösenden Schätzproblems aus unvollständigen Beobachtungsdaten.

Der zur Beobachtung y_i gehörende vollständige Beobachtungsvektor ergibt sich unter Einbeziehung der unzugänglichen Beobachtungen $\boldsymbol{x}_i$ zu

$$\boldsymbol{z}_i = \begin{bmatrix} y_i \\ \boldsymbol{x}_i \end{bmatrix} = [y_i\ ,\ x_{i1}\ ,\ x_{i2}]^T\ . \tag{421.8}$$

Für den vollständigen Beobachtungsvektor des Schätzproblems aus unvollständigen Daten gilt damit

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 \\ \ldots \\ \boldsymbol{z}_N \end{bmatrix} = [y_1\ ,\ x_{11}\ ,\ x_{12}\ ,\ \ldots\ ,\ y_N\ ,\ x_{N1}\ ,\ x_{N2}]^T\ . \tag{421.9}$$

Die bisherigen Ausführungen zusammenfassend, kann die im Rahmen der Parameterschätzung zu lösende Aufgabe wie folgt als Schätzproblem aus unvollständigen Beobachtungsdaten formuliert werden:

Schätzproblem aus unvollständigen Beobachtungsdaten:

Von den in (421.9) spezifizierten vollständigen Beobachtungsdaten $\boldsymbol{z}$ sind nur die gemäß (421.7) in dem Vektor $\boldsymbol{y}$ enthaltenen Beobachtungen tatsächlich gemessen, es fehlen die Beobachtungen x_{i1} und x_{i2} für $i \in \{1, \ldots, N\}$ aus (421.6).
Ausgehend hiervon sind die unbekannten Parameter $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$ der Modelle 1 und 2 (vgl. (421.3) bzw. (421.4)) sowie die zugehörigen Varianzen σ_1^2 und σ_2^2 der Beobachtungen zu schätzen.

Dieses Schätzproblem läßt sich mittels des EM-Algorithmus lösen. Im nächsten Abschnitt wird aufgezeigt, welche Berechnungen hierfür im E-Schritt und im M-Schritt erforderlich sind.

422 Ableitung der EM-Iterationsschritte

Wie in Kapitel 2 ausführlich erläutert, wird im Rahmen des EM-Algorithmus die Kullback-Leibler-Statistik iterativ berechnet und maximiert(vgl. 324.20). Die Berechnung der Kullback-Leibler-Statistik setzt voraus, daß die von den unbekannten Parametern $\boldsymbol{\beta}$ abhängige Dichtefunktion $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ des vollständigen Beobachtungsvektors $\boldsymbol{z}$ bekannt ist. Daher wird zunächst diese Dichtefunktion angegeben.

Herleitung der Dichte $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ des vollständigen Beobachtungsvektors

Zur Ableitung der Dichtefunktion $f(\boldsymbol{z} \mid \boldsymbol{\beta})$ wird zunächst mit $p(\boldsymbol{x}_i \mid \boldsymbol{\beta})$ Wahrscheinlichkeitsverteilung der diskreten Zufallsvariablen $\boldsymbol{x}_i \in \{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ bezeichnet, die a priori, d.h.

ohne Kenntnis von der numerischen Realisierug des Beobachtungsvektors $\boldsymbol{y}$, jeder Beobachtung y_i die Wahrscheinlichkeiten $P(\boldsymbol{x}_i = \boldsymbol{e}_1 \mid \boldsymbol{\beta}) = p(\boldsymbol{e}_1 \mid \boldsymbol{\beta})$ bzw. $P(\boldsymbol{x}_i = \boldsymbol{e}_2 \mid \boldsymbol{\beta}) = p(\boldsymbol{e}_2 \mid \boldsymbol{\beta})$ zuordnet, daß sie zu Modell 1 (d.h. zur Hochspannungsleitung) bzw. zu Modell 2 (d.h. zur Straße) gehört, sofern mit $\boldsymbol{\beta}$ die Parameter der beiden Modelle gegeben sind. Es wird hier angenommen, daß diese Dichtefunktion für alle Beobachtungen y_i dieselbe ist.

Ginge aus der Aufgabenstellung ein Zusammenhang zwischen den Modellparametern $\boldsymbol{\beta}$ und den Wahrscheinlichkeiten $P(\boldsymbol{x}_i = \boldsymbol{e}_1)$ bzw. $P(\boldsymbol{x}_i = \boldsymbol{e}_2)$ hervor, so würde sich hieraus für $p(\boldsymbol{x}_i \mid \boldsymbol{\beta})$ auch tatsächlich eine von den Parametern $\boldsymbol{\beta}$ abhängige Funktion ergeben. Da aus der hier vorliegenden Aufgabenstellung ein solcher Zusammenhang jedoch nicht ersichtlich ist, werden die Wahrscheinlichkeiten $P(\boldsymbol{x}_i = \boldsymbol{e}_1)$ und $P(\boldsymbol{x}_i = \boldsymbol{e}_2)$ nach dem erwarteten Anteil der zur Hochspannungsleitung (Modell 1) bzw. zur Straße (Modell 2) gehörenden Beobachtungen an der Gesamtzahl N der Beobachtungen bemessen. Es wird angenommen, daß 1/3 der Beobachtungen zur Hochspannungsleitung und 2/3 der Beobachtungen zur Straße gehören. Daher gilt

$$P(\text{Beobachtung } y_i \text{ gehört zu Objekt } \mathcal{O}_1 \text{ (HS-Leitung)}) = P(\boldsymbol{x}_i = \boldsymbol{e}_1 \mid \boldsymbol{\beta}) = p(\boldsymbol{e}_1 \mid \boldsymbol{\beta}) = \frac{1}{3} \quad (422.1)$$

und

$$P(\text{Beobachtung } y_i \text{ gehört zu Objekt } \mathcal{O}_2 \text{ (Straße)}) = P(\boldsymbol{x}_i = \boldsymbol{e}_2 \mid \boldsymbol{\beta}) = p(\boldsymbol{e}_2 \mid \boldsymbol{\beta}) = \frac{2}{3}. \quad (422.2)$$

Die Beobachtungen y_i werden unabhängig davon, zu welchem Objekt sie gehören (d.h. unabhängig davon, welches Beobachtungsmodell ihnen zugrunde liegt), als normalverteilt vorausgesetzt. Je nachdem, ob eine Beobachtung y_i zu Objekt $\mathcal{O}_1$ oder zu Objekt $\mathcal{O}_2$ gehört, liegen ihrer Normalverteilung allerdings unterschiedliche Parameter zugrunde (vgl. Gl. (421.4)). Gehört die Beobachtung y_i zu der Hochspannungsleitung (Modell 1), so gilt für die Wahrscheinlichkeitsdichte $f(y_i \mid \boldsymbol{x}_i = \boldsymbol{e}_1, \boldsymbol{\beta})$ in Abhängigkeit von den Parametern $\boldsymbol{\beta}_1$ der linearen Approximation und der Standardabweichung σ_1 der zur Hochspannungsleitung gehörenden Beobachtungen

Modell 1 (lineare Approximation):

$$f(y_i \mid \boldsymbol{x}_i = \boldsymbol{e}_1, \boldsymbol{\beta}) = f(y_i \mid \boldsymbol{e}_1, \boldsymbol{\beta}_1, \sigma_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}[y_i - \overset{1}{\boldsymbol{m}}_i^T \boldsymbol{\beta}_1]^2\right) \quad (422.3)$$

Gehört die Beobachtung y_i hingegen zur Straße (Modell 2), so gilt für $f(y_i \mid \boldsymbol{x}_i = \boldsymbol{e}_2, \boldsymbol{\beta})$ in Abhängigkeit von den Parametern $\boldsymbol{\beta}_2$ der Approximation mittels eines Polynoms 2. Grades und der Standardabweichung σ_2 der Beobachtungen, die zur Straße gehören:

Modell 2 (Approximation durch Polynom 2ten Grades):

$$f(y_i \mid \boldsymbol{x}_i = \boldsymbol{e}_2, \boldsymbol{\beta}) = f(y_i \mid \boldsymbol{e}_2, \boldsymbol{\beta}_2, \sigma_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2\sigma_2^2}[y_i - \overset{2}{\boldsymbol{m}}_i^T \boldsymbol{\beta}_2]^2\right) \quad (422.4)$$

Mit (422.1), (422.2), (422.3) und (422.4) lässt sich die Dichtefunktion $f(\boldsymbol{z}_i \mid \boldsymbol{\beta}) = f(\boldsymbol{z}_i \mid \boldsymbol{\beta}_1, \sigma_1, \boldsymbol{\beta}_2, \sigma_2)$ der vollständigen Beobachtung $\boldsymbol{z}_i$ ableiten. Aufgrund der Bayes - Formel

gilt

$$f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \boldsymbol{\beta}) = p(\boldsymbol{e}_1 \mid \boldsymbol{\beta}) \cdot f(y_i \mid \boldsymbol{e}_1, \boldsymbol{\beta}) = \frac{1}{3\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}[y_i - \boldsymbol{m}_i^{1T}\boldsymbol{\beta}_1]^2} \tag{422.5}$$

und

$$f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_2 \mid \boldsymbol{\beta}) = p(\boldsymbol{e}_2 \mid \boldsymbol{\beta}) \cdot f(y_i \mid \boldsymbol{e}_2, \boldsymbol{\beta}) = \frac{2}{3\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}[y_i - \boldsymbol{m}_i^{2T}\boldsymbol{\beta}_2]^2}. \tag{422.6}$$

Hiermit erhält man unter Ausnutzung der Tatsache, daß $\boldsymbol{x}_i$ in Abhängigkeit von dem für die Beobachtung y_i zutreffenden Modell entweder die Elemente $x_{i1} = 1$ und $x_{i2} = 0$ oder $x_{i1} = 0$ und $x_{i2} = 1$ enthält, die Dichte

$$\begin{aligned} f(\boldsymbol{z}_i \mid \boldsymbol{\beta}) &= f(y_i, \boldsymbol{x}_i \mid \boldsymbol{\beta}) \\ &= x_{i1} \cdot p(\boldsymbol{e}_1 \mid \boldsymbol{\beta}) \cdot f(y_i \mid \boldsymbol{e}_1, \boldsymbol{\beta}) + x_{i2} \cdot p(\boldsymbol{e}_2 \mid \boldsymbol{\beta}) \cdot f(y_i \mid \boldsymbol{e}_2, \boldsymbol{\beta}) \\ &= [p(\boldsymbol{e}_1 \mid \boldsymbol{\beta}) \cdot f(y_i \mid \boldsymbol{e}_1, \boldsymbol{\beta})]^{x_{i1}} \cdot [p(\boldsymbol{e}_2\boldsymbol{\beta}) \cdot f(y_i \mid \boldsymbol{e}_2, \boldsymbol{\beta})]^{x_{i2}}, \end{aligned} \tag{422.7}$$

so daß folgt

$$f(\boldsymbol{z}_i \mid \boldsymbol{\beta}) = \left(\frac{1}{3\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}[y_i - \boldsymbol{m}_i^{1T}\boldsymbol{\beta}_1]^2}\right)^{x_{i1}} \cdot \left(\frac{2}{3\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}[y_i - \boldsymbol{m}_i^{2T}\boldsymbol{\beta}_2]^2}\right)^{x_{i2}}. \tag{422.8}$$

Die tatsächlichen Beobachtungen y_i und die nicht zugänglichen Beobachtungsvektoren $\boldsymbol{x}_i$ werden als paarweise statistisch voneinander unabhängig aufgefaßt. Daher erhält man nach (vgl. [KOCH 1998], Gl. (215.1)) die Dichtefunktion des vollständigen Beobachtungsvektors $\boldsymbol{z}$ aus dem Produkt der Einzeldichten $f(\boldsymbol{z}_i \mid \boldsymbol{\beta})$, es ergibt sich also die **Dichtefunktion des vollständigen Beobachtungsvektors**

$$\begin{aligned} f(\boldsymbol{z} \mid \boldsymbol{\beta}) = f(\boldsymbol{z} \mid \boldsymbol{\beta}_1, \sigma_1, \boldsymbol{\beta}_2, \sigma_2) &= \prod_{i=1}^{N} f(\boldsymbol{z}_i \mid \boldsymbol{\beta}) \\ &= \prod_{i=1}^{N} \left(\frac{1}{3\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}[y_i - \boldsymbol{m}_i^{1T}\boldsymbol{\beta}_1]^2}\right)^{x_{i1}} \cdot \left(\frac{2}{3\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}[y_i - \boldsymbol{m}_i^{2T}\boldsymbol{\beta}_2]^2}\right)^{x_{i2}}. \end{aligned} \tag{422.9}$$

Durch Logarithmieren dieser Dichtefunktion erhält man den **Logarithmus der Dichte des vollständigen Beobachtungsvektors**

$$\begin{aligned} \log f(\boldsymbol{z} \mid \boldsymbol{\beta}) &= \sum_{i=1}^{N} \log f(\boldsymbol{z}_i \mid \boldsymbol{\beta}_1, \sigma_1, \boldsymbol{\beta}_2, \sigma_2) \\ &= \sum_{i=1}^{N} \left\{ x_{i1} \log \left(\frac{1}{3\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}[y_i - \boldsymbol{m}_i^{1T}\boldsymbol{\beta}_1]^2}\right) + x_{i2} \log \left(\frac{2}{3\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}[y_i - \boldsymbol{m}_i^{2T}\boldsymbol{\beta}_2]^2}\right) \right\}. \end{aligned} \tag{422.10}$$

Auf Grundlage der Gleichungen (422.9) und (422.10) soll nun die beim EM-Algorithmus in jedem Iterationsschritt zu maximierende Kullback-Leibler-Statistik abgeleitet werden.

