Likelihood-basierte Methoden für das Testen einer endlichen Anzahl von Zuständen in unabhängigen und Markov abhängigen Mischungsmodellen

Diplomarbeit

vorgelegt von

Florian Schwaiger

angefertigt am

Fachbereich Mathematik

der Philipps-Universität Marburg

2010

betreut von

Prof. Dr. Holzmann

Inhaltsverzeichnis

Einleitung

Mischungsmodelle bieten die Möglichkeit Datensätze bestehend aus mehreren Populationen über ein gemeinsames Modell zu beschreiben. Dabei wählt man für jede Population (wir sprechen ab hier von Komponenten der Mischung) die Wahrscheinlichkeit für ihr Auftreten und eine spezielle Verteilung. Die Verteilungsfunktion des Mischungsmodells ergibt sich dann schließlich über die gewichteten Verteilungsfunktionen der einzelnen Komponenten. Für die Modellwahl sind bei vorliegenden Stichproben drei zentrale Fragestellungen zu klären bzw. zu beachten:

1. Wieviele Komponenten besitzt das Mischungsmodell?
2. Welcher Abhängigkeitstruktur unterliegen die Stichproben?
3. Welche Verteilungsannahmen werden für die Komponenten zugrunde gelegt?

Für die Verteilungen der Komponenten betrachten wir parametrische Verteilungsfamilien mit einem eindimensionalen Parameter, wobei die Verteilungen aller Komponenten aus der selben Verteilungsfamilie stammen sollen. Beispiele sind die Poisson-, Binomial- (mit fester Versuchsanzahl) oder Normalverteilung (mit fester Standardabweichung oder festem Erwartungswert). Unter dieser Annahme werden wir uns in dieser Arbeit mit der ersten Fragestellung beschäftigen. Genauer betrachten wir Testverfahren, die die Hypothese einer gewissen Anzahl von Komponenten untersuchen. Bezüglich der Abhängigkeit betrachten wir dabei unabhängige sowie markovabhängige Stichproben.
Im Kontext der Modellwahl wurde der Likelihood Quotienten Test (*engl. likelihood-ratio test* LRT) entwickelt. Dieses Verfahren berechnet unter dem größeren und dem kleineren Modell die Maximum Likelihoodschätzer (*engl. maximum likelihood estimators* MLE) und betrachtet anschließend die Differenz der logarithmierten Dichten beider Schätzungen. Unter gewissen Bedingungen lässt sich zeigen (vgl. [Fer96]), dass wenn die Stichprobengröße gegen ∞ geht, die Teststatistik eines LRT unter der Nullhypothese χ^2-verteilt ist. Eine Voraussetzung dafür ist, dass der Parameter des wahren Modells im Alternativraum identifizierbar ist. Dies ist in unserem Fall nicht gewährleistet, da wir

unter der Alternative z.B. ein Gewicht gleich Null und den Parameter beliebig wählen können. Ein standardmäßiger LRT hilft uns also bei der Untersuchung unserer Fragestellung nicht weiter.
Um diese Problematik zu lösen, haben Chen et. al für das Testen auf eine bzw. zwei Komponenten einen modified LRT (MLRT) entwickelt (vgl. [CCK01] bzw. [CCK04]). Grundidee ist es beliebig kleine Gewichte für die Komponenten zu verhindern und so Konsistenzaussagen über die Parameter der Mischung zu ermöglichen. Im zweiten Kapitel dieser Arbeit werden wir diesen Test für das Testen zweier Komponenten betrachten. Wir werden dabei auf die Herleitung der asymptotischen Verteilung der Teststatistik eingehen und die praktische Umsetzung des Tests beschreiben. Ferner führen wir für Mischungen aus normalverteilten Zufallsvariablen mit festem Erwartungswert Simulationen durch. Eine wesentliche Annahme dieses Tests ist, dass die Zufallsvariablen unabhängig voneinander verteilt sind. Diese Annahme verhindert die Anwendung des Tests in allen Situationen, bei denen wir Abhängigkeiten zwischen den einzelnen Zufallsvariablen beobachten. Da Hidden Markov Modelle die Möglichkeit bieten Mischungen mit Abhängigkeit zu modellieren, verallgemeinerten Dannemann und Holzmann den MLRT für diese Modellklasse (vgl. [DH08]). Sie zeigten, dass die Resultate bezüglich der asymptotischen Verteilung der Teststatistik des MLRT auch noch beim Zugrundelegen eines Hidden Markov Modells gelten. Diese Verallgemeinerung werden wir im zweiten Teil von Kapitel 2 betrachten. Auch hier gehen wir auf die Herleitung der asymptotischen Verteilung der Teststatistik ein und führen Simulationen durch.
Mit Hilfe des MLRT können wir nun die Hypothese von zwei gegen mehrere Komponenten untersuchen. Falls wir hier die Nullhypothese verwerfen, können wir erstmal keine weiteren Aussagen über die Anzahl der Komponenten treffen, außer dass sie größer ist als zwei. Um diese Lücke zu schließen, wurde ein EM-Test entwickelt (vgl. [CL09]). Hiermit können wir für ein beliebiges $m_0 \in \mathbb{N}$ die Nullhypothese von m_0 Komponenten untersuchen. Wir werden im dritten Kapitel dieser Arbeit näher auf dieses Verfahren eingehen. Auch bei diesem Test ist die Unabhängigkeit wieder eine Annahme. Es stellt sich daher die Frage, ob wir auch hier eine Verallgemeinerung auf Hidden Markov Modelle durchführen können. Dieser Fragestellung werden wir im zweiten Teil von Kapitel 3 simulationsbasiert nachgehen.
Im nun folgenden ersten Kapitel gehen wir zunächst formal auf Mischungsmodelle ein. Außerdem betrachten wir Verfahren für die Berechnung von Maximum Likelihoodschätzern in dieser Modellklasse und erläutern die Grundlagen von Hidden Markov Modellen.

1 Grundlagen

1.1 Mischungmodelle

Wir beginnen diesen Abschnitt mit der Definition von Mischungsmodellen:

Definition 1.1 (Mischungsmodell *engl. mixture model*)**.** Sei $f(\cdot, \theta)$ eine Dichte bezüglich dem σ-endlichen Maß μ mit Parameter $\theta \in \Theta$ und $G : \Theta \to [0, 1]$ eine diskrete Verteilungsfunktion mit

$$G(\theta) := \sum_{j=1}^{k} \pi_j \, I(\theta_j \leq \theta),$$

wobei $\theta_1, \ldots, \theta_k \in \Theta$ und $\pi_j \geq 0$ mit $\sum_{j=1}^{k} \pi_j = 1$ die jeweils zugehörigen Gewichte sind. Das eigentliche Mischungsmodell wird nun durch die Dichte

$$f(x; G) := \int_{\Theta} f(x, \theta) \, dG(\theta) = \sum_{j=1}^{k} \pi_j \, f(x, \theta_j) \tag{1.1}$$

beschrieben. Im Folgenden bezeichnen wir G als mixing distribution, f als Komponentendichte und $\theta_1, \ldots, \theta_k$ als support points.

Im Rahmen dieser Arbeit werden wir uns für die einzelnen Komponenten generell auf Verteilungsfamilien beschränken, die nur einen reellen Parameter besitzen ($\Theta \subset \mathbb{R}$) und bezeichnen stets mit $f(x, \theta)$ die Dichte der Komponenten. Diese wird als beliebig (unter gewissen kontextabhängigen Regularitätsbedingungen) angenommen. Ein Mischungsmodell mit k Komponenten wird also durch Angabe der mixing distribution G charakterisiert. Die Menge aller mixing distibutions mit k Komponenten bezeichnen wir mit

$$\mathcal{M}_k := \left\{ G(\theta) = \sum_{j=1}^{k} \pi_j \, I(\theta_j \leq \theta) : \theta_1 \leq \ldots \leq \theta_k, , \pi_j \geq 0, \sum_{j=1}^{k} \pi_j = 1 \right\}.$$

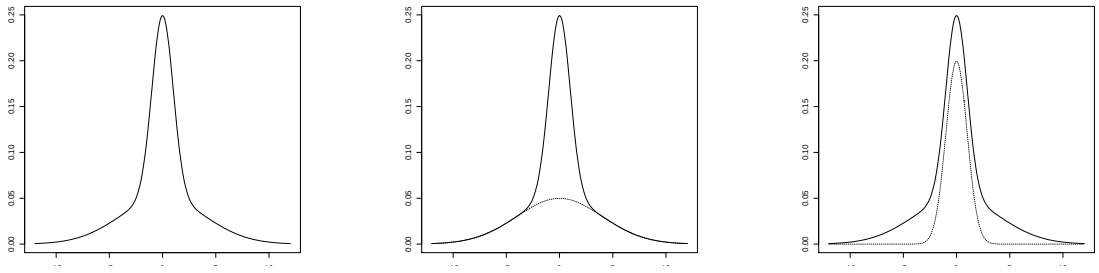
Folglich entspricht die Menge aller mixing distributions $\mathcal{M}$ der Vereinigung aller Mengen $\mathcal{M}_k$ mit $k \geq 1$.

Da wir uns mit Testverfahren bzgl. der Anzahl der Komponenten beschäftigen, sind wir formal an folgender Testsitution interessiert:

$$H_0 : G \in \mathcal{M}_{m_0} \text{ gegen } H_1 : G \in \mathcal{M} \backslash \mathcal{M}_{m_0} \text{ für } m_0 \in \mathbb{N}.$$

Das folgende Beispiel veranschaulicht die Modellklasse der Mischungsmodelle.

Beispiel 1.1. Wir betrachten eine Mischung von zwei normalverteilten Komponenten mit festem Erwartungswert Null und variabler Standardabweichung. Beide Komponenten sollen mit gleicher Wahrscheinlichkeit auftreten. Für die Standardabweichungen wählen wir die Werte $\sigma_1 := 1$ und $\sigma_2 := 2$. Eine Zufallsvariable, die nach einem solchen Modell verteilt ist, folgt also jeweils mit Wahrscheinlichkeit 0.5 entweder einer Verteilung mit geringerer oder höherer Standardabweichung.
Wenn $f_\sigma(\cdot)$ die Dichte einer normalverteilten Zuvallsvariable mit Erwartungswert Null und Standardabweichung σ beschreibt, so besitzt unser Mischungsmodell die Dichte $0.5\ f_1(x) + 0.5\ f_2(x)$. Diese Dichte wird in Abbildung 1.1 (a) veranschaulicht. Die Ab-



(a) (b) (c)

Abbildung 1.1: Dichte des Mischungsmodells zweier normalverteilter Komponenten

bildungen 1.1 (b) und (c) zeigen die Dichte des Mischungsmodells und jeweils gepunktet eine der beiden Komponentendichten.

Mischungsmodelle von Normalverteilungen mit festem Erwartungswert und variabler Standardabweichung werden wir ebenfalls für die Simulationen in den Kapiteln zwei und drei benutzen. Außerdem werden wir mit diesem Modell die Renditen der Siemens AG in Abschnitt 3.2.3 untersuchen.

1.2 Parameterschätzung bei unabhängigen Stichproben

Angenommen es liegen n unabhängig identisch verteilte Zufallsvariablen $X_1, \ldots, X_n$ eines Mischungsmodells mit mixing distribution $G(\cdot) = \sum_{j=1}^{k} \pi_j \; I(\theta_j \leq \cdot) \in \mathcal{M}_k$ und bekannter parametrischer Verteilungsfamilie der Komponentendichte $f(\cdot, \theta)$ vor. Sind wir an einer Parameterschätzung interessiert, so ist die Maximum Likelihoodschätzung eine Methode dafür. Die Idee dieser Vorgehensweise ist die gemeinsame Dichte der betrachteten Zufallsvariablen gegeben einer Stichprobe $x_1, \ldots, x_n$ über die Parameter der mixing distribution zu maximieren. Wir betrachten nun zwei verschiedene Verfahren um diesen Schätzer zu berechnen: Zum einen können wir ihn durch direktes Maximieren der Likelihoodfunktion ermitteln, zum anderen können wir auf den EM Algorithmus zurückgreifen.

Direktes Maximieren der Likelihoodfunktion

Im hier betrachteten Fall unabhängiger Zufallsvariablen ergibt sich die Likelihoodfunktion

$$L_n(G \mid \boldsymbol{x}) := f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; G) = \prod_{i=1}^{n} f(x_i; G) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j \; f(x_i; \theta_j),$$

wobei $\boldsymbol{x} = (x_1, \ldots, x_n)$ die Stichprobe ist. Da diese Werte typischerweise klein sind und das Multiplizieren zu numerischen Problemen führen kann, wird eher die logarithmierte Likelihoodfunktion

$$l_n(G \mid \boldsymbol{x}) := \log\left(L_n(G)\right) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \pi_j \; f(x_i; \theta_j)\right)$$

unter den Nebenbedingungen, die durch die Einschränkungen der Komponentenparameter und Gewichte gegeben werden, maximiert. Formal erhalten wir also

$$\widehat{G}_{ML} := \underset{G \in \mathcal{M}_k}{\operatorname{argmax}} \left(l_n(G \mid \boldsymbol{x})\right).$$

Im Falle von z.B. normalverteilten Komponenten mit festem Erwartungswert wären die Parameter θ_j die Standardabweichungen und somit die Nebenbedingungen $\theta_j > 0$ für $j = 1, \ldots, k$ und $\pi_j \in (0, 1)$ mit $\sum_{j=1}^{k} \pi_j = 1$.

EM Algorithmus

Der EM Algorithmus (EM steht für „expectation-maximization“ und beschreibt die beiden Schritte eines Iterationsdurchlaufs) ist ein iteratives Verfahren zur Parameterschätzung

über die Maximum Likelihoodmethode für Modelle mit unbeobachtbaren (latenten) Variablen. Wir möchten daher zunächst kurz erklären wie man über die Einführung einer latenten Variable ebenfalls eine Zufallsvariable X erhalten kann, die einem Mischungsmodell mit mixing distribution $G \in \mathcal{M}_k$ folgt. Zunächst definieren wir die diskrete Zufallsvariable S, die Werte in $\{1, \ldots, k\}$ mit $P(S = j) = \pi_j$ für $j = 1, \ldots, k$ annimmt. Diese Zufallsvariable ist i.A. nicht beobachtbar. Als nächstes setzen wir nun die Verteilung X gegeben der Zustand $S = j$ ist eingetreten, indem wir definieren $(X \mid S = j) :\sim f(\cdot, \theta_j)$. Somit erhalten wir

$$X = \sum_{j=1}^{k} I\left(S = j\right) \cdot \left(X \mid S = j\right) \sim \sum_{j=1}^{k} \pi_j f(\cdot, \theta_j) = f(\cdot; G).$$

Diese Sichtweise lässt sich natürlich auch auf die Situation übertragen, in der wir n unabhängige Stichproben des Mischungsmodells vorliegen haben. Hier betrachtet man dann den bivariaten Prozess $(S_i, X_i)_{i=1,\ldots,n}$, wobei S_i jeweils die zu X_i gehörige latente Variable ist und die S_i voneinander unabhängig sind (beim HMM sind die S_i abhängig und folgen einer Markov Kette). Die folgende Ausführung (ausgenommen die Zerlegung der Probleme im M-Schritt) orientiert sich an [FS06] Kapitel 2. Wir gehen nun auf die Funktionsweise des EM Algorithmus in diesem Zusammenhang ein und geben zunächst die Likelihoodfunktion der kompletten Daten (beobachtbar und latent) an

$$\begin{aligned} L_n(G \mid \boldsymbol{x}, \boldsymbol{s}) &= \prod_{i=1}^{n} P(S_i = s_i; G)\ f(x_i \mid S_i = s_i; G) \\ &= \prod_{i=1}^{n} \pi_{s_i}\ f(x_i, \theta_{s_i}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[\pi_j\ f(x_i, \theta_j)\right]^{I(s_i = j)}, \end{aligned}$$

wobei $\boldsymbol{x} = (x_1, \ldots, x_n)$ bzw. $\boldsymbol{s} = (s_1, \ldots, s_n)$ die beobachtbaren bzw. nicht beobachtbaren Werte sind. Für den EM Algorithmus ist die logarithmierte Likelihoodfunktion relevant. Diese ergibt sich zu

$$l_n(G \mid \boldsymbol{x}, \boldsymbol{s}) = \log\left(\prod_{i=1}^{n} \prod_{j=1}^{k} \left[\pi_j\ f(x_i, \theta_j)\right]^{I(s_i = j)}\right) = \sum_{i=1}^{n} \sum_{j=1}^{k} I(s_i = j)\ \log\left(\pi_j\ f(x_i, \theta_j)\right),$$

Diese Funktion kann natürlich nicht wie oben direkt maximiert werden, da die Beobachtung der latenten Variablen nicht möglich und somit $I(s_i = j)$ unbekannt ist. Der EM Algorithmus löst dieses Problem wie folgt: Zunächst müssen wir Startwerte für die zu schätzenden Parameter $(\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k)$ von G wählen. Wir bezeichnen diese mit $\pi_j^{(0)}$ und $\theta_j^{(0)}$ für $j = 1, \ldots, k$ bzw. die zugehörige mixing distribution mit $G^{(0)}$. Von

diesen Parametern ausgehend starten wir nun ein iteratives Verfahren, welches in jedem Schritt den Log-Likelihoodwert vergrößert. Grundidee dabei ist es die unbeobachtbaren Werte $I(s_i = j)$ durch ihre bedingten Erwartungswerte gegeben den aktuellen Parametern und dem beobachtbaren Datensatz zu schätzen. Von hier ausgehend maximieren wir anschließend jeweils die Parameter der Mischung. Wir geben nun den m-ten Schritt des Verfahrens an:

1. Berechne für $i = 1, \ldots, n$ und $j = 1, \ldots, k$

$$D_{ij}^{(m)} := E\left(I(S_i = j) \mid \boldsymbol{x},\ G^{(m-1)}\right) = P\left(S_i = j \mid \boldsymbol{x},\ G^{(m-1)}\right) = \frac{\pi_j^{(m-1)}\ f(x_i, \theta_j^{(m-1)})}{f(x_i; G^{(m-1)})}.$$

   Wir berechnen daher in diesem Schritt gegeben die Stichprobe und den aktuellen Parametern die Wahrscheinlichkeit dafür dass $S_i = j$. Da dies mit dem Erwartungswert über die Indikatorfunktion übereinstimmt, nennt man diesen Teil der Iteration E-Schritt. Somit haben wir (insbesondere zu Beginn bei m=1) eine Möglichkeit die unbekannten Werte der Log-Likelihoodfunktion durch ihre bedingten Erwartungswerte zu ersetzen.

2. Maximiere den bedingten Erwartungswert der Log-Likelihoodfunktion über die Parameter der mixing distribution, d.h.

$$G^{(m)} := \underset{G \in \mathcal{M}_k}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} D_{ij}^{(m)}\ \log\left(\pi_j\ f(x_i, \theta_j)\right) \right].$$

   ist zu berechnen.

   Die Maximierung über $G \in \mathcal{M}_k$ bedeutet dabei, dass wir k Gewichte, die sich zu Eins addieren, und k zulässige support points optimal wählen. Wir beschreiben im Folgenden die dazu nötige Vorgehensweise. Aufgrund der Struktur der Zielfunktion können wir dieses Optimierungsproblem in einzelne Teilprobleme zerlegen, was uns das Berechnen erleichtert:

$$\sum_{i=1}^{n} \sum_{j=1}^{k} D_{ij}^{(m)}\ \log\left(\pi_j\ f(x_i, \theta_j)\right) = \sum_{i=1}^{n} \sum_{j=1}^{k} D_{ij}^{(m)}\ \log\left(f(x_i, \theta_j)\right) + \sum_{j=1}^{k} \log\left(\pi_j\right) \underbrace{\left(\sum_{i=1}^{n} D_{ij}^{(m)} \right)}_{:= a_j^{(m)}}. \quad (1.2)$$

   Diese Aufteilung ist beim direkten Maximieren (vgl. vorherigen Abschnitt) nicht zulässig, da der Logarithmus außerhalb der inneren Summe steht. Daher mussten wir dort die Parameter gemeinsam in einem Schritt schätzen. Die Gewichte sind somit nur abhängig von der zweiten Summe und es gilt

$$\left(\pi_1^{(m)}, \ldots, \pi_k^{(m)}\right) := \operatorname{argmax} \left[\sum_{j=1}^{k} a_j^{(m)} \log\left(\pi_j\right) \right] \text{ s.t. } \pi_j > 0,\ \sum_{j=1}^{k} \pi_j = 1.$$

Löst man dieses Problem erhält man, dass sich die optimalen Werte über

$$\pi_j^{(m)} = n^{-1}\, a_j^{(m)} = n^{-1} \sum_{i=1}^{n} D_{ij}^{(m)} \text{ für } j = 1, \ldots, k$$

berechnen. Sie ergeben sich also über den Mittelwert der aktuell geschätzten Wahrscheinlichkeiten dafür, dass $S_i = j$ $(i = 1, \ldots, n)$. Für die Schätzung der neuen support points ist nur noch der vordere Teil aus der Zerlegung in Gleichung (1.2) relevant. Diesen können wir schreiben als

$$\sum_{i=1}^{n}\sum_{j=1}^{k} D_{ij}^{(m)}\ \log\left(f(x_i, \theta_j)\right) = \sum_{i=1}^{n} D_{i1}^{(m)}\ \log\left(f(x_i, \theta_1)\right) + \cdots + \sum_{i=1}^{n} D_{ik}^{(m)}\ \log\left(f(x_i, \theta_k)\right).$$

Der optimale support point $\theta_j^{(m)}$ ermittelt sich also unabhängig von den restlichen über

$$\theta_j^{(m)} := \underset{\theta \in \Theta}{\operatorname{argmax}} \left[\sum_{i=1}^{n} D_{ij}^{(m)}\ \log\left(f(x_i, \theta_j)\right) \right] \text{ für } j = 1, \ldots, k.$$

Die Menge Θ beschreibt dabei die zulässige Region der support points. Für diese Schätzung der support points ergeben sich je nach parametrischer Verteilungsfamilie unterschiedliche Lösungen.

Neben dem Zweck der Maximum Likelihoodschätzung ist der EM Algorithmus für uns interessant, da er im EM-Test eine wichtige Rolle spielt.

1.3 Hidden Markov Modelle

Wie bereits in der Einleitung erwähnt, sind der MLRT und der EM-Test Verfahren für unabhängig identisch verteilte Zufallsgrößen. Um diese Tests auf abhängige Mischungsmodelle übertragen zu können, betrachten wir Hidden Markov Modelle (HMM). Wir definieren daher zunächst, was wir unter einem solchen verstehen:

Definition 1.2 (Hidden Markov Modell)**.** Sei $(S_i)_{i\in\mathbb{N}}$ ein stochastischer Prozess mit Werten in $\mathcal{S} := \{1, \ldots, k\}$. Gilt für $i \in \mathbb{N}$ und beliebige $s_1, \ldots, s_i, s_{i+1} \in \mathcal{S}$ mit $P(S_i = s_i, \ldots, S_1 = s_1) > 0$

$$P\left(S_{i+1} = s_{i+1} \mid S_i = s_i, \ldots, S_1 = s_1\right) = P\left(S_{i+1} = s_{i+1} \mid S_i = s_i\right), \tag{1.3}$$

so nennen wir diesen Prozess eine Markov Kette. In allgemeineren Modellen würde man einen solchen Prozess als diskrete Markov Kette erster Ordnung mit endlichem Zustandsraum bezeichnen. Da wir ausschließlich obigen Fall betrachten, sprechen wir nur von einer Markov Kette.

Sei $(S_i)_{i\in\mathbb{N}}$ nun eine Markov Kette und $(X_i)_{i\in\mathbb{N}}$ ein Prozess von Mischungsmodellen gemäß

$$X_i := \sum_{j=1}^{k} I(S_i = j)\ Z_{ji} \text{ für } i \in \mathbb{N},$$

wobei $(Z_{ji})_{i\in\mathbb{N}}$ für $j \in \{1,\dots,k\}$ ein stochastischer Prozess ist, der unabhängig identisch verteilt ist gemäß der Dichte $f(\cdot,\theta_j)$ bezüglich dem σ-endlichen Maß μ. Wie im Überblick der Arbeit erwähnt, lassen wir für θ_j nur eindimensionale, reellwertige Parameter zu. Wir nennen den bivariaten Prozess $(X_i, S_i)_{i\in\mathbb{N}}$ ein Hidden Markov Modell, wobei nur der Prozess der X_i beobachtbar ist.

Anmerkungen und Annahmen zur Markov Kette des Modells

Generell möchten wir uns vom iid Mischungsfall nicht zu weit entfernen. Genauer soll nur die Unabhängigkeit der beobachtbaren Variablen aufgegeben, aber die Stationarität beibehalten werden. Daher betrachten wir nun zunächst einige Eigenschaften der Markov Kette, die uns dann Rückschlüsse bezüglich der beobachtbaren Variablen des HMM zulassen.
Wir gehen davon aus, dass die Übergangswahrscheinlichkeiten der Markov Kette unabhängig von Index $i \in \mathbb{N}$ sind und sie somit zeitlich homogen ist. Für die Markov Kette $(S_i)_{i\in\mathbb{N}}$ bezeichnen wir die Einperiodenübergangsmatrix mit

$$(\Gamma)_{lj} := P\left(S_{i+1} = l \mid S_i = j\right) \text{ für } l, j \in \mathcal{S},$$

und mit $\boldsymbol{u}(i) = (P(S_i = 1),\dots,P(S_i = k))$ die unbedingte Verteilung des Zustands zum Zeitpunkt i. Legen wir nun noch die Verteilung des ersten Zustands $\boldsymbol{u}(1)$ fest, so haben wir den Prozess eindeutig charakterisiert. Denn da per Annahme die Markov Kette zeitlich homogen ist und endlichen Zustandsraum besitzt, gilt

$$\boldsymbol{u}(i+1) = \boldsymbol{u}(i)\ \Gamma. \tag{1.4}$$

Wir nehmen weiter an, dass die Markov Kette irreduzibel ist. Diese Eigenschaft bedeutet anschaulich, dass von einem beliebigen Zustand aus jeder Zustand erreichbar ist und wir somit die Markov Kette nicht in zwei getrennte Ketten teilen können.

Beispiel 1.2 (Irreduzibilität)**.** Betrachten wir die zwei Markov Ketten mit den Übergangsmatrizen

$$\Gamma_1 := \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.6 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix}, \Gamma_2 := \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

und beliebigen Startverteilungen, so ist die Markov Kette mit Übergangsmatrix Γ_1 irreduzibel, da wir von Zustand eins jeden direkt und von Zustand zwei und drei jeweils über den ersten Zustand jeden erreichen können. Im Gegensatz dazu ist die zweite Markov Kette nicht irreduzibel.

Da wir nur Markov Ketten mit endlichem Zustandsraum betrachten, liefert uns die Annahme der Irreduzibilität, dass jeder Zustand und somit die Markov Kette positiv rekurrent ist (vgl. [MS05, Satz 9.28]). Dabei bedeutet der Zustand s ist positiv rekurrent: Mit Wahrscheinlichkeit eins kehren wir unendlich oft zu s zurück (rekurrent) und die erwartete Rückkehrzeit ist endlich (positiv rekurrent).
Die Irreduzibilität und positive Rekurrenz stellen nun die eindeutige Existenz einer stationären Verteilung $\boldsymbol{\delta}$ mit

$$\boldsymbol{\delta} = \boldsymbol{\delta}\,\Gamma, \text{ wobei } \boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)\,, \ \sum_{j=1}^{k} \delta_j = 1$$

sicher (vgl. [MS05, Satz 9.36]).
Ferner nehmen wir an, dass die Markov Kette aperiodisch ist, d.h. jeder Zustand besitzt eine positive Wahrscheinlichkeit im Zustand zu verweilen ($\Gamma_{ll} > 0$ für alle $l \in \mathcal{S}$). Diese zusätzliche Annahme liefert uns einerseits, dass die Markov Kette ergodisch ist und andererseits konvergiert die Verteilung der Markov Kette für eine beliebige Startverteilung gegen die stationäre Verteilung δ (vgl. [MS05, Satz 9.47]).
Wir starten die Markov Kette mit dieser Verteilung δ, d.h. $\boldsymbol{u}(1) := \boldsymbol{\delta}$. Dies ist wegen angesprochener Konvergenz keine starke Einschränkung. Wir erhalten nun wegen (1.4) induktiv, dass die unbedingte Verteilung für jedes $i \in \mathbb{N}$ konstant gleich $\boldsymbol{\delta}$ ist. Überdies hinaus ist die Markov Kette sogar ein stationärer Prozess (vgl. [MS05, Satz 9.32]). Insgesamt ist somit die Verteilung des Prozesses in unserem Fall eindeutig durch Angabe von Γ charakterisiert.

