Multimodale Likelihood-Funktionen in Mischverteilungsmodellen

Dissertation

zur Erlangung des akademischen Grades
eines *Doktors der Naturwissenschaften*
der Technischen Universität Dortmund

Der Fakultät Statistik
der Technischen Universität Dortmund
vorgelegt von

Malte Jastrow

Dortmund, Juli 2021

Erstgutachter: Prof. Dr. Claus Weihs
Zweitgutachter: Dr. Uwe Ligges

Datum der mündlichen Prüfung: 26. August 2021

Inhaltsverzeichnis

1 Einleitung

Mischverteilungsmodelle (*Mixture Models*) dienen allgemein zur Anpassung zusammengesetzter Verteilungen an Daten, in denen einzelne Gruppen von Beobachtungen unterschiedlichen Verteilungen folgen. Durch die Modellierung der Gruppenzugehörigkeiten als latente Variable sind diese Modelle darüber hinaus ein populäres Verfahren zur Clusteranalyse (unüberwachtes Lernen). Dabei werden die Gruppen, denen Beobachtungen zugeordnet werden sollen, durch unterschiedlich parametrisierte Verteilungskomponenten repräsentiert. Die Verteilungsparameter der einzelnen Komponenten, sowie deren Mischungsverhältnis können mittels Maximum-Likelihood-Prinzip geschätzt werden. Wie durch Day (1969) beschrieben, kann die Likelihood-Funktion bereits für die Mischung zweier Normalverteilungskomponenten zahlreiche Optima aufweisen, wenn sich die zugrunde liegenden Varianzen stark unterscheiden.

Ziel dieser Arbeit ist es, genauer zu untersuchen, wie sich die Parameter der Verteilungskomponenten auf das Entstehen von Multimodalität auswirken und wie gut die multimodalen Likelihood-Funktionen mit dem üblicherweise verwendeten EM-Algorithmus optimiert werden können.

Dazu werden in Kapitel 2, nach einer theoretischen Einführung zu Mischverteilungsmodellen (Kap. 2.1), zunächst Likelihood-Funktionen verschiedener Mischungen grafisch dargestellt, um die vorliegende Multimodalität zu demonstrieren (Kap. 2.2). Da aus der Theorie zum EM-Algorithmus (Kap. 2.3) hervorgeht, dass es keine Garantie zum Auffinden des globalen Optimums bei Multimodalität gibt, wird in Kapitel 2.4 eine systematische Untersuchung des Einflusses der Mischungsparameter auf die Komplexität der Likelihood-Funktionen mittels Heuristiken durchgeführt.

Daran anschließend werden in Kapitel 3 Simulationsstudien zum Vergleich verschiedener Optimierungsalgorithmen auf entsprechenden Likelihood-Funktionen von Normalverteilungsmischungen durchgeführt. In Kapitel 3.1 wird der Aufbau der Studien für die Simulationen mit zwei und fünf zu optimierenden Parametern der Modelle vorgestellt. Neben EM werden mit CMA-ES und MBO zwei Verfahren mit vielversprechenden Ansätzen zur globalen Optimierung verwendet (Kap. 3.2). Die Beschreibung und Auswertung der Studien findet in den Kapiteln 3.3 (2D) und 3.4 (5D) statt. Zum Abschluss wird anhand einzelner Iterationsschritte und der Darstellung der internen Optimierungsfunktion analysiert, wie der EM-Algorithmus trotz Multimodalität zu guten Optimierungsergebnissen kommt (Kap. 3.5).

Der letzte Teil dieser Arbeit beschäftigt sich mit dem Auffinden möglichst vieler lokaler Optima (Kap. 4). In einem Beweis von Améndola Cerón (2017) (Kap. 4.1)

wird für eine aus beliebig vielen Clustern zusammengesetzte Datensituation eine untere Schranke vorhandener Optima angegeben. Dabei werden Startpunkte für den EM-Algorithmus so gewählt, dass die initiale Aufteilung jeweils ein einzelnes Cluster von allen anderen Beobachtungen trennt, womit schließlich für jedes Cluster ein spezifisches Optimum erreicht werden kann. Da die zugrunde liegenden Daten nicht als normalverteilt angesehen werden können, wird das Vorgehen auf Datensituationen adaptiert, die Annäherungen an normalverteilte Daten darstellen. In diesen weiterhin auf Clustern basierenden Szenarien wird der EM-Algorithmus analog verwendet, um mit clusterbasierten Startpunkten eine große Zahl unterschiedlicher Optima aufzufinden. In einem ersten Schritt werden die Annäherungen an die Normalverteilung aus gleichverteilten Clustern zusammengesetzt (Kap. 4.2), in einem weiteren aus normalverteilten Clustern (Kap. 4.3). Danach erfolgt das Hinzufügen von Beobachtungen einer zusätzlichen Mischungskomponente (Kap. 4.4), bevor das Vorgehen abschließend auf normalverteilte Daten angewendet wird (Kap. 4.5). Dabei werden die Startpunkte für EM auf Basis von zufällig bestehenden Clustern in den Daten gebildet.

Ein zentrales Ergebnis ist, dass der EM-Algorithmus zur Optimierung aller fünf Parameter einer Mischung zweier normalverteilter Komponenten mit multimodaler Likelihood deutlich bessere Ergebnisse liefert als die Vergleichsalgorithmen aus dem Bereich der *Black-Box*-Optimierung. Darüber hinaus bietet sich mit den vorgeschlagenen Clusterstartpunkten für EM eine für den Anwendungsfall relevante Methode, um möglichst viele lokale Optima einer multimodalen Likelihood-Funktion zu identifizieren. Letzteres gelingt deutlich besser als mit der häufig praktizierten Verwendung zufälliger Startpunkte für EM und kann einen entscheidenden Beitrag zur Bewertung eines globalen Optimierungsergebnisses liefern.

2 Likelihood-Funktionen von Mischverteilungsmodellen

Auf Basis des Mischverteilungsmodells, welches in Kapitel 2.1 eingeführt wird, werden in Abschnitt 2.2 multimodale Likelihood-Funktionen generiert und vorgestellt. Um die Eigenschaften dieser Funktionen weiter zu analysieren, folgt anschließend eine Beschreibung des üblicherweise zur Optimierung der Parameter des Mischverteilungsmodells verwendeten EM-Algorithmus (Kap. 2.3) inklusive relevanter Konvergenzeigenschaften. Da keine globale Konvergenz bei vorhandener Multimodalität garantiert werden kann, wird in Kapitel 2.4 mithilfe von einfachen Heuristiken die Komplexität der Likelihood-Funktionen in Abhängigkeit der zugrunde liegenden Parameter charakterisiert.

Sämtliche Berechnungen in dieser Arbeit wurden mit der freien Software R (R Core Team, 2021) durchgeführt, dies gilt insbesondere auch für die erzeugten Grafiken. Für einen Teil der enthaltenen Grafiken wurde das R-Paket *ggplot2* (Wickham, 2016) verwendet. Weitere zum Einsatz gekommene Zusatzpakete werden an den entsprechenden Stellen der Analyse erwähnt. Mithilfe des R-Pakets *knitr* (Xie, 2020) wurden die Berechnungen reproduzierbar in dieses LaTeX-Dokument eingebettet.

2.1 Mischverteilungsmodell (Mixture Model)

Die klassische Arbeit von Pearson (1894) zum Anpassen zweier Normalverteilungskomponenten an Daten zu Körpermaßen von Krabben kann als eine der ersten Anwendungen von Mischverteilungsmodellen angesehen werden. Vermutlich aufgrund des hohen Komplexitätsgrades der verwendeten Momentenmethode zur Parameterschätzung wurden Mischverteilungsmodelle zur damaligen Zeit noch nicht besonders populär. Der erste Einsatz von Maximum-Likelihood-Schätzungen in Mischverteilungsmodellen ist belegt durch Rao (1948), allerdings dauerte es bis zur Veröffentlichung des EM-Algorithmus durch Dempster et al. (1977), dass Mischverteilungsmodelle ein weit verbreitetes Verfahren in der Statistik wurden (McLachlan & Peel, 2000, S. 35-37). Die Grundlagen dieser Modelle werden im Folgenden erläutert.

Sei $\boldsymbol{X}_j$ ein p-dimensionaler Zufallsvektor mit Wahrscheinlichkeitsdichte $f(\boldsymbol{x}_j)$ im $\mathbb{R}^p$, wobei $\boldsymbol{x}_j$ eine Realisierung des Zufallsvektors bezeichnet. In einem finiten

Mischverteilungsmodell (*Finite Mixture Model*) bestehend aus g Komponenten kann die Wahrscheinlichkeitsdichte geschrieben werden als

$$f(\boldsymbol{x}_j|\,\boldsymbol{\Psi}) = \sum_{i=1}^{g} \lambda_i f_i(\boldsymbol{x}_j|\boldsymbol{\theta}_i),$$

mit den Dichten der einzelnen Komponenten $f_i(\boldsymbol{x}_j|\boldsymbol{\theta}_i)$ für $i = 1, \ldots, g$ und den Mischungsanteilen λ_i, welche die Bedingungen $0 \leq \lambda_i \leq 1$ und $\sum_{i=1}^{g} \lambda_i = 1$ erfüllen (McLachlan & Peel, 2000, S. 6).

Der Vektor $\boldsymbol{\Psi} = (\lambda_1, \ldots, \lambda_{g-1}, \boldsymbol{\xi}^T)^T$ enthält alle unbekannten Parameter des Modells, wobei $\boldsymbol{\xi}$ alle a priori verschiedenen Parameter aus $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g$ enthält (McLachlan & Peel, 2000, S. 22). Um $\boldsymbol{\Psi}$ in der Dichte $f(\boldsymbol{x}_j|\,\boldsymbol{\Psi})$ zu schätzen wird die Likelihood-Funktion $L(\boldsymbol{\Psi})$ maximiert. Sie ist gegeben durch

$$L(\boldsymbol{\Psi}) = \prod_{j=1}^{n} f(\boldsymbol{x}_j|\,\boldsymbol{\Psi})$$

für Realisationen $\boldsymbol{x}_j$ des unabhängig verteilten Zufallsvektors $\boldsymbol{X}_j, j = 1, \ldots, n$ (McLachlan & Peel, 2000, S. 40-41). $\boldsymbol{\Psi}$ enthält lediglich $g - 1$ Mischungsanteils-Schätzer, da $\lambda_g = 1 - \sum_{i=1}^{g-1} \lambda_i$ gilt.

Als Beispiel ist die Mischung zweier univariater Normalverteilungen $U \sim \mathcal{N}(\mu_1, \sigma_1^2)$ und $V \sim \mathcal{N}(\mu_2, \sigma_2^2)$ gegeben durch

$$L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda|\boldsymbol{x}) = \prod_{j=1}^{n} \left[\lambda\, \phi(x_j|\mu_1, \sigma_1^2) + (1-\lambda)\, \phi(x_j|\mu_2, \sigma_2^2)\right],$$

wobei λ den Mischungsanteil der ersten Komponente und ϕ die Dichtefunktion der Normalverteilung bezeichnet. Solche Mischungen zweier Normalverteilungen werden im folgenden Unterkapitel betrachtet, um eine erste Analyse der Multimodalität bei Mischverteilungsmodellen durchzuführen. Das standardmäßige Verfahren zur Optimierung der Likelihood-Parameter wird in Kapitel 2.3 vorgestellt.

2.2 Multimodale Likelihood-Funktionen

Multimodalität der Likelihood-Funktion kann das Auffinden der optimalen Parameter eines Mischverteilungsmodells erschweren. Wie von Day (1969) beschrieben, kann eine Mischung zweier univariater Normalverteilungen mit ungleichen Varianzen bereits zu einer Likelihood mit zahlreichen lokalen Optima führen. Diese Optima werden durch nah aneinander gruppierte Beobachtungen hervorgerufen. Diese sogenannten *spurious maximizers* (McLachlan & Peel, 2000, S. 17 ff.) existieren zusätzlich zum Optimum mit dem besten beschränkten Funktionswert und

können dazu führen, dass EM oder Quasi-Newton Algorithmen das wahre Optimum nicht finden (Hathaway, 1985). In einer neueren Arbeit zeigt Améndola Cerón (2017), dass die Anzahl nicht-trivialer kritischer Punkte für bestimmte (aus Clustern zusammengesetzte) Stichproben in normalverteilten Mischmodellen nach oben unbeschränkt ist und mindestens ein Optimum pro Cluster existiert. Die darin verwendete Vorgehensweise wird in Kapitel 4.1 detailliert vorgestellt und anschließend auf weitere Stichproben adaptiert.

An dieser Stelle ist das erste Ziel herauszufinden, wie die Parameter der Mischungen das Auftreten von lokalen Optima beeinflussen und ob es neben der Normalverteilung weitere Verteilungen gibt, bei denen das Problem der Multimodalität in vergleichbarer Form auftritt.

Um einen ersten Eindruck der Likelihood-Funktionen zu gewinnen, ist eine grafische Darstellung wünschenswert. Mischungen zweier univariater Normalverteilungen verfügen allerdings über insgesamt fünf freie Parameter, was sie im Allgemeinen nicht grafisch darstellbar macht. Um die Anzahl der Parameter zu reduzieren, werden zunächst nur die Lageparameter der beiden Verteilungen betrachtet. Im Rahmen dieser Arbeit wird die Optimierung immer als Minimierungsproblem der logarithmierten Likelihood betrachtet. Gegebene Varianzen und Mischungsanteile führen zum folgenden vereinfachten Problem:

$$\min_{\boldsymbol{\mu}} \; -\log L(\mu_1, \mu_2 | \boldsymbol{x}, \sigma_1^2, \sigma_2^2, \lambda).$$

Zur Generierung solcher Funktionen müssen die als Vektor $\boldsymbol{x}$ gegebenen Beobachtungen zufällig gezogen werden. Die Anzahl gegebener Beobachtungen n ist ein Parameter des Generierungsprozesses, der ebenso zu Beginn festgelegt werden muss wie die gegebenen Parameter der Likelihood-Funktion λ und $\boldsymbol{\sigma^2} = (\sigma_1^2, \sigma_2^2)$. Die wahren Mittelwerte der beiden Normalverteilungen $\boldsymbol{\mu} = (\mu_1, \mu_2)$ werden zufällig gleichverteilt aus dem Intervall $[-5, 5]^2$ gezogen. Danach werden die Beobachtungen in zwei aufeinander folgenden Schritten ermittelt. Zuerst wird mithilfe einer diskreten Zufallsstichprobe aus $Z \sim \mathcal{B}(\lambda)$ bestimmt, aus welcher der beiden Komponenten jede einzelne Beobachtung stammt. Im zweiten Schritt wird jede Beobachtung aus einer der beiden Zufallsvariablen $U \sim \mathcal{N}(\mu_1, \sigma_1^2)$ oder $V \sim \mathcal{N}(\mu_2, \sigma_2^2)$ gezogen, je nach Ausgang des ersten Schrittes. Für alle auf diese Weise erzeugten Funktionen gilt o. B. d. A. $\sigma_1 < \sigma_2$. Der ersten Verteilungskomponente liegt also immer die kleinere Standardabweichung zugrunde.

Als erstes Beispiel ist in Abbildung 1 die Likelihood-Funktion für die Mittelwertparameter einer Mischung zweier normalverteilter Komponenten mit $n = 10$ Beobachtungen und festen Standardabweichungen $\sigma_1 = 0.01$ und $\sigma_2 = 1$ sowie festem

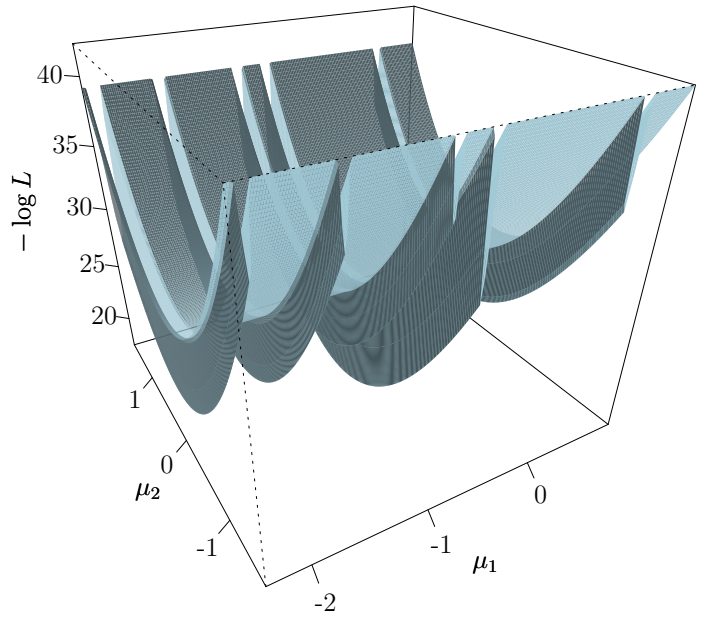

Abbildung 1: Likelihood für $n = 10$ mit $\sigma_1 = 0.01$ und $\sigma_2 = 1$, sowie $\lambda = 0.5$

Mischungsverhältnis $\lambda = 0.5$ dargestellt. Es ist deutlich zu erkennen, dass in Richtung von μ_1 insgesamt fünf lokale Optima existieren, von denen das bei μ_1 knapp über -1 den erkennbar geringsten Wert aufweist.

In Abbildung 2 sind einige Beispiele für heteroskedastische Normalverteilungsmischungen enthalten. Die Zeile der Grafikmatrix gibt die Kombination der Varianzparameter $\boldsymbol{\sigma}$ an, während die Spalten die verschiedenen Stichprobengrößen repräsentieren. In allen Grafiken sind μ_1 und μ_2 jeweils im Bereich $[-10, 10]$ dargestellt, um einen Vergleich zwischen den Parameterkombinationen zu ermöglichen. Aus Gründen der Übersichtlichkeit sind an den Achsen für μ_1 und μ_2 keine Werte abgetragen und für die Funktionswerte lediglich Minimum und Maximum angegeben. Um einen Eindruck über die Ausdehnung der Funktionswerte im betrachteten Ausschnitt zu erhalten, sind einzig die beiden Extremwerte zur Orientierung an den Enden der entsprechenden Achse angegeben. Bei der links oben enthaltenen Funktion handelt es sich um die Funktion aus Abbildung 1, dargestellt auf dem größeren Bereich der Mittelwerte. Da die vorherige Darstellung auf die wahren Mittelwerte zentriert und auf die wahren Standardabweichungen skaliert ist, sind hier große Bereiche dargestellt, die keine weiteren Optima enthalten. Es bestehen lediglich wenige zusätzliche Optima in der Nähe des zuvor betrachteten Ausschnittes. Für $\boldsymbol{\sigma} = (0.01, 1)$ sind auch die weiteren Funktionen multimodal: Bei Erhöhung auf $n = 100$ existieren ein deutlich erkennbares globales Optimum und separiert davon für kleinere Werte von μ_1 zahlreiche lokale Optima. Für $n = 1000$ existiert ebenfalls ein gut erkennbares globales Optimum. Aufgrund sehr ähnlicher, zufällig

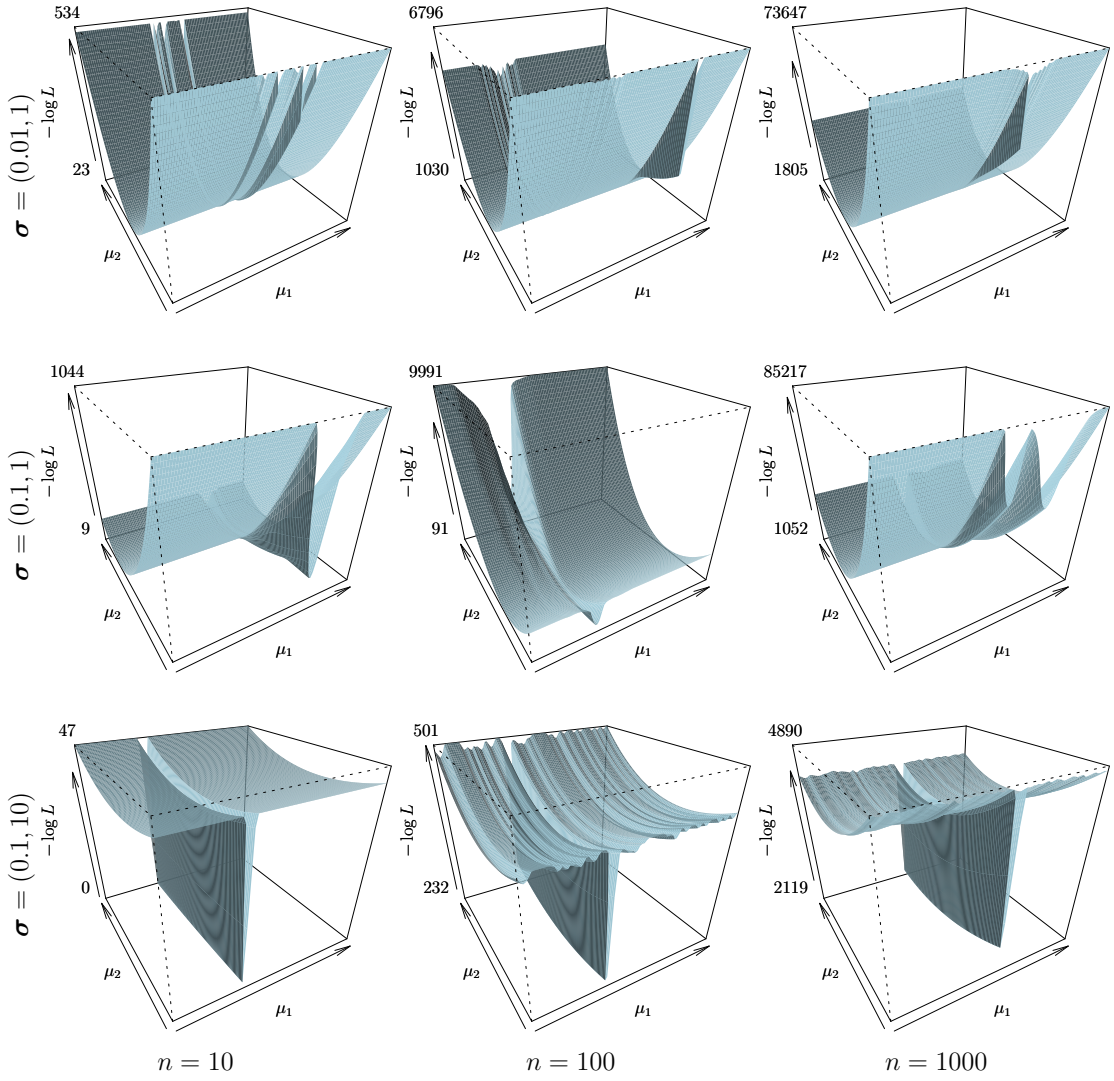

Abbildung 2: Mischungen univariater Normalverteilungen

bestimmter Mittelwertparameter der Verteilungen sind hier in unmittelbarer Nähe zahlreiche lokale Optima vorhanden, deren Funktionswerte sich im Verhältnis zum globalen Optimum allerdings kaum vom allgemeinen Niveau der Funktion unterscheiden. Für die in der Mitte dargestellte Varianzkombination $\boldsymbol{\sigma} = (0.1, 1)$ sind für alle drei Stichprobengrößen deutlich weniger lokale Optima zu erkennen, trotzdem ist auch hier keine der drei Funktionen auf dem betrachteten Bereich unimodal.

Für $\boldsymbol{\sigma} = (0.1, 10)$ hingegen wird die Multimodalität am deutlichsten sichtbar, da die lokalen Optima für 100 und 1000 Beobachtungen über den gesamten dargestellten Bereich von μ_1 verteilt sind. Dies hängt damit zusammen, dass der gewählte Ausschnitt $[10, 10]$ besser zur größeren Standardabweichung passt als in den anderen Fällen. Es ist daher denkbar, dass ähnlich viele Optima vorhanden sind wie im Fall $\boldsymbol{\sigma} = (0.01, 1)$ und lediglich eine andere Skalierung vorliegt. Bei der unteren

Varianzkombination ist sogar davon auszugehen, dass außerhalb des betrachteten Bereiches noch weitere Optima vorhanden sind. Entweder aus diesem Grund oder weil zufällig keine lokalen Optima entstanden sind, ist für $n = 10$ nur das globale Optimum erkennbar. Außerdem fällt insgesamt auf, dass alle Optima offenbar immer von der Verteilungskomponente mit geringerer Varianz verursacht werden und die andere Verteilung die Multimodalität demnach nicht beeinflusst. Diese vermuteten Zusammenhänge zwischen Multimodalität und Parametern bei Mischungen von Normalverteilungen werden in Kapitel 2.4 mittels Simulationen systematisch überprüft.

An dieser Stelle soll zunächst untersucht werden, ob auch für Mischungen aus anderen Verteilungen multimodale Likelihood-Funktionen entstehen können. Um Multimodalität im $\mu_1 \times \mu_2$-Raum durch Heteroskedaszität hervorzurufen, ist es notwendig Verteilungen mit separaten Lage- und Streuungsparametern zu verwenden. Daher werden im Folgenden sogenannte *location-scale*-Verteilungsfamilien betrachtet.

Nach Casella & Berger (2002, S. 119) ist die Familie der Wahrscheinlichkeitsdichtefunktionen

$$\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

mit Lageparameter $-\infty < \mu < \infty$ und Streuungsparameter $\sigma > 0$ eine *location-scale*-Familie mit Standarddichtefunktion $f(x)$.

Bei der bereits verwendeten univariaten Normalverteilung mit Wahrscheinlichkeitsdichtefunktion (vgl. Johnson et al., 1995a, S. 80)

$$f_{\text{normal}}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

handelt es sich um eine *location-scale*-Verteilungsfamilie, wobei die Standarddichtefunktion mit $\mu = 0$ und $\sigma = 1$ gegeben ist.

Weitere Beispiele solcher Verteilungsfamilien mit separaten Parametern μ für die Lage und σ für die Streuung sind die logistische Verteilung mit Dichtefunktion (vgl. Johnson et al., 1995b, S. 116)

$$f_{\text{logist}}(x|\mu,\sigma) = \frac{1}{\sigma} \cdot \frac{e^{-\frac{x-\mu}{\sigma}}}{\left(1 + e^{-\frac{x-\mu}{\sigma}}\right)^2},$$

die Cauchy-Verteilung mit Dichtefunktion (vgl. Johnson et al., 1995a, S. 299)

$$f_{\text{Cauchy}}(x|\mu,\sigma) = \frac{1}{\pi\sigma} \cdot \frac{1}{1 + \left(\frac{x-\mu}{\sigma}\right)^2}$$

und die Laplace-Verteilung mit Dichtefunktion (vgl. Johnson et al., 1995b, S. 164)

$$f_{\text{Laplace}}(x|\mu,\sigma) = \frac{1}{2\sigma} \cdot e^{-\frac{|x-\mu|}{\sigma}},$$

wobei sich die Standarddichten auch hier jeweils mit den Parametern $\mu = 0$ und $\sigma = 1$ ergeben.

Um zu überprüfen, ob die Multimodalität der Likelihood-Funktion auch für diese weiteren *location-scale*-Familien auftritt, werden im Folgenden univariate Mischverteilungsmodelle mit zwei Komponenten aus logistischen, Cauchy- und Laplace-Verteilungen vorgestellt. Dabei werden analog zum Normalverteilungsbeispiel die identischen Kombinationen von Streuungsparametern und Beobachtungsanzahl verwendet. In Abbildung 3 sind die Likelihood-Funktionen für Mischungen zweier logistischer Verteilungen dargestellt. Bezüglich Anzahl und Lage der lokalen Optima zeigt sich ein sehr ähnliches Bild zur Normalverteilung: Für $\boldsymbol{\sigma} = (0.01, 1)$ liegen die lokalen Optima in einem kleinen Bereich bezüglich μ_1, während sie für $\boldsymbol{\sigma} = (0.1, 10)$ über den gesamten Bereich verteilt sind. Für $\boldsymbol{\sigma} = (0.1, 1)$ sind auch

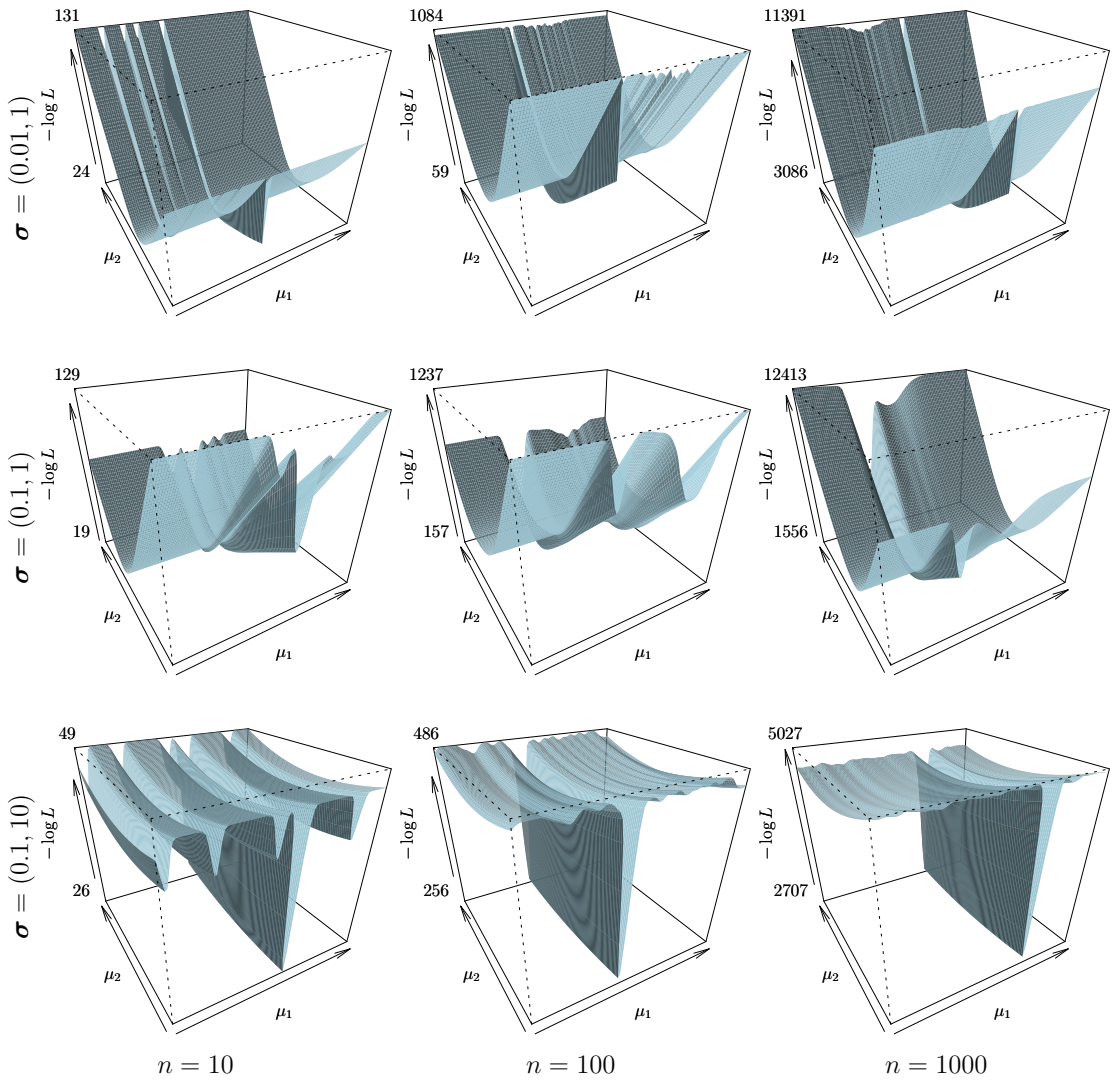

Abbildung 3: Mischungen univariater logistischer Verteilungen

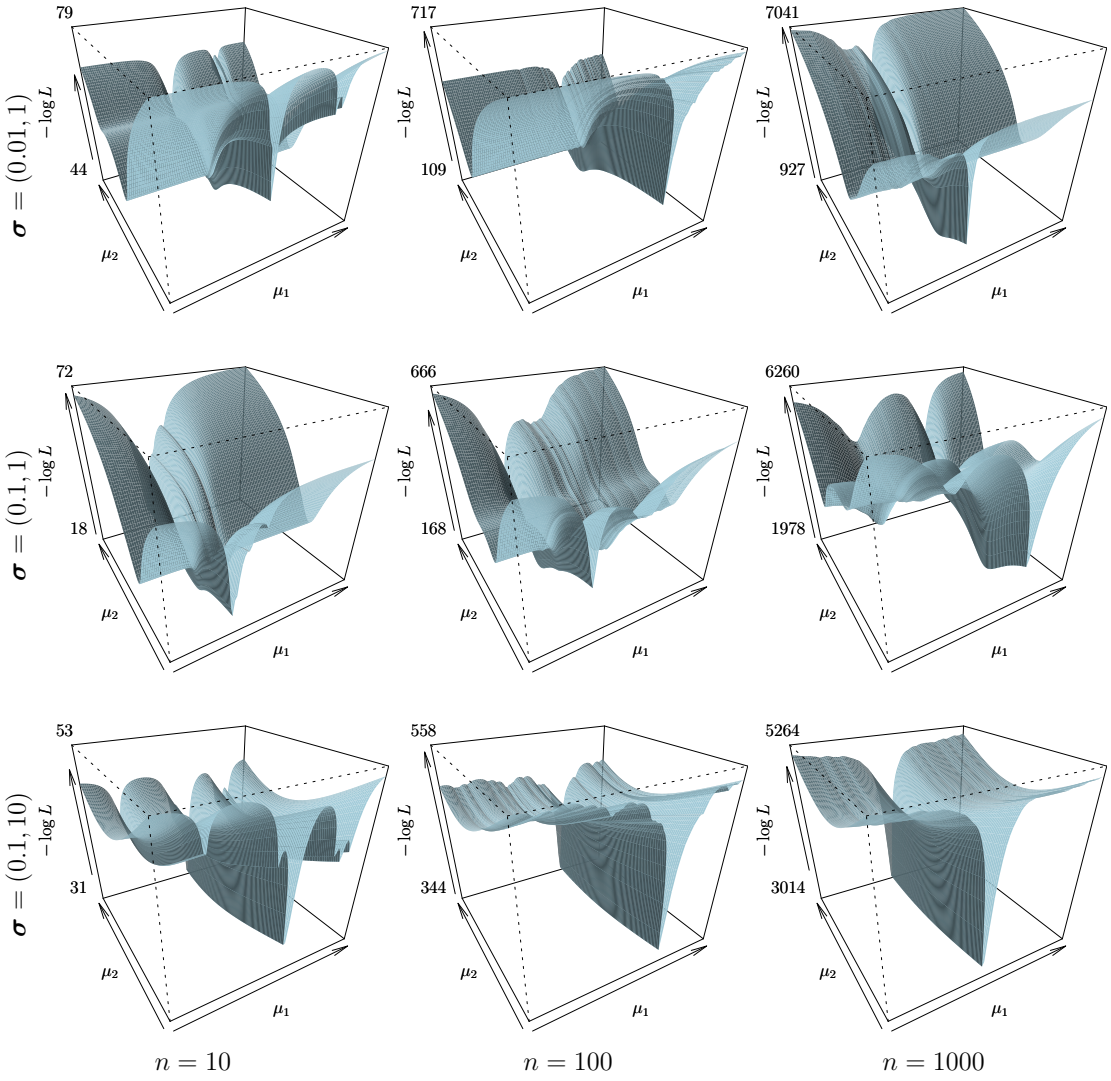

Abbildung 4: Mischungen univariater Cauchy-Verteilungen

hier deutlich weniger Optima erkennbar. Ein Unterschied ist, dass für $n = 10$ in allen Fällen deutlichere bzw. mehr lokale Optima zu sehen sind. Auf Grund des geringen Stichprobenumfangs handelt es sich vermutlich um zufallsbedingte Unterschiede durch die konkreten Realisationen der Zufallsstichproben. Insgesamt zeigen die Funktionsoberflächen keine Auffälligkeiten, die sie deutlich von denen der Normalverteilungsmischungen unterscheiden lassen.

Im Gegensatz dazu zeigt sich für die Likelihood-Funktionen von Mischungen der Cauchy-Verteilungen in Abbildung 4 eine andere Charakteristik: Während der Funktionsverlauf in μ_1-Richtung im Bereich der Optimalwerte tendenziell etwas spitzer als in den vorherigen Beispielen ist, gehen die Funktionen von einem Optimum aus mit sehr großen Radien auf das allgemeine Funktionsniveau über, was dazu führt, dass der Einzugsbereich, in dem die Steigung der Funktion in Richtung

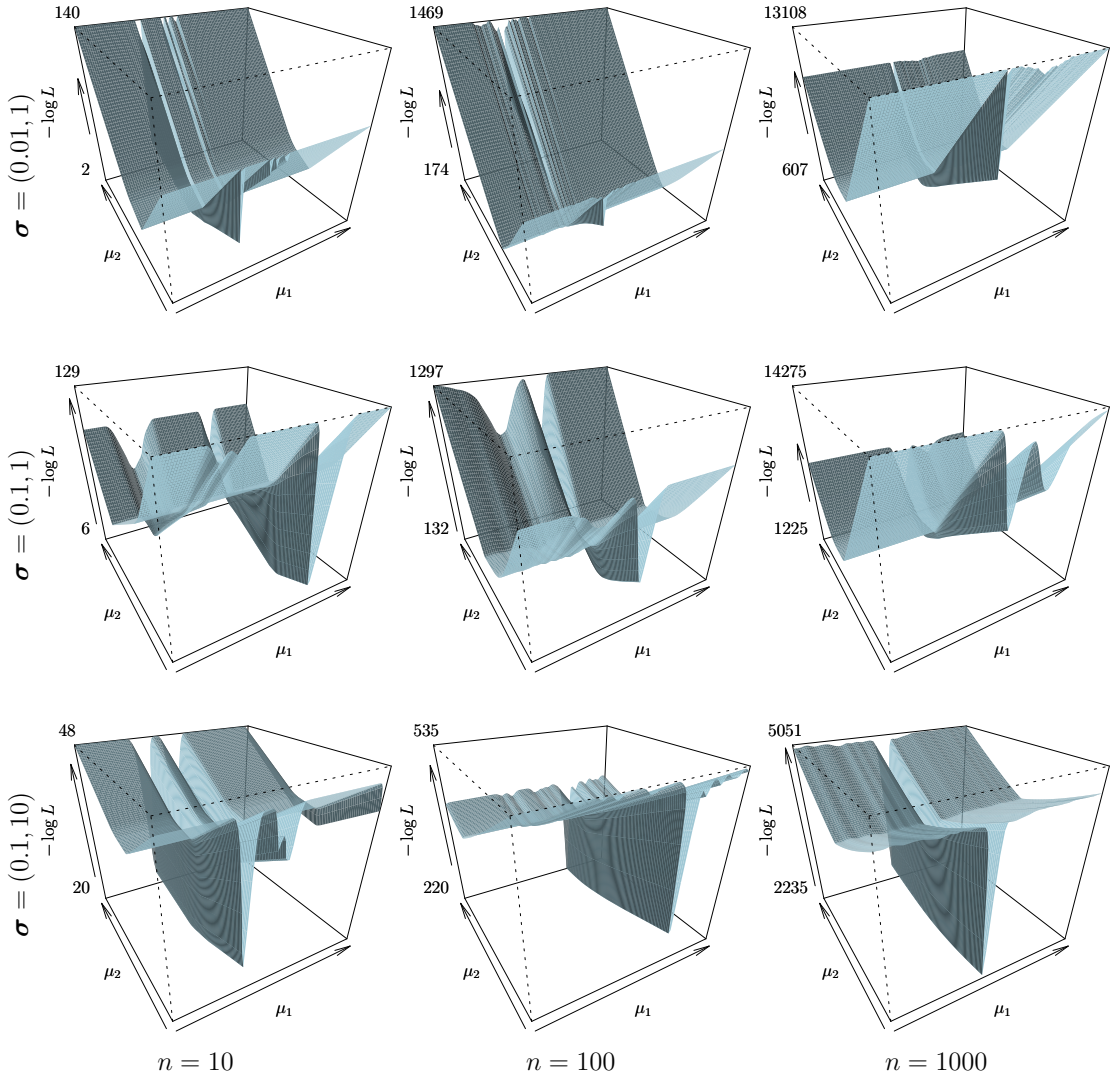

Abbildung 5: Mischungen univariater Laplace-Verteilungen

des Optimums weist, deutlich größer ist. Dennoch bestätigen sich die bisherigen Beobachtungen bezüglich Anzahl und Lage der lokalen Optima auch in diesem Fall.

Die Beispiele zur Laplace-Verteilung in Abbildung 5 stellen das andere Extrem bezüglich des Funktionsverlaufes im Bereich der Optima dar, da die Werte hier eher linear vom Optimum zum allgemeinen Funktionsniveau ansteigen und so sehr kleine Einzugsbereiche in μ_1-Richtung entstehen. Darüber hinaus sind auch hier lokale Optima in ähnlichem Maße wie bei den bisher betrachteten Verteilungen vorhanden.

Da deutlich geworden ist, dass die lokalen Optima üblicherweise nur in Richtung des Mittelwertes der Verteilungskomponente mit kleinerer Varianz bestehen, werden zusätzlich noch Querschnittsdarstellungen zweier Beispiele für alle vier Vertei-

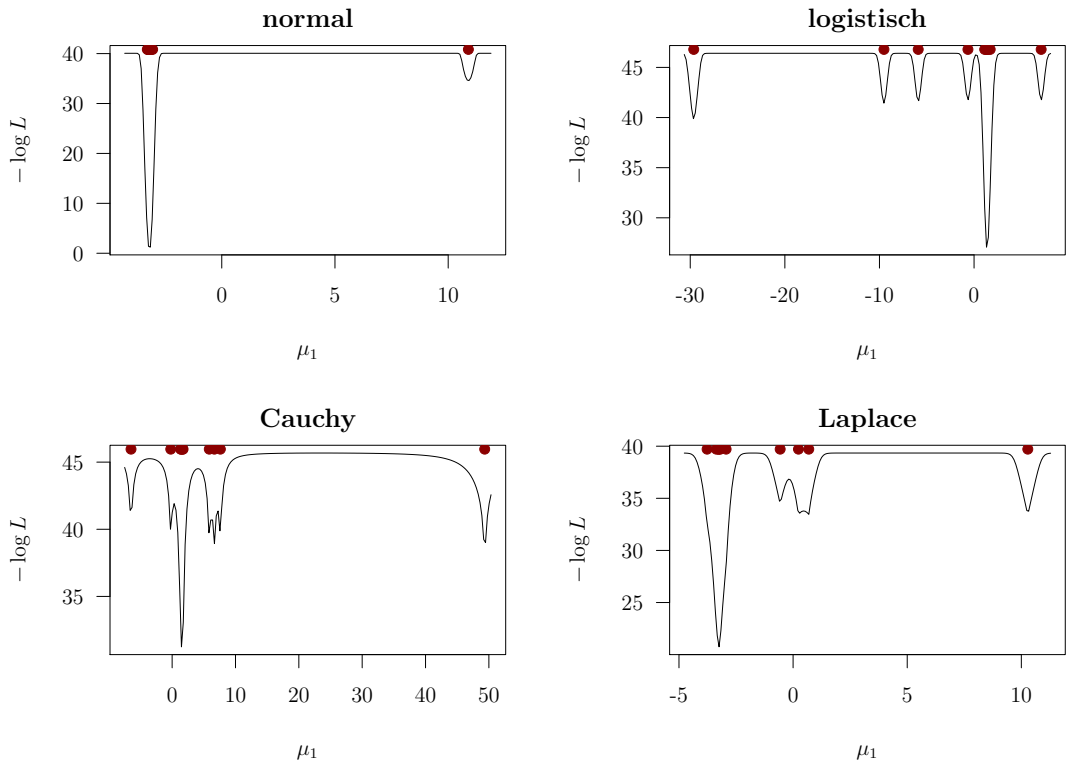

Abbildung 6: Querschnittsdarstellungen mit $n = 10$

lungen betrachtet. In Abbildung 6 sind die Funktionen für $\boldsymbol{\sigma} = (0.1, 10)$ im Querschnitt am theoretischen Wert für μ_2 dargestellt. Zusätzlich sind die 10 gegebenen Beobachtungen am oberen Rand der Grafiken eingezeichnet. Da hier kein festes Intervall betrachtet wird, sondern alle Beobachtungen aus der jeweiligen Zufallsstichprobe dargestellt werden, zeigt sich eine weitere Auswirkung der unterschiedlichen Verteilungscharakteristiken: Da die beobachtete Form der Optima allgemein die Form einer einzelnen Dichte widerspiegelt, lässt sich leicht erklären, dass es für die Cauchy-Verteilung zu deutlich weiter entfernten Ausreißern kommen kann als für die übrigen Verteilungen.