E-Schritt: Berechnung der Kullback-Leibler-Statistik

Sind mit $\hat{\boldsymbol{\beta}}^{(j)} = \left[\hat{\boldsymbol{\beta}}_1^{(j)T}, \hat{\sigma}_1^{(j)}, \hat{\boldsymbol{\beta}}_2^{(j)T}, \hat{\sigma}_2^{(j)}\right]^T$ vorläufige Schätzwerte für die unbekanten Parameter bekannt, so ergibt sich mit $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_1^T, \sigma_1, \boldsymbol{\beta}_2^T, \sigma_2\right]^T$ die Kullback-Leibler-Statistik

$Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(j)}) = E[\log f(\boldsymbol{z} \mid \boldsymbol{\beta}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}]$ als Erwartungswert der logarithmierten Likelihoodfunktion (422.10). Im Falle einer überabzählbaren Menge $\mathcal{Z}(\boldsymbol{y})$ wäre $Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(j)})$ nach (231.8) durch Integration zu berechnen. In unserem Beispiel handelt es sich bei $\mathcal{Z}(\boldsymbol{y})$ wegen

$$\begin{aligned}\mathcal{Z}(\boldsymbol{y}) &= \mathcal{Z}([y_1\ ,\ y_2\ ,\ \ldots\ ,\ y_N]^T) \\ &= \left\{ [y_1, x_{11}, x_{12}, \ldots, y_N, x_{N1}, x_{N2}]^T \,\middle|\, \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} \in \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \ \forall\, i \in \{1, \ldots, N\} \right\}\end{aligned} \tag{422.11}$$

allerdings um eine diskrete Menge, so daß anstelle der Integration (231.8) eine Summation über alle Elemente $\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y})$ durchzuführen ist. Mit (422.10) ergibt sich

$$\begin{aligned}Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(j)}) &= E[\log f(\boldsymbol{z} \mid \boldsymbol{\beta}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] = \sum_{\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y})} \log f(\boldsymbol{z} \mid \boldsymbol{\beta}) \cdot f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) \\ &= \sum_{x_{N1}=0}^{1} \cdots \sum_{x_{21}=0}^{1} \sum_{x_{11}=0}^{1} \sum_{i=1}^{N} \Bigg\{ x_{i1} \log \left(\frac{1}{3\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}[y_i - \boldsymbol{m}_i^{1T}\boldsymbol{\beta}_1]^2} \right) \\ &\qquad + (1 - x_{i1}) \log \left(\frac{2}{3\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}[y_i - \boldsymbol{m}_i^{2T}\boldsymbol{\beta}_2]^2} \right) \Bigg\} \cdot f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}).\end{aligned} \tag{422.12}$$

Man erkennt in dieser Gleichung, daß der Erwartungswert $Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(j)})$ linear in den Indikatoren x_{i1} und $x_{i2} = 1 - x_{i1}$ ist. Wegen der Linearität des Erwartungswertoperators (vgl. [Koch 1998], Gl. 231.5) gilt daher

$$\begin{aligned}Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(j)}) &= E[\log f(\boldsymbol{z} \mid \boldsymbol{\beta}) \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] \\ &= \sum_{i=1}^{N} \Bigg\{ E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] \log \left(\frac{1}{3\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}[y_i - \boldsymbol{m}_i^{1T}\boldsymbol{\beta}_1]^2} \right) \\ &\qquad + E[x_{i2} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] \log \left(\frac{2}{3\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}[y_i - \boldsymbol{m}_i^{2T}\boldsymbol{\beta}_2]^2} \right) \Bigg\}.\end{aligned} \tag{422.13}$$

Der Expectation - Schritt des hier vorliegenden EM-Algorithmus (d.h. die Berechnung der Kullback-Leibler-Statistik) besteht also im wesentlichen in der Berechnung der Erwartungswerte $E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}]$ und $E[x_{i2} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}]$. Daher werden nun Gleichungen zur Berechnung dieser Erwartungswerte hergeleitet.

Die Indikatoren x_{i1} und x_{i2} für $i \in \{1, \ldots, N\}$ sind diskrete Zufallsvariablen, die nur die Werte 0 und 1 annehmen können. Der Erwartungswert einer solchen Zufallsvariable X ist nach ([Koch 1998], Gl. (231.1)) definiert als die Summe

$$E(X) = \sum_{x=0}^{1} x \cdot P(X = x) = 1 \cdot p(X = 1) + 0 \cdot p(X = 0) = P(X = 1). \tag{422.14}$$

Daher gilt für die Erwartungswerte von x_{i1} und x_{i2} mit $i \in \{1, \ldots n\}$

$$E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] = P(x_{i1} = 1 \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) \quad \text{und} \tag{422.15}$$

$$E[x_{i2} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] = P(x_{i2} = 1 \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}), \tag{422.16}$$

d.h. der Erwartungswert $E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}]$ gibt die Wahrscheinlichkeit dafür an, daß der Beobachtung y_i das Modell 1 (Hochspannungsleitung) zugrundeliegt, sofern insgesamt die Beobachtungen $\boldsymbol{y}$ vorliegen und die vorläufigen Parameter $\hat{\boldsymbol{\beta}}^{(j)}$ der beiden Modelle 1 und 2 gegeben sind. Entsprechend gibt der Erwartungswert $E[x_{i2} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}]$ die Wahrscheinlichkeit dafür an, daß der Beobachtung y_i das Modell 2 (Straße) zugrundeliegt, falls $\boldsymbol{y}$ und $\hat{\boldsymbol{\beta}}^{(j)}$ gegeben sind. Da die Zugehörigkeit der Beobachtung y_i zur Hochspannungsleitung (Modell 1) bzw. zur Straße (Modell 2) zwei sich gegenseitig ausschließende Ereignisse darstellen, ergibt sich nach ([KOCH 1998], Gl. (213.3))

$$E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] + E[x_{i2} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] = 1. \tag{422.17}$$

Nach der Bayes-Formel gilt zunächst

$$P(x_{i1} = 1 \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) = P(\boldsymbol{x}_i = \boldsymbol{e}_1 \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) = \frac{f(\boldsymbol{y}, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)})}{f(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(j)})}. \tag{422.18}$$

Hierin gilt mit (422.1) wiederum aufgrund der Bayes-Formel

$$f(\boldsymbol{y}, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) = p(\boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) \cdot f(\boldsymbol{y} \mid \boldsymbol{x}_i = \boldsymbol{e}_1, \hat{\boldsymbol{\beta}}^{(j)}), \tag{422.19}$$

woraus wegen der statistischen Unabhängigkeit der Beobachtungen y_i folgt

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) &= p(\boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) \cdot \prod_{k=1}^{N} f(y_k \mid \boldsymbol{x}_i = \boldsymbol{e}_1, \hat{\boldsymbol{\beta}}^{(j)}) \\ &= p(\boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) \cdot f(y_i \mid \boldsymbol{x}_i = \boldsymbol{e}_1, \hat{\boldsymbol{\beta}}^{(j)}) \cdot \prod_{\substack{k=1 \\ k \neq i}}^{N} f(y_k \mid \boldsymbol{x}_i = \boldsymbol{e}_1, \hat{\boldsymbol{\beta}}^{(j)}). \end{aligned} \tag{422.20}$$

Nach (422.3) und (422.4) sind die Dichtefunktionen $f(y_k \mid \boldsymbol{x}_k = \boldsymbol{e}_1, \hat{\boldsymbol{\beta}}^{(j)})$ und $f(y_k \mid \boldsymbol{x}_k = \boldsymbol{e}_2, \hat{\boldsymbol{\beta}}^{(j)})$ für $k \neq i$ unabhängig vom Indikator $\boldsymbol{x}_i$. Daher gilt für alle $k \neq i$ die Beziehung $f(y_k \mid \boldsymbol{x}_i = \boldsymbol{e}_1, \hat{\boldsymbol{\beta}}^{(j)}) = f(y_k \mid \hat{\boldsymbol{\beta}}^{(j)})$ und es folgt

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) &= p(\boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) \cdot f(y_i \mid \boldsymbol{x}_i = \boldsymbol{e}_1, \hat{\boldsymbol{\beta}}^{(j)}) \cdot \prod_{\substack{k=1 \\ k \neq i}}^{N} f(y_k \mid \hat{\boldsymbol{\beta}}^{(j)}) \\ &= f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) \cdot \prod_{\substack{k=1 \\ k \neq i}}^{N} f(y_k \mid \hat{\boldsymbol{\beta}}^{(j)}). \end{aligned} \tag{422.21}$$

Wegen der statistischen Unabhängigkeit der Beobachtungen y_i gilt außerdem

$$f(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(j)}) = \prod_{k=1}^{N} f(y_k \mid \hat{\boldsymbol{\beta}}^{(j)}). \tag{422.22}$$

Durch Einsetzen von (422.21) und (422.22) in (422.18) erhält man schließlich

$$\begin{aligned} P(x_{i1} = 1 \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) &= \frac{f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) \cdot \prod_{\substack{k=1 \\ k \neq i}}^{N} f(y_k \mid \hat{\boldsymbol{\beta}}^{(j)})}{\prod_{k=1}^{N} f(y_k \mid \hat{\boldsymbol{\beta}}^{(j)})} \\ &= \frac{f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)})}{f(y_i \mid \hat{\boldsymbol{\beta}}^{(j)})}. \end{aligned} \tag{422.23}$$

Bei der Zugehörigkeit einer Beobachtung y_i zur Hochspannungsleitung bzw. zur Straße handelt es sich um zwei sich gegenseitig ausschließende Ereignisse. Daher gilt (vgl. [KOCH 1998], Gl.(213.3))

$$f(y_i \mid \hat{\boldsymbol{\beta}}^{(j)}) = f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) + f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_2 \mid \hat{\boldsymbol{\beta}}^{(j)}), \tag{422.24}$$

Hiermit ergibt sich

$$P(x_{i1} = 1 \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) = \frac{f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)})}{f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) + f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_2 \mid \hat{\boldsymbol{\beta}}^{(j)})} = E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}], \tag{422.25}$$

worin die Dichten $f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)})$ und $f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_2 \mid \hat{\boldsymbol{\beta}}^{(j)})$ nach (422.3) und (422.4) berechnet werden. Ganz analog läßt sich zeigen, daß gilt

$$P(x_{i2} = 1 \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) = \frac{f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_2 \mid \hat{\boldsymbol{\beta}}^{(j)})}{f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_1 \mid \hat{\boldsymbol{\beta}}^{(j)}) + f(y_i, \boldsymbol{x}_i = \boldsymbol{e}_2 \mid \hat{\boldsymbol{\beta}}^{(j)})} = E[x_{i2} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}]. \tag{422.26}$$

Mit (422.5) und (422.6) ergeben sich schließlich die **Erwartungswerte der fehlenden Beobachtungen:**

$$\mu_{i1}^{(j)} := E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] = \frac{\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T}\hat{\boldsymbol{\beta}}_1^{(j)}]^2}}{\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T}\hat{\boldsymbol{\beta}}_1^{(j)}]^2} + \frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \boldsymbol{m}_j^{2T}\hat{\boldsymbol{\beta}}_2^{(j)}]^2}} \tag{422.27}$$

und

$$\mu_{i2}^{(j)} := E[x_{i2} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] = \frac{\frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \boldsymbol{m}_j^{2T}\hat{\boldsymbol{\beta}}_2^{(j)}]^2}}{\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T}\hat{\boldsymbol{\beta}}_1^{(j)}]^2} + \frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \boldsymbol{m}_j^{2T}\hat{\boldsymbol{\beta}}_2^{(j)}]^2}} \tag{422.28}$$

Der E-Schritt des zur Lösung der Aufgabe anzuwendenden EM-Algorithmus besteht in der Berechnung der Erwartungswerte $\mu_{i1}^{(j)}$ und $\mu_{i2}^{(j)}$ für $i \in \{1, \ldots, N\}$ nach diesen beiden Gleichungen. Im j-ten Iterationsschritt ergibt sich dann nach (422.13) die **Kullback-Leibler-Statistik** in Abhängigkeit von dem Parametervektor $\boldsymbol{\beta}$ zu

$$Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}^{(j)}) = \sum_{i=1}^{N} \left\{ \mu_{i1}^{(j)} \log \left(\frac{1}{3\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}[y_i - \boldsymbol{m}_i^{1T}\boldsymbol{\beta}_1]^2} \right) + \mu_{i2}^{(j)} \log \left(\frac{2}{3\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}[y_i - \boldsymbol{m}_i^{2T}\boldsymbol{\beta}_2]^2} \right) \right\}. \tag{422.29}$$

Kontrolle der Gleichungen (422.27) und (422.28):

Zum gleichen Ergebnis wie in (422.27) und (422.28) kommt man *ohne* die Betrachtung der Erwartungswerte als Wahrscheinlichkeiten, wenn man ausgehend von (422.13) beispielsweise den Erwarungswert $E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}]$ nach

$$E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] = \sum_{\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y})} x_{i1} \cdot f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}),$$

berechnet. Nach (231.1) gilt $f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) = f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(j)}) \left[g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(j)}) \right]^{-1}$. Hierin ist die Dichte $f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(j)})$ mit (422.9) gegeben, die Dichte $g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(j)})$ ergibt sich aus (222.1), wobei an die Stelle der Integration wieder eine Summation über alle $\boldsymbol{x} \in \mathcal{Z}(\boldsymbol{y})$ tritt:

$$\begin{aligned}
g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(j)}) &= \sum_{\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y})} f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(j)}) \qquad \text{mit } f(\boldsymbol{z} \mid \hat{\boldsymbol{\beta}}^{(j)}) \overset{(422.9)}{=} \prod_{i=1}^{N} f(\boldsymbol{z}_i \mid \hat{\boldsymbol{\beta}}^{(j)}) \\
&= \sum_{\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y})} \prod_{i=1}^{N} f(\boldsymbol{z}_i \mid \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)}) \\
&= \sum_{\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y})} \prod_{i=1}^{N} \underbrace{\left(\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_i - \boldsymbol{m}_i^{1T}\hat{\beta}_1^{(j)}]^2} \right)^{x_{i1}}}_{=: W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})^{x_{i1}}} \cdot \underbrace{\left(\frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_i - \boldsymbol{m}_i^{2T}\hat{\beta}_2^{(j)}]^2} \right)^{x_{i2}}}_{=: W(y_i, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})^{x_{i2}}} \\
&= \sum_{x_{N1}=0}^{1} \cdots \sum_{x_{21}=0}^{1} \sum_{x_{11}=0}^{1} \prod_{i=1}^{N} W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})^{x_{i1}} \cdot W(y_i, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})^{1-x_{i1}} \\
&= \sum_{x_{N1}=0}^{1} \cdots \sum_{x_{21}=0}^{1} \prod_{i=2}^{N} W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})^{x_{i1}} \cdot W(y_i, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})^{1-x_{i1}} \\
&\qquad \cdot \Big(W(y_1, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_1, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)}) \Big) \\
&= \sum_{x_{N1}=0}^{1} \cdots \sum_{x_{31}=0}^{1} \prod_{i=3}^{N} W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})^{x_{i1}} \cdot W(y_i, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})^{1-x_{i1}} \\
&\qquad \cdot \Big(W(y_1, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_2, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)}) \Big) \cdot \Big(W(y_2, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_2, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)}) \Big) \\
&= \ldots \\
g(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}^{(j)}) &= \prod_{j=1}^{N} \left\{ \frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T}\hat{\beta}_1^{(j)}]^2} + \frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \boldsymbol{m}_j^{2T}\hat{\beta}_2^{(j)}]^2} \right\} \qquad (422.30)
\end{aligned}$$