Eigenschaften eines HMM

Beschreiben wir ein HMM wie in Definition 1.2 festgelegt, so beobachten wir zunächst, dass X_i nur abhängt vom aktuellen Zustand der Markov Kette und unabhängig ist von der restlichen Vergangenheit bis zum Zeitpunkt $i-1$, d.h. es gilt

$$(X_i \mid S_1, \ldots, S_i) \overset{(V.)}{=} (X_i \mid S_i)\,.$$

Dieser Zusammenhang ist gültig, weil X_i so gesetzt wurde, dass $(X_i \mid S_i)$ in Verteilung mit $Z_{S_i i}$ übereinstimmt. Genauer gilt

$$(X_i \mid S_i = j) \sim f(\cdot, \theta_j) \text{ mit } j \in \mathcal{S} \text{ und } \theta_j \in \Theta. \tag{1.5}$$

Ferner wurden die Z_{ji} als unabhängig angenommen und daher sind die beobachtbaren Variablen bis $X_1, \ldots, X_i$ gegeben der Vergangenheit der Markov Kette $\sigma(S_1, \ldots, S_i)$ ebenfalls unabhängig voneinander, gemäß Gleichung (1.5), verteilt. Diese Beobachtung bedeutet, dass wir in unserem Modell die komplette Abhängigkeitsstruktur der beobachtbaren Variablen nur über die Markov Kette modellieren.
Als nächstes gehen wir noch kurz auf die unbedingte Verteilung einer einzelnen beobachtbaren Variable ein. Generell können wir

$$X_i = \sum_{j=1}^{k} \mathbf{1}_{\{S_i = j\}} \cdot (X_i \mid S_i = j)$$

schreiben und erhalten somit, dass X_i die Dichte $\sum_{j=1}^{m} P(S_i = j) \cdot f(\cdot, \theta_j)$ besitzt. Da wir die Markov Kette mit der stationären Verteilung $\boldsymbol{\delta}$ starten, gilt nun

$$X_i \sim \sum_{j=1}^{m} \delta_j \cdot f(\cdot, \theta_j) \text{ für alle } i \in \mathbb{N}.$$

Somit ist X_i für alle $i \in \mathbb{N}$ nach demselben Mischungsmodell verteilt. Ferner überträgt sich sogar die Stationarität der Markov Kette aufgrund der zeitlichen Homogenität auf die beobachtbaren Variablen und macht auch sie zu einem stationären Prozess.

2 Test auf zwei Komponenten

In diesem Kapitel wird für den Fall $m_0 = 2$, d.h. Test auf zwei Komponenten, zunächst ein modifizierter Likelihood Ratio Test für unabhängige Stichproben vorgestellt. Anschließend betrachten wir eine Erweiterung für Hidden Markov Modelle. Als Ergebnis wird sich herausstellen, dass trotz der Abhängigkeitsstruktur des HMM die Teststatistiken unter H_0 asymptotisch die gleiche Verteilung aufweisen.

2.1 Modified Likelihood Ratio Test

Die folgenden Ausführungen orientieren sich maßgeblich an einer Arbeit von Chen et. al (vgl. [CCK04]). Beim MLRT sind die vorliegenden Zufallsvariablen $X_1, \ldots, X_n$ voneinander unabhängig gemäß eines Mischungsmodells (vgl. Definition 1.1) verteilt. Der Test dient dazu die Hypothese von zwei Komponenten zu untersuchen, so dass hier $H_0 : G \in \mathcal{M}_2$ gegen $H_1 : G \in \mathcal{M} \backslash \mathcal{M}_2$ getestet wird. Grundsätzliches Ziel ist es nun die asymptotische Verteilung einer Teststatistk zu berechnen, welche der eines gewöhnlichen LRT ähnelt. Bevor wir auf die asymptotischen Eigenschaften des Tests eingehen, treffen wir einige Annahmen, die für deren Herleitung benötigt werden.

2.1.1 Annahmen

Als erstes gehen wir davon aus, dass der Parameterraum der Komponentendichten f eine kompakte Teilmenge der reellen Zahlen ist, d.h. $\Theta \overset{kp}{\subset} \mathbb{R}$. Weiterhin treffen wir die Annahme, dass eine mixing distribution $G \in \mathcal{M}_2$ identifizierbar ist, so dass für $G_1, G_2 \in \mathcal{M}_2$ gilt:

$$f(x; G_1) = f(x; G_2) \ \mu\text{-f.s.} \implies G_1(\theta) = G_2(\theta) \ \forall \ \theta \in \Theta. \tag{2.1}$$

Um später die asymptotische Verteilung der (bisher nicht definierten) Teststatistik, gegeben H_0, berechnen zu können, gehen wir weiterhin davon aus, dass die wahre mixing

distribution G_0 in $\mathcal{M}_2$ liegt und definieren sie durch

$$G_0(\theta) := \pi_0 \, I(\theta_{01} \leq \theta) + (1 - \pi_0) \, I(\theta_{02} \leq \theta) \tag{2.2}$$

mit $\theta_{01} < \theta_{02}$ und $\pi_0 \in (0, 1)$. Sämtliche Erwartungswerte werden in diesem Kapitel (wenn nicht explizit anders erwähnt) bzgl. der Dichte

$$f(x; G_0) = \pi_0 \, f(x, \theta_{01}) + (1 - \pi_0) \, f(x, \theta_{02})$$

berechnet.

Annahme 2.1 (Integrierbarkeits- bzw. Konsistenzbedingungen)**.** Die Komponentendichte $f(x, \theta)$ bzw. die wahre Dichte der Mischung erfüllen die Bedingungen, die Leroux (vgl. [Ler92a]) fordert. Somit gilt insbesondere

1. $f(x, \theta) \leq h(x)$, wobei h stetig ist, $\theta \in \Theta$ beliebig und es gilt $E\left(|\log(h(X)|\right) < \infty$
2. $E\left([\log(f(X, \theta))]^-\right) < \infty$ für $\theta \in \Theta$ beliebig und $[x]^- := \max\{-x, 0\}$.

Mit der Annahme der Stetigkeit (vgl. Annahme 2.2) erhalten wir dank Annahme 2.1 später die Konsistenz des Maximumlikelihood Schätzers.

Annahme 2.2 (Glattheit)**.** Der Träger von $f(x, \theta)$ ist unabhängig von θ und $f(x, \theta)$ ist dreimal stetig differenzierbar nach θ.

Wir definieren nun für $i = 1, \ldots, n$ einige Größen, die später für die Herleitung der asymptotischen Verteilung der Teststatistik benötigt werden. Die Ableitungen sind dabei wieder nach θ zu verstehen.

$$Y_{ij}(\theta) := \frac{f(X_i, \theta) - f(X_i, \theta_{0j})}{f(X_i; G_0)} \text{ für } j = 1, 2 \tag{2.3}$$

$$Y_i'(\theta) := \frac{f'(X_i, \theta)}{f(X_i; G_0)}, \quad Y_i''(\theta) := \frac{f''(X_i, \theta)}{f(X_i; G_0)}, \quad Y_i'''(\theta) := \frac{f'''(X_i, \theta)}{f(X_i; G_0)} \tag{2.4}$$

$$\Delta_i := \frac{f(X_i, \theta_{01}) - f(X_i, \theta_{02})}{f(X_i; G_0)}$$

Da die X_i unabhängig identisch verteilt sind, gilt dies auch für die Zufallsvektoren

$$\left(\Delta_i, Y_i'(\theta_{01}), Y_i'(\theta_{02}), Y_i''(\theta_{01}), Y_i''(\theta_{02})\right).$$

Wir treffen nun für die zugehörige Kovarianzmatrix die folgende Annahme:

Annahme 2.3 (Positive Definitheit der Kovarianzmatrix)**.**

$$\Sigma := \mathrm{COV}\left(\Delta_1, Y_1'(\theta_{01}), Y_1'(\theta_{02}), Y_1''(\theta_{01}), Y_1''(\theta_{02})\right)$$

ist positiv definit.

Diese Annahme stellt sicher, dass die quadratische Form über diese Matrix (d.h. $\sqrt{t'\Sigma t}$ für $t \in \mathbb{R}^5$) eine Norm definiert.

Annahme 2.4 (Gleichmäßige Beschränktheit)**.** Es existiert eine integrierbare Funktion g (d.h. $E(|g(X)|) < \infty$) und ein $\delta > 0$, so dass $|Y_{ij}(\theta)|^{4+\delta} \leq g(X_i)$, $|Y_i'(\theta)|^3 \leq g(X_i)$,$|Y_i''(\theta)|^3 \leq g(X_i)$ und $|Y_i'''(\theta)|^3 \leq g(X_i)$ für alle $\theta \in \Theta$.

Nach Chen und Li (vgl. [CL09]) impliziert Annahme 2.4, dass für eine Folge von zufälligen Parametern θ_n, die stochastisch gegen ein $\theta_0 \in \Theta$ konvergiert

$$n^{-1/2} \sum_{i=1}^{n} Y_i'''(\theta_n) = O_P(1) \tag{2.5}$$

gilt.
Wir gehen nun in diesem Abschnitt davon aus, dass die bisher formulierten Annahmen stets erfüllt sind. Die Erste wird für die Konsistenz des modifizierten Maximum Likelihoodschätzers benötigt. Die Differenzierbarkeit ist notwendig, da wir später in einer Taylorentwicklung im Beweis zu Lemma 2.7 die Ableitungen, wie in (2.3) und (2.4) definiert, benötigen. Die vierte Annahme spielt eine entscheidende Rolle bei der Herleitung der Ordnung des Fehlers der Taylorentwicklung in Lemma 2.7.

2.1.2 Teststatistik

Im Gegensatz zu einem gewöhnlichen LRT wird bei diesem Test nicht der Logarithmus der Likelihoodfunktion optimiert, sondern eine modifizierte Version, bei der kleine Werte für die Gewichte π_i bestraft werden. Wir definieren also zunächst die modifizierte Likelihoodfunktion und anschließend die Teststatistik.

Definition 2.1 (modifizierte Log-Likelihoodfunktion)**.** Zu einer mixing distribution $G \in \mathcal{M}_k$ mit $k \in \mathbb{N}$ ist die gewöhnliche Log-Likelihoodfunktion gegeben durch

$$l_n(G) = \sum_{i=1}^{n} \log(f(X_i; G)).$$

Wir bezeichnen weiter mit

$$\text{pen}(G) := C_k \cdot \sum_{j=1}^{k} \log(\pi_k) \tag{2.6}$$

den Penaltyterm und definieren hiermit schließlich die modifizierte Version der Log-Likelihoodfunktion über

$$\widetilde{l}_n(G) := l_n(G) + \text{pen}(G), \tag{2.7}$$

wobei C_k eine beliebige positive Konstante ist. Entsprechend dieser Definition bezeichnen wir mit

$$\widehat{G}^{(k)} := \underset{G \in \mathcal{M}_k}{\text{argmax}}\ \widetilde{l}_n(G)$$

den modifizierten Maximum Likelihoodschätzer für G über $\mathcal{M}_k$.

Es stellt sich nun die Frage, warum man eine modifizierte Version der Log-Likelihoodfunktion betrachten sollte. Der Sinn dahinter ist, dass eine Schätzung $\widehat{\pi}_j$, welche beliebig klein werden kann, für den zugehörigen Schätzwert des Parameters $\widehat{\theta}_j$ beliebige (zulässige) Werte ermöglicht und somit die Konsistenz von $\widehat{\theta}_j$ verhindert. Im folgenden Abschnitt über Konsistenz werden wir sehen, dass die modifizierte Log-Likelihoodfunktion und die zusätzlichen getroffenen Annahmen zu konsistenten Schätzern führen. Gerade diese Konsistenz wird uns ermöglichen den fünften Teil von Satz 2.5 zu folgern (stochastische Konvergenz der Momente). Dieses Resultat wird letztendlich eine wichtige Bedeutung für die Herleitung der asymptotischen Verteilung der Teststatistik haben. Diese definieren wir nun wie folgt (das bisher unbekannte k^* wird direkt nach der Definiton näher erläutert):

Definition 2.2 (Teststatistik)**.** Für ein bel. $k \geq k^*$, die Stichprobengröße $n \in \mathbb{N}$ und die modifizierten Maximumlikelihood Schätzer $\widehat{G} := \widehat{G}^{(k)}$ und $\widehat{G}_0 := \widehat{G}^{(2)}$ definieren wir die Teststatistik durch

$$R_n := 2 \cdot \left(l_n(\widehat{G}) - l_n(\widehat{G}_0) \right). \tag{2.8}$$

Wir werden nun begründen, wie die untere Schranke k^* für k zu wählen ist. Dazu betrachten wir Darstellungen von G_0 im Raum $\mathcal{M}_k$. Unter einer Darstellung von G_0 verstehen wir eine mixing distribution aus $\mathcal{M}_k$, die für alle $\theta \in \Theta$ mit G_0 übereinstimmt. Generell findet man unendlich viele solcher Darstellungen: Es muss $r_1 \in \{1, \cdots, k-1\}$ mal der support point θ_{01} vorkommen mit Gewichten, die in Summe mit π_0 übereinstimmen. Außerdem müssen $r_2 \in \{1, \cdots, k - r_1\}$ mit θ_{02} übereinstimmen und aufsummierte Gewichte von $1 - \pi_0$ aufweisen. Die restlichen support points können beliebig sein, müssen

aber alle Gewichte von Null haben. Eine Darstellung ist für uns optimal, wenn sie einen betragsmäßig möglichst kleinen Penaltyterm besitzt. Wie wir dazu die support points und Gewichte wählen müssen, ist erstmal unklar und wird daher im Folgenden besprochen.
Unser Ziel ist es, für festes natürliches k größer zwei (später sogar größer gleich vier - für diese erstmal allgemeine Betrachtung genügt uns $k > 2$) eine Darstellung zu finden, die einen betragsmäßig möglichst kleinen Penaltyterm besitzt. Penaltyterme sind immer negativ und somit bedeutet diese Wahl eine minimale Strafe. Eine betragsmäßig möglichst kleine Wahl bedeutet also hier den größtmöglichen Wert zu finden. Wir können also direkt Gewichte, die gleich Null sind, ausschließen, da diese der Maximierung des Penaltyterms im Wege stünden. Eine optimale Darstellung von G_0 muss also r support points $\theta_i^{(rep)}$ haben, die gleich θ_{01} sind und $k-r$ die mit θ_{02} übereinstimmen. Um die Notation übersichtlich zu halten gehen wir davon aus, dass die ersten r support points gleich θ_{01} sind. Ferner müssen somit die ersten r Gewichte in Summe mit π_0 übereinstimmen. Um den Penalty der Darstellung zu maximieren, haben wir also zwei Fragen zu beantworten:

1. Welcher Wert für r ist optimal?

2. Wie sind dann die r Gewichte der support points θ_{01} bzw. die $k-r$ von θ_{02} zu wählen?

Um dieses Problem zu lösen, halten wir zunächst r fest und minimieren den Betrag des Penaltyterms über die Wahl der Gewichte. Wir erhalten somit für jedes mögliche $r \in \{1, \dots, k-1\}$ eine optimale Aufteilung der Gewichte. Anschließend wählen wir als optimales r dasjenige, welches den betragsmäßig kleinsten Penalty erzielt.

Lemma 2.3. *Sei $k \in \mathbb{N}$ beliebig größer zwei.*

1. *Für $r \in \{1, \dots, k-1\}$ hat die Darstellung $G_0^{(k,r)}$ mit Parameter*

$$\left(\frac{\pi_0}{r}, \dots, \frac{\pi_0}{r}, \frac{1-\pi_0}{k-r}, \dots, \frac{1-\pi_0}{k-r}, \theta_{01}, \dots, \theta_{01}, \theta_{02}, \dots, \theta_{02}\right) \tag{2.9}$$

   den betragsmäßig minimal möglichen Penalty unter Darstellungen, die r-mal den support point θ_{01} besitzen. Der Penaltyterm hat den Wert

$$pen\left(G_0^{(k,r)}\right) = C_k \left[r \, \log\left(\frac{\pi_0}{r}\right) + (k-r) \, \log\left(\frac{1-\pi_0}{k-r}\right)\right].$$

2. *Für*

$$r_0(k) := \underset{r \in \{1,\dots,k-1\}}{\operatorname{argmax}} \; pen\left(G_0^{(k,r)}\right)$$

besitzt die Darstellung $G_0^{(k)} := G_0^{(k,r_0(k))}$ *den betragsmäßig minimal möglichen Penalty unter allen Darstellungen von* G_0 *in* $\mathcal{M}_k$.

Beweis. Der Penaltyterm

$$\text{pen}(G) = C_k \sum_{i=1}^{k} \log(\pi_i) = C_k \log(\pi_1 \cdot \ldots \cdot \pi_k)$$

wird betragsmäßig minimal genau dann wenn er maximal wird (da er kleiner gleich Null ist), was genau dann der Fall ist, wenn

$$\exp(\log(\pi_1 \cdot \ldots \cdot \pi_k)) = \pi_1 \cdot \ldots \cdot \pi_k$$

maximal wird. Als Nebenbedingung haben wir $\pi_1 + \cdots + \pi_r = \pi_0$ und $\pi_{r+1} + \cdots + \pi_k = 1 - \pi_0$. Wir optimieren nun diese zwei Gruppen von Gewichten mit der jeweiligen Nebenbedingung getrennt. Also ist das erste Optimierungsproblem: $\max(\pi_1 \cdot \ldots \cdot \pi_r)$ s.t. $\pi_1 + \cdots + \pi_r = \pi_0$. Diese Funktion wird unter der Nebenbedingung maximal, wenn wir die Gewichte gleichmäßig aufteilen, d.h. $\pi_1 = \ldots \pi_r = \pi_0/r$. Analog erhalten wir für das zweite Optimierungsproblem $\pi_{r+1} = \ldots \pi_k = (1-\pi_0)/(k-r)$. Für diese Gewichte ergibt sich direkt der obige Penaltyterm. □

Die optimale Wahl der Darstellung ist also bei festem k nur abhängig von π_0. Mit diesem Lemma haben wir nun die Möglichkeit zu jedem $k \in \mathbb{N}$ eine entsprechende optimale Darstellung $G_0^{(k)}$ anzugeben.
k^* aus Definition (2.2) soll nun der kleinstmögliche Wert k sein, so dass bei der optimalen Darstellung $G_0^{(k)}$ jeder der beiden wahren support points mindestens zwei mal vorkommt. Da $r_0(k)$ gerade die Anzahl von support points θ_{01} beschreibt, definieren wir

$$k^* := \min\{k : 2 \leq r_0(k) \leq k-2\}. \tag{2.10}$$

$r_0(k)$ und somit auch k^* hängen vom wahren Gewicht π_0 ab, welches später beim Testen unbekannt ist. Daher muss k^* dann auch geschätzt werden.
Im weiteren Verlauf gehen wir davon aus, dass $k \geq k^*$ ist und $G_0^{(k)}$ gewählt ist wie in Lemma 2.3 beschrieben, d.h. den Parameter

$$\left(\frac{\pi_0}{r_0(k)}, \ldots, \frac{\pi_0}{r_0(k)}, \frac{1-\pi_0}{k-r_0(k)}, \ldots, \frac{1-\pi_0}{k-r_0(k)}, \theta_{01}, \ldots, \theta_{01}, \theta_{02}, \ldots, \theta_{02}\right)$$

besitzt. Wir bezeichnen die Gewichte der optimalen Darstellung ab jetzt mit $\pi_j^{(0)}$. Mit $\widehat{G}$ und $\widehat{G}_0$ meinen wir die modifizierten Maximum Likelihoodschätzer aus Definition 2.2 und mit $\left(\widehat{\pi}_1, \ldots, \widehat{\pi}_k, \widehat{\theta}_1, \ldots, \widehat{\theta}_k\right)$ bzw. $\left(\widehat{\pi}_0, \widehat{\theta}_{01}, \widehat{\theta}_{02}\right)$ bezeichnen wir den zu $\widehat{G}$ bzw. $\widehat{G}_0$ gehörigen Parameter.

2.1.3 Konsistenz

Wie oben bereits erwähnt, hat die modifizierte Likelihoodfunktion das Ziel kleine Werte für $\widehat{\pi}_j$ zu verhindern. Das folgende Lemma sagt nun, dass mit beliebig großer Sicherheit alle $\widehat{\pi}_j$ von $\widehat{G}$ asymptotisch größer Null sind.

Lemma 2.4. *Unter den getroffenen Annahmen in Abschnitt 2.1.1 gibt es für jedes $\delta > 0$ ein $\varepsilon_\delta > 0$, so dass*

$$P\left(\widehat{\pi}_1 \geq \varepsilon_\delta, \ldots, \widehat{\pi}_k \geq \varepsilon_\delta\right) \geq 1 - \delta \ \forall \ n \geq n_0(\delta)$$

ist.

Beweis. Der Beweis gliedert sich in zwei Schritte: Zuerst begründen wir, warum der Penaltyterm $\operatorname{pen}(\widehat{G})$, wie in Formel (2.6) definiert, stochastisch beschränkt ist und anschließend leiten wir daraus die Aussage des Lemmas ab.
Mit $\overline{G}$ bezeichnen wir den gewöhnlichen Likelihoodschätzer aus $\mathcal{M}_k$ und setzen weiter $Y_n := l_n(\overline{G}) - l_n(G_0)$. Nach Dacunha-Castelle und Gassiat (vgl. [DCG99]) folgt, dass Y_n stochastisch beschränkt ist. $\operatorname{pen}(G_0^{(k)})$ kann wegen (2.9) ausgerechnet werden, liegt in $(-\infty, 0)$ und ist nicht zufällig. Somit gilt auch

$$Y_n - \operatorname{pen}(G_0^{(k)}) = O_P(1).$$

Da $\pi_j \leq 1$ sind alle Logarithmen negativ und somit die Penaltyterme kleiner gleich Null. Es genügt somit für die stochastische Beschränktheit $-\operatorname{pen}(\widehat{G})$ nach oben abzuschätzen. $\widehat{G}$ maximiert die modifizierte Likelihoodfunktion, so dass $\widetilde{l}_n^{(k)}(\widehat{G}) \geq \widetilde{l}_n^{(k)}(G_0^{(k)})$ gilt. Einfaches Auflösen gemäß Definition (2.1) liefert

$$-\operatorname{pen}(\widehat{G}) \leq l_n(\widehat{G}) - l_n(G_0) - \operatorname{pen}(G_0^{(k)}).$$

Da $\overline{G}$ der Maximum Likelihoodschätzer ist, gilt $l_n(\widehat{G}) \leq l_n(\overline{G})$ so dass wir

$$-\operatorname{pen}(\widehat{G}) \leq Y_n - \operatorname{pen}(G_0^{(k)})$$

erhalten, was nun $\operatorname{pen}(\widehat{G}) = O_P(1)$ liefert. Die Penaltyterme sind negativ und daher gilt $|\operatorname{pen}(\widehat{G})| = -C_k \sum_{j=1}^k \log(\widehat{\pi}_j)$. Aufgrund der stochastischen Beschränktheit existieren nun für beliebiges $\delta > 0$ Konstanten $M(\delta) > 0$ und $n_0(\delta) \in \mathbb{N}$, so dass

$$P\left(-\sum_{i=1}^k \log(\widehat{\pi}_i) \geq M(\delta)\right) < \delta \text{ für alle } n \geq n_0(\delta).$$

Aus dieser Ungleichung lässt sich folgern, dass

$$P\left(\widehat{\pi}_1 \geq e^{-M(\delta)}, \ldots, \widehat{\pi}_k \geq e^{-M(\delta)}\right) \geq 1 - \delta.$$

Setzt man nun $\varepsilon(\delta) := e^{-M(\delta)}$, so erhält man die Behauptung. □

Als nächstes zerlegen wir nun unsere geschätzte mixing distribution $\widehat{G}$. Dazu definieren wir zunächst $\theta_{0,mid} := (\theta_{01}+\theta_{02})/2$ und setzen $\widehat{\pi} := \widehat{G}(\theta_{0,mid})$. Somit ist $\widehat{\pi}$ die Summe von Gewichten $\widehat{\pi}_j$, deren Komponentenparameter $\widehat{\theta}_j$ kleiner gleich $\theta_{0,mid}$ sind. Die zugehörige Indexmenge bezeichnen wir mit J und können nun $\widehat{G}$ zerlegen über

$$\widehat{G}(\theta) = \widehat{\pi}\ \widehat{G}_1(\theta) + (1-\widehat{\pi})\ \widehat{G}_2(\theta), \tag{2.11}$$

wobei

$$\widehat{G}_1(\theta) := \sum_{j \in J} \widehat{\pi}_j^*\ I(\widehat{\theta}_j \leq \theta) \tag{2.12}$$

also eine mixing distribution bestehend aus Komponentenparametern $\widehat{\theta}_j \leq \theta_{0,mid}$ ist. Die entsprechenden Gewichte $\widehat{\pi}_j^*$ müssen dabei umnormiert werden, so dass $\widehat{G}_1(\theta_{0,mid}) = 1$ gilt, d.h. $\widehat{\pi}_j^* := \widehat{\pi}_j/\widehat{\pi}$. Für $\widehat{G}_2(\theta)$ wählt man dementsprechend die restlichen Komponentenparameter $\widehat{\theta}_j$ und normiert entsprechend die zugehörigen Gewichte um.
Wir werden nun zeigen, dass $\widehat{G}$ gegen die optimale Darstellung von G_0 konvergiert. Anschaulich bedeutet das, dass $\widehat{G}$ asymptotisch nur drei entscheidende Parameter besitzt. Es gibt erstens eine Gruppe von support points, die in Wahrscheinlichkeit gegen θ_{01} konvergieren, zweitens haben sie zugehörige Gewichte, welche in Summe gegen π_0 konvergieren und drittens laufen die restlichen support points gegen θ_{02} (gemeint ist hier jeweils Konvergenz in Wahrscheinlichkeit).