Vor allem ist aber deutlich zu erkennen, dass lokale Optima bei allen Funktionen immer an Stellen bestehen, wo sich eine oder mehrere Beobachtungen befinden. Das deckt sich mit der eingangs erwähnten Problematik der *spurious maximizers* (vgl. McLachlan & Peel, 2000, S. 17 ff.). In Abbildung 7 sind die entsprechenden Grafiken für $n = 100$ zu sehen. Auch hier ist zu erkennen, dass die lokalen Optima immer mit den Werten einzelner gegebener Beobachtungen korrespondieren. Da es hier mehr Beobachtungen gibt, gibt es folglich auch mehr lokale Optima.

Damit ist die grafische Vorstellung der multimodalen Likelihood-Funktionen abgeschlossen. Bevor in Kapitel 2.4 eine systematische Analyse der Optima folgt, wird im folgenden Unterkapitel zunächst der üblicherweise verwendete EM-Algorithmus

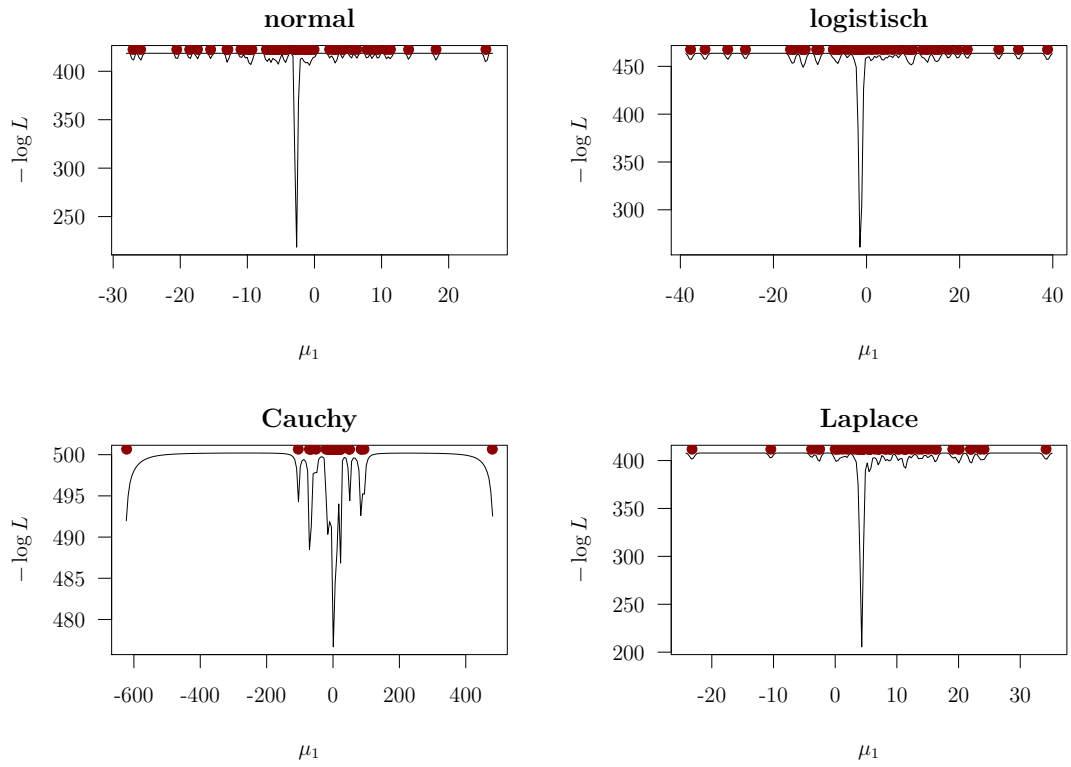

Abbildung 7: Querschnittsdarstellungen mit $n = 100$

zum Optimieren der Likelihood vorgestellt. Mit einigen theoretischen Eigenschaften wird dabei verdeutlicht, dass Multimodalität prinzipiell ein Problem für die Optimierung darstellen kann.

2.3 EM-Algorithmus

Das populärste Verfahren zur Likelihood-Optimierung in Mischverteilungsmodellen ist *expectation maximization* (EM), das von Dempster et al. (1977) vorgestellt wurde.

Das Logarithmieren der Likelihood eines Mischverteilungsmodells, wie in Kapitel 2.1 vorgestellt, führt zu

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{g} \lambda_i f_i(\boldsymbol{x}_j | \boldsymbol{\theta}_i) \right).$$

Da im Allgemeinen keine direkte Lösung für dieses Likelihood-Problem besteht, approximiert der EM-Algorithmus die Maximum-Likelihood-Schätzung $\hat{\boldsymbol{\Psi}}$ iterativ. Dazu wird die *complete-data*-Likelihood $L_c(\boldsymbol{\Psi})$ betrachtet:

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{ij} (\log \lambda_i + \log f_i(\boldsymbol{x}_j | \boldsymbol{\theta}_i)),$$

wobei $z_{ij} = (\boldsymbol{z}_j)_i \in \{0, 1\}$ angibt, dass $\boldsymbol{x}_j$ aus der i-ten Komponente des Modells stammt (McLachlan & Peel, 2000, S. 47-48). Natürlich sind diese Klassenzugehörigkeiten im Bereich des unüberwachten Lernens bzw. der Clusteranalyse nicht bekannt.

Der EM-Algorithmus besteht aus zwei Schritten, die iterativ wiederholt werden. Im E-Schritt wird die bedingte Erwartung von $\log L_c(\boldsymbol{\Psi})$ für die aktuellen Parameterwerte aus $\boldsymbol{\Psi}$ und die gegebenen Beobachtungen $\boldsymbol{x}$ bestimmt. Da $\log L_c(\boldsymbol{\Psi})$ linear in z_{ij} ist, reicht es aus, die bedingte Erwartung der Zufallsvariable Z_{ij} zu berechnen, die durch die Klassenzugehörigkeiten z_{ij} realisiert wird. In der $(k+1)$-ten Iteration führt das zu

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Psi}^{(k)}}(Z_{ij}|\boldsymbol{x}) &= P_{\boldsymbol{\Psi}^{(k)}}(Z_{ij} = 1|\boldsymbol{x}) \\ &= \tau_i\left(\boldsymbol{x}_j|\,\boldsymbol{\Psi}^{(k)}\right),\end{aligned}$$

wobei

$$\tau_i\left(\boldsymbol{x}_j|\,\boldsymbol{\Psi}^{(k)}\right) = \frac{\lambda_i^{(k)} f_i\left(\boldsymbol{x}_j|\boldsymbol{\theta}_i^{(k)}\right)}{\sum_{h=1}^{g} \lambda_h^{(k)} f_h\left(\boldsymbol{x}_j|\boldsymbol{\theta}_h^{(k)}\right)}$$

die A-posteriori-Wahrscheinlichkeit ist, dass $\boldsymbol{x}_j$ aus der i-ten Mischungskomponente stammt. Das führt zum bedingten Erwartungswert

$$Q\left(\boldsymbol{\Psi}|\,\boldsymbol{\Psi}^{(k)}\right) = \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i\left(\boldsymbol{x}_j|\,\boldsymbol{\Psi}^{(k)}\right)(\log\lambda_i + \log f_i(\boldsymbol{x}_j|\boldsymbol{\theta}_i)),$$

der sich durch das Ersetzen der z_{ij} mit den A-posteriori-Wahrscheinlichkeiten aus der *complete-data*-Likelihood ergibt (McLachlan & Peel, 2000, S. 48 f.).

Der M-Schritt in der $(k+1)$-ten Iteration des Algorithmus besteht aus der Maximierung von $Q\left(\boldsymbol{\Psi}|\,\boldsymbol{\Psi}^{(k)}\right)$. Wenn die Klassenzugehörigkeiten bekannt wären, könnten die Mischungsanteile durch die relativen Häufigkeiten geschätzt werden. Allerdings können hier, wie im E-Schritt, wieder die A-posteriori-Wahrscheinlichkeiten verwendet werden, was zu folgendem Ausdruck führt:

$$\lambda_i^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_i\left(\boldsymbol{x}_i|\,\boldsymbol{\Psi}^{(k)}\right)}{n}.$$

Die verbleibenden Parameter $\boldsymbol{\xi}^{(k+1)}$ lassen sich ermitteln, indem die A-posteriori-Wahrscheinlichkeiten mit dem Gradienten multipliziert und für alle Beobachtungen und Komponenten aufsummiert gleich Null gesetzt werden:

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i\left(\boldsymbol{x}_j|\,\boldsymbol{\Psi}^{(k)}\right)\frac{\partial \log f_i(\boldsymbol{x}_j|\boldsymbol{\theta}_i))}{\partial\boldsymbol{\xi}} = \mathbf{0}.$$

Für Mischungen von Normalverteilungen existieren ML-Schätzer für die Mittelwerte $\boldsymbol{\mu}_i$ und die Kovarianzmatrizen $\boldsymbol{\Sigma}_i$ in geschlossener Form (McLachlan & Peel, 2000, S. 82):

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i \left(\boldsymbol{x}_j | \, \boldsymbol{\Psi}^{(k)} \right) \boldsymbol{x}_j}{\sum_{j=1}^{n} \tau_i \left(\boldsymbol{x}_j | \, \boldsymbol{\Psi}^{(k)} \right)}$$

und

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i \left(\boldsymbol{x}_j | \, \boldsymbol{\Psi}^{(k)} \right) \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k+1)} \right) \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k+1)} \right)^T}{\sum_{j=1}^{n} \tau_i \left(\boldsymbol{x}_j | \, \boldsymbol{\Psi}^{(k)} \right)}.$$

Die Berechnungen von E- und M-Schritt werden alternierend wiederholt bis die Differenz

$$L\left(\boldsymbol{\Psi}^{(k+1)} \right) - L\left(\boldsymbol{\Psi}^{(k)} \right)$$

hinreichend gering ist. Natürlich ist ein solches Abbruchkriterium nur sinnvoll, wenn Konvergenz zu einem Optimum erreicht wird (McLachlan & Peel, 2000, S. 49 f.).

Einige Konvergenzeigenschaften des EM-Algorithmus wurden von Wu (1983) zusammengetragen. Grundsätzlich ist garantiert, dass

$$L\left(\boldsymbol{\Psi}^{(k+1)} \right) \geq L\left(\boldsymbol{\Psi}^{(k)} \right),$$

der erreichte Likelihood-Wert sich also im Laufe der Iterationen nicht verschlechtern kann. Darüber hinaus gilt

$$L\left(\boldsymbol{\Psi}^{(k)} \right) \overset{k \to \infty}{\longrightarrow} L^*,$$

wenn L nach oben beschränkt ist. L^* ist dabei ein stationärer Wert von L, wenn $Q\left(\boldsymbol{\Psi} | \, \boldsymbol{\Psi}^{(k)} \right)$ stetig in $\boldsymbol{\Psi}$ und $\boldsymbol{\Psi}^{(k)}$ ist. Damit konvergiert der Algorithmus unter diesen Voraussetzungen entweder zu einem Sattelpunkt oder zu einem lokalen Optimum.

Allerdings ist die Beschränktheitsbedingung für Mischungen univariater Normalverteilungen mit unterschiedlichen Varianzen nicht erfüllt. Wie von Day (1969) beschrieben, entsteht ein unendlicher Likelihood-Wert, wenn eine Komponente als nur aus einem der gegebenen Punkte bestehend aufgefasst wird, da die zugehörige Varianz in diesem Fall 0 ist. Der EM-Algorithmus findet dann in jedem Schritt für eine weitere Annäherung von σ_1 an 0 eine Verbesserung des Likelihood-Wertes und terminiert theoretisch nie. Praktisch kommt es zu einem Fehler, sobald σ_1 nicht mehr numerisch von 0 zu unterscheiden ist und der Funktionswert damit unendlich wird.

Für die bisher beschriebenen zweidimensionalen Optimierungsprobleme für die Lageparameter der Verteilungen besteht diese Problematik nicht, da die festen Varianzparameter sich nicht gegen 0 entwickeln können. Von Hathaway (1985) wird vorgeschlagen, das Verhältnis der beiden Varianzen durch eine zu wählende Konstante zu beschränken, um eine Entwicklung gegen 0 bei gleichzeitiger Beibehaltung der anderen Varianz zu verhindern. Die festen Varianzen in den zweidimensionalen Beispielen können mit ihrem festen Verhältnis als Spezialfall solch einer Restriktion aufgefasst werden. Für Optimierungsprobleme, in denen weitere Parameter optimiert werden sollen, wird in den entsprechenden Kapiteln auf die bestehende Problematik der Unbeschränktheit detailliert eingegangen.

An dieser Stelle bleibt festzuhalten, dass es für den EM-Algorithmus keine Garantie gibt, auf den betrachteten multimodalen Problemen das globale Optimum zu finden und daher eine genauere Analyse sinnvoll erscheint. Bevor die Optimierungsleistung von EM in Kapitel 3 mit anderen Algorithmen verglichen wird, erfolgt im folgenden Unterkapitel eine Charakterisierung der Komplexität der darin zu verwendenden Likelihood-Funktionen.

2.4 Heuristische Analyse multimodaler Likelihood-Funktionen

Um festzustellen, wie die Komplexität der Likelihood-Funktionen beeinflusst werden kann, werden zwei Eigenschaften ermittelt: die Anzahl der lokalen Optima und die Größe des Einzugsbereiches des globalen Optimums. Für beide Eigenschaften wird ein Querschnitt der Funktion am theoretischen Mittelwertparameter μ_2 der Verteilung mit größerer Varianz gebildet. Die resultierende eindimensionale Funktion besitzt μ_1 als alleinigen Parameter. Es ist möglich diese Vereinfachung für die betrachteten Beispiele anzuwenden, da davon ausgegangen wird, dass die Existenz der Optima nicht von μ_2 abhängt.

Der Einzugsbereich wird für den eindimensionalen Querschnitt im Folgenden auch als *Breite des globalen Optimums* bezeichnet. Als erstes wird ein beliebiges Optimierungsverfahren für konvexe Funktionen am theoretischen Wert von μ_1 gestartet, um das empirische Optimum in der vorliegenden Realisation der Likelihood-Funktion zu approximieren. Vom resultierenden Punkt ausgehend wird die Funktion in beiden Richtungen schrittweise ausgewertet bis das nächste Maximum (oder Sattelpunkt) auf jeder Seite erreicht ist. Die Breite des globalen Optimums ergibt sich dann als Differenz der μ_1-Werte beider Maxima. Um die Anzahl lokaler Maxima zu ermitteln, wird die Funktion auf dem gesamten Bereich äquidistant abgetastet. Wenn der Wert an einem Punkt kleiner ist als die Werte seiner beiden Nachbarn,

wird der Punkt als lokales Optimum gezählt. In Abbildung 8 sind beide Heuristiken für die Funktion im Zentrum von Abbildung 2 dargestellt. Die Schrittweite beträgt in beiden Fällen 0.1. Die Ergebnisse sind eine Breite von 1.1 und eine Anzahl von 9 lokalen Optima.

Um diese Eigenschaften systematisch zu analysieren, wird eine Simulationsstudie auf den Likelihood-Funktionen von Mischungen zweier univariater Normalverteilungen mit unterschiedlich stark voneinander abweichenden Varianzparametern durchgeführt. Es sind sechs verschiedene Varianzkombinationen aus $\sigma_1, \sigma_2 \in \{0.01, 0.1, 1, 10\}$ enthalten, von denen drei bereits im Eingangsbeispiel gezeigt wurden. Entsprechend sind auch die drei Stichprobengrößen $n \in \{10, 100, 1000\}$ enthalten. In den Tabellen 1 und 2 sind die Ergebnisse für die Breite des globalen Optimums und die Anzahl lokaler Optima dargestellt. Die angegebenen Werte entsprechen

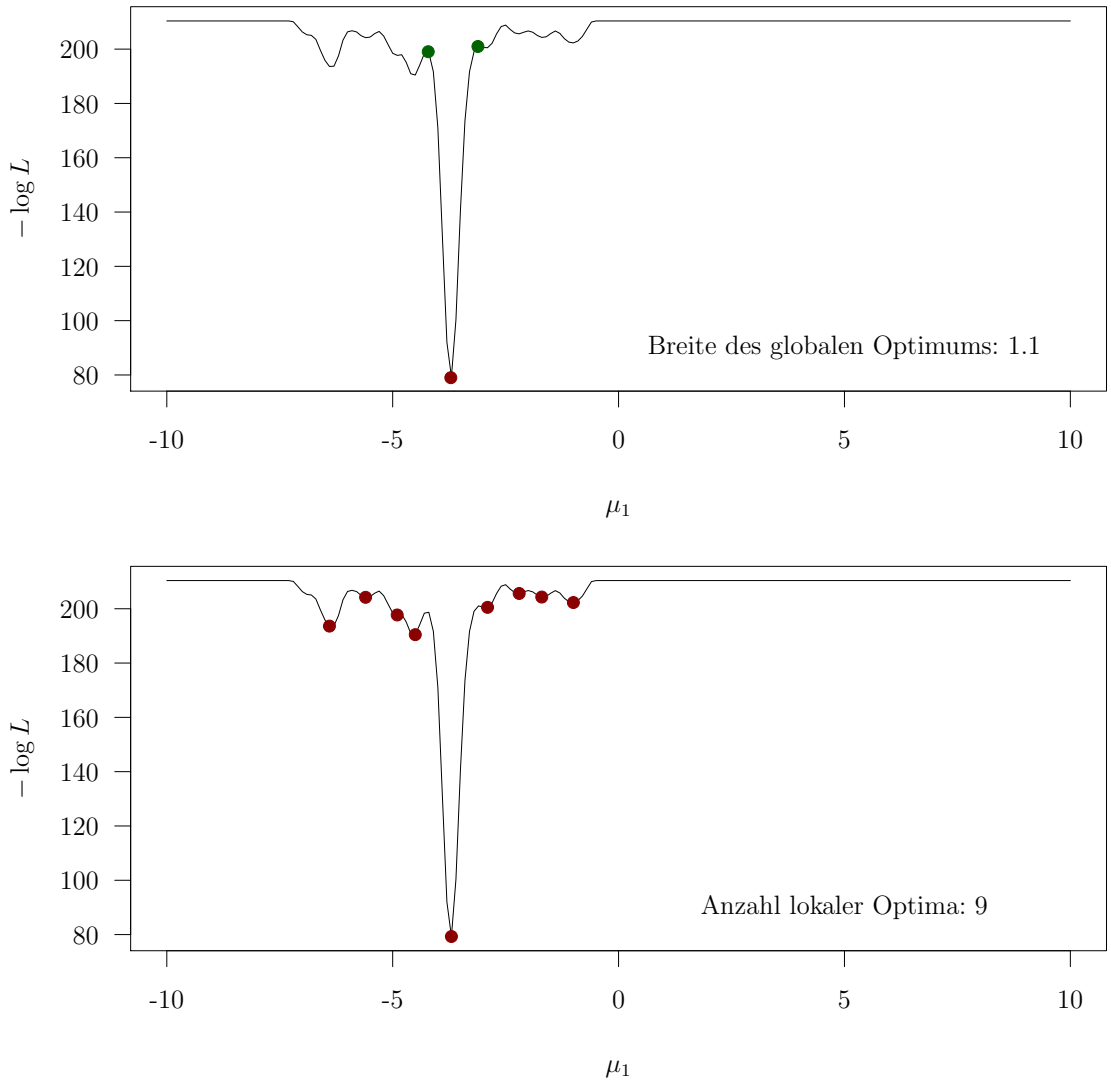

Abbildung 8: Breite des globalen Optimums und Anzahl der lokalen Optima

dem arithmetischen Mittel über die Ergebnisse von jeweils 100 Zufallsstichproben aus der entsprechenden Mischverteilung. Es ist ersichtlich, dass die Breite des globalen Optimums in Tabelle 1 stark mit der Größe von σ_1 korreliert ist. Kleine Werte von σ_1 führen zu einem schmalen globalen Optimum, während größere Werte den Einzugsbereich des Optimums breiter werden lassen. σ_2 scheint keinen Einfluss auf die Breite zu haben. Darüber hinaus kann eine leichte zusätzliche Verbreiterung des Optimums für große Stichproben beobachtet werden.

Die Anzahl lokaler Optima in Tabelle 2 hängt hingegen nicht von den Werten der Standardabweichungen selbst, sondern vom Abstand zwischen ihnen ab. Der Abstand ist definiert als $\sigma_{\Delta_{10}} := |\log_{10}\sigma_1 - \log_{10}\sigma_2|$. Für den Abstand von einer Zehnerpotenz ($\boldsymbol{\sigma} \in \{(0.01, 0.1), (0.1, 1), (1, 10)\}$) treten die geringsten Anzahlen lokaler Optima auf und es ist kein Effekt von n beobachtbar. Für den Abstand von drei Zehnerpotenzen ($\boldsymbol{\sigma} = (0.01, 10)$) hingegen wird die größte Anzahl lokaler Optima gezählt. Hier kommt es durch zusätzliche Erhöhung von n zu einem weiteren Anstieg der Anzahl lokaler Optima.

Die geschilderten Zusammenhänge sollen nun durch Modellschätzungen bestätigt werden. Dazu wird für die Breite des globalen Optimums sowie die Anzahl lokaler Optima jeweils ein lineares Regressionsmodell angepasst. In den Tabellen 3 und 4 sind die Koeffizienten der beiden Modelle dargestellt. Neben den Modellkoeffizienten für die enthaltenen Variablen wird immer ein Koeffizient β_0 für den Achsenabschnitt mitgeschätzt. Zur Verfügung stehende Einflussgrößen sind σ_1, σ_2, n und alle Interaktionen zweiter Ordnung. Für die Breite des globalen Optimums

Tabelle 1: Mittlere Breite des globalen Optimums (Normalverteilungen)

	$\boldsymbol{\sigma} = (0.01, 0.1)$	(0.01, 1)	(0.01, 10)	(0.1, 1)	(0.1, 10)	(1, 10)
$n = 10$	0.637	0.226	0.201	1.593	1.448	11.028
$n = 100$	0.578	0.289	0.243	1.805	1.590	13.108
$n = 1000$	0.634	0.356	0.351	2.043	1.613	18.997

Tabelle 2: Mittlere Anzahl lokaler Optima (Normalverteilungen)

	$\boldsymbol{\sigma} = (0.01, 0.1)$	(0.01, 1)	(0.01, 10)	(0.1, 1)	(0.1, 10)	(1, 10)
$n = 10$	1.76	5.04	4.29	4.11	3.77	2.15
$n = 100$	1.98	14.92	26.58	6.81	17.40	2.35
$n = 1000$	1.57	20.56	65.72	6.16	22.70	2.09

B wird das beste adjustierte R^2 von 0.996 von dem Modell (Tabelle 3) erreicht, das σ_1, n und die zugehörige Interaktion enthält:

$$B_i = \beta_0 + \beta_{\sigma_1}\sigma_{1i} + \beta_n n_i + \beta_{\sigma_1 \times n}\sigma_{1i} n_i + \varepsilon_i.$$

Dass die größere Standardabweichung σ_2 nicht enthalten ist, kann dadurch erklärt werden, dass die Form des betrachteten Optimums im Querschnitt gerade die Dichtefunktion der Verteilungskomponente mit kleiner Varianz widerspiegelt und die Breite dieses Optimums somit unabhängig von der Varianz der anderen Komponente ist. Mittels t-Test erweisen sich β_0, β_{σ_1} und $\beta_{\sigma_1 \times n}$ als signifikant (z. N. $\alpha = 0.05$), was die obigen Beobachtungen bestätigt: Die Breite des globalen Optimums hängt von σ_1 ab und wenn σ_1 groß ist, zeigt sich für steigendes n eine zusätzliche Verbreiterung des Optimums.

Für die Anzahl lokaler Optima A enthält das beste Modell (Tabelle 4) $\sigma_{\Delta_{10}}$, n und den zugehörigen Interaktionsterm:

$$A_i = \beta_0 + \beta_{\sigma_{\Delta_{10}}}\sigma_{\Delta_{10}i} + \beta_n n_i + \beta_{\sigma_{\Delta_{10}} \times n}\sigma_{\Delta_{10}i} n_i + \varepsilon_i.$$

$\beta_{\sigma_{\Delta_{10}} \times n}$ und β_n stellen sich dabei als signifikant heraus. Zusammengefasst lautet die Interpretation hier: Je größer der Abstand zwischen den Standardabweichungen, desto mehr Optima treten auf. Außerdem erhöhen große Beobachtungszahlen

Tabelle 3: Regressionskoeffizienten zur Erklärung der Breite des Optimums (adj. $R^2 = 0.996$)

	Schätzung	Standardfehler	p-Wert
β_0	0.3253	0.1190	1.62×10^{-02}
β_{σ_1}	11.3244	0.2886	1.01×10^{-15}
β_n	−0.0002	0.0002	3.83×10^{-01}
$\beta_{\sigma_1 \times n}$	0.0076	0.0005	4.24×10^{-10}

Tabelle 4: Regressionskoeffizienten zur Erklärung der Anzahl Optima (adj. $R^2 = 0.855$)

	Schätzung	Standardfehler	p-Wert
β_0	−1.2711	4.4926	0.7814
$\beta_{\sigma_{\Delta_{10}}}$	4.7780	2.4607	0.0726
β_n	−0.0270	0.0077	0.0036
$\beta_{\sigma_{\Delta_{10}} \times n}$	0.0243	0.0042	0.0001

die Anzahl der lokalen Optima zusätzlich, wenn große Abstände zwischen den Standardabweichungen vorliegen.

Zum Abschluss der heuristischen Analyse werden die Mischungen von Normalverteilungen mit denen der weiteren in Kapitel 2.2 vorgestellten Verteilungen verglichen. In Abbildung 9 sind die Breiten der globalen Optima für Cauchy-, Laplace- und Normalverteilung sowie die logistische Verteilung dargestellt. Es ist zu erkennen, dass sich die Breite des Optimums bei Mischungen von Cauchy-Verteilungen deutlich von denen der weiteren Verteilungen unterscheidet. Durch den runden Verlauf der Likelihood-Funktion sind die Optima mit der hier verwendeten Methode grundsätzlich breiter als für die übrigen Verteilungen. Zwei benachbarte Optima grenzen im Gegensatz zu den anderen Verteilungen immer direkt aneinander, da dazwischen kein Plateau mit konstanten Funktionswerten besteht (vgl. Abb. 4, S. 10). Dadurch ist insbesondere bei wenigen vorhandenen Optima für $n = 10$ die grundsätzlich größere Breite zu erklären. Für größere Werte von n nähert sich die Breite den übrigen Verteilungen an, da deutlich mehr Optima existieren und das globale Optimum gar nicht mehr eine so große Breite einnehmen kann. Die einzige Ausnahme davon bildet der Fall $\boldsymbol{\sigma} = (0.01, 0.1)$: Hier steigt die Breite für $n = 100$ sogar deutlich an und ist für $n = 1000$ nur geringfügig kleiner als für $n = 10$. In diesem Szenario ist auch die größere Standardabweichung mit 0.1 relativ klein, weswegen die Heuristik mit fester Abtastrate hier zu Ungenauigkeiten führen kann. Erkennbare Unterschiede der übrigen Verteilungen sind darüber hinaus lediglich im Fall $\boldsymbol{\sigma} = (1, 10)$ zu erkennen, wo die Breite für $n = 1000$ bei allen Verteilungen deutlich ansteigt. Hier ist zu beachten, dass es sich um den für die Optimierung einfachen Fall mit sehr niedriger Multimodalität handelt. Dementsprechend besteht für Mischungen aller Verteilungen ein relativ deutliches, breites Optimum. Dass die Breiten sich erkennbar unterscheiden, liegt erneut an den unterschiedlichen Charakteristika der Verteilungen. Im Gegensatz zu den rund auslaufenden Optima bei der Cauchy-Verteilung verläuft die Normalverteilungs-Likelihood vom Optimum aus sehr steil bis sie abrupt in ein Plateau übergeht und bildet damit das andere Extrem unter den betrachteten Verteilungen, während Laplace- und logistische Verteilung dazwischen einzusortieren sind.

Für die Anzahl lokaler Optima ergibt sich auch bezüglich der Cauchy-Verteilung ein einheitlicheres Bild. In Abbildung 10 ist zu erkennen, dass sich die Anzahl Optima bei den meisten Varianzkombinationen für steigendes n erhöht. Einzig für $\boldsymbol{\sigma} = (1, 10)$ aufgrund nicht vorhandener Multimodalität und für $\boldsymbol{\sigma} = (0.01, 0.1)$, was erneut auf die zu kleine größere Varianz zurückzuführen sein kann, bleibt die Anzahl Optima konstant. Für $\sigma_1 = 0.01$ zeigt die Cauchy-Verteilung zweimal leichte Abweichungen bei $n = 1000$, einmal gilt dies für die Normalverteilung.

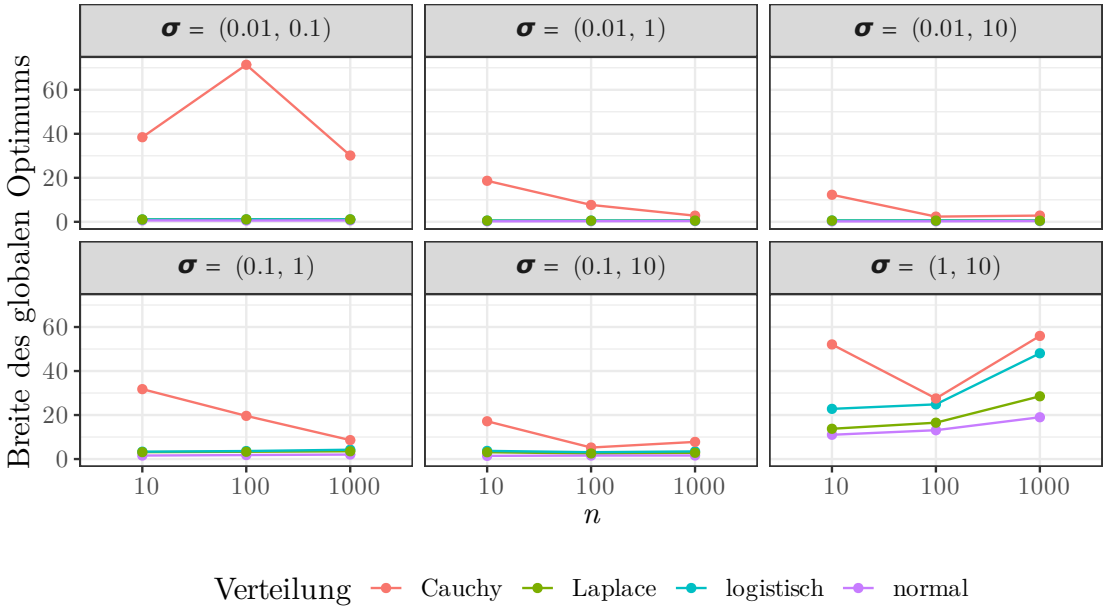

Abbildung 9: Mittlere Breite des globalen Optimums

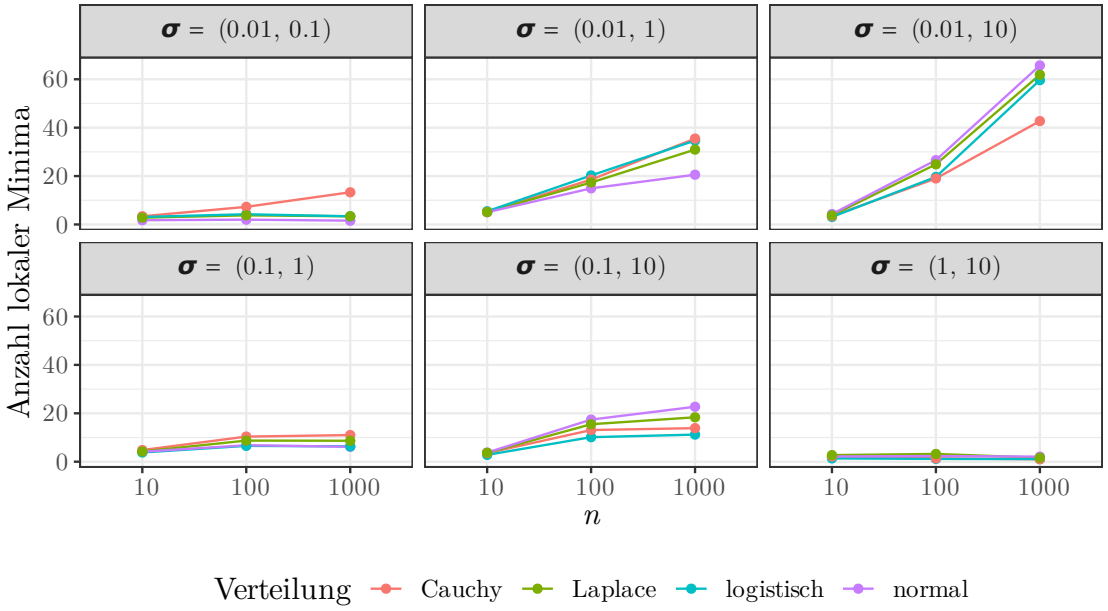

Abbildung 10: Mittlere Anzahl lokaler Optima

Mit Ausnahme der Breite des Optimums bei Cauchy-Verteilungen lässt sich damit insgesamt erkennen, dass die Likelihood-Funktionen von Mischungen der weiteren Verteilungen bezüglich Multimodalität und Breite des Optimums sehr ähnlich zu denen von Normalverteilungsmischungen sind. Dementsprechend beschränken sich die weiteren Analysen auf Mischungen von Normalverteilungen, da für die übrigen Verteilungen keine grundsätzlich anderen Ergebnisse erwartet werden.

Damit ist die heuristische Analyse der multimodalen Likelihood-Funktionen von Mischungen zweier univariater Verteilungen abgeschlossen. Die beiden betrachteten Aspekte werden in den folgenden Analysen erneut aufgegriffen: Während in Kapitel 3 das Auffinden des (globalen) Optimums behandelt wird, geht es in Kapitel 4 darum, möglichst viele lokale Optima zu identifizieren.

3 Optimierung multimodaler Likelihood-Funktionen

Um herauszufinden, wie gut die Optimierung der multimodalen Likelihood-Funktionen gelingt und um speziell den EM-Algorithmus mit anderen Optimierungsverfahren zu vergleichen, wird eine Simulationsstudie durchgeführt. Zunächst wird in Kapitel 3.1 der Aufbau der Simulation beschrieben. In Kapitel 3.2 werden die verwendeten Optimierungsalgorithmen vorgestellt, bevor in den Kapiteln 3.3 und 3.4 die Ergebnisse des zweidimensionalen und des fünfdimensionalen Teils der Simulation präsentiert werden. Das letzte Unterkapitel 3.5 beschäftigt sich mit einer detaillierten Analyse der Optimierungsverläufe.

3.1 Simulationsdesign

Um die Leistungsfähigkeit verschiedener Optimierungsalgorithmen auf den vorgestellten Likelihood-Funktionen aus Mischverteilungsmodellen zu vergleichen, wird eine Simulationsstudie durchgeführt. Enthaltene Algorithmen sind EM, CMA-ES, MBO sowie eine Zufallssuche (*Zufall*). Letztere gilt als Basismethode, die von den aufwändigeren Verfahren geschlagen werden sollte. Als Alternativen zum üblicherweise verwendeten EM werden mit CMA-ES und MBO zwei allgemeine Optimierungsalgorithmen in den Vergleich aufgenommen, die als *Black-Box-Optimierer* keine Vorkenntnisse über die Likelihood-Funktion verwenden. Dabei werden zwei verschiedene Ansätze genutzt, um im Laufe der Optimierung lokale Optima wieder verlassen zu können. Die zugrunde liegenden Prinzipien beider Verfahren werden im folgenden Unterkapitel 3.2 kurz vorgestellt. Alle Algorithmen werden auf 1500 verschiedene Instanzen von Likelihood-Funktionen angewendet. Die 1500 Instanzen ergeben sich durch Kombination von fünf verschiedenen Wertepaaren für $\boldsymbol{\sigma}$, drei verschiedenen Stichprobengrößen und 100 Wiederholungen für jede Kombination. Zu Beginn (Kap. 3.3) werden wie im vorherigen Kapitel zweidimensionale Likelihood-Funktionen, mit denen lediglich die Lageparameter der beiden Mischungskomponenten bestimmt werden sollen, betrachtet. Während die übrigen Algorithmen in jedem Lauf von einem einzelnen Punkt starten, benötigt MBO ein initiales Design bestehend aus mehreren Punkten. Dazu wird ein *maximin latin hypercube sample* (LHS) der Größe 10 im Bereich $[-10, 10]^2$ gezogen. Für die Zufallssuche werden diese Punkte als erste zehn Iterationen verwendet, für CMA-ES und EM wird der Designpunkt mit dem besten Funktionswert verwendet. Da es keine Möglichkeit gibt, identische Startvorraussetzungen zwischen Startpunkt und Startdesign herzustellen, soll den beiden letztgenannten Algorithmen durch Wahl

des besten Punktes aus dem Startdesign möglichst wenig relevante Information gegenüber den anderen Algorithmen vorenthalten werden. Zum Abschluss der 5D-Analyse in Kapitel 3.4 wird ein zusätzlicher Vergleich zwischen dem besten und einem zufälligen Startpunkt aus dem Startdesign für EM betrachtet, um mögliche Auswirkungen der Wahl des Startpunkts zu diskutieren. Jedem Optimierungslauf stehen insgesamt 100 Auswertungen der Zielfunktion zur Verfügung und es gibt insgesamt 100 Wiederholungen der Läufe mit anderen Startpunkten.

Da das theoretische Optimum von $\boldsymbol{\mu}$ nicht mit dem empirischen Optimum in der vorliegenden Realisation der Funktion übereinstimmt, wird das empirische Optimum durch einen EM-Optimierungslauf startend am theoretischen Optimum (*EM Theorie*) approximiert.

In den vorangegangenen Kapiteln wurden ausschließlich 2D-Likelihood-Funktionen mit festen Varianzen und Mischungsverhältnissen betrachtet, insbesondere weil dort auch eine grafische Betrachtung möglich ist. Diese Funktionen werden im ersten Teil der Simulationsstudie (Kap. 3.3) optimiert. Der Fall, dass die unterschiedlichen Varianzen zweier gemischter Verteilungen bekannt sind, deren Mittelwerte allerdings nicht, ist allerdings kein realistischer Anwendungsfall. Daher werden im zweiten Teil (Kap. 3.4) Likelihood-Funktionen verwendet, bei denen zusätzlich zu den Mittelwerten auch die Varianzen und das Mischungsgewicht optimiert werden sollen. Dazu werden neun verschiedene Mischungsgewichte $\lambda \in \{0.1, ..., 0.9\}$ und alle neun möglichen Kombinationen von $\mu_1, \mu_2 \in \{-5, 0, 5\}$ als Mittelwerte betrachtet. Die restlichen Simulationsparameter werden aus dem 2D-Fall übernommen, was zu insgesamt 121 500 Funktionsinstanzen führt. Sofern im Zusammenhang mit diesen Funktionen die Bezeichnung *globales Optimum* verwendet wird, ist damit grundsätzlich das beste beschränkte Optimum gemeint, da die Funktionen allgemein unbeschränkt sind (vgl. Kap. 2.3, S. 15).

Zum Aufbau und zur parallelisierten Durchführung der Simulationsstudie auf dem Linux-HPC-Rechencluster der TU Dortmund (LiDO3) wird das R-Paket *batchtools* (Lang et al., 2017) verwendet.

3.2 Optimierungsverfahren

EM

Die Theorie zum EM-Algorithmus wurde bereits in Kapitel 2.3 (ab S. 13) vorgestellt. Die verwendete Implementierung basiert auf dem R-Paket *mixtools* (Benaglia et al., 2009). Aufteilungen mit nur einer Beobachtung in einer der Verteilungskomponenten werden in dieser Simulationsstudie als fehlgeschlagene Optimierungsversuche betrachtet, da sie aufgrund der Unbeschränktheit keinen sinnvoll interpre-

tierbaren Likelihoodwert aufweisen (vgl. Kap. 2.3, S. 15). Ein weiteres Problem entsteht, wenn im Laufe der Optimierung alle Beobachtungen vollständig einer Klasse zugewiesen werden, der Mischungsparameter numerisch für eine Komponente also 1 und für die andere 0 wird. Dann schlägt die Optimierung fehl, da zur Bestimmung der Verteilungsparameter einer Komponente eine Division durch 0 erfolgen müsste. Zu Beginn der 5D-Studie (ab S. 26) wird das Auftreten fehlgeschlagener Optimierungen genauer dargestellt, für das zunächst betrachtete Teilproblem der Mittelwerte existieren diese Probleme auf Grund der festen Mischungsverhältnisse und Varianzen nicht.

CMA-ES

Bei der *Covariance Matrix Adaption Evolutionary Strategy* (*kurz:* CMA-ES) handelt es sich um einen numerischen Optimierungsalgorithmus aus der Klasse der evolutionären Algorithmen. Allgemein basieren diese Algorithmen auf der Idee, eine Population von Kandidatenpunkten durch Variation und Selektion zu verbessern. Konkret werden einzelne Individuen aus der vorhandenen Population dazu zufällig kleinen Veränderungen unterzogen (Mutation) oder mehrere Individuen zu einem neuen Kandidatenpunkt zusammengesetzt (Rekombination). Durch Variation wird üblicherweise eine deutlich größere Menge an Individuen generiert, von denen nur diejenigen mit den besten Funktionswerten in die nächste Generation selektiert werden. Dadurch können kontinuierlich Lösungen mit besseren Funktionswerten generiert werden, bis idealerweise die gesamte Population zu einem optimalen Wert konvergiert.

Evolutionäre Strategien zeichnen sich dadurch aus, dass neue Individuen immer zufällig aus einer multivariaten Normalverteilung gezogen werden. Dabei wird im Falle der Rekombination ein neuer Mittelwert der Verteilung aus bestimmten Individuen generiert, während zur Mutation Zufallszahlen auf vorhandene Individuen aufaddiert werden. Die Grundidee von CMA-ES ist es auch die Kovarianzmatrix dieser multivariaten Verteilung auf Basis der vorhandenen Population zu adaptieren, um möglichst häufig gute Funktionswerte zu finden. Dazu wird die Kovarianz vom aktuellen Mittelwert aus in Richtung der besten Funktionswerte vergrößert bzw. in den orthogonalen Richtungen zusammengezogen. Durch diese flexible Anpassung soll ein höherer Anteil guter Kandidaten in der Variationsmenge erreicht werden als durch von der Population unabhängige Kovarianzmatrizen. Für die mathematischen Details der Kovarianzmatrix-Adaption wird auf Hansen & Ostermeier (1996) und Hansen (2006) verwiesen, da die folgenden Analysen keinen Bezug darauf nehmen.

CMA-ES wird in dieser Analyse verwendet, da die stochastische Veränderung der Population auf nicht-konvexen Optimierungsproblemen dazu führt, dass der Ein-

zugsbereich eines lokalen Optimums verlassen werden kann und so im Verlaufe der Optimierung Optima mit besseren Funktionswerten gefunden werden können. Zur Durchführung der Optimierungsläufe wird das R-Paket *cmaesr* (Bossek, 2016) verwendet.

MBO

Mit *modellbasierter Optimierung* (*kurz:* MBO) wird allgemein das Vorgehen bezeichnet, eine Zielfunktion durch ein geeignetes Surrogatmodell zu ersetzen, welches stattdessen optimiert wird und damit auch eine Lösung für die Zielfunktion liefern soll. Prinzipiell eignet sich dieses Vorgehen besonders, wenn Auswertungen der Zielfunktion sehr kosten- oder zeitaufwändig sind. Auf die hier betrachteten Likelihoodfunktionen trifft dies zwar nicht zu, dennoch kommt MBO hier als alternatives Verfahren zum Einsatz, da Lösungen über rein lokale Konvergenz hinaus erreicht werden können.