Hiermit erhält man schließlich

$$f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) = \prod_{j=1}^{N} \frac{\left(\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T}\hat{\beta}_1^{(j)}]^2} \right)^{x_{j1}} \cdot \left(\frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \boldsymbol{m}_j^{2T}\hat{\beta}_2^{(j)}]^2} \right)^{x_{j2}}}{\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T}\hat{\beta}_1^{(j)}]^2} + \frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \boldsymbol{m}_j^{2T}\hat{\beta}_2^{(j)}]^2}}.$$

Diese Dichtefunktion wird zur Berechnung des Erwartungswertes $E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}]$ verwendet. Man erhält

$$\begin{aligned}
E[x_{i1} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}] &= \sum_{\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{y})} x_{i1} \cdot f(\boldsymbol{z} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}^{(j)}) \\
&= \sum_{x_{N1}=0}^{1} \cdots \sum_{x_{11}=0}^{1} x_{i1} \cdot \prod_{j=1}^{N} \frac{W(y_j, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})^{x_{j1}} \cdot W(y_j, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})^{1-x_{j1}}}{W(y_j, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_j, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})} \quad \text{für } i = 1, 2, \ldots, N \\
&= \sum_{x_{N1}=0}^{1} \cdots \sum_{x_{11}=0}^{1} \left\{ x_{i1} \cdot \frac{W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})^{x_{i1}} \cdot W(y_i, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})^{1-x_{i1}}}{W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_i, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})} \right. \\
&\qquad \left. \cdot \prod_{\substack{j=1 \\ j \neq i}}^{N} \frac{W(y_j, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})^{x_{j1}} \cdot W(y_j, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})^{1-x_{j1}}}{W(y_j, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_j, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})} \right\}
\end{aligned}$$

$$= \sum_{x_{N1}=0}^{1} \cdots \sum_{x_{(i-1),1}=0}^{1} \sum_{x_{(i+1),1}=0}^{1} \cdots \sum_{x_{11}=1}^{N} \left\{ \frac{W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})}{W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_i, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})} \right.$$

$$\left. \cdot \prod_{\substack{j=1 \\ j \neq i}}^{N} \frac{W(y_j, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})^{x_{j1}} \cdot W(y_j, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})^{1-x_{j1}}}{W(y_j, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_j, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})} \right\}$$

$$= \frac{W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})}{W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_i, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})} \cdot \frac{1}{\prod_{\substack{j=1 \\ j \neq i}}^{N} [W(y_j, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_j, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})]}$$

$$\cdot \underbrace{\sum_{x_{N1}=0}^{1} \cdots \sum_{x_{(i-1),1}=0}^{1} \sum_{x_{(i+1),1}=0}^{1} \cdots \sum_{x_{11}=1}^{N} \prod_{\substack{j=1 \\ j \neq i}}^{N} [W(y_j, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})^{x_{j1}} \cdot W(y_j, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})^{1-x_{j1}}]}_{= \prod_{\substack{j=1 \\ j \neq i}}^{N} [W(y_j, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_j, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})] \qquad \text{vgl.(422.30)}}$$

$$= \frac{W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)})}{W(y_i, \hat{\beta}_1^{(j)}, \hat{\sigma}_1^{(j)}) + W(y_i, \hat{\beta}_2^{(j)}, \hat{\sigma}_2^{(j)})}$$

$$\Rightarrow E[x_{i1} \mid \boldsymbol{y}, \hat{\beta}^{(j)}] = \frac{\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T} \hat{\beta}_1^{(j)}]^2}}{\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T} \hat{\beta}_1^{(j)}]^2} + \frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \boldsymbol{m}_j^{2T} \hat{\beta}_2^{(j)}]^2}}.$$

Analog gilt

$$E[x_{i2} \mid \boldsymbol{y}, \hat{\beta}^{(j)}] = \frac{\frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T} \hat{\beta}_2^{(j)}]^2}}{\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \boldsymbol{m}_j^{1T} \hat{\beta}_1^{(j)}]^2} + \frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \boldsymbol{m}_j^{2T} \hat{\beta}_2^{(j)}]^2}},$$

es werden also auf diese Weise die gleichen Ergebnisse erhalten wie in (422.27) und (422.28), womit diese Ergebnisse kontrolliert sind.

M-Schritt: Maximierung der Kullback-Leibler-Statistik

Nachdem im E-Schritt des EM-Algorithmus die Erwartungswerte $\mu_{ik}^{(j)}$ der Indikatoren x_{ik} für $i \in \{1, \ldots N\}$ und $k \in \{1, 2\}$ nach (422.27) und (422.28) berechnet worden sind, ist im M-Schritt muß die Kullback-Leibler-Statistik (422.29) durch Variation der unbekannten Parameter $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_1^T,\ \sigma_1,\ \boldsymbol{\beta}_2^T,\ \sigma_2\right]^T$ zu maximieren. Maximiert werden muß also die Funktion

$$Q(\boldsymbol{\beta} \mid \hat{\beta}^{(j)}) = \sum_{i=1}^{N} \left\{ \mu_{i1}^{(j)} \log \left(\frac{1}{3\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2(\sigma_1)^2}[y_i - \boldsymbol{m}_i^{1T} \boldsymbol{\beta}_1]^2} \right) \right.$$

$$\left. + \mu_{i2}^{(j)} \log \left(\frac{2}{3\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2(\sigma_2)^2}[y_i - \boldsymbol{m}_i^{2T} \boldsymbol{\beta}_2]^2} \right) \right\}$$

$$= -\sum_{i=1}^{N} \left[\mu_{i1}^{(j)} \log(3\sqrt{2\pi}\sigma_1) + \mu_{i2}^{(j)} \log(3/2\sqrt{2\pi}\sigma_2) \right]$$

$$- \sum_{i=1}^{N} \left(\frac{1}{2(\sigma_1)^2} \mu_{i1}^{(j)} [y_i - \boldsymbol{m}_i^{1T} \boldsymbol{\beta}_1]^2 + \frac{1}{2(\sigma_2)^2} \mu_{i2}^{(j)} [y_i - \boldsymbol{m}_i^{2T} \boldsymbol{\beta}_2]^2 \right).$$

Wie man sieht, hängt hierin nur die zweite Summe von den Modellparametern $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$ ab, so daß es zur Bestimmung neuer Schätzwerte $\hat{\boldsymbol{\beta}}_1^{(j+1)}$ und $\hat{\boldsymbol{\beta}}_2^{(j+1)}$ für diese Parameter

zunächst ausreicht, die zweite Summe separat zu maximieren. Gesucht werden also jetzt Schätzwerte $\hat{\boldsymbol{\beta}}_1^{(j+1)}$ und $\hat{\boldsymbol{\beta}}_2^{(j+1)}$, die die Summe

$$-\sum_{i=1}^{N}\left(\frac{1}{2\sigma_1^2}\mu_{i1}^{(j)}[y_i-\overset{1}{\boldsymbol{m}}_i^T\boldsymbol{\beta}_1]^2+\frac{1}{2\sigma_2^2}\mu_{i2}^{(j)}[y_i-\overset{2}{\boldsymbol{m}}_i^T\boldsymbol{\beta}_2]^2\right)$$

für $\boldsymbol{\beta}_1=\hat{\boldsymbol{\beta}}_1^{(j+1)}$ und $\boldsymbol{\beta}_2=\hat{\boldsymbol{\beta}}_2^{(j+1)}$ maximieren. Diese Summe kann wieder in zwei Teilsummen zerlegt werden, die getrennt voneinander zu maximieren sind:

$$-\frac{1}{2\sigma_1^2}\sum_{i=1}^{N}\mu_{i1}^{(j)}[y_i-\overset{1}{\boldsymbol{m}}_i^T\boldsymbol{\beta}_1]^2\rightarrow\max\qquad\text{und}\qquad-\frac{1}{2\sigma_2^2}\sum_{i=1}^{N}\mu_{i2}^{(j)}[y_i-\overset{2}{\boldsymbol{m}}_i^T\boldsymbol{\beta}_2]^2\rightarrow\max \tag{422.31}$$

Fasst man die Erwartungswerte $\mu_{i1}^{(j)}$ bzw. $\mu_{i2}^{(j)}$ der unzugänglichen Beobachtungen in den Gewichtsmatrizen $\boldsymbol{P}_1^{(j)}$ bzw. $\boldsymbol{P}_2^{(j)}$ zusammen, also

$$\boldsymbol{P}_1^{(j)}=diag\left(\mu_{11}^{(j)}\ ,\ \mu_{21}^{(j)}\ ,\ldots,\ \mu_{N1}^{(j)}\right)\ ,\ \boldsymbol{P}_2=diag\left(\mu_{12}^{(j)}\ ,\ \mu_{22}^{(j)}\ ,\ldots,\ \mu_{N2}^{(j)}\right) \tag{422.32}$$

und entsprechend die Koeffizienvektoren $\overset{1}{\boldsymbol{m}}_i^T$ bzw. $\overset{2}{\boldsymbol{m}}_i^T$ der Beobachtungsgleichungen des Modells 1 (421.3) bzw. des Modells 2 (421.4) in den Matrizen $\overset{1}{\boldsymbol{M}}$ und $\overset{2}{\boldsymbol{M}}$,

$$\overset{1}{\boldsymbol{M}}=\begin{bmatrix}\overset{1}{\boldsymbol{m}}_1^T\\ \overset{1}{\boldsymbol{m}}_2^T\\ \cdots\\ \overset{1}{\boldsymbol{m}}_N^T\end{bmatrix}\qquad,\qquad\overset{2}{\boldsymbol{M}}=\begin{bmatrix}\overset{2}{\boldsymbol{m}}_1^T\\ \overset{2}{\boldsymbol{m}}_2^T\\ \cdots\\ \overset{2}{\boldsymbol{m}}_N^T\end{bmatrix}, \tag{422.33}$$

so ergeben sich die Schätzwerte $\hat{\boldsymbol{\beta}}_1^{(j+1)}$ und $\hat{\boldsymbol{\beta}}_2^{(j+1)}$ aufgrund der Beziehungen

$$\hat{\boldsymbol{\beta}}_1^{(j+1)}=\arg\max_{\boldsymbol{\beta}_1}-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\overset{1}{\boldsymbol{M}}\boldsymbol{\beta}_1\right)^T\boldsymbol{P}_1\left(\boldsymbol{y}-\overset{1}{\boldsymbol{M}}\boldsymbol{\beta}_1\right)\text{ und} \tag{422.34}$$

$$\hat{\boldsymbol{\beta}}_2^{(j+1)}=\arg\max_{\boldsymbol{\beta}_1}-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\overset{2}{\boldsymbol{M}}\boldsymbol{\beta}_2\right)^T\boldsymbol{P}_2\left(\boldsymbol{y}-\overset{1}{\boldsymbol{M}}\boldsymbol{\beta}_2\right). \tag{422.35}$$

Aus den Gleichungen (422.31) und (422.35) ist ersichtlich, daß sich die neuen Schätzwerte $\hat{\boldsymbol{\beta}}_1^{(j+1)}$ und $\hat{\boldsymbol{\beta}}_2^{(j+1)}$ für die unbekannten Parameter $\boldsymbol{\beta}_1$ und $\boldsymbol{\beta}_2$ als Schätzwerte zweier Parameterschätzungen in den Modellen 1 und 2 nach der Methode der kleinsten Quadrate ergeben, wenn die Beobachtungen y_i bei der Ausgleichung im Modell 1 mit den Gewichten $\mu_{i1}^{(j)}$ und bei der Ausgleichung im Modell 2 mit den Gewichten $\mu_{i2}^{(j)}$ versehen werden. Gemäß [Koch 1998], Satz (323.3) ergeben sich die Schätzungen

$$\boxed{\hat{\boldsymbol{\beta}}_1^{(j+1)}=(\overset{1}{\boldsymbol{M}}^T\boldsymbol{P}_1^{(j)}\overset{1}{\boldsymbol{M}})^{-1}\overset{1}{\boldsymbol{M}}^T\boldsymbol{P}_1^{(j)}\boldsymbol{y}\quad\text{und}\quad\hat{\boldsymbol{\beta}}_2^{(j+1)}=(\overset{2}{\boldsymbol{M}}^T\boldsymbol{P}_2^{(j)}\overset{2}{\boldsymbol{M}})^{-1}\overset{2}{\boldsymbol{M}}^T\boldsymbol{P}_2^{(j)}\boldsymbol{y}.} \tag{422.36}$$

Für die unbekannten Standardabweichungen σ_1 und σ_2 der zu den Modellen 1 und 2 gehörenden Beobachtungen sollen nun erwartungstreue Schätzungen $\hat{\sigma}_1^{(j+1)}$ und $\hat{\sigma}_2^{(j+1)}$ abgeleitet werden. Nach [Koch 1998], Gl. (325.6) gilt

$$(\hat{\sigma}_1^{(j+1)})^2=\frac{\Omega_1^{(j)}}{N-u_1}\qquad\text{und}\qquad(\hat{\sigma}_2^{(j+1)})^2=\frac{\Omega_2^{(j)}}{N-u_2}, \tag{422.37}$$

worin $\Omega_1^{(j)}$ und $\Omega_2^{(j)}$ die Quadratsummen der geschätzten Residuen in den Modellen 1 und 2 bezeichnen, die sich im j-ten Iterationsschritt mit den Schätzungen $\hat{\boldsymbol{\beta}}_1^{(j+1)}$ bzw. $\hat{\boldsymbol{\beta}}_2^{(j+1)}$ ergeben. u_1 und u_2 bezeichnen jeweils die Anzahl der unbekannten Parameter in Modell 1 und Modell 2; hier gilt $u_1 = 2$ und $u_2 = 3$.