Satz 2.5 (Konsistenz)**.** *Setzt man $G_l(\theta) := I(\theta_{0l} \leq \theta)$, $l = 1, 2$, so gelten unter den getroffenen Annahmen in Abschnitt 2.1.1 die folgenden Aussagen:*

1. $|\widehat{\pi} - \pi_0| = o_P(1)$
2. $|\widehat{G}_l(\theta) - G_l(\theta)| = o_P(1)$, $l = 1, 2$ *und* $\theta_{01} \neq \theta \neq \theta_{02}$
3. $|\widehat{\theta}_j - \theta_{01}| = o_P(1)$ *für* $j \in J$ *und* $|\widehat{\theta}_j - \theta_{02}| = o_P(1)$ *für* $j \in \{1, \ldots, k\} \setminus J$.
4. $|\widehat{\pi}_j - \pi_j^{(0)}| = o_P(1)$ *für alle* $j = 1, \ldots, k$
5. $\int_\Theta |\theta - \theta_{0l}|^r\, d\widehat{G}_l(\theta) = o_P(1)$ *mit* $r > 0$ *und* $l = 1, 2$

Beweis. Aufgrund der getroffenen Annahme 2.1 ergibt sich die Konsistenz des modifizierten Maximum Likelihoodschätzers $\widehat{G}$ in dem Sinne, wie sie von Leroux in [Ler92a] definiert wird, d.h. mit Wahrscheinlichkeit eins konvergiert die geschätzte mixing distribution $\widehat{G}$ punktweise an jeder Stetigkeitsstelle von G_0 gegen die wahre mixing distribution G_0. Wir erhalten also, dass fast sicher gilt:

$$\widehat{G}(\theta) \xrightarrow{P} G_0(\theta) \text{ für } \theta_{01} \neq \theta \neq \theta_{02}. \tag{2.13}$$

$\theta_{0,mid}$, wie im Kontext der Zerlegung von $\widehat{G}$ definiert, ist eine Stetigkeitsstelle von G_0, so dass wir umittelbar

$$\widehat{\pi} = \widehat{G}(\theta_{0,mid}) \xrightarrow{P} G_0(\theta_{0,mid}) = \pi_0$$

beobachten, womit die erste Behauptung gezeigt ist.
Ferner gilt für $\theta < \theta_{0,mid}$, $\theta \neq \theta_{01}$

$$\widehat{G}_1(\theta) = \widehat{\pi}^{-1}\, \widehat{\pi}\, \widehat{G}_1(\theta) = \widehat{\pi}^{-1}\, \widehat{G}(\theta) \xrightarrow{P} \pi_0^{-1}\, G_0(\theta) = \pi_0^{-1}\, \pi_0\, G_1(\theta) = G_1(\theta).$$

Da außerdem

$$\widehat{G}_1(\theta) = G_1(\theta) = 1 \text{ für } \theta \geq \theta_{0,mid}$$

ist die zweite Behauptung für $l = 1$ gezeigt. Der Fall $l = 2$ verläuft völlig analog, und somit können wir an dieser Stelle die zweite Behauptung insgesamt folgern.
Dank Lemma 2.4 können wir davon ausgehen, dass alle Gewichte $\widehat{\pi}_j$ von $\widehat{G}$ echt größer Null sind. Aufgrund der Gültigkeit von Behauptung zwei bzw. (2.13) wäre es also nur möglich, dass ein support point $\widehat{\theta}_j$ nicht gegen einen der wahren Parameter θ_{01} oder θ_{02} konvergiert, falls das zugehörige Gewicht gegen Null konvergiert. Dieser Fall wir nun aber gerade durch Lemma 2.4 ausgeschlossen. Daher folgt aus der zweiten Behauptung bzw. aus (2.13) direkt die Dritte.
Die erste und dritte Aussage bedeuten also, dass $\widehat{G}$ gegen eine Darstellung von G_0 konvergiert. Erstmal unklar ist, wie viele support points von $\widehat{G}$ gegen θ_{01} konvergieren, d.h. wie die Indexmenge J asymptotisch aussieht. Da $\widehat{G}$ die modifizierte Likelihoodfunktion $l_n(\cdot) + \text{pen}(\cdot)$ optimiert und alle Darstellungen denselben Likelihoodwert aufweisen, wird $\widehat{G}$ gegen eine Darstellung mit betragsmäßig möglichst kleinem Penaltyterm konvergieren. Diese ist unsere optimale Darstellung $G_0^{(k)}$. Die Anzahl der Elemente in J konvergiert also gegen $r_0(k)$ und alle Gewichte von $\widehat{G}$ konvergieren somit stochastisch gegen die der Darstellung $G_0^{(k)}$, d.h. $\widehat{\pi}_j \xrightarrow{P} \pi_j^{(0)}$.
Gleichung (2.14) zeigt die fünfte Behauptung exemplarisch für $l = 1$ und beendet somit den Beweis

$$\int_\Theta |\theta - \theta_{01}|^r \, d\widehat{G}_1(\theta) = \sum_{j \in J} \widehat{\pi}_j^* \cdot \underbrace{|\widehat{\theta}_j - \theta_{01}|^r}_{=o_P(1) \text{ wg. 3.}} = o_P(1). \tag{2.14}$$

Der Stern an $\widehat{\pi}_j^*$ deutet dabei an, dass die Gewichte umnormiert werden mussten. □

Besonders die fünfte Aussage des Satzes wird im folgenden Kapitel eine wichtige Rolle spielen, denn genau solche Terme tauchen hier in Restgliedern von Taylorentwicklungen auf und helfen somit die asymptotische Vernachlässigbarkeit zu beweisen.

2.1.4 Asymptotische Verteilung der Teststatistik

In diesem Abschnitt werden wir die asymptotische Verteilung der Teststatistik R_n herleiten. Dazu zerlegen wir zunächst R_n (vgl. (2.8)) gemäß

$$R_n = R_{1n} - R_{0n},$$

wobei $R_{1n} := 2 \cdot \Big(l_n(\widehat{G}) - l_n(G_0) \Big)$ und $R_{0n} := 2 \cdot \Big(l_n(\widehat{G}_0) - l_n(G_0) \Big)$.
Im ersten Schritt werden wir nur R_{1n} betrachten und mit Hilfe einer Taylorentwicklung dessen Grenzverhalten näher untersuchen.

Definition 2.6. Wir definieren nun die folgenden Größen:

$$\begin{aligned}
\delta_i &:= \frac{f(X_i;\widehat{G}) - f(X_i;G_0)}{f(X_i;G_0)},\ i = 1,\dots,n \\
\widehat{m}_{lj} &:= \int_\Theta (\theta - \theta_{0j})^l \, d\widehat{G}_j(\theta),\ l = 1,2,3 \text{ und } j = 1,2 \\
\hat{\boldsymbol{t}} &:= \left(\widehat{\pi} - \pi_0, \widehat{\pi}\ \widehat{m}_{11}, (1-\widehat{\pi})\ \widehat{m}_{12}, \widehat{\pi}\ \frac{\widehat{m}_{21}}{2}, (1-\widehat{\pi})\ \frac{\widehat{m}_{22}}{2} \right)' \\
\Delta_i &:= \frac{f(X_i,\theta_{01}) - f(X_i,\theta_{02})}{f(X_i;G_0)},\ i = 1,\dots,n \\
\boldsymbol{b}_i &:= (\Delta_i, Y_i'(\theta_{01}), Y_i'(\theta_{02}), Y_i''(\theta_{01}), Y_i''(\theta_{02}))',\ i = 1,\dots,n \\
\boldsymbol{\beta} &:= \sum_{i=1}^n \boldsymbol{b}_i \text{ und } \boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2') \text{ mit } \boldsymbol{\beta}_1 \in \mathbb{R}^3 \\
\boldsymbol{B} &:= \sum_{i=1}^n \boldsymbol{b}_i\, \boldsymbol{b}_i' \text{ und } \boldsymbol{B} = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array} \right) \text{ mit } B_{11} \in \mathbb{R}^{3\times 3} \\
C_n &:= \frac{2}{3} \sum_{i=1}^n \left(\boldsymbol{b}_i' \hat{\boldsymbol{t}} \right)^3
\end{aligned}$$

Die Größen $\widehat{m}_{lj}$ und $\hat{\boldsymbol{t}}$ wurden in Abhängigkeit von $\widehat{G}$ definiert. Sprechen wir von $m_{lj}(G)$ bzw. von $t(G)$ meinen wir obige Definitionen, jedoch für ein beliebiges $G \in \mathcal{M}_k$, d.h.

$$t(G) = \left(\pi(G) - \pi_0, \pi_0\ m_{11}(G), (1-\pi_0)\ m_{12}(G), \pi_0\ \frac{m_{21}(G)}{2}, (1-\pi_0)\ \frac{m_{22}(G)}{2} \right)',$$

wobei wir $m_{lj}(G)$ erhalten, indem wir bzgl. G_j integrieren und G wieder durch G_1, G_2 darstellen wie in (2.11). Ferner bezeichnen wir die Menge aller $t(G)$ mit

$$T := \{t(G) | G \in \mathcal{M}_k\}.$$

Von diesen Definitionen ausgehend werden wir nun in einem Lemma zuerst eine asymptotische Grenze von R_{1n} herleiten und anschließend zeigen, dass diese Grenze auch angenommen wird.

Lemma 2.7. *Unter den getroffenen Annahmen in Abschnitt 2.1.1 gilt:*

$$R_{1n} = \boldsymbol{\beta}_1' B_{11}^{-1} \boldsymbol{\beta}_1 + \sup_{t \in \mathbb{R}^2_{\geq}} \left(2\, \widetilde{\boldsymbol{\beta}}_2' t - t' \widetilde{B}_{22} t \right) + o_P(1),$$

wobei $\widetilde{\boldsymbol{\beta}}_2' := \boldsymbol{\beta}_2' - \boldsymbol{\beta}_1' B_{11}^{-1} B_{12}$ *und* $\widetilde{B}_{22} := B_{22} - B_{21} B_{11}^{-1} B_{12}$.

Überblick über die Vorgehensweise des Beweises:

1. Ausgangspunkt des Beweises ist die Abschätzung $R_{1n} \leq 2 \sum_{i=1}^n \delta_i - \sum_{i=1}^n \delta_i^2 + \frac{2}{3} \sum_{i=1}^n \delta_i^3$.
2. Eine Taylorentwicklung bis zur Ordnung zwei liefert $2 \sum_{i=1}^n \delta_i = 2 \left(\boldsymbol{\beta}' \hat{\boldsymbol{t}} + \varepsilon_n \right)$ mit $\varepsilon_n = o_P(\hat{\boldsymbol{t}}' \boldsymbol{B} \hat{\boldsymbol{t}}) + o_P(1)$.
3. Die Fehler bei einer Entwicklung für den quadratischen bzw. kubischen Teil von (2.15) sind höchstens von der Ordnung $O_P(\varepsilon_n)$, s.d. $R_{1n} \leq 2\, \boldsymbol{\beta}' \hat{\boldsymbol{t}} - \hat{\boldsymbol{t}}' \boldsymbol{B} \hat{\boldsymbol{t}} + C_n + O_P(\varepsilon_n)$.
4. Weil $C_n = O_P(\hat{\boldsymbol{t}}' \boldsymbol{B} \hat{\boldsymbol{t}})$ gilt $R_{1n} \leq \sup_{t \in T} (2\, \boldsymbol{\beta}' t - t' \boldsymbol{B} t) + o_P(1)$. Hiermit folgern wir $R_{1n} \leq \widetilde{\boldsymbol{\beta}}_1' B_{11}^{-1} \widetilde{\boldsymbol{\beta}}_1 + \sup_{t \in \mathbb{R}^2_{\geq}} \left(2\, \widetilde{\boldsymbol{\beta}}_2' t - t' \widetilde{B}_{22} t \right) + o_P(1)$.
5. Im letzten Schritt zeigen wir dann die Gleichheit.

Beweis. Schritt 1
Zuerst begründen wir, warum δ_i bzw. deren Summe eine wichtige Rolle für die Untersuchung von R_{1n} spielt. Mit obiger Definition von δ_i ergibt sich

$$\begin{aligned} R_{1n} &= 2 \cdot \left(l_n(\widehat{G}) - l_n(G_0) \right) = 2 \cdot \left(\sum_{i=1}^n \log(f(X_i; \widehat{G})) - \sum_{i=1}^n \log(f(X_i; G_0)) \right) \\ &= 2 \cdot \sum_{i=1}^n \log \left(\frac{f(X_i; \widehat{G})}{f(X_i; G_0)} \right) = 2 \cdot \sum_{i=1}^n \log(1 + \delta_i). \end{aligned}$$

Da weiter $\log(1+x) \leq x - x^2/2 + x^3/3$ erhalten wir zusammen

$$R_{1n} = 2\sum_{i=1}^{n} \log(1+\delta_i) \leq 2\sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \delta_i^2 + \frac{2}{3}\sum_{i=1}^{n} \delta_i^3. \tag{2.15}$$

Schritt 2

Als nächstes betrachten wir δ_i. Setzen wir für $\widehat{G}(\theta)$ die Zerlegung $\widehat{\pi}\ \widehat{G}_1 + (1-\widehat{\pi})\ \widehat{G}_2$ und für G_0 die wahre Form $\pi_0\ G_1 + (1-\pi_0)\ G_2$ ein, so erhalten wir

$$\delta_i = \frac{\widehat{\pi}\ f(X_i;\widehat{G}_1) + (1-\widehat{\pi})\ f(X_i;\widehat{G}_2) - \pi_0\ f(X_i,\theta_{01}) - (1-\pi_0)\ f(X_i,\theta_{02})}{f(X_i;G_0)}.$$

Anschließend addieren wir $\pm\ \widehat{\pi}\ f(X_i,\theta_{01})$ sowie $\pm\ (1-\widehat{\pi})\ f(X_i,\theta_{02})$ und erreichen durch Ausklammern

$$\delta_i = (\widehat{\pi}-\pi_0)\ \Delta_i + \widehat{\pi}\ \frac{f(X_i;\widehat{G}_1) - f(X_i,\theta_{01})}{f(X_i;G_0)} + (1-\widehat{\pi})\ \frac{f(X_i;\widehat{G}_2) - f(X_i,\theta_{02})}{f(X_i;G_0)}. \tag{2.16}$$

Wie (2.15) andeutet sind wir an der Summe über die δ_i interessiert. Diese können wir nun direkt mittels (2.16) charakterisieren über

$$\sum_{i=1}^{n}\delta_i = (\widehat{\pi}-\pi_0)\sum_{i=1}^{n}\Delta_i + \widehat{\pi}\sum_{i=1}^{n}\frac{f(X_i;\widehat{G}_1) - f(X_i,\theta_{01})}{f(X_i;G_0)} + (1-\widehat{\pi})\sum_{i=1}^{n}\frac{f(X_i;\widehat{G}_2) - f(X_i,\theta_{02})}{f(X_i;G_0)}. \tag{2.17}$$

Als nächstes sehen wir uns nun $\sum_{i=1}^{n}\frac{f(X_i;\widehat{G}_1)-f(X_i,\theta_{01})}{f(X_i;G_0)}$ an und erinnern zunächst an die Gültigkeit von $f(x;\widehat{G}_1) = \int_\Theta f(x,\theta)\ d\widehat{G}_1(\theta)$. Diese Gleichheit überträgt sich natürlich auch auf die Summation an verschiedenen Stellen X_i, so dass

$$\sum_{i=1}^{n}\frac{f(X_i;\widehat{G}_1) - f(X_i,\theta_{01})}{f(X_i;G_0)} = \int_\Theta \sum_{i=1}^{n}\frac{f(X_i,\theta) - f(X_i,\theta_{01})}{f(X_i;G_0)}\ d\widehat{G}_1(\theta). \tag{2.18}$$

Für den Integranden führen wir nun eine Taylorentwicklung um die Stelle θ_{01} durch (per Definition ist dieser nichts anderes als die Summe über alle Y_{i1}) und erhalten

$$\sum_{i=1}^{n}\frac{f(X_i,\theta) - f(X_i,\theta_{01})}{f(X_i;G_0)} = \sum_{i=1}^{n} Y_{i1}(\theta)$$

$$= (\theta-\theta_{01})\sum_{i=1}^{n} Y_i'(\theta_{01}) + \frac{(\theta-\theta_{01})^2}{2}\sum_{i=1}^{n} Y_i''(\theta_{01}) + r_n^{(1)}(\theta),$$

wobei

$$r_n^{(1)}(\theta) = \frac{1}{6}(\theta-\theta_{01})^3\sum_{i=1}^{n} Y_i'''(\eta_1(\theta)) \text{ mit } \eta_1(\theta) \text{ zwischen } \theta \text{ und } \theta_{01}$$

das Restglied der Taylorentwicklung ist. Setzen wir dieses Ergebnis in (2.18) ein, folgt

$$\sum_{i=1}^{n} \frac{f(X_i;\widehat{G}_1) - f(X_i,\theta_{01})}{f(X_i;G_0)} = \int_\Theta \left((\theta-\theta_{01}) \sum_{i=1}^{n} Y_i'(\theta_{01}) + \frac{(\theta-\theta_{01})^2}{2} \sum_{i=1}^{n} Y_i''(\theta_{01}) + r_n^{(1)}(\theta) \right) d\widehat{G}_1(\theta)$$
$$= \widehat{m}_{11} \sum_{i=1}^{n} Y_i'(\theta_{01}) + \frac{\widehat{m}_{21}}{2} \sum_{i=1}^{n} Y_i''(\theta_{01}) + \int_\Theta r_n^{(1)}(\theta)\, d\widehat{G}_1(\theta) \tag{2.19}$$

Die Messbarkeit von $r_n^{(1)}$ lässt sich mit der Integraldarstellung des Restglieds begründen: Da $r_n^{(1)}(\theta) = \int_{\theta_{01}}^{\theta} (t-\theta_{01})^2 1/2 \sum_{i=1}^{n} Y_i'''(t)dt$ mit stetigen Integranden, ist $r_n^{(1)}(\theta)$ stetig, also insbesondere messbar.
Völlig analog zur Vorgehensweise, die zu (2.19) geführt hat, folgern wir für $\widehat{G}_2$

$$\sum_{i=1}^{n} \frac{f(X_i;\widehat{G}_2) - f(X_i,\theta_{02})}{f(X_i;G_0)} = \widehat{m}_{21} \sum_{i=1}^{n} Y_i'(\theta_{02}) + \frac{\widehat{m}_{22}}{2} \sum_{i=1}^{n} Y_i''(\theta_{02}) + \int_\Theta r_n^{(2)}(\theta)\, d\widehat{G}_2(\theta). \tag{2.20}$$

Wir setzen nun (2.19) und (2.20) in die Summendarstellung der δ_i also (2.17) ein. Damit ergibt sich

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} \Bigg[(\widehat{\pi}-\pi_0)\Delta_i + \widehat{\pi}\widehat{m}_{11} Y_i'(\theta_{01}) + \widehat{\pi}\frac{\widehat{m}_{21}}{2} Y_i''(\theta_{01}) +$$
$$(1-\widehat{\pi})\widehat{m}_{12} Y_i'(\theta_{02}) + (1-\widehat{\pi})\frac{\widehat{m}_{22}}{2} Y_i''(\theta_{02}) \Bigg] +$$
$$\widehat{\pi} \int_\Theta r_n^{(1)}(\theta)\, d\widehat{G}_1(\theta) + (1-\widehat{\pi}) \int_\Theta r_n^{(2)}(\theta)\, d\widehat{G}_2(\theta).$$

Die Definition von $\varepsilon_n := \widehat{\pi} \int_\Theta r_n^{(1)}(\theta)\, d\widehat{G}_1(\theta) + (1-\widehat{\pi}) \int_\Theta r_n^{(2)}(\theta)\, d\widehat{G}_2(\theta)$ liefert uns schließlich

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} \boldsymbol{b}_i'\hat{\boldsymbol{t}} + \varepsilon_n = \boldsymbol{\beta}'\hat{\boldsymbol{t}} + \varepsilon_n. \tag{2.21}$$

Für die Untersuchung der Ordnung des Fehlers ε_n betrachten wir zunächst für $l = 1, 2$ die Integrale der Restglieder $r_n^{(l)}(\theta)$. Für $l = 1$ gilt

$$\int_\Theta r_n^{(1)}(\theta)\, d\widehat{G}_1(\theta) = \sum_{j\in J} \left(\frac{1}{6} \left(\widehat{\theta}_j - \theta_{01} \right)^3 \sum_{i=1}^{n} Y_i''' \left(\eta_1(\widehat{\theta}_j) \right) \right). \tag{2.22}$$

Da $\eta_1(\widehat{\theta}_j)$ zwischen $\widehat{\theta}_j$ und θ_{01} liegt und $\widehat{\theta}_j \xrightarrow{P} \theta_{01}$ nach Lemma 2.5, muss auch $\eta_1(\widehat{\theta}_j)$ konsistent sein für θ_{01}. Daher ist nach Gleichung (2.5) $n^{-1/2} \sum_{i=1}^{n} Y_i''' \left(\eta_1(\widehat{\theta}_j) \right) = O_P(1)$ und folglich gilt

$$\int_\Theta r_n^{(1)}(\theta)\, d\widehat{G}_1(\theta) = O_P(n^{1/2}) \sum_{j\in J} \left(\frac{1}{6} \left(\widehat{\theta}_j - \theta_{01} \right)^3 \right) = O_P(n^{1/2}) \underbrace{\int_\Theta (\theta-\theta_{01})^3\, d\widehat{G}_1(\theta)}_{=:\widehat{m}_{31}}.$$

Für das Integral über $r_n^{(2)}(\theta)$ ergibt sich diese Formulierung analog, so dass wir nun für ε_n die folgende Abschätzung durchführen können:

$$\begin{aligned}
|\varepsilon_n| &\leq \left| \int_\Theta r_n^{(1)}(\theta)\ d\widehat{G}_1(\theta) \right| + \left| \int_\Theta r_n^{(2)}(\theta)\ d\widehat{G}_2(\theta) \right| \\
&\leq O_P(n^{1/2}) \left(\left| \int_\Theta (\theta - \theta_{01})^3\ d\widehat{G}_1(\theta) \right| + \left| \int_\Theta (\theta - \theta_{02})^3\ d\widehat{G}_1(\theta) \right| \right) \\
&\leq O_P(n^{1/2}) \left(\int_\Theta |\theta - \theta_{01}|^3\ d\widehat{G}_1(\theta) + \int_\Theta |\theta - \theta_{02}|^3\ d\widehat{G}_2(\theta) \right) \\
&= O_P\left(n^{1/2}\right)\ \ (\|\widehat{m}_{31}\| + \|\widehat{m}_{32}\|)\,, \text{ wobei } \|\widehat{m}_{3j}\| := \int_\Theta |\theta - \theta_{0j}|^3\ d\widehat{G}_j(\theta). \qquad (2.23)
\end{aligned}$$

Als nächstes möchten wir $\|\widehat{m}_{3l}\|$ jeweils über $\widehat{m}_{2l}$ (für $l = 1, 2$) darstellen. Es gilt

$$\begin{aligned}
\|\widehat{m}_{3l}\| &= \sum_{j\in J} \widehat{\pi}_j^*\ \left|\widehat{\theta}_j - \theta_{0l}\right|^3 = \sum_{j\in J} \widehat{\pi}_j^*\ \left|\widehat{\theta}_j - \theta_{0l}\right|^2\ \underbrace{\left|\widehat{\theta}_j - \theta_{0l}\right|^1}_{=o_P(1)} \\
&= o_P(1)\ \sum_{j\in J} \widehat{\pi}_j^*\ \left|\widehat{\theta}_j - \theta_{0l}\right|^2 = o_P(1)\ \sum_{j\in J} \widehat{\pi}_j^*\ \left(\widehat{\theta}_j - \theta_{0l}\right)^2 \\
&= o_P(1)\ \widehat{m}_{2l}. \qquad (2.24)
\end{aligned}$$

Setzen wir nun dieses Ergebnis in (2.23) ein, erhalten wir

$$|\varepsilon_n| \leq n^{1/2}\ (\widehat{m}_{21} + \widehat{m}_{22})\ o_P(1) \leq \left[1 + n\left(\widehat{m}_{21}^2 + \widehat{m}_{22}^2\right)\right]\ o_P(1). \qquad (2.25)$$

Da $\{X_i\}_{i=1,\ldots,n}$ eine Folge von unabhängigen identisch verteilten Zufallsvariablen ist, sind auch die $\boldsymbol{b}_i \in \mathbb{R}^5$ iid. Ferner haben alle Komponenten von $\boldsymbol{b}_i$ den Erwartungswert Null und die Kovarianzmatrix Σ, so dass nach dem starken Gesetz der großen Zahlen $n^{-1}\boldsymbol{B}$ fast sicher gegen $\Sigma = \mathrm{Cov}(b_1)$ konvergiert. Nach Annahme 2.3 ist Σ positiv definit. Dies gilt aufgrund folgender Begründung asymptotisch ebenfalls für $n^{-1}\boldsymbol{B}$:
Σ ist positiv definit und somit gilt $\det(\Sigma) \geq \varepsilon > 0$. Da sich die fast sichere Konvergenz auf die Determinante überträgt, liegt $\det(n^{-1}\boldsymbol{B})$ ab einem gewissen $n_0 \in \mathbb{N}$ fast sicher in einer δ Umgebung (mit $\delta < \varepsilon$) von $\det(\Sigma)$. Somit ist die Determinante von $n^{-1}\boldsymbol{B}$ asymptotisch positiv und daher $n^{-1}\boldsymbol{B}$ asymptotisch positiv definit.
Die quadratische Form über $n^{-1}\boldsymbol{B}$ definiert also für hinreichend große n eine Norm im $\mathbb{R}^5$. Da $\sum_{j=1}^5 t_j^2$ ebenfalls eine Norm definiert, erhalten wir aufgrund der Äquivalenz von Normen

$$\begin{aligned}
n(\widehat{m}_{21}^2 + \widehat{m}_{22}^2) &\leq n\left((\widehat{\pi} - \pi_0)^2 + \widehat{m}_{21}^2 + \widehat{m}_{22}^2 + \widehat{m}_{12}^2 + \widehat{m}_{11}^2\right) \\
&\leq n\left(c\ \hat{\boldsymbol{t}}'\ (n^{-1}\boldsymbol{B})\ \hat{\boldsymbol{t}}\right) = O_P(\hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}}). \qquad (2.26)
\end{aligned}$$

Mit (2.25) folgt schließlich

$$|\varepsilon_n| \leq \left[1 + O_P(\hat{\boldsymbol{t}}' \boldsymbol{B} \hat{\boldsymbol{t}})\right] \; o_P(1) = o_P(1) + o_P(\hat{\boldsymbol{t}}' \boldsymbol{B} \hat{\boldsymbol{t}}). \tag{2.27}$$

Insgesamt haben wir in diesem Schritt gezeigt, dass

$$\sum_{i=1}^{n} \delta_i = (\boldsymbol{\beta}' \hat{\boldsymbol{t}} + \varepsilon_n) \text{ mit } \varepsilon_n = o_P(1) + o_P(\hat{\boldsymbol{t}}' \boldsymbol{B} \hat{\boldsymbol{t}}).$$

Schritt 3

Bisher haben wir uns nur mit der Summe der δ_i beschäftigt. Um aber unsere Ausgangsungleichung (2.15) benutzen zu können, müssen wir auch noch $\sum_{i=1}^{n} \delta_i^2$ bzw. $\sum_{i=1}^{n} \delta_i^3$ betrachten. Wir werden nun exemplarisch für die Summe der Quadrate begründen, warum die Fehler bei der Taylorentwicklung höchstens von der Ordnung von ε_n sind.
Im zweiten Schritt des Beweises führten wir die Taylorentwicklung direkt für $\sum_{i=1}^{n} \delta_i$ durch und erhielten einen Fehler ε_n. Wir hätten auch für jedes δ_i die beschriebene Vorgehenweise wählen können und hätten dann n Fehler ε_{in} erhalten, die in Summe mit ε_n übereinstimmen. Somit gilt

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} \left(\boldsymbol{b}_i' \hat{\boldsymbol{t}} + \varepsilon_{in}\right)$$

mit

$$\varepsilon_{in} = \widehat{\pi} \int_{\Theta} \frac{1}{6} (\theta - \theta_{01})^3 Y_i'''(\eta_{1i}(\theta)) \; d\widehat{G}_1(\theta) + (1 - \widehat{\pi}) \int_{\Theta} \frac{1}{6} (\theta - \theta_{02})^3 Y_i'''(\eta_{2i}(\theta)) \; d\widehat{G}_2(\theta).$$