Konkret basiert der verwendete Algorithmus auf dem Prinzip der *Efficient Global Optimization (EGO)* (Jones et al., 1998). Dabei wird als Surrogatmodell ein Kriging- bzw. Gaußprozess-Modell verwendet, welches die Zielfunktion durch Interpolation weniger initialer Auswertungen aus einem Latin-Hypercube-Design annähert. Anschließend werden auf Basis des Modells iterativ Kandidatenpunkte für ein Optimum identifiziert, auf der Zielfunktion ausgewertet und in das Modell aufgenommen. Dabei wird nicht einfach das Optimum des Modells verwendet, sondern der Punkt an dem das *Expected Improvement* am größten ist. Diese erwartete Verbesserung gegenüber dem besten bekannten Wert lässt sich als Gewichtung zwischen Optimalität und Unsicherheit des Surrogatmodells an einem Punkt formulieren. Prinzipiell ist die Unsicherheit in Bereichen, in denen bisher kaum Auswertungen vorliegen, größer als in der Nähe bekannter Punkte, wodurch die globale Exploration der Zielfunktion ermöglicht wird.

Für eine formale Beschreibung des Verfahrens wird auf (Jones et al., 1998) verwiesen, da auch in diesem Fall in den weiteren Analysen kein Bezug darauf genommen wird. Angewendet wird das Verfahren mit Hilfe des R-Pakets *mlrMBO* (Bischl et al., 2017), welches auf Kriging-Modelle aus *DiceKriging* (Roustant et al., 2012) als Surrogatmodell zurückgreift. Im Unterschied zu *EGO* wird dort mit dem *Lower Confidence Bound* standardmäßig eine direkte Gewichtung von Punkt- und Unsicherheitsschätzer des Modells verwendet.

3.3 2D-Simulationsstudie

In dieser ersten Simulation wird das Optimierungsproblem, wie in den bisherigen Vorbetrachtungen auch, auf den Raum der Mittelwerte der beiden Verteilungskom-

ponenten beschränkt. Die Simulationsergebnisse für das zweidimensionale Teilproblem mit festen Varianzen und Mischungsverhältnissen sind in Abbildung 11 dargestellt. In jeder Teilgrafik sind die erreichten Funktionswerte auf 100 Instanzen nach Algorithmus getrennt als parallele Boxplots dargestellt. Die festen Parameter ergeben sich für jede Teilgrafik aus Spalten- und Zeilenbeschriftung. Für die verschiedenen Varianzkombinationen ist basierend auf den Ergebnissen der vorausgehenden Analyse (Kap. 2.4) zusätzlich angegeben, wie groß der Einzugsbereich des globalen Optimums ist („*global*“) und wie viele lokale Optima existieren („*lokal*“). Es ist anzumerken, dass insgesamt sechs der 1500 MBO-Läufe kein Ergebnis geliefert haben und folglich nicht in den Grafiken enthalten sind. Aufgrund des äußerst geringen Anteils und da keine Struktur bezüglich der Simulationsparameter dieser Läufe erkennbar ist, wurde auf eine weitere Analyse dieser fehlgeschlagenen Läufe verzichtet.

Den Ergebnissen entsprechend ist die Optimierung für $\boldsymbol{\sigma} \in \{(0.01, 1), (0.01, 10)\}$ generell herausfordernder als für $\boldsymbol{\sigma} \in \{(0.1, 1), (0.1, 10)\}$, obwohl die Kombinationen $\boldsymbol{\sigma} = (0.01, 1)$ und $\boldsymbol{\sigma} = (0.1, 10)$ relativ ähnliche Ergebnisse für die Anzahl lokaler Optima in der vorausgehenden Analyse geliefert haben. Die besten Ergebnisse werden für die Standardabweichungen $\boldsymbol{\sigma} = (1, 10)$ erreicht, da die resultierenden Funktionen nahezu unimodal sind. Generell bestätigt sich hier die Vermutung aus dem vorherigen Kapitel, dass die Breite des globalen Optimums Einfluss auf die Komplexität des Optimierungsproblems hat. Eine direkte Auswirkung der Multimodalität ist hier nicht erkennbar.

Insgesamt liefert MBO hier die besten Ergebnisse, für kleines n teilweise sogar bessere als der EM-Algorithmus mit Startpunkt am theoretischen Optimum. Der EM-Algorithmus liefert deutlich schlechtere Ergebnisse als die übrigen Algorithmen. Hier liegt die Vermutung nahe, dass der EM-Algorithmus mit festen Varianzen und festem Mischungsverhältnis nicht wie erwartet funktioniert.

In Abbildung 12 ist statt des erreichten Funktionswertes der Abstand der Optimierungsergebnisse zum theoretischen Optimum auf Parameterebene dargestellt. Grundsätzlich haben die Ergebnisse des EM-Algorithmus mit Start am theoretischen Optimum auch die geringste euklidische Distanz zu diesem. MBO ist teilweise ähnlich nah dran, im Fall von $\boldsymbol{\sigma} = (0.1, 1)$ und $n = 10$ sogar besser. Bei nur zehn Beobachtungen ist es auch durchaus möglich, dass theoretisches Optimum und empirischer Optimalwert bezüglich der gegebenen Beobachtungen voneinander abweichen und somit von den theoretischen Werten ausgehend nur ein schlechteres lokales Optimum erreicht werden kann. Für steigendes n werden diese Abweichungen kleiner, sodass *EM Theorie*, wie schon bei den Funktionswerten zu erkennen, stabiler gute Ergebnisse erreicht. CMA-ES und die Zufallssuche landen üblicherweise weit

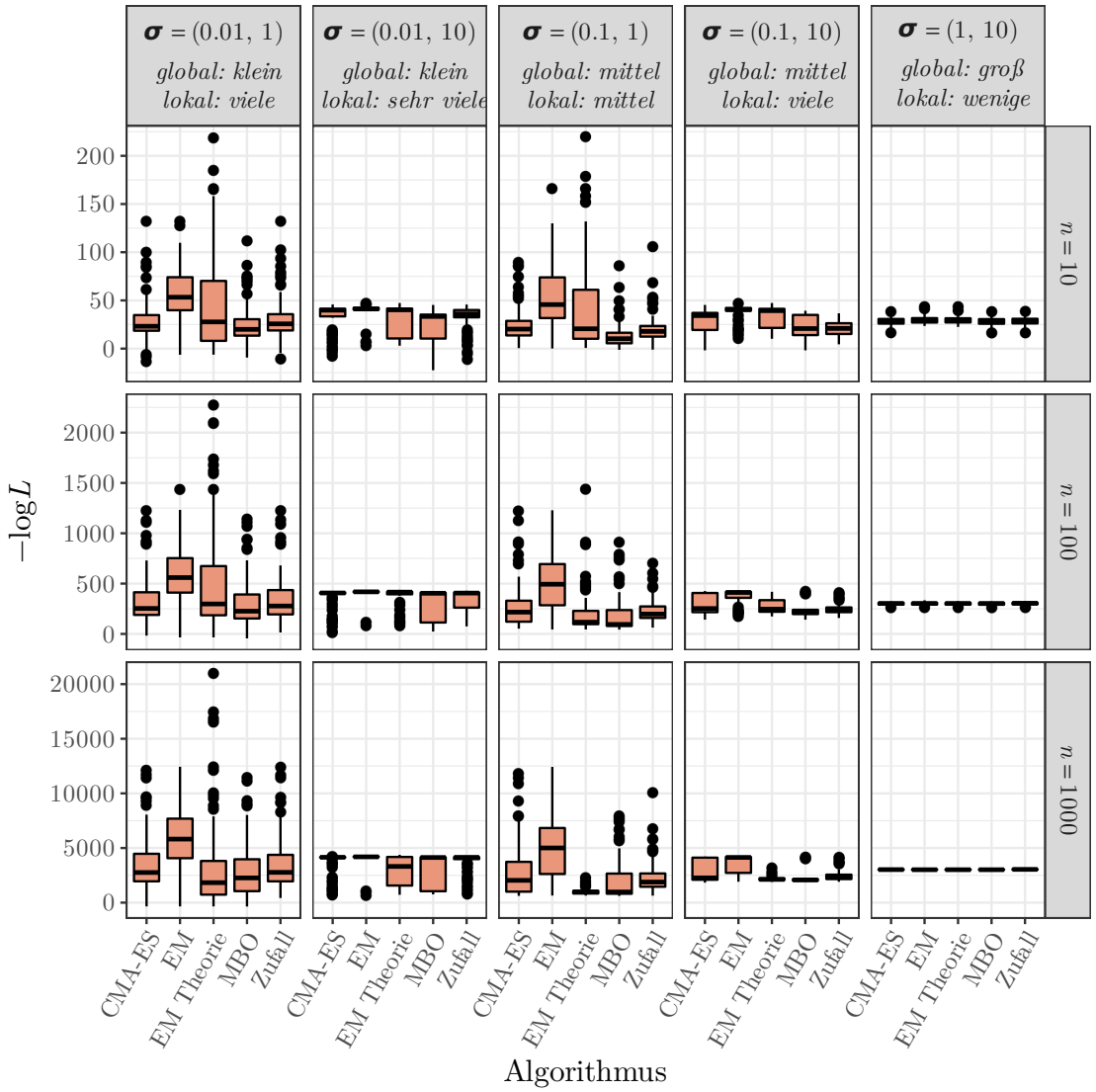

Abbildung 11: Erreichte Funktionswerte in 2D

vom theoretischen Wert entfernt. Interessant ist, dass EM und MBO häufig einen ähnlich geringen Abstand zu den theoretischen Optimalparametern erreichen, der erreichte Funktionswert von EM aber deutlich schlechter ist (vgl. Abbildung 11).

Insgesamt ist also zu erkennen, dass EM auf diesem Teilproblem nur funktioniert, wenn bereits in unmittelbarer Nähe des Optimums gestartet wird (*EM Theorie* für großes n). Die Vermutung liegt nahe, dass der Algorithmus durch die festen Parameter zu unflexibel ist, sich aus einem lokalen Optimum herauszubewegen, während die Vergleichsalgorithmen üblicherweise zumindest lokale Optima mit besseren Funktionswerten erreichen. Im Kapitel 3.5 wird auch eine Instanz dieses 2D-Optimierungsproblems noch etwas detaillierter betrachtet, im folgenden Unterkapitel soll aber zunächst in der Simulation zur Optimierung aller fünf Parameter der Likelihood überprüft werden, ob EM dort besser funktioniert.

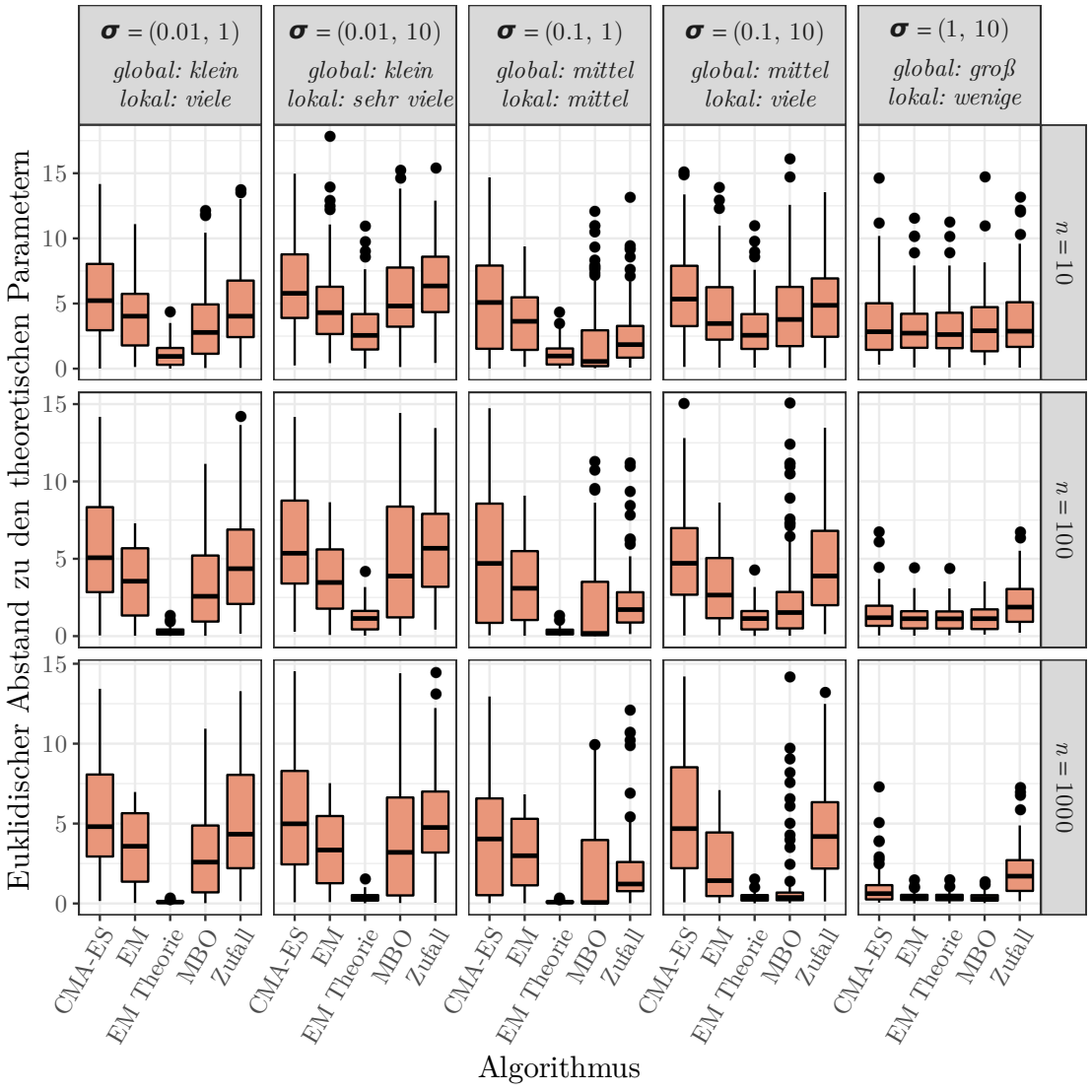

Abbildung 12: Euklidischer Abstand der Parameter in 2D

3.4 5D-Simulationsstudie

Bevor die Optimierungsergebnisse analog zur 2D-Simulation grafisch dargestellt werden, wird ein kurzer Überblick über fehlgeschlagenen Läufe des EM-Algorithmus gegeben. Im Gegensatz zur zweidimensionalen Optimierung, treten bei Optimierung aller Parameter der Mischung die in Abschnitt 3.2 besprochenen Fehler auf. In Tabelle 5 ist ein klarer Einfluss der Beobachtungsanzahl auf die Anzahl abgebrochener Optimierungsläufe zu erkennen: Für $n = 10$ liegt der Anteil bei ca. 15 % in den Fällen mit $\sigma_2 = 1$ und ca. 12 %, wenn $\sigma_2 = 10$ gilt. Für $n = 100$ bzw. $n = 1000$ sind es ca. 1.6 % bzw. 1.25 % (bei $\sigma_2 = 1$) und maximal 0.1 % (bei $\sigma_2 = 10$). Bei wenigen Beobachtungen kann es offensichtlich häufiger vorkommen, dass einzelne Beobachtungen separiert werden und zu einer unendlichen Likelihood führen, bei mehr Beobachtungen gibt es weniger große Lücken innerhalb der gegebenen Punkte. Die Unterscheidung für σ_2 ist ebenfalls plausibel: Die möglichen Mittelwerte der Komponenten liegen bei -5, 0 und 5. Bei sehr kleinem Mischungsanteil der Komponente mit kleinerer Varianz (σ_1) kann es vorkommen, dass dieser zufällig nur eine Beobachtung zugrunde liegt. Bei $\sigma_2 = 1$ ist die Wahrscheinlichkeit relativ hoch, dass diese eine Beobachtung separiert wird und zu einer unendlichen Likelihood führt. Für $\sigma_2 = 10$ hingegen ist zumindest für $n \in \{100, 1000\}$ wahrscheinlich, dass weitere Beobachtungen in der Nähe liegen und das Auftreten des Problems verhindern.

Darüber hinaus ist der Einfluss des Mischungsverhältnisses λ relevant. Denn insbesondere für $\lambda = 0.1$ und $\lambda = 0.9$ ist anzunehmen, dass das Problem der Separierung einer einzelnen Beobachtung bei $n = 10$ häufiger auftritt als bei gleichmäßiger aufgeteilten Beobachtungen. In Tabelle 6 sind die Anteile abgebrochener Läufe aufgeteilt nach n und λ dargestellt. Für $n = 10$ bestätigt sich die Vermutung deutlich: Während der Anteil für Werte von λ zwischen 0.3 und 0.6 ca. 10 % bis 11.5 % beträgt, liegt dieser für $\lambda = 0.1$ bei 15 % und für Mischungsparameter von 0.8 bzw. 0.9 sogar bei ca. 19 % bzw. 23 %. Für kleines λ wird der Komponente mit der kleineren Varianz tendenziell häufiger nur eine Beobachtung zugeordnet, wodurch es vermehrt zum Problem der Separierung kommt. Dadurch erhöht sich

Tabelle 5: Anteil abgebrochener Optimierungsläufe bzgl. σ und n

	$\boldsymbol{\sigma} = (0.01, 1)$	$(0.01, 10)$	$(0.1, 1)$	$(0.1, 10)$	$(1, 10)$
$n = 10$	0.150	0.120	0.162	0.125	0.142
$n = 100$	0.016	0.001	0.016	0.001	0.001
$n = 1000$	0.013	0.000	0.012	0.000	0.000

die Anzahl der abgebrochenen Läufe für $\lambda \to 0$. Für $\lambda \to 1$ hingegen, werden der Komponente mit der kleineren Varianz viele Beobachtungen und der anderen Komponente erwartungsgemäß genauso häufig nur eine Beobachtung zugeordnet. Hier ist die Wahrscheinlichkeit, dass diese Beobachtung zufällig in der Nähe einer Beobachtung der anderen Komponente liegt, aber deutlich geringer, da alle anderen Beobachtungen aufgrund der kleinen Varianz konzentriert an einem Punkt liegen. Sollte dieser unwahrscheinliche Fall eintreten, scheint es zudem plausibel, dass die sehr dicht liegenden Beobachtungen dann einer Komponente zugeordnet werden und die eingangs beschriebene zweite Art des Optimierungsabbruchs auftritt (vgl. Kap. 3.2, S. 24 f.).

Im Fall, dass die Verteilung mit der großen Varianz aus nur einer Beobachtung besteht, kehrt sich zudem der Einfluss der größeren Varianz um: Wie in Tabelle 7 zu erkennen ist, treten für die Instanzen mit $\sigma_2 = 10$ kaum größere Probleme für $\lambda = 0.1$ im Vergleich zu $\lambda = 0.5$ auf, während die Optimierungsabbrüche für die größeren Werte von λ deutlich zunehmen. Im Gegensatz dazu nehmen der Anteil abgebrochener Läufe für $\sigma_2 = 1$ in Richtung beider Extreme von λ eher gleichmäßig zu.

Der in Tabelle 5 zu erkennende leichte Vorteil für $\sigma_2 = 10$ besteht also nur für kleine Werte von λ, dort allerdings sehr deutlich. Demgegenüber wirkt sich die größere Varianz der zweiten Komponente für großes λ negativ aus, da hier der mögliche Bereich, in dem eine einzelne Beobachtung aus dieser Komponente realisiert werden kann, deutlich größer ist als für $\sigma_2 = 1$ und dadurch die kaum streuenden Beobachtungen der ersten Komponente seltener genau getroffen werden. Dementsprechend tritt der Abbruch aufgrund einer einzelnen separierten Beobachtung hier deutlich häufiger auf. Insgesamt ergibt sich dadurch der leichte Vorteil für $\sigma_2 = 10$ in Tabelle 5, aber gleichzeitig auch der höhere Anteil für $\lambda \to 1$ in Tabelle 6.

Die Optimierungsergebnisse für die vier Optimierungsalgorithmen sind in Abbildung 13 dargestellt. Für $n = 10$ traten drei einzelne EM-Läufe mit extrem hohen Werten für $-\log L$ auf, die aus Gründen der Übersichtlichkeit nicht in der Grafik enthalten sind. Die Funktionswerte an den entsprechenden Stellen sind durchaus

Tabelle 6: Anteil abgebrochener Läufe bzgl. λ und n

	$\lambda = 0.1$	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$n = 10$	0.147	0.127	0.114	0.113	0.104	0.101	0.129	0.192	0.234
$n = 100$	0.006	0.008	0.008	0.007	0.006	0.007	0.007	0.007	0.009
$n = 1000$	0.006	0.005	0.005	0.005	0.004	0.006	0.005	0.005	0.006

Tabelle 7: Anteil abgebrochener Läufe für $n = 10$

	$\boldsymbol{\sigma} = (0.01, 1)$	$(0.01, 10)$	$(0.1, 1)$	$(0.1, 10)$	$(1, 10)$
$\lambda = 0.1$	0.213	0.104	0.216	0.103	0.099
$\lambda = 0.2$	0.174	0.096	0.172	0.092	0.100
$\lambda = 0.3$	0.133	0.104	0.136	0.099	0.099
$\lambda = 0.4$	0.122	0.109	0.116	0.103	0.113
$\lambda = 0.5$	0.109	0.099	0.108	0.096	0.111
$\lambda = 0.6$	0.103	0.088	0.118	0.090	0.104
$\lambda = 0.7$	0.122	0.116	0.140	0.117	0.151
$\lambda = 0.8$	0.176	0.168	0.204	0.182	0.231
$\lambda = 0.9$	0.199	0.199	0.252	0.247	0.271

plausibel. Warum die Optimierungsläufe dorthin konvergieren konnten, wurde aufgrund der zu vernachlässigenden Anzahl solcher Fälle nicht weiter untersucht.

Bei den Ergebnissen fällt auf, dass EM hier immer sehr nah an die Ergebnisse von *EM Theorie* herankommt. Die übrigen Algorithmen erreichen nur vereinzelt ähnlich gute Ergebnisse und schneiden im Mittel deutlich schlechter ab. Die einzige Ausnahme ist wieder die Varianzkombination $\boldsymbol{\sigma} = (1, 10)$, für die von allen Algorithmen ähnlich gute Ergebnisse erreicht werden, da die Funktion offensichtlich keine große Herausforderung für die Algorithmen darstellt. Auch für die verschiedenen Werte von n ergibt sich ein sehr ähnliches Gesamtbild.

In Abbildung 14 ist der Abstand zum theoretischen Optimum im Parameterraum dargestellt. Grundsätzlich zeigt sich ein ähnliches Bild wie bei den Funktionswerten. Der EM-Algorithmus erreicht überwiegend deutlich näher an den theoretischen Optimalparametern liegende Ergebnisse als die Vergleichsalgorithmen und ist auch hier in den meisten Fällen nur geringfügig schlechter als die EM-Variante mit Start am theoretischen Optimum. Auffällig ist jedoch, dass für die Varianzkombinationen $\boldsymbol{\sigma} = (0.01, 10)$, $\boldsymbol{\sigma} = (0.1, 10)$ und $\boldsymbol{\sigma} = (1, 10)$ die deutlich am weitesten entfernten Ergebnisse (eukl. Abstand > 20) auch allesamt von EM erreicht werden. Hinzu kommt, dass für 10 Beobachtungen bei diesen 3 Varianzkombinationen auch im Mittel die am weitesten entfernten Ergebnisse erreicht werden, obwohl die zugehörigen Funktionswerte in Abbildung 13 deutlich besser als die der drei Vergleichsalgorithmen sind. Dass weiter entfernt ein besserer Funktionswert gefunden werden kann, ist auf die Ungenauigkeit des theoretischen Optimums bei so wenigen Beobachtungen zurückzuführen. Dass die übrigen Algorithmen allerdings grundsätzlich näher

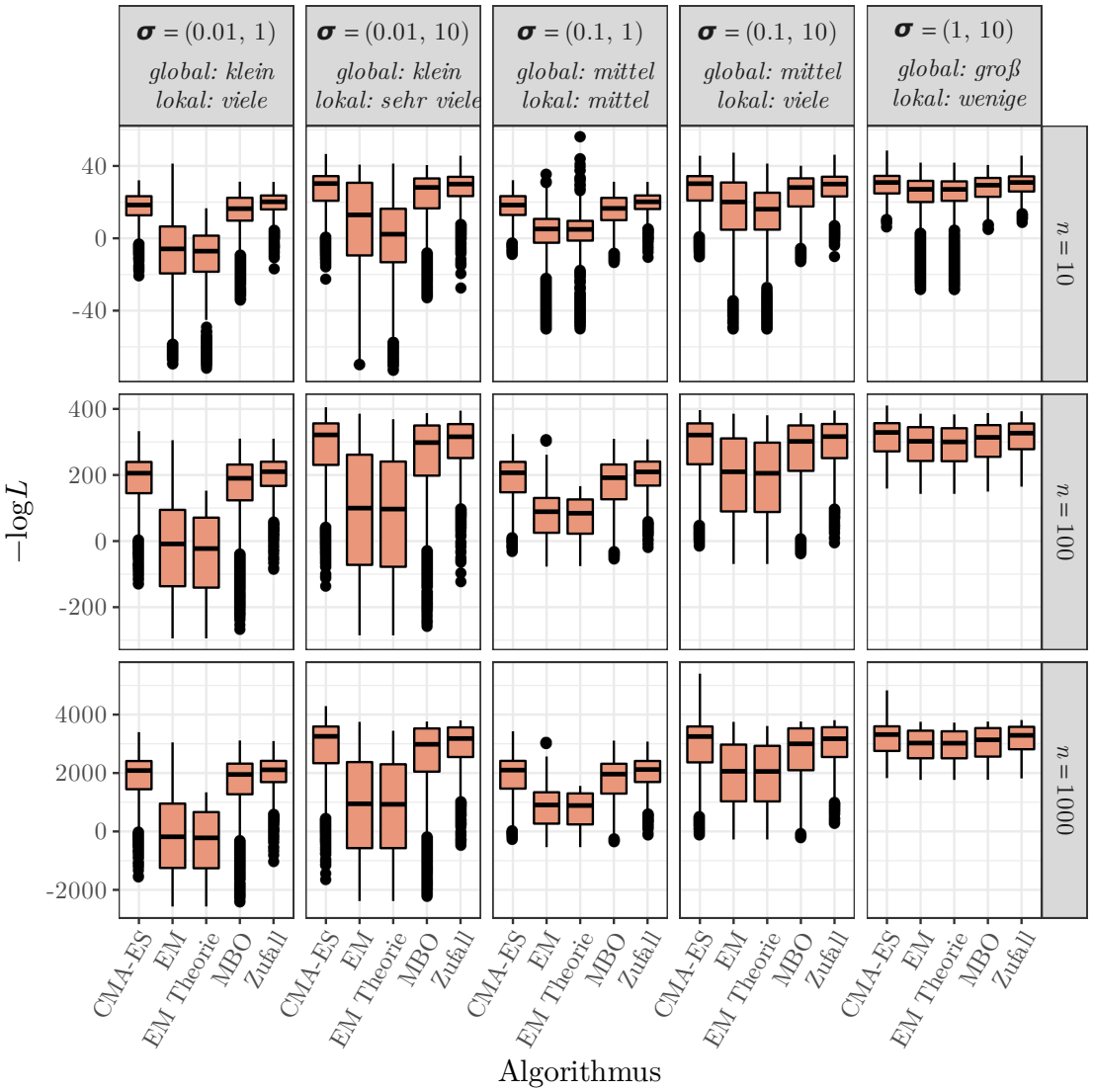

Abbildung 13: Erreichte Funktionswerte in 5D

an die Parameter des theoretischen Optimums herankommen, ist dadurch nicht zu erklären.

Da die Parameter Werte auf unterschiedlichen Skalen annehmen (z. B. $\mu_1, \mu_2 \in [-5, 5]$ und $\lambda \in (0, 1)$), wird als Alternative zu euklidischen Distanz die Gower-Distanz betrachtet, welche auf dem Gower-Unähnlichkeitsmaß (Gower, 1971) basiert. Dabei wird die relative Abweichung bezüglich des vorhandenen Wertebereiches in allen Dimensionen einzeln bestimmt und zu einem Distanzwert gemittelt. Für zwei Vektoren $\boldsymbol{y}, \boldsymbol{z} \in M \subset \mathbb{R}^p$, mit Spannweiten r_i der einzelnen Dimensionen aller Vektoren aus M ist die Gower-Distanz gegeben durch:

$$d_g(\boldsymbol{y}, \boldsymbol{z}) = \frac{1}{p} \sum_{i=1}^{p} \frac{|y_i - z_i|}{r_i}.$$

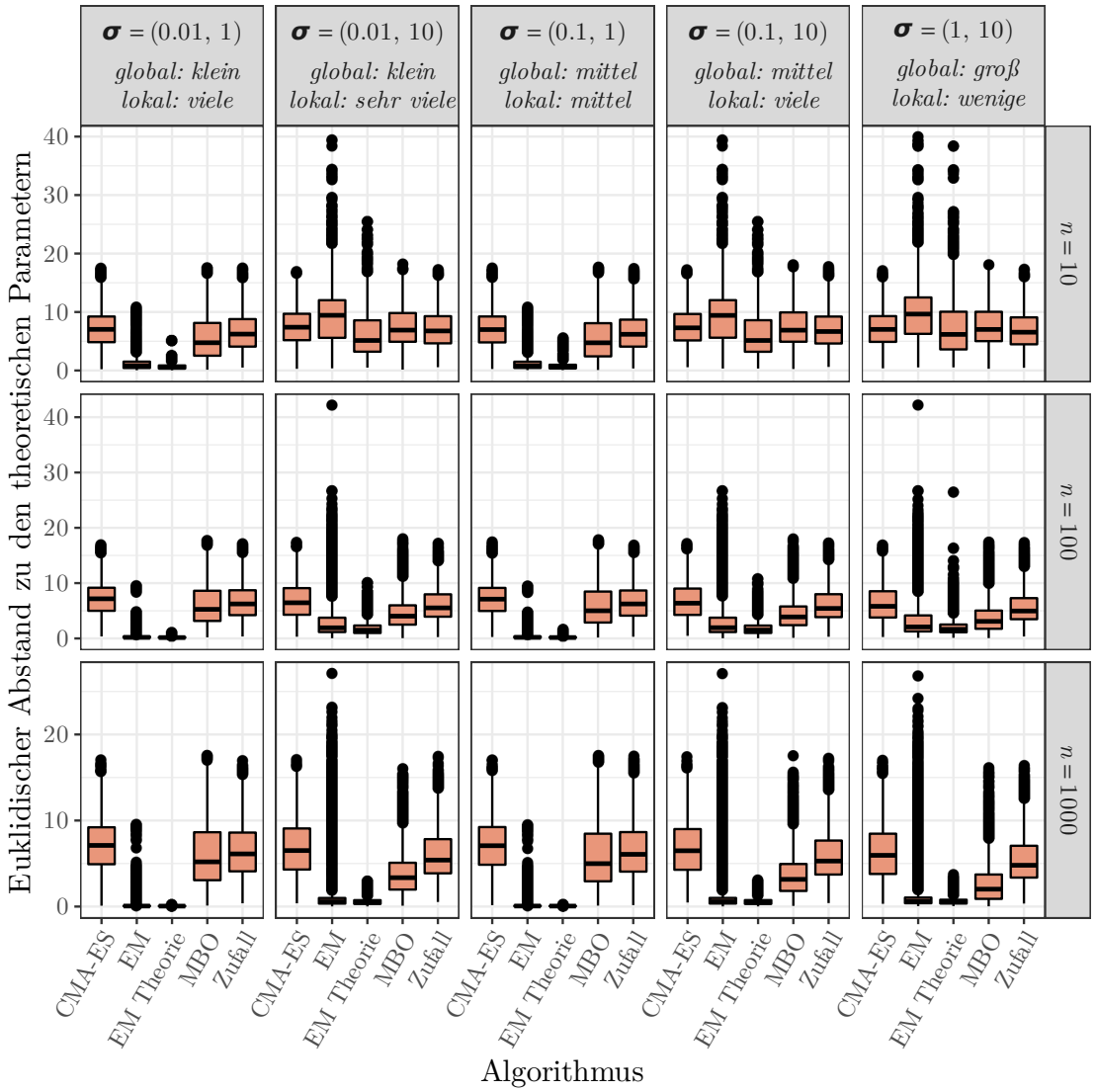

Abbildung 14: Euklidischer Abstand der Parameter in 5D

Für die Gower-Distanz zum theoretischen Optimum in Abbildung 15 ergeben sich gegenüber der euklidischen Distanz keine deutlichen Unterschiede im Verhältnis der Algorithmen untereinander. Allerdings fällt auf, dass die Struktur der Ausreißer bei beiden EM-Varianten zeilenweise über alle Varianzkombinationen deutlich einheitlicher ausfällt als in Abbildung 14, wo ein deutlicher Unterschied zwischen den Kombinationen mit $\sigma_2 = 1$ und $\sigma_2 = 10$ besteht. Daher wird im weiteren Verlauf der Analyse die Gower-Distanz verwendet.

Um den Zusammenhang zwischen erreichtem Funktionswert und Abstand zum theoretischen Optimum weiter zu analysieren, werden die beiden Größen in Streudiagrammen gegeneinander aufgetragen. In Abbildung 16 sind die Streudiagramme von $-\log L$ und Gower-Distanz getrennt nach Algorithmen und n dargestellt. Im Gegensatz zu den vorherigen Grafiken findet hier zunächst keine Unterscheidung

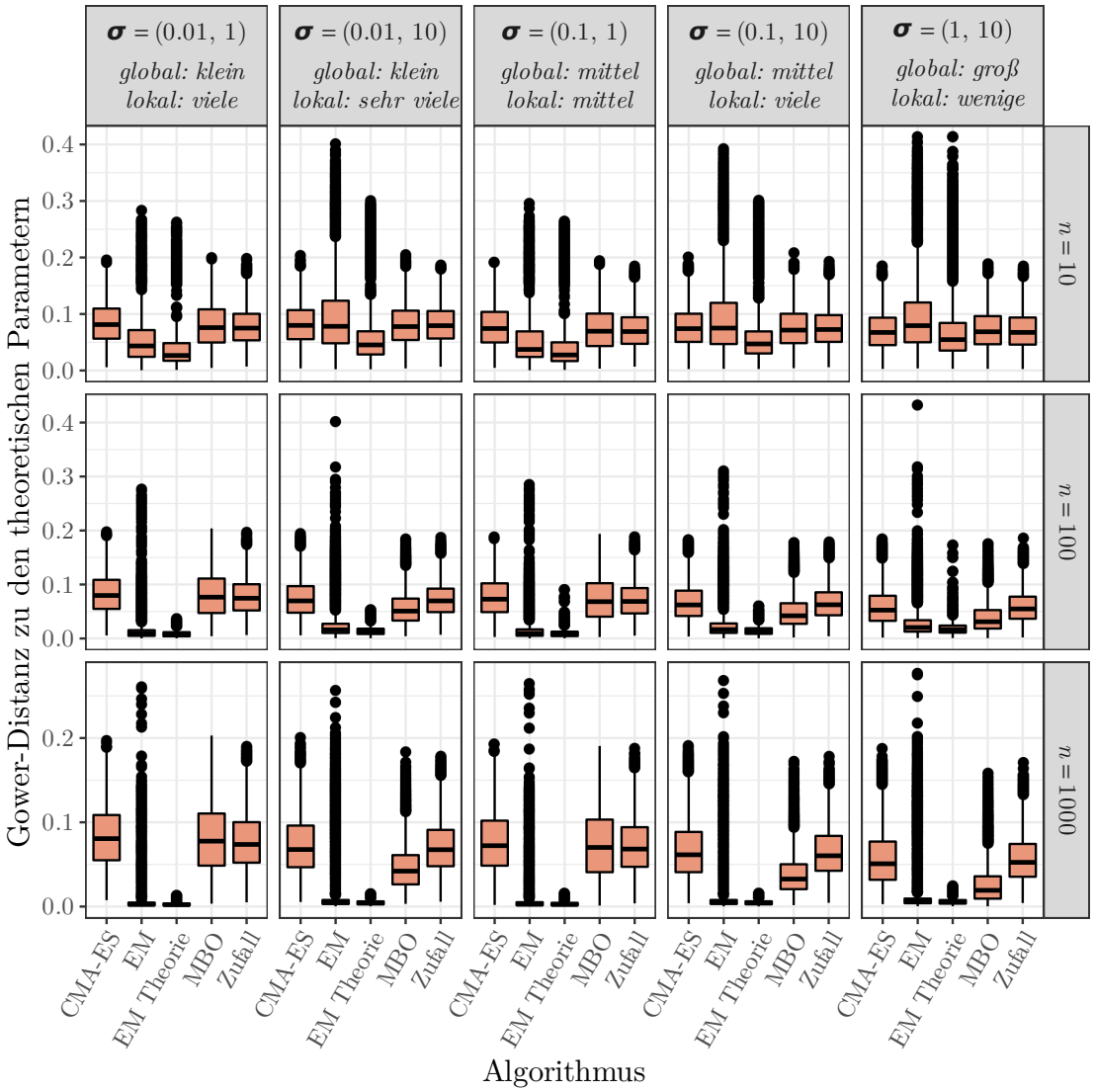

Abbildung 15: Gower-Distanz der Parameter in 5D

bezüglich der Varianzkombinationen statt. Es ist zu erkennen, dass bei den Algorithmen CMA-ES, MBO und der Zufallssuche sehr dicht besetzte Punktewolken entstehen. Für den EM-Algorithmus (und auch die Variante mit theoretischem Startwert) sind hingegen deutlichere Strukturen zu erkennen. Für $n = 10$ existieren neben einer zu den anderen Algorithmen vergleichbaren Punktewolke, mehrere kleinere Bereiche, in denen sich die Ergebnisse der Optimierungsläufe häufen. Diese Optima verfügen überwiegend über bessere Likelihoodwerte, liegen aber zum Teil auch weiter entfernt von den theoretischen Parametern. Für EM mit theoretischen Startwerten liegen die Ergebnisse mit den besten Funktionswerten ausschließlich etwas weiter entfernt. Wie bereits erwähnt, ist die Bestimmung der Parameter auf Basis von 10 Beobachtungen sehr ungenau, weswegen das Optimum der Zielfunktion mitunter deutlich von den theoretischen Optimalwerten abweichen kann. Für

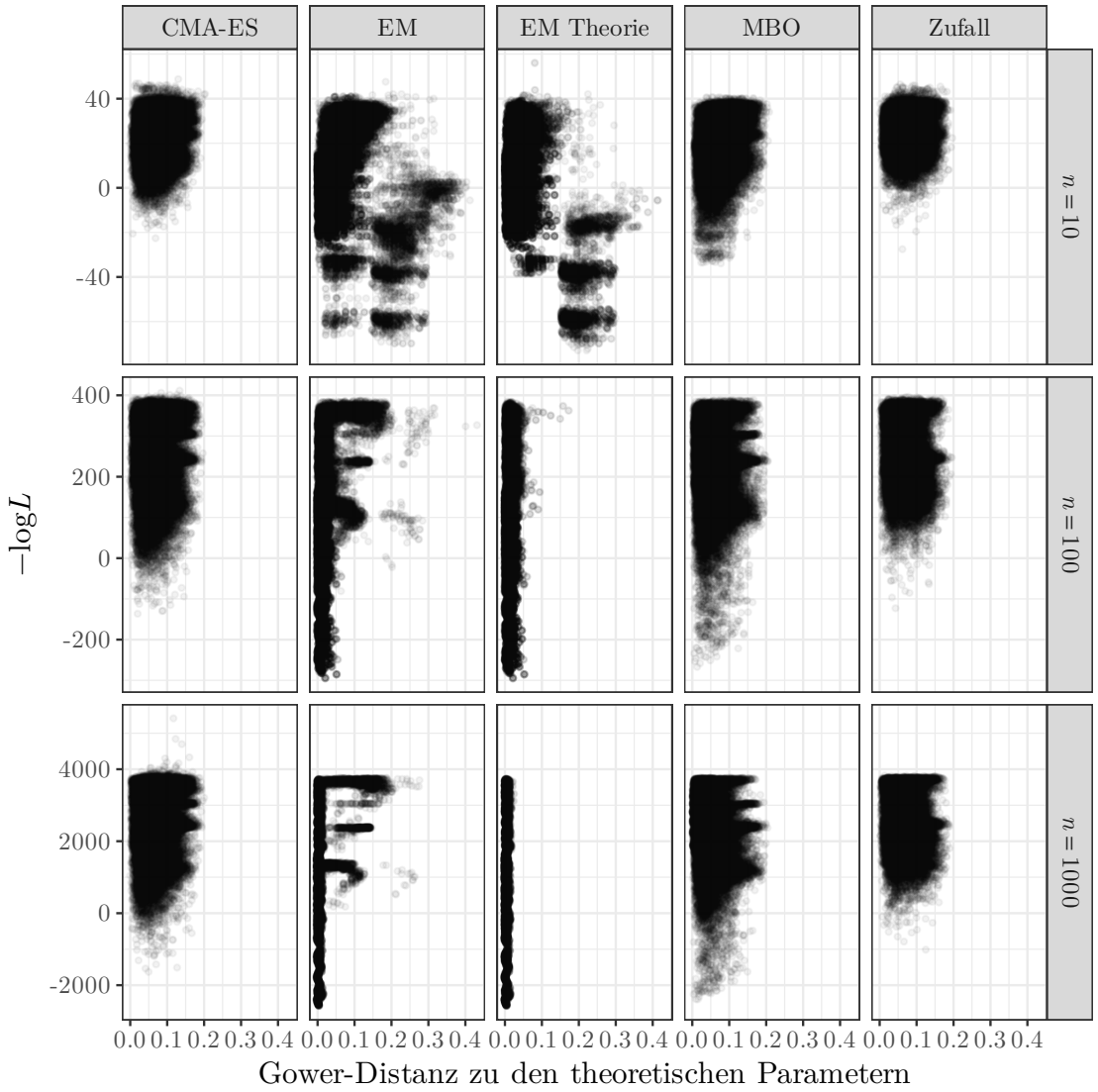

Abbildung 16: Funktionswert und Gower-Distanz 5D

$n = 100$ und $n = 1000$ ist dieser Effekt dementsprechend nicht mehr zu beobachten. Die erreichten Werte des EM-Algorithmus liegen fast ausschließlich sehr nah an den theoretischen Parameterwerten. Lediglich für bestimmte Likelihoodwerte existieren Ergebnisse, die weiter entfernt von den theoretischen Parametern liegen. Für $n = 1000$ gibt es drei solche Bereiche in etwa bei den negativen log-Likelihoodwerten 1250, 2500 und 3750. Diese Häufungen sind auch für die übrigen Algorithmen zu erahnen, durch die breite Streuung innerhalb der Punktewolken allerdings nicht deutlich zu erkennen. Die EM-Ergebnisse, die bessere Funktionswerte als die übrigen Algorithmen liefern, besitzen ausschließlich Parameterwerte nah an den theoretischen Werten.

Für weitere Klarheit bezüglich der weiter entfernt liegenden Ergebnisse sorgt die Darstellung in Abbildung 17. Hier sind nur die Ergebnisse des EM-Algorithmus

dargestellt, wodurch die verschiedenen Varianzkombinationen einzeln betrachtet werden können. Hier ist für $n = 100$ und $n = 1000$ zu erkennen, dass die weiter entfernten lokalen Optima bei den einzelnen Varianzkombinationen nur an einer Stelle gehäuft auftreten. Der Wert entspricht immer dem in unmittelbarer Nähe der theoretischen Werte gefundenen Optimum mit dem schlechtesten Funktionswert. Für $n = 1000$ liegt dieser Wert bei ca. 1750, wenn die größere der beiden Standardabweichungen 1 ist und bei ca. 3750, wenn diese 10 beträgt. Im ersten Fall existieren zusätzlich einige Ausreißer mit größerem Funktionswert bei ca. 2500. Für $n = 100$ ist ein ähnliches Bild, allerdings mit größerer Ungenauigkeit, zu beobachten, während die lokalen Optima bei $n = 10$ für die einzelnen Varianzkombinationen nicht mehr so deutlich zu erkennen sind, wie es in der gemeinsamen Darstellung den Anschein hatte. Allerdings werden in allen fünf Fällen die besten Funktionswerte jeweils einmal sehr nah an den theoretischen Funktionswerten erreicht und einmal auch in einem lokalen Optimum, welches etwas weiter entfernt liegt. Im Bereich zwischen diesen beiden Optima und der großen Punktewolke liegen zahlreiche verstreute Ergebnisse, die erneut auf die Ungenauigkeit bei $n = 10$ zurückzuführen sind. Hingegen sind für $n = 1000$ sogar innerhalb der Werte mit kleiner Gower-Distanz zu den theoretischen Parametern deutlich einzelne Häufungen zu erkennen.