Die Vektoren $\hat{\boldsymbol{r}}_1^{(j)}$ und $\hat{\boldsymbol{r}}_2^{(j)}$ der geschätzten Beobachtungsresiduen in den Modellen 1 und 2 ergeben sich in der j-ten EM-Iteration zu

$$\hat{\boldsymbol{r}}_1^{(j)} = \overset{1}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_1^{(j+1)} - \boldsymbol{y} \qquad \text{und} \qquad \hat{\boldsymbol{r}}_2^{(j)} = \overset{2}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_2^{(j+1)} - \boldsymbol{y}. \tag{422.38}$$

Hiermit erhält man in der j-ten EM-Iteration die Quadratsummen Ω_1 und Ω_2 der Beobachtungsresiduen

$$\Omega_1^{(j)} = (\hat{\boldsymbol{r}}_1^{(j)})^T \cdot \boldsymbol{P}_1^{(j)} \cdot \hat{\boldsymbol{r}}_1^{(j)} \qquad \text{und} \qquad \Omega_2^{(j)} = (\hat{\boldsymbol{r}}_2^{(j)})^T \cdot \boldsymbol{P}_2^{(j)} \cdot \hat{\boldsymbol{r}}_2^{(j)}. \tag{422.39}$$

Durch Einsetzen von (422.39) und (422.38) in (422.21) ergeben sich die Schätzwerte $\hat{\sigma}_1^{(j+1)}$ und $\hat{\sigma}_2^{(j+1)}$ für die Standardabweichungen der zu den Modellen 1 (Hochspannungsleitung) und 2 (Straße) gehörenden Beobachtungen

$$\boxed{\hat{\sigma}_1^{(j+1)} = \left(\frac{(\overset{1}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_1^{(j+1)} - \boldsymbol{y})^T \cdot \boldsymbol{P}_1^{(j)} \cdot (\overset{1}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_1^{(j+1)} - \boldsymbol{y})}{N-2} \right)^{\frac{1}{2}}} \tag{422.40}$$

und

$$\boxed{\hat{\sigma}_2^{(j+1)} = \left(\frac{(\overset{2}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_2^{(j+1)} - \boldsymbol{y})^T \cdot \boldsymbol{P}_2^{(j)} \cdot (\overset{2}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_2^{(j+1)} - \boldsymbol{y})}{N-3} \right)^{\frac{1}{2}}} \tag{422.41}$$

Mit (422.36), (422.40) und (422.41) sind alle im M-Schritt des EM-Algorithmus zu berechnenden Größen bestimmt.

Abbruchkriterium, Kovarianzmatrizen der Schätzwerte und Klassifikation der Beobachtungen

Abbruchkriterium: Nachdem die im E-Schritt und im M-Schritt des zur Lösung der in Abschnitt 41 Schätz- und Klassifikationsaufgabe zu implementierenden EM-Algorithmus abgeleitet wurden, muß noch ein Kriterium eingeführt werden, aufgrund dessen am Ende jeder Iteration entschieden wird, ob die EM-Sequenz nochmals durchlaufen wird oder nicht. Zur Definition eines solchen Abbruchkriteriums wird hier die Quadratwurzel

$$\delta = \left(\|\hat{\boldsymbol{\beta}}_1^{(j+1)} - \hat{\boldsymbol{\beta}}_1^{(j)}\|^2 + \|\hat{\boldsymbol{\beta}}_2^{(j+1)} - \hat{\boldsymbol{\beta}}_2^{(j)}\|^2 \right)^{1/2} \tag{422.42}$$

aus der Quadratsumme der euklidschen Abstände der Schätzungen $\hat{\boldsymbol{\beta}}_1^{(j+1)}$ und $\hat{\boldsymbol{\beta}}_2^{(j+1)}$ der j-ten Iteration von den entsprechenden Schätzungen $\hat{\boldsymbol{\beta}}_1^{(j)}$ und $\hat{\boldsymbol{\beta}}_2^{(j)}$ der $j-1$ten Iteration betrachtet. Bleibt diese Quadratsumme kleiner als ein vorzugebender Schwellwert ϵ, was gleichbedeutend damit ist, daß sich die Schätzungen der unbekannten Parameter

zwischen zwei Iterationen um weniger als ϵ ändern, so sollen die EM-Iterationen abgebrochen werden. Sonst wird mit der nächsten Iteration begonnen. Es ergibt sich also das **Abbruchkriterium:**

$$\text{Falls } \delta > \epsilon : \text{Übergang zur nächsten EM-Iteration} \tag{422.43}$$

$$\text{sonst : Abbruch der EM-Iterationen} \tag{422.44}$$

Für ϵ wird im folgenden der Wert $1 \cdot 10^{-9}$ gewählt. Die sich ind er letzten EM-Iteration ergebende Schätzung $\hat{\boldsymbol{\beta}}^{(j+1)} = \left[(\hat{\boldsymbol{\beta}}_1^{(j+1)})^T, \hat{\sigma}_1^{(j+1)}, (\hat{\boldsymbol{\beta}}_2^{(j+1)})^T, \hat{\sigma}_2^{(j)}\right]$ für die unbekannten Parameter $\hat{\boldsymbol{\beta}} = \left[\hat{\boldsymbol{\beta}}_1^T, \sigma_1, \hat{\boldsymbol{\beta}}_2^T, \sigma_2\right]$ wird im folgenden mit $\hat{\boldsymbol{\beta}}^* = \left[(\hat{\boldsymbol{\beta}}_1^{(*)})^T, \hat{\sigma}_1^{(*)}, (\hat{\boldsymbol{\beta}}_2^{(*)})^T, \hat{\sigma}_2^{(*)}\right]^T$ bezeichnet.

Kovarianzmatrizen der Schätzungen: Schätzwerte für unbekannte Parameter, die im Rahmen einer Ausgleichung bestimmt werden, sind für sich allein nur beschränkt aussagekräftig, da aus den Schätzwerten zunächst nicht hervorgeht, wie sicher bzw. wie unsicher die Schätzung ist. Daher werden zum Abschluß der EM-Iterationen noch Schätzungen $\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_1^{(*)})$ und $\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_2^{(*)})$ für die Kovarianzmatrizen $\boldsymbol{D}(\hat{\boldsymbol{\beta}}_1^{(*)})$ und $\boldsymbol{D}(\hat{\boldsymbol{\beta}}_2^{(*)})$ der geschätzten Parameter $\hat{\boldsymbol{\beta}}_1^{(*)}$ und $\hat{\boldsymbol{\beta}}_2^{(*)}$ angegeben. Aus den Quadratwurzeln der Diagonalelemente dieser Matrizen gehen die geschätzten Standardabweichungen der Schätzwerte hervor, die hier als Maße für die Genauigkeit der einzelnen Schätzwerte dienen. Aus den Nebendiagonalelementen lassen sich Korrelationen zwischen den einzelnen Parametern ableiten, also Maße für statistische Abhängikeiten zwischen den Schätzwerten. Nach [KOCH 1998], Gl. (325.7) ergeben sich die Schätzungen

$$\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_1^{(*)}) = (\hat{\sigma}_1^{(*)})^2 \cdot (\overset{1}{\boldsymbol{M}}{}^T \boldsymbol{P}_1^{(*)} \overset{1}{\boldsymbol{M}})^{-1} \qquad \text{und} \tag{422.45}$$

$$\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_2^{(*)}) = (\hat{\sigma}_2^{(*)})^2 \cdot (\overset{2}{\boldsymbol{M}}{}^T \boldsymbol{P}_2^{(*)} \overset{2}{\boldsymbol{M}})^{-1} \tag{422.46}$$

für die beiden Kovarianzmatrizen der in den Modellen 1 und 2 erhaltenen Schätzwerte $\hat{\boldsymbol{\beta}}_1^{(*)}$ und $\hat{\boldsymbol{\beta}}_2^{(*)}$.

Klassifikation der Beobachtungen: Die angestrebte Klassifikation der Beobachtungen, also die Zuordnung der einzelnen Beobachtungen jeweils entweder zur Hochspannungsleitung (Objekt $\mathcal{O}_1$) oder zur Straße (Objekt $\mathcal{O}_2$ kann mittels der sich nach der letzten EM-Iteration im E-Schritt ergebenden Erwartungswerte $\mu_{i1}^{(*)}$ und $\mu_{i2}^{(*)}$, $i = 1, \ldots, N$ auf einfache Weise erfolgen: Nach (422.25) und (422.26) sind $\mu_{i1}^{(*)}$ und $\mu_{i2}^{(*)}$ die Wahrscheinlichkeiten, daß die Beobachtung y_i zu Modell 1 (d.h. zur Hochspannungsleitung) bzw. zu Modell 2 (d.h. zur Straße) gehört. Es liegt nun nahe, jede Beobachtung y_i dem Modell M_k, $k \in \{1, 2\}$ zuzuordnen, für das die Wahrscheinlichkeit $\mu_{ik}^{(*)}$ am größten ist[7]. Ist also y_i eine Beobachtung und ist o_i die Bezeichnung des Objektes, zu dem y_i gehört ($o_i = 1$, falls y_i zur Hochspannungsleitung gehört, $o_i = 2$, falls y_i zur Straße gehört), so ergeben sich die geschätzte Objektbezeichnungen $\hat{o}_i$ für $i = 1, \ldots, N$ mit

$$\hat{o}_i = \begin{cases} 1 & , \text{ falls } \mu_{i1}^{(*)} > \mu_{i2}^{(*)} \\ 2 & , \text{ falls } \mu_{i1}^{(*)} \leq \mu_{i2}^{(*)}. \end{cases} \qquad , \qquad i = 1, \ldots, N \tag{422.47}$$

[7] Diese Vorgehensweise entspricht einer Maximum-Likelihood-Schätzung der Objektbezeichnung o_i, $o_i \in \{1, 2\}$ des zur Beobachtung y_i gehörenden Objektes

Die in diesem Abschnitt abgeleiteten Rechenformeln, die zur Lösung der zu Beginn formulierten Aufgabe mittels des EM-Algorithmus benötigt werden, sollen nun zusammengefasst werden.

Zusammenfassung der Ergebnisse

In Übersicht (4.1) ist der zur Lösung der in Abschnitt 41 gestellten Aufgabe zu implementierende Algorithmus mit allen erforderlichen Rechenschritten dargestellt. Aus dem dargestellten Ablaufschema wird klar, daß es sich bei dem hier vorliegenden EM-Algorithmus um eine Parameterschätzung mit *iterativer Regewichtung der Beobachtungen* handelt: Im E-Schritt der j-ten Iteration werden auf Grundlage der in der vorhergehenden Iteration berechneten Schätzwerte $\hat{\boldsymbol{\beta}}^{(j)}$ die Beobachtungen für die nachfolgenden Parameterschätzungen in den Modellen 1 und 2 neu gewichtet: Eine Beobachtung y_i, die im j-ten Iterationsschritt besser durch das sich mit den Schätzwerten $\hat{\boldsymbol{\beta}}^{(j)}$ ergebende Modell 1 als durch das entsprechende Modell 2 beschrieben wird, wird bei der nächsten Parameterschätzung in Modell 1 herauf- und in Modell 2 herabgewichtet. Analog dazu wird sie bei der Neuschätzung in Modell 2 herauf- und in Modell 1 herabgewichtet, wenn sie besser durch das sich aus der vorhergehenden Iteration ergebende Modell 2 als durch das entsprechende Modell 1 beschrieben wird.

Auf Dauer führt diese iterative Neugewichtung der Beobachtungen dazu, daß Beobachtungen, die nicht zu dem Modell 1 (d.h. zur Hochspannungsleitung) gehören, bei der Parameterschätzung in diesem Modell ein solch geringes Gewicht erhalten, daß sie praktisch keinen Einfluß mehr auf die Schätzwerte für die unbekannten Parameter dieses Modells nehmen. Dann bestimmen nur noch die tatsächlich zur Hochspannungsleitung gehörenden Beobachtungen die Schätzwerte für die unbekannten Parameter des Modells, das die zur Hochspannungsleitung gehörenden Beobachtungen beschreibt. Genau so erhalten die nicht zur Straße gehörenden Beobachtungen bei den Parameterschätzungen in Modell 2 ein solch geringes Gewicht, daß sie sich auf die Schätzungen in diesem Modell kaum mehr auswirken. Somit ist der sich hier ergebende EM-Algorithmus eng verwandt mit Methoden zur robusten Parameterschätzung[8] (vgl. [KOCH 1998],Kap.3.8), bei denen Ausreißer in den Beobachtungsdaten durch iterative Regewichtung der Beobachtungen sukzessive herabgewichtet werden, bis ihr Einfluß auf das Schätzergebnis vernachlässigbar gering ist.

Der in diesem Abschnitt beschriebene Algorithmus wurde in der in Übersicht 4.1 dargestellten Form mittels der Programmiersprache C in einem Programm mit dem Namen `EM.c` umgesetzt. (Der Quellcode dieses Programms befindet sich in Anhang A1.) Das Programm liest eine Messreihe der Form

`40` *(Anzahl der Koordinatenpaare)*	
(u-Wert)	*(y-Wert)*
1	15.18094175
2	13.87552839
3	6.03249538
4	7.420570315
5	9.613662336
⋮	⋮

aus der Datei `Linien.dat` ein und bestimmt aus dem Datenmaterial in der oben beschrie-

[8] Es sei darauf hingewiesen, daß dies nur für das hier vorliegende Beispiel gilt.

benen Weise Schätzwerte für die unbekannten Parameter einer ausgleichenden Gerade und eines ausgleichenden Polynoms 2ten Grades. Hierbei erfolgt eine Klassifikation der Beobachtungen nach der Zugehörigkeit zum linearen Modell oder zum Polynommodel 2ten Grades. Die Ergebnisse werden in die Datei `Erg.dat` ausgegeben.