Diese Darstellung ermöglicht es nun direkt

$$\sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} \left(\left(\boldsymbol{b}_i' \hat{\boldsymbol{t}}\right)^2 + \boldsymbol{b}_i' \hat{\boldsymbol{t}} \; \varepsilon_{in} + (\varepsilon_{in})^2 \right) = \hat{\boldsymbol{t}}' \boldsymbol{B} \hat{\boldsymbol{t}} + \sum_{i=1}^{n} \left(\boldsymbol{b}_i' \hat{\boldsymbol{t}} \; \varepsilon_{in} + (\varepsilon_{in})^2 \right) =: \hat{\boldsymbol{t}}' \boldsymbol{B} \hat{\boldsymbol{t}} + \varepsilon_n^{(Q)}$$

anzugeben. Ferner gilt

$$\varepsilon_{in} = \frac{1}{6} \, \widehat{\pi} \sum_{j \in J} (\theta_{01} - \widehat{\theta}_j)^3 Y_i'''(\eta_{1i}(\widehat{\theta}_j)) + \frac{1}{6} \, (1 - \widehat{\pi}) \sum_{j \in J \setminus \{1, \ldots, k^*\}} (\theta_{02} - \widehat{\theta}_j)^3 Y_i'''(\eta_{2i}(\widehat{\theta}_j)).$$

Wir können also die folgende Abschätzung vornehmen:

$$\begin{aligned}
|\varepsilon_{in}| &\leq \left| \sum_{j \in J} (\theta_{01} - \widehat{\theta}_j)^3 Y_i'''(\eta_{1i}(\widehat{\theta}_j)) \right| + \left| \sum_{j \in J \setminus \{1, \ldots, k^*\}} (\theta_{02} - \widehat{\theta}_j)^3 Y_i'''(\eta_{2i}(\widehat{\theta}_j)) \right| \\
&\leq \sum_{j \in J} \left| \theta_{01} - \widehat{\theta}_j \right|^3 \left| Y_i'''(\eta_{1i}(\widehat{\theta}_j)) \right| + \sum_{j \in J \setminus \{1, \ldots, k^*\}} \left| \theta_{02} - \widehat{\theta}_j \right|^3 \left| Y_i'''(\eta_{2i}(\widehat{\theta}_j)) \right| \\
&\leq \sum_{j \in J} \left| \theta_{01} - \widehat{\theta}_j \right|^3 g\left(X_i\right)^{1/3} + \sum_{j \in J \setminus \{1, \ldots, k^*\}} \left| \theta_{02} - \widehat{\theta}_j \right|^3 g\left(X_i\right)^{1/3} \\
&= g\left(X_i\right)^{1/3} \; \left(\|\widehat{m}_{31}\| + \|\widehat{m}_{32}\| \right)
\end{aligned} \tag{2.28}$$

Um über die Ordnung von $\varepsilon_n^{(Q)}$ eine Aussage treffen zu können, betrachten wir nun zuerst die Summe über die quadrierten Fehler und benutzen dabei Abschätzung (2.28):

$$\begin{aligned}\sum_{i=1}^{n}(\varepsilon_{in})^2 = \sum_{i=1}^{n}\left(|\varepsilon_{in}|\ |\varepsilon_{in}|\right) &\overset{(2.28)}{\leq} \left(\|\widehat{m}_{31}\| + \|\widehat{m}_{32}\|\right)^2 \sum_{i=1}^{n} g\left(X_i\right)^{2/3} \\ &\leq \left(\|\widehat{m}_{31}\| + \|\widehat{m}_{32}\|\right)^2 \sum_{i=1}^{n}\left(1 + g\left(X_i\right)\right) = \left(\|\widehat{m}_{31}\| + \|\widehat{m}_{32}\|\right)^2 O_P(n).\end{aligned}$$

Dabei gilt das letzte Gleichheitszeichen, da nach dem starken Gesetz der großen Zahlen $n^{-1}\ \sum_{i=1}^{n} g\left(X_i\right)$ fast sicher gegen $E\left(g\left(X_1\right)\right) < \infty$ konvergiert. Wir schätzen den resultierenden Ausdruck wie folgt weiter ab:

$$\begin{aligned}\sum_{i=1}^{n}(\varepsilon_{in})^2 &\leq O_P(n)\ \left(\|\widehat{m}_{31}\| + \|\widehat{m}_{32}\|\right)^2 \overset{(2.24)}{=} o_P(n)\ \left(\widehat{m}_{21} + \widehat{m}_{22}\right)^2 \\ &= o_P(n)\ \left(\widehat{m}_{21}^2 + \widehat{m}_{22}^2\right).\end{aligned}$$

Mit (2.26) erhalten wir nun schließlich

$$\sum_{i=1}^{n}(\varepsilon_{in})^2 = o_P(\hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}}).$$

Um die Ordnung des Fehlers $\varepsilon_n^{(Q)}$ angeben zu können, müssen wir zusätzlich noch $\sum_{i=1}^{n}\boldsymbol{b}_i'\hat{\boldsymbol{t}}\ \varepsilon_{in}$ betrachten. Hierbei handelt es sich um fünf Summen, die sich aus

$$\begin{aligned}\boldsymbol{b}_i'\hat{\boldsymbol{t}} &= \left(\Delta_1, Y_i'(\theta_{01}), Y_i'(\theta_{02}), Y_i''(\theta_{01}), Y_i''(\theta_{02})\right)\ \left(\hat{t}_1, \ldots, \hat{t}_5\right)' \\ &= \hat{t}_1\ \Delta_i + \hat{t}_2\ Y_i'(\theta_{01}) + \hat{t}_3\ Y_i'(\theta_{02}) + \hat{t}_4\ Y_i''(\theta_{01}) + \hat{t}_5\ Y_i''(\theta_{02})\end{aligned}$$

ergeben. Wir schätzen exemplarisch die vierte Summe ab ($\hat{t}_4 = 0.5\ \widehat{\pi}\ \widehat{m}_{21}$):

$$\begin{aligned}\left|\sum_{i=1}^{n}\hat{t}_4\ Y_i''(\theta_{01})\ \varepsilon_{in}\right| &\leq \widehat{m}_{21}\ \sum_{i=1}^{n}|Y_i''(\theta_{01})|\ |\varepsilon_{in}| \leq \widehat{m}_{21}\ \sum_{i=1}^{n} g\left(X_i\right)^{1/3}|\varepsilon_{in}| \\ &\overset{(2.28)}{\leq} \widehat{m}_{21}\ \left(\|\widehat{m}_{31}\| + \|\widehat{m}_{32}\|\right)\ \sum_{i=1}^{n} g\left(X_i\right)^{2/3} \\ &= \widehat{m}_{21}\ \left(\|\widehat{m}_{31}\| + \|\widehat{m}_{32}\|\right)\ O_P(n) \overset{(2.24)}{=} o_P(n)\ \widehat{m}_{21}\left(\widehat{m}_{21} + \widehat{m}_{22}\right) \\ &\leq o_P(n)\ \left(2\ \max\left(\widehat{m}_{21}^2, \widehat{m}_{22}^2\right)\right) \leq o_P(n)\ \left(2\ \left(\widehat{m}_{21}^2 + \widehat{m}_{22}^2\right)\right).\end{aligned}$$

Es ergibt sich nun erneut mit (2.26)

$$\sum_{i=1}^{n}\hat{t}_4\ Y_i''(\theta_{01})\ \varepsilon_{in} = o_P(\hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}})$$

und mit gleichen Überlegungen folgt für die restlichen vier Summen

$$\sum_{i=1}^{n} \boldsymbol{b}_i'\hat{\boldsymbol{t}}\,\varepsilon_{in} = o_P(\hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}}).$$

Somit erhalten wir insgesamt, dass

$$\sum_{i=1}^{n} \delta_i^2 = \hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}} + o_P(\hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}}).$$

Eine analoge Argumentation bestätigt, dass der Fehler bei $\sum_{i=1}^{n} \delta_i^3$ ebenfalls höchstens von dieser Ordnung ist. Mit Hilfe dieser Ergebnisse können wir nun (2.15) schreiben als

$$R_{1n} \leq 2\,\boldsymbol{\beta}'\hat{\boldsymbol{t}} - \hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}} + C_n + o_P(\hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}}) + o_P(1). \tag{2.29}$$

Schritt 4

Zuerst begründen wir nun, warum C_n höchstens von der Ordnung $o_P(\hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}})$ ist. Es gilt

$$\boldsymbol{b}_i'\hat{\boldsymbol{t}} = (\widehat{\pi} - \pi_0)\Delta_i + \widehat{\pi}\widehat{m}_{11}Y_i'(\theta_{01}) + \ldots + (1-\widehat{\pi})\frac{\widehat{m}_{22}}{2}Y_i''(\theta_{02}) = o_P(1).$$

und somit ist

$$C_n = \sum_{i=1}^{n} \left(\boldsymbol{b}_i'\hat{\boldsymbol{t}}\right)^3 = \sum_{i=1}^{n} o_P(1)\ \left(\boldsymbol{b}_i'\hat{\boldsymbol{t}}\right)^2 = o_P(1)\,\hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}} = o_P(\hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}}).$$

Daher wird unser Endergebnis aus Schritt 3 zu

$$R_{1n} \leq 2\,\boldsymbol{\beta}'\hat{\boldsymbol{t}} - \hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}}\,[1 + o_p(1)] + o_P(1). \tag{2.30}$$

Als nächstes definieren wir die Funktionen

$$\varphi_1(t) := 2\,\boldsymbol{\beta}'t - t'\boldsymbol{B}t\,[1 + o_p(1)] \text{ und } \varphi_2(t) := 2\,\boldsymbol{\beta}'t - t'\boldsymbol{B}t$$

für $t \in \mathbb{R}^5$. Da $\varphi_1(\hat{\boldsymbol{t}}) = 2\,\boldsymbol{\beta}'\hat{\boldsymbol{t}} - \hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}}\,[1 + o_p(1)]$ und $\hat{\boldsymbol{t}} \in T$ erhalten wir unmittelbar

$$R_{1n} \leq 2\,\boldsymbol{\beta}'\hat{\boldsymbol{t}} - \hat{\boldsymbol{t}}'\boldsymbol{B}\hat{\boldsymbol{t}}\,[1 + o_p(1)] + o_P(1) \leq \sup_{t\in T} \varphi_1(t) + o_P(1). \tag{2.31}$$

Unser nächstes Ziel ist es nun zu folgern, dass

$$R_{1n} \leq \sup_{t\in T} \varphi_2(t) + o_P(1) = \sup_{t\in T} (2\,\boldsymbol{\beta}'t - t'\boldsymbol{B}t) + o_P(1).$$

Dies ist nicht unmittelbar klar, denn B hat die Ordnung n (da $n^{-1}\boldsymbol{B} \overset{f.s.}{\to} \Sigma$) und somit ist $t'\boldsymbol{B}t\ o_P(1) = t'o_P(n)t$ und nicht offensichtlich für beliebige $t \in T$ vernachlässigbar. Wir

müssen daher über die Ordnung der Optimalstelle von φ_1 argumentieren und betrachten diese zuerst unrestringiert. Aufgrund der notwendigen Bedingung erster Ordnung erhalten wir mit $1 + o_P(1) = [1 + o_P(1)]^{-1} =: c$

$$t^*_{ur,\varphi_1} = c\, \boldsymbol{B}^{-1}\boldsymbol{\beta} = c\, O_P(n^{-1})\, O_P(\sqrt{n}) = c\, O_P(n^{-1/2}) = O_P(n^{-1/2}). \tag{2.32}$$

Dabei ist nach dem Zentralen Grenzwertsatz $\boldsymbol{\beta} = \sum_{i=1}^{n} \boldsymbol{b}_i = O_P(n^{1/2})$. Es gilt nun insbesondere für die auf T beschränkte Optimalstelle t^*_{r,φ_1}

$$\begin{aligned}
\varphi_1(t^*_{r,\varphi_1}) \leq \varphi_1(t^*_{ur,\varphi_1}) &= 2\,\boldsymbol{\beta}' t^*_{ur,\varphi_1} - {t^*_{ur,\varphi_1}}'\boldsymbol{B} t^*_{ur,\varphi_1}\,[1 + o_p(1)] \\
&= O_P(n^{1/2})O_P(n^{-1/2}) - O_P(n^{-1/2})O_P(n)O_P(n^{-1/2})(1 + o_P(1)) \\
&= O_P(1).
\end{aligned}$$

φ_1 an den Optima ausgewertet, ist also durch $O_P(1)$ beschränkt. Diese Tatsache ermöglicht uns nun eine Aussage über die Ordnung der Optimalstelle von φ_1 unter der Einschränkung auf T. Wir können direkt folgern, dass die beschränkte Optimalstelle $t^*_{r,\varphi_1} \in T$ ebenfalls höchstens von der Ordnung $O_P(n^{-1/2})$ ist, da sonst $\varphi_1(t^*_{r,\varphi_1}) > O_P(1)$ wäre, was einen Widerspruch darstellen würde. Mit $o_P(1)\, c^2 = o_P(1)$ erhalten wir also

$$\begin{aligned}
\sup_{t\in T} \varphi_1(t) = \varphi_1(t^*_{r,\varphi_1}) &= 2\,\boldsymbol{\beta}' t^*_{r,\varphi_1} - {t^*_{r,\varphi_1}}'\boldsymbol{B} t^*_{r,\varphi_1}\,[1 + o_p(1)] \\
&= 2\,\boldsymbol{\beta}' t^*_{r,\varphi_1} - {t^*_{r,\varphi_1}}'\boldsymbol{B} t^*_{r,\varphi_1} + o_P(1)\, c\, O_P(n^{-1/2})\, O_P(n)\, c\, O_P(n^{-1/2}) \\
&= 2\,\boldsymbol{\beta}' t^*_{r,\varphi_1} - {t^*_{r,\varphi_1}}'\boldsymbol{B} t^*_{r,\varphi_1} + o_P(1)\, c^2\, O_P(1) \\
&= 2\,\boldsymbol{\beta}' t^*_{r,\varphi_1} - {t^*_{r,\varphi_1}}'\boldsymbol{B} t^*_{r,\varphi_1} + o_P(1) = \varphi_2(t^*_{r,\varphi_1}) + o_P(1) \\
&\leq \sup_{t\in T} \varphi_2(t) + o_P(1) = \sup_{t\in T} (2\,\boldsymbol{\beta}' t - t'\boldsymbol{B} t) + o_P(1).
\end{aligned}$$

Aufgrund von (2.31) können wir jetzt folgern, dass

$$R_{1n} \leq \sup_{t\in T} (2\,\boldsymbol{\beta}' t - t'\boldsymbol{B} t) + o_P(1). \tag{2.33}$$

Wir zerlegen nun $t = (t_1, t_2)$, wobei t_1 drei Komponenten besitzt und definieren $\tilde{t}_1 := t_1 + B_{11}^{-1} B_{12} t_2$. Unsere Funktion φ_2 können wir über $\tilde{t}_1$ und t_2 ausdrücken:

$$\varphi_2(t) = 2\,\boldsymbol{\beta}' t - t'\boldsymbol{B} t = 2\boldsymbol{\beta}_1' \tilde{t}_1 - \tilde{t}_1' B_{11} \tilde{t}_1 + 2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2' \widetilde{B}_{22} t_2 := \tilde{\varphi}_2(\tilde{t}_1, t_2).$$

Wir betrachten die transformierte Variable $\tilde{t}_1$ als frei, d.h. wählen als zulässigen Bereich den kompletten $\mathbb{R}^3$. Somit vergrößern wir evtl. das Supremum, falls der tatsächliche

zulässige Bereich kleiner ist. Es gilt also die Abschätzung

$$\begin{aligned}\sup_{t\in T}(2\,\boldsymbol{\beta}'t - t'\boldsymbol{B}t) &\leq \sup_{\tilde{t}_1\in\mathbb{R}^3}\left(2\boldsymbol{\beta}_1'\tilde{t}_1 - \tilde{t}_1'B_{11}\tilde{t}_1\right) + \sup_{t_2\in\mathbb{R}^2_{\geq}}\left(2\widetilde{\boldsymbol{\beta}}_2't_2 - t_2'\widetilde{B}_{22}t_2\right)\\ &= \boldsymbol{\beta}_1'B_{11}^{-1}\boldsymbol{\beta}_1 + \sup_{t_2\in\mathbb{R}^2_{\geq}}\left(2\widetilde{\boldsymbol{\beta}}_2't_2 - t_2'\widetilde{B}_{22}t_2\right).\end{aligned}$$

Somit erhalten wir mit (2.33) schließlich das gewünschte Ergebnis dieses Schrittes:

$$R_{1n} \leq \boldsymbol{\beta}_1'B_{11}^{-1}\boldsymbol{\beta}_1 + \sup_{t_2\in\mathbb{R}^2_{\geq}}\left(2\widetilde{\boldsymbol{\beta}}_2't_2 - t_2'\widetilde{B}_{22}t_2\right) + o_P(1).$$

Schritt 5

Wir möchten nun zeigen, dass die obere Grenze von R_{1n} aus Schritt 4 sogar angenommen wird. Dazu wählen wir zunächst $\tilde{t}^* = (\tilde{t}^{(1)}, \tilde{t}^{(2)}) \in \mathbb{R}^3\times\mathbb{R}^2_{\geq}$, welches gerade diese annimmt, d.h.

$$\tilde{t}^{(1)} := B_{11}^{-1}\boldsymbol{\beta}_1 \text{ und } \tilde{t}^{(2)} \text{ s.d. } 2\widetilde{\boldsymbol{\beta}}_2'\tilde{t}^{(2)} - \tilde{t}^{(2)'}\widetilde{B}_{22}\tilde{t}^{(2)} = \sup_{t_2\in\mathbb{R}^2_{\geq}}\left(2\widetilde{\boldsymbol{\beta}}_2't_2 - t_2'\widetilde{B}_{22}t_2\right).$$

Mit der Argumentation aus Schritt vier erhalten wir auch hier, dass $\tilde{t}^{(1)}$ und $\tilde{t}^{(2)}$ von der Ordnung $O_P(n^{-1/2})$ sind. Ferner gilt nun für das zugehörige rücktransformierte $t^* = \left(\tilde{t}^{(1)} - B_{11}^{-1}B_{12}\tilde{t}^{(2)}, \tilde{t}^{(2)}\right) = O_P(n^{-1/2})$. Besonders die stochastische Konvergenz der ersten Komponente gegen Null wird nun von uns benutzt.

Als nächstes zeigen wir, dass ein $G^* \in \mathcal{M}_k$ existiert mit $t(G^*) = t^*$ (indem wir es konstruieren). Als erstes legen wir die Anzahl von support points kleiner gleich $\theta_{0,mid}$ auf $r_0(k)$ fest und wählen sie somit wie bei $G_0^{(k)}$. Da $t_1(G^*) = G^*(\theta_{0,mid}) - \pi_0$, was mit der ersten Komponenten von t^* übereinstimmen soll, müssen wir bei G^* die ersten $r_0(k)$ Gewichte so wählen, dass sie in Summe mit $\pi_0 + t_1^*$ übereinstimmen. Wir setzen

$$\pi_j(G^*) := \pi_j^{(0)} + \frac{t_1^*}{r_0(k)} \text{ für } j = 1,\ldots,r_0(k)$$

und erhalten das gewünschte Verhalten. Die restlichen Gewichte definieren wir analog über $\pi_j(G^*) := \pi_j^{(0)} - \frac{t_1^*}{k-r_0(k)}$ für $j = r_0(k)+1,\ldots,k$. Weil $t_1^* = o_P(1)$ ist unsere Definition somit für große n zulässig. Außerdem erhalten wir mit dieser Definition

$$\pi_j(G^*) = \pi_j^{(0)} + o_P(1) \text{ für } j = 1,\ldots,k. \tag{2.34}$$

und für $\pi^* := G^*(\theta_{0,mid}) = \sum_{j=1}^{r_0(k)}\pi_j(G^*)$ gilt somit

$$|\pi_0 - \pi^*| = o_P(1). \tag{2.35}$$

Als nächstes betrachten wir die support points von G^*. Wir sehen uns zuerst die ersten $r_0(k)$ Stück an, d.h. diejenigen, die kleiner gleich $\theta_{0,mid}$ sind. Diese spielen für die zweite und vierte Komponente von t^* eine Rolle: Sowohl $t_2(G^*) = t_2^*$ als auch $t_4(G^*) = t_4^*$ muss erfüllt werden, so dass sich die Bedingungen

$$\pi_0 m_{11}(G^*) = \pi_0 \sum_{j=1}^{r_0(k)} (\theta_j(G^*) - \theta_{01})\, \pi_j(G^*) \stackrel{!}{=} t_2^*$$

und

$$\frac{\pi_0}{2} m_{21}(G^*) = \frac{\pi_0}{2} \sum_{j=1}^{r_0(k)} (\theta_j(G^*) - \theta_{01})^2\, \pi_j(G^*) \stackrel{!}{=} t_4^*$$

ergeben. $r_0(k)$ wurde größer gleich zwei gewählt und somit sind genügend support points zum Anpassen an die beiden Bedingungen vorhanden. An dieser Stelle geht außerdem die Forderung der Nichtnegativität für t_4^* (bzw. bei den restlichen support points die von t_5^*) entscheidend ein, denn ohne diese könnten wir das Moment über die support points nicht anpassen, da das zweite Moment immer nicht negativ ist. Für die restlichen support points ergeben sich entsprechende Bedingungen mit t_3^* und t_5^*. Diese können auch wieder erfüllt werden, da wir außerdem $r_0(k) \leq k-2$ gewählt haben und uns somit auch hier wieder mindestens zwei support points zur Verfügung stehen. Die so gewählte mixing distribution $G^* \in \mathcal{M}_k$ erfüllt also die fünf Bedingungen, damit $t(G^*) = t^*$.
R_{1n} und δ_i wurden in Abhängigkeit von $\widehat{G}$ und G_0 definiert. Wir setzen nun beide entsprechend für G^*, d.h.

$$R_{1n}^* := 2\ (l_n(G^*) - l_n(G_0)) = 2 \sum_i^n \log(1+\delta_i^*)\,, \text{ wobei } \delta_i^* := \frac{f(X_i;G^*) - f(X_i;G_0)}{f(X_i;G_0)}. \quad (2.36)$$

Die Durchführung einer Taylorentwicklung der Funktion $h(x) := 2\log(1+x)$ um $x_0 = 0$ liefert uns $h(x) = 2x - (1+\gamma_i)^{-2}x^2$ mit γ_i zwischen 0 und x. Werten wir diesen Ausdruck nun an jeder Stelle δ_i^* aus und setzen die Ergebnisse in (2.36) ein, so erhalten wir

$$R_{1n}^* = 2 \sum_i^n \delta_i^* - \sum_i^n \delta_i^* (1+\gamma_i)^{-2} \text{ mit } |\gamma_i| < |\delta_i^*|\,. \quad (2.37)$$

Wir zeigen nun, dass $\gamma_i = o_P(1)$ für $i = 1, \ldots, n$.
Da $t^* = o_P(1)$, sind die Momente von G^* $o_P(1)$ und somit konvergieren insbesondere die

ersten $r_0(k)$ support points von G^* in Wahrscheinlichkeit gegen θ_{01}. Daher gilt

$$f(X_i; G_1^*) = \frac{1}{\pi^*} \sum_{j=1}^{r_0(k)} \pi_j(G^*)\ f(X_i, \theta_j(G^*)) = \frac{1}{\pi^*} \sum_{j=1}^{r_0(k)} \pi_j(G^*)\ (f(X_i, \theta_{01}) + o_P(1))$$
$$= \frac{f(X_i, \theta_{01}) + o_P(1)}{\pi^*} \sum_{j=1}^{r_0(k)} \pi_j(G^*) = f(X_i, \theta_{01}) + o_P(1)$$

und wir können

$$\left| \frac{f(X_i; G_1^*) - f(X_i, \theta_{01})}{f(X_i; G_0)} \right| = o_P(1)$$

folgern. Diese Rechnung funktioniert für G_2^* und θ_{02} völlig analog. Wir stellen nun δ_i^* dar wie in (2.16) und können schließlich mit (2.35) folgern:

$$|\gamma_i| \leq |\delta_i^*| \leq |(\pi^* - \pi_0)\ \Delta_i| + \pi^* \left| \frac{f(X_i; G_1^*) - f(X_i, \theta_{01})}{f(X_i; G_0)} \right| + (1 - \pi^*) \left| \frac{f(X_i; G_2^*) - f(X_i, \theta_{02})}{f(X_i; G_0)} \right| = o_P(1).$$

Unsere Entwicklung (2.37) für R_{1n}^* wird somit zu

$$R_{1n}^* = 2 \sum_{i=1}^{n} \delta_i^* - \sum_{i=1}^{n} \delta_i^{*2} \left[1 + o_P(1)\right]. \tag{2.38}$$

An dieser Stelle angelangt, können wir nun wie in den Schritten zwei und drei eine Taylorentwicklung für $\sum_{i=1}^n \delta_i^*$ durchführen. Wir erhalten wieder

$$2 \sum_{i=1}^{n} \delta_i^* = 2\boldsymbol{\beta}' t^* + \varepsilon_n^* \text{ mit } \varepsilon_n^* = o_P(t^{*\prime} \boldsymbol{B} t^*) + o_P(1)$$

und

$$\sum_{i=1}^{n} \delta_i^{*2} = t^{*\prime} \boldsymbol{B} t^* + O_P(\varepsilon_n^*) = t^{*\prime} \boldsymbol{B} t^* + o_P(t^{*\prime} \boldsymbol{B} t^*) + o_P(1).$$

Setzen wir diese Ergebnisse in (2.38) ein, so erhalten wir mit $t^* = O_P(n^{-1/2})$ nach Ausmultiplizieren

$$\begin{aligned} R_{1n}^* &= 2\boldsymbol{\beta}' t^* + t^{*\prime} \boldsymbol{B} t^* + o_P(t^{*\prime} \boldsymbol{B} t^*) + o_P(1) \\ &= 2\boldsymbol{\beta}' t^* + t^{*\prime} \boldsymbol{B} t^* + o_P(1) O_P(n^{-1/2}) O_P(n) O_P(n^{-1/2}) + o_P(1) \\ &= 2\boldsymbol{\beta}' t^* + t^{*\prime} \boldsymbol{B} t^* + o_P(1) = \tilde{\varphi}_2(\tilde{t}^*) + o_P(1) \\ &= \boldsymbol{\beta}_1' B_{11}^{-1} \boldsymbol{\beta}_1 + \sup_{t_2 \in \mathbb{R}^2_{\geq}} \left(2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2' \widetilde{B}_{22} t_2 \right) + o_P(1). \end{aligned} \tag{2.39}$$

Als letztes betrachten wir nun die Differenz von R_{1n} und R_{1n}^* und stellen fest, dass

$$R_{1n} - R_{1n}^* = \underbrace{\widetilde{l}_n(\widehat{G}) - \widetilde{l}_n(G^*)}_{\geq 0} - \underbrace{C_k \left[\text{pen}(\widehat{G}) - \text{pen}(G^*) \right]}_{= o_P(1)} \geq o_P(1).$$

Dabei ist der vordere Teil größer gleich Null, da wir $\widehat{G}$ als Maximum der modifizierten Likelihoodfunktion gewählt haben. Die Differenz der Penaltyterme konvergiert stochastisch gegen Null, weil nach (2.34) die Gewichte von G^* und nach Satz 2.5 die von $\widehat{G}$ jeweils gegen die von $G_0^{(k)}$ konvergieren. Wir können also $R_{1n} \geq R_{1n}^* + o_P(1)$ und somit nach (2.39)

$$R_{1n} \geq \boldsymbol{\beta}_1' B_{11}^{-1} \boldsymbol{\beta}_1 + \sup_{t_2 \in \mathbb{R}^2_{\geq}} \left(2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2' \widetilde{B}_{22} t_2 \right) + o_P(1)$$

festhalten. Kombinieren wir dies mit Schritt vier, so erhalten wir die zu beweisende Aussage

$$R_{1n} = \boldsymbol{\beta}_1' B_{11}^{-1} \boldsymbol{\beta}_1 + \sup_{t_2 \in \mathbb{R}^2_{\geq}} \left(2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2' \widetilde{B}_{22} t_2 \right) + o_P(1).$$

□

Wir wollen nun in einem weiteren Lemma das Grenzverhalten von

$$R_{0n} = 2 \cdot \left(l_n(\widehat{G}_0) - l_n(G_0) \right),$$

also der Teststatistik eines LRT mit der Einpunkthypothese $G = G_0$ gegen $G \in \mathcal{M}_2$, betrachten.