Im nächsten Schritt soll das Auftreten dieser verschiedenen Optima weiter erklärt werden. Dazu werden die drei Grafiken mit der Varianzkombination $\boldsymbol{\sigma} = (0.01, 10)$ aus Abbildung 17 bezüglich der zugrunde liegenden Mischungsproportion λ aufgeteilt. Insbesondere für $n = 1000$ ist in Abbildung 18 zu sehen, dass zu jedem Wert von λ genau eines der lokalen Optima gehört. Mit steigendem λ, also mit erhöhtem Anteil von Beobachtungen in der Verteilung mit kleinerer Varianz, wird der Funktionswert am Optimum geringer und auch die Zahl der Optimierungsläufe, die einen schlechteren Funktionswert erreichen, nimmt deutlich ab. Stammen nur wenige Beobachtungen aus der Verteilung mit der kleineren Varianz (z. B. bei $\lambda = 0.1$), erreichen viele Läufe das Optimum nicht. Der erreichte Funktionswert unterscheidet sich in diesen Fällen allerdings auch deutlich weniger vom Optimum als für größeres λ.

Im Fall $n = 100$ ist das gleiche Gesamtbild zu erkennen, der einzige Unterschied ist die etwas größere Streuung aufgrund der geringeren Anzahl zur Verfügung stehender Beobachtungen. Daher sind die einzelnen Optima in Abbildung 17 auch nicht so deutlich zu erkennen wie für $n = 1000$. Auch für $n = 10$ wirkt sich der Parameter λ entsprechend aus, allerdings werden die Ergebnisse hier stark von der Unsicherheit der kleinen Stichprobe überlagert. So sind die Optima auch für einzelne Werte von λ nicht klar von den Punkten der übrigen Läufe abzugrenzen. Mit steigendem λ ist dies jedoch etwas besser möglich. Allerdings ist auch offensichtlich, dass ein

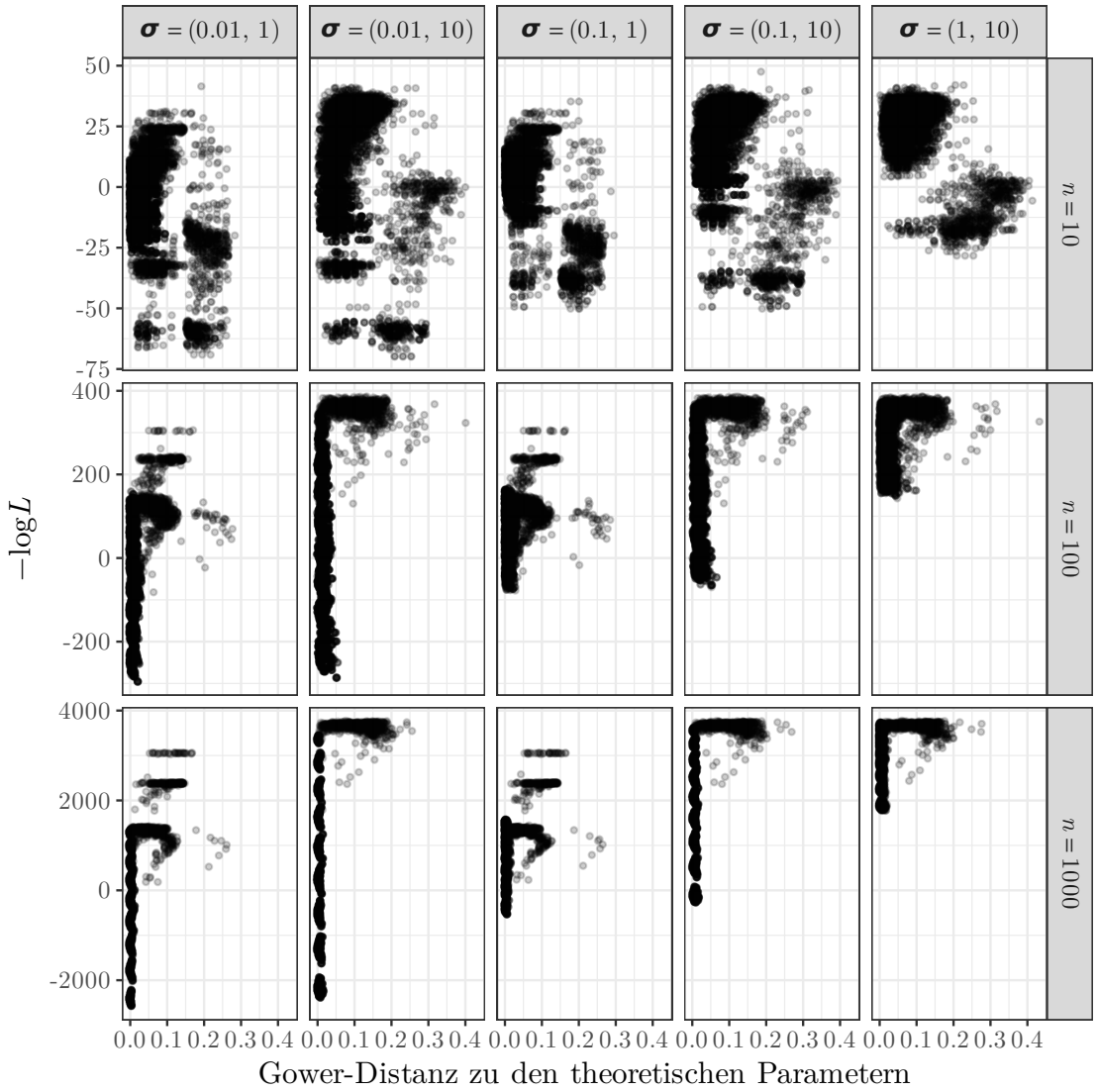

Abbildung 17: Funktionswert und Gower-Distanz bei EM

deutlich größerer Teil der Läufe als bei den größeren Beobachtungsanzahlen das jeweilige Optimum nicht erreicht.

Für die übrigen Algorithmen sind vergleichbare Strukturen bezüglich erreichtem Funktionswert und Parameter-Abstand nicht zu erkennen. Die Werte der Läufe bilden jeweils eine relativ homogene Punktewolke. Abbildung 19 enthält zusätzlich zur Darstellung der EM-Läufe aus Abbildung 18 die Ergebnisse von MBO, CMA-ES und der Zufallssuche. Die bereits erwähnten Punktewolken dieser drei Algorithmen überdecken sich untereinander größtenteils, während der Bereich der guten EM-Läufe deutlich heraussticht. Lediglich MBO reicht mit einer nennenswerten Anzahl von Läufen an diesen heran, CMA-ES und der Zufallssuche gelingt dies nur sehr vereinzelt. Auch bezüglich der Distanz zum Optimum bestätigen sich die Eindrücke aus den vorherigen Darstellungen: MBO landet tendenziell etwas näher an den

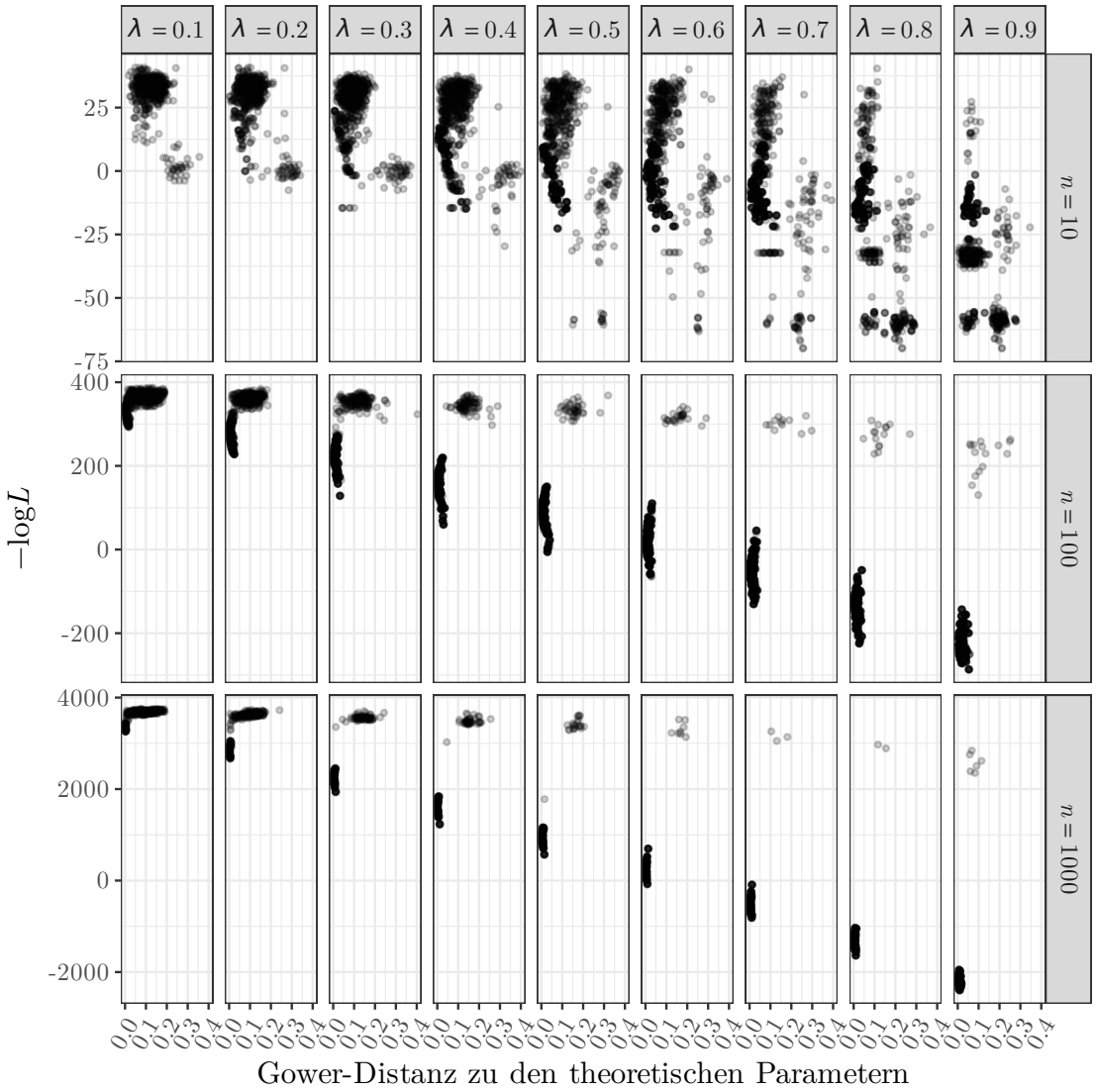

Abbildung 18: EM mit $\boldsymbol{\sigma} = (0.01, 10)$

theoretischen Werten als die beiden anderen Vergleichsalgorithmen, insbesondere bei größeren Werten des Mischunsgsverhältnisses λ. Für CMA-ES fällt auf, dass bei ungefähr gleichen Beobachtungszahlen die Läufe mit schlechten Funktionswerten sogar etwas weiter entfernt von den theoretischen Optimalparametern enden als bei der Zufallssuche. Insgesamt ist jedoch auch bezüglich der Distanz das Cluster der erfolgreichen EM-Läufe deutlich im Vorteil gegenüber den Punktewolken der anderen Algorithmen. Das gilt auch für den daraus hervortretenden Teil der erfolgreicheren MBO-Läufe. Es ist also erkennbar, dass keiner der Vergleichsalgorithmen ähnlich systematisch wie EM gute Optimierungsergebnisse erzielt.

Bisher nicht betrachtet wurde der Einfluss der Mittelwertparameter μ_1 und μ_2, welche in insgesamt neun verschiedenen Kombinationen in die Simulation eingegangen sind. In Abbildung 20 ist beispielhaft für den EM-Algorithmus mit $n = 1000$,

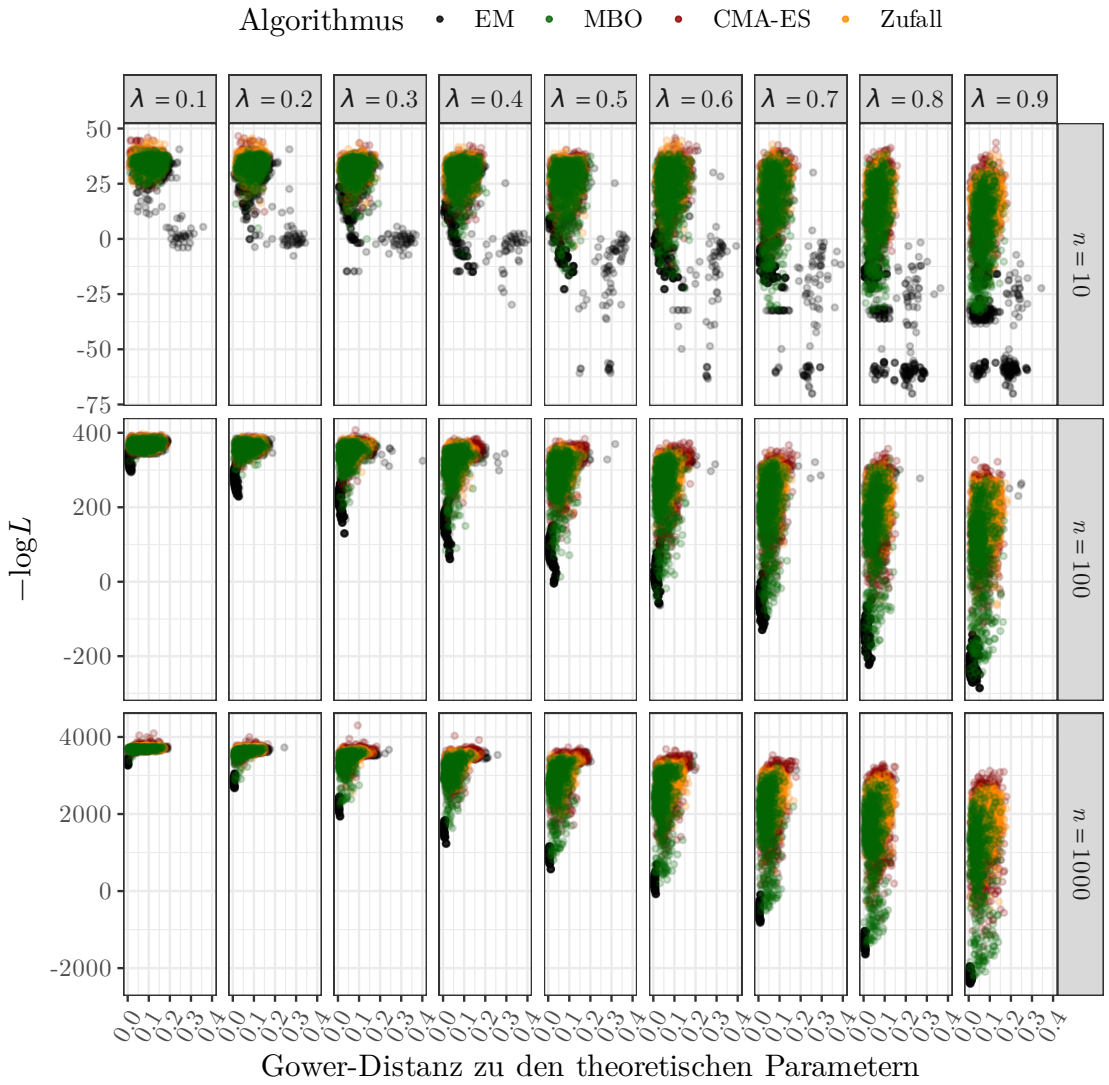

Abbildung 19: Vergleich der Algorithmen für mit $\boldsymbol{\sigma} = (0.01, 10)$

$\boldsymbol{\sigma} = (0.01, 10)$ und $\lambda = 0.5$ dargestellt, dass die Lageparameter der Verteilungen keinen Einfluss auf den erreichten Funktionswert, bzw. den Abstand zu den theoretischen Optimalparametern haben. In allen neun Fällen liegt die gleich geformte Punktewolke am Optimum bei einem Funktionswert von ca. 1000 und mit einer Gower-Distanz von ca. 0.01 von den theoretischen Werten entfernt. Vereinzelt existieren Läufe, die in beiden Kriterien deutlich schlechtere Ergebnisse liefern.

Abschließend soll der Einfluss des Startwerts für EM in dieser Simulationsstudie betrachtet werden. Wie in Kapitel 3.1 beschrieben, werden die bisher betrachteten EM-Läufe der Simulationsstudie immer an dem Designpunkt des MBO-Startdesigns gestartet, der den besten Funktionswert aufweist. Die Intention dabei ist, dem Algorithmus die relevante Information aus dem Startdesign zur Verfügung zu stellen, um ihn nicht gegenüber den anderen Algorithmen zu benachteiligen. Um

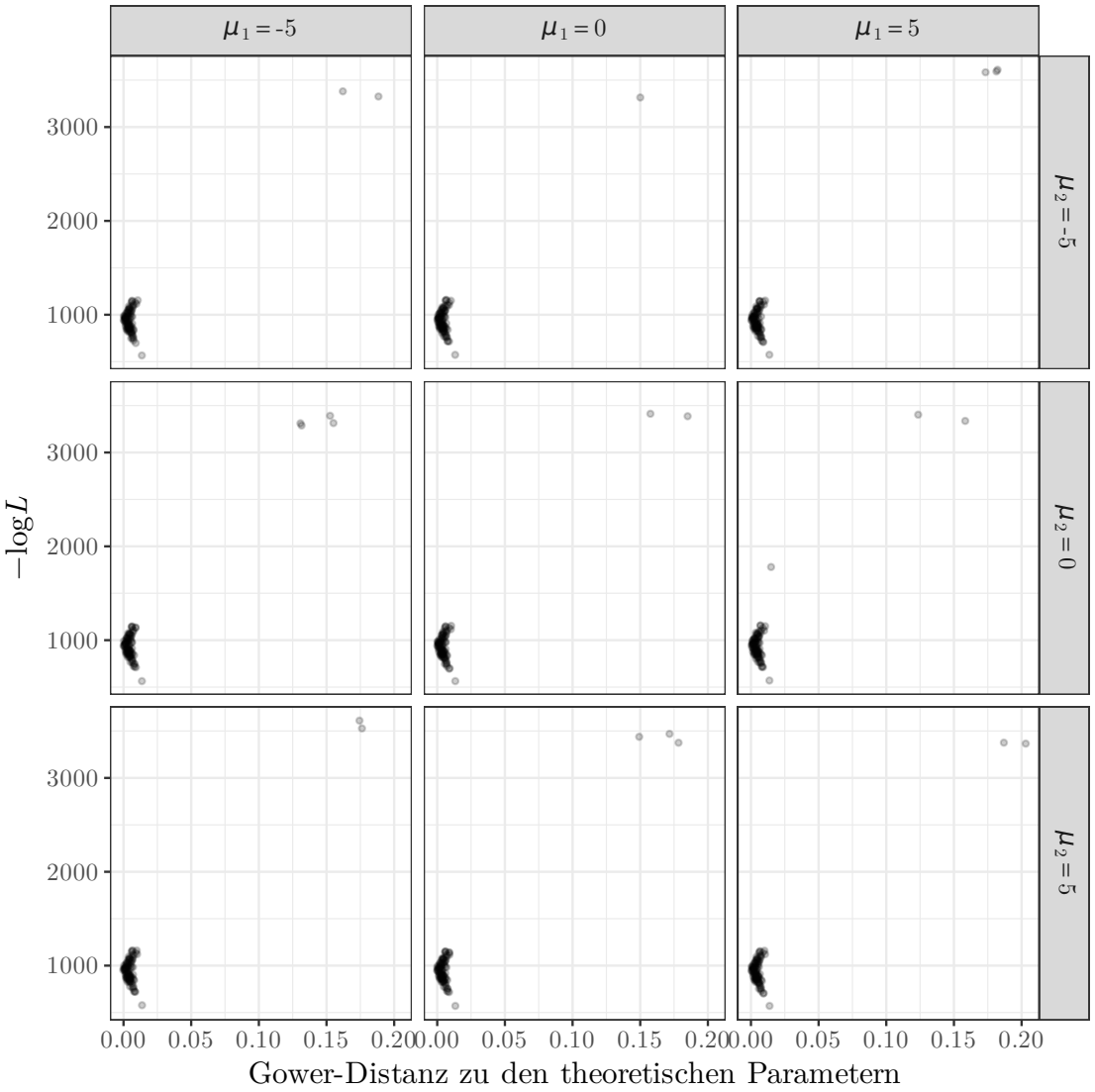

Abbildung 20: EM mit $\boldsymbol{\sigma} = (0.01, 10)$, $\lambda = 0.5$ und $n = 1000$

auszuschließen, dass die festgestellten Vorteile des EM-Algorithmus nicht in erster Linie auf diese Wahl zurückzuführen sind, werden in Abbildung 21 zusätzliche EM-Läufe mit Start an einem zufällig gleichverteilt aus den Punkten des Startdesigns gezogenen Startpunkt *EM Zufall* betrachtet. Aus Gründen der Übersichtlichkeit ist neben den beiden bereits bekannten EM-Varianten nur noch die Zufallssuche enthalten, gegenüber der die beiden anderen Varianten kaum Vorteile aufweisen konnten. Dem EM-Algorithmus gelingt dies offensichtlich auch ohne vom besten Punkt des Startdesigns aus zu starten. Alles in allem sind die erreichten Funktionswerte von *EM Zufall* zwar leicht schlechter als die der EM-Variante mit bestem Startpunkt, aber trotzdem in allen Fällen *EM Theorie* deutlich ähnlicher als den Ergebnissen der Zufallssuche. Damit bleibt die festgestellte Überlegenheit gegenüber CMA-ES und MBO auch bei zufälligem Startwert nahezu ungemindert

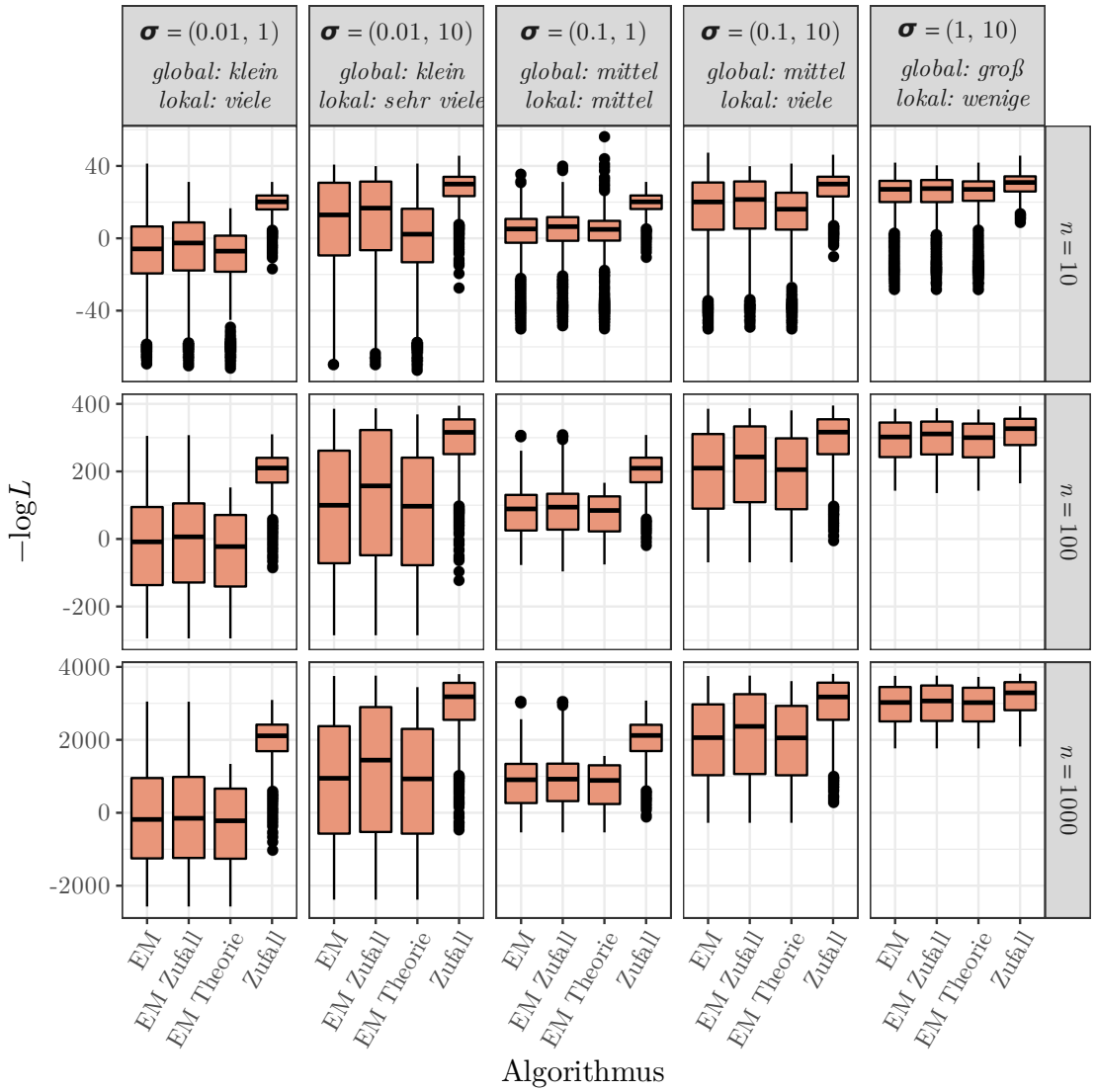

Abbildung 21: Erreichte Funktionswerte in 5D

bestehen, sodass durch Wahl des Startpunktes in der ursprünglichen Simulation keine maßgebliche Bevorteilung des EM-Algorithmus stattgefunden hat.

Insgesamt liefert der EM-Algorithmus in 5 Dimensionen also bessere Ergebnisse als die übrigen Algorithmen. Im folgenden Abschnitt werden die Optimierungsverläufe analysiert, um mögliche Erklärungen dafür zu erhalten.

3.5 Analyse der Optimierungsverläufe

Im Folgenden wird eine Likelihood-Funktionsinstanz mit den theoretischen Parametern $\boldsymbol{\mu} = (-5, 5)$, $\boldsymbol{\sigma} = (0.1, 10)$ und $\lambda = 0.5$ basierend auf 100 Beobachtungen betrachtet. Diese Funktion wird sowohl als fünfdimensionales Optimierungsproblem, als auch als zweidimensionales Optimierungsproblem (σ_1, σ_2 und λ fest)

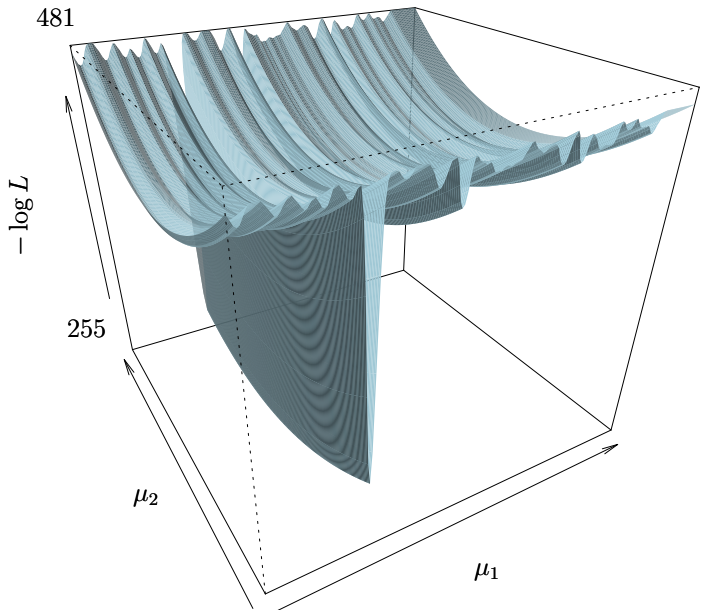

Abbildung 22: Likelihood-Funktion in 2D

betrachtet. In Abbildung 22 ist die entsprechende Funktionsinstanz im Raum der Mittelwerte analog zu den Grafiken aus Kapitel 2.2 dargestellt. Im Folgenden wird der Zielfunktionswert in Graustufen in einer zweidimensionalen Darstellung von μ_1 und μ_2 hinzugefügt, um zusätzlich den Optimierungsverlauf einzeichnen zu können. In Abbildung 23 sind mehrere Grafiken zum zweidimensionalen Fall enthalten. Oben ist die veränderte Darstellung der Funktion enthalten, in der die hier vertikal verlaufenden Optima weiterhin erkennbar sind (vgl. Abb. 22). Die Beschriftung im oberen Teil beinhaltet die Parameterwerte am Start und am theoretischen Optimum sowie die erreichten Ergebnisse von CMA-ES und EM. Die Reihenfolge der enthaltenen Parameter ist in der ersten Zeile angegeben. Zusätzlich sind die Optimierungsläufe für das 2D-Problem in der Grafik eingezeichnet. Ausgehend von einem zufällig gewählten Startpunkt ist die direkte Verbindung zum Optimierungsergebnis des jeweiligen Algorithmus als Linie dargestellt. Es ist zu erkennen, dass CMA-ES in diesem Beispiel nicht gut funktioniert, da das Ergebnis sehr nahe am Startwert und sehr weit entfernt vom theoretischen globalen Optimum liegt. Der EM-Algorithmus erreicht zwar ein Ergebnis, welches in Richtung des globalen Optimums liegt, es jedoch nicht erreicht. Es ist zusätzlich der zurückgelegte Weg über die einzelnen EM-Iterationen eingezeichnet, der sich in diesem Beispiel allerdings nur geringfügig unterscheidet, da die erste Iteration schon sehr nahe am späteren Ergebnis liegt.

In der unten links enthaltenen Grafik in Abbildung 23 sind die Funktionswerte entlang der direkten Verbindungen und dem Weg über die EM Iterationen aus

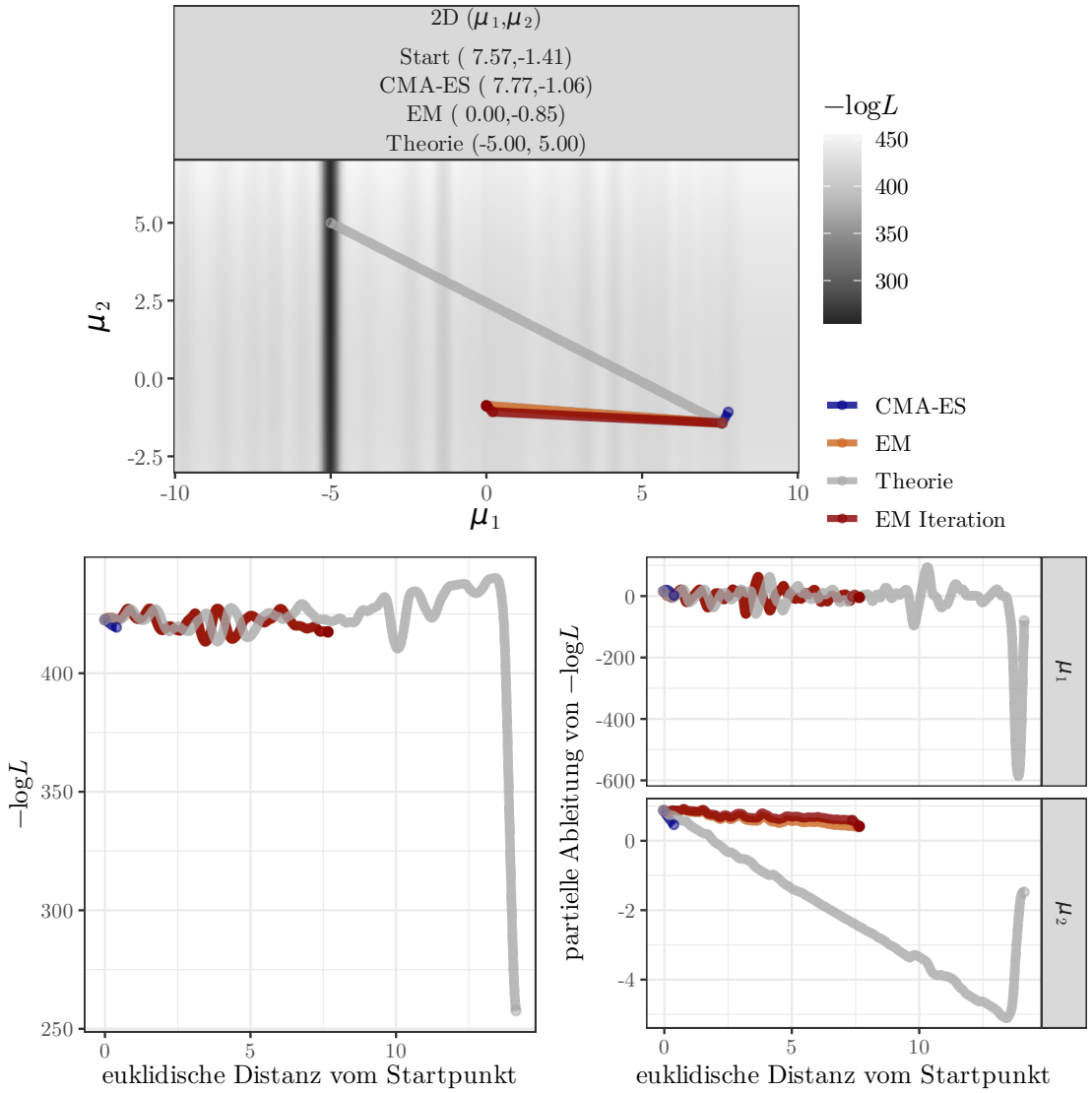

Abbildung 23: Grafische Darstellung der Optimierungspfade im 2D-Fall

der oberen Grafik dargestellt. Es ist zu erkennen, dass sich der direkte Pfad zum EM-Ergebnis und der Pfad entlang der Iterationen kaum unterscheiden, obwohl sie zunächst in eine leicht versetzte Richtung verlaufen. Die zurückgelegte Distanz über die einzelnen Iterationen ist insgesamt etwas länger, kann aber auf dem letzten Stück den Funktionswert auch nur minimal verringern. Anhand des Theorie-Pfades kann man erkennen, dass noch weitere lokale Minima in μ_1-Richtung zu überwinden sind, bis das globale Optimum erreicht wird. Unten rechts in Abbildung 23 sind die partiellen Ableitungen für die beiden Parameter dargestellt. Es ist zu erkennen, dass der EM-Algorithmus nicht exakt in einem Kandidatenpunkt für ein lokales Optimum terminiert. Auch das theoretische Optimum stimmt nicht exakt mit dem empirischen Optimum auf dieser Funktionsinstanz überein, da die partiellen Ableitungen im Endpunkt nicht exakt 0 sind. Es bestätigt sich also die Vermutung,

dass der auf das zweidimensionale Problem adaptierte EM-Algorithmus nicht wie erwartet funktioniert.

Im fünfdimensionalen Beispiel ist ein deutlicher Unterschied zwischen der direkten Verbindung zum Endpunkt des EM-Laufs und dem detaillierten Iterationsverlauf zu erkennen. In Abbildung 24 sind die Grafiken analog zum zweidimensionalen Fall enthalten. Für die obere Abbildung ist hier zu beachten, dass ausschließlich der (bezüglich der Multimodalität relevante) Raum der Mittelwertparameter dargestellt wird, obwohl mit den Varianzen und dem Mischungsverhältnis in diesem Fall drei weitere Parameter optimiert werden. Die der Darstellung zugrunde liegenden Werte der drei Parameter entsprechen dem Ergebnis des Optimierungslaufes. Darüber hinaus ist anzumerken, dass es hier zu einer Vertauschung der beiden Verteilungskomponenten im Verlauf der Optimierung kommt. Das ist grundsätzlich

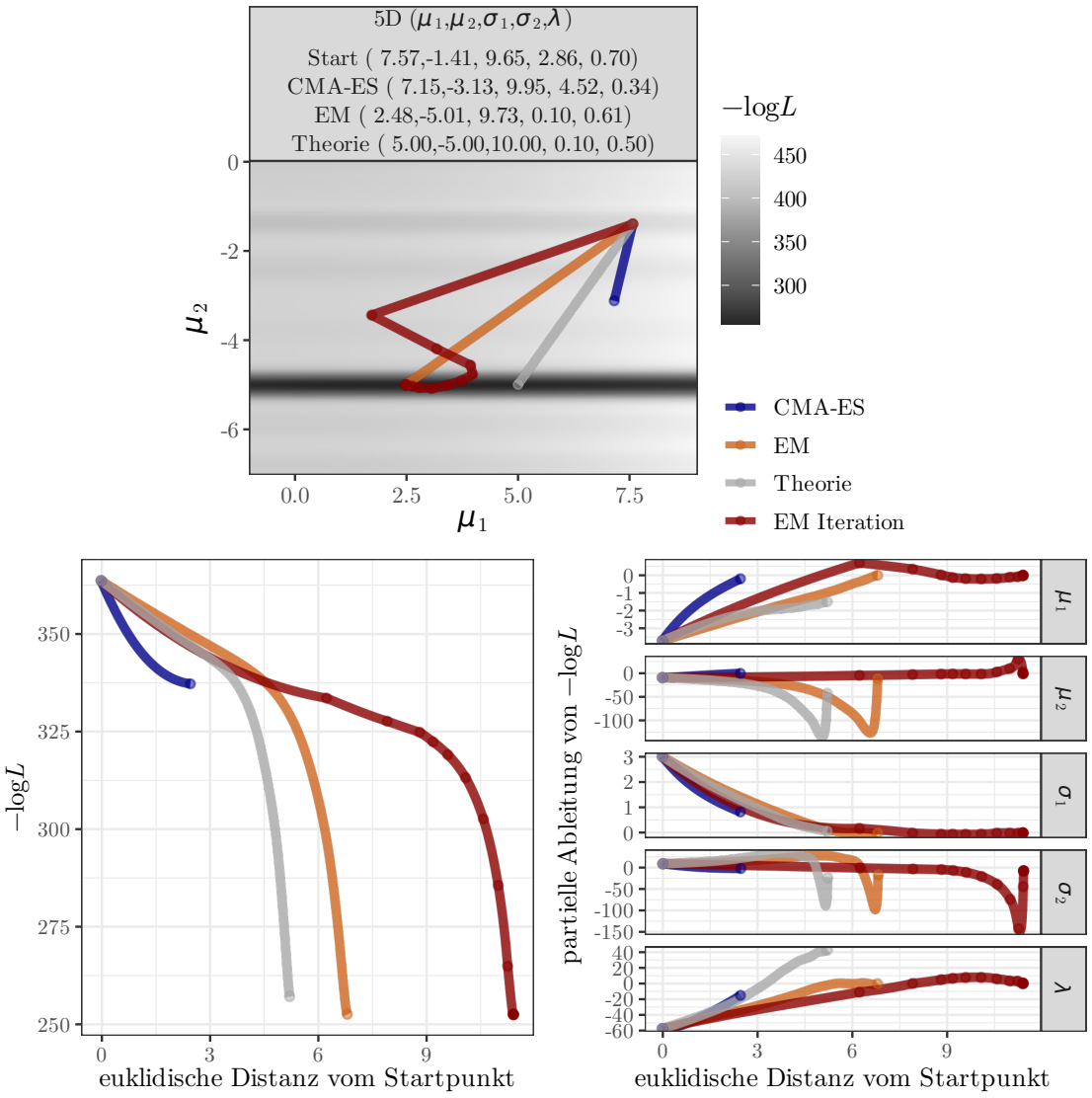

Abbildung 24: Grafische Darstellung der Optimierungspfade im 5D-Fall

möglich, da es im fünfdimensionalen Fall keine festen Parameter mehr gibt und die Bezeichnung der beiden Komponenten somit nicht fest vorgegeben ist. Dementsprechend ist im Ergebnis dieses Optimierungsbeispiels $\sigma_2 < \sigma_1$. Für das theoretische Optimum wurde die Bezeichnung der Komponenten hier ebenfalls getauscht, sodass die theoretische Lösung passend zum EM-Ergebnis dargestellt wird.

In der oberen Grafik (Abb. 24) ist zu erkennen, dass die erste EM-Iteration deutlich abseits der Geraden zwischen Start und Zielpunkt liegt. Von dort verläuft der Pfad wieder auf die direkte Verbindungsgerade zu und schneidet sie im Laufe der dritten Iteration. Danach wird erstmals der Bereich mit deutlich geringeren Funktionswerten in der Nähe des globalen Optimums betreten und der Zielpunkt anschließend in immer kleiner werden Schritten erreicht. Es ist klar zu erkennen, dass im Verlauf einige lokale Optima im Ergebnisraum der Mittelwertparameter überschritten werden, sowohl von EM als auch von den anderen eingezeichneten Verläufen.

Im Gegensatz dazu sind alle Verläufe der Funktionswerte in der Grafik unten links in Abbildung 24 monoton fallend, das heißt es müssen keine lokalen Optima überwunden werden. Das gilt auch für den hier deutlich weiteren Weg über die einzelnen EM-Iterationen. Außerdem erreicht der EM-Algorithmus einen etwas besseren Likelihoodwert als das theoretische Optimum, welches auch beim Blick auf die partiellen Ableitungen (unten rechts in Abbildung 24) nicht das exakte Optimum dieser konkreten Funktion sein kann, da bis auf σ_1 alle partiellen Ableitungen deutlich ungleich 0 sind. Das Ergebnis des EM-Algorithmus ist in allen partiellen Ableitungen 0, der Algorithmus hat also einen Kandidatenpunkt für ein Optimum gefunden. Aufgrund des Funktionswertes und der Lage im Raum der Mittelwertparameter ist davon auszugehen, dass es sich um das globale Optimum handelt. Um zu analysieren, warum auf dem Weg vom Start zum Optimum keine lokalen Optima zu überwinden sind, obwohl der Raum der Mittelwertparameter zahlreiche Optima aufweist, ist in Abbildung 25 der aktuelle μ_1, μ_2-Raum in jeder Iteration des EM-Algorithmus dargestellt.