423 Ergebnisse

Das Programm `EM.c` wurde auf den Datensatz `Linien.dat` (vgl. Anhang A2) angewendet. Abbildung (2) visualisiert die eingelesenen Daten:

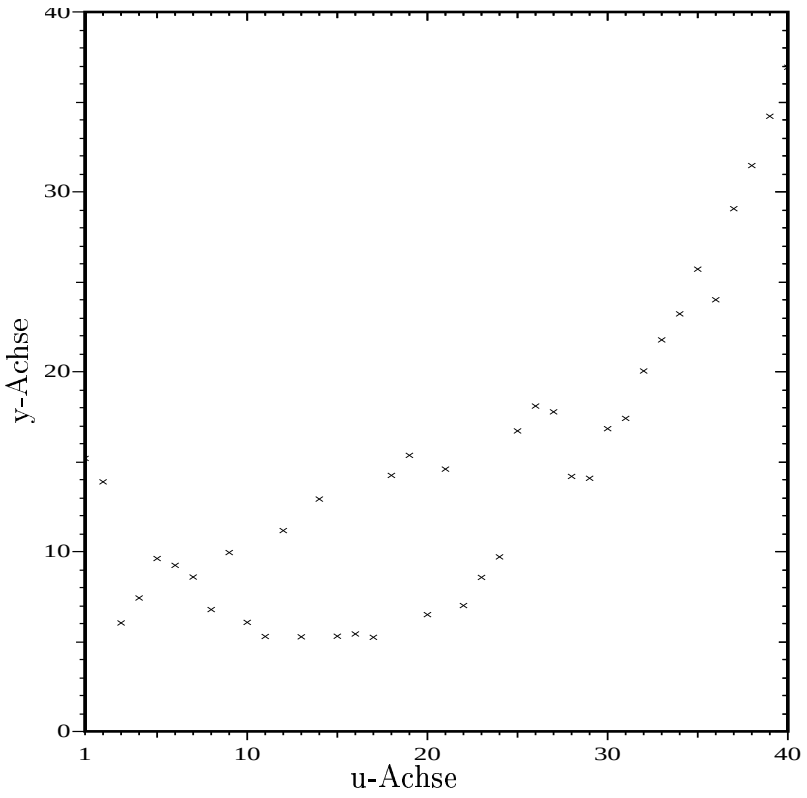


Abbildung 2: Visualisierung der Koordinatenpaare aus dem Datensatz `Linien.dat`

Mit diesen Eingangsdaten wurden mit dem Programm `EM.c` folgende Ergebnisse erhalten, wobei für das Abbruchkriterium der Schwellwert $\epsilon = 1 \cdot 10^{-9}$ gewählt wurde:

Schätzwerte des Modells 1 (Hochspannungsleitung):

Als *Schätzwerte für die Geradenparameter* wurden erhalten:

$$\hat{\boldsymbol{\beta}}_1^{(*)} = \begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \end{bmatrix} = \begin{bmatrix} 5.1330 \\ 0.4985 \end{bmatrix} \qquad \text{mit} \qquad \begin{matrix} \hat{\sigma}_{a_0} = 0.214983 \\ \hat{\sigma}_{a_1} = 0.010525 \end{matrix}. \tag{423.1}$$

Hierbei wurden die Standardabweichungen mittels der unten aufgeführten Kovarianzmatrix berechnet. Aufgrund der kleinen Standardabweichungen können die Schätzwerte als relativ sicher und die Modellparameter a_0 und a_1 als signifikant gelten[9]. Mit den Schätzwerten ergibt sich die *ausgleichende Gerade* im Bild, d.h. die Gerade, die die Mittellinie der Hochspannungsleitung im Bild approximiert, zu

$$\hat{y}_1(u) = 5.1330 + 0.4985 \cdot u \tag{423.2}$$

Als *geschätzte Kovarianzmatrix des Vektors* $\hat{\boldsymbol{\beta}}_1^{(*)}$ wurde die Matrix

$$\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_1^{(*)}) = \begin{bmatrix} \hat{\sigma}_{a_0}^2 & \hat{\sigma}_{a_0,a_1} \\ \hat{\sigma}_{a_1,a_0} & \hat{\sigma}_{a_1}^2 \end{bmatrix} = \begin{bmatrix} 0.046218 & -0.001987 \\ -0.001987 & 0.000111 \end{bmatrix} \tag{423.3}$$

erhalten. Die geschätzte Standardabweichung der Gewichtseinheit der zur Hochspannungsleitung gehörenden Beobachtungen ergibt sich zu $\hat{\sigma}_1^{(*)} = 0.3687$

Schätzwerte des Modells 2 (Straße):

Es ergaben sich folgende *Schätzwerte für die Koeffizienten des Polynoms 2ten Grades*:

$$\hat{\boldsymbol{\beta}}_2^{(*)} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} 16.3266 \\ -1.5097 \\ 0.0505 \end{bmatrix} \quad \text{mit} \quad \begin{aligned} \hat{\sigma}_{b_0} &= 0.214657 \\ \hat{\sigma}_{b_1} &= 0.024097 \\ \hat{\sigma}_{b_2} &= 0.000560 \end{aligned} \tag{423.4}$$

Auch hier sind die erhaltenen Standardabweichungen der Schätzwerte im Vergleich zu den Schätzwerten selbst klein, was ein Beleg für die Signifikanz der Modellparameter ist.[9] Weiter können aufgrund dieser kleinen Standardabweichungen die Schätzungen als relativ sicher gelten. Mit den Schätzwerten erhält man das *ausgleichende Polynom 2ten Grades* im Bild, d.h. das Polynom 2ten Grades, das die Mittellinie der Straße im Bild approximiert, zu

$$\hat{y}_2(u) = 16.3266 - 1.5097 \cdot u + 0.0505 \cdot u^2 \tag{423.5}$$

Als *Geschätzte Kovarianzmatrix des Vektors* $\hat{\boldsymbol{\beta}}_2^{(*)}$ wurde die Matrix

$$\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_2^{(*)}) = \begin{bmatrix} \hat{\sigma}_{b_0}^2 & \hat{\sigma}_{b_0,b_1} & \hat{\sigma}_{b_0,b_2} \\ \hat{\sigma}_{b_1,b_0} & \hat{\sigma}_{b_1}^2 & \hat{\sigma}_{b_1,b_2} \\ \hat{\sigma}_{b_2,b_0} & \hat{\sigma}_{b_2,b_1}^2 & \hat{\sigma}_{b_2}^2 \end{bmatrix} = \begin{bmatrix} 0.046078 & -0.004504 & 0.000091 \\ -0.004504 & 0.000581 & -0.000013 \\ 0.000091 & -0.000013 & 0.0000003 \end{bmatrix} \tag{423.6}$$

erhalten. Die geschätzte Standardabweichung der Gewichtseinheit der zur Straße gehörenden Beobachtungen beträgt ebenfalls $\hat{\sigma}_2^{(*)} = 0.3687$. Dies deutet auf eine einheitliche Meßgenauigkeit der zur Hochspannungsleitung und zur Straße gehörenden Beobachtungen hin.

Die Ergebnisse der Klassifikation der Beobachtungen gehen aus Übersicht 423 hervor.Hierin gelten folgende Bezeichnungen:

u : (feste) Ordinatenwerte der Koordinatenpaare des Eingabedatensatzes

y : gemessene Abszissenwerte des Eingabedatensatzes

$\mu_{i1}^{(*)}$,$\mu_{i2}^{(*)}$: Wahrscheinlichkeiten der Zugehörigkeit eines Koordinatenpaars zur HS-Leitung bzw. zur Straße

$\hat{o}_i$: Klassifikationsergebnis: Geschätzte Objektbezeichnung des Objektes, zu der eine Beobachtung y_i gehört. (HS-Leitung:$o_i = 1$, Straße: $o_i = 2$)

$\hat{r}_{i1}$, $\hat{r}_{i2}$: Beobachtungsresiduen im Modell 1 bzw. im Modell 2. Für jede Beobachtung ist das Residuum nur in dem Modell angegeben, dem die Beobachtung bei der Klassifikation zugeordnet wird.

$\hat{y}_{i1}$, $\hat{y}_{i2}$: geschätzte Erwartungswerte der Beobachtungen. Der Erwartungswert wird nur in dem Modell angegeben, dem die Beobachtung bei der Klassifikation zugeordnet wird.

Es wird deutlich, daß alle Beobachtungsresiduen klein bleiben, was darauf hindeutet, daß keine Fehlklassifikationen von Beobachtungen vorliegen.

Die Ergebnisse von Parmeterschätzung und Klassifikation sind in Abbildung 3 visualisiert. Hieraus wird offensichtlich, daß alle Koordinatenpaare des Eingangsdatensatzes den

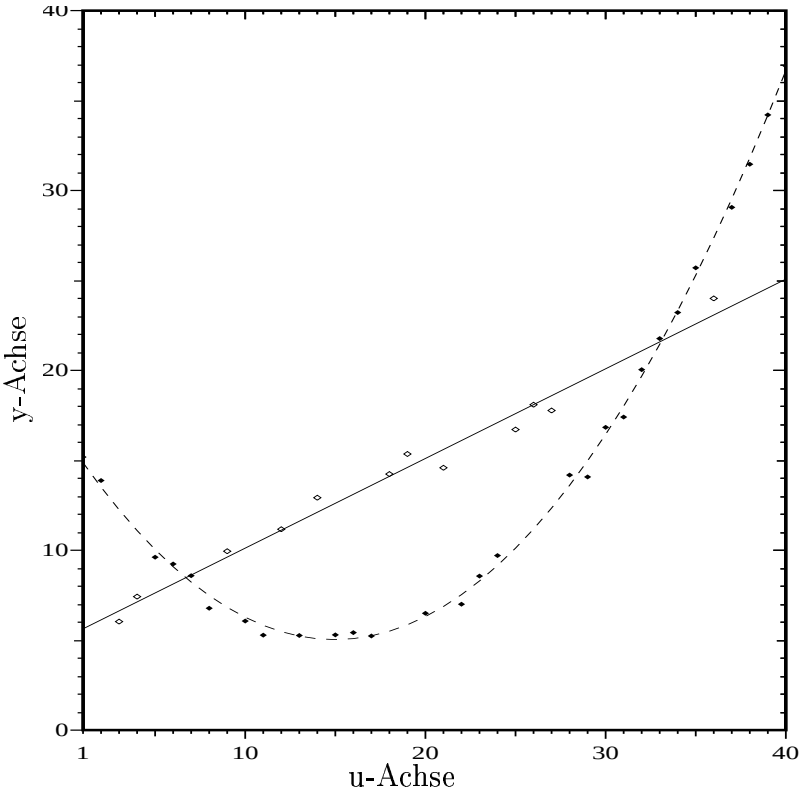


Abbildung 3: Visualisierung der Ergebnisse von Parameterschätzung und Klassifikation: Die durchgezogene Linie stellt die mittels des EM-Algorithmus bestimmte ausgleichende Gerade, die gestrichelte Linie das ausgleichende Polynom 2ten Grades dar. Die voll gezeichneten Punkte stellen Datenpunkte des Eingangsdatensatzes dar, die vom Programm als zur Straße (also zum Modell 2) gehörig klassifiziert wurden; die nicht ausgefüllten Punkte stellen Datenpunkte dar, die als zur Hochspannungsleitung (Objekt 1, Gerade) gehörig klassifiziert wurden.

Objekten 1 und 2 korrekt zugeordnet werden. Die beiden ausgleichenden Kurven approximieren das Datenmaterial augenscheinlich optimal.

Dieses optimale Ergebnis der Parameterschätzung und Klassifikation erklärt sich z.T. dadurch, daß sich die Objekte 1 und 2 in ihrer Form klar unterscheiden und daß die Messwertstreuung im Beispiel relativ klein gewählt wurde, so daß in Abbildung 2 die Beobachtungen klar dem einen oder anderen Objekt zugeschrieben werden können. Es wäre also noch zu testen,wie der Algorithmus auf stark verrauschten Daten arbeitet, bzw. was passiert, wenn die Objekte im Bild nicht so deutlich voneinander unterschieden werden können. (Z.B. wenn sich die zwei gerade Objektmittellinien im Bild schleifend schneiden.) Diese Untersuchungen sollen jedoch nicht zum weiteren Inhalt dieser Arbeit gemacht werden, da es hier nur darum geht, eine mögliche Anwendung des EM-Algorithmus *vorzustellen.*

[9]Die 3fache Standardabweichung jedes Schätzwertes ist kleiner als der Schätzwert selbst, so daß mit einer Irrtumswahrscheinlichkeit von weniger als 0,001 die Parameter des Modells von 0 verschieden sind.