Lemma 2.8. *Unter den getroffenen Annahmen in Abschnitt 2.1.1 gilt:*

$$R_{0n} = \boldsymbol{\beta}_1' B_{11}^{-1} \boldsymbol{\beta}_1 + o_P(1),$$

wobei $\boldsymbol{\beta}_1$ wie in Lemma 2.7 definiert ist.

Beweis. Der entscheidende Unterschied zum vorherigen Lemma ist hier, dass

$$m_{2l}(G^*) = m_{1l}(G^*)^2 = o_P(1) m_{1l}(G^*) \text{ für } l = 1, 2, \tag{2.40}$$

d.h. die zweiten Momente entsprechen jeweils den quadrierten ersten Momenten (welche $o_P(1)$ sind). Diese Gleichheit liegt daran, dass die beiden aufgeteilten mixing distributions jeweils nur einen support point besitzen.
Auch hier gilt das Ergebnis des vorherigen Lemmas nur für R_{0n}, d.h.

$$R_{0n} = \boldsymbol{\beta}_1' B_{11}^{-1} \boldsymbol{\beta}_1 + \sup_{t_2 \in \mathbb{R}^2_{\geq}} \left(2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2' \widetilde{B}_{22} t_2 \right) + o_P(1).$$

Die Optimalstelle des Supremums t^* (also alle fünf Komponenten) hat nach vorherigem Beweis die Ordnung $n^{-1/2}$, d.h. insbesondere ist $m_{1l}(G^*) = O_P(n^{-1/2})$. Beide Komponenten der Optimalstelle des hinteren Supremums haben mit (2.40) die Form:

$$t_2^* = (\pi^*/2\, m_{21}(G^*), (1-\pi^*)/2\, m_{22}(G^*)) = (o_P(m_{11}(G^*)), o_P(m_{12}(G^*))) = o_P(n^{-1/2}).$$

Setzen wir nun dies in das Supremum ein, so erhalten wir

$$\begin{aligned}\sup_{t_2 \in \mathbb{R}^2_{\geq}} \left(2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2' \widetilde{B}_{22} t_2\right) &= 2\widetilde{\boldsymbol{\beta}}_2' t_2^* - t_2^{*\prime} \widetilde{B}_{22} t_2^* \\ &= O_P(n^{1/2}) o_P(n^{-1/2}) - o_P(n^{-1/2}) O_P(n) o_P(n^{-1/2}) \\ &= o_P(1).\end{aligned}$$

Diese Beobachtung beendet bereits den Beweis. □

Bevor wir schließlich zum Satz über die asymptotische Verteilung von R_n kommen können, benötigen wir ein letztes Lemma. Hier betrachten wir unter anderem die folgenden Kovarianzmatrizen:

$$\Sigma := \mathrm{COV}\left[\Delta_1, Y_1'(\theta_{01}), Y_1'(\theta_{02}), Y_1''(\theta_{01}), Y_1''(\theta_{02})\right] \text{ mit } \Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right),\ \Sigma_{11} \in \mathbb{R}^{3\times 3} \tag{2.41}$$

sowie

$$\widetilde{\Sigma}_{22} := \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \tag{2.42}$$

Lemma 2.9. *Unter den getroffenen Annahmen in Abschnitt 2.1.1 gilt:*

1. $n^{-1}\boldsymbol{B} \xrightarrow{f.s.} \Sigma$, $B_{11}^{-1}B_{12} \xrightarrow{f.s.} \Sigma_{11}^{-1}\Sigma_{12}$ *und* $n^{-1}\widetilde{B}_{22} \xrightarrow{f.s.} \widetilde{\Sigma}_{22}$
2. $\frac{\widetilde{\boldsymbol{\beta}}_2}{\sqrt{n}} \xrightarrow{V.} \mathcal{N}(0, \widetilde{\Sigma}_{22})$

Beweis. Die $\boldsymbol{b}_i \in \mathbb{R}^5$ sind identisch unabhängig verteilt mit Erwartungswert Null (da X_i iid). Nach dem starken Gesetz der großen Zahlen gilt

$$n^{-1}\boldsymbol{B} = n^{-1}\sum_{i=1}^{n} \boldsymbol{b}_i \boldsymbol{b}_i' \xrightarrow{f.s.} E(\boldsymbol{b}_1\boldsymbol{b}_1') = \mathrm{COV}(\boldsymbol{b}_1) = \Sigma.$$

Somit gilt natürlich auch für die einzelnen Blockmatrizen $n^{-1}B_{ml} \xrightarrow{f.s.} \Sigma_{ml}$, $m, l = 1, 2$. Wir erhalten also

$$B_{11}^{-1}B_{12} = \left(\frac{B_{11}}{n}\right)^{-1} \frac{B_{12}}{n} \xrightarrow{f.s.} \Sigma_{11}^{-1}\Sigma_{12}$$

und entsprechend

$$n^{-1}\widetilde{B}_{22} = n^{-1}B_{22} - n^{-1}B_{21}B_{11}^{-1}B_{12} \xrightarrow{f.s.} \widetilde{\Sigma}_{22}.$$

Für den Nachweis der zweiten Aussage definieren wir zunächst für $i = 1, \ldots, n$

$$\boldsymbol{\alpha}_i' := \left(Y_i''(\theta_{01}), Y_i''(\theta_{02})\right) - \left(\Delta_i, Y_i'(\theta_{01}), Y_i'(\theta_{02})\right)\Sigma_{11}^{-1}\Sigma_{12}.$$

Auch die $\boldsymbol{\alpha}_i$ sind identisch unabhängig verteilt mit Erwartungswert Null und Kovarianzmatrix

$$\mathrm{COV}(\boldsymbol{\alpha}_1) = \widetilde{\Sigma}_{22}.$$

Nach dem Zentralen Grenzwertsatz gilt also

$$n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\alpha}_i \xrightarrow{V.} \mathcal{N}(0, \widetilde{\Sigma}_{22}).$$

Weil

$$\widetilde{\boldsymbol{\beta}}_2' = \boldsymbol{\beta}_2' - \boldsymbol{\beta}_1' B_{11}^{-1} B_{12} = \sum_{i=1}^{n} (Y_i''(\theta_{01}), Y_i''(\theta_{02})) - \sum_{i=1}^{n} (\Delta_i, Y_i'(\theta_{01}), Y_i'(\theta_{02}))\, B_{11}^{-1} B_{12}$$

und nach dem ersten Teil des Lemmas $B_{11}^{-1} B_{12} \xrightarrow{f.s.} \Sigma_{11}^{-1} \Sigma_{12}$ folgt

$$\left| n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - n^{-1/2} \widetilde{\boldsymbol{\beta}}_2' \right| \xrightarrow{f.s.} 0.$$

Nach dem Satz von Slutzky (vgl. [Kle06, Satz 13.18]) folgt nun die zweite Aussage des Lemmas. □

Wir haben nun mit den bisherigen Lemmata die Voraussetzung geschaffen, um den Satz über die asymptotische Verteilung der Teststatistik R_n zu formulieren.

Satz 2.10 (Asymptotische Verteilung der Teststatistik)**.** *Falls die Annahmen aus Abschnitt 2.1.1 erfüllt sind, $k \geq k^*$ und $X_i \overset{iid}{\sim} f(x; G_0)$ ist die asymptotische Verteilung der Teststatistik die χ^2-Mischung*

$$\left(\frac{1}{2} - \frac{\alpha}{2\pi}\right) \chi_0^2 + \frac{1}{2} \chi_1^2 + \frac{\alpha}{2\pi} \chi_2^2.$$

Dabei ist χ_0^2 die Einpunktverteilung bei Null und χ_1^2 bzw. χ_2^2 sind χ^2-Verteilungen mit einem bzw. zwei Freiheitsgraden. ρ ist der Korrelationskoeffizient der Kovarianzmatrix $\widetilde{\Sigma}_{22}$ und $\alpha := \cos^{-1}(\rho)$.

Beweis. Nach den Lemmata 2.7 und 2.8 gilt

$$\left| R_n - \sup_{t_2 \in \mathbb{R}^2_{\geq}} \left(2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2' \widetilde{B}_{22} t_2 \right) \right| = o_P(1). \tag{2.43}$$

Daher bestimmen wir nun die Verteilung von diesem Supremum für große n. Wir nehmen ohne Beschränkung der Allgemeinheit an, dass

$$\widetilde{\Sigma}_{22} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

und erinnern daran, dass dies (fast sicher für n groß) die Kovarianzmatrix von $n^{-1/2}\widetilde{\boldsymbol{\beta}}_2$ ist. Substituieren wir $\tau := n^{-1/2}t_2$, so ist der zulässige Bereich für τ ebenfalls der $\mathbb{R}^2_{\geq}$. Wir erhalten somit:

$$\begin{aligned}\sup_{t_2\in\mathbb{R}^2_{\geq}}\left(2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2'\widetilde{B}_{22}t_2\right) &= \sup_{n^{-1/2}t_2\in\mathbb{R}^2_{\geq}}\left(2\widetilde{\boldsymbol{\beta}}_2'\frac{t_2}{\sqrt{n}} - \frac{t_2'}{\sqrt{n}}\widetilde{B}_{22}\frac{t_2}{\sqrt{n}}\right)\\ &= \sup_{t_2\in\mathbb{R}^2_{\geq}}\left(2\frac{\widetilde{\boldsymbol{\beta}}_2'}{\sqrt{n}}t_2 - t_2'\frac{\widetilde{B}_{22}}{n}t_2\right) \xrightarrow{(V.)} \sup_{\xi_1\geq 0,\xi_2\geq 0}\left(2\boldsymbol{Z}'\xi - \xi'\widetilde{\Sigma}_{22}\xi\right),\end{aligned}$$

wobei $\xi' = (\xi_1, \xi_2)$ und $\boldsymbol{Z} = (Z_1, Z_2)'$ ein bivariat normalverteiler Zufallsvektor ist mit Erwartungswert Null und Kovarianzmatrix $\widetilde{\Sigma}_{22}$. Wir setzen zur Abkürzung

$$A := \begin{pmatrix} 1 & 0 \\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{pmatrix}$$

und definieren durch

$$\boldsymbol{W} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} := A \cdot \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

einen bivariat normalverteilten Zufallsvektor mit Kovarianzmatrix

$$\mathrm{COV}(\boldsymbol{W}) = \mathrm{COV}(A \cdot \boldsymbol{Z}) = A \cdot \widetilde{\Sigma}_{22} \cdot A'.$$

Da

$$A \cdot \widetilde{\Sigma}_{22} \cdot A' = \begin{pmatrix} 1 & \rho \\ 0 & \frac{1-\rho^2}{\sqrt{1-\rho^2}} \end{pmatrix} \cdot A' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

ist $\boldsymbol{W}$ bivariat normalverteilt mit Kovarianzmatrix I_2. W_1 und W_2 sind somit unabhängige standardnormalverteilte Zufallsvariablen.
Aufgrund der Definitionen von W_1 und W_2 gilt, dass $Z_2 = \sqrt{1-\rho^2}W_2 + \rho W_1$. Setzen wir diese Umformung in das betrachtete Supremum ein, so erhalten wir:

$$\sup_{\xi_1\geq 0,\xi_2\geq 0}\left(2\boldsymbol{Z}'\xi - \xi'\widetilde{\Sigma}_{22}\xi\right) = \sup_{\xi_1\geq 0,\xi_2\geq 0}\left(2(W_1, \sqrt{1-\rho^2}W_2 + \rho W_1)\xi - \xi'\widetilde{\Sigma}_{22}\xi \pm W_1^2 \pm W_2^2\right)$$

$$= W_1^2 + W_2^2 - \inf_{\xi_1\geq 0,\xi_2\geq 0}\left(W_1^2 - 2W_1(\xi_1 + \rho\xi_2) + W_2^2 - 2W_2\sqrt{1-\rho^2}\xi_2 + \xi'\widetilde{\Sigma}_{22}\xi\right)$$

Ferner folgt die Gleichung

$$\begin{aligned}&W_1^2 - 2W_1(\xi_1 + \rho\xi_2) + W_2^2 - 2W_2\sqrt{1-\rho^2}\xi_2 + \xi'\widetilde{\Sigma}_{22}\xi\\ &= [W_1 - (\xi_1 + \rho\xi_2)]^2 + \left[W_2 - (1-\rho^2)^{1/2}\xi_2\right]^2\end{aligned}$$

unmittelbar durch Ausmultiplizieren. Dabei gilt $[\xi_1 + \rho\xi_2]^2 + \left[(1-\rho^2)^{1/2}\xi_2\right]^2 = \xi'\widetilde{\Sigma}_{22}\xi$. Wir erhalten somit

$$\sup_{\xi_1\geq 0,\xi_2\geq 0}\left(2\boldsymbol{Z}'\xi - \xi'\widetilde{\Sigma}_{22}\xi\right) = W_1^2 + W_2^2 - \inf_{\xi_1\geq 0,\xi_2\geq 0}\left([W_1 - \underbrace{(\xi_1+\rho\xi_2)}_{=:\eta_1}]^2 + [W_2 - \underbrace{(1-\rho^2)^{1/2}\xi_2}_{=:\eta_2}]^2\right)$$
$$= W_1^2 + W_2^2 - \inf_{(\eta_1,\eta_2)\in S}\left([W_1-\eta_1]^2 + [W_2-\eta_2]^2\right).$$

Natürlich müssen wir bei der Transformation von ξ_l zu η_l den zulässigen Bereich beachten. Bei der zweiten Komponente ändert sich nichts, denn $\eta_2 \geq 0 \Leftrightarrow \xi_2 \geq 0$. Unsere Definition von η_2 liefert $\xi_2 = \eta_2(1-\rho^2)^{-1/2}$. Setzen wir dies in die Definition von η_1 ein, erhalten wir

$$\eta_1 = \xi_1 + \rho\xi_2 = \xi_1 + \rho\eta_2(1-\rho^2)^{-1/2}$$

und lösen nach ξ_1 auf, so folgt $\xi_1 = \eta_1 - \rho\eta_2(1-\rho^2)^{-1/2}$. Somit ist also die Bedingung $\xi_1 \geq 0$ äquivalent zu $\eta_1 \geq \rho\eta_2(1-\rho^2)^{-1/2}$. Daher ist der zulässige Bereich S von (η_1,η_2)

$$S = \left\{(\eta_1,\eta_2) \;:\; \eta_2 \geq 0, \eta_1 \geq \rho\eta_2(1-\rho^2)^{-1/2}\right\}.$$

Den bisherigen Beweis zusammenfassend gilt also $(W_1,W_2) \sim \mathcal{N}(0,I_2)$ und

$$\sup_{t_2\in\mathbb{R}^2_{\geq}}\left(2\widetilde{\boldsymbol{\beta}}_2't_2 - t_2'\widetilde{B}_{22}t_2\right) \xrightarrow{(V.)} W_1^2 + W_2^2 - \inf_{(\eta_1,\eta_2)\in S}\left([W_1-\eta_1]^2 + [W_2-\eta_2]^2\right). \tag{2.44}$$

Die Ergebnisse (2.43) und (2.44) liefern gemeinsam nach dem Satz von Slutzky:

$$R_n \xrightarrow{(V.)} W_1^2 + W_2^2 - \inf_{(\eta_1,\eta_2)\in S}\left([W_1-\eta_1]^2 + [W_2-\eta_2]^2\right). \tag{2.45}$$

Weil $\rho \in [-1,1]$, ist $\alpha = \cos^{-1}(\rho) \in [0,\pi]$ als Winkel im Bogenmaß. Im Gradmaß liegt der Winkel α somit zwischen Null und 180 Grad. Die Menge S ist anschaulich ein Ausschnitt des ersten Quadranten mit Winkel α des $\eta_1 - \eta_2$ Koordinatensystems.
Die Optimalstelle des Infimums in Gleichung (2.45) ist nichts anderes als derjenige Punkt aus S, welcher den minimalen Abstand zu (W_1,W_2) hat. Falls $(W_1,W_2) \notin S$, so ist die Optimalstelle die Projektion von (W_1,W_2) auf S. Abbildung 2.1 veranschaulicht den Kegel S, den dualen Kegel S^* und die jeweiligen Projektionswege. Diese Veranschaulichung hilft uns die Verteilung in einer Fallunterscheidung abzuleiten. Wir betrachten die drei Fälle (W_1,W_2) liegt im Kegel S, im dazu dualen Kegel S^* oder im Rest des $\mathbb{R}^2$. Für jeden dieser Fälle leiten wir eine Verteilung ab. Die Wahrscheinlichkeit für die Fälle wird sich jeweils aus dem Winkel der Kegel ergeben. Wir können so zum Schluß die asymptotische Verteilung von R_n als Mischung dieser drei hergeleiteten Verteilungen mit den entsprechenden Wahrscheinlichkeiten angeben.

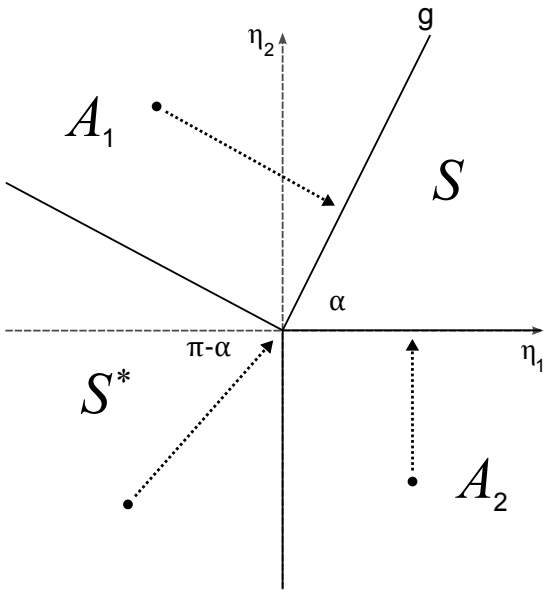


Abbildung 2.1: Veranschaulichung Kegel und Projektionswege. Quelle: [CCK04]

1. Falls $(W_1, W_2) \in S$, so ist die Optimalstelle $(\eta_1^*, \eta_2^*) = (W_1, W_2)$ und das Infimum nimmt den Wert Null an. Daher ist hier nach (2.45)

$$(R_n \,|\, (W_1, W_2) \in S) \xrightarrow{(V.)} \left(W_1^2 + W_2^2 \,|\, (W_1, W_2) \in S\right).$$

Wir werden nun zeigen, dass die Bedingung $(W_1, W_2) \in S$ nichts daran ändert, dass $W_1^2 + W_2^2$ χ_2^2-verteilt ist. Generell gilt für die bivariat standardnormalverteilte Zufallsvariable

$$(W_1, W_2) \overset{(V.)}{=} T \cdot (\cos(\nu), \sin(\nu)), \text{ wobei } \nu \sim U([0, 2\pi]), \; T^2 \sim \chi_2^2$$

und ν und T von einander unabhängig sind. Dabei beschreibt ν die Richtung und T die Länge von (W_1, W_2). Richtung und Länge sind nach obiger Formulierung also unabhängig voneinander. Die Einschränkung $(W_1, W_2) \in S$ betrifft nur die Richtung und somit ν. Wir erhalten daher

$$\begin{aligned}(R_n \,|\, (W_1, W_2) \in S) &\xrightarrow{(V.)} \left(W_1^2 + W_2^2 \,|\, (W_1, W_2) \in S\right) \\ &\overset{(V.)}{=} \left(T^2(\cos^2(\nu) + \sin^2(\nu)) \,|\, \nu \in [0, \alpha]\right) \overset{(V.)}{=} T^2 \sim \chi_2^2.\end{aligned}$$

Da (W_1, W_2) $\mathcal{N}(0, I_2)$-verteilt ist, nimmt es Werte in S mit Wahrscheinlichkeit $\frac{\alpha}{2\pi}$ an.

2. Falls (W_1, W_2) im dualen Kegel von S liegt, so ist der Punkt aus S mit dem minimalen Abstand zu (W_1, W_2) der Nullpunkt, d.h. $(\eta_1^*, \eta_2^*) = (0, 0)$. Das Infimum nimmt also den Wert $W_1^2 + W_2^2$ an und somit ist nach (2.45)

$$(R_n \,|\, (W_1, W_2) \in S^*) \xrightarrow{(V.)} 0.$$

Da S durch den Winkel α beschrieben wird, hat der duale Kegel S^* den Winkel $\pi - \alpha$. Somit tritt dieser Fall mit der Wahrscheinlichkeit $\frac{\pi-\alpha}{2\pi} = \frac{1}{2} - \frac{\alpha}{2\pi}$ auf.

3. Falls (W_1, W_2) in der restlichen Region liegt, befindet sich (W_1, W_2) entweder in A_1 oder in A_2 (wie in Abbildung 2.1 angegeben). Wir behandeln nun zuerst den Fall $(W_1, W_2) \in A_2$. Der Punkt (η_1^*, η_2^*) ist in diesem Fall einfach die orthogonale Projektion auf die η_1-Achse und somit der Punkt $(W_1, 0)$. Hier nimmt das Infimum den Wert $(W_1 - W_1)^2 + (W_2 - 0)^2 = W_2^2$ an und folglich ist wiederum nach (2.45)

$$(R_n \,|\, (W_1, W_2) \in A_2) \xrightarrow{(V.)} \left(W_1^2 \,|\, (W_1, W_2) \in A_2\right).$$

Die Einschränkung $(W_1, W_2) \in A_2$ bedeutet für die erste Komponente W_1, dass sie nur positive Werte annehmen kann. Das ursprünglich normalverteilte W_1 ist gegeben der Einschränkung nicht mehr normalverteilt. Wir zeigen nun, dass dennoch $(W_1^2 \,|\, (W_1, W_2) \in A_2)$ χ^2-verteilt ist mit einem Freiheitsgrad. Wir setzen $V := |W_1|$ sowie $\widetilde{W_1} := (W_1 \,|\, (W_1, W_2) \in A_2) = (W_1 \,|\, W_1 \geq 0)$ und beobachten für $y \in \mathbb{R}$ bel.

$$\begin{aligned} F_{\widetilde{W_1}}(y) &= P(W_1 \leq y \,|\, W_1 \geq 0) = P(W_1 \geq 0)^{-1}\, P(0 \leq W_1 \leq y) \\ &= 2\, P(0 \leq W_1 \leq y) = P(-y \leq W_1 \leq y) = P(V \leq y) = F_V(y). \end{aligned} \tag{2.46}$$

Bei Gleichung (2.46) geht die Standardnormalverteilung von W_1 und die daraus resultierende Symmetrie um Null ein. V und $\widetilde{W_1}$ sind somit gleich in Verteilung und daher auch ihre Quadrate, d.h. $V^2 \overset{(V.)}{=} \widetilde{W_1}^2$. Da V als der Betrag von W_1 definiert wurde, stimmen auch V^2 und W_1^2 überein. Diese beiden Aussagen zusammen ergeben, dass $\widetilde{W_1}^2$ und W_1^2 in Verteilung übereinstimmen. Daher erhalten wir schließlich

$$(R_n \,|\, (W_1, W_2) \in A_2) \xrightarrow{(V.)} \left(W_1^2 \,|\, (W_1, W_2) \in A_2\right) \overset{(V.)}{=} \widetilde{W_1}^2 \overset{(V.)}{=} W_1^2 \sim \chi_1^2.$$

Im Fall $(W_1, W_2) \in A_1$ wird der Punkt nicht auf die η_2-Achse projiziert, sondern auf die Gerade

$$g(\eta_1) := m \cdot \eta_1 \text{ mit } m := \frac{(1 - \rho^2)^{1/2}}{\rho}.$$

Diese Gerade entsteht aus der Charakterisierung von S $\eta_1 \geq \rho\eta_2/(1-\rho^2)^{1/2}$, indem man hier Gleichheit fordert und nach η_2 auflöst. Wir können jedoch diesen Fall analog zu $(W_1, W_2) \in A_2$ lösen. Sei h die Gerade, die senkrecht auf g steht und durch den Nullpunkt geht. Wir betrachten nun diese als die neue x-Achse und die Gerade g als die neue y-Achse. Somit rotieren wir also das Koordinatensystem um den Winkel $\frac{\pi}{4} - \alpha$. Mit S_r bezeichnen wir die Menge S im rotierten Koordinatensystem bzw. mit (V_1, V_2) den Punkt (W_1, W_2). Wir erhalten also mit (2.45)

$$\begin{aligned}
(R_n \mid (W_1, W_2) \in A_1) &\xrightarrow{(V.)} \left(W_1^2 + W_1^2 - \inf_{(\eta_1,\eta_2)\in S} \left([W_1 - \eta_1]^2 + [W_2 - \eta_2]^2 \right) \mid (W_1, W_2) \in A_1 \right) \\
&= \left(V_1^2 + V_1^2 - \inf_{(\eta_1,\eta_2)\in S_r} \left([V_1 - \eta_1]^2 + [V_2 - \eta_2]^2 \right) \mid V_1 \leq 0, V_2 \geq 0 \right) \\
&= \left(V_1^2 + V_1^2 - \left([V_1 - 0]^2 + [V_2 - V_2]^2 \right) \mid V_1 \leq 0, V_2 \geq 0 \right) \\
&= \left(V_2^2 \mid V_2 \geq 0 \right).
\end{aligned}$$

Somit befinden wir uns wieder in der Situation wie im ersten Fall und erhalten analog

$$(R_n \mid (W_1, W_2) \in A_1) \xrightarrow{(V.)} \chi_1^2.$$

Die asymptotische Verteilung von R_n ergibt sich schließlich als Mischung der einzelnen Verteilungen der drei betrachteten Fälle:

a) Mit Wahrscheinlichkeit $\frac{\alpha}{2\pi}$ liegt (W_1, W_2) in S, dann ist $R_n \sim \chi_2^2$.

b) Mit Wahrscheinlichkeit $\frac{1}{2} - \frac{\alpha}{2\pi}$ liegt (W_1, W_2) in S^*, dann ist $R_n = 0$.

c) Mit Wahrscheinlichkeit $\frac{1}{2}$ liegt (W_1, W_2) in $A_1 \cup A_2$, dann ist $R_n \sim \chi_1^2$.

□

2.1.5 Durchführung des Tests

Im letzten Abschnitt dieses Kapitels möchten wir kurz die Vorgehensweise beim Testen darstellen. Wir nehmen grundsätzlich an, dass wir einen Datensatz vorliegen haben, welcher aus einem unabhängig identisch verteilten Mischungsmodel mit bekannter Verteilungsfamilie der Komponentendichte (z.B. normalverteilt mit $\mu = 0$ und beliebigem σ) stammt. Die Anzahl der Komponenten und Ausprägungen der Parameter sind unbekannt. Es soll getestet werden, ob wir einen Datensatz aus einem Mischungsmodell mit höchstens zwei Komponenten vorliegen haben.

Überblick der Vorgehensweise

1. Wir bestimmen den modifizierten Maximum Likelihoodschätzer $\widehat{G}_0 \in \mathcal{M}_2$ mit den Parametern $(\widehat{\pi}_0, \widehat{\theta}_1, \widehat{\theta}_2)$. Somit schätzen wir die wahren Komponentenparameter $(\theta_{01}, \theta_{02})$ über die support points $(\widehat{\theta}_1, \widehat{\theta}_2)$ und das wahre Gewicht π_0 durch $\widehat{\pi}_0$. Dabei verwenden wir an dieser Stelle keinen Penalty (d.h. $C_2 = 0$), denn unter der Nullhypothese werden wir auch ohne Penalty das Gewicht und die support points konsistent schätzen. Falls π_0 deutlich von 0.5 verschieden ist (z.B. 0.75), würde eine Penalisierung an dieser Stelle bewirken, dass wir π_0 schlechter schätzen würden als eigentlich möglich, da Gewichte nahe 0.5 von der Optimierung bevorzugt behandelt werden. Dies würde zu einer Reduktion des Log-Likelihoodwertes unter der Hypothese führen und unser Test würde (bei finitem n) leicht anti konservativ werden.