Wie in Kapitel 2.4 herausgestellt wurde, wird die Multimodalität maßgeblich durch den Abstand der beiden Standardabweichungen beeinflusst. Im Startpunkt sind diese mit $\sigma_1 = 9.6$ und $\sigma_2 = 3.9$ noch in der gleichen Größenordnung, sodass keine Multimodalität besteht. Der Algorithmus macht einen großen ersten Iterationsschritt, die Varianzparameter verändern sich jedoch nur geringfügig und es sind keine größeren Differenzen zwischen den Funktionswerten zu erkennen. Im weiteren Verlauf werden die Abstände zwischen den Iterationen kleiner und ca. ab der sechsten Iteration ($\sigma_1 = 9.8$ und $\sigma_2 = 1.0$) wird deutlich, dass im Bereich des globalen Optimums ($\mu_2 = -5$) bessere Funktionswerte vorliegen als im Bereich des Startpunktes. Der Unterschied in den Varianzen entspricht hier ungefähr dem Punkt,

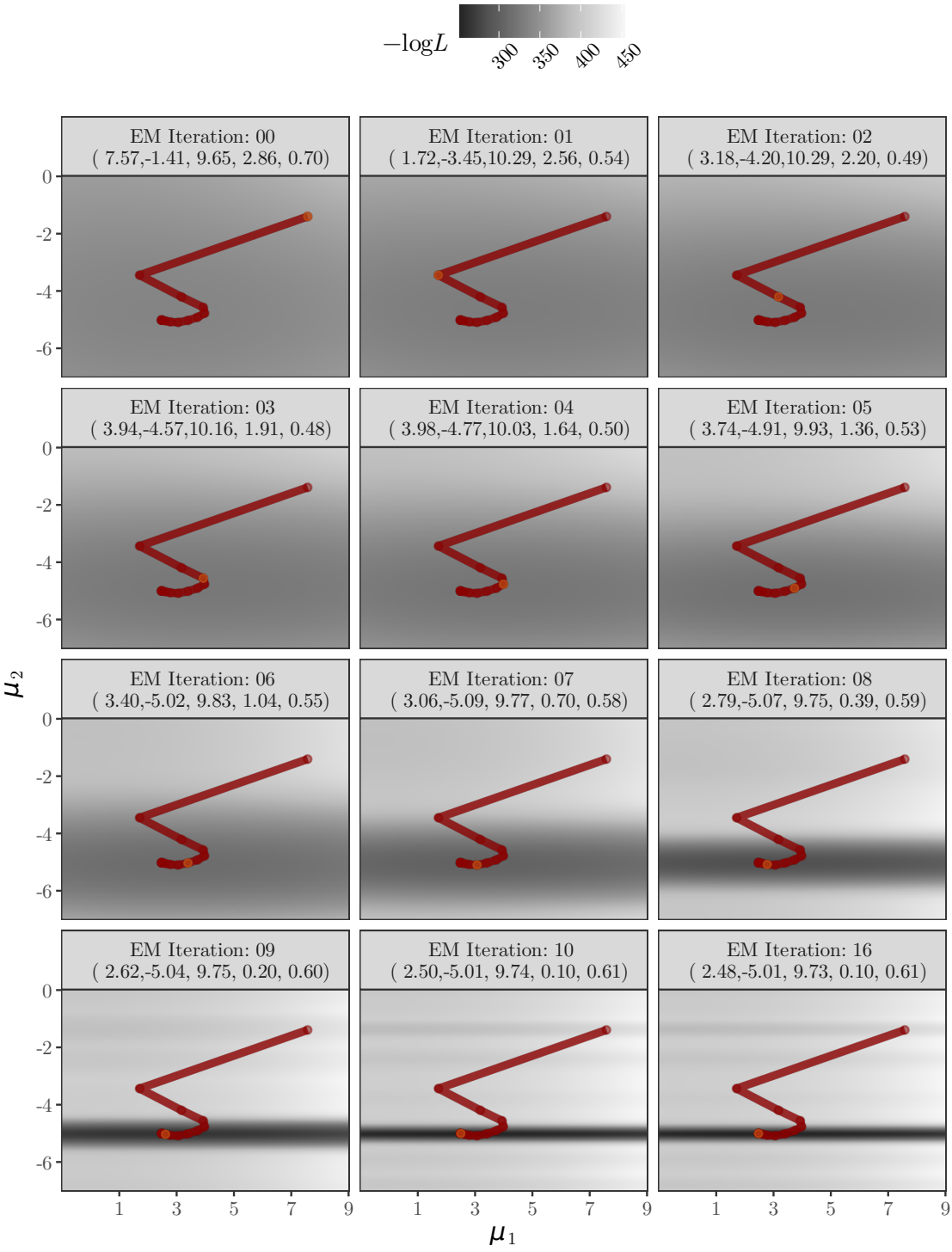

Abbildung 25: Raum der Mittelwerte im Verlauf der 5D-Optimierung mittels EM

ab dem in den Vorversuchen (Kap. 2.4) Multimodalität festgestellt werden konnte. Allerdings gilt in dieser sechsten Iteration bereits $\mu_2 = -5$, sodass die lokalen Optima, die in dieser Richtung ab jetzt bestehen, nicht mehr überwunden werden müssen, da der Wert des theoretischen Optimums in diesem Parameter bereits erreicht wurde. In den weiteren Iterationen werden nur noch sehr kleine Distanzen zwischen den einzelnen Iterationen überbrückt und die kleinere Varianz nähert sich immer weiter 0.1 an, wodurch die Multimodalität im μ_1, μ_2-Raum immer stärker wird und sich schließlich in Iteration 16 die bereits in Abbildung 24 dargestellte finale Funktion im Raum der Mittelwerte ergibt.

Bei Veränderung des Startwertes der kleineren Varianz σ_2 von 2.86 auf den theoretischen Wert 0.1 gelingt es dem EM-Algorithmus nicht mehr die zugrunde liegenden Mittelwerte zu erreichen. In der grafischen Darstellung in Abbildung 26 wird deutlich, dass im Laufe der Optimierung nur μ_1 variiert wird, während μ_2 in unmittelbarer Nähe des Startwertes verbleibt. Durch die kleinere Varianz liegt hier die Vermutung nahe, dass am Startwert bereits lokale Optima im Raum der Mittelwerte existieren und der Algorithmus zum bzgl. μ_2 nächstgelegenen Optimum konvergiert. Die partiellen Ableitungen in Abbildung 26 belegen die (lokale) Konvergenz des EM-Laufs.

Analog zum vorherigen Beispiel ist der Raum der Mittelwertparameter in Abbildung 27 in einzelnen Iterationsschritten dargestellt. Die bereits am Startwert bestehenden lokalen Optima sind in der Darstellung für Iteration 0 deutlich zu erkennen. In der ersten Iteration findet die Veränderung von μ_1 statt, die den bereits in Abbildung 26 dargestellten Verlauf der Optimierung im Raum der Mittelwerte hervorruft. Außerdem wird das Mischungsverhältnis von 0.7 auf 0.96 zugunsten der ersten Komponente erhöht. Für das erreichte lokale Optimum ist das nachvollziehbar, da dort nur sehr wenige Beobachtungen mit einer hohen Zugehörigkeitswahrscheinlichkeit zur zweiten Komponente vorhanden seien können. Diese stammen tatsächlich aus der Mischungskomponente mit der großen Varianz und liegen nur zufällig in der Nähe des Ergebnisses für μ_2. Die 50 % der Beobachtungen, die eigentlich zur zweiten Komponente mit $\mu_2 = -5$ gehören, werden hier vollständig der ersten Komponente zugerechnet. Durch diese Veränderung ist das tatsächliche Optimum in den weiteren Iterationen weniger stark ausgeprägt, aber trotzdem noch deutlich als Bereich mit den besten Likelihood-Werten zu erkennen. Die weiteren Iterationen verursachen nur noch minimale Veränderungen der Parameter ohne erkennbare Auswirkungen. Das erreichte lokale Optimum bzgl. μ_2 ist durchgehend vorhanden.

Durch die Betrachtung der Likelihood bei lokaler Konvergenz wird deutlich, dass das globale Optimum teilweise nicht erreicht werden kann. Unklar bleibt bis hierhin

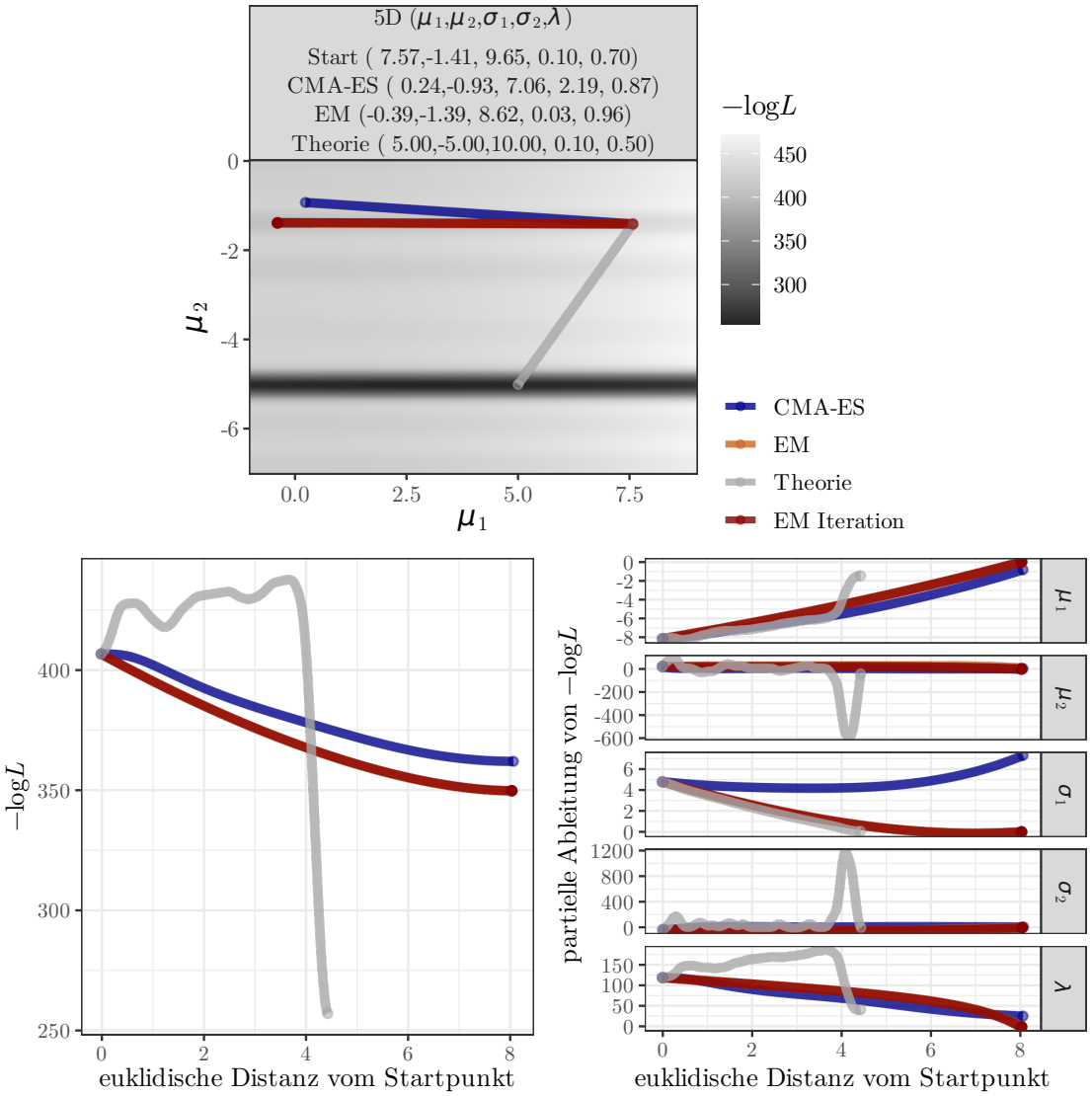

Abbildung 26: Optimierungspfade im 5D-Fall mit verändertem Startwert

trotzdem, wie der EM-Algorithmus zu den einzelnen Iterationsschritten kommt, die letztlich zu einem Optimum führen. Dazu wird die Funktionsweise des Algorithmus im Folgenden noch etwas genauer analysiert. Wie in Kapitel 2.3 beschrieben, optimiert der EM-Algorithmus nicht die bisher betrachtete Likelihood-Funktion direkt, sondern die *complete-data*-Likelihood für gegebene Klassenzuordnungen. EM verwendet also Vorwissen gegenüber einer rein numerischen Optimierung auf der Likelihood-Funktion. Zwar sind vorab keine Zugehörigkeiten zu den $g = 2$ Komponenten bekannt, jedoch können die A-posteriori-Wahrscheinlichkeiten τ des vorherigen Schrittes k eingesetzt werden, wodurch als interne Optimierungsfunktion im EM-Algorithmus der bedingte Erwartungswert

$$Q\left(\boldsymbol{\Psi}|\,\boldsymbol{\Psi}^{(k)}\right) = \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i\left(x_j|\,\boldsymbol{\Psi}^{(k)}\right)\left(\log\lambda_i + \log f_i(x_j|\boldsymbol{\theta}_i)\right) \tag{3.1}$$

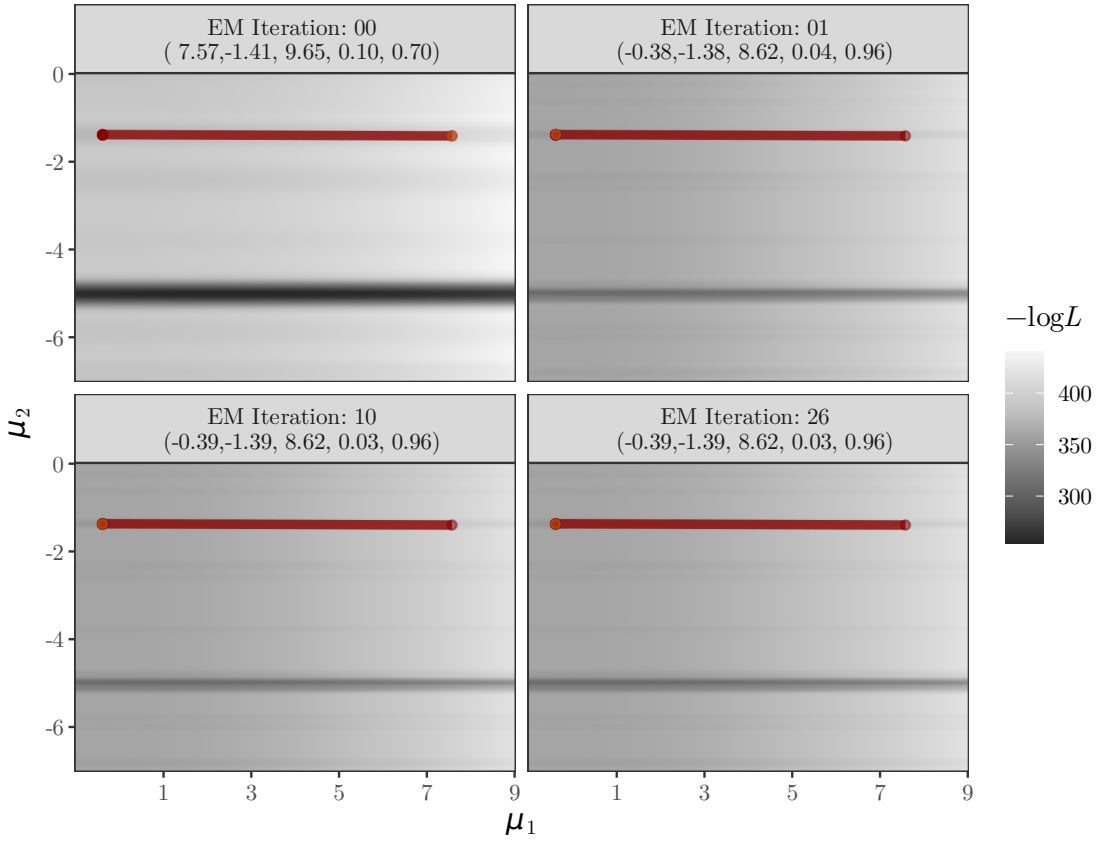

Abbildung 27: Raum der Mittelwerte bei verändertem Startwert

verwendet wird. Statt der Likelihood soll nun diese interne Funktion des E-Steps dargestellt werden, um Vorteile des EM-Algorithmus aufzuzeigen.

Um das absolute Optimum in den einzelnen Iterationsschritten besser sichtbar zu machen, wird jedoch zunächst auch für die Likelihood die Darstellung der Funktionsoberfläche angepasst. Die Farbskala verläuft auf den oberen 99 % von dunkelgrau nach weiß und auf dem kleinsten Prozent nochmal von weiß nach schwarz. Dies führt dazu, dass der Bereich des absoluten Minimums im dargestellten Bereich sichtbar wird. Gleichzeitig muss natürlich beachtet werden, dass kein so deutliches Optimum vorliegt, wie es auf den ersten Blick scheint. Ganz im Gegenteil handelt es sich nur um äußerst geringe Differenzen, weshalb diese Art der Darstellung überhaupt erst verwendet wird. Um das Optimum in jeder Iteration sichtbar machen zu können, muss außerdem eine eigene Farbskala pro Iteration verwendet werden. Ansonsten wäre das Optimum unter Umständen nur in der Iteration mit der stärksten Ausprägung eines Optimums zu erkennen.

Für das bekannte Beispiel aus den Abbildungen 24 und 25 sind in Abbildung 28 vier Iterationen auf der Likelihood-Funktion dargestellt. Es ist deutlich zu erkennen, dass das Optimum bereits bei der ersten Iteration in der Nähe der später erreichten Mittelwerte liegt, ohne dass der Algorithmus in die unmittelbare Nähe gelangt. Es ist also nicht der Fall, dass in jedem Schritt die optimalen Mittelwerte bedingt auf die übrigen Parameter gefunden werden und sich durch die aktualisierten Parameter im nächsten Schritt eine deutliche Veränderung des Optimums ergibt. Stattdessen wird das sich leicht verändernde Optimum mit immer kleiner werden Schritten angenähert und irgendwann eingeholt. So ist der Algorithmus in Iteration 5 schon deutlich näher am Optimum, in Iteration 10 ist das mittlerweile in μ_2-Richtung stark zusammengezogene Optimum nahezu erreicht, während sich für weiter entfernte Werte von μ_2 die bekannten lokalen Optima gebildet haben.

Im Vergleich dazu lassen sich die einzelnen Iterationsschritte durch Betrachten des internen Erwartungswertes genau erklären. In Abbildung 29 ist die Funktion Q im Raum der Mittelwerte für die gleichen Iterationen dargestellt. Für die Iterationen 1 und 5 ist zu erkennen, dass der Algorithmus immer genau die optimalen Mittelwerte, bedingt auf die Parameter des vorherigen Schrittes, erreicht. Das ist dadurch zu erklären, dass für jeden einzelnen Schritt die eindeutige analytische Lösung auf Basis der Zugehörigkeitswahrscheinlichkeiten der vorherigen Iteration verwendet wird (vgl. Kap. 2.3). Die hier dargestellten Mittelwerte der Normalverteilungskomponenten ergeben sich im $(k+1)$-ten Schritt durch

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i \left(x_j | \, \boldsymbol{\Psi}^{(k)} \right) x_j}{\sum_{j=1}^{n} \tau_i \left(x_j | \, \boldsymbol{\Psi}^{(k)} \right)}. \tag{3.2}$$

Aufgrund der Eindeutigkeit der analytischen Lösungen ist Q damit grundsätzlich unimodal, was zu einem deutlich vereinfachten Optimierungsproblem für EM führt. Jede Darstellung einer einzelnen Iteration ist damit eine direkte Darstellung der Formeln 3.1 und 3.2.

In Iteration 10 kann der Erwartungswert numerisch nicht dargestellt werden, da alle Werte in dem betrachteten Bereich den Wert ∞ annehmen. Q ist analog zur log-Likelihood bereits logarithmiert, sodass $\exp(Q)$ mit der nicht-logarithmierten Likelihood verglichen werden kann. Im Gegensatz zu Letzterer ist $\exp(Q)$ allerdings ein reines Produkt. Dadurch verursacht bereits eine einzelne Zugehörigkeitswahrscheinlichkeit von (numerisch) 0 insgesamt einen Wert von 0. Um dies zu verdeutlichen ist in dem entsprechenden Fall $-\exp(Q)$ dargestellt, welches konstant 0 ist. Eine numerische Optimierung des vereinfachten Problems Q wäre demnach nicht möglich. Ein weiterer Vorteil von EM ist jedoch, dass zur Optimierung von Q

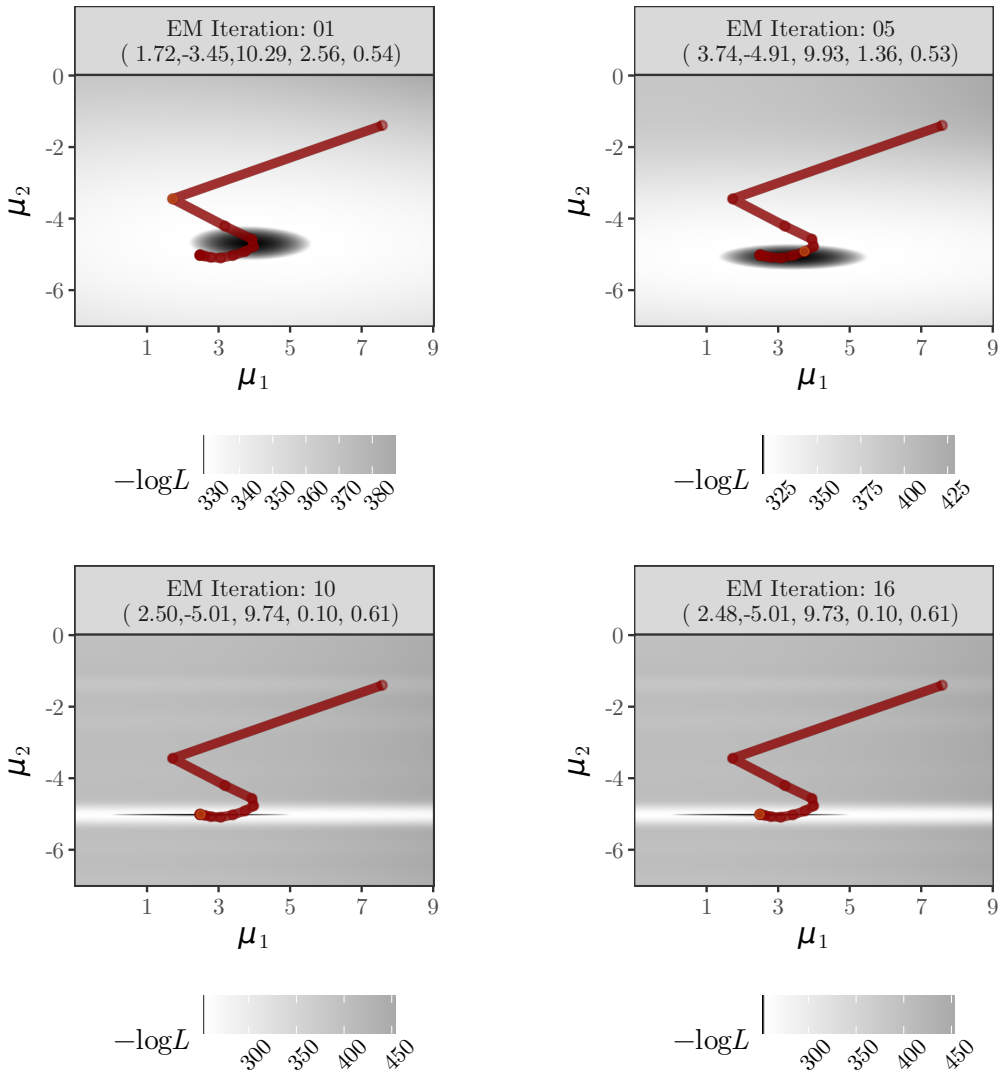

Abbildung 28: Raum der Mittelwerte der Likelihood im Verlauf der 5D-Optimierung mittels EM

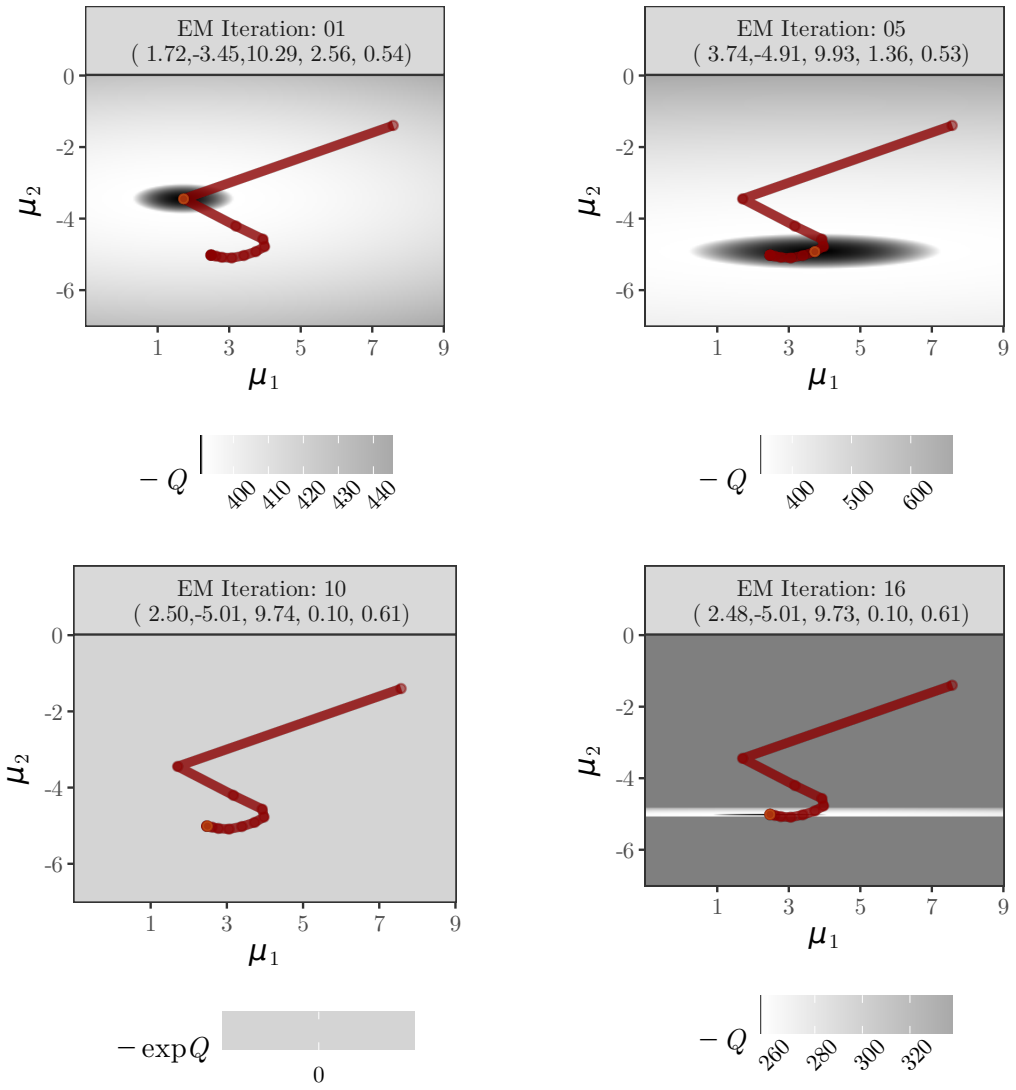

Abbildung 29: Raum der Mittelwerte des bedingten Erwartungswertes Q im Verlauf der 5D-Optimierung mittels EM

für Normalverteilungsmischungen analytische Lösungen existieren, die diese numerische Problematik umgehen, weil keine verschiedenen Werte von Q miteinander verglichen werden müssen. In späteren Iterationen ist daher zu erkennen, dass auch wieder Funktionswerte in der Nähe des Optimums darstellbar sind. Der Bereich, in dem die lokalen Minima der Likelihood existieren, nimmt aber weiterhin unendliche Funktionswerte an. Dementsprechend gibt es dort auch keine lokalen Minima, die erreicht werden könnten. Der bedingte Erwartungswert ist also immer unimodal. Das erklärt auch, warum es für die Normalverteilung überhaupt eine eindeutige analytische Lösung über die einzige Nullstelle der Ableitung geben kann.

Auch bei Konvergenz zu einem lokalen Optimum existiert im bedingten Erwartungswert nur das entsprechende Optimum. Für das Beispiel aus Abbildungen 26 und 27 sind in Abbildung 30 ebenfalls vier Iterationen auf der Likelihood-Funktion

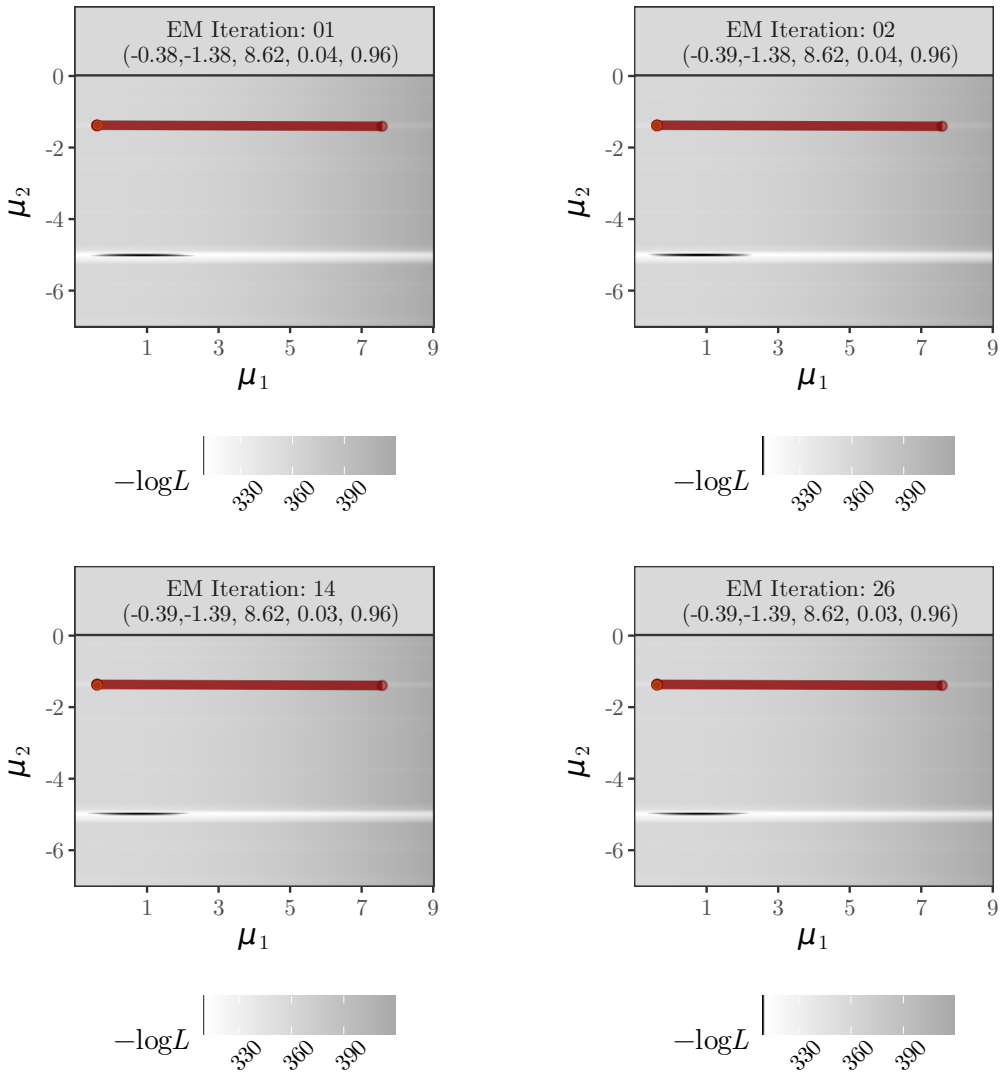

Abbildung 30: Raum der Mittelwerte der Likelihood bei verändertem Startwert

mit der veränderten Darstellung des Optimums enthalten. Da nach der ersten Iteration keine nennenswerten Veränderungen stattfinden, sind keine Unterschiede in der Likelihood-Funktion zwischen den dargestellten Iterationen zu erkennen. Wie schon für das vorherige Beispiel erkennt man hier auch den μ_1-Wert des tatsächlichen Optimums bei ca. 1 deutlich, während diese Information in der ursprünglichen Darstellung von den starken Unterschieden in μ_2 überlagert wurde. In Abbildung 31 ist zu erkennen, dass in diesem Beispiel bereits nach der ersten Iteration Q konstant unendlich ist. In den weiteren Iterationen existieren nur im Bereich des erreichten lokalen Optimums wieder darstellbare Werte, sodass der bedingte Erwartungswert Q auch in diesem Fall unimodal ist.

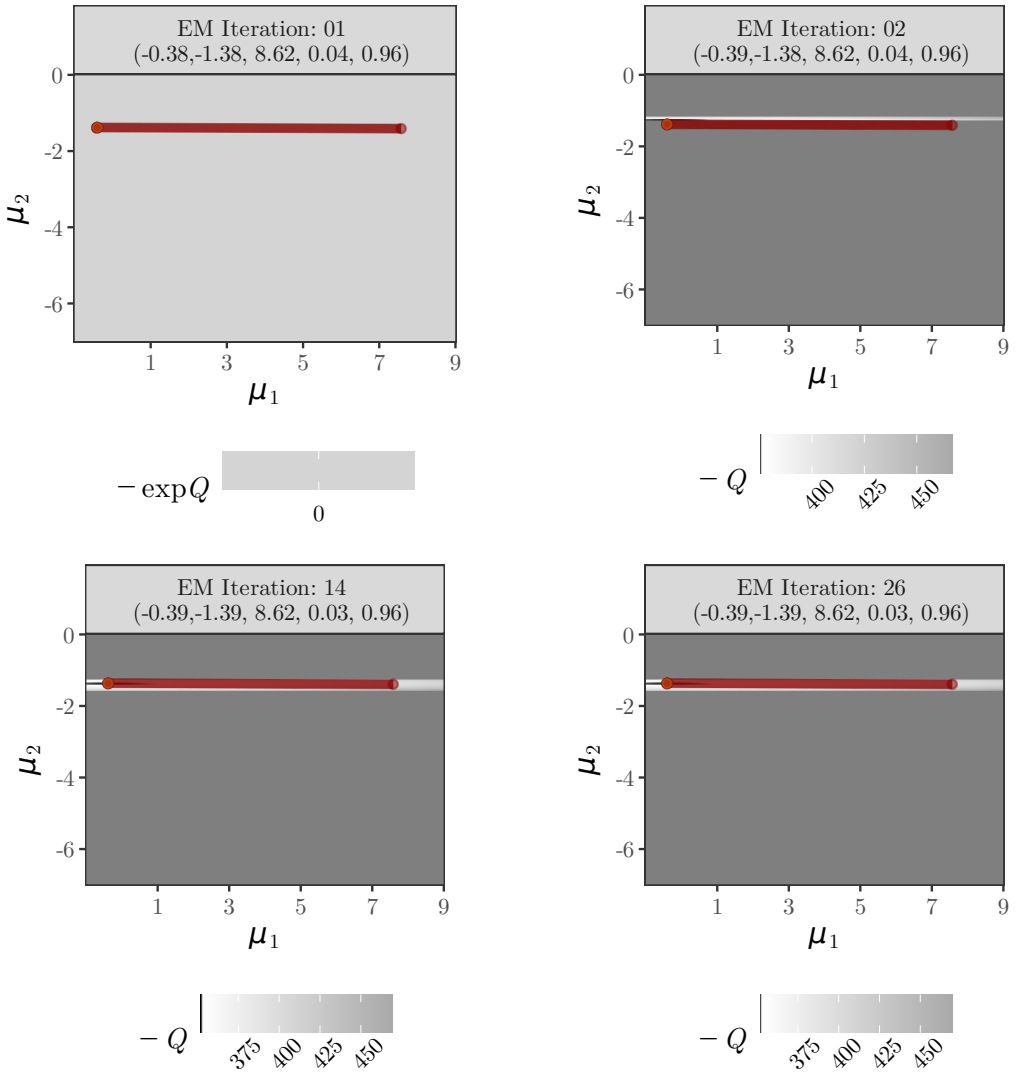

Abbildung 31: Raum der Mittelwerte des bedingten Erwartungswertes Q bei verändertem Startwert

Der entscheidende Vorteil von EM ist also die Verwendung von Zugehörigkeitswahrscheinlichkeiten für die einzelnen Beobachtungen, wodurch sich eine unimodale Zielfunktion ergibt. Durch diese Vereinfachung des Problems gegenüber den Vergleichsalgorithmen MBO und CMA-ES, die *Black-Box*-Optimierung direkt auf der Likelihood-Funktion betreiben, gelingt es regelmäßig das zugrunde liegende Optimum der Likelihood-Funktionen zu finden.

4 Anzahl Optima multimodaler Likelihood-Funktionen

Wie zu Beginn von Kapitel 2.2 erläutert, entstehen lokale Optima an Stellen, an denen einzelne oder gruppierte Beobachtungen isoliert auftreten. Dieser Eindruck soll durch theoretische Ergebnisse bestätigt werden. Als Ausgangspunkt dient ein Beweisszenario von Améndola Cerón (2017). Dort wird gezeigt, dass für eine systematisch aufgebaute Stichprobe aus K Clustern mit jeweils zwei Beobachtungen mindestens K lokale Optima in der Likelihood-Funktion eines daran angepassten Mischverteilungsmodells aus zwei Normalverteilungen existieren. Dieses Vorgehen wird im folgenden Unterkapitel 4.1 genauer erläutert, bevor daran anschließend untersucht wird, ob damit auch in Szenarien, die eine normalverteilte Datengrundlage annähern, spezifische Optima für einzelne Cluster in den Daten erreicht werden können. Dazu werden in einem ersten Schritt aus Gleichverteilungen zusammengesetzte Stichproben betrachtet (Kap. 4.2), bevor in einem weiteren Schritt Cluster aus normalverteilten Beobachtungen zusammengesetzt (Kap. 4.3) und schließlich Beobachtungen einer zusätzlichen Verteilungskomponente hinzugefügt werden (Kap. 4.4). In einer abschließenden Anwendung auf normalverteilte Zufallsstichproben in Kapitel 4.5 wird praxisrelevant demonstriert, wie auch darin systematisch einzelne lokale Optima identifiziert werden können.

4.1 Gleichverteilte Cluster mit zwei Beobachtungen

Die betrachtete Stichprobe besteht für $n = 2K$ Beobachtungen und $K > 2$ aus den Werten

$$(x_1, \ldots, x_{2K}) = (1, 1.2, 2, 2.2, \ldots, K, K + 0.2),$$

sodass jedes enthaltene Cluster $k \in \{1, \ldots, K\}$ aus zwei Punkten mit Abstand 0.2 besteht, während der Abstand zur nächsten nicht enthaltenen Beobachtung immer 0.8 beträgt. Für jedes dieser Cluster wird ein EM-Lauf gestartet, wobei $\mu_1 = k + 0.1$ und $\sigma_1 = 0.1$ als Startwerte für die erste Mischungskomponente gewählt werden. Die Startparameter der zweiten Komponente werden aus den übrigen Beobachtungen bestimmt, wobei die unkorrigierte Stichprobenvarianz verwendet wird. Das Mischungsverhältnis ist $\lambda = 1/K$. Für diese Läufe kann gezeigt werden, dass $0.09 \leq \mu_1 - k \leq 0.11$ und damit für jedes Cluster ein spezifisches Optimum erreicht wird. Dementsprechend verfügt die Likelihood-Funktion über mindestens K lokale Optima (Améndola Cerón, 2017, S. 30 ff.).

In Abbildung 32 ist das Vorgehen von Améndola Cerón (2017) für $K = 7$ dargestellt. In der oberen Grafik handelt es sich um einen Querschnitt bezüglich μ_1 am Optimierungsergebnis für $k = 1$, in der Mitte für $k = 4$ und unten für $k = 7$. Am oberen Rand sind die Beobachtungen eingezeichnet, die vertikalen Linien repräsentieren den μ_1-Wert der einzelnen EM-Läufe, der Punkt darauf gibt den erreichten Funktionswert an. Da in der oberen Darstellung die weiteren Parameter auf das Ergebnis des ersten Clusters festgesetzt sind, stimmt nur dort der gelb eingekreiste, erreichte Wert mit dem Funktionswert des Querschnitts überein. Bei den weiteren Clustern sind veränderte Werte der übrigen Parameter nötig, um den Funktionswert des jeweiligen Optimierungslaufes im Querschnitt exakt zu erreichen, wie in den beiden weiteren Darstellungen zu erkennen ist. In jeder einzelnen Grafik ist bereits deutlich zu erkennen, dass für jedes Ausgangscluster aus der Stichprobe ein lokales Optimum existiert und dieses auch bezüglich des μ_1-Wertes im jeweiligen Optimierungslauf erreicht wird. Zudem liegt der μ_1-Wert für jedes Optimum mittig zwischen den beiden Punkten des jeweiligen Clusters und die Optimierungsergebnisse in Tabelle 8 bestätigen die Aussage $0.09 \leq \mu_1 - k \leq 0.11$ für dieses Beispiel. Darüber hinaus ist in der Tabelle zu erkennen, dass auch die weiteren Parameter der beiden Normalverteilungen in der Nähe des jeweiligen Startwertes verbleiben, während das Mischungsverhältnis λ stark durch den Abstand der beiden Mittelwerte beeinflusst ist. Da es sich um konvergente Lösungen des EM-Algorithmus handelt, liegen hier lokale Optima vor. Bei ähnlichen Mittelwerten passen die Beobachtungen des kleinen Clusters auch sehr gut zum Rest, sodass der ersten Komponente ein kleineres Gewicht zufällt, während die Beobachtungen bei größtmöglichem Abstand der Mittelwerte nur noch geringe Werte in der Dichte der zweiten Komponente besitzen und der Mischungsanteil demnach in der Nähe des Startwerts von 1/7 verbleibt. Grundsätzlich genügt natürlich schon der Verbleib am Startwert für einen Parameter, um für jeden Startwert ein spezifisches und damit von den anderen Läufen verschiedenes Optimum zu erhalten.

Die Anzahl der Beobachtungen pro Cluster und die Abstände zwischen den Beobachtungen sind hier fest vorgegeben. Durch das Ausnutzen dieser Eigenschaften der Stichprobe gelingt es, die Aussage, dass jedes einzelne Cluster ein spezifisches lokales Optimum besitzt, für eine beliebige Anzahl Cluster K zu beweisen (vgl. Améndola Cerón, 2017, S. 31 ff.). Aufgrund dieser Eigenschaften ist die betrachtete Stichprobe offensichtlich nicht normalverteilt, bzw. aus zwei Normalverteilungen zusammengesetzt. Für die folgenden Annäherungsschritte an eine Normalverteilung (Kap. 4.2 und 4.3) wird daher die gesamte Datensituation als Repräsentation einer einzigen Verteilungskomponente aufgefasst. Das Erreichen lokaler Optima funktioniert mit Daten, die ausschließlich aus der Verteilung mit größerer Varianz stam-

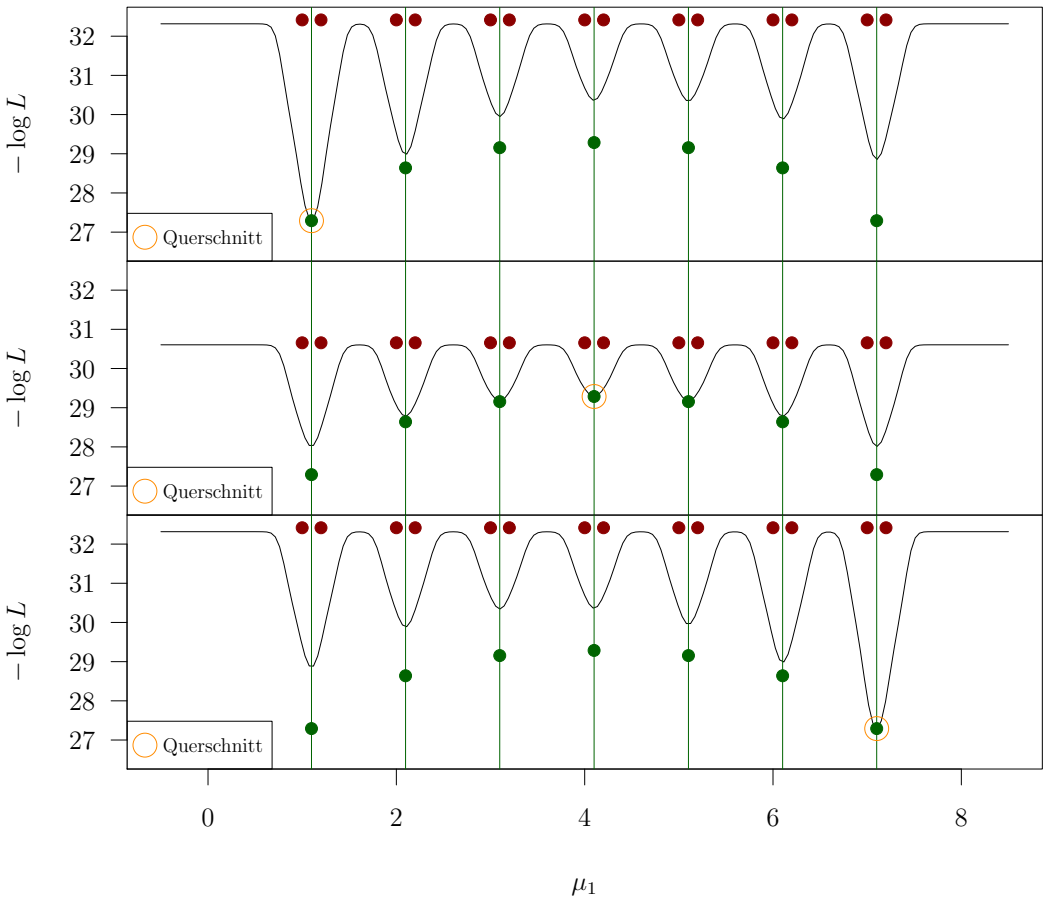

Abbildung 32: Durchführung für $K = 7$

Tabelle 8: Optimierungsergebnisse für $K = 7$

	λ	μ_1	μ_2	σ_1	σ_2	$-\log L$	Konv.
$k = 1$	0.131	1.099	4.553	0.1	1.746	27.292	ja
$k = 2$	0.103	2.098	4.330	0.1	1.989	28.640	ja
$k = 3$	0.079	3.098	4.186	0.1	2.064	29.155	ja
$k = 4$	0.069	4.100	4.100	0.1	2.075	29.286	ja
$k = 5$	0.079	5.102	4.014	0.1	2.064	29.155	ja
$k = 6$	0.103	6.102	3.870	0.1	1.989	28.640	ja
$k = 7$	0.131	7.101	3.647	0.1	1.746	27.292	ja

men, da die Verteilungskomponente mit der für die Multimodalität entscheidenden kleineren Varianz an einzelne Beobachtungscluster in diesen Daten angepasst werden kann. Eine zusätzlich zugrunde liegende Komponente mit kleinerer Varianz würde in diesen Fällen ohnehin nicht korrekt identifiziert. In Kapitel 4.4 werden schließlich Beobachtungen, die aus einer zusätzlichen Verteilungskomponente mit kleinerer Varianz stammen, hinzugefügt.