Abschließend soll die iterative Regewichtung der Beobachtungen im Rahmen des hier angewandten EM-Algorithmus noch veranschaulicht werden: In Abbildung 423 sind beispielhaft die Gewichte $\mu_{31}^{(j)}$ und $\mu_{32}^{(j)}$ der Beobachtung y_3 bei den Parameterschätzungen in den Modellen 1 und 2 gegen die laufenden Nummern j der Iterationen aufgetragen. Man er-

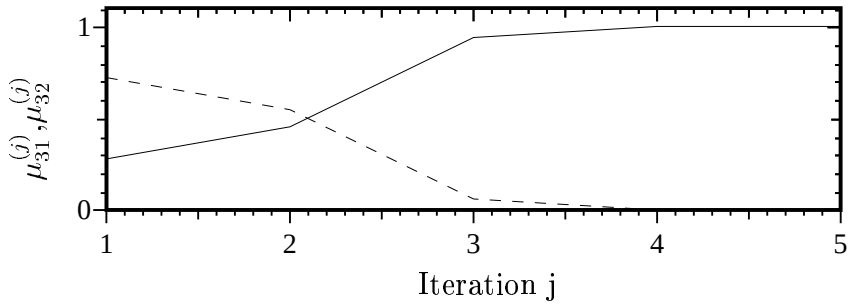


Abbildung 4: Visualisierung der iterativen Regewichtung der Beobachtungen im Rahmen des hier angewandten EM-Algorithmus. Die durchgezogene Linie stellt den Verlauf der Gewichtung der Beobachtung y_3 bei den Parameterschätzungen im Modell 1 dar, die gestrichelte Linie zeigt den Verlauf der Gewichtung dieser Beobachtung bei den Schätzungen in Modell 2.

kennt,daß die Beobachtung y_3 in Modell 1 zunächst ein geringes Gewicht hat. Im Rahmen des hier angewendeten EM-Algorithmus wird sie in diesem Modell jedoch von Iteration zu Iteration z.T. drastisch heraufgewichtet und entsprechend im Modell 2 herabgewichtet, so daß sie sich am Ende nur noch auf die Schätzwerte der Parameterschätzungen in Modell 1 auswirkt. In Modell 2 konvergiert ihr Gewicht gegen Null. Die Gewichte konvergieren sehr schnell (nach 5 Iterationsschritten ist keine Veränderung mehr erkennbar), so daß der hier implementierte EM-Algorithmus relativ schnell konvergiert.

Übersicht 4.1: Zusammenfassung der Rechenschritte

EM-Algorithmus zur Lösung der Aufgabe:

1. **Bestimmung von Näherungswerten:** Näherungswerte $\hat{\boldsymbol{\beta}}^{(0)}$ für die unbekannten Parameter $\boldsymbol{\beta}$ der beiden Modelle 1 und 2 werden durch zwei initiale Parameterschätzungen in diesen Modellen nach der Methode der kleinsten Quadrate erhalten. Als Gewichtsmatrix der Beobachtungen $\boldsymbol{y}$ wird dabei jeweils die Einheitsmatrix $\boldsymbol{P}_1^{(0)} = \boldsymbol{P}_2^{(0)} = \boldsymbol{I}$ gewählt. Es ergeben sich also für $k = 1, 2$ die Näherungswerte

$$\hat{\boldsymbol{\beta}}_k^{(0)} = (\overset{k}{\boldsymbol{M}}{}^T \overset{k}{\boldsymbol{M}})^{-1} \overset{k}{\boldsymbol{M}}{}^T \boldsymbol{y} \quad , \quad \sigma_k^{(0)} = \left[(\overset{k}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_k^{(0)} - \boldsymbol{y})^T \cdot (\overset{k}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_k^{(0)} - \boldsymbol{y})/(N-k-1)\right]^{1/2}$$

2. **EM-Iterationen:**

   i.) **E-Schritt:** Berechnung der Erwartungswerte $\mu_{ik}^{(j)}$ für $k = 1, 2$ und $i = 1, \ldots, N$ mittels

$$\mu_{ik}^{(j)} = \frac{\frac{k}{3\sqrt{2\pi}\hat{\sigma}_k^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_k^{(j)})^2}[y_j - \overset{k}{\boldsymbol{m}}_j^T \hat{\boldsymbol{\beta}}_k^{(j)}]^2}}{\frac{1}{3\sqrt{2\pi}\hat{\sigma}_1^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_1^{(j)})^2}[y_j - \overset{1}{\boldsymbol{m}}_j^T \hat{\boldsymbol{\beta}}_1^{(j)}]^2} + \frac{2}{3\sqrt{2\pi}\hat{\sigma}_2^{(j)}} e^{-\frac{1}{2(\hat{\sigma}_2^{(j)})^2}[y_j - \overset{2}{\boldsymbol{m}}_j^T \hat{\boldsymbol{\beta}}_2^{(j)}]^2}}$$

   ii.) **M-Schritt:** Aufstellen der Matrizen $\boldsymbol{P}_k^{(j+1)} = diag(\mu_{1k}^{(j+1)}, \mu_{2k}^{(j+1)}, \ldots, \mu_{Nk}^{(j+1)})$ für $k = 1, 2$ und Berechnung der Schätzungen

$$\hat{\boldsymbol{\beta}}_k^{(j+1)} = (\overset{k}{\boldsymbol{M}}{}^T \boldsymbol{P}_k^{(j+1)} \overset{k}{\boldsymbol{M}})^{-1} \overset{k}{\boldsymbol{M}}{}^T \boldsymbol{P}_k^{(j+1)} \boldsymbol{y} \qquad \text{und}$$
$$\hat{\sigma}_k^{(j+1)} = \left[(\overset{k}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_k^{(j+1)} - \boldsymbol{y})^T \boldsymbol{P}_k^{(j+1)} (\overset{k}{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_k^{(j+1)} - \boldsymbol{y})/(N-k-1)\right]^{1/2}$$

   für $k = 1, 2$

   iii.) **Abbruchkriterium:** Falls $\left(\hat{\boldsymbol{\beta}}_1^{(k+1)} - \hat{\boldsymbol{\beta}}_1^{(k)}\right)^2 + \left(\hat{\boldsymbol{\beta}}_2^{(k+1)} - \hat{\boldsymbol{\beta}}_2^{(k)}\right)^2 < \epsilon$ gilt, wird der Algorithmus bei 3. fortgesetzt, sonst ist eine weitere EM-Iteration erforderlich, so daß der Algorithmus bei i.) fortgesetzt wird.

3. **Berechnung der Kovarianzmatrizen und Klassifikation der Beobachtungen:** Für $k = 1, 2$ ergeben sich die Kovarianzmatrizen der Schätzungen $\hat{\boldsymbol{\beta}}_1^{(*)}$ und $\hat{\boldsymbol{\beta}}_2^{(*)}$ zu

$$\hat{\boldsymbol{D}}(\hat{\boldsymbol{\beta}}_k^{(*)}) = (\hat{\sigma}_k^{(*)})^2 \cdot (\overset{k}{\boldsymbol{M}}{}^T \boldsymbol{P}_k^{(*)} \overset{k}{\boldsymbol{M}})^{-1}.$$

   Die geschätzten Bezeichnungen $\hat{o}_i$ der Objekte, zu denen die Beobachtungen y_i gehören, ergeben sich mit

$$\hat{o}_i = \begin{cases} 1(\text{ HS-Leitung}) & , \text{ falls } \mu_{i1}^{(*)} > \mu_{i2}^{(*)} \\ 2\ (\text{ Straße}) & , \text{ falls } \mu_{i1}^{(*)} \leq \mu_{i2}^{(*)}. \end{cases} \quad , \qquad i = 1, \ldots, N$$

Übersicht 4.2: Klassifikationsergebnisse und Beobachtungsresiduen

u	y	$\mu_{i1}^{(*)}$	$\mu_{i2}^{(*)}$	$\hat{o}_i$	$\hat{r}_{i1}$	$\hat{r}_{i2}$	$\hat{y}_{i1}$	$\hat{y}_{i2}$
1.0	15.181	0.000	1.000	2	–	0.000	–	14.867
2.0	13.876	0.000	1.000	2	–	0.000	–	13.509
3.0	6.032	1.000	0.000	1	0.596	–	6.628	–
4.0	7.421	1.000	0.000	1	0.000	–	7.127	–
5.0	9.614	0.000	1.000	2	–	0.426	–	10.040
6.0	9.236	0.006	0.994	2	–	0.000	–	9.085
7.0	8.585	0.440	0.560	2	–	0.000	–	8.231
8.0	6.777	0.000	1.000	2	–	0.701	–	7.478
9.0	9.945	1.000	0.000	1	0.000	–	9.619	–
10.0	6.065	0.000	1.000	2	–	0.210	–	6.275
11.0	5.280	0.000	1.000	2	–	0.545	–	5.825
12.0	11.168	1.000	0.000	1	0.000	–	11.115	–
13.0	5.261	0.000	1.000	2	–	0.000	–	5.227
14.0	12.919	1.000	0.000	1	0.000	–	12.111	–
15.0	5.295	0.000	1.000	2	–	0.000	–	5.033
16.0	5.421	0.000	1.000	2	–	0.000	–	5.087
17.0	5.234	0.000	1.000	2	–	0.009	–	5.243
18.0	14.235	1.000	0.000	1	0.000	–	14.105	–
19.0	15.350	1.000	0.000	1	0.000	–	14.604	–
20.0	6.495	0.000	1.000	2	–	0.000	–	6.314
21.0	14.583	1.000	0.000	1	1.017	–	15.601	–
22.0	7.000	0.000	1.000	2	–	0.532	–	7.532
23.0	8.564	0.000	1.000	2	–	0.000	–	8.293
24.0	9.705	0.000	1.000	2	–	0.000	–	9.154
25.0	16.707	1.000	0.000	1	0.887	–	17.595	–
26.0	18.089	1.000	0.000	1	0.004	–	18.093	–
27.0	17.768	1.000	0.000	1	0.824	–	18.591	–
28.0	14.179	0.000	1.000	2	–	0.000	–	13.609
29.0	14.075	0.000	1.000	2	–	0.900	–	14.975
30.0	16.832	0.000	1.000	2	–	0.000	–	16.442
31.0	17.404	0.000	1.000	2	–	0.606	–	18.010
32.0	20.039	0.014	0.986	2	–	0.000	–	19.679
33.0	21.767	0.388	0.612	2	–	0.000	–	21.448
34.0	23.214	0.004	0.996	2	–	0.105	–	23.319
35.0	25.699	0.000	1.000	2	–	0.000	–	25.290
36.0	23.998	1.000	0.000	1	0.000	–	23.078	–
37.0	29.062	0.000	1.000	2	–	0.474	–	29.536
38.0	31.459	0.000	1.000	2	–	0.351	–	31.810
39.0	34.200	0.000	1.000	2	–	0.000	–	34.185
40.0	36.924	0.000	1.000	2	–	0.000	–	36.661

Kapitel 5

Zusammenfassung

Die Photogrammetrie ist untrennbar mit Methoden der Ausgleichungsrechnung und Statistik verbunden; dies gilt sowohl für die klassische Photogrammetrie als auch für moderne Verfahren der digitalen Bildverarbeitung. Klassische Methoden der geodätischen Ausgleichungsrechnung reichen allerdings zur Lösung bestimmter Probleme aus der Photogrammetrie nicht immer aus, wie in Abschnitt 122 anhand eines einfachen Beispiels gezeigt wurde. Zur Lösung solcher Probleme müssen diese Parameterschätzverfahren an die konkrete Aufgabe angepaßt, erweitert, ergänzt oder durch andere Parameterschätzverfahren ersetzt werden. Bestimmte Aufgaben, bei denen die klassischen Schätzverfahren versagen, können mittels Schätzverfahren aus unvollständigen Beobachtungsdaten - wenn auch z.T. unter erheblichem Rechenaufwand – gelöst werden. Ein Beispiel hierfür sind Parameterschätzungen, bei denen die zugrundeliegenden Beobachtungen hinsichtlich ihrer Zugehörigkeit zu einem von mehreren alternativ gültigen Modellen unbestimmt sind (vgl. Abschnitt 122 und Kapitel 4).

Der in dieser Arbeit diskutierte EM-Algorithmus stellt ein solches Parameterschätzverfahren dar, mit dem Parameter aus unvollständigen Beobachtungsdaten geschätzt werden können. Er wurde erstmals im Jahre 1968 von den amerikanischen Professoren A. Dempster, N. Laird und D.Rubin unter dem Namen ,,EM-Algorithmus" vorgestellt (vgl. [Dempster et al. 1968]) und ergab sich damals als verallgemeinerte Zusammenfassung mehrerer Einzellösungen von Parameterschätzproblemen, die von verschiedenen Autoren entwickelt worden waren. Zwar gestaltet sich die Umsetzung des EM-Algorithmus in bezug auf eine konkrete Aufgabe häufig einfach, eine allgemeine Diskussion der zugrundeliegenden Theorie stellt jedoch eine sehr komplexe Aufgabe dar[1]. Die überwiegend englischsprachige Literatur zu diesem Thema richtet sich überwiegend an Mathematiker und Informatiker, deren Fachtermini sich z.T. von der im Rahmen der geodätischen Ausgleichungsrechnung verwendeten Terminologie unterscheiden.

Um dem aus dem Bereich der Geodäsie stammenden Leser den Zugang zum EM-Algorithmus zu erleichtern, wurde in Abschnitt 13 dieser Arbeit ausgehend von den klassischen geodätischen Parameterschätzverfahren das Prinzip der Parameterschätzung aus unvollständigen Beobachtungsdaten erläutert. Anschließend wurde in Kapitel 2 schrittweise der EM-Algorithmus abgeleitet: Ausgehend von der in Abschnitt 22 beschriebenen Maximum-Likelihood-Methode, die die Grundlage des EM-Algorithmus darstellt, wurde

[1] Dies zeigt sich u.a. daran, daß –wie [Wu 1982] nachwies – der Aufsatz der oben genannten und offensichtlich mit dem Thema vertrauten Professoren Dempster, Laird und Rubin Fehler enthält, die dazu führen, daß die in [Dempster et al. 1968] getroffenen Aussagen zur Konvergenz des EM-Algorithmus ungültig sind.

zunächst in Abschnitt 222 gezeigt, wie die Maximum-Likelihood-Methode zur Schätzung von Parametern aus unvollständigen Beobachtungsdaten genutzt werden kann. Dann wurde in Abschnitt 231 die Schlüsselgleichung des (G)EM-Algorithmus abgeleitet und daraus schließlich der in Abschnitt 232 definierte (G)EM-Algorithmus entwickelt.