2. Wie im Abschnitt über die Teststatistik verdeutlicht wurde, hängt der wahre Wert für k^* vom unbekannten wahren Wert für π_0 ab. Da wir π_0 über $\widehat{\pi}_0$ konsistent schätzen können, schätzen wir nun k^* in Abhängigkeit von $\widehat{\pi}_0$. Für das in (2.10) definierte k^* geben Chen et. al (vgl. [CCK04]) die Berechnungsformel

$$k^* = \max\left(\left\lfloor \frac{1.5}{\pi_0} \right\rfloor, \left\lfloor \frac{1.5}{1-\pi_0} \right\rfloor, 4\right)$$

   an. Wir schätzen also $\widehat{k}^* := \max\left(\lfloor 1.5/\widehat{\pi}_0 \rfloor, \lfloor 1.5/(1-\widehat{\pi}_0) \rfloor, 4\right)$.

3. Nun können wir für $k \geq \widehat{k}^*$ den modifizierten Maximum Likelihoodschätzer $\widehat{G} \in \mathcal{M}_k$ schätzen und anschließend die Teststatistik $R_n = l_n(\widehat{G}) - l_n(\widehat{G}_0)$ berechnen. Diese gleichen wir dann mit der Verteilung aus Satz 2.10 ab und verwerfen die Nullhypothese, falls R_n einen größeren Wert annimmt als das $(1-a)$ Quantil (wobei a unser gewähltes Signifikanzlevel ist). Hierbei ist zu beachten, dass die

asymptotische Verteilung von R_n abhängig ist vom Korrelationskoeffizienten ρ, welcher sich aus der Kovarianzmatrix $\widetilde{\Sigma}_{22}$ ergibt. Diese Matrix und somit ρ kann nach Lemma 2.9 konsistent geschätzt werden.

Schätzung von ρ allgemein

Bevor wir für ausgewählte parametrische Familien von Verteilungen der X_i einen Weg zur Schätzung von ρ angeben, betrachten wir die zur Schätzung benötigten Größen allgemein. Es gilt $\boldsymbol{b}_i = b(X_i, \pi_0, \theta_{01}, \theta_{02}) = (\Delta_i, Y_i'(\theta_{01}), Y_i'(\theta_{02}), Y_i''(\theta_{01}), Y_i''(\theta_{02}))$, wobei

$$b(x, \pi, \theta_1, \theta_2) := c(x, \pi, \theta_1, \theta_2) \cdot \begin{pmatrix} f(x,\theta_1) - f(x,\theta_2) \\ \frac{\partial f(x,\theta)}{\partial \theta} \mid_{\theta=\theta_1} \\ \frac{\partial f(x,\theta)}{\partial \theta} \mid_{\theta=\theta_2} \\ \frac{\partial^2 f(x,\theta)}{\partial \theta^2} \mid_{\theta=\theta_1} \\ \frac{\partial^2 f(x,\theta)}{\partial \theta^2} \mid_{\theta=\theta_2} \end{pmatrix}$$

und

$$c(x, \pi, \theta_1, \theta_2) := \frac{1}{\pi\ f(x,\theta_1) + (1-\pi) f(x,\theta_2)}.$$

Wir können also $\boldsymbol{b}_i$ nicht exakt berechnen, aber es schätzen durch

$$\widehat{\boldsymbol{b}}_i := b(X_i, \widehat{\pi}_0, \widehat{\theta}_1, \widehat{\theta}_2). \tag{2.47}$$

Schließlich schätzen wir $\Sigma = \mathrm{COV}[\boldsymbol{b}_1]$ über

$$\widehat{\Sigma} := n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{b}}_i \widehat{\boldsymbol{b}}_i', \tag{2.48}$$

weil $n^{-1}\boldsymbol{B}$ nach Lemma 2.9 fast sicher gegen Σ konvergiert. Zerlegen wir nun $\widehat{\Sigma}$ analog zu (2.41) (d.h. $\widehat{\Sigma}_{11}$ ist die obere linke 3×3 Teilmatrix von $\widehat{\Sigma}$), können wir schließlich $\widetilde{\Sigma}_{22}$ über

$$A := \widehat{\Sigma}_{22} - \widehat{\Sigma}_{21} \widehat{\Sigma}_{11}^{-1} \widehat{\Sigma}_{12} \tag{2.49}$$

schätzen. Die Matrix $A = (\alpha_{lj})_{l,j=1,2}$ ist asymptotisch eine 2×2 Kovarianzmatrix. Wir können nun den zu A gehörigen Korrelationskoeffizienten und somit unseren Schätzer für ρ berechnen über

$$\widehat{\rho} := \frac{\alpha_{12}}{\sqrt{\alpha_{11}\ \alpha_{22}}}. \tag{2.50}$$

Wir geben nun für die Poissonverteilung, die Binomialverteilung und die Normalverteilung jeweils eine Berechnungsformel für die Funktion $b(x, \pi, \theta_1, \theta_2)$ an. Weiterhin verweisen wir darauf, dass sich anschließend aus den Gleichungen (2.47) bis (2.50) die restlichen benötigten Größen zur Schätzung von ρ ergeben.

Schätzung von ρ bei Poissonverteilung

Für $x \in \mathbb{N}_0$ und $\theta > 0$ beschreibt

$$f_{\text{poi}}(x, \theta) := \frac{\theta^x}{x!} e^{-\theta}$$

die Dichte einer poissonverteilten Zufallsvariablen bzgl. dem Zählmaß. In diesem Fall gilt

$$b(x, \pi, \theta_1, \theta_2) = c_{\text{poi}}(x, \pi, \theta_1, \theta_2) \cdot \begin{pmatrix} f_{\text{poi}}(x, \theta_1) + f_{\text{poi}}(x, \theta_2) \\ f_{\text{poi}}(x, \theta_1) \cdot \left(\frac{x}{\theta_1} - 1 \right) \\ f_{\text{poi}}(x, \theta_2) \cdot \left(\frac{x}{\theta_2} - 1 \right) \\ f_{\text{poi}}(x, \theta_1) \cdot \left[\left(\frac{x}{\theta_1} - 1 \right)^2 - \frac{x}{\theta_1^2} \right] \\ f_{\text{poi}}(x, \theta_2) \cdot \left[\left(\frac{x}{\theta_2} - 1 \right)^2 - \frac{x}{\theta_2^2} \right] \end{pmatrix},$$

wobei $c_{\text{poi}}(x, \pi, \theta_1, \theta_2) := \left[\pi f_{\text{poi}}(x, \theta_1) + (1 - \pi) f_{\text{poi}}(x, \theta_2) \right]^{-1}$.

Schätzung von ρ bei Normalverteilung mit festem Erwartungswert

Für $x \in \mathbb{R}$ und $\theta > 0$ beschreibt

$$f_{\text{nor}}(x, \theta) := \frac{1}{\theta \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x - \mu_0}{\theta} \right)^2 \right)$$

die Dichte einer normalverteilten Zufallsvariable bzgl. dem Lebesgue-Maß mit festem Erwartungswert μ_0 und beliebiger Standardabweichung θ. In diesem Fall gilt

$$b(x, \pi, \theta_1, \theta_2) = c_{\text{nor}}(x, \pi, \theta_1, \theta_2) \cdot \begin{pmatrix} f_{\text{nor}}(x, \theta_1) + f_{\text{nor}}(x, \theta_2) \\ f_{\text{nor}}(x, \theta_1) \cdot \left(\frac{(x-\mu_0)^2}{\theta_1^3} - \frac{1}{\theta_1} \right) \\ f_{\text{nor}}(x, \theta_2) \cdot \left(\frac{(x-\mu_0)^2}{\theta_2^3} - \frac{1}{\theta_2} \right) \\ f_{\text{nor}}(x, \theta_1) \cdot \left(\frac{(x-\mu_0)^4}{\theta_1^6} - 5 \, \frac{(x-\mu_0)^2}{\theta_1^4} + 2 \, \frac{1}{\theta_1^2} \right) \\ f_{\text{nor}}(x, \theta_2) \cdot \left(\frac{(x-\mu_0)^4}{\theta_2^6} - 5 \frac{(x-\mu_0)^2}{\theta_2^4} + 2 \frac{1}{\theta_2^2} \right) \end{pmatrix},$$

wobei $c_{\text{nor}}(x, \pi, \theta_1, \theta_2) := \left[\pi f_{\text{nor}}(x, \theta_1) + (1 - \pi) f_{\text{nor}}(x, \theta_2) \right]^{-1}$.

Schätzung von ρ bei Binomialverteilung mit fester Anzahl von Versuchen

Im Gegensatz zur üblichen Schreibweise verwenden wir hier für die Anzahl der Versuche m und betrachten es als fest (die übliche Parameterbezeichnung durch n ist hier nicht

möglich, da wir mit n immer die Größe der Stichprobe der X_i bezeichnen). Für $x \in \mathbb{N}_0$ und $\theta \in [0,1]$ beschreibt

$$f_{\text{bin}}(x,\theta) := \binom{m}{x} \theta^x \ (1-\theta)^{m-x}$$

die Dichte einer binomialverteilten Zufallsvariable bzgl. dem Zählmaß. In diesem Fall gilt

$$b(x,\pi,\theta_1,\theta_2) = c_{\text{bin}}(x,\pi,\theta_1,\theta_2) \cdot \begin{pmatrix} f_{\text{bin}}(x,\theta_1) + f_{\text{bin}}(x,\theta_2) \\ f_{\text{bin}}(x,\theta_1)\left(\frac{x}{\theta_1} - \frac{m-x}{1-\theta_1}\right) \\ f_{\text{bin}}(x,\theta_2)\left(\frac{x}{\theta_2} - \frac{m-x}{1-\theta_2}\right) \\ f_{\text{bin}}(x,\theta_1)\left[\left(\frac{x}{\theta_1} - \frac{m-x}{1-\theta_1}\right)^2 + \left(\frac{m-x}{(1-\theta_1)^2} - \frac{x}{\theta_1^2}\right)\right] \\ f_{\text{bin}}(x,\theta_2)\left[\left(\frac{x}{\theta_2} - \frac{m-x}{1-\theta_2}\right)^2 + \left(\frac{m-x}{(1-\theta_2)^2} - \frac{x}{\theta_2^2}\right)\right] \end{pmatrix},$$

wobei $c_{\text{bin}}(x,\pi,\theta_1,\theta_2) := \left[\pi f_{\text{bin}}(x,\theta_1) + (1-\pi) f_{\text{bin}}(x,\theta_2)\right]^{-1}$.

2.1.6 Simulationen

In diesem Abschnitt untersuchen wir, wie sich die Teststatistik bei unterschiedlich großen Stichprobengrößen n verhält. Speziell werden wir uns unter der Nullhypothese ansehen, inwiefern für $n = 100, 250, 500, 1000$ ein gewähltes Signifikanzniveau eingehalten wird und wie die Wahl des Penaltyparameters C_k eine Rolle dabei spielt. Für die Untersuchung der Testgüte wählen wir Mischungen mit drei Komponenten.
Wir betrachten zunächst das Verhalten unter der Nullhypothese und simulieren dazu Zweikomponentenmischungen von Normalverteilungen mit festem Erwartungswert ($\mu_1 = \mu_2 = 0$), verschiedenen Standardabweichungen σ_1, σ_2 und dem Mischungsverhältnis π_0. Damit die Ergebnisse nicht zu sehr von der speziellen Wahl der Parameter abhängen, betrachten wir vier verschiedene Situationen. Tabelle 2.1 zeigt die Wahl der Parameter der vier Szenarien N_1, N_2, N_3 und N_4.
Wir führen Simulationen für verschiedene Stichprobengrößen mit den Penaltyparametern $C_2 = 0$ und $C_k = 1$ durch. Dazu simulieren wir 10000 Durchläufe, berechnen jeweils die Teststatistik und zählen die Anzahl der Verwerfungen der Nullhypothese bei drei verschiedenen Signifikanzniveaus $\alpha_1 = 0.1, \alpha_2 = 0.05$ und $\alpha_3 = 0.025$. Die simulierten Testlevels ergeben sich dann aus dem Quotient von Anzahl der Verwerfungen und 10000. Bei den nun folgenden Simulationen schätzen wir k^* nicht mit, sondern wählen es wie durch die entsprechenden Mischungsverhältnisse π_0 gegeben: Bei N_1 $k^* = 5$ und bei den

	σ_1	σ_2	π_0
N_1	1	4	0.67
N_2	2	6	0.50
N_3	1	6	0.57
N_4	2	8	0.55

Tabelle 2.1: Parameter der iid normalverteilten Zweikomponentenmischungen

restlichen $k^* = 4$. Die folgenden Ergebnisse stammen aus einer Implementierung[1] des MLRT der Programmiersprache R. Zur Berechnung der Maximumlikelihood Schätzer benutzen wir die in R vorhandene Optimierungsfunktion `nlm`. Diese Methode benutzt den Newton-Algorithmus und ist eigentlich für unbeschränkte Optimierungsprobleme gedacht. Offensichtlich haben wir hier beschränkte Optimierungsprobleme vorliegen, denn unsere Gewichte müssen in $(0, 1)$ liegen und sich zu Eins aufsummieren. Außerdem müssen die support points positiv sein. Wir benutzen daher innerhalb der modifizierten Log-Likelihoodfunktion (unsere Zielfunktion) Transformationen, die die unbeschränkten Optimierungsparameter jeweils vor der Funktionsauswertung in den zulässigen Bereich transformieren.
Für die support points ist prinzipiell jede stetige Transformation geeignet, die als Bildmenge die positiven reellen Zahlen hat und umkehrbar ist. Eine mögliche Wahl wäre die Exponentialfunktion. Um die Optimierung zu verbessern, benutzen wir nur Transformationen, die in das Intervall $[\varepsilon, M]$ abbilden, wobei M hinreichend groß und $\varepsilon > 0$ hinreichend klein gewählt werden. Konkret wählen wir

$$f_{\text{Trafo}}(s) := M \cdot h(s) + \varepsilon, \text{wobei } h : \mathbb{R} \to (0, 1)\,.$$

Für die Funktion h können wir z.B. die Umkehrfunktion der Logitfunktion oder eine stetige Verteilungsfunktion wählen. Für die vorliegenden Ergebnisse benutzten wir eine Normalverteilung mit Erwartungswert Null und Standardabweichung 100. Die Begründung ist, dass diese Funktion wesentlich flacher verläuft und nicht wie bei der inversen Logitfunktion Werte größer bzw. kleiner 10 bereits fast auf Eins bzw. Null abgebildet werden.
Als Transformation für die Gewichte $\pi_1, \ldots, \pi_{m_0}$ benutzen wir eine Wahl, die von Zuc-

[1] Vielen Dank an Jörn Dannemann, der mir seine Implementierungen zur Verfügung stellte, die ich benutzen und erweitern konnte.

chini (vgl. [ZM09]) vorgeschlagen wird:

$$\pi_j := \frac{\exp(\tau_i)}{1 + \sum_{k=1}^{m_0-1} \exp(\tau_i)} \text{ für } j = 1, \ldots, m_0 - 1.$$

Dabei beschreiben die Variablen $\tau_1, \ldots, \tau_{m_0-1}$ die unbeschränkten Parameter (der Gewichte) der Optimierung. Beim MLRT ist $m_0 = 2$. Diese Transformation funktioniert auch später beim EM-Test, wo m_0 bel. ist.
Tabelle 2.2 zeigt die Ergebnisse der Simulation für die Stichprobengröße $n = 100$.

Level	0.1	0.05	0.025
N_1	0.060	0.029	0.014
N_2	0.029	0.016	0.008
N_3	0.081	0.039	0.018
N_4	0.054	0.025	0.013

Tabelle 2.2: Simulierte Testlevels bei $n = 100$ und $C_k = 1$

Es ist zu erkennen, dass die Testlevels bei $n = 100$ zumindest bei N_3 einigermaßen gut eingehalten werden. In den anderen Fällen testen wir deutlich konservativ. Als nächstes erhöhen wir nun die Stichprobengröße auf $n = 250$. Die Ergebnisse finden sich in der Tabelle 2.3.

Level	0.1	0.05	0.025
N_1	0.092	0.047	0.024
N_2	0.063	0.032	0.015
N_3	0.098	0.052	0.026
N_4	0.079	0.041	0.020

Tabelle 2.3: Simulierte Testlevels bei $n = 250$ und $C_k = 1$

Hier nähern sich in allen Szenarien die simulierten Testlevels den Vorgegebenen an. In allen Fällen bis auf N_2 testen wir mit der asymptotischen Verteilung fast gemäß den wahren (für $n = 250$ gültigen) Levels. Um ein noch deutlicheres Bild zu erhalten, erhöhen wir nun n weiter auf 500 und abschließend auf 1000.
Im dritten Szenario, in dem die support points weit voneinander entfernt sind, testen wir im Fall $n = 500$ und $n = 1000$ anti konservativ (vgl. Tabellen 2.4 und 2.5). Wir haben uns zwar im Vergleich zu $n = 250$ eher verschlechtert, jedoch ist von 500 zu 1000

Level	0.1	0.05	0.025
N_1	0.098	0.052	0.027
N_2	0.077	0.040	0.021
N_3	0.113	0.061	0.034
N_4	0.090	0.047	0.023

Tabelle 2.4: Simulierte Testlevels bei $n = 500$ und $C_k = 1$

Level	0.1	0.05	0.025
N_1	0.100	0.052	0.029
N_2	0.089	0.048	0.027
N_3	0.109	0.059	0.031
N_4	0.097	0.051	0.030

Tabelle 2.5: Simulierte Testlevels bei $n = 1000$ und $C_k = 1$

wieder eine Verbesserung zu erkennen. Um sicher zu gehen, dass hier kein strukturelles Problem der Implementierung vorliegt, simulieren wir in diesem Fall zusätzlich für $n = 5000$. Es ergeben sich dabei die simulierten Levels $0.095, 0.052$ und 0.029. Wir können also erkennen, dass für größer werdende Stichproben die Levels auch hier immer besser übereinstimmen.

Als letztes möchten wir nun in diesem Abschnitt auf die Wahl des Parameters C_k der Penaltyfunktion eingehen. Je größer wir diesen Wert wählen, desto wichtiger wird der Wert der Penaltyfunktion für den modifizierten Log-Likelihoodwert. Die Penaltyfunktion wird betragsmäßig minimal, wenn wir eine gleichmäßige Aufteilung der Gewichte wählen. Je größer wir also C_k wählen, desto vorteilhafter wird es für die modifizierte Log-Likelihoodfunktion die Gewichte nahe 0.5 zu setzen. Wählen wir also C_k groß, so schränken wir praktisch den Alternativraum ein, denn wir können die Gewichte nicht mehr optimal anpassen. Somit verkleinern wir durch eine große Wahl von C_k den Log-Likelihoodwert unter der Alternative und machen somit den Test konservativer.

Mit Hilfe dieser Überlegung können wir z.B. versuchen die Testlevels im Szenario N_3 und der Stichprobengröße $n = 500$ besser einzuhalten (wir testeten hier anti konservativ (vgl. Tabelle 2.4)), indem wir C_k von 1 auf 5 erhöhen. Führen wir nun diese Simulation durch, so erhalten wir die Testlevels $0.08, 0.039$ und 0.02. Wir haben also unseren Test durch die Wahl $C_k = 5$ etwas zu konservativ gemacht und sehen, dass wir durch eine entsprechende Wahl von C_k unsere Levels bei Bedarf anpassen können.

Insgesamt können wir nun aus der Simulation folgende Schlüsse ziehen:

- In unseren betrachteten Fällen stimmen bereits ab $n = 250$ die Levels der Teststatistiken mit denen der asymptotischen Verteilungen jeweils recht gut überein.
- Es kann vorkommen, dass man sich bei steigender Stichprobengröße leicht in den anti konservativen Bereich verschlechtert. Erhöht man jedoch n weiter, so lässt sich eine gute Übereinstimmung erzielen.
- Durch Erhöhen des Parameters C_k kann man den Test konservativer machen und ggf. die Levels besser anpassen. Asymptotisch hat diese Wahl jedoch keinen Einfluss (wir konnten auch z.B. bei N_3 für $C_k = 1$ und $n = 5000$ eine Übereinstimmung der Levels feststellen).

Simulationen unter der Alternative

In diesem Teil möchten wir auf die Güte des Tests eingehen. Dazu simulieren wir drei Szenarien von Dreikomponentenmischungen mit support points wie in Tabelle 2.6 gegeben. Für die Gewichte benutzen wir in allen Szenarien $(0.25, 0.5, 0.25)$.

	σ_1	σ_2	σ_3
A_1	1	4	10
A_2	2	4	8
A_3	1	3	9

Tabelle 2.6: Parameter der Dreikomponentenmischungen

Generell benötigen wir für die iterative Optimierung der Zielfunktion Startwerte. Unter der Hypothese müssen wir einen für das Gewicht und zwei für die support points festlegen. Wir wählen für das Gewicht den Wert 0.5 und für die support points jeweils den Mittelwert der ersten beiden bzw. der letzten beiden wahren support points. Für die Optimierung unter der Alternative setzen wir $k^* = 6$ und benutzen als Startwerte eine gleichmäßige Aufteilung der Ergbnisse der Optimierung unter der Hypothese. Die Tabellen 2.7 bis 2.9 zeigen die Ergebnisse unserer Simulationen.
Wir können erkennen, dass die Szenarien A_1 und A_3 ähnlich gut funktionieren. Bei A_2 liegen die support points enger zusammen als bei den anderen Szenarien und die Testgüte ist für alle Stichprobengrößen deutlich geringer.

Level	0.1	0.05	0.025
A_1	0.773	0.667	0.558
A_2	0.198	0.109	0.059
A_3	0.806	0.695	0.576

Tabelle 2.7: Simulierte Verwerfungsraten bei $n = 250$ und $C_k = 1$

Level	0.1	0.05	0.025
A_1	0.948	0.909	0.855
A_2	0.356	0.231	0.148
A_3	0.973	0.947	0.908

Tabelle 2.8: Simulierte Verwerfungsraten bei $n = 500$ und $C_k = 1$

Level	0.1	0.05	0.025
A_1	0.961	0.913	0.860
A_2	0.570	0.439	0.330
A_3	0.991	0.985	0.976

Tabelle 2.9: Simulierte Verwerfungsraten bei $n = 1000$ und $C_k = 1$

2.2 Erweiterung für Hidden Markov Modelle

Der folgende Abschnitt beschäftigt sich mit einer Abwandlung des MLRT von Dannemann und Holzmann (vgl. [DH08]). Grob zusammengefasst geben sie hier die Unabhängigkeit zwischen den Zufallsvariablen $X_1, \dots X_n$ auf und legen ein stationäres Hidden Markov Modell zu Grunde. Sie behalten so die identische Verteilung der X_i nach einem Mischungsmodell bei. Ziel dieses Tests ist es erneut zu überprüfen, ob das Mischungsmodell höchstens zwei Komponenten besitzt.
Unsere Ausführungen gliedern sich nun in die folgenden Schritte: Zuerst betrachten wir die hier nötigen Annahmen, welche im Wesentlichen mit denen des MLRT übereinstimmen. Im zweiten Teil zeigen wir, dass die asymptotische Verteilung der hier vorliegenden Teststatistik mit der des MLRT übereinstimmt. Abschließend betrachten wir die Performance des Tests bei verschiedenen Stichprobengrößen n.

2.2.1 Annahmen und Notation

Beim MLRT gingen wir davon aus, dass die Daten $X_1, \dots X_n$ unabhängig identisch verteilt einem Mischungsmodell mit m_0 Komponenten folgen. Unter der Nullhypothese $m_0 = 2$ berechneten wir dann die asymptotische Verteilung der Teststatistik (vgl. Abschnitte 2.1.2 bis 2.1.4). Hier gehen wir nun davon aus, dass $X_1, \dots, X_n$ die beobachtbaren Variablen eines Hidden Markov Modells $(S_i, X_i)_{i\in\mathbb{N}}$ gemäß Definition 1.2 sind. Die Markov Kette $(S_i)_{i\in\mathbb{N}}$ kann dabei unter der Nullhypothese $m_0 = 2$ Zustände annehmen und ist per Annahme ergodisch. Diese Annahme stellt die eindeutige Existenz einer strikt positiven stationären Verteilung der Markov Kette sicher (vgl. Abschnitt 1.3).
Für den beobachtbaren Prozess $(X_i)_{i\in\mathbb{N}}$ hatte sich in Abschnitt 1.3 ergeben: Gegeben den Zuständen der Markov Kette $(S_i)_{i\in\mathbb{N}}$ sind die X_i unabhängig voneinander. Die Verteilung von X_i hängt nur von S_i ab und wird über die Dichte $f(\cdot, \theta_{S_i})$ beschrieben. Dabei ist $\theta_{S_i} \in \Theta \overset{kp}{\subset} \mathbb{R}$ ein eindimensionaler reellwertiger Parameter und f wie bei Definition 1.1 eine Dichte bzgl. dem σ-endlichen Maß μ. Das heißt für $i = 1, \dots, n$ gilt

$$(X_i \mid S_i = j) \sim f(\cdot, \theta_j) \text{ mit } j \in \{1, \dots, m\} \text{ und } \theta_j \in \Theta.$$

Da wir nun aber davon ausgehen, dass die Startverteilung der Markov Kette die stationäre Verteilung ist, ist nach Abschnitt 1.3 die Markov Kette ein stationärer Prozess. Diese Annahme ist keine starke Einschränkung, denn für jede beliebige Startverteilung konvergiert unter unseren Annahmen die Verteilung von S_i gegen $\boldsymbol{\pi}^{(s)}$ (vgl. [MS05, Theorem 9.47]). Ferner sind die X_i ebenfalls stationär und gegeben durch das Mischungsmo-

dell mit dem Parameter

$$\left(\pi_1^{(s)}, \ldots, \pi_m^{(s)}, \theta_1, \ldots, \theta_m\right),$$

wobei $\left(\pi_1^{(s)}, \ldots, \pi_m^{(s)}\right)$ die stationäre Verteilung der Markov Kette ist. Wir bezeichnen dieses Mischungsmodell im Folgenden als Marginalverteilung der beobachtbaren Variablen. Der grundlegende Unterschied zu dem bisher betrachteten MLRT ist also, dass die beobachtbaren Variablen X_i zwar identisch verteilt, aber nicht unabhängig sind.
Es soll nun wieder die Nullhypothese $H_0 : G \in \mathcal{M}_2$ gegen $H_1 : G \in \mathcal{M} \backslash \mathcal{M}_2$ getestet werden. Dabei ist $\mathcal{M}_k$ die Menge aller Mischungen mit k Komponenten. Um die asymptotische Verteilung der Teststatistik unter H_0 bestimmen zu können, gehen wir davon aus, dass die Markov Kette zwei Zustände annehmen kann und die stationäre Verteilung $(\pi_0, 1-\pi_0)$ besitzt. Ferner bezeichnen wir mit θ_{01}, θ_{02} die wahren support points. Die Marginalverteilung wird nun gegeben durch die mixing distribution

$$G_0^{\mathrm{HMM}}(\theta) := \pi_0 \, I(\theta_{01} \leq \theta) + (1-\pi_0) \, I(\theta_{02} \leq \theta) \text{ mit } \theta_{01} < \theta_{02} \text{ und } \pi_0 \in (0,1).$$

Völlig analog zum MLRT gehen wir davon aus, dass die Annahmen 2.1 bis 2.4 erfüllt sind. Die Größen

$$Y_{ij}(\theta), Y_i'(\theta), Y_i''(\theta), Y_i'''(\theta) \text{ für } i, 1, \ldots, n \text{ und } j = 1, 2$$

definieren wir ebenfalls entsprechend.