4.2 Normalverteilte Clustergrößen mit gleichverteilten Beobachtungen

Eine erste Möglichkeit ist es, der Normalverteilung entsprechend, die Anzahl der Beobachtungen in den Clustern in der Mitte deutlich zu erhöhen und davon ausgehend zu beiden Rändern hin schrittweise zu verringern. Für eine vorgegebene Intervallbreite 2^{-p}, mit $p \in \mathbb{N}_0$, wird nun die Anzahl der Beobachtungen n_k für K Cluster aus der Dichtefunktion der Standardnormalverteilung an den Intervallmittelpunkten m_k bestimmt. Diese Beobachtungen liegen gleichverteilt auf den mittleren 50 % eines jeden Intervalls und bilden so die einzelnen Cluster. Um wie im Ausgangsbeispiel eine deterministische Datensituation zu erhalten, liegen die Beobachtungen jeweils äquidistant im Cluster-Intervall. Das Histogramm mit entsprechender Intervallbreite stimmt dadurch mit dem einer perfekt normalverteilten Stichprobe überein. Die Anzahl Cluster ergibt sich dabei durch $K = K_0 \cdot 2^p$, wobei K_0 die Anzahl vorhandener Cluster im Fall $p = 0$ darstellt. K_0 ergibt sich wiederum aus der Gesamtanzahl an Beobachtungen n_0 für $p = 0$. Im Fall $p = 0$ ist die Intervallbreite $2^{-0} = 1$, daraus folgt, dass als Clustermittelpunkte alle ganzzahligen Werte in Frage kommen. Nur an den ganzzahligen Stellen x mit Standardnormalverteilungsdichte $\varphi(x) \geq 1/n_0$ wird ein Cluster durch mindestens eine Beobachtung realisiert. Beispielsweise ergibt sich für $n_0 = 100$ eine Clusteranzahl von $K_0 = 5$, da $\varphi(2) > 0.01 > \varphi(3)$, sodass für $|x| = 3$ keine Beobachtungen realisiert werden und an den Stellen -2 und 2 die äußersten besetzten Cluster existieren.

Um wie im Ausgangsbeispiel systematisch verschiedene Clusteranzahlen K zuzulassen, kann die Potenz p variiert werden. Gleichzeitig führt eine Erhöhung der Anzahl Cluster auch zu einer Erhöhung der Beobachtungsanzahl $n := n_0 \cdot 2^p$. Dementsprechend führt eine Halbierung der Intervallbreite nicht nur zu einer Verdopplung der Clusteranzahl K, sondern auch zu einer Verdopplung von n. Damit ist die Stichprobe für dieses Szenario durch Wahl von n_0 und p eindeutig bestimmt. Mit den Grenzen $a_{1,k} = m_k - \frac{2^{-(p+1)}}{2}$ und $b_{1,k} = m_k + \frac{2^{-(p+1)}}{2}$ der Cluster-Intervalle innerhalb

der Histogramm-Intervalle ergibt sich folgende Dichte $f_{1,p}(x)$ einer Mischung aus K Gleichverteilungen für das beschriebene Szenario:

$$\begin{aligned}
f_{1,p}(x) &= \sum_{k=1}^{K} \alpha_k \cdot f_k(x) \\
&= \sum_{k=1}^{K} \underbrace{\varphi(m_k) \cdot 2^{-p}}_{\text{Histogramm-Dichte}} \cdot f_{\mathcal{U}_{[a_{1,k},b_{1,k}]}}(x) \\
&= \sum_{k=1}^{K} \varphi(m_k) \cdot 2^{-p} \cdot \mathbb{1}_{x\in[a_{1,k},b_{1,k}]} \cdot \frac{1}{b_{1,k}-a_{1,k}} \\
&= \sum_{k=1}^{K} \varphi(m_k) \cdot \frac{1}{2^p} \cdot \mathbb{1}_{x\in[a_{1,k},b_{1,k}]} \cdot \frac{2}{2^{-p}} \\
&= \sum_{k=1}^{K} 2\varphi(m_k) \cdot \mathbb{1}_{x\in[a_{1,k},b_{1,k}]}.
\end{aligned}$$

Aufgrund der letzten Zeile liegt die Vermutung nahe, dass auch für größer werdendes p neben jedem von einem Cluster besetzten Bereich ein genauso großer Bereich vorhanden ist, in dem die Dichtefunktion 0 ist. Durch den Faktor 2 vor der Normalverteilungsdichte wird dies innerhalb der einzelnen Histogramm-Intervalle jeweils ausgeglichen.

Um eine normalverteilte Stichprobe noch weiter anzunähern, wird ein zweites Szenario konstruiert, das asymptotisch gegen die Normalverteilung konvergieren soll. Dazu wachsen die Cluster-Intervalle von anfänglich der halben Histogramm-Intervallbreite mit steigendem p auf die gesamte Breite an. Dementsprechend werden die Grenzen dieser inneren Intervalle zu $a_{2,k} = m_k - \frac{2^{-(p+1)}}{1+2^{-p}}$ und $b_{2,k} = m_k + \frac{2^{-(p+1)}}{1+2^{-p}}$ verändert. Für die zugehörige Dichtefunktion $f_{2,p}(x)$ gilt:

$$\begin{aligned}
f_{2,p}(x) &= \sum_{k=1}^{K} \alpha_k \cdot f_k(x) \\
&= \sum_{k=1}^{K} \underbrace{\varphi(m_k) \cdot 2^{-p}}_{\text{Histogramm-Dichte}} \cdot f_{\mathcal{U}_{[a_{2,k},b_{2,k}]}}(x) \\
&= \sum_{k=1}^{K} \varphi(m_k) \cdot 2^{-p} \cdot \mathbb{1}_{x\in[a_{2,k},b_{2,k}]} \cdot \frac{1}{b_{2,k}-a_{2,k}} \\
&= \sum_{k=1}^{K} \varphi(m_k) \cdot \frac{1}{2^p} \cdot \mathbb{1}_{x\in[a_{2,k},b_{2,k}]} \cdot \frac{1+2^{-p}}{2^{-p}} \\
&= \sum_{k=1}^{K} (1+2^{-p})\varphi(m_k) \cdot \mathbb{1}_{x\in[a_{2,k},b_{2,k}]}.
\end{aligned}$$

In Abbildung 33 ist $f_{1,p}$ für die drei niedrigsten Werte von p dargestellt. Es ist zu erkennen, dass sich die besetzen Bereiche bei Erhöhung von p um 1 halbieren und in den bisher unbesetzten Bereichen jeweils ein zusätzlicher Bereich mit $f_{1,p} > 0$ entsteht. Umgekehrt gilt dies ebenso für die unbesetzten Bereiche, sodass $f_{1,p}$ insgesamt für jedes p auf 50 % des betrachteten Bereiches Werte größer Null annimmt und für die übrigen 50 % gleich Null ist. Die Höhe der Bereiche bleibt bei Halbierung der Breite jeweils gleich. Insgesamt ist keine Annäherung an die eigentlich zugrunde liegende Dichte der Standardnormalverteilung zu sehen. Im Gegensatz dazu ist in Abbildung 34 zu erkennen, dass für $f_{2,p}$ mit jeder Erhöhung von p um eins die Bereiche mit $f_{1,p} > 0$ nicht exakt halbiert werden, sondern kontinuierlich etwas größer werden, während sich gleichzeitig die Höhe verringert. Das hat zur Konsequenz, dass hier der Anteil des unbesetzten Bereiches von anfänglich 50 % mit jeder Erhöhung von p kleiner wird und eine Annäherung an die Standardnormalverteilung stattfindet.

Die Verteilungsfunktionen zu beiden Szenarien sind in Abbildung 35 approximativ dargestellt. Dazu werden die Dichtefunktionen mit einer feinen Sequenz von Punkten ausgewertet und die kumulative Summe der Dichten an jedem Wert durch die Gesamtsumme normiert. Es ist zu erkennen, dass $\tilde{F}_{2,p}$ sich deutlich besser einem glatten Verlauf annähert, da die Steigung entlang der besetzten Bereiche mit steigendem p abnimmt und gleichzeitig die Bereiche mit Steigung 0 immer kleiner werden. Für $\tilde{F}_{1,p}$ hingegen findet eine solche Annäherung nicht statt, stattdessen ist die kontinuierliche Halbierung deutlich zu erkennen.

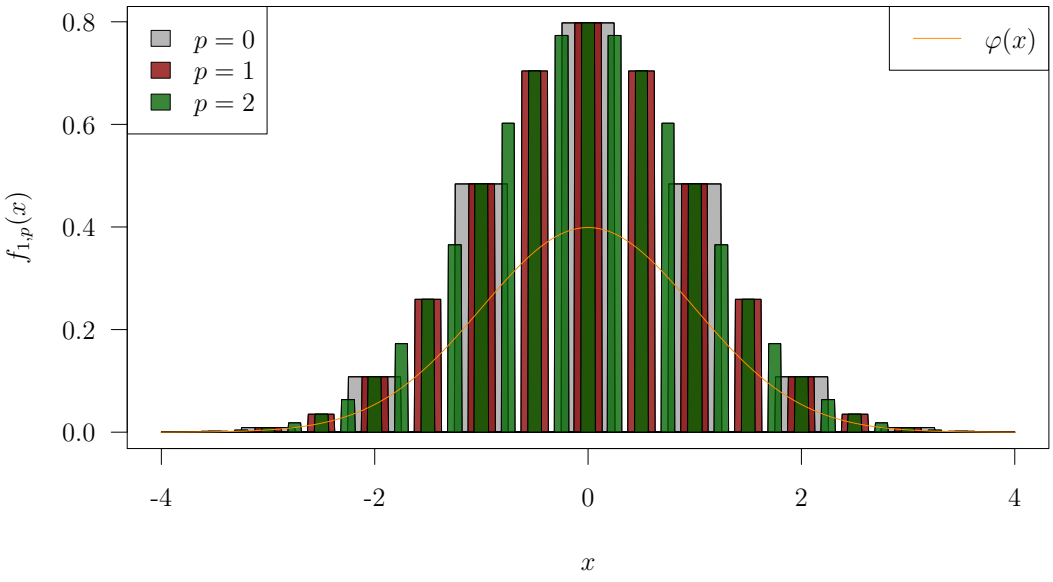

Abbildung 33: Dichte $f_{1,p}$ für $p = 0, 1, 2$

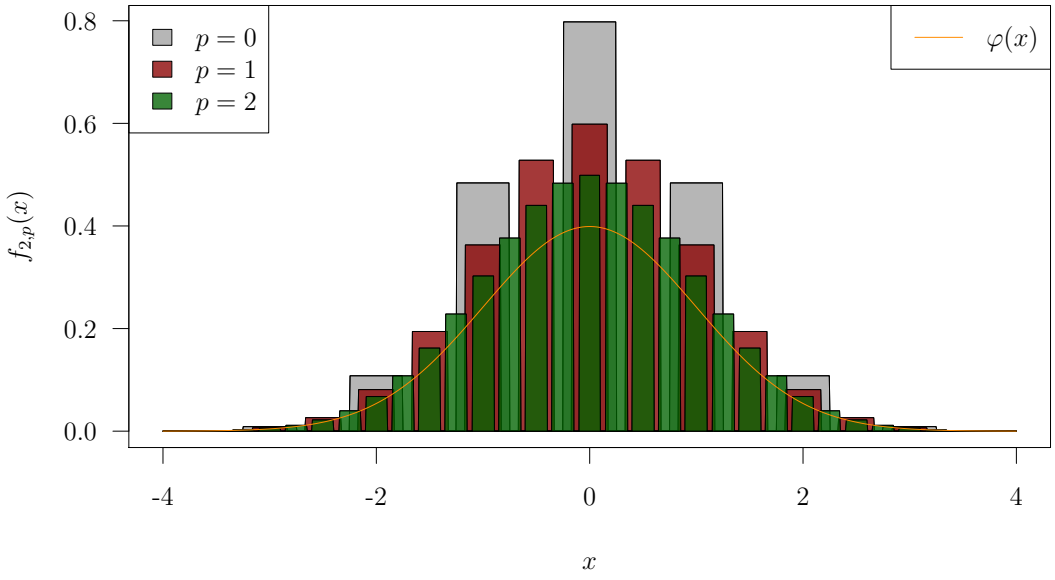

Abbildung 34: Dichte $f_{2,p}$ für $p = 0, 1, 2$

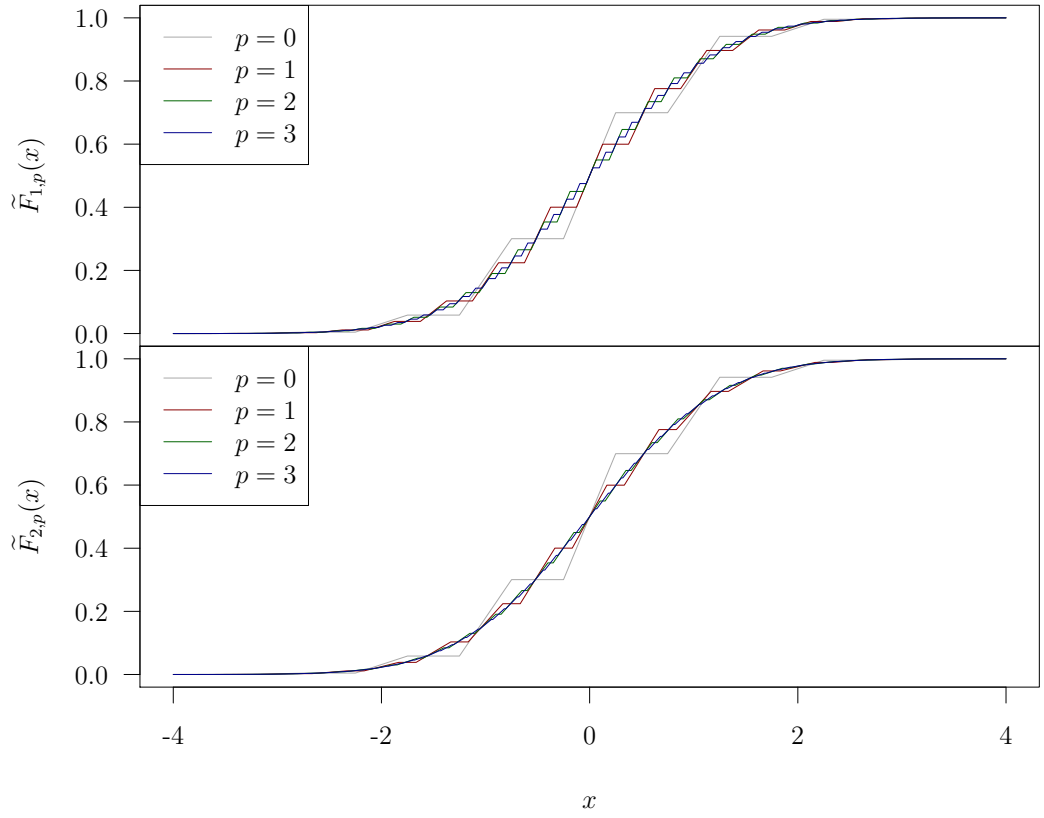

Abbildung 35: Approximierte Verteilungsfunktionen für $p = 0, 1, 2, 3$

Um den entscheidenden Unterschied zwischen beiden Varianten der Dichtefunktion herauszustellen, werden die Bereiche zwischen den Cluster-Intervallen betrachtet, in denen keine Dichte angenommen wird. Sei $p \in \mathbb{N}_0$ beliebig und seien m_k und $m_{k+1} = m_k + 2^{-p}$ Mittelpunkte der benachbarten Cluster-Intervalle $[a_k, b_k]$ und $[a_{k+1}, b_{k+1}]$ in einer allgemeinen Dichtefunktion f_p, dann gilt für den Bereich zwischen den Intervallen:

$$\begin{aligned}(b_k, a_{k+1}) &= (b_k, a_k + 2^{-p}) \\ &= (m_k + z(p), m_k - z(p) + 2^{-p}),\end{aligned}$$

mit verallgemeinerten Intervallgrenzen $a_k = m_k - z(p)$ und $b_k = m_k + z(p)$, wobei die Funktion $z(p)$ für $f_{1,p}$ spezifiziert wird durch $z_1(p) := \frac{2^{-(p+1)}}{2}$ und für $f_{2,p}$ durch $z_2(p) := \frac{2^{-(p+1)}}{1+2^{-p}}$. Für die Länge dieses Intervalls gilt:

$$m_k - z(p) + 2^{-p} - (m_k + z(p)) = 2^{-p} - 2z(p). \tag{4.1}$$

Setzt man nun die konkreten Funktionen für die Intervallgrenzen ein, ergibt sich für $f_{1,p}$ eine Intervalllänge von

$$\begin{aligned}2^{-p} - 2z_1(p) &= 2^{-p} - \frac{2^{-p}}{2} \\ &= \frac{2^{-p}}{2}\end{aligned}$$

und für $f_{2,p}$ eine Intervalllänge von

$$2^{-p} - 2z_2(p) = 2^{-p} - \frac{2^{-p}}{1 + 2^{-p}}.$$

Multipliziert man diese Intervalllängen mit der Gesamtzahl vorhandener Cluster in der Dichte, erhält man für $f_{1,p}$

$$\begin{aligned}K \cdot \frac{2^{-p}}{2} &= K_0 \cdot 2^p \cdot \frac{2^{-p}}{2} \\ &= \frac{K_0}{2}.\end{aligned}$$

Da K_0 die Anzahl nebeneinander liegender Intervalle mit Breite $2^0 = 1$ beschreibt, handelt es sich dabei gleichzeitig um die Länge des betrachteten Gesamtbereichs. Beispielsweise ergibt sich für das in den Abbildungen 33 und 34 erkennbare $K_0 = 7$ mit den äußersten Clustermittelpunkten -3 und 3, dass dieser Bereich der Länge 7 zwischen -3.5 und 3.5 liegt. Unabhängig von p liegt für $f_{1,p}$ also immer auf der Hälfte des betrachteten Gesamtbereiches ein Dichtewert von 0 vor.

Für $f_{2,p}$ ergibt sich entsprechend

$$K \cdot \left(2^{-p} - \frac{2^{-p}}{1+2^{-p}}\right) = K_0 \cdot 2^p \cdot \left(2^{-p} - \frac{2^{-p}}{1+2^{-p}}\right)$$
$$= K_0 - \frac{K_0}{1+2^{-p}},$$

wobei in diesem Fall gilt:

$$\lim_{p\to\infty} K_0 - \frac{K_0}{1+\underbrace{2^{-p}}_{\to 0}} = K_0 - K_0 = 0.$$

Der Teil des betrachteten Gesamtbereichs mit $f_{2,p} = 0$ wird also für steigendes p kleiner und ist asymptotisch für $p \to \infty$ nicht mehr vorhanden. Dementsprechend findet für $f_{2,p}$ mit steigendem p eine Annäherung an die Dichte der Standardnormalverteilung statt, da die Bereiche mit Dichtewert 0 in den einzelnen Histogramm-Intervallen asymptotisch nicht mehr ausgeglichen werden müssen.

Im Folgenden werden dem Vorgehen von Améndola Cerón (2017) entsprechend Optimierungsläufe für beide Varianten mit verschiedenen Werte von p durchgeführt. In Abbildung 36 ist der Fall mit Halbierung der Cluster-Intervalle für p von 0 bis 4 dargestellt. Für $p = 0$ ist zu erkennen, dass die Optimierungsläufe bezüglich μ_1 jeweils in der Nähe des Startwerts im jeweiligen Clusterzentrum verbleiben, bei den dünner besetzten Clustern jedoch etwas nach außen verschoben sind. Offensichtlich ist jedoch, dass von jedem Cluster aus ein spezifisches Optimum erreicht wird.

Für $p = 1$ ist zu erkennen, dass grundsätzlich auch für jedes Cluster ein spezifisches Optimum erreicht wird. Allerdings deutet sich bereits die Problematik an, die sich bei höheren Werten von p noch verstärkt: Die Optimierungsläufe, die am Start jeweils eines der Cluster am Rand mit zwei Beobachtungen vom Rest trennen, entwickeln sich zu einer Aufteilung, bei der nur noch der äußerste Punkt von den restlichen Beobachtungen separiert ist. Hier tritt das bereits in Kapitel 2 erwähnte Problem der sich gegen unendlich entwickelnden Likelihood für ein Cluster mit nur einem Punkt auf (vgl. McLachlan & Krishnan, 2008, S. 81). Der Algorithmus konvergiert dementsprechend nicht und bricht üblicherweise mit einem Fehler ab. Für diese Analyse wurde entschieden, das jeweils letzte Ergebnis vor dem Abbruch zu erhalten und als nicht konvergentes Ergebnis zurückzugeben. Natürlich ist eine entsprechende Normalverteilungskomponente mit nur einem Punkt nicht bestimmbar, sodass kein Mischverteilungsmodell angegeben werden kann. Dennoch kann es prinzipiell ein valides Clusterergebnis sein, nur einen Punkt vom Rest zu separieren. Der bis dahin erreichte Likelihood-Wert kann sich allerdings schon deutlich vom Niveau der tatsächlich erreichten Optima unterschieden. Für $p = 3$ ist zudem zu erkennen, dass derartige Ergebnisse teilweise mehrfach auftreten, da es im

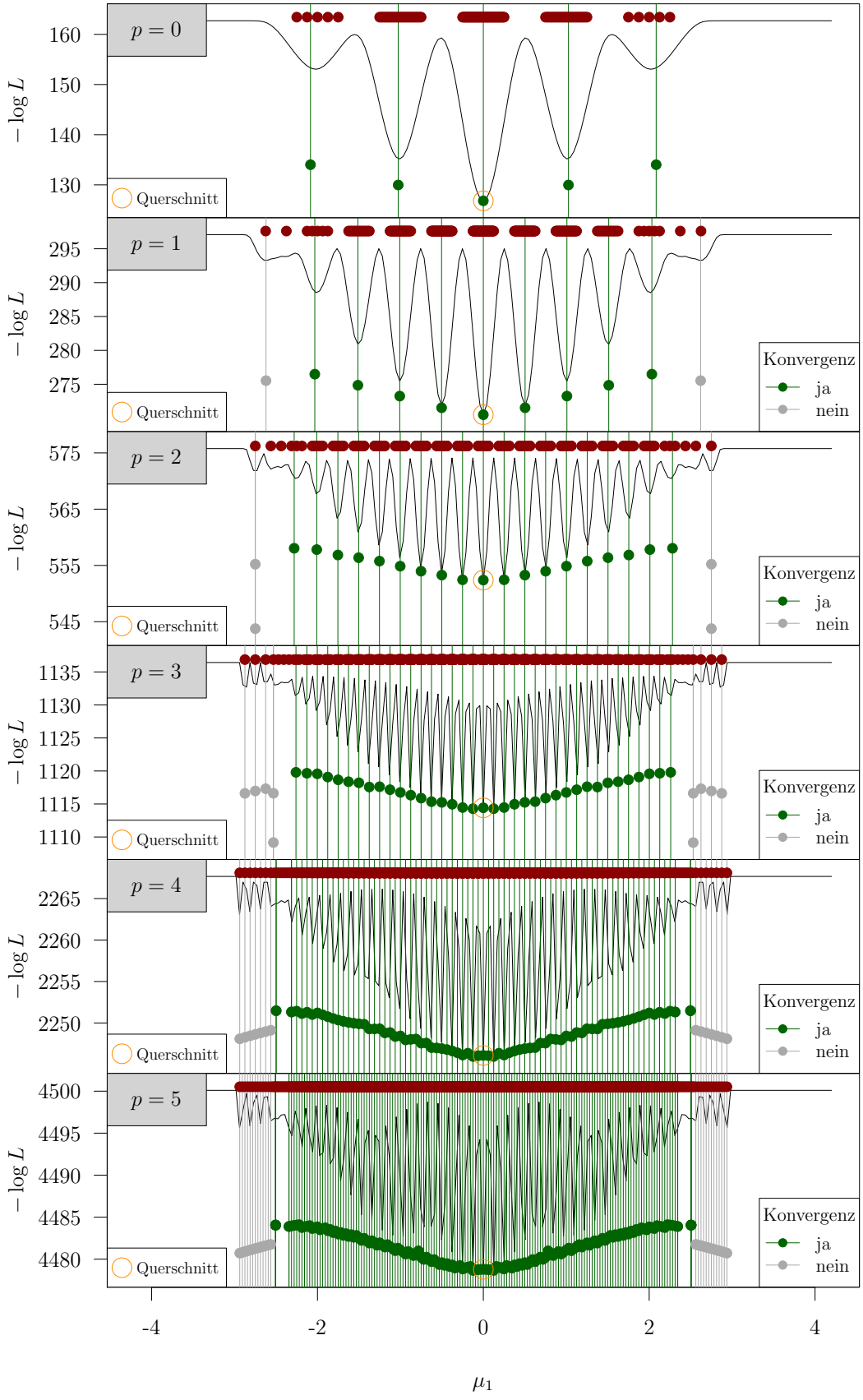

Abbildung 36: Fall 1: Gleichverteilungen mit Halbierung für $n_0 = 100$

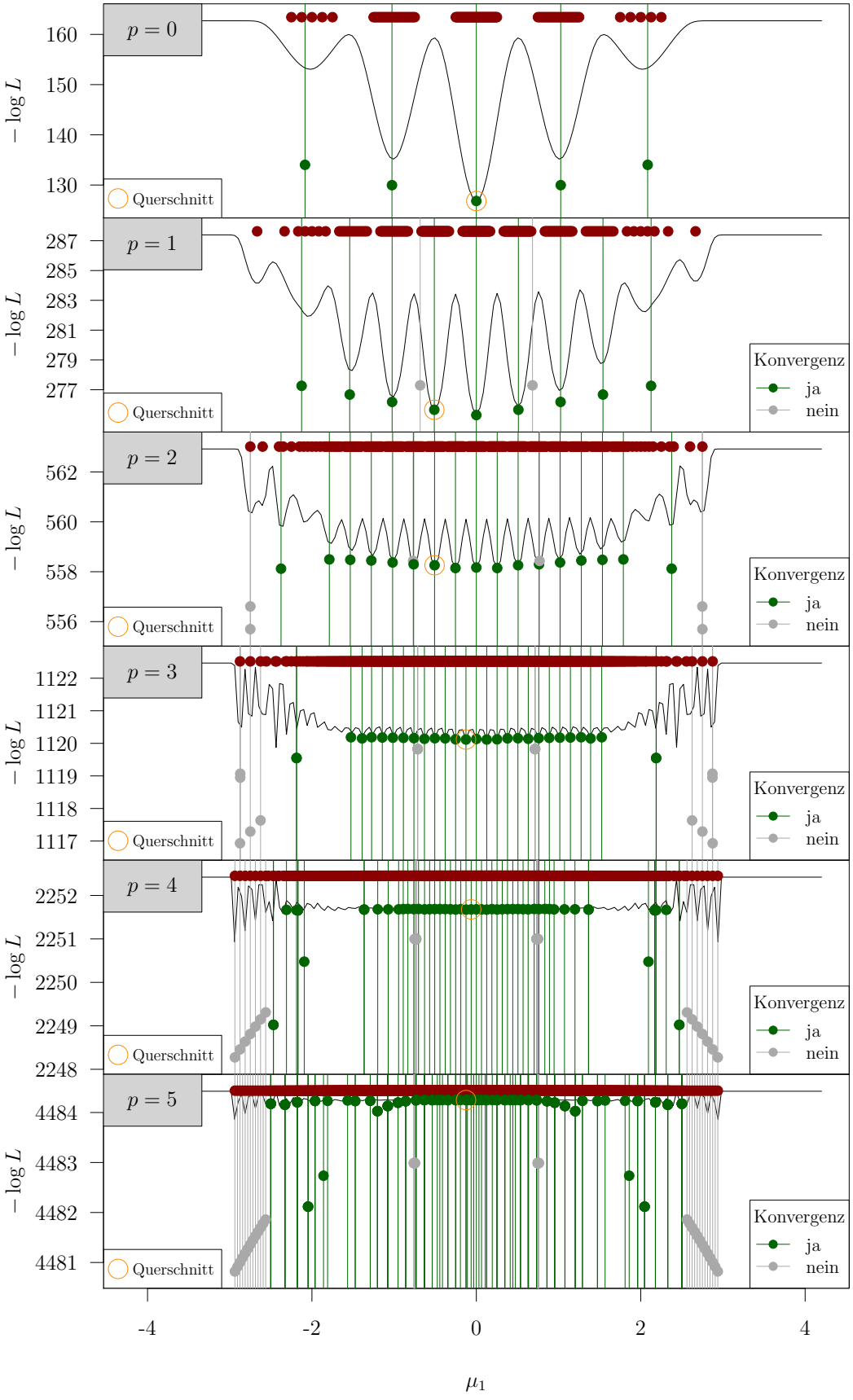

Abbildung 37: Fall 2: Gleichverteilungen mit Asymptotik für $n_0 = 100$

Randbereich Startwerte gibt, von denen sich der Algorithmus zur Separierung des nächstäußeren Clusters bzw. Einzelpunktes hin entwickelt. Vernachlässigt man die Randbereiche, wird jedoch für einen großen inneren Bereich für jedes Cluster ein spezifisches Optimum erreicht. Beim Vergleich der verschiedenen Werte von p ist zudem zu erkennen, dass dieser innere Bereich für steigendes p nicht kleiner wird.

Die Durchführung der Optimierungsläufe für den Fall mit asymptotisch wachsenden Clusterbreiten ist in Abbildung 37 dargestellt. Für $p = 0$ stimmen die Dichtefunktionen $f_{1,0}$ und $f_{2,0}$ überein, sodass das gleiche Optimierungsergebnis erzielt wird. Für die weiteren Werte von p werden allerdings deutliche Unterschiede sichtbar. Es fällt auf, dass sich von eher am Rand gelegenen Startpunkten nicht nur Aufteilungen mit Einzelpunkten ergeben, sondern auch nicht-konvergente Lösungen mit zentraleren Werten von μ_1 erreicht werden. Ab $p = 2$ entstehen diese Läufe ungefähr in Bereichen mit $1.5 < |\mu_1| < 2$, in denen keine oder nur vereinzelte spezifische Optima zu den Startclustern erreicht werden können. In Abbildung 38 ist beispielhaft der Optimierungsverlauf für $p = 3$ mit Startcluster $k = 10$ dargestellt.

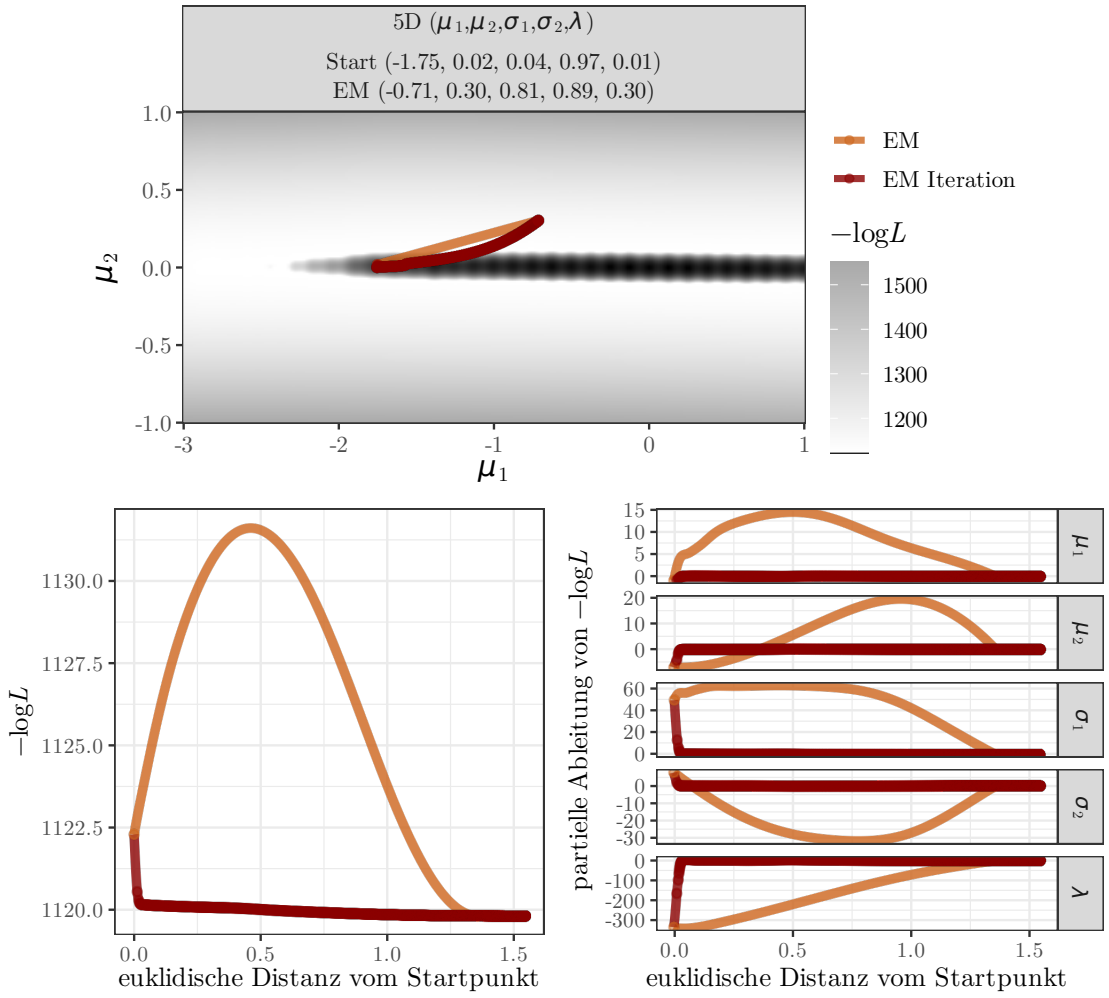

Abbildung 38: Optimierungspfade für $k = 10$ bei $p = 3$

Es ist deutlich zu erkennen, dass der Algorithmus mit wenigen Iterationen zu Beginn die Likelihood deutlich verbessern kann, bis alle partiellen Ableitungen ungefähr 0 erreichen. Das entspricht der Erwartung, dass das nächstgelegene Optimum für diesen Startpunkt erreicht wird. Der Algorithmus terminiert jedoch nicht, sondern findet mit jeder weiteren der 10000 betrachteten Iterationen eine weitere sehr kleine Verbesserung der Likelihood. Dabei nähert sich die kleinere Varianz der größeren an. Entsprechend der partiellen Ableitungen befinden sich alle diese Iterationen in unmittelbarer Nähe eines Kanditatenpunktes für ein lokales Optimum. Dies bestätigt auch die Darstellung der Likelihood-Werte im Raum der Mittelwerte in den einzelnen Iterationen in Abbildung 39. Dort ist deutlich zu erkennen, dass es sich in den einzelnen Iterationen um ein Optimum handelt, welches sich offensichtlich kontinuierlich weiter verschiebt. Da der EM-Algorithmus üblicherweise

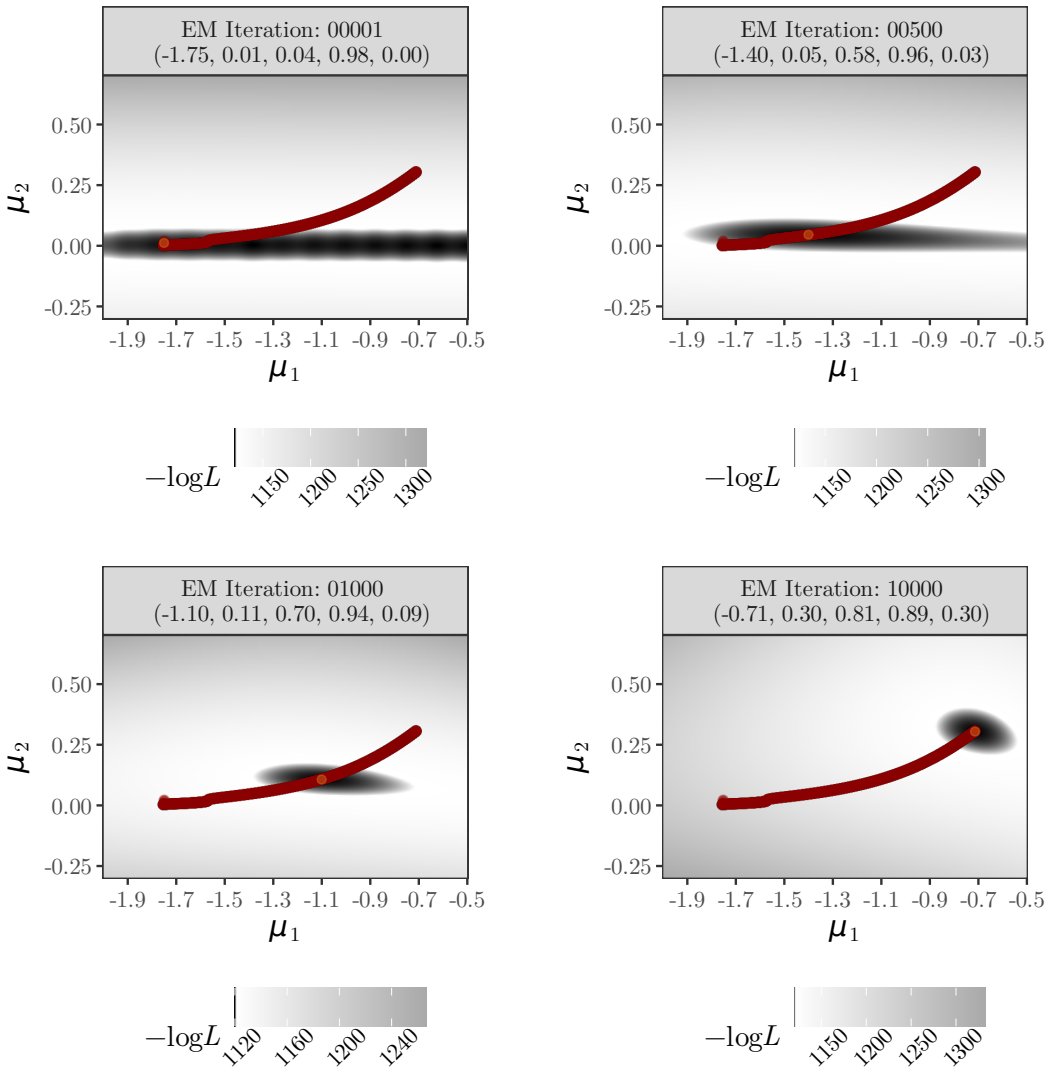

Abbildung 39: Mittelwerte im Verlauf der Optimierung für $k = 10$ bei $p = 3$

in relativ wenigen immer kleiner werdenden Schritten ein Optimum annähert, ist dieses Iterationsverhalten durchaus ungewöhnlich. Die Iterationsanzahl wurde hier bereits auf 10000 erhöht, trotzdem kommt es nicht zur Konvergenz. Es gibt keine Anzeichen, dass sich dies bei weiterer Erhöhung ändern könnte. Relevant für die Untersuchung der spezifischen Optima ist in diesem Zusammenhang, dass dieses Iterationsverhalten für einen ganzen Bereich von Startpunkten mit $1.5 < |\mu_1| < 2$ auftritt, und somit unabhängig von der Frage der Konvergenz für die zugehörigen Cluster keine spezifischen Optima erreicht werden.

Darüber hinaus kommt es in Abbildung 37 für $|\mu_1| < 1.5$ zwar zu überwiegend konvergierenden Optimierungsläufen, für $p = 5$ wird aber deutlich, dass auch im inneren Bereich nicht für alle Startcluster spezifische Optima erreicht werden können, da häufiger mehrere Läufe am gleichen Optimum enden. Dies gilt zum Beispiel für die Läufe mit $k = 77$ und $k = 78$, deren Optimierungsverläufe in Abbildung 40 dargestellt sind. Insgesamt werden hier von 133 konvergenten Läufen lediglich 71 verschiedene Optima erreicht, während es für $p = 5$ bei f_1 153 von 163 sind. Es ist also davon auszugehen, dass die einsetzende Asymptotik hier dem Erreichen spezifischer Optima für jedes Cluster entgegenwirkt.

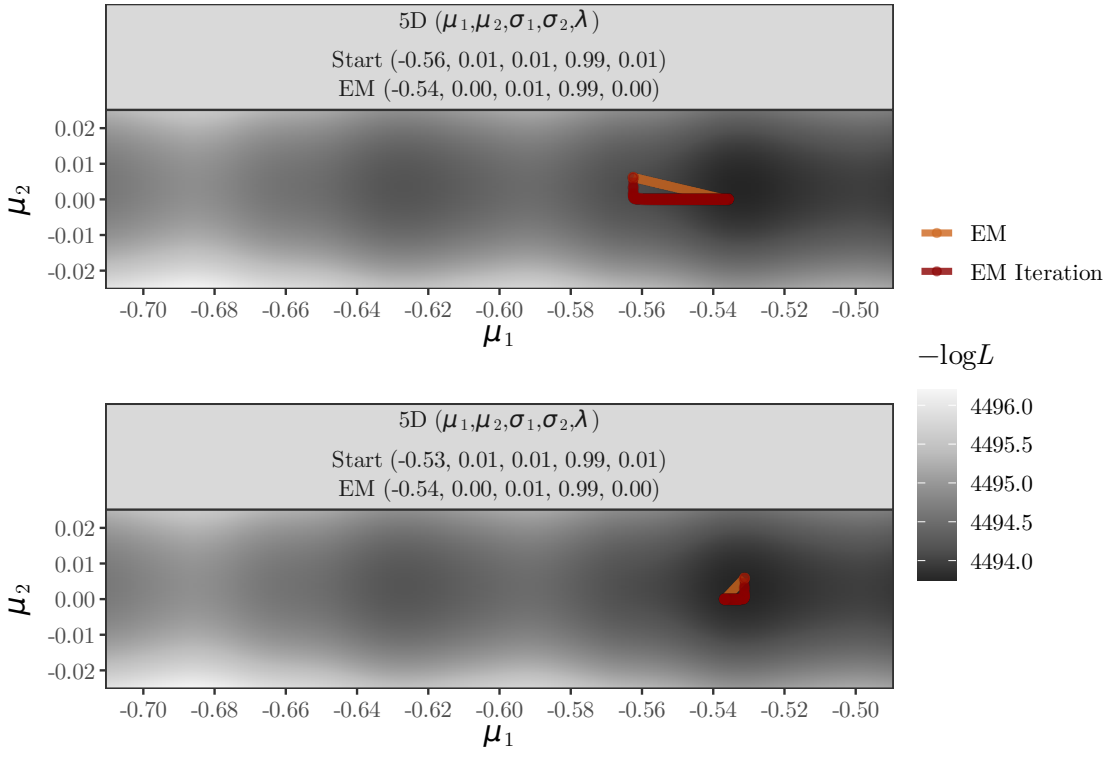

Abbildung 40: Optimierung für $k = 77$ (oben) und $k = 78$ (unten) bei $p = 5$

Für die aus gleichverteilten Beobachtungen zusammengesetzten Cluster ist es – ohne für wachsendes p asymptotisch gegen die Standardnormalverteilung zu konvergieren – also möglich, in einem relativ großen inneren Bereich für jedes Cluster ein spezifisches Optimum der Likelihood-Funktion zu erreichen. Im Fall mit asymptotischer Annäherung an die Normalverteilung können spezifische Cluster hingegen nur vereinzelt erreicht werden. Im folgenden Unterkapitel wird daher versucht durch Verwendung von normal- statt gleichverteilten Beobachtungen innerhalb der Cluster ein Normalverteilungsszenario weiter anzunähern.