In Kapitel 3 wurden die Eigenschaften des (G)EM-Algorithmus in allgemeiner Form diskutiert, wobei Betrachtungen zur Konvergenz der logarithmierten Likelihoodfunktion bzw. deren Argument sowie die Konvergenzgeschwindigkeit im Vordergrund standen. Da in der diesbezüglichen Literatur einige Autoren verschiedene Auffassungen vertreten, wurden hier im wesentlichen die allgemein als richtig anerkannten Ergebnisse von [Wu 1982] wiedergegeben. Die Darstellung wurde dabei relativ ausführlich gehalten, um dem nicht mit der englischsprachigen Terminologie der mathematischen Statistik bzw. Informatik vertrauten Leser den Zugang hierzu zu erleichtern.
Zunächst wurde in Abschnitt 321 das allgemeine Verhalten der logarithmierten Likelihoodfunktion im Rahmen der EM-Iterationen erklärt. Anschließend wurden in Abschnitt 322 die Bedingungen erarbeitet, unter denen die logarithmierte Likelihoodfunktion im Rahmen der EM-Iterationen gegen einen stationären Punkt bzw. gegen ein lokales Maximum konvergiert. In Abschnitt 323 wurden schließlich die Bedingungen genannt, unter denen die Folge der (vorläufigen) Schätzwerte für die unbekannten Parameter beim (G)EM-Algorithmus konvergiert.

Von besonderer Bedeutung für die Einsetzbarkeit eines Algorithmus ist nach der Frage, ob er zum Ziel führt, die Geschwindigkeit, mit dem er dies tut. In der Literatur (vgl. [Dempster et al. 1968],[Wu 1982]) wird dem EM-Algorithmus i.a. eine nur langsame Konvergenz zugeschrieben. Die Konvergenzgeschwindigkeit hängt dabei vom Anteil der fehlenden Beobachtungen an der Gesamtzahl der Beobachtungen ab. In Abschnitt 324 wurde ein Maß für die Konvergenzgeschwindigkeit des EM-Algorithmus abgeleitet.

In Kapitel 4 wurde das in der Einleitung skizzierte Beispiel für ein Schätz- und Klassifikationsproblem, aufgegriffen. Zur Lösung dieses Problems wurde es als Parameterschätzproblem aus unvollständigen Beobachtungsdaten formuliert und anschließend wurden die im E-Schritt und im M-Schritt durchzuführenden Berechnungen erläutert. Dies führte auf ein Ablaufschema zur Lösung der Schätz- gegebenen und Klassifikationsaufgabe, das in einem Programm zur Umsetzung gelangte (Vgl. Anhang A). Das Programm wurde auf einen Testdatensatz angewendet,wobei sehr gute Ergebnisse erzielt wurden. Dies kann u.a. neben der Qualität des Algorithmus auch dadurch begründet werden, daß das zugrundeliegende Datenmaterial eine nur gerige Streuung aufwies und daß die alternativ zur Auswahl stehenden Beobachtungsmodelle zur Parameterschätzung voneinander gut unterscheidbar waren. Es wäre noch zu testen, wie der Algorithmus auf stark verrauschten Meßdaten und sich ähnelnden Schätzmodellen arbeitet.

Diese Arbeit zeigt u.a., daß der EM-Algorithmus eine interessante Alternative zu herkömmlichen Verfahren der geodätischen Ausgleichungsrechnung darstellt, falls diese zur Lösung einer Aufgabe ungeeignet ist. Aufgrund der im Allgemeinen nur langsamen Konvergenz des EM-Algorithmus sollte vor dessen Einsatz allerdings geprüft werden, ob sich das gewünschte Ergebnis nicht auch mit einer weniger rechenintensiven Methode erreichen läßt.

Literaturverzeichnis

[Dempster et al. 1968] A. Dempster, N. Laird, D.Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Jounal of the Royal Statistical Society, Series B (Methodological), Bd. 39, Nr. 1, 1977, S. 1-38.

[Förstner 1991] W.Förstner. Statistische Kenngrößen, Institut für Photogrammetrie der Universität Bonn, 1991.

[Fuchs 1998] C.Fuchs. *Extraktion polymorpher Bildstrukturen und ihre topologische und geometrische Gruppierung*, Dissertation am Institut für Photogrammetrie der Universität Bonn, 1998.

[Helfrich I 1995] H.-P. Helfrich. *Mathematik I.* Manuskript zur Vorlesung Mathematik I für Geodäten, Mathematisches Seminar der Landwirtschaftlichen Fakultät der Universität Bonn, 1995.

[Helfrich II 1996] H.-P. Helfrich. *Mathematik II.* Manuskript zur Vorlesung Mathematik II für Geodäten, Mathematisches Seminar der Landwirtschaftlichen Fakultät der Universität Bonn, 1996.

[Hinton 1997]] G.E. Hinton, M.Revow. *Using Mixtures of Factor Analyzers for Segmentation and Pose Estimation.* Department of Computer Science, University of Toronto, Canada, 1997.

[Hornegger 1996] J. Hornegger. *Statistische Modellierung, Klassifikation und Lokalisierung von Objekten*, Shaker Verlag, Aachen, 1996.

[Jordan et al. 1994] M.I. Jordan, R.A. Jacobs. *Hierarchical mixtures of experts and the EM-algorithm*, Neural Computation, 6, S. 181 - 214, 1994.

[Kerner et al. 1995] O. Kerner, J. Maurer, J.Steffens, T. Thode, R. Voller. *Vieweg Mathematik Lexikon*,3. Aufl., Vieweg, 1995.

[Koch 1998] K.R. Koch. *Parameterschätzung und Hypothesentests*, Dümmler Verlag, Bonn, 1998.

[Kraus 1994] K. Kraus. *Photogrammetrie, Band1, Grundlagen und Standardverfahren*, Dümmler Verlag, Bonn, 1994.

[Redner 1984] R.A.Redner, H.F.Walker. *Mixture Densities, Maximum Likelihood And The EM-Algorithm*, Society for Industrial and Applied Mathematics Review, Bd. 26, Nr. 2, S.195-239, 1984.

[Rao 1965] C.R. Rao. *Linear Statistical Inference and its Applications.* Wiley-Verlag, New York, 1968.

[Tanner 1993] M.A.Tanner. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer Series in Statistics, Springer, Heidelberg, 1993.

[Wu 1982] C.F.J. Wu. *On The Convergence Properties Of The EM Algorithm.* The Annals of Statistics, Vol. 11, No.1, S. 95-103, 1983.

[Zangwill 1969] W.I. Zangwill. *Nonlinear Programming: A Unified Approach.* Prentice Hall, Englewood Cliffs, New Jersey.

Anhang A

Anhang

A1 C-Quellcode des Programms EM.c (Auszug)