2.2.2 Teststatistik

Die modifizierte Log-Likelihoodfunktion definieren wir wie bei dem MLRT in Definition 2.1 für $G \in \mathcal{M}_k$ über

$$\widetilde{l}_n(G) := l_n(G) + \mathrm{pen}(G),$$

wobei

$$l_n(G) := \sum_{i=1}^{n} \log(f(X_i; G)) \text{ und } \mathrm{pen}(G) := C_k \cdot \sum_{j=1}^{k} \log(\pi_k).$$

Es ist hier zu beachten, dass es sich bei $l_n(G_0^{\mathrm{HMM}})$ nicht um den tatsächlichen Log-Likelihoodwert der beobachtbaren Variablen X_i handelt, sondern die Log-Likelihoodfunktion unabhängiger Zufallsvariablen ist. Die Abhängigkeitsstruktur der X_i wird also an dieser Stelle nicht berücksichtigt (vgl. Abschnitt 2.2.3). Daher bezeichnen wir $l_n(\cdot)$ in diesem Abschnitt als Log-Likelihoodfunktion unter Unabhängigkeit. Wir können nun die Teststatistik R_n wie im vorherigen Test definieren gemäß

$$R_n^{\mathrm{HMM}} := 2 \cdot \left(l_n(\widehat{G}^{\mathrm{HMM}}) - l_n(\widehat{G}_0^{\mathrm{HMM}})\right),$$

wobei $\widehat{G}^{\text{HMM}}$ bzw. $\widehat{G}_0^{\text{HMM}}$ die modifizierten Maximum Likelihoodschätzer über $\mathcal{M}_2$ bzw. $\mathcal{M}_k$ sind und $k \geq \max\left(\left\lfloor \frac{1.5}{\pi_0} \right\rfloor, \left\lfloor \frac{1.5}{1-\pi_0} \right\rfloor, 4\right)$ ist.

2.2.3 Konsistenz und asymptotische Verteilung der Teststatistik

Wir verwenden in den folgenden Abschnitten die Größen aus Definition 2.6 des MLRT. Bei der Herleitung der asymptotischen Verteilung des MLRT spielt die zuvor gezeigte Konsistenz von $\widehat{G} \in \mathcal{M}_k$ (unter H_0) eine entscheidende Rolle. Zuerst folgerten wir nach Leroux (vgl. [Ler92a]), dass $\widehat{G}$ konsistent ist für G_0 (in dem Sinne wie Leroux in [Ler92a] Konsistenz definiert). Unter H_0 sind die Parameter einer Darstellung von G_0 in $\mathcal{M}_k$ nicht identifizierbar und wir könnten somit ohne die Penalisierung keine weitere Aussage treffen. Da wir aber kleine Werte der Gewichte bestrafen, folgerten wir zusätzlich die stochastische Konvergenz der Parameter von $\widehat{G}$ gegen die der optimalen Darstellung $G_0^{(k)}$ (vgl. die Abschnitte 2.1.2 und 2.1.3).
Hier haben wir zusätzlich das Problem, dass wir die Parameter nicht über die echte Log-Likelihoodfunktion schätzen, sondern mittels der Log-Likelihoodfunktion unter Unabhängigkeit. Auf Parameterebene (also unter Annahme eines identifizierbaren Parameters der Marginalverteilung) zeigt Lindgren (vgl. [Lin78]) die Konsistenz der Parameter des Hidden Markov Modells, obwohl er über die Likelihoodfunktion unter Unabhängigkeit optimiert. Diese Aussage können wir so erstmal nicht treffen, jedoch gilt hier

$$\widehat{G}^{\text{HMM}}(\theta) \xrightarrow{P} G_0^{\text{HMM}}(\theta) \text{ für } \theta_{01} \neq \theta \neq \theta_{02}. \tag{2.51}$$

Auf dieser Tatsache aufbauend folgern wir nun analog die Gültigkeit aller Aussagen des Satzes über die Konsistenz (vgl. Satz 2.5) in unserem hier betrachteten Kontext. Insbesondere folgt wegen (2.51) und der Penalisierung, dass $\widehat{G}^{\text{HMM}}$ gegen die optimale Darstellung von G_0^{HMM} konvergiert.
Ein wesentlicher Zwischenschritt zur Herleitung der asymptotischen Verteilung der Teststatistik des MLRT waren die Lemmata 2.7 und 2.8, die zu

$$\left| R_n - \sup_{t_2 \in \mathbb{R}^2_{\geq}} \left(2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2' \widetilde{B}_{22} t_2 \right) \right| = o_P(1)$$

führten (vgl. Gleichung (2.43)). Da die Markov Kette $(S_i)_{i\in\mathbb{N}}$ stationär und ergodisch ist, folgt nach Leroux (vgl. [Ler92b]), dass $(X_i)_{i\in\mathbb{N}}$ ebenfalls ergodisch ist. Es gilt daher nach dem Ergodensatz auch hier

$$n^{-1}\widetilde{B}_{22} \xrightarrow{f.s.} \widetilde{\Sigma}_{22}, \tag{2.52}$$

wobei $\widetilde{\Sigma}_{22}$ in Gleichung (2.42) und $\widetilde{B}_{22}$ in der Formulierung von Lemma 2.7 definiert werden. Da in den Beweisen der Lemmata 2.7 und 2.8 die Unabhängigkeit nur für die Begründung von $\boldsymbol{B} = O_P(n)$ benutzt wurde, was wir hier durch Gleichung (2.52) erhalten, können wir direkt

$$R_n^{\text{HMM}} = \sup_{t_2 \in \mathbb{R}^2_{\geq}} \left(2\widetilde{\boldsymbol{\beta}}_2' t_2 - t_2' \widetilde{B}_{22} t_2 \right) + o_P(1) \tag{2.53}$$

folgern. Im Beweis von Lemma 2.9, welcher die asymptotische Normalverteilung von $n^{-1/2}\widetilde{\boldsymbol{\beta}}_2$ nachweist, ging entscheidend die Unabhängigkeit der X_i ein. Das folgende Lemma zeigt, dass wir diese Aussage auch beim zu Grundelegen eines HMM treffen können.

Lemma 2.11. *Wie in Lemma 2.9 gilt unter den getroffenen Annahmen in Abschnitt 2.2.1 auch hier*

$$\frac{\widetilde{\boldsymbol{\beta}}_2}{\sqrt{n}} \xrightarrow{V.} \mathcal{N}(0, \widetilde{\Sigma}_{22})$$

Beweis. Die X_i sind die beobachtbaren Variablen eines Hidden Markov Modells und wie wir bereits gesehen haben stationär und ergodisch. Ferner ist jedes

$$\boldsymbol{b}_i = (\Delta_i, Y_i'(\theta_{01}), Y_i'(\theta_{02}), Y_i''(\theta_{01}), Y_i''(\theta_{02}))$$

die stetige Transformation des zuhörigen X_i und somit überträgt sich die Stationarität bzw. Ergodizität auf die Folge dieser Zufallsvariablen bzw. auch auf die von

$$\boldsymbol{\alpha}_i' := (Y_i''(\theta_{01}), Y_i''(\theta_{02})) - (\Delta_i, Y_i'(\theta_{01}), Y_i'(\theta_{02}))\, \Sigma_{11}^{-1}\Sigma_{12},$$

welche völlig identisch im Beweis zu Lemma 2.9 definiert wurden. Im Kontext von Lemma 2.9 waren die $\boldsymbol{\alpha}_i$ iid und wir konnten direkt mit dem Zentralen Grenzwertsatz bei unabhängiger identischer Verteilung argumentieren. Stattdessen zeigen wir hier, dass $(\boldsymbol{\alpha}_i)_{i\in\mathbb{N}}$ eine stationäre Martingaldifferenzenfolge (MDF) ist und argumentieren dann über den hier gültigen Zentralen Grenzwertsatz.
Wir bezeichnen für $k = 1, 2$ mit λ_k den bedingten Erwartungswert von $\boldsymbol{b}_1$ gegeben $S_1 = k$, d.h. $\boldsymbol{\lambda}_k := E(\boldsymbol{b}_1 \mid S_1 = k)$. Unter der Nullhypothese kann die Markov Kette zwei Zustände annehmen. Somit erhalten wir

$$E(\boldsymbol{b}_1) = P(S_1 = 1) \cdot \boldsymbol{\lambda}_1 + P(S_1 = 2) \cdot \boldsymbol{\lambda}_2,$$

und da $E(\boldsymbol{b}_1) = 0$ ist, folgt nun

$$\boldsymbol{\lambda}_2 = c_1 \cdot \boldsymbol{\lambda}_1 \tag{2.54}$$

mit $c_1 := -\frac{P(S_1=1)}{P(S_1=2)} = -\frac{\pi_0}{1-\pi_0}$. Als nächstes betrachten wir $\Delta_1 \boldsymbol{b}_1$. Es gilt

$$E\left(\Delta_1 \boldsymbol{b}_1\right) = E\left(\left[\frac{f(X_1,\theta_{01}) - f(X_1,\theta_{02})}{f(X_1; G_0^{\mathrm{HMM}})}\right] \boldsymbol{b}_1\right) \in \mathbb{R}^5.$$

Wir berechnen die k-te Komponente ($k = 1, \ldots, 5$) dieses Vektors. Dabei bezeichnet b_{1k} die k-te Komponente von $\boldsymbol{b}_1$ und bei $b_{1k}(y)$ (für $y \in \mathbb{R}$) setzen wir in die Definition von z.B. $Y_1'(\theta) = \frac{f'(X_1,\theta)}{f(X_1;G_0^{\mathrm{HMM}})}$ y statt X_1 ein. Wir erhalten

$$\begin{aligned} E\left(\Delta_1 b_{1k}\right) &= \int_{\mathbb{R}} \left(\frac{f(y,\theta_{01}) - f(y,\theta_{02})}{f(y; G_0^{\mathrm{HMM}})}\right) b_{1k}(y) f(y; G_0^{\mathrm{HMM}}) dy \\ &= \int_{\mathbb{R}} b_{1k}(y) f(y,\theta_{01}) dy - \int_{\mathbb{R}} b_{1k}(y) f(y,\theta_{02}) dy \\ &= E\left(b_{1k} \mid S_1 = 1\right) - E\left(b_{1k} \mid S_1 = 2\right) \end{aligned}$$

und somit gilt

$$\Sigma\, \mathbf{1}_5 = E\left(\Delta_1 \boldsymbol{b}_1\right) = E\left(\boldsymbol{b}_1 \mid S_1 = 1\right) - E\left(\boldsymbol{b}_1 \mid S_1 = 2\right) = \boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2 \overset{(2.54)}{=} \left(1 - c_1\right) \boldsymbol{\lambda}_1, \quad (2.55)$$

wobei $\Sigma = E\left(\boldsymbol{b}_1 \boldsymbol{b}_1'\right)$ und $\mathbf{1}_5 := (1,0,0,0,0)'$. Schreiben wir $\boldsymbol{\lambda}_1 = (\lambda_{11}, \ldots, \lambda_{15})'$ und setzen $\mathbf{1}_3 := (1,0,0)'$, so folgt aus (2.55)

$$\Sigma_{11}\, \mathbf{1}_3 = \left(1 - c_1\right) \left(\lambda_{11}, \lambda_{12}, \lambda_{13}\right)' \text{ bzw. } \Sigma_{21}\, \mathbf{1}_3 = \left(1 - c_1\right) \left(\lambda_{14}, \lambda_{15}\right)' \quad (2.56)$$

und somit

$$\left(1 - c_1\right)^{-1} \mathbf{1}_3 = \Sigma_{11}^{-1} \left(\lambda_{11}, \lambda_{12}, \lambda_{13}\right)' \text{ bzw. } \left(1 - c_1\right)^{-1} \mathbf{1}_3 = \Sigma_{21}^{-1} \left(\lambda_{14}, \lambda_{15}\right)'. \quad (2.57)$$

Σ_{11} und Σ_{21} stammen dabei aus der Zerlegung von Σ in Blockmatrizen gemäß Gleichung (2.41).
Wir betrachten nun die Filtration $(\mathcal{F}_i)_{i \in \mathbb{N}}$ gegeben durch

$$\mathcal{F}_i := \sigma\left(S_j, \boldsymbol{b}_j;\ j \leq i\right) \text{ für } i \in \mathbb{N}.$$

Mit dieser Definition sind $(\boldsymbol{b}_i \mid \mathcal{F}_{i-1})$ und $(\boldsymbol{b}_i \mid S_{i-1})$ gleich in Verteilung, da die Verteilung von $\boldsymbol{b}_i$ nur abhängig ist vom vorherigen Zustand der Markov Kette. Da $\boldsymbol{\alpha}_i$ sich per Definition direkt über $\boldsymbol{b}_i$ ergibt, ist $\boldsymbol{\alpha}_i$ bzgl. $\mathcal{F}_i$ messbar und es gilt ebenfalls $(\boldsymbol{\alpha}_i \mid \mathcal{F}_{i-1}) \overset{(V.)}{=} (\boldsymbol{\alpha}_i \mid S_{i-1})$. Wir erhalten somit

$$E\left(\boldsymbol{b}_i \mid \mathcal{F}_{i-1}\right) = E\left(\boldsymbol{b}_i \mid S_{i-1}\right) \text{ bzw. } E\left(\boldsymbol{\alpha}_i \mid \mathcal{F}_{i-1}\right) = E\left(\boldsymbol{\alpha}_i \mid S_{i-1}\right).$$

Wir zeigen nun, dass $(\boldsymbol{\alpha}_i)_{i\in\mathbb{N}}$ eine MDF ist und weisen zuerst nach, dass

$$E\left(\boldsymbol{\alpha}_i \mid S_{i-1} = k\right) = 0 \text{ für } k = 1, 2.$$

Zunächst definieren wir $p_{lk} := P(S_i = l \mid S_{i-1} = k)$ für $l, k = 1, 2$ und beobachten für $k = 1, 2$:

$$\begin{aligned} E\left(\boldsymbol{b}_i \mid S_{i-1} = k\right) &= p_{1k}\, E\left(\boldsymbol{b}_i \mid S_i = 1\right) + p_{2k}\, E\left(\boldsymbol{b}_i \mid S_i = 2\right) \\ &= p_{1k}\, \boldsymbol{\lambda}_1 + p_{2k}\, \boldsymbol{\lambda}_2 \overset{(2.54)}{=} \left(p_{1k} + c_1\, p_{2k}\right)\, \boldsymbol{\lambda}_1 \end{aligned}$$

Das zweite Gleichheitszeichen gilt dabei aufgrund der Stationarität von $\boldsymbol{b}_i$. Setzen wir $d_k := p_{1k} + c_1\, p_{2k}$, so gilt also

$$E\left(\boldsymbol{b}_i \mid S_{i-1} = k\right) = d_k\, \boldsymbol{\lambda}_1 \text{ für } k = 1, 2. \tag{2.58}$$

Mit diesem Zwischenergebnis können wir nun die folgende Aussage treffen (wir bezeichnen mit b_{ij} dabei die j-te Komponente von $\boldsymbol{b}_i$):

$$\begin{aligned} E\left(\boldsymbol{\alpha}_i' \mid S_{i-1} = k\right) &= E\left((b_{i4}, b_{i5}) \mid S_{i-1} = k\right) - E\left((b_{i1}, b_{i2}, b_{i3}) \mid S_{i-1} = k\right)\, \Sigma_{11}^{-1}\Sigma_{12} \\ &\overset{(2.58)}{=} d_k\, \left(\lambda_{14}, \lambda_{15}\right) - d_k\, \left(\lambda_{11}, \lambda_{12}, \lambda_{13}\right)\, \Sigma_{11}^{-1}\Sigma_{12} \\ &\overset{(2.57)}{=} d_k \left(\left(\lambda_{14}, \lambda_{15}\right) - (1 - c_1)^{-1}\, \mathbf{1}_3'\Sigma_{12}\right) \\ &\overset{(2.56)}{=} d_k \left(\left(\lambda_{14}, \lambda_{15}\right) - (1 - c_1)^{-1}\, (1 - c_1)\, \left(\lambda_{14}, \lambda_{15}\right)\right) = 0 \end{aligned}$$

Wir erhalten also

$$E\left(\boldsymbol{\alpha}_i \mid \mathcal{F}_{i-1}\right) = E\left(\boldsymbol{\alpha}_i \mid S_{i-1}\right) = 0.$$

Da $\sigma\left(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{i-1}\right) \subset \mathcal{F}_{i-1}$, gilt nach der Iterationseigenschaft für bedingte Erwartungen

$$E\left(\boldsymbol{\alpha}_i \mid \sigma\left(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{i-1}\right)\right) = E\left(E\left(\boldsymbol{\alpha}_i \mid \mathcal{F}_{i-1}\right) \mid \sigma\left(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{i-1}\right)\right) = 0$$

und somit ist $(\boldsymbol{\alpha}_i)_{i\in\mathbb{N}}$ eine ergodische, stationäre MDF.
Nach dem Zentralen Grenzwertsatz für Martingaldifferenzenfolgen (vgl. [Bil61]) gilt

$$n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\alpha}_i \xrightarrow{(V.)} \mathcal{N}(0, \widetilde{\Sigma}_{22}),$$

wobei $\widetilde{\Sigma}_{22} = \mathrm{COV}(\boldsymbol{\alpha}_1)$. Wie im Beweis zu Lemma 2.9 folgt nun hier die Behauptung mit dem Satz von Slutzky. □

Wir haben mit Lemma 2.11 und Gleichung (2.53) dieselben Vorarbeiten geleistet, die zur Herleitung der asymptotischen Verteilung der Teststatistik R_n des MLRT nötig waren und können daher den folgenden Satz formulieren:

Satz 2.12 (Asymptotische Verteilung der Teststatistik)**.** *Falls die Annahmen 2.1 bis 2.4 erfüllt sind, $k \geq k^*$ und X_i die beobachtbaren Variablen des bisher betrachteten Hidden Markov Modells sind, ist die asymptotische Verteilung der Teststatistik die χ^2-Mischung*

$$\left(\frac{1}{2} - \frac{\alpha}{2\pi}\right)\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\alpha}{2\pi}\chi_2^2.$$

Dabei ist χ_0^2 die Einpunktverteilung bei Null und χ_1^2 bzw. χ_2^2 sind χ^2-Verteilungen mit einem bzw. zwei Freiheitsgraden. ρ ist der Korrelationskoeffizient der Kovarianzmatrix $\widetilde{\Sigma}_{22}$ und $\alpha := \cos^{-1}(\rho)$.

Beweis. Wir können hier völlig analog argumentieren wie im Beweis zu Satz 2.10 und verweisen daher auf diesen. □

2.2.4 Simulationen

In diesem Abschnitt führen wir im HMM-Fall zuerst Simulationen unter der Hypothese durch und betrachten anschließend Dreikomponentenmischungen um die Testgüte zu untersuchen. Unter der Hypothese von zwei Komponenten betrachten wir wieder vier Szenarien. Für die support points wählen wir dieselben wie im MLRT unter Unabhängigkeit (vgl. Tabelle 2.1). Das Mischungsverhältnis π_0 ergibt sich hier durch die stationäre Verteilung, welche durch die Matrix der Übergangswahrscheinlichkeiten festgelegt wird. Wir betrachten hier für das Szenario N_l die Übergangsmarix Γ_l (für $l = 1, \ldots, 4$) und definieren diese über

$$\Gamma_1 := \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix},\ \Gamma_2 := \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix},\ \Gamma_3 := \begin{pmatrix} 0.85 & 0.15 \\ 0.2 & 0.8 \end{pmatrix},\ \Gamma_4 := \begin{pmatrix} 0.84 & 0.16 \\ 0.19 & 0.81 \end{pmatrix}.$$

Die stationären Verteilungen dieser Übergangsmatrizen lauten $\pi_{1,stat} = \left(\frac{2}{3}, \frac{1}{3}\right)$, $\pi_{2,stat} = (0.5, 0.5)$,$\pi_{3,stat} \approx (0.57, 0.43)$ und $\pi_{4,stat} \approx (0.55, 0.45)$. Somit stimmen also für jedes Szenario die stationären Verteilungen unserer Stichproben mit denen unter Unabhängigkeit überein (vgl. Tabelle 2.1). Wir können also k^* völlig analog wählen, d.h. bei N_1 ist $k^* = 5$ und sonst $k^* = 4$. Auch hier verwenden wir bei der Schätzung unter der Hypothese keinen Penalty und unter der Alternative den Parameter $C_k = 1$. Die Tabellen 2.10 bis 2.12 zeigen die Ergebnisse der Simulationen.

Auch hier können wir erkennen, dass sich mit steigender Stichprobengröße die simulierten Levels den theoretischen der asymptotischen Verteilung annähern. Bei N_1 und N_3 beobachten wir einen ähnlichen Effekt wie im iid Fall bei N_3 (d.h. für $n = 1000$ testen wir leicht anti-konservativ). Ebenfalls wie im iid Fall gilt für das Szenario N_2, dass wir für alle Stichprobengrößen konservativ testen, uns aber mit steigender Stichprobengröße den vorgegebenen Levels annähern. Generell können wir zwischen den Simulationen unter der Hypothese im iid- und HMM-Fall keine strukturellen Unterschiede feststellen, außer dass wir bei $n = 1000$ im iid Fall die Levels etwas besser einhalten.

Level	0.1	0.05	0.025
N_1	0.087	0.045	0.024
N_2	0.065	0.036	0.019
N_3	0.101	0.054	0.028
N_4	0.078	0.040	0.021

Tabelle 2.10: Simulierte Testlevels bei $n = 250$ und $C_k = 1$

Level	0.1	0.05	0.025
N_1	0.095	0.051	0.029
N_2	0.071	0.036	0.019
N_3	0.113	0.063	0.033
N_4	0.093	0.048	0.025

Tabelle 2.11: Simulierte Testlevels bei $n = 500$ und $C_k = 1$

Level	0.1	0.05	0.025
N_1	0.112	0.061	0.031
N_2	0.084	0.042	0.023
N_3	0.118	0.064	0.036
N_4	0.104	0.056	0.028

Tabelle 2.12: Simulierte Testlevels bei $n = 1000$ und $C_k = 1$

Simulationen unter der Alternative

Wir betrachten nun die Testgüte für die Stichprobengrößen $250, 500$ und 1000. Dazu simulieren wir drei Szenarien von Dreikomponentenmischungen mit den support points wie im iid Fall (vgl. Tabelle 2.6). Für alle Szenarien legen wir die Übergangsmatrix

$$\Gamma_0 := \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.05 & 0.9 & 0.05 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

zugrunde und erhalten so die stationäre Verteilung $(0.25, 0.5, 0.25)$. Unter gültiger Nullhypothese konnten wir die Gewichte der asymptotischen Verteilung der Teststatistik für jedes Szenario berechnen. Dies ist hier nicht möglich, da keine Zweikomponentenmischung vorliegt. Wir schätzen diese Gewichte daher in Abhängigkeit der geschätzten Parameter unter der Hypothese. Die Tabellen 2.13 bis 2.15 zeigen die Ergebnisse unserer Simulationen.
Abschließend möchten wir kurz die Simulationsergebnisse der Testgüte zwischen dem iid- und HMM-Fall vergleichen (für Ergebnisse iid-Fall siehe Tabellen 2.7 bis 2.9): Die Ergebnisse im Szenario A_1 sind in beiden Fällen nahezu identisch (es kommt höchstens zu einer Abweichung von 0.02). Für $n = 1000$ unterscheidet sich bei A_1 und A_3 im iid- und HMM-Fall die Testgüte kaum. Für geringere Stichprobengrößen ist die Testgüte bei Unabhängigkeit etwas größer (Abweichung im Bereich $0.02 - 0.05$). Der beobachtete Unterschied wird jedoch von $n = 250$ zu $n = 500$ geringer. Insgesamt können wir also festhalten, dass in zwei von drei betrachteten Fällen die Testgüte bei Unabhängigkeit für kleine Stichprobengrößen geringfügig besser ist. Mit steigender Stichprobengröße verschwindet jedoch dieser Unterschied.

Level	0.1	0.05	0.025
A_1	0.719	0.606	0.498
A_2	0.186	0.109	0.062
A_3	0.744	0.632	0.520

Tabelle 2.13: Simulierte Verwerfungsraten bei $n = 250$ und $C_k = 1$

Level	0.1	0.05	0.025
A_1	0.938	0.895	0.837
A_2	0.349	0.230	0.145
A_3	0.950	0.916	0.865

Tabelle 2.14: Simulierte Verwerfungsraten bei $n = 500$ und $C_k = 1$

Level	0.1	0.05	0.025
A_1	0.962	0.924	0.875
A_2	0.571	0.433	0.318
A_3	0.992	0.986	0.976

Tabelle 2.15: Simulierte Verwerfungsraten bei $n = 1000$ und $C_k = 1$

3 Test auf mehrere Komponenten

Bisher haben wir mit dem MLRT ein Verfahren vorgestellt, welches uns die Möglichkeit gibt für ein Mischungsmodell zu testen, ob höchstens zwei Komponenten vorliegen. Dieses Testverfahren ließ sich sowohl beim Vorliegen unabhängiger Stichproben als auch auf die beobachtbaren Variablen eines HMM anwenden. Speziell für den Fall, falls man mit dem MLRT die Nullhypothese verwirft, ist man daran interessiert die Hypothese von Mischungen höherer Ordnung zu untersuchen. Chen und Li (Vgl. [CL09]) stellen ein Verfahren namens EM-Test vor, welches genau zur Untersuchung solcher Fragestellungen konzipiert wurde. Es kann hiermit nicht nur die Hypothese $m_0 = 3$ getestet werden, sondern es lässt sich für den Test einer beliebigen Ordnung m_0 anwenden. Daher wird hier getestet

$$H_0 : G \in \mathcal{M}_{m_0} \text{ gegen } H_1 : G \in \mathcal{M} \backslash \mathcal{M}_{m_0}.$$

Zwischen dem MLRT und dem EM-Test besteht ein enger Zusammenhang, so wird sich beispielsweise herausstellen, dass der EM-Test für $m_0 = 2$ die gleiche asymptotische Verteilung wie der MLRT besitzt. Ferner folgt die Teststatistik des EM-Tests für ein beliebiges m_0 asymptotisch einer Mischung aus $m_0 + 1$ χ^2-verteilten Zufallsvariablen. Das folgende Kapitel gliedert sich in zwei Abschnitte: Im ersten Teil werden wir für unabhängige Stichproben den EM-Test vorstellen. Anschließend untersuchen wir simulationsbasiert, inwiefern sich dieser Test auf Hidden Markov Modelle übertragen lässt.

3.1 EM-Test

3.1.1 Notation und Annahmen

Zunächst legen wir die Verteilung der Stichprobe unter der Nullhypothese fest. Wir gehen dabei in diesem Teil davon aus, dass wir die unabhängigen Zufallsvariablen $X_1, \ldots, X_n$ des folgenden Mischungsmodells vorliegen haben: Mit $f(\cdot, \theta)$ bezeichnen wir die Kom-

ponentendichte und mit

$$\Psi_0(\theta) := \sum_{h=1}^{m_0} \pi_{0h}\, I(\theta_{0h} \leq \theta)$$

die mixing distribution für $\theta \in \Theta \subset \mathbb{R}$. Dabei sei $\theta_{01} < \cdots < \theta_{0m_0}$ sowie $\pi_{0h} \in (0,1)$ für $h = 1, \ldots, m_0$ mit $\sum_{h=1}^{m_0} \pi_{0h} = 1$. Ferner fassen wir zur Abkürzung der Schreibweise die Parameter zu den beiden Vektoren $\boldsymbol{\pi}_0 := (\pi_{01}, \ldots, \pi_{0m_0})$, $\boldsymbol{\theta}_0 := (\theta_{01}, \ldots, \theta_{0m_0})$ zusammen. X_i folgt also einem m_0-Komponenten Mischungsmodell mit Dichte

$$f(x, \Psi_0) = \int_\Theta f(x, \theta)\, d\Psi_0(\theta) = \sum_{h=1}^{m_0} \pi_{0h} f(x, \theta_{0h}).$$

Die gewöhnliche Log-Likelihoodfunktion ist hier gegeben durch

$$l_n(\Psi) = \sum_{i=1}^{n} \log\left(f(X_i, \Psi)\right) \text{ für } \Psi \in \mathcal{M}_{m_0}.$$

Wir benötigen später eine stetige Funktion, die kleine bzw. große Gewichte bestraft. Diese bezeichnen wir mit $\boldsymbol{p}(\gamma)$ für $\gamma \in (0,1)^{m_0}$ und wählen sie so, dass für $\gamma_h \to 0$ oder 1 $\boldsymbol{p}(\gamma) \to -\infty$. Chen und Li wählen für $\boldsymbol{p}$ z.B. die Funktion

$$\boldsymbol{p}(\gamma) := C \sum_{h=1}^{m_0} p(\gamma_h) := C \sum_{h=1}^{m_0} \log(1 - |1 - 2\gamma_h|),$$

wobei $C > 0$ eine beliebige positive Konstante ist.
Als nächstes geben wir nun die für diesen Test benötigten Annahmen an.Zunächst müssen wir wie beim MLRT in Gleichung (2.1) fordern, dass zwei mixing distributions (hier aus $\mathcal{M}_{m_0}$) identifizierbar sind. Außerdem benötigen wir eine Annahme, welche die Konsistenz des Maximum Likelihoodschätzers sicher stellt:

Annahme 3.1 (Integrierbarkeits- bzw. Konsistenzbedingungen)**.** Völlig analog zum MLRT fordern wir die Annahme 2.1.