4.3 Normalverteilte Clustergrößen mit normalverteilten Beobachtungen

Um eine Dichte mit Normalverteilungen innerhalb der Cluster konstruieren zu können, müssen die bei den Gleichverteilungen vorgegebenen Intervallgrenzen durch eine geeignete Wahl der Standardabweichung ersetzt werden. Diese soll zunächst mit jeder Erhöhung von p halbiert werden. Um nicht bereits im Fall von $p = 0$ mit einer Standardabweichung von 1 stark überlappende Cluster zu erzeugen ist ein zusätzlicher Verringerungsfaktor s_0 nötig. Im ersten Fall wird $s_0 = 2^{-3}$ gewählt. Für die Dichte aus K Normalverteilungen mit Mittelwerten $m_k = (k - \frac{K-1}{2} - 1) \cdot \frac{1}{2^p}$ und Standardabweichung $s = s_0 \cdot 2^{-p} = 2^{-(p+3)}$ gilt:

$$
\begin{aligned}
f_{3,p}(x) &= \sum_{k=1}^{K} \alpha_k \cdot f_k(x) \\
&= \sum_{k=1}^{K} \underbrace{\varphi\left(m_k\right) \cdot 2^{-p}}_{\text{Histogramm-Dichte}} \cdot f_{\mathcal{N}(m_k,s)}(x) \\
&= \sum_{k=1}^{K} \varphi\left(m_k\right) \cdot 2^{-p} \cdot \varphi\left(\frac{x - m_k}{s}\right) \cdot \frac{1}{s} \\
&= \sum_{k=1}^{K} \varphi\left(m_k\right) \cdot \varphi\left(\frac{x - m_k}{s}\right) \cdot 2^{-p} \cdot \frac{1}{s_0 \cdot 2^{-p}} \\
&= \sum_{k=1}^{K} \varphi\left(m_k\right) \cdot \varphi\left(\frac{x - m_k}{s}\right) \cdot \frac{1}{s_0} \\
&= \sum_{k=1}^{K} \varphi\left(m_k\right) \cdot \varphi\left(\frac{x - m_k}{s}\right) \cdot \frac{1}{2^{-3}}.
\end{aligned}
$$

Angelehnt an das Vorgehen für gleichverteilte Beobachtungen innerhalb der Cluster wird auch hier eine weitere Dichte konstruiert, bei der sich s_0 asymptotisch gegen

1 entwickelt und somit keine exakte Halbierung der Standardabweichung für jede Erhöhung von p erfolgt. Mit $s_0 = 1 - \sqrt[4]{\frac{1}{p+2}}$ ergibt sich analog zu $f_{3,p}$:

$$\begin{aligned} f_{4,p}(x) &= \sum_{k=1}^{K} \alpha_k \cdot f_k(x) \\ &= \sum_{k=1}^{K} \varphi(m_k) \cdot \varphi\left(\frac{x - m_k}{s}\right) \cdot \frac{1}{s_0} \\ &= \sum_{k=1}^{K} \varphi(m_k) \cdot \varphi\left(\frac{x - m_k}{s}\right) \cdot \underbrace{\frac{1}{1 - \sqrt[4]{\frac{1}{p+2}}}}_{\xrightarrow{p\to\infty} 1}. \end{aligned}$$

In Abbildung 41 ist $f_{3,p}$ für $p = 0, 1, 2$ dargestellt. Es ist wie bei den Gleichverteilungen in den Clustern zu erkennen, dass zwischen den einzelnen Maxima der Funktion immer Werte von nahezu 0 angenommen werden, während die Maxima für alle größeren Werte von p den gleichen Funktionswert beibehalten. Durch die exakte Halbierung der Standardabweichung bei Erhöhung von p ist der relative Abstand der beiden äußersten Beobachtungen zweier benachbarter Cluster unabhängig von p. Dadurch ist zwischen den Clustern auch asymptotisch immer ein Dichtewert von annähernd 0 vorhanden, sodass keine Annäherung an die Standardnormalverteilung stattfindet.

Für $f_{4,p}$ in Abbildung 42 hingegen ist deutlich zu erkennen, dass für steigendes p keine Funktionswerte von 0 zwischen den Clustern mehr angenommen werden und

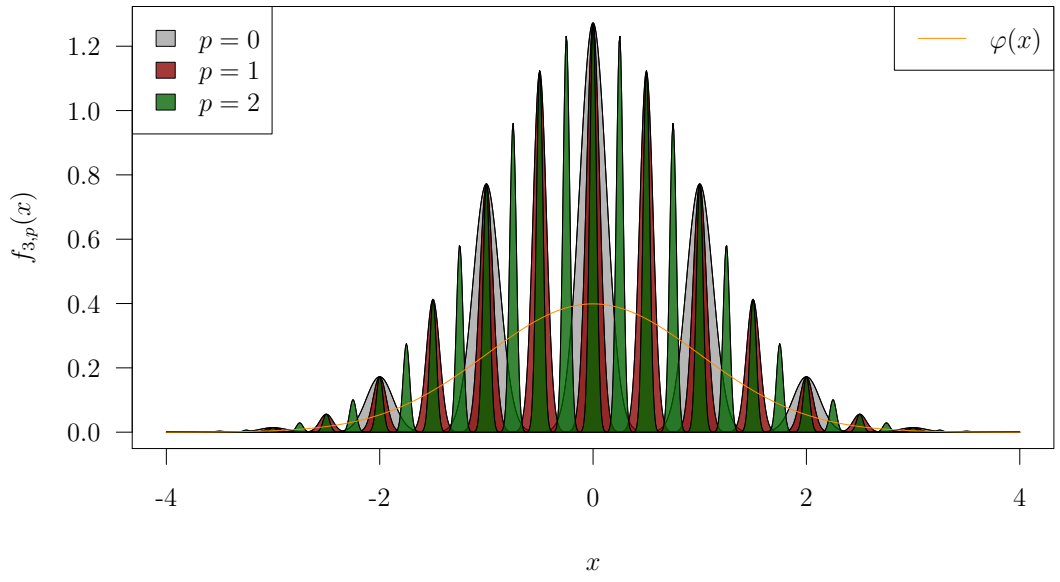

Abbildung 41: Dichte $f_{3,p}$ für $p = 0, 1, 2$

sich insgesamt die Maxima von oben und die Minima von unten an die Standardnormalverteilung annähern. Dies ist ein Unterschied zum vorherigen Unterkapitel. Die Maximalwerte der Kombination von Gleichverteilungen nähern sich dort ebenfalls von oben an die Dichte der Standardnormalverteilung an, die asymptotisch nicht mehr auftretenden Minima hingegen haben immer einen Funktionswert von 0 (vgl. Abb. 34, S. 63).

Der sehr viel glattere Verlauf der beiden zugehörigen Verteilungsfunktionen in Abbildung 43 ist dadurch zu erklären, dass hier Normalverteilungen verwendet wurden, während die Gleichverteilungen im vorherigen Unterkapitel deutlich erkennbare Stufen in den Verläufen verursachen. Darüber hinaus ist auch hier zu erkennen, dass die asymptotische Variante $\tilde{F}_{4,p}$ eine stärkere Angleichung an den Verlauf der Standardnormalverteilung besitzt, während die Variante ohne Asymptotik wie im Gleichverteilungs-Fall zwar mit steigendem p feiner segmentiert, jedoch immer Abweichungen zur Standardnormalverteilung bestehen bleiben.

Im Gegensatz zu den vorherigen Szenarien nehmen die Normalverteilungsdichten nie einen Wert von exakt 0 an, dementsprechend existiert auch kein Intervall zwischen zwei Clustern, auf dem die zusammengesetzte Dichte durchgängig 0 ist. Trotzdem kann untersucht werden, ob die Verteilungen sich asymptotisch annähern, indem statt der Intervallgrenzen der Gleichverteilungen symmetrische Intervalle um die Mittelwerte der Normalverteilungen betrachtet werden. Der Einfachheit halber wird daher das Intervall $[m_k - s/2, m_k + s/2]$ gewählt. Analog zu Kapitel 4.2 kann mithilfe dieser Intervalle untersucht werden, ob benachbarte Verteilungen sich

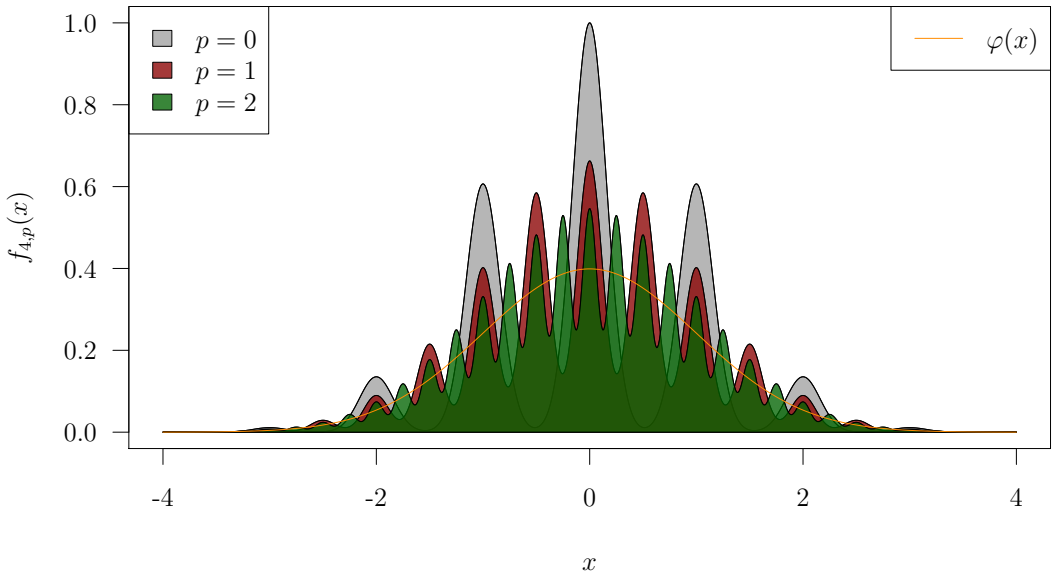

Abbildung 42: Dichte $f_{4,p}$ für $p = 0, 1, 2$

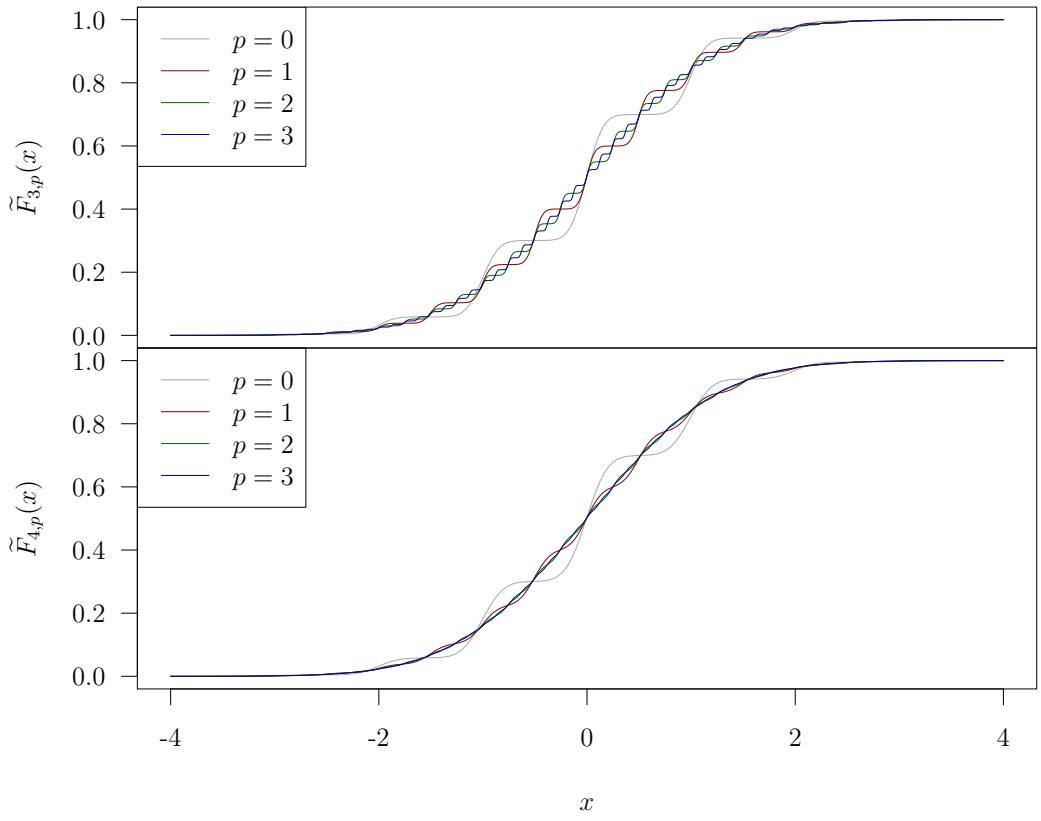

Abbildung 43: Approximierte Verteilungsfunktionen für $p = 0, 1, 2, 3$

asymptotisch annähern bzw. die Bereiche zwischen den Intervallen asymptotisch kleiner werden.

Für den Abstand zwischen zwei benachbarten Intervallen gilt dann mit Ersetzen von $z(p)$ durch $s/2$ in der allgemeinen Formel 4.1 (S. 64):

$$\begin{aligned} 2^{-p} - 2z(p) &= 2^{-p} - 2\frac{s}{2} \\ &= 2^{-p} - s_0 \cdot 2^{-p}. \end{aligned}$$

Für $f_{3,p}$ ergibt sich durch Einsetzen von s_0 eine Intervalllänge von

$$\begin{aligned} 2^{-p} - 2^{-3} \cdot 2^{-p} &= 2^{-p} - \frac{1}{8} \cdot 2^{-p} \\ &= \frac{7}{8} \cdot 2^{-p} \end{aligned}$$

und für $f_{4,p}$ eine Intervalllänge von

$$2^{-p} - \left(1 - \sqrt[4]{\frac{1}{p+2}}\right) \cdot 2^{-p}.$$

Multipliziert man hier ebenfalls die Intervalllängen mit der Anzahl vorhandener Cluster K, erhält man für $f_{3,p}$

$$K \cdot \frac{7}{8} \cdot 2^{-p} = K_0 \cdot 2^p \cdot \frac{7}{8} \cdot 2^{-p} = \frac{7}{8} K.$$

Das bedeutet 87.5 % des betrachteten Bereiches der Dichte liegen zwischen den Intervallen. Entscheidend ist hier, dass dies unabhängig von p der Fall ist, also keine asmyptotische Annäherung stattfindet.

Für $f_{4,p}$ ergibt sich im Gegensatz dazu

$$\begin{aligned} K \cdot \left(2^{-p} - \left(1 - \sqrt[4]{\frac{1}{p+2}}\right) \cdot 2^{-p}\right) &= K_0 \cdot 2^p \left(2^{-p} - \left(1 - \sqrt[4]{\frac{1}{p+2}}\right) \cdot 2^{-p}\right) \\ &= K_0 - K_0 \cdot \left(1 - \sqrt[4]{\frac{1}{p+2}}\right), \end{aligned}$$

wobei hier im Grenzwert gilt:

$$\lim_{p\to\infty} K_0 - K_0 \cdot \left(1 - \underbrace{\sqrt[4]{\frac{1}{p+2}}}_{\to 0}\right) = K_0 - K_0 = 0.$$

Wie bei den Gleichverteilungen (Kap. 4.2) bestätigt sich also, dass bei der zweiten Variante eine Annäherung der einzelnen Verteilungen stattfindet und die dünn besetzten Bereiche asymptotisch nicht mehr existieren. Asymptotisch grenzen benachbarte Intervalle mit Breite s und Mittelpunkt m_k unmittelbar aneinander. Da diese Intervalle nur etwa 38 % der Gesamtdichte der Verteilungen enthalten, wird also deutlich, dass es zu erheblichen Überlappungen der einzelnen Verteilungen kommt. Genau durch diese Überlappungen mehrerer Clusterverteilungen ist es überhaupt möglich, dass der glatte Verlauf der Standardnormalverteilungsdichte angenähert werden kann.

Analog zu den äquidistanten Beobachtungen für die Gleichverteilung, wird hier eine deterministische Anordnung der Beobachtungen verwendet, um konkrete Szenarien zu generieren. Dazu wird eine Folge von n_k äquidistanten Werten zwischen 0 und 1 verwendet, um die entsprechenden Quantile der Normalverteilung mit Mittelwert m_k und Standardabweichung s zu erhalten. Diese Quantilswerte dienen als perfekt verteilte Stichprobe der entsprechenden Normalverteilung. Diese Vorgehensweise wird in den folgenden Beispielen ebenfalls zur Bestimmung der Beobachtungsanzahlen der Cluster verwendet. Anstatt den (gerundeten) Dichtewert am Intervallmittelpunkt m_k zu verwenden, werden n perfekt standardnormalverteilte Beobachtungen generiert. Anschließend wird für jedes Histogramm-Intervall der Breite 2^{-p} die Clustergröße n_k als Anzahl der darin enthaltenen Beobachtungen bestimmt.

In Abbildung 44 sind die Optimierungsläufe für $f_{3,p}$ mit p von 0 bis 5 dargestellt. Ein grundsätzlicher Unterschied zu den vorherigen Szenarien ist, dass hier die unbesetzten Bereiche zwischen den einzelnen Clustern im mittleren, stärker besetzten Bereich kleiner sind als zwischen den äußeren Clustern. Dementsprechend handelt

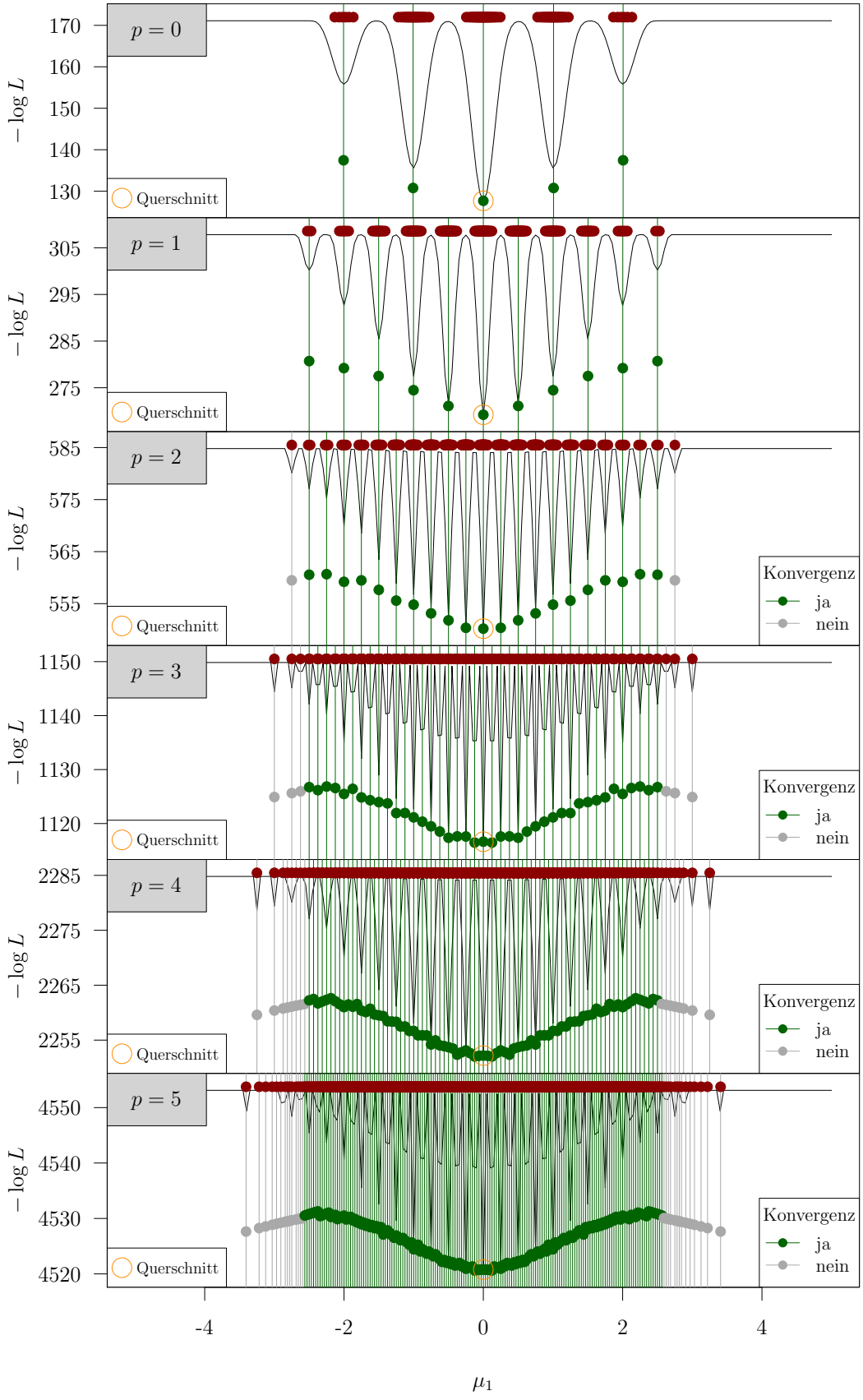

Abbildung 44: Fall 3: Normalverteilungen mit Halbierung für $n_0 = 100$

es sich hier um einen zusätzlichen Schritt der Annäherung an die Normalverteilung. Es ist zu erkennen, dass auch hier für einen großen Bereich in der Mitte für jedes Cluster ein spezifisches Optimum erreicht werden kann. Durch das Ersetzen der Gleichverteilungen durch Normalverteilungen, bei strikter Halbierung der Clusterbreiten, können auf Basis der Grafiken keine nennenswerten Veränderungen der Optimierungsergebinsse festgestellt werden (vgl. Abb. 36, S. 66). Betrachtet man die Anzahl konvergenter Läufe und voneinander verschiedener Ergebnisse, ist allerdings ein Unterschied feststellbar: während es für $f_{1,p}$ vereinzelt zu gleichen Ergebnissen bei den konvergenten Läufen kam, z. B. 155 verschiedene Ergebnisse bei 163 konvergenten Optimierungsläufen, stimmt in diesem neuen Szenario die Anzahl konvergenter Läufe immer exakt mit der Anzahl verschiedener Optima überein. Das bedeutet, dass trotz weiterer Annäherung an eine Datensituation aus gemischten Normalverteilungen mehr lokale Optima identifiziert werden können.

Für die Variante mit $s_0 \to 1$ in Abbildung 45 zeigt sich hingegen ein offensichtlicher Vorteil gegenüber der asymptotischen Variante im vorherigen Unterkapitel (vgl. 37, S. 67). Nicht-konvergente Optimierungsläufe, die nicht am Rand liegen und somit größere Lücken im Bereich der konvergenten Läufe hinterlassen, treten hier nicht auf. Stattdessen konvergieren die Läufe für dieses vierte Szenario auf einem großen inneren Bereich, vergleichbar mit dem vorherigen Fall ohne Asymptotik. Für $p = 5$ wird allerdings deutlich, dass wie im vorherigen Unterkapitel bei der asymptotischen Variante mit steigendem p häufiger gleiche Ergebnisse von verschiedenen Startclustern erzielt werden.

Da die grafische Darstellung für Werte von $p > 5$ nicht geeignet ist, sind in Abbildung 46 die Anteile konvergenter Läufe und verschiedener erreichter Optima für alle vier Szenarien bis $p = 10$ dargestellt. Es ist zu erkennen, dass in allen vier Fällen ab $p = 6$ mehr als 80 % der EM-Läufe konvergieren, insbesondere auch für $f_{2,p}$, wo die beobachteten nicht-konvergenten Läufe im inneren Bereich für die größeren Werte von p offensichtlich nicht mehr auftreten. Noch deutlicher fällt auf, dass die beiden asymptotischen Varianten nicht für alle konvergenten Läufe ein eigenes Optimum erreichen. Für niedrigere Werte von p ist dieses Problem für die Variante mit Gleichverteilung deutlicher, mit größerem p findet allerdings eine Annäherung auf einem Niveau von ca. 40 % statt. Das bedeutet, dass mit Asymptotik nur etwa jeder zweite konvergente Lauf ein bisher nicht erreichtes lokales Optimum liefert. Für die Varianten ohne Asymptotik bestätigt sich, dass bei Normalverteilungen innerhalb der Cluster auch bis $p = 10$ jeder konvergente Lauf ein spezifisches Optimum erreicht. Mit Gleichverteilungen konvergieren im Gegensatz dazu etwas mehr Läufe, jedoch existieren etwas weniger unterschiedliche Lösungen.

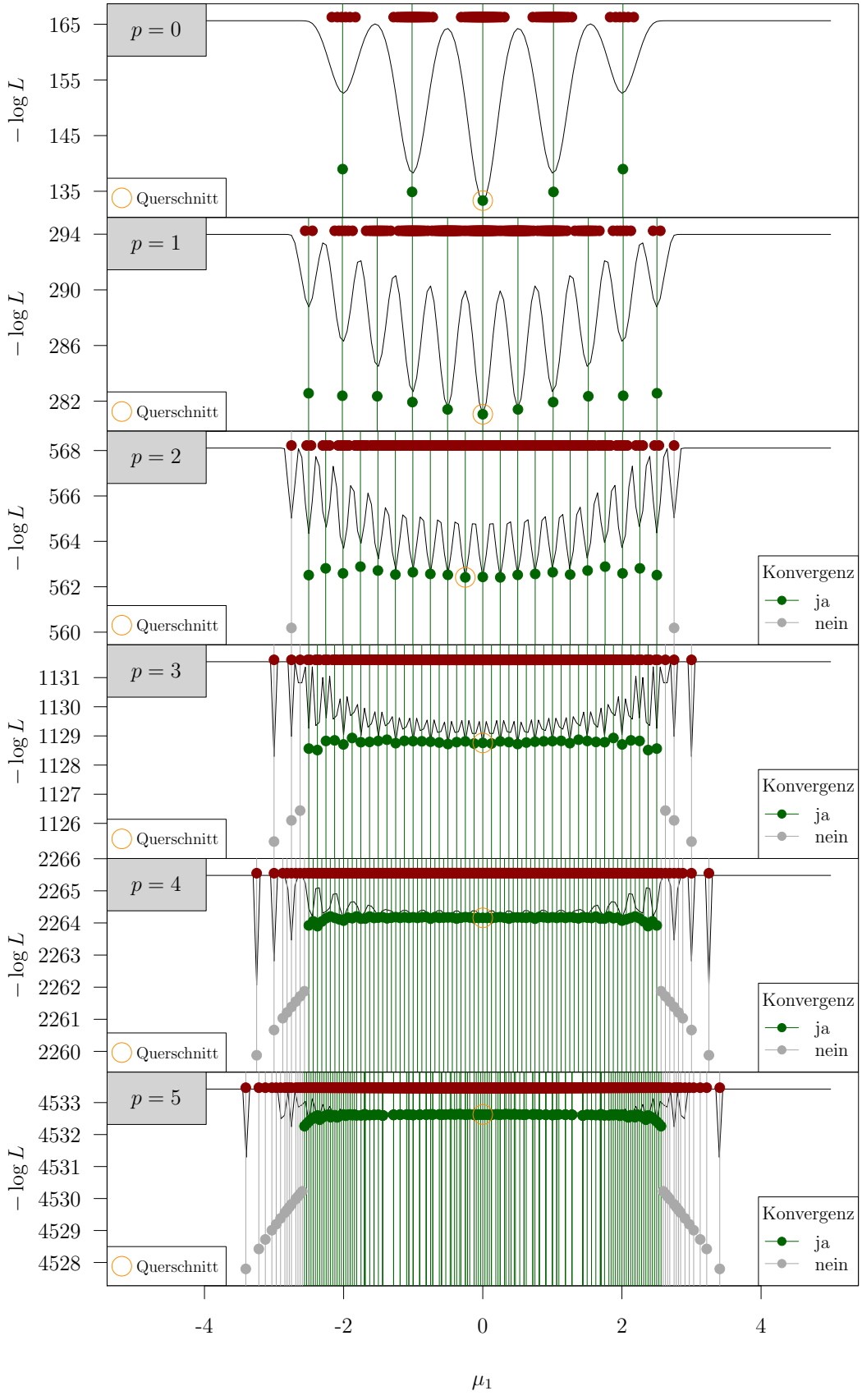

Abbildung 45: Fall 4: Normalverteilungen mit Asymptotik für $n_0 = 100$

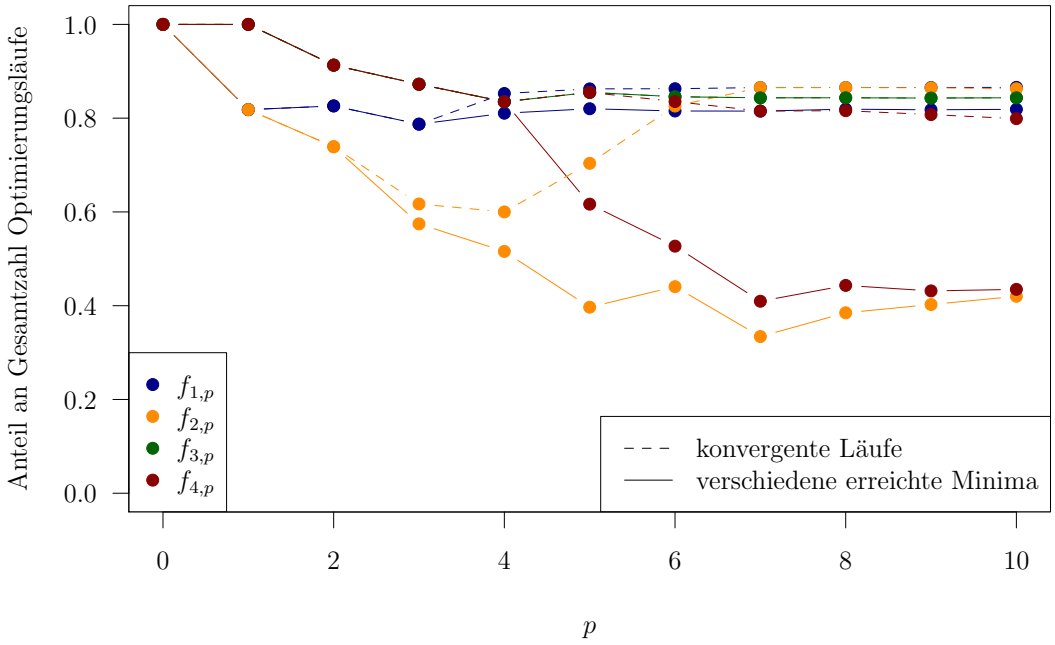

Abbildung 46: Anteil konvergenter und verschiedener Läufe in den vier Szenarien

Es kann also auch bei Verwendung von Clustern mit normalverteilten Beobachtungen bei einsetzender Asymptotik nicht grundsätzlich ein spezifisches Optimum für jedes Startcluster gefunden werden. Prinzipiell ist diese Erkenntnis auch nachvollziehbar, schließlich liegt die Annahme zugrunde, dass in den Beobachtungen einzelne Cluster identifizierbar sind. Wird insgesamt eine Normalverteilung angenähert, sind einzelne Cluster irgendwann nicht mehr erkennbar, auch wenn man die theoretischen Clustermittelpunkte in den hier betrachteten Szenarien kennt und an ihnen Optimierungsläufe starten kann. Letztlich ist dann ohnehin nur noch eine Mischungskomponente vorhanden, sodass die Modellierung eines Mischverteilungsmodells nicht notwendig bzw. zielführend ist.

Solange Cluster erkennbar sind, ist es also möglich für einen Großteil der vorhandenen Cluster spezifische Optima mit dem EM-Algorithmus zu erreichen. Der Analyse dieser kontrollierbaren Cluster-Szenarien liegt zugrunde, dass zufällig gruppierte Beobachtungen in realen Datensätzen auftreten, die lokale Optima hervorrufen. Daher soll in einem abschließenden Schritt die Anwendung auf normalverteilte Zufallsdaten aus zwei verschiedenen Verteilungskomponenten vorgenommen werden (Kap. 4.5). Als Zwischenschritt dorthin wird zunächst allerdings beispielhaft gezeigt, wie sich das Hinzufügen von Daten einer zusätzlichen Verteilungskomponente mit kleiner Varianz in den hier betrachteten kontrollierbaren Szenarien auswirkt.

4.4 Hinzufügen von Beobachtungen einer zusätzlichen Komponente

Ein weiterer Aspekt, der die vorherigen Szenarien trotz Annäherung an eine Normalverteilung von einem geeigneten Anwendungsszenario für ein Mischungsmodell mit zwei Komponenten unterscheidet, ist das Fehlen der zweiten Komponente. Die bisher betrachteten Situationen bestehen aus vielen Clustern, von denen jedes einzelne als Komponente mit kleiner Varianz betrachtet werden kann. Das reicht auch aus, um das Entstehen der lokalen Optima zu demonstrieren. Allerdings entstammen alle betrachteten Cluster, auch das jeweils als Komponente mit kleiner Varianz angenommene, aus der Komponente mit der größeren Varianz. Daher soll im Folgenden betrachtet werden, wie sich das Hinzufügen eines zusätzlichen Clusters mit kleiner Varianz als echte zusätzliche Komponente auswirkt. Dazu wird das vierte Szenario aus dem vorherigen Kapitel verwendet, wobei die neu hinzukommende Komponente analog zu den einzelnen Clustern ideal normalverteilt aus den Quantilen bestimmt wird. Der Mittelwert μ_1 gehört dabei, wie in allen bisherigen Analysen, zur Komponente mit der kleineren Varianz. Die zusätzliche Komponente wird folglich als erste Komponente bezeichnet. Desweiteren sind in den folgenden Beispielen immer $\sigma_1 = 0.01$ und $\lambda = 0.5$ bei $n_0 = 100$ in der zweiten Komponente mit $\mu_2 = 0$ und $\sigma_2 = 1$.

In Abbildung 47 ist das Hinzufügen von Beobachtungen einer Komponente mit $\mu_1 = 0$ dargestellt. Das ohnehin bereits größte Cluster mit $\mu_1 = 0$ wird dadurch zusätzlich mit der Gesamtzahl vorhandener Beobachtungen in allen bisherigen Clustern verstärkt. Die Beobachtungszahl in diesem Cluster wird also um ein Vielfaches erhöht. Dementsprechend unterscheidet sich das globale Optimum in seinem Funktionswert deutlich von den lokalen Optima, der Einzugsbereich bleibt jedoch klein. Dadurch werden für $p = 0$ und $p = 1$ weiterhin alle lokalen Optima gefunden, für $p = 2$ konvergieren allerdings die beiden Läufe von den direkt benachbarten Clustern zum globalen Optimum und die beiden äußersten Cluster führen zu einem nicht-konvergenten Lauf. Der Großteil der lokalen Optima wird weiterhin erreicht. Da diese initialen Fälle nicht auf gravierende Änderungen gegenüber den vorherigen Analysen hindeuten, wird auch aus Gründen der Übersichtlichkeit eine Betrachtung bis $p = 2$ als ausreichend angesehen.

Zusätzlich wird das Gesamtergebnis mit gleich vielen Wiederholungen des EM-Algorithmus an zufälligen Startpunkten verglichen. Dabei gibt es zwei verschiedene Varianten: Bei *Zufall A* werden alle Parameter zufällig bestimmt, die Mittelwerte im Bereich der vorhandenen Beobachtungen, die Standardabweichungen zwischen 0.001 und 1 und das Mischungsverhältnis zwischen 0 und 1. Bei *Zufall B* sind

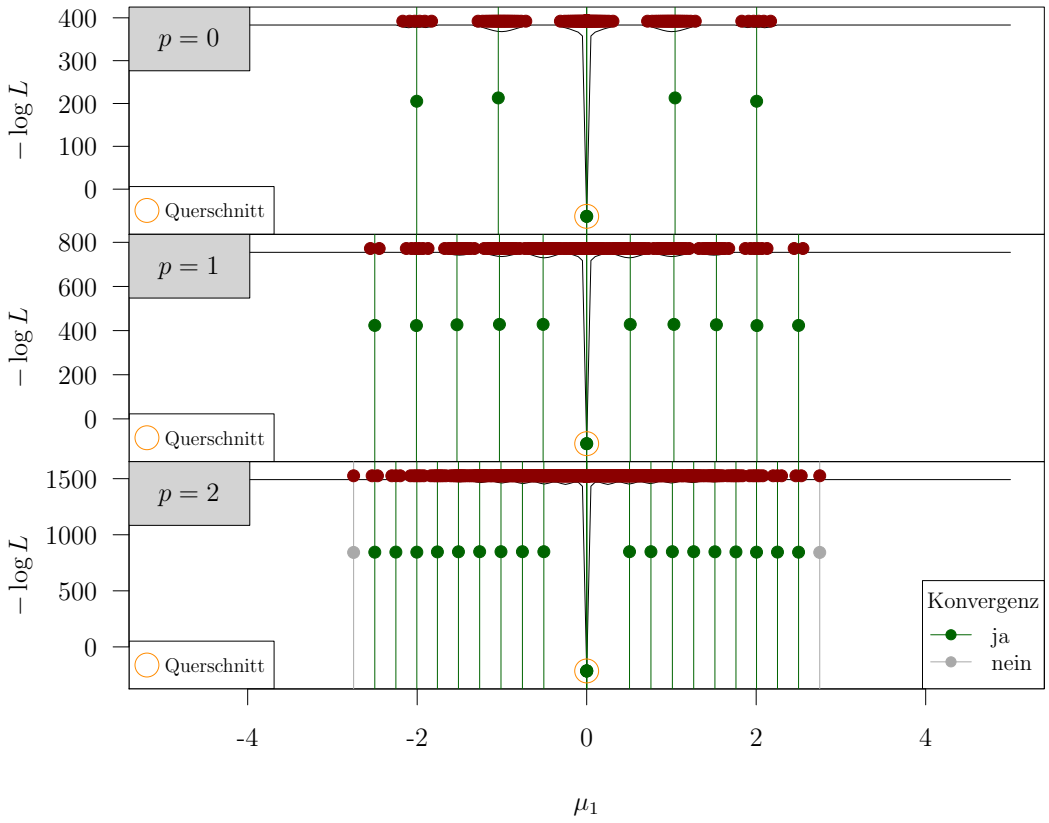

Abbildung 47: Fall 4 und zusätzliches Cluster mit $\mu_1 = 0$

$\mu_2 = 0$, $\boldsymbol{\sigma} = (0.1, 1)$ und $\lambda = 0.5$ fest und nur μ_1 wird aus dem Bereich der vorhandenen Beobachtungen zufällig bestimmt. *Zufall B* enthält daher zusätzliches Vorwissen und damit deutlich eingeschränkten Zufall. In Tabelle 9 ist dargestellt, wie viele der durchgeführten Läufe jeweils konvergieren bzw. das absolute Optimum erreichen und wie viele verschiedene lokale Optima insgesamt erreicht werden. Für $p = 0$ sind jeweils sechs Läufe angegeben, da sowohl einmal am zusätzlichen Cluster um 0 gestartet wird als auch einmal an dem bereits dort vorhandenen. Dementsprechend wird in diesem Beispiel der kleinste Wert auch mit der Cluster-Methode immer mindestens zweifach erreicht. Es ist zu erkennen, dass durch das Starten an den Clustermitten für steigendes p mehr lokale Optima gefunden werden können, als mit den zufälligen Startpunkten, im Fall $p = 3$ sogar fast viermal so viele.

In Abbildung 48 wird das zusätzliche Cluster an $\mu_1 = -1.5$ hinzugefügt. Es ergibt sich ein sehr ähnliches Bild zum vorherigen Beispiel, mit dem Unterschied, dass ab $p = 2$ zunehmend Läufe im Bereich von $-\mu_1$ zum zentralen Optimum konvergieren und eine Lücke hinterlassen. Zur Verdeutlichung werden hier, wie in den vorherigen Analysen, Werte bis $p = 5$ betrachtet. Die Erklärung für das Fehlen dieser lokalen Optima ist, dass für die entsprechenden Startpunkte offensichtlich eine Umkehrung

Tabelle 9: Fall 4 und zusätzliche Komponente mit $\mu_1 = 0$

$\boldsymbol{p = 0}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−63.4915	6	6	2	5
Zufall A	−63.4915	6	6	1	4
Zufall B	−63.4915	6	6	1	3
$\boldsymbol{p = 1}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−112.2346	12	12	2	11
Zufall A	−112.2346	12	12	3	6
Zufall B	−112.2346	12	12	2	6
$\boldsymbol{p = 2}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−214.4369	24	22	4	19
Zufall A	−214.4369	24	23	5	5
Zufall B	−214.4369	24	24	5	5

der Komponenten im Verlauf der Optimierung stattfindet: Die Parameter des kleinen Startclusters nähern sich der Standardnormalverteilung an, während die Parameter, die am Start aus den übrigen Beobachtungen bestimmt wurden, zu denen der neuen Komponente konvergieren. Die mit steigendem p zunehmende Häufung von $\mu_1 = 0$ ist also zusätzlich als Erreichen des absoluten Minimums zu betrachten. Darüber hinaus wird deutlich, dass auch für größeres p weiterhin ein großer Teil der lokalen Optima erreicht wird. Im Vergleich mit den beiden Zufallsstart-Varianten in Tabelle 10 zeigt sich, dass mit reinen Zufallsstartpunkten für höhere Werte von p nur sehr wenige verschiedene Optima identifiziert werden können (13 von 194 Läufen bei $p = 5$). Die stark eingeschränkten Zufallsstartpunkte erreichen in allen Fällen mindestens in jedem dritten Lauf ein neues lokales Optimum und sind damit deutlich näher an den Clusterstartpunkten, die immer für mehr als 50 % der Läufe ein eigenes Optimum erreichen. Die beiden Zufallsvarianten erreichen für höheres p dagegen fast doppelt so häufig das absolute Minimum.