Im Folgenden wird ein Auszug aus dem C-Quelltext des Programms EM.c wiedergegeben. Von den Ein- und Ausgaberoutinen, die hier aus Platzgründen nicht aufgeführt sind, werden nur die Funktionsprototypen angegeben.

```
/*  Programm EM.c, Marc Luxen, 11.2000 */

#include<stdio.h>
#include<math.h>


/*Konstanten */
#define MaxDim 50               /*Maximale Matrixdimension */
#define pi 3.1415926
#define Epsilon 0.000000001     /*Schwellwert fuer Abbruchkriterium */


/*Datentypen*/
typedef struct
  {                             /*Datentyp Matrix */
    int N_r, N_c;
    double Elem[MaxDim][MaxDim];
  }
t_Matrix;


/* Funktionsprototypen */
void     Matrix_ausgeben (FILE * Disk, t_Matrix A);
t_Matrix Inverse         (t_Matrix A);
t_Matrix Transponierte   (t_Matrix A);
t_Matrix Produkt         (t_Matrix A, t_Matrix B);
t_Matrix Differenz       (t_Matrix A, t_Matrix B);
t_Matrix Einheitsmatrix  (int n);
double   Euklidnorm      (t_Matrix X);
t_Matrix cProdukt        (double c, t_Matrix A);
t_Matrix Schaetzung      (t_Matrix X, t_Matrix y, t_Matrix P);
t_Matrix Residuen        (t_Matrix X, t_Matrix y, t_Matrix beta);
t_Matrix Koeffizientenmatrix (int c, t_Matrix x);
void     Messreihe_einlesen  (char Dateiname[], t_Matrix * A, t_Matrix * B);
double   W (double x, double y, double sigma, t_Matrix beta);
```

```
double mu (int c,
           double x,
           double y,
           double sigma1,
           t_Matrix beta1,
           double sigma2,
           t_Matrix beta2);

t_Matrix P (int c,
            t_Matrix u,
            t_Matrix y,
            double sigma1,
            t_Matrix beta1,
            double sigma2,
            t_Matrix beta2);

double Omega (t_Matrix X,
              t_Matrix y,
              t_Matrix P,
              t_Matrix beta);

void Plotdateien_schreiben (t_Matrix P1,    t_Matrix P2,
                            t_Matrix u,     t_Matrix y,
                            t_Matrix beta1, t_Matrix beta2);

void Ergebnisdatei_schreiben (int j,            t_Matrix u,
                              t_Matrix y,       double sigma1,
                              t_Matrix beta1,   t_Matrix M1,
                              t_Matrix P1,      double sigma2,
                              t_Matrix beta2,   t_Matrix P2,
                              t_Matrix M2,      t_Matrix Cov1,
                              t_Matrix Cov2);


/* Routinen:*/

void
Messreihe_einlesen (char Dateiname[], t_Matrix * A, t_Matrix * B)

  /* Einlesen der Messreihe aus der Datei <Dateiname>.
     Die u-Werte der Messreihe werden in die Matrix A und
     die y-Werte in die Matrix B (beides N_r*1-Matrizen)
     geschrieben */

{
  FILE *Disk;
  int r;

  Disk = fopen (Dateiname, "r");

  /* Anzahl der Messwerte einlesen */
  fscanf (Disk, "%d", &((*A).N_r));

  /* Matrizenformate festlegen */
  (*A).N_c = 1;
  (*B).N_r = (*A).N_r;
  (*B).N_c = 1;

  /*Messwerte einlesen: */
  for (r = 0; r <= ((*A).N_r - 1); r++)
```

```
    {
      fscanf (Disk, "%lf", &((*A).Elem[r][0]));
      fscanf (Disk, "%lf", &((*B).Elem[r][0]));
    }
  fclose (Disk);
}


void
Matrix_ausgeben (FILE * Disk, t_Matrix A)

  /* Gibt die Matrix A in den Ausgabestrom Disk aus */

{
  int i, j;
  for (i = 0; i <= (A.N_r - 1); i++)
    {
      for (j = 0; j <= (A.N_c - 1); j++)
        fprintf (Disk, "%10.6lf    ", A.Elem[i][j]);
      fprintf (Disk, "\n");
    };
}


t_Matrix
Inverse (t_Matrix A)

  /* Berechnet die Inverse der Matrix A nach der Gauss - Jordan - Methode
     (vgl. [Koch 1997], Gl. (133.21) */

{
  int i, j, k, r, c;            /* Laufvariablen */
  t_Matrix Inv;                 /* Return - Variable

  /* Format der Ausgabematrix festlegen */
  Inv.N_r = A.N_r;
  Inv.N_c = A.N_c;

  /*Eliminationen */
  for (i = 0; i <= (A.N_r - 1); i++)
    {
      for (j = 0; j <= (A.N_r - 1); j++)
        for (k = 0; k <= (A.N_r - 1); k++)
          {
            if (k != i)
              {
                if (j != i)
                  Inv.Elem[j][k] = A.Elem[j][k] - A.Elem[j][i]
                    * A.Elem[i][k] / A.Elem[i][i];
                else
                  {
                    Inv.Elem[i][k] = A.Elem[i][k] / A.Elem[i][i];
                    Inv.Elem[k][i] = -A.Elem[k][i] / A.Elem[i][i];
                  }
              }
            Inv.Elem[i][i] = 1 / A.Elem[i][i];
          }

      if (i < A.N_r - 1)
        for (r = 0; r <= (A.N_r - 1); r++)
```

```
          for (c = 0; c <= (A.N_c - 1); c++)
            A.Elem[r][c] = Inv.Elem[r][c];
    };
  return (Inv);                 /* Rueckgabe an aufrufende Routine */
}


t_Matrix
Transponierte (t_Matrix A)

  /*Transponiert die Matrix A */

{
  t_Matrix Trans;               /*Return - Matrix */
  int i, j;                     /*Laufvariablen */

  /*Format der Ausgabematrix festlegen */
  Trans.N_r = A.N_c;
  Trans.N_c = A.N_r;

  /*Element [i,j] der Return - Matrix wird das
    Element [j,i] der Matrix A zugewiesen: */
  for (i = 0; i <= (Trans.N_r - 1); i++)
    for (j = 0; j <= (Trans.N_c - 1); j++)
      Trans.Elem[i][j] = A.Elem[j][i];
  return (Trans);               /* Rueckgabe an aufrufende Routine */
}


t_Matrix
Produkt (t_Matrix A, t_Matrix B)

  /*Berechnet das Matrizenprodukt A*B der beiden Matrizen
     A und B und gibt es an die aufrufende Routine zurueck */

{
  int r, c, k;                  /* Laufvariablen */
  t_Matrix Prod;                /* Return - Matrix */

  /* Festlegung des Formates der Return - Matrix: */
  Prod.N_r = A.N_r;
  Prod.N_c = B.N_c;

  /* Berechnung der Elemente der Return - Matrix: */
  for (r = 0; r <= (Prod.N_r - 1); r++)
    for (c = 0; c <= (Prod.N_c - 1); c++)
      {
        Prod.Elem[r][c] = 0;
        for (k = 0; k <= (A.N_c - 1); k++)
          Prod.Elem[r][c] = Prod.Elem[r][c] + A.Elem[r][k] * B.Elem[k][c];
      };
  return (Prod);                /* Rueckgabe an aufrufende Routine */
}


t_Matrix
cProdukt (double c, t_Matrix A)

  /* Multipliziert die Matrix A mit dem Skalar c und gibt das
     Produkt an die aufrufende Routine zurueck */
```

```
{
  int z, s;                     /*Laufvariablen */
  t_Matrix Prod;                /* Return - Matrix */

  /* Format der Return - Matrix festlegen: */
  Prod.N_r = A.N_r;
  Prod.N_c = A.N_c;

  /* Multiplikation der Matrixelemente mit c, Prod[i][j]=c*A[i][j] */
  for (z = 0; z <= (A.N_r - 1); z++)
    for (s = 0; s <= (A.N_r - 1); s++)
      Prod.Elem[z][s] = c * A.Elem[z][s];

  return (Prod);                /* Rueckgabe an aufrufende Routine */
}


double
W (double u, double y, double sigma, t_Matrix beta)

  /* Berechnet die Dichtefunktionen f(y_i|x_i=e_i,beta) und gibt sie
     als double-Wert an die aufrufende Routine zurueck
     y: Beobachtung
     u: zu y gehoerender Abszissenwert*/

{
  if (beta.N_r == 3)            /* Entweder Berechnung der
                                   Dichte des Modells 2... */
    return (2 / (3 * sqrt (2 * pi) * sigma)
            * exp (-pow (y - beta.Elem[0][0]
                         - beta.Elem[1][0] * u
                         - beta.Elem[2][0] * pow (u, 2), 2)
                   / (2 * pow (sigma, 2))));
  else                          /* ... oder Berechnung der
                                   Dichte des Modells 1 */
    return (1 / (3 * sqrt (2 * pi) * sigma)
            * exp (-pow (y - beta.Elem[0][0]
                         - beta.Elem[1][0] * u, 2)
                   / (2 * pow (sigma, 2))));
}


double
mu (int c,          double x,
    double y,       double sigma1,
    t_Matrix beta1, double sigma2,
    t_Matrix beta2)

  /* Berechnung des Erwartungswertes mu_c im E-Schritt */

{
  double W1, W2;
  W1 = W (x, y, sigma1, beta1);
  W2 = W (x, y, sigma2, beta2);
  if (c == 1)
    return (W1 / (W1 + W2));
  else
    return (W2 / (W1 + W2));
}
```

```
t_Matrix
P (int c,       t_Matrix u,
   t_Matrix y,  double sigma1,
   t_Matrix beta1,
   double sigma2,
   t_Matrix beta2)

  /* Baut die Matrix P1 bzw P2 auf  und gibt sie an die aufrufende
     Routine zuruechk
     falls c=1: Aufbauen der Matrix P1
     sonst    : Aufbauen der Matrix P2*/

{
  int i, j;                     /* Laufvariablen */
  t_Matrix PMat;                /*Return - Matrix */

  /* Festlegung der Dimension der Return - Matrix */
  PMat.N_r = y.N_r;
  PMat.N_c = y.N_r;

  for (i = 0; i <= (PMat.N_r - 1); i++)
    for (j = 0; j <= (PMat.N_r - 1); j++)
      if (i == j)               /*Nur die Diagonalelemente
                                  sind von 0 verschieden! */

   /*Falls P1 aufgebaut werden Soll... */
        if (c == 1)

          /*Zuweisung der Erwartungswerte: */
          PMat.Elem[i][j] = mu (1, u.Elem[i][0], y.Elem[i][0], sigma1,
                                beta1, sigma2, beta2);

  /*... sonst wird P2 aufgebaut */
        else
          PMat.Elem[i][j] = mu (2, u.Elem[i][0], y.Elem[i][0], sigma1,
                                beta1, sigma2, beta2);
      else
        PMat.Elem[i][j] = 0;
  return (PMat);                /* Rueckgabe an aufrufende Routine */
}


t_Matrix
Schaetzung (t_Matrix X, t_Matrix y, t_Matrix P)

  /*Berechnet die Schaetzwerte eines Gauss-Markoff-Modells mit der
     Koeffizientenmatrix X, dem Beobachtungsvektor y und der
     Gewichtsmatrix P */

{
  t_Matrix XtP;                 /*Dummy-Matrix */
  t_Matrix Inv;                 /*Return - Matrix */
  XtP = Produkt (Transponierte (X), P);         /*Bekannte */
  Inv = Inverse (Produkt (XtP, X));             /*  Glei- */
  return (Produkt (Produkt (Inv, XtP), y));     /* chungen */
}
```

```
t_Matrix
Differenz (t_Matrix A, t_Matrix B)

  /* Berechnet die Differenz A-B zwischen den Matrizen A und B */

{
  int i, j;                     /* Laufvariablen */
  t_Matrix Diff;                /*Return - Matrix */

  /*Festlegung der Dimension der Return - Matrix */
  Diff.N_r = A.N_r;
  Diff.N_c = A.N_c;

  /*Berechnen der Differenzen A[i,j]-B[i,j] */
  for (i = 0; i <= (A.N_r - 1); i++)
    for (j = 0; j <= (A.N_r - 1); j++)
      Diff.Elem[i][j] = A.Elem[i][j] - B.Elem[i][j];
  return (Diff);                /* Rueckgabe an aufrufende Routine */
}


t_Matrix
Residuen (t_Matrix X, t_Matrix y, t_Matrix beta)

  /* Berechnet die Beobachtungsresiduen eines GMM mit der
     Koeffizientenmatrix X, dem Beobachtungsvektor y und
     der Schaetzung beta */

{
  return (Differenz (Produkt (X, beta), y));    /*bekannte Formel */
}


double
Omega (t_Matrix X, t_Matrix y, t_Matrix P, t_Matrix beta)

  /*Berechnet die Quadratsumme der Residuen eines GMM mit der
     Koeffizientenmatrix X, dem Beobachtungsvektor y, der Gewichtsmatrix P
     und den Schaetzwerten beta */

{
  t_Matrix e;                   /*Return - Matrix */

  e = Residuen (X, y, beta);    /*Bekannte */
  e = Produkt (Produkt (Transponierte (e), P), e);      /*Gleichung */
  return (e.Elem[0][0]);        /* Rueckgabe an aufrufende Routine */
}


t_Matrix
Koeffizientenmatrix (int c, t_Matrix u)

  /*Baut die Koeffizientenmatrix 1 (fuer c=1)bzw. 2 (fuer c=2) auf */
  /*u:Abszissenwert */

{
  int i, j;                     /* Laufvariable */
  t_Matrix Koeff;               /* Return - Matrix */

  /*Dimension der Return - Matrix festlegen */
```

```
  Koeff.N_c = c + 1;
  Koeff.N_r = u.N_r;

  /* Berechnung der Matrixelemente */
  for (i = 0; i <= (Koeff.N_r - 1); i++)
    {
      /* Die ersten beiden Koeffizienten sind in beiden
         Koeffizientenmatrizen dieselben.... */
      Koeff.Elem[i][0] = 1;
      Koeff.Elem[i][1] = u.Elem[i][0];

      /*Bei der Koeffizientenmatrix 2 kommt noch ein dritter
         Koeffizient hinzu */
      if (c == 2)
        Koeff.Elem[i][2] = pow (u.Elem[i][0], 2);
    };
  return (Koeff);               /* Rueckgabe an aufrufende Routine */
}


t_Matrix
Einheitsmatrix (int n)

  /* Erzeugt eine n*n - Einheitsmatrix */

{
  int r, c;                     /*Laufvariablen */
  t_Matrix I;                   /* Return - Matrix */

  /* Dimensionen der Return - Matrix festlegen*/
     I.N_r = n;
     I.N_c = n;

     /*Berechnung der Matrixelemente */
  for (r = 0; r < n; r++)
    for (c = 0; c < n; c++)
      if (r == c)
        I.Elem[r][c] = 1;
      else
        I.Elem[r][c] = 0;
  return (I);                   /* Rueckgabe an aufrufende Routine */
}


double
Euklidnorm (t_Matrix X)

  /*Berechnet die euklidsche Vektornorm eines Vektors X */

{
  double N = 0.0;               /*Return - Variable wird initialisiert */
  int i;                        /* Laufvariable */

  /* Quadratsumme der Vektorelemente bilden: */
  for (i = 0; i <= (X.N_r - 1); i++)
    N = N + pow (X.Elem[i][0], 2);
  return (sqrt (N));            /* Rueckgabe an aufrufende Routine */
}
```

```
int
Klasse (double a, double b)

  /* Gibt aus, welche von den beiden Zahlen a und b die groessere ist
     (Zahl 1 oder Zahl 2)
     Von der aufrufenden Routine aus sind a und b hier
     Wahrscheinlichkeiten des Zutreffens des Modells 1 oder 2 fuer
     eine Beobachtung. Daher nimmt diese Funktion hier einer
     Klassifikationsentscheidung vor. */

{
  if (a > b)
    return (1);
  else
    return (2);

}


void
main (void)
{
  t_Matrix u;   /*Vektor der Abszissenwerte */
  t_Matrix y;   /*Vektor der Ordinatenwerte (Beobachtungsvektor) */
  t_Matrix M1;  /*Koeffizientenmatrix des Modells 1 */
  t_Matrix M2;  /*Koeffizientenmatrix des Modells 2 */
  t_Matrix P1;  /*Gewichtsmatrix des Modells 1 */
  t_Matrix P2;  /*Gewichtsmatrix des Modells 2 */
  t_Matrix beta1;       /*Geschaetzter Parametervektor des Modells 1 */
  t_Matrix Hbeta1;      /*Dummy-Parametervektor (Hilfsvektor) */
  t_Matrix beta2;       /*Geschaetzter Parametervektor des Modells 2 */
  t_Matrix Hbeta2;      /*Dummy-Parametervektor (Hilfsvektor) */
  t_Matrix Cov1;        /*Geschaetzte Kovarianzmatrix der Parameter beta1 */
  t_Matrix Cov2;        /*Geschaetzte Kovarianzmatrix der Parameter beta2 */

  int i, j;                     /*Laufvariablen */
  double sigma1, sigma2;        /*Standardabweichungen der Gewichtseinheit */
  FILE *Disk;                   /*Ausgabestrom fuer Zwischenergebnisse */

  Messreihe_einlesen ("/home/Marc/Arbeit/Software/Linien.dat", &u, &y);
  Disk = fopen ("/home/Marc/Arbeit/Software/P1.dat", "w");

  /* Naeherungswerte */

  /* Als Gewichtsmatrizen werden zunaechst Einheitsmatrizen gewaehlt */
  P1 = Einheitsmatrix (u.N_r);
  P2 = P1;

  /* Aufbauen der Koeffizientenmatrizen : */
  M1 = Koeffizientenmatrix (1, u);
  M2 = Koeffizientenmatrix (2, u);

  /* Initiale Parameterschaetzungen in den Modellen 1 und 2 */
  beta1 = Schaetzung (M1, y, P1);
  beta2 = Schaetzung (M2, y, P2);
  /* Bestimmung der Varianzen der Gewichtseinheit */
  sigma1 = sqrt (Omega (M1, y, P1, beta1) / (y.N_r - 2));
  sigma2 = sqrt (Omega (M2, y, P2, beta2) / (y.N_r - 3));

  j = 0;                        /* Iterationszaehler j wird auf 0 gesetzt */
```

```
  do
    {
      j++;
      /*Um einen Vergleich der neuen Schaetzwerte mit den alten
         Schaetzwerten zu ermoeglichen, muessen die alten
         Schaetzwerte zwischengespeichert werden */
      Hbeta1 = beta1;
      Hbeta2 = beta2;

      /* E-Schritt: Aufbau der Gewichtsmatrizen P1 und P2 */
      P1 = P (1, u, y, sigma1, beta1, sigma2, beta2);
      P2 = P (2, u, y, sigma1, beta1, sigma2, beta2);

      /* M-Schritt: Neuschaetzung der unbekannten Parameter mit
         der neuen Gewichtsmatrix */
      beta1 = Schaetzung (M1, y, P1);
      beta2 = Schaetzung (M2, y, P2);
      /* Neuschaetzungen der Varianzen der Gewichtseinheit */
      sigma1 = sqrt (Omega (M1, y, P1, beta1) / (y.N_r - 2));
      sigma2 = sqrt (Omega (M2, y, P2, beta2) / (y.N_r - 3));

      /*Ausgabe von Zwischenergebnissen fuer Visualisierung */
      fprintf (Disk, "%d  ", j);
      for (i = 0; i <= (y.N_r - 1); i++)
        fprintf (Disk, "%8.6lf  ", P1.Elem[i][i]);
      fprintf (Disk, "\n");
    }

  /* Abbruchbedingung: */
  while (sqrt (pow (Euklidnorm (Differenz (beta1, Hbeta1)), 2)
               + pow (Euklidnorm (Differenz (beta2, Hbeta2)), 2))
         >= Epsilon);

  fclose (Disk);  /* Schliessen des Ausgabestroms fuer Zwischenergebnisse */

  /*Berechnung der Kovarianzmatrizen der finalen Schaetzungen beta1
     und beta2: */
  Cov1 = cProdukt (pow (sigma1, 2),
                   Inverse (Produkt (Produkt (Transponierte (M1), P1), M1)));

  Cov2 = cProdukt (pow (sigma2, 2),
                   Inverse (Produkt (Produkt (Transponierte (M2), P2), M2)));

  /*Ausgabe der Ergebnisse */
  Ergebnisdatei_schreiben (j, u, y, sigma1, beta1, M1, P1,
                           sigma2, beta2, P2, M2, Cov1, Cov2);

  Plotdateien_schreiben (P1, P2, u, y, beta1, beta2);
}
```

A2 Datensatz Linien.dat

Den Berechnungen in Abschnitt 423 liegen folgende Daten zugrunde:

```
40
1       15.18094175
2       13.87552839
3       6.03249538
4       7.420570315
5       9.613662336
6       9.235540291
7       8.584545608
8       6.776858258
9       9.945421688
10      6.064615363
11      5.280097939
12      11.16776634
13      5.261352972
14      12.91940046
15      5.29507723
16      5.420944694
17      5.233561958
18      14.2346887
19      15.34965322
20      6.49533126
21      14.58344748
22      6.999892587
23      8.564241033
24      9.705304727
25      16.70748849
26      18.0886264
27      17.76763562
28      14.17874028
29      14.07531348
30      16.8324473
31      17.40444817
32      20.03882833
33      21.76675137
34      23.21420071
35      25.69938612
36      23.9982072
37      29.06170133
38      31.45851914
39      34.19962462
40      36.9238597
```