Bei der Taylorentwicklung im Beweis der asymptotischen Verteilung der Teststatistik müssen wir auch hier die Ableitungen nach den Parametern der Komponentendichte bilden. Die folgende Annahme stellt deren Existenz sicher.

Annahme 3.2 (Glattheit)**.** Der Träger von $f(x, \theta)$ ist unabhängig von θ und $f(x, \theta)$ ist viermal stetig differenzierbar nach θ.

Wir können nun für $i = 1, \ldots, n$ und $h = 1, \ldots, m_0$ die Größen

$$Y_i(\theta) := \frac{f'(X_i, \theta)}{f(X_i; \Psi_0)},\ Z_i(\theta) := \frac{f''(X_i, \theta)}{2\, f(X_i; \Psi_0)},\ \Delta_{ih} := \frac{f(X_i, \theta_{0h}) - f(X_i, \theta_{0m_0})}{f(X_i; \Psi_0)}$$

definieren. Ferner fassen wir diese Zufallsvariablen ausgewertet an den wahren support points zu den Vektoren

$$\boldsymbol{b}_{1i} := (\Delta_{i1}, \ldots, \Delta_{im_0-1}, Y_i(\theta_{01}), \ldots, Y_i(\theta_{0m_0}))\,,\ \boldsymbol{b}_{2i} := (Z_i(\theta_{01}), \ldots, Z_i(\theta_{0m_0}))$$

zusammen. Als nächstes treffen wir eine Annahme, die die gleichmäßige Beschränktheit der soeben definierten Größen nicht für alle $\theta \in \Theta$, sondern nur in allen ϵ-Umgebungen um die wahren support points fordert.

Annahme 3.3 (Gleichmäßige Beschränktheit)**.** Sei $N(\theta, \epsilon) := \{\theta' \in \Theta : |\theta' - \theta| \leq \epsilon\}$ für $\theta \in \Theta$ und $\epsilon > 0$. Es existiert eine integrierbare Funktion $g(\cdot)$ (d.h. $E(|g(X)|) < \infty$, wobei der Erwartungswert bzgl. der wahren Dichte $f(\cdot, \Psi_0)$ berechnet wird) und ein $\epsilon_0 > 0$, s.d.

$$|\Delta_{ih}|^3 \leq g(X_i),\ |Y_i(\theta)|^3 \leq g(X_i),\ |Z_i^{(k)}(\theta)|^3 \leq g(X_i),$$

für $\theta \in N(\theta_{0h}, \epsilon_0)$ mit $h \in \{1, \ldots, m_0\}$ und $k = 0, 1, 2$. Dabei beschreibt $Z_i^{(k)}(\theta)$ die k-te Ableitung von $Z_i(\theta)$.

Die Annahme der gleichmäßigen Beschränktheit impliziert die Straffheit von z.B.

$$n^{-1/2} \sum_{i=1}^{n} Z_i^{(1)}(\theta)$$

in einer ε-Umgebung der wahren support points und ist somit eine stärkere Annahme. Da die X_i unabhängig identisch verteilt sind, gilt dies auch für den Zufallsvektor $\boldsymbol{b}_i := (\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i})$. Die folgende Annahme richtet sich an die zugehörige Kovarianzmatrix.

Annahme 3.4 (Positive Definitheit der Kovarianzmatrix)**.** $\boldsymbol{\Sigma} := \mathrm{COV}(\boldsymbol{b}_1)$ ist positiv definit.

3.1.2 Iterative Berechnung der Teststatistik

Die Teststatistik des EM-Tests wird in mehreren Schritten berechnet. Wir geben nun einen kurzen Überblick, wie diese aussehen. Anschließend werden wir die einzelnen Schritte im Detail betrachten.

1. Berechnung des Maximum Likelihoodschätzers unter der Hypothese ($\widehat{\Psi}_0 \in \mathcal{M}_{m_0}$).

Für verschiedene $\boldsymbol{\beta} \in (0, 1)^{m_0}$ folgen dann die Schritte zwei und drei.

2. Berechnung des Maximum Likelihoodschätzers $\Psi^{(1)}(\boldsymbol{\beta})$ aus einer Teilmenge $\Omega_{2m_0}(\boldsymbol{\beta})$ des Alternativenraumes $\mathcal{M}_{2m_0}$. Diese Einschränkung ist, dass um jeden der geschätzten support points von $\widehat{\Psi}_0$ hier zwei support points liegen müssen, wobei die Aufteilung von $\boldsymbol{\beta}$ abhängt.

3. Durchführung mehrerer spezieller EM Iterationen von den Parametern von $\Psi^{(1)}(\boldsymbol{\beta})$ ausgehend. Der Zweck dieses Schrittes ist es, die Parameter speziell unter verletzter Nullhypothese besser anpassen zu könnnen. Ergebnis dieses Schrittes ist die logarithmierte Likelihoodratio $M(\boldsymbol{\beta})$. Im ausführlichen Teil erhält die logarithmierte Likelihoodratio noch die Indizes n und k, die für die Größe der Stichprobe und die Anzahl der EM Iterationen stehen.

4. Berechnung der Teststatistik als Maximum aller $M(\boldsymbol{\beta})$.

Erster Schritt: Berechnung Schätzer unter der Hypothese

Als erstes bestimmen wir den Maximum Likelihoodschätzer unter der Hypothese. Dieser ergibt sich über

$$\widehat{\Psi}_0 := \underset{\Psi \in \mathcal{M}_{m_0}}{\operatorname{argmax}}\, l_n(\Psi).$$

Dabei bezeichnen wir mit $(\hat{\theta}_{01}, \ldots, \hat{\theta}_{0h})$ die zugehörigen geschätzten support points und mit $(\hat{\pi}_{01}, \ldots, \hat{\pi}_{0h})$ die entsprechenden Gewichte. Praktisch können wir diese durch direktes Maximieren der logarithmierten Likelihoodfunktion oder den EM Algorithmus bestimmen (vgl. Abschnitt 1.2).
Es sei nun $\boldsymbol{\beta} \in (0,1)^{m_0}$. Wir werden die beiden folgenden Schritte in Abhängigkeit von $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{m_0})$ erläutern. Anschließend gehen wir darauf ein, welche Werte wir für $\boldsymbol{\beta}$ wählen.

Zweiter Schritt: Berechnung Startwerte für Schätzer unter der Alternative

Zunächst zerlegen wir den Parameterraum in m_0 Intervalle, welche von den support points von $\widehat{\Psi}_0$ abhängen. Wir definieren

$$I_h := \left(\frac{\hat{\theta}_{0h-1} + \hat{\theta}_{0h}}{2}, \frac{\hat{\theta}_{0h} + \hat{\theta}_{0h+1}}{2} \right] \text{ für } h = 2, \ldots, m_0 - 1$$

und

$$I_1 := \left(\eta_1, \frac{\hat{\theta}_{01} + \hat{\theta}_{02}}{2} \right], \; I_{m_0} := \left(\frac{\hat{\theta}_{0m_0-1} + \hat{\theta}_{0m_0}}{2}, \eta_2 \right),$$

wobei $\eta_1 := \inf \Theta$ und $\eta_2 := \sup \Theta$. Falls wir einen kompakten Parameterraum vorliegen haben, würden wir die Intervalle I_1 bzw. I_{m_0} links bzw. rechts abgeschlossen definieren. Anschaulich bedeutet diese Definition, dass wir Intervalle bilden, die jeweils einen support point von $\widehat{\Psi}_0$ enthalten und die Randpunkte gerade dem Mittelpunkt zwischen zwei support points entsprechen. Von diesen Intervallen ausgehend definieren wir nun durch

$$\Omega_{2m_0}(\boldsymbol{\beta}) := \left\{ \sum_{h=1}^{m_0} \pi_h \beta_h \; I(\theta_{1h} \leq \theta) + \pi_h (1 - \beta_h) \; I(\theta_{2h} \leq \theta) : \theta_{1h}, \theta_{2h} \in I_h, \; \sum_{h=1}^{m_0} \pi_h = 1 \right\}$$

eine Teilmenge von $\mathcal{M}_{2m_0}$. D.h. wir wählen spezielle $2m_0$-Komponenten mixing distributions aus. Bei dieser Definition ist zu beachten, dass $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{m_0})$ bereits vor der Auswahl möglicher Gewichte π_h festgelegt wurde und somit nur spezielle Aufteilungen der Gewichte erreicht werden können. Präziser können wir nur für $h = 1, \ldots, m_0$ das gemeinsame Gewicht von θ_{1h} und θ_{2h} frei wählen. Ihre Aufteilung wird bereits vorher durch β_h gewählt.
Diese Einschränkung hat einen entscheidenden Vorteil, wie das folgende Beispiel zeigt.

Beispiel 3.1. Wir betrachten die m_0-Komponenten mixing distribution $G(\theta) \in \mathcal{M}_{m_0}$ mit den support points $\left(\theta_1^{(G)}, \ldots, \theta_{m_0}^{(G)}\right)$ und den Gewichten $\left(\pi_1^{(G)}, \ldots, \pi_{m_0}^{(G)}\right)$. Weiterhin seien alle $\beta_h \in (0, 1)$. Möchten wir nun diese mixing distribution in $\Omega_{2m_0}(\boldsymbol{\beta})$ darstellen, so müssen wir jeweils zwei support points der Darstellung gleich $\theta_h^{(G)}$ und die Gewichte gleich $\pi_h^{(G)}$ wählen. Wir vermeiden somit durch diese Einschränkung die prinzipielle Möglichkeit einzelne Gewichte der Darstellung gleich Null und den zugehörigen support point beliebig zu setzen. Diese Beobachtung wird für die Konsistenz des (modifizierten) Maximum Likelihoodschätzers in $\Omega_{2m_0}(\boldsymbol{\beta})$, der nun definiert wird, eine entscheidende Rolle spielen.

Wir definieren mittels

$$pl_n(\Psi) := l_n(\Psi) + p(\boldsymbol{\beta})$$

eine modifizierte Version der Log-Likelihoodfunktion und setzen über

$$\Psi^{(1)}(\boldsymbol{\beta}) := \underset{\Psi \in \Omega_{2m_0}(\boldsymbol{\beta})}{\operatorname{argmax}} \; pl_n(\Psi)$$

den modifizierten Maximum Likelihoodschätzer unter der eingeschränkten Alternative. Bezüglich der Optimierung sei nochmals angemerkt, dass hier (wegen Beschränkung auf $\Omega_{2m_0}(\boldsymbol{\beta})$) nur m_0 Gewichte $(\pi_1, \ldots, \pi_{m_0})$ und $2m_0$ support points gewählt werden können. Diese definieren dann zusammen mit $\boldsymbol{\beta}$ die eigentlichen Gewichte der mixing distribution. $\boldsymbol{\beta}$ kann in diesem Schritt nicht verändert werden. $\left(\theta_{11}^{(1)}, \theta_{21}^{(1)}, \ldots, \theta_{1m_0}^{(1)}, \theta_{2m_0}^{(1)}\right)$

sind die durch die Optimierung gewählten support points von $\Psi^{(1)}(\boldsymbol{\beta})$, wobei $\theta_{1h}^{(1)}, \theta_{2h}^{(1)} \in I_h$. Die gewählten Gewichte der Optimierung bezeichnen wir mit $\boldsymbol{\pi}^{(1)} := \left(\pi_1^{(1)}, \ldots, \pi_{m_0}^{(1)}\right)$. Somit sind die eigentlichen Gewichte von $\Psi^{(1)}(\boldsymbol{\beta})$ gegeben durch

$$\left(\pi_1^{(1)}\beta_1, \pi_1^{(1)}(1-\beta_1), \ldots, \pi_{m_0}^{(1)}\beta_{m_0}, \pi_{m_0}^{(1)}(1-\beta_{m_0})\right).$$

Wir fassen die $2m_0$ support points zu den zwei Vektoren $\boldsymbol{\theta}_1^{(1)} := (\theta_{11}^{(1)}, \ldots, \theta_{1m_0}^{(1)})$ und $\boldsymbol{\theta}_2^{(1)} := (\theta_{21}^{(1)}, \ldots, \theta_{2m_0}^{(1)})$ zusammen. Diese Wahl hat den Hintergrund, dass von beiden Vektoren die h-te Komponente in I_h, dem Intervall um den wahren support point θ_{0h} liegt und somit im Falle von Konsistenz beide Vektoren gegen den wahren support point Vektor $\boldsymbol{\theta}_0$ konvergieren.
Die eigentlichen Gewichte der mixing distribution konnten nicht frei gewählt werden (lediglich der Vektor der π_h). Diese Tatsache könnte unter der Alternative zu schlechten Anpassungen führen und so die Teststärke verschlechtern. Der folgende Schritt dient dazu diesen Effekt abzumildern.

Dritter Schritt: EM Iteration

In diesem Schritt möchten wir die Schätzung unter der Alternative weiter verbessern. Speziell lassen wir nun auch eine Optimierung nach $\boldsymbol{\beta}$ zu. Dafür führen wir EM-Iterationen aus, die die penalisierten Log-Likelihoodfunktion weiter verbessern.
Die Mischungen in $\Omega_{2m_0}(\boldsymbol{\beta})$ haben m_0 Zustände, die sich jeweils wieder in zwei aufteilen (insgesamt also $2m_0$ Zustände). Für $h \in \{1, \ldots, m_0\}$ bezeichnen wir diese beiden Zustände mit $1h$ und $2h$. Insgesamt ergeben sich somit (passend zur Indizierung der support points) die Zustände $11, 21, \ldots, 1m_0, 2m_0$. Für die Berechnung der Iterationsvorschrift benötigen wir die penalisierte Log-Likelihoodfunktion der kompletten Daten. Diese lautet

$$pl_n^{(C)}(\Psi \mid \boldsymbol{x}, \boldsymbol{s}) = \sum_{i=1}^{n}\sum_{h=1}^{m_0} I(s_i = 1h)\ \log(\pi_h\beta_h f(x_i, \theta_{1h})) + I(s_i = 2h)\ \log(\pi_h(1-\beta_h)f(x_i, \theta_{2h})) + p(\beta_h),$$

wobei $\boldsymbol{x}$ bzw. $\boldsymbol{s}$ die Vektoren der Beobachtungen bzw. der realisierten Zustände bezeichnen. Wie im normalen EM-Algorithmus (vgl. Abschnitt 1.2) benötigen wir Startwerte für die Iteration. Für diese wählen wir die Parameter von $\Psi^{(1)}(\boldsymbol{\beta})$. Die mixing distribution resultierend aus dem k-ten Schritt bezeichnen wir mit $\Psi^{(k)}(\boldsymbol{\beta})$ und ihre Parameter entsprechend mit $\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\pi}^{(k)}$, $\boldsymbol{\theta}_1^{(k)}$ und $\boldsymbol{\theta}_2^{(k)}$. Dabei ist insbesondere $\boldsymbol{\beta}^{(1)} := \boldsymbol{\beta}$.
Wir geben nun den k-ten Schritt ($k \geq 2$) der Iteration an und berechnen zunächst die Wahrscheinlichkeit dafür, dass die Zustandsvariable im Zeitpunkt i gegeben der

Stichprobe und den aktuellen Parametern einen gewissen Zustand angenommen hat. Wir definieren daher für $h = 1, \ldots, m_0$

$$w_{i1h}^{(k)} := P\left(S_i = 1h \mid \boldsymbol{X}, \Psi^{(k-1)}(\boldsymbol{\beta})\right) = \frac{\pi_h^{(k-1)} \beta_h^{(k-1)} f(X_i, \theta_{1h}^{(k-1)})}{f(X_i; \Psi^{(k-1)}(\boldsymbol{\beta}))}$$

und

$$w_{i2h}^{(k)} := P\left(S_i = 2h \mid \boldsymbol{X}, \Psi^{(k-1)}(\boldsymbol{\beta})\right) = \frac{\pi_h^{(k-1)} (1-\beta_h^{(k-1)}) f(X_i, \theta_{2h}^{(k-1)})}{f(X_i; \Psi^{(k-1)}(\boldsymbol{\beta}))}.$$

Wie in Abschnitt 1.2 setzen wir in die penalisierte Log-Likelihoodfunktion der kompletten Daten für die nicht beobachtbaren $I(s_i = 1h)$ bzw. $I(s_i = 2h)$ die Werte $w_{i1h}^{(k)}$ bzw. $w_{i2h}^{(k)}$ ein. Wir erhalten den Erwartungswert der penalisierten Log-Likelihoodfunktion gegeben unserer Stichprobe und den aktuellen Parametern damit über

$$pl_n^{(C,k)}(\Psi) = \sum_{i=1}^{n} \sum_{h=1}^{m_0} \left[w_{i1h}^{(k)} \ \log(\pi_h \beta_h f(X_i, \theta_{1h})) + w_{i2h}^{(k)} \ \log(\pi_h (1-\beta_h) f(X_i, \theta_{2h})) + p(\beta_h) \right].$$

Die Updates für die support points und die Gewichte erhalten wir aus der Maximierung dieser Funktion nach der mixing distribution bzw. nach den Parametern, d.h.

$$\Psi^{(k)}(\boldsymbol{\beta}) := \underset{\Psi \in \Omega_{2m_0}(\boldsymbol{\beta})}{\operatorname{argmax}} \left(pl_n^{(C,k)} \right).$$

Für die Berechnung zerlegen wir diese Funktion in $2m_0$ Teile, die jeweils von den support points θ_{1h} bzw. θ_{2h} abhängen, einen Teil der von den Gewichten $(\pi_1, \ldots, \pi_{m0})$ bestimmt wird und weitere m_0 Teile, die jeweils von β_h abhängen. Zur Abkürzung der Schreibweise definieren wir zunächst

$$a_{1h}^{(k)} := \sum_{i=1}^{n} w_{i1h}^{(k)} \text{ und } a_{2h}^{(k)} := \sum_{i=1}^{n} w_{i2h}^{(k)}.$$

Es gilt

$$\begin{aligned}
pl_n^{(C,k)}(\Psi) &= \sum_{i=1}^{n} \sum_{h=1}^{m_0} w_{i1h}^{(k)} \ \log(f(X_i, \theta_{1h})) + \sum_{i=1}^{n} \sum_{h=1}^{m_0} w_{i2h}^{(k)} \ \log(f(X_i, \theta_{2h})) \\
&+ \sum_{i=1}^{n} \sum_{h=1}^{m_0} w_{i1h}^{(k)} \ \log(\pi_h) + \sum_{i=1}^{n} \sum_{h=1}^{m_0} w_{i2h}^{(k)} \ \log(\pi_h) \\
&+ \sum_{i=1}^{n} \sum_{h=1}^{m_0} w_{i1h}^{(k)} \ \log(\beta_h) + \sum_{i=1}^{n} \sum_{h=1}^{m_0} w_{i2h}^{(k)} \ \log(1-\beta_h) + \sum_{h=1}^{m_0} p(\beta_h) \\
&= \sum_{h=1}^{m_0} \sum_{i=1}^{n} w_{i1h}^{(k)} \ \log(f(X_i, \theta_{1h})) + \sum_{h=1}^{m_0} \sum_{i=1}^{n} w_{i2h}^{(k)} \ \log(f(X_i, \theta_{2h})) \\
&+ \sum_{h=1}^{m_0} \log(\pi_h) \left(a_{1h}^{(k)} + a_{2h}^{(k)} \right) + \sum_{h=1}^{m_0} \left(\log(\beta_h) a_{1h}^{(k)} + \log(1-\beta_h) a_{2h}^{(k)} \right) + \sum_{h=1}^{m_0} p(\beta_h)
\end{aligned}$$

Wir können somit zunächst jeden support point einzeln für $h = 1, \ldots, m_0$ über

$$\theta_{1h}^{(k)} = \underset{\theta \in \Theta}{\operatorname{argmax}} \left(\sum_{i=1}^{n} w_{i1h}^{(k)} \, \log(f(X_i, \theta)) \right)$$

bzw.

$$\theta_{2h}^{(k)} = \underset{\theta \in \Theta}{\operatorname{argmax}} \left(\sum_{i=1}^{n} w_{i2h}^{(k)} \, \log(f(X_i, \theta)) \right)$$

ermitteln. Ferner berechnen sich die Gewichte $\boldsymbol{\pi}^{(k)}$ über

$$\boldsymbol{\pi}^{(k)} = \operatorname{argmax} \left(\sum_{h=1}^{m_0} \log(\pi_h) \left(a_{1h}^{(k)} + a_{2h}^{(k)} \right) \right), \text{ s.t. } \pi_h > 0, \; \sum_{h=1}^{m_0} \pi_h = 1$$

und damit gilt

$$\pi_h^{(k)} = n^{-1} \left(a_{1h}^{(k)} + a_{2h}^{(k)} \right) \text{ für } h = 1, \ldots, m_0.$$

Abschließend ist nur noch für $h = 1, \ldots, m_0$ das Gewicht $\beta_h^{(k)}$ zu ermitteln. Dieses ergibt sich wegen obiger Darstellung von $pl_n^{(C,k)}(\Psi)$ über

$$\beta_h^{(k)} = \underset{\beta \in (0,1)}{\operatorname{argmax}} \left(\log(\beta) a_{1h}^{(k)} + \log(1-\beta) a_{2h}^{(k)} + p(\beta) \right).$$

Zu Beginn des Algorithmus geben wir $K \in \mathbb{N}$ vor und führen obige Iteration K mal durch. Somit erhalten wir am Ende die Schätzung unter der Alternative $\Psi^{(K+1)}(\boldsymbol{\beta})$. Als Abschluss dieses Schrittes definieren wir die Log-Likelihoodratio

$$M_n^{(K+1)}(\boldsymbol{\beta}) := 2 \left(pl_n \left(\Psi^{(K+1)}(\boldsymbol{\beta}) \right) - l_n \left(\widehat{\Psi}_0 \right) \right).$$

Vierter Schritt: Berechnung der Teststatistik

Da wir in den Schritten zwei und drei von einem bel. $\boldsymbol{\beta} \in (0,1)^{m_0}$ ausgegangen sind, ist $M_n^{(K+1)}(\boldsymbol{\beta})$ (wie die Notation schon andeutet) abhängig von dieser Wahl. Um diesen Effekt abzuschwächen und unter der Alternative eine möglichst gute Anpassung an die vorliegende mixing distribution zu gewährleisten, starten wir die Schritte zwei und drei von mehereren $\boldsymbol{\beta}$ aus. Genauer definieren wir die Menge $B := \{0.1, 0.3, 0.5\}$ und starten obige Schritte für jedes $\boldsymbol{\beta} \in B^{m_0}$. Insgesamt führt dies zu 3^{m_0} Durchläufen. Abschließend definieren wir die Teststatistik des EM-Tests über

$$EM_n^{(K)} := \max_{\boldsymbol{\beta} \in B^{m_0}} \left\{ M_n^{(K+1)}(\boldsymbol{\beta}) \right\}.$$

3.1.3 Asymptotische Eigenschaften

In diesem Abschnitt halten wir die beiden Hauptresultate bezüglich der Asymptotik fest. Einerseits sind die Schätzer von obiger Iteration konsistent und andererseits konvergiert die Teststatistik gegen eine Verteilung, welche stark an den MLRT erinnert. Im Fall $m_0 = 2$ konvergieren $EM_n^{(K)}$ und die Teststatistik des MLRT sogar gegen dieselbe Verteilung. Zunächst gehen wir auf die Konsistenz ein:

Satz 3.1 (Konsistenz)**.** *Unter der Nullhypothese und unseren getroffenen Annahmen gilt für jedes $\boldsymbol{\beta} \in B^{m_0}$ und $k \in \{1, \ldots, K+1\}$: Die Schätzer $\boldsymbol{\pi}^{(k)}, \boldsymbol{\theta}_1^{(k)}$ und $\boldsymbol{\theta}_2^{(k)}$ sind konsistent und bezüglich der Ordnung gilt*

$$\boldsymbol{\pi}^{(k)} - \boldsymbol{\pi}_0 = O_P(n^{-1/2}),\ \boldsymbol{\theta}_1^{(k)} - \boldsymbol{\theta}_0 = O_P(n^{-1/4}) \text{ und } \boldsymbol{\theta}_2^{(k)} - \boldsymbol{\theta}_0 = O_P(n^{-1/4}).$$

Beweisidee. Zunächst begründet man unter der Nullhypothese, dass die geschätzte mixing distribution $\Psi^{(1)}(\boldsymbol{\beta})$ konsistent ist an allen Stetigkeitsstellen θ. Da die Menge B die Null nicht enthält, ist dieses Resultat nur möglich, falls die beiden support points $\theta_{1h}^{(1)}, \theta_{2h}^{(1)}$ aus I_h gegen den wahren support point θ_{0h} konvergieren. Somit muss nun auch ihr gemeinsames Gewicht gegen das des wahren support points konvergieren. Wir erhalten also die Konsistenz der Parameter für $k = 1$, d.h. diese gilt bereits ohne EM-Iteration. Als nächstes definieren wir die Momente für $h = 1, \ldots, m_0$

$$m_{1h}^{(k)} := \beta_h^{(k)} \left(\theta_{1h}^{(k)} - \theta_{0h}\right) + \left(1 - \beta_h^{(k)}\right) \left(\theta_{2h}^{(k)} - \theta_{0h}\right)$$

und

$$m_{2h}^{(k)} := \beta_h^{(k)} \left(\theta_{1h}^{(k)} - \theta_{0h}\right)^2 + \left(1 - \beta_h^{(k)}\right) \left(\theta_{2h}^{(k)} - \theta_{0h}\right)^2.$$

Hiermit definieren wir den Vektor

$$\boldsymbol{t}^{(k)} := \left(\pi_1^{(k)} - \pi_{01}, \ldots, \pi_{m_0-1}^{(k)} - \pi_{0m_0-1}, \pi_1^{(k)} m_{11}^{(k)}, \ldots, \pi_{m_0}^{(k)} m_{1m_0}^{(k)}, \pi_1^{(k)} m_{21}^{(k)}, \ldots, \pi_{m_0}^{(k)} m_{2m_0}^{(k)}\right)'.$$

Mit $\boldsymbol{b}_i$ bezeichnen wir den Vektor $(\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i})$ für $i = 1, \ldots, n$. Diese Vektoren sind unabhängig identisch verteilt und mit $\boldsymbol{\Sigma}$ bezeichnen wir die zugehörige Kovarianzmatrix. Mit Hilfe dieser Größen können wir (wie im MLRT) mit einer Taylorentwicklung die folgende obere Schranke für $R_{1n}\left(\Psi^{(1)}(\boldsymbol{\beta})\right) := 2\left(pl_n\left(\Psi^{(1)}(\boldsymbol{\beta})\right) - l_n\left(\Psi_0\right)\right)$ herleiten:

$$R_{1n}\left(\Psi^{(1)}(\boldsymbol{\beta})\right) \leq 2\boldsymbol{t}^{(k)'} \sum_{i=1}^{n} \boldsymbol{