Im dritten Beispiel wird die zusätzliche Komponente mit $\mu_1 = -3$ sehr weit außen hinzugefügt. Bezüglich des Erreichens der lokalen Optima sind in der zugehörigen Abbildung 49 keine qualitativen Unterschiede zum vorherigen Beispiel festzustellen, allerdings liegt der Bereich der Läufe, in dem keine spezifischen Optima erreicht werden hier nicht auf der anderen Seite der größeren Verteilung ganz außen, sondern innen in der Nähe von 0. Da im vorherigen Beispiel -1.5 zentral in der linken Hälfte der größeren Verteilung liegt, war diese offensichtlich inverse Spiegelung auf der

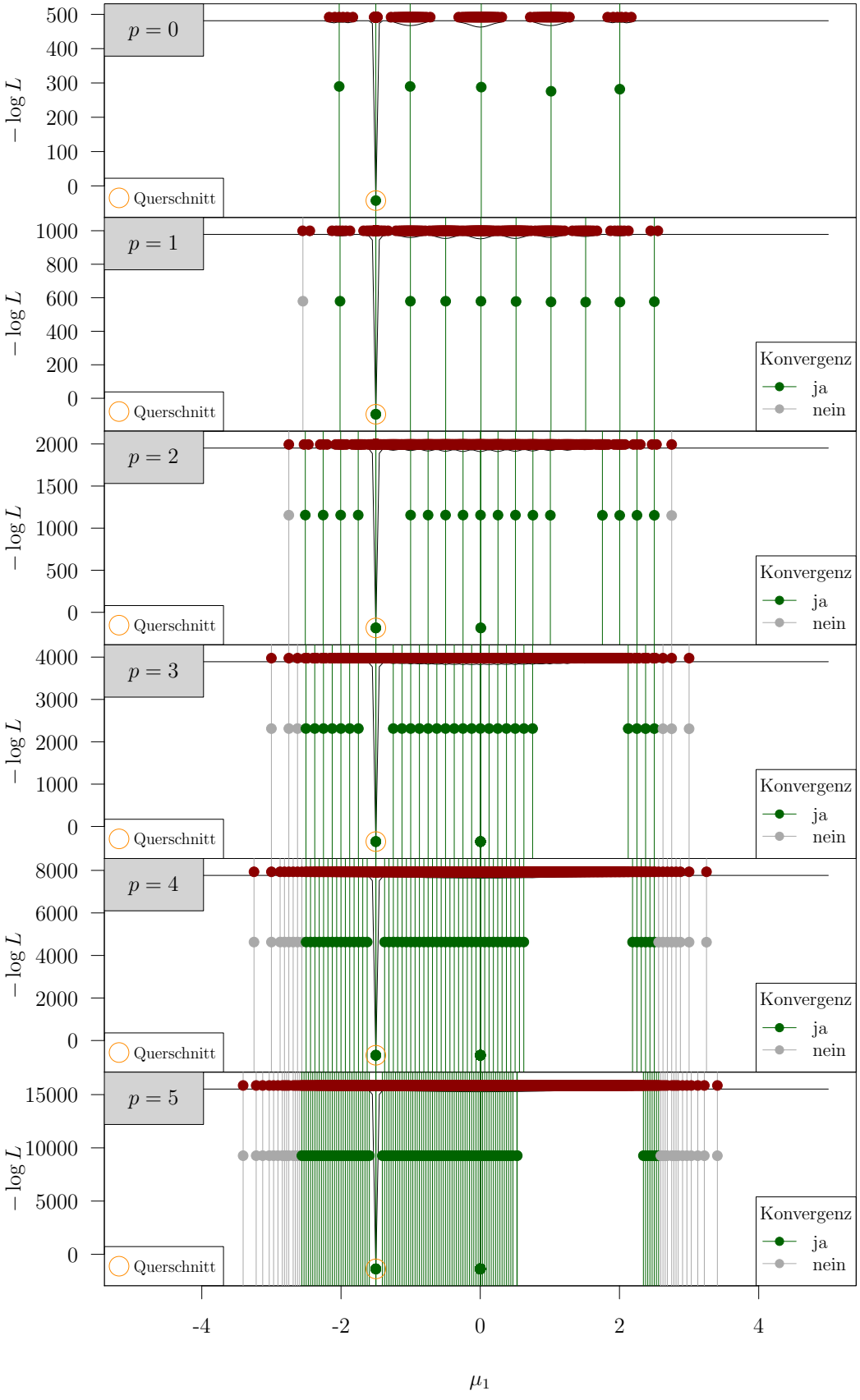

Abbildung 48: Fall 4 und zusätzliches Cluster mit $\mu_1 = -1.5$

Tabelle 10: Fall 4 und zusätzliches Cluster mit $\mu_1 = -1.5$

$\boldsymbol{p = 0}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−42.849	6	6	1	6
Zufall A	−42.849	6	6	2	5
Zufall B	−42.849	6	6	1	4
$\boldsymbol{p = 1}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−95.1963	12	11	2	10
Zufall A	−95.1963	12	12	8	5
Zufall B	−95.1963	12	12	1	8
$\boldsymbol{p = 2}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−185.1969	24	22	3	19
Zufall A	−185.1969	24	23	6	2
Zufall B	−185.1969	24	24	11	11
$\boldsymbol{p = 3}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−356.9582	48	42	10	30
Zufall A	−356.9582	48	47	24	3
Zufall B	−356.9582	48	48	23	24
$\boldsymbol{p = 4}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−692.6706	98	82	24	56
Zufall A	−692.6706	98	97	45	9
Zufall B	−692.6706	98	98	50	37
$\boldsymbol{p = 5}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−1372.1561	194	166	57	105
Zufall A	−1372.1561	194	193	97	13
Zufall B	−1372.1561	194	194	90	77

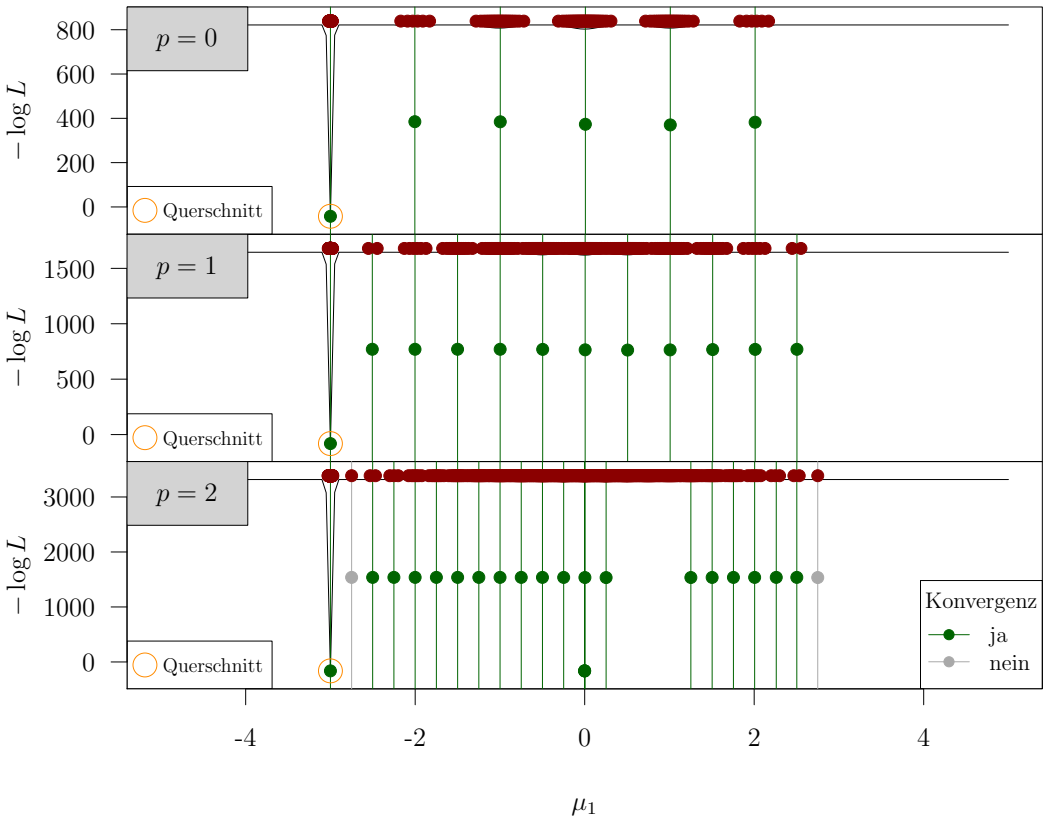

Abbildung 49: Fall 4 und zusätzliches Cluster mit $\mu_1 = -3$

Tabelle 11: Fall 4 und zusätzliches Cluster mit $\mu_1 = -3$

$\boldsymbol{p = 0}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−42.1637	6	6	1	6
Zufall A	−42.1637	6	6	3	3
Zufall B	−42.1637	6	6	3	3
$\boldsymbol{p = 1}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−81.2367	12	12	1	12
Zufall A	−81.2367	12	12	7	3
Zufall B	−81.2367	12	12	2	8
$\boldsymbol{p = 2}$	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−162.1708	24	22	3	20
Zufall A	−162.1708	24	23	13	4
Zufall B	−162.1708	24	24	14	11

anderen Seite der Verteilung dort noch nicht zu erkennen. Auch der Vergleich mit den Zufallsstartpunkten in Tabelle 11 stimmt in den hier betrachteten Beispielen bis $p = 2$ weitgehend mit dem des vorherigen Beispiels überein.

Insgesamt lässt sich feststellen, dass durch das Hinzufügen von Beobachtungen einer zusätzlichen, gleichmächtigen Komponente, zwar ein im Funktionswert deutlich dominierendes Optimum entsteht, die Erreichbarkeit für einen Großteil der lokalen Optima allerdings gewährleistet ist. Auch im Vergleich mit Zufallsstartpunkten, zeigt sich, dass das Starten an Clustern in den Beobachtungen hilfreich sein kann, um möglichst viele lokale Optima der Likelihood-Funktion zu erreichen.

Ob diese Aussage für den praktischen Anwendungsfall, dass die Beobachtungen der beiden Komponenten aus normalverteilten Zufallsstichproben bestehen, Bestand hat, soll im Folgenden überprüft werden.

4.5 Anwendung auf normalverteilte Daten

Der gesamten bisherigen Betrachtung von aus Clustern zusammengesetzten Stichproben lag die Annahme zugrunde, dass in realen Datensätzen zufällige Cluster aus einzelnen Beobachtungen bestehen. Das bisherige Vorgehen soll nun auf den praxisrelevanten Fall normalverteilter Zufallsdaten aus zwei Komponenten übertragen werden. Das Auffinden lokaler Optima der Likelihood-Funktion soll dabei möglichst vergleichbar zu den vorherigen konstruierten Szenarien erfolgen. Der grundlegende Unterschied ist jedoch, dass keine deterministischen Beobachtungen vorliegen. Daher müssen für jede Realisierung einer solchen Verteilungsmischung vorab einzelne Cluster identifiziert werden, von denen aus dann die lokalen Optima erreicht werden sollen. Um diese Vorbereitung der Modellierung nicht aufwändiger zu machen als das Mischverteilungsmodell selbst, wird eine simple Regel zur Erkennung von zufällig entstandenen Clustern in der eindimensionalen Stichprobe mit n Beobachtungen gewählt: Zwei benachbarte Beobachtungen werden unterschiedlichen Clustern zugeordnet, wenn der Abstand zwischen ihnen größer als die Spannweite aller Beobachtungen geteilt durch n ist, andernfalls gehören sie zum selben Cluster. Bei komplett äquidistanten Beobachtungen würde dementsprechend nur ein Cluster mit allen Beobachtungen gebildet, da der Mindestabstand an keiner Stelle überschritten wird. Mithilfe dieser Vorverarbeitung der Beobachtungen kann die Analyse aus dem vorherigen Kapitel analog angewendet werden. Die Komponente mit größerer Varianz hat auch hier $\mu_2 = 0$ und $\sigma_2 = 1$ als Parameter. Außerdem gilt in den folgenden Beispielen $\mu_1 = -1.5$ und $\lambda = 0.5$, während σ_1 und die Anzahl der Beobachtungen n variiert werden.

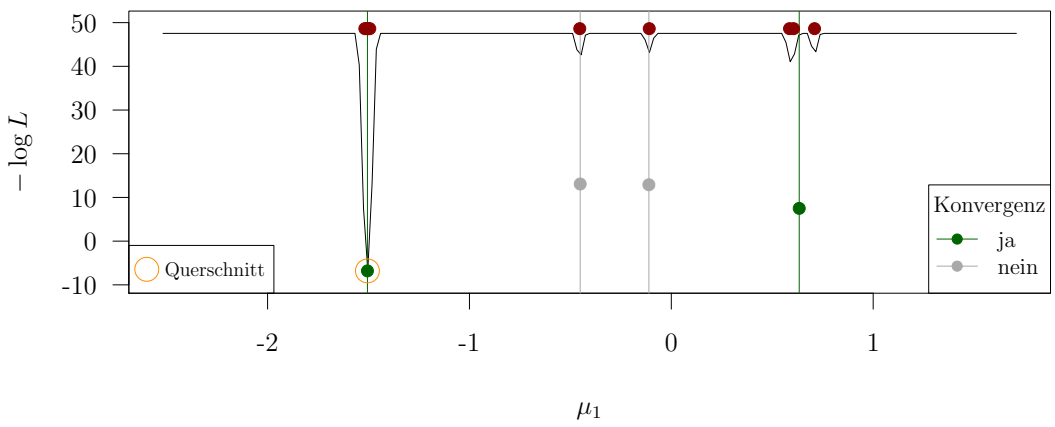

Abbildung 50: $n = 10$, $\sigma_1 = 0.01$

Für das betrachtete Beispiel mit $n = 10$ und $\sigma_1 = 0.01$ werden vier Cluster identifiziert. Die Ergebnisse der zugehörigen EM-Läufe sind in Abbildung 50 dargestellt. Es ist zu erkennen, dass zwei Cluster jeweils nur aus einer Beobachtung bestehen und die Optimierung dementsprechend fehlschlägt. Die drei weiteren Beobachtungen aus der zweiten Komponente bilden ein gemeinsames Cluster, für das ein spezifisches Optimum erreicht wird. Das deutlichere Optimum liegt jedoch für $\mu_1 = -1.5$ vor, da sich dort die fünf Beobachtungen aus der ersten Komponente befinden. In der zugehörigen Tabelle 12 ist zu erkennen, dass für die beiden Varianten der Zufallstartpunkte jeweils alle vier Läufe konvergieren, dabei erreicht die vollständig zufällige Variante *Zufall A* zweimal das Minimum und zwei weitere lokale Optima, während die Variante mit zufälligem Startwert für μ_1 viermal den niedrigsten Wert erreicht.

Im zweiten Beispiel wird σ_1 auf 0.1 erhöht. In Abbildung 51 ist zu sehen, dass hier mit den Startpunkten auf Basis der identifizierten Cluster ebenfalls vier EM-Läufe durchgeführt werden, von denen zwei nicht konvergieren und die übrigen zwei verschiedene Optima erreichen. Dabei fällt auf, dass die einzelne Beobachtung mit einem Wert von ca. 0.1 dem Cluster mit zwei Beobachtungen bei ca. -0.2 zugeordnet wird, da der Abstand kleiner als ein Zehntel der Spannweite der Beobachtungen

Tabelle 12: $n = 10$, $\sigma_1 = 0.01$

	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	-6.8062	4	2	1	2
Zufall A	-6.8062	4	4	2	3
Zufall B	-6.8062	4	4	4	1

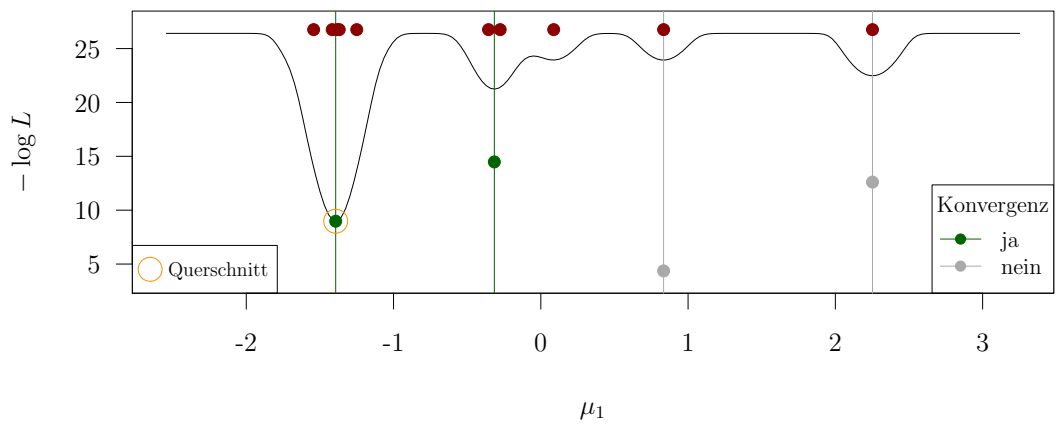

Abbildung 51: $n = 10$, $\sigma_1 = 0.1$

von ca. 4 ist. Dementsprechend ist allerdings auch sichergestellt, dass die fünf Beobachtungen der Komponente mit kleiner Varianz als ein Cluster erkannt werden, obwohl sie breiter streuen als das für die entsprechende Komponente im vorherigen Beispiel der Fall war. Durch diese Vergrößerung von σ_1, ergibt sich auch für den Vergleich mit den Zufallsstartpunkten in Tabelle 13 ein ähnliches Bild: Lediglich ein Lauf von *Zufall A* verhält sich anders und konvergiert nicht. An der Anzahl erreichter verschiedener Optima ergeben sich keine Änderungen.

Um einen etwas differenzierteren Vergleich zu ermöglichen, wird in einem weiteren Schritt die Beobachtungszahl auf $n = 100$ erhöht, wodurch sich mehr Cluster und damit mehr Optimierungsläufe für den Vergleich ergeben sollen. Wie in Abbildung 52 zu erkennen ist, gelingt dies für $\sigma_1 = 0.01$, wobei mehrheitlich Cluster aus Einzelbeobachtungen entstehen, von denen aus die Optimierung scheitert. Der Vergleich in Tabelle 14 zeigt allerdings, dass alle 10 konvergierenden EM-Läufe unterschiedliche Optima erreichen, was den Zufallsstart-Varianten in 7 von 26 bzw. 9 von 28 Fällen gelingt. Die Quote ist also deutlich besser, die absolute Anzahl jedoch nur geringfügig.

Für die Erhöhung auf $\sigma_1 = 0.1$ ergeben sich erneut kaum Unterschiede. In Abbildung 53 ist lediglich zu erkennen, dass um das globale Optimum herum in einem

Tabelle 13: $n = 10$, $\sigma_1 = 0.1$

	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	8.9747	4	2	1	2
Zufall A	8.9747	4	3	1	3
Zufall B	8.9747	4	4	4	1

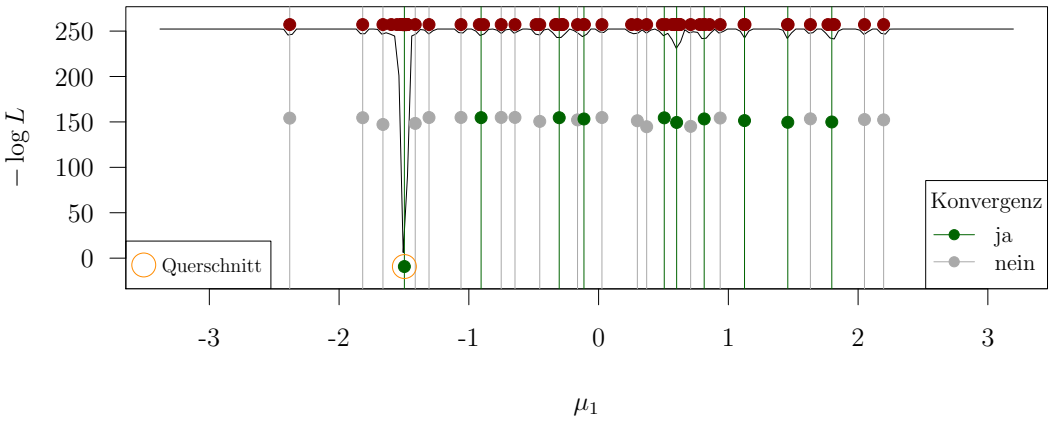

Abbildung 52: $n = 100$, $\sigma_1 = 0.01$

Tabelle 14: $n = 100$, $\sigma_1 = 0.01$

	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−9.2412	28	10	1	10
Zufall A	−9.2412	28	26	11	7
Zufall B	−9.2412	28	28	4	9

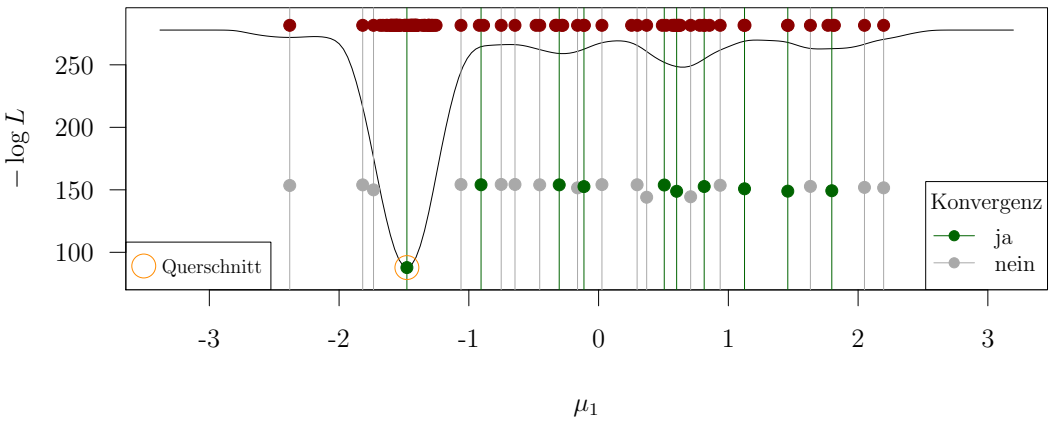

Abbildung 53: $n = 100$, $\sigma_1 = 0.1$

etwas größeren Bereich keine Einzelcluster identifiziert werden. In Tabelle 15 zeigt sich, dass erneut 10 konvergierende Läufe 10 verschiedene Optima erreichen, für die Zufallsstartpunkte sind es mit 6 bzw. 7 noch einmal leicht weniger als zuvor.

Als nächstes soll überprüft werden, ob sich der leichte Vorteil der Clustermethode beim Auffinden möglichst vieler lokaler Maxima für eine Erhöhung von n auf 1000 bestätigt. Anhand der grafischen Darstellungen in den Abbildungen 54 und 55 für beide Werte von σ_1 lässt sich aufgrund der Vielzahl entstandener Optimierungsläufe nur noch erahnen, dass der Anteil konvergierender Läufe in etwa mit den vorherigen Beispielen übereinstimmen könnte. Die Vergleichstabellen offenbaren, dass die verschiedenen ermittelten Anzahlen von Läufen sehr nahe am Zehnfachen der Beispiele mit $n = 100$ liegen, einzig die Anzahl verschiedener erreichter Optima für

Tabelle 15: $n = 100$, $\sigma_1 = 0.1$

	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	87.7925	26	10	1	10
Zufall A	87.7925	26	24	9	6
Zufall B	87.7925	26	25	5	7

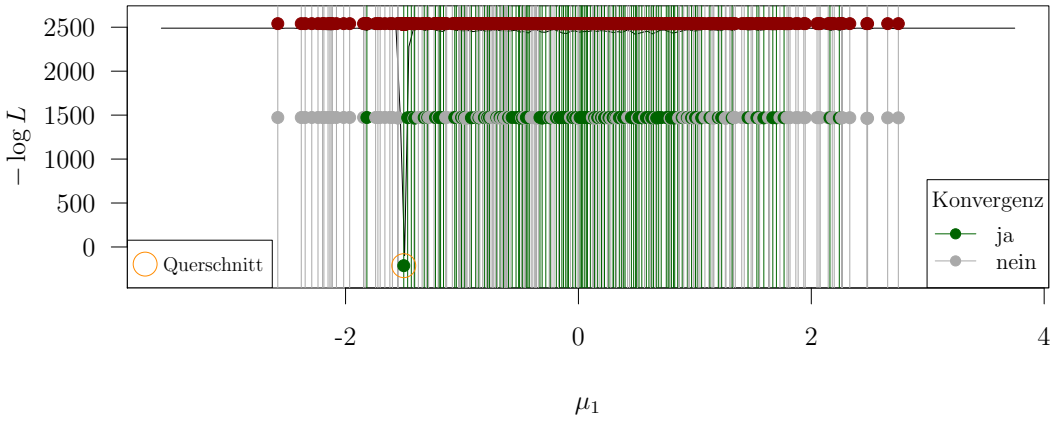

Abbildung 54: $n = 1000$, $\sigma_1 = 0.01$

Tabelle 16: $n = 1000$, $\sigma_1 = 0.01$

	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	−212.7104	251	104	1	103
Zufall A	−212.7104	251	249	113	21
Zufall B	−212.7104	251	251	89	72

die vollständigen Zufallsstartpunkte (*Zufall A*) weicht in beiden Fällen deutlich ab. Für $\sigma_1 = 0.01$ in Tabelle 16 sind es 21 in 249 Läufen und für $\sigma_1 = 0.1$ in Tabelle 17 17 von 242, was in etwa einem Drittel des Anteils bei $n = 100$ entspricht. Dementsprechend scheint sich hier ein deutlicher Vorteil der Clusterstartpunkte gegenüber den reinen Zufallsstartpunkten zu ergeben. Darüber hinaus zeigt sich auch für die Erhöhung von σ_1 bei $n = 1000$ erneut eine leichte Verbesserung gegenüber beiden Zufallsstart-Varianten.

Um diesen Eindruck unabhängig von Einzelbeispielen zu überprüfen, wird abschließend eine Simulation der sechs betrachteten Fälle mit jeweils 1000 Wiederholungen durchgeführt. In Tabelle 18 ist jeweils das arithmetische Mittel der verschiedenen Laufanzahlen enthalten. Für $n = 10$ und $n = 100$ ist zu erkennen, dass die Gesamtzahl durchgeführter Läufe und damit die Zahl vorab ermittelter Cluster in den Stichproben sehr genau mit den vorherigen Beispielen übereinstimmt. Für die beiden Fälle mit $n = 1000$ sind es jeweils ca. 30 Läufe weniger als in den Einzelbeispielen. Für die Clustermethode zeigt sich auch mit 1000 Wiederholungen, dass fast jeder konvergente EM-Lauf zu einem eigenen lokalen Optimum führt. Bei mehr als 100 konvergenten Läufen in den Fällen mit $n = 1000$ gibt es im Schnitt weniger als 2 Läufe, die das gleiche Optimum erreichen. Bei den geringeren Beobachtungs-

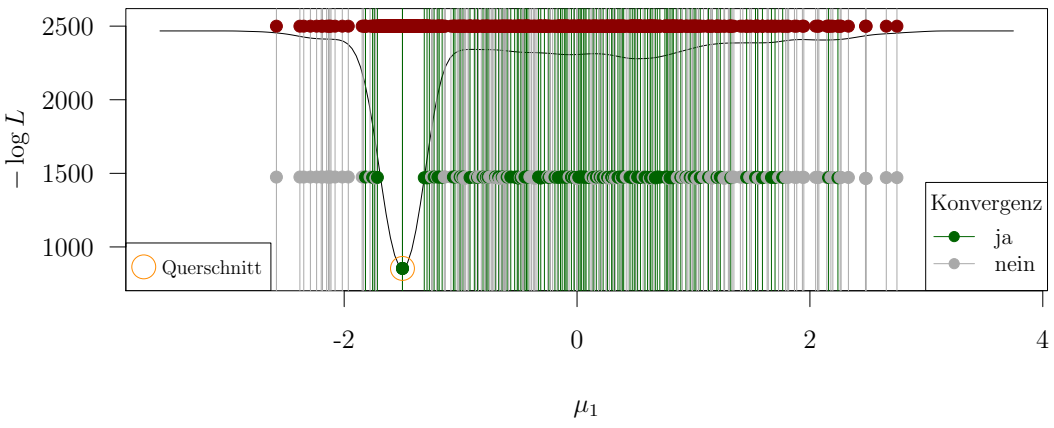

Abbildung 55: $n = 1000$, $\sigma_1 = 0.1$

Tabelle 17: $n = 1000$, $\sigma_1 = 0.1$

	Optimum	Läufe	konvergent	optimal	verschiedene Optima
Cluster	853.9985	244	107	3	104
Zufall A	853.9985	244	242	102	17
Zufall B	853.9985	244	244	48	55

Tabelle 18: Gemittelte Simulationsergebnisse mit 1000 Wiederholungen

$\boldsymbol{n = 10, \sigma_1 = 0.01}$	Läufe	konvergent	optimal	verschiedene Optima
Cluster	4.314	2.086	1.057	1.978
Zufall A	4.314	3.423	1.652	2.349
Zufall B	4.314	2.170	1.378	1.474
$\boldsymbol{n = 10, \sigma_1 = 0.1}$	Läufe	konvergent	optimal	verschiedene Optima
Cluster	4.196	2.003	1.033	1.937
Zufall A	4.196	3.329	1.668	2.277
Zufall B	4.196	2.165	1.391	1.508
$\boldsymbol{n = 100, \sigma_1 = 0.01}$	Läufe	konvergent	optimal	verschiedene Optima
Cluster	27.949	11.516	1.025	11.415
Zufall A	27.949	26.401	11.654	5.514
Zufall B	27.949	24.704	6.009	7.807
$\boldsymbol{n = 100, \sigma_1 = 0.1}$	Läufe	konvergent	optimal	verschiedene Optima
Cluster	27.036	11.655	1.055	11.518
Zufall A	27.036	25.679	11.143	5.376
Zufall B	27.036	24.405	6.582	7.517
$\boldsymbol{n = 1000, \sigma_1 = 0.01}$	Läufe	konvergent	optimal	verschiedene Optima
Cluster	224.749	104.009	1.030	103.019
Zufall A	224.749	219.756	97.915	23.597
Zufall B	224.749	214.783	70.889	57.250
$\boldsymbol{n = 1000, \sigma_1 = 0.1}$	Läufe	konvergent	optimal	verschiedene Optima
Cluster	215.24	105.666	1.699	103.948
Zufall A	215.24	210.869	95.508	20.446
Zufall B	215.24	205.608	61.370	47.574

zahlen beträgt die Abweichung bei deutlich weniger konvergenten Läufen ca. 0.1, d.h. nur in etwa jeder zehnten Wiederholung kommt es vor, dass zwei Läufe zum gleichen Optimum konvergieren. Außerdem geht mit der Verringerung der Anzahl Läufe für die Clusterstartpunkte keine entsprechende Verringerung der Anzahl konvergenter Läufe gegenüber dem vorausgegangenen Beispiel einher, während für die beiden Zufallsstart-Varianten ein entsprechender Rückgang zu erkennen ist.

Auch die beobachteten Zusammenhänge zwischen den Parametern und der Anzahl gefundener Optima bestätigen sich in der Simulation. Während die vollständige Zufallsstartmethode *Zufall A* für $n = 10$ mit mehr als zwei gefundenen Optima am besten abschneidet, werden für $n = 100$ nur noch ca. die Hälfte der mehr als 11 gefundenen Optima der Clustermethode gefunden und damit auch weniger als mit *Zufall B*, welches auf ca. zwei Drittel kommt. Dieser Trend setzt sich für $n = 1000$ fort: Hier erreicht *Zufall A* in beiden Fällen weniger als 25 % der mehr als 100 gefundenen Optima der Clustermethode, während es für *Zufall B* etwa die Hälfte ist. Die auf μ_1 beschränkten Zufallsstartpunkte erzielen demnach mit steigendem n deutlich höhere Anzahlen verschiedener Optima als die vollständigen Zufallspunkte, verschlechtern sich aber ebenfalls im Vergleich mit den Clusterstartpunkten. Eine leicht verringerte Anzahl gefundener Optima beider Zufallsmethoden bestätigt sich auch für die Erhöhung von σ_1 von 0.01 auf 0.1 bei den höheren Beobachtungszahlen. Im Gegensatz dazu erhöht sich die entsprechende Anzahl bei den Clusterstartpunkten für $n = 100$ von 11.42 auf 11.52 und bei $n = 1000$ von 103.02 auf 103.95 sogar geringfügig.

Insgesamt bestätigt sich mit diesem abschließenden Teil der Analyse, dass auch für normalverteilte Zufallsdaten aus zwei Komponenten mit Startpunkten auf Basis von Clustern in den Beobachtungen verschiedene lokale Optima einer multimodalen Likelihood-Funktion mit dem EM-Algorithmus erreicht werden können. Dabei zeigt sich im betrachteten Fall mit starker Multimodalität, dass mit steigender Beobachtungszahl deutlich mehr voneinander verschiedene Lösungen erreicht werden können als mit gleich vielen Zufallsstartpunkten für den EM-Algorithmus.

Da allgemein kein Verfahren existiert, das Konvergenz zum globalen Optimum garantiert, kann es grundsätzlich sinnvoll sein, möglichst viele lokale Optima miteinander zu vergleichen. Zum Anpassen von Mischverteilungsmodellen bietet sich mit den Clusterstartpunkten für EM eine in der Praxis anwendbare Methode, mit der das Auffinden möglichst vieler Optima besser gelingt als mit Zufallsstartpunkten.

5 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde die Multimodalität von Likelihood-Funktionen in Mischverteilungsmodellen aus unterschiedlichen Perspektiven betrachtet und analysiert. In Kapitel 2 wurden, neben einer grafischen Betrachtung von Mischungen verschiedener Verteilungen, Heuristiken verwendet, um die Komplexität und den Einfluss der Mischungsparameter darauf zu beschreiben. In Kapitel 3 wurden Vergleichsstudien zur Optimierung der betrachteten Likelihood-Funktionen mit verschiedenen Algorithmen durchgeführt, während in Kapitel 4 auf Basis von aus Clustern zusammengesetzten Datensituationen ein Vorgehen zum Auffinden möglichst vieler verschiedener lokaler Optima der Likelihood von Normalverteilungsmischungen entwickelt wurde.

Nach einer allgemeinen Einführung in das Mischverteilungsmodell (Kap. 2.1) wurden in Kapitel 2.2 die konkreten Auswirkungen des bekannten Auftretens von Multimodalität bei Mischung zweier Normalverteilungskomponenten mit unterschiedlichen Varianzen grafisch dargestellt. Es wurde deutlich, dass die zugehörige Likelihood-Funktion, je nach Verhältnis der Komponentenvarianzen zueinander, im Raum der Mittelwertparameter einen hohen Grad an Multimodalität aufweisen kann und damit offensichtlich Optimierungsprobleme von hoher Komplexität entstehen. Über die Darstellung der bekannten Problematik für Normalverteilungen (vgl. Day, 1969) hinaus konnte ein entsprechendes Auftreten der Multimodalität auch für Mischungen weiterer Verteilungen erreicht werden, bei denen sich Lage- und Streuungsparameter separat voneinander variieren lassen. Konkrete Beispiele für solche *location-scale*-Verteilungsfamilien wurden neben der Normalverteilung mit logistischer Verteilung, sowie Cauchy- und Laplace-Verteilung angegeben.

Nach einer generellen Einführung des üblicherweise verwendeten EM-Algorithmus zur Optimierung der Likelihood (Kap. 2.3) und der Schlussfolgerung, dass dieser im Falle von Multimodalität grundsätzlich keine guten Lösungen garantieren kann, wurde eine systematische Analyse der Komplexität der Funktionen durchgeführt. Dazu wurden in Kapitel 2.4 zwei Heuristiken zur Bestimmung der Anzahl Optima sowie der Breite des globalen Optimums entwickelt. Diese Heuristiken basieren jeweils auf feinen Abtastungen eines Querschnittes der Likelihood-Funktion. Mittels einer Simulation für verschiedene Kombinationen der Varianzparameter und der Anzahl gegebener Beobachtungen konnte herausgefunden werden, dass die Breite hauptsächlich von der kleineren der beiden Varianzen sowie der Anzahl gegebener Beobachtungen abhängt. Wird beides größer, so steigt die Breite des globalen Optimums an. Im Gegensatz dazu gilt für die Anzahl lokaler Optima: Je stärker sich die beiden Varianzen unterscheiden, desto mehr Optima treten auf. Zusätzlich

wird die Anzahl noch durch höhere Beobachtungsanzahlen gesteigert. Für Mischungen aus den weiteren Verteilungen ergaben sich sehr ähnliche Ergebnisse, sodass insgesamt von den gleichen Entstehungsfaktoren für Multimodalität ausgegangen werden kann.

Auf Basis dieser Erkenntnisse wurden in Kapitel 3 Simulationsstudien auf Likelihood-Funktionen unterschiedlicher Komplexität durchgeführt, um verschiedene Optimierungsalgorithmen systematisch zu vergleichen. Der Aufbau der Studien ist in Kapitel 3.1 beschrieben. Es sind 1500 verschiedene Funktionsinstanzen für das zweidimensionale Optimierungsproblem im Raum der Mittelwerte enthalten und 121 500 für die Optimierung aller fünf Parameter der Mischverteilungsmodelle mit zwei Normalverteilungskomponenten. Die grundlegenden Funktionsprinzipien der neben EM in den Vergleich einbezogenen Algorithmen CMA-ES und MBO wurden in Kapitel 3.2 erläutert. Als zusätzliche Vergleichswerte wurden auch eine Zufallssuche und EM-Läufe, die am theoretisch zugrunde liegenden Optimum starten, betrachtet. Ergebnisse der 2D-Studie (Kap. 3.3) sind, dass der auf die Mittelwertparameter eingeschränkte EM-Algorithmus das globale Optimum üblicherweise nicht findet, da durch die festen übrigen Parameter die nötige Flexibilität verloren geht, den Einzugsbereich eines Optimums zu verlassen. Die Vergleichsalgorithmen finden üblicherweise etwas bessere lokale Optima.

Für den anwendungsrelevanteren Simulationsteil in fünf Dimensionen (Kap. 3.4) wurden zuerst fehlgeschlagene EM-Läufe analysiert. Für kleine Beobachtungszahlen und für ein Mischungsverhältnis nahe 1 bzw. 0 ergaben sich deutlich häufiger Abbrüche, da es in diesen Fällen leichter zur Separierung einzelner Beobachtungen kommen kann, wodurch das bekannte Problem der sich gegen unendlich entwickelnden Likelihood auftritt. In der Optimierung aller Dimensionen erreichte der EM-Algorithmus insgesamt deutlich bessere Ergebnisse, üblicherweise in der Nähe der EM-Läufe mit Start am theoretischen Optimum. Insbesondere auf den komplexeren Problemen ergaben sich deutliche Vorteile gegenüber den Vergleichsalgorithmen CMA-ES und MBO, die sich wiederum nicht deutlich von der Zufallssuche absetzen konnten. In einer Analyse der Streudiagramme von erreichtem Funktionswert und Parameter-Abstand zum theoretischen Optimum ergab sich, dass EM beständig mit einem Großteil der Läufe gute Likelihood-Werte in unmittelbarer Nähe der theoretischen Parameter erreicht, während die Ergebnisse der übrigen Algorithmen deutlich breiter streuen und nur vereinzelt vergleichbare Likelihood-Werte liefern (vgl. Abb. 19, S. 40).

Um die Vorteile von EM genauer zu analysieren, wurden in Kapitel 3.5 abschließend die Iterationsverläufe während der Optimierung betrachtet. Das wesentliche Fazit ist hier, dass der EM-Algorithmus häufig bereits in die unmittelbare Nähe des

Optimums gelangt, bevor sich die Varianzparameter so stark unterscheiden, dass Multimodalität im Raum der Mittelwerte entsteht. Um einzelne Iterationsschritte genauer zu erklären, wurde schließlich die interne Zielfunktion zusammen mit den analytischen Mittelwertupdates dargestellt. Dabei zeigte sich, dass die Verwendung von Zugehörigkeitswahrscheinlichkeiten das Optimierungsproblem stark vereinfacht und der unimodale Erwartungswert im E-Step sogar dann erfolgreich optimiert werden kann, wenn es numerisch zu Darstellungsproblemen kommt.

In Kapitel 4 dieser Arbeit wurde die Anzahl vorhandener Optima in den Likelihood-Funktionen analysiert. Zunächst wurde in Kapitel 4.1 ein aus Clustern zusammengesetztes Datenszenario vorgestellt, für welches Améndola Cerón (2017) zeigen konnte, dass mindestens für jedes Cluster ein lokales Optimum in der Likelihood einer Mischung aus zwei Normalverteilungskomponenten existiert. Dazu wurden Startpunkte für den EM-Algorithmus so gewählt, dass jeweils ein einzelnes Cluster initial eine Komponente der Mischung bildet und alle übrigen Beobachtungen zusammen die zweite Komponente. Die deterministischen Ergebnisse der Optimierungsläufe befanden sich dann jeweils in unmittelbarer Nähe der Startpunkte. Diese Vorgehensweise wurde in den folgenden Unterkapiteln auf Datensituationen adaptiert, die normalverteilte Daten annähern.

In Kapitel 4.2 wurden die Beobachtungszahlen für die einzelnen Cluster so angepasst, dass das Histogramm mit passender Intervallbreite dem standardnormalverteilter Beobachtungen entspricht. In den einzelnen Clustern sind die Beobachtungen dabei stetig gleichverteilt und die Clusteranzahl lässt sich schrittweise verdoppeln. In einer ersten Variante wurde bei jeder Verdoppelung eine exakte Halbierung der Clusterbreiten durchgeführt, während in einer zweiten Variante die Breiten asymptotisch anwachsen und so erkennbar eine Normalverteilung annähern. Für die erste Variante konnte für einen inneren Bereich von ca. 80 % der Cluster ein spezifisches Optimum erreicht werden. Für die asymptotische Variante waren es nach wenigen Verdopplungsschritten nur noch weniger als 50 %, da im inneren Bereich von benachbarten Clustern aus identische Optima erreicht wurden.

In zwei weiteren Szenarien (Kap. 4.3) fand durch Verwendung von Normalverteilungen auch innerhalb der einzelnen Cluster ein weiterer Annäherungsschritt statt, was sowohl mit als auch ohne eingebaute Asymptotik nicht zu einer Verringerung der Anzahl gefundener Optima führte. In Kapitel 4.4 wurden Beobachtungen einer zweiten Normalverteilungskomponente mit kleinerer Varianz hinzugefügt und herausgefunden, dass mit Start an den einzelnen Clustern mehr verschiedene lokale Optima erreicht werden können als mit gleich vielen Zufallsstartpunkten für EM-Läufe. Abschließend wurden für normalverteilte Daten aus zwei Komponenten in Kapitel 4.5 zufällig bestehende Cluster identifiziert, von denen aus EM-Läufe ge-

startet werden können. Es zeigte sich hier, dass im Vergleich zu Zufallsstartpunkten insbesondere für steigende Beobachtungs- und Clusteranzahlen mehr als viermal so viele voneinander verschiedene lokale Optima der Likelihood-Funktion gefunden werden konnten. Damit stellen die Clusterstartpunkte ein geeignetes Vorgehen zum Identifizieren möglichst vieler Optima dar, was allgemein zur Beurteilung eines Ergebnisses der globalen Optimierung von praktischem Nutzen sein kann.

Zusammenfassend sind die zentralen Ergebnisse dieser Arbeit, dass stark multimodale Likelihood-Funktionen über den bekannten Fall von Mischungen für Normalverteilungen hinaus auch für Mischungen von Cauchy-, Laplace- und logistischen Verteilungen existieren und dass der Grad der Multimodalität in allen Fällen vom Abstand der Varianzparameter und der Anzahl gegebener Beobachtungen abhängt. Darüber hinaus zeigte sich, dass moderne *Black-Box*-Algorithmen mit Ansätzen, die das Auffinden eines globalen Optimums begünstigen sollen, auf diesen Funktionen deutlich schlechtere Ergebnisse erreichen als der EM-Algorithmus, der häufig in die unmittelbare Nähe der theoretisch zugrunde liegenden Parameter konvergiert. Abschließend wurde mit den Clusterstartpunkten für EM eine Methode vorgeschlagen, mit der es auch auf normalverteilten Daten gelingt, eine große Anzahl lokaler Optima einer Likelihood-Funktion zu identifizieren. Die Bewertung des globalen Optimierungsergebnisses in der praktischen Anwendung von Mischverteilungsmodellen kann damit erheblich erleichtert werden, da im Vergleich zu reinen Zufallsstartpunkten zum Teil mehr als doppelt so viele Vergleichswerte zur Verfügung stehen.

Für weitere Analysen wäre es möglicherweise von Interesse, die Betrachtungen auf Mischungen von mehr als zwei Komponenten bzw. Mischungen multivariater Komponenten auszuweiten. Die grundsätzliche Vermutung dazu wäre, dass es bereits ausreicht, in zwei Verteilungskomponenten bzw. zwei einzelnen Dimensionen zweier Komponenten die ursächlichen Varianzunterschiede herbeizuführen, um dort die Multimodalität zu verursachen. Eine systematische Analyse würde aufgrund einer Vielzahl zu betrachtender Kombinationsmöglichkeiten den Umfang dieser Arbeit um ein Vielfaches übersteigen und auch die Komplexität der Optimierungsprobleme könnte sich durch die zusätzlichen Dimensionen massiv erhöhen.

Mittels eines möglichst simplen multivariaten Clusterverfahrens zum Identifizieren der Clusterstartpunkte ließe sich auch das Auffinden möglichst vieler lokaler Optima zum Bewerten des globalen Optimierungsergebnisses auf mehrdimensionale Verteilungskomponenten erweitern. Beispielsweise könnte mit dem *k-means*-Algorithmus eine vorgegebene Anzahl von k Startpunkten generiert werden, von denen aus dann die EM-Läufe gestartet werden.

Literaturverzeichnis

Aménd ola Cerón, C. E. (2017). *Algebraic Statistics of Gaussian Mixtures*. Dissertation, TU Berlin.

Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. (2009). mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, 32(6):1–29.

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J. & Lang, M. (2017). mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. arXiv:1703.03373 [stat.ML].

Bossek, J. (2016). *cmaesr: Covariance Matrix Adaptation Evolution Strategy*. R package version 1.0.3.

Casella, G. & Berger, R. L. (2002). *Statistical Inference*. Duxbury Pacific Grove, CA, second edition.

Day, N. E. (1969). Estimating the Components of a Mixture of Normal Distributions. *Biometrika*, 56(3):463–474.

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

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27(4):857–871.

Hansen, N. (2006). The CMA Evolution Strategy: A Comparing Review. In Lozano, J. A., Larrañaga, P., Inza, I. & Bengoetxea, E., editors, *Towards a New Evolutionary Computation: Advances in the Estimation of Distribution Algorithms*, 75–102. Springer, Berlin, Heidelberg.

Hansen, N. & Ostermeier, A. (1996). Adapting Arbitrary Normal Mutation Distributions in Evolution Strategies: The Covariance Matrix Adaptation. In *Proceedings of IEEE International Conference on Evolutionary Computation*, 312–317.

Hathaway, R. J. (1985). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *The Annals of Statistics*, 13(2):795–800.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995a). *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, New York, second edition.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995b). *Continuous Univariate Distributions*, volume 2. John Wiley & Sons, New York, second edition.

Jones, D. R., Schonlau, M. & Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492.

Lang, M., Bischl, B. & Surmann, D. (2017). batchtools: Tools for R to Work on Batch Systems. *Journal of Open Source Software*, 2(10):135.

McLachlan, G. & Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley & Sons, New York, second edition.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, New York.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. (1948). The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203.

Roustant, O., Ginsbourger, D. & Deville, Y. (2012). DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization. *Journal of Statistical Software*, 51(1):1–55.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wu, C. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103.

Xie, Y. (2020). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.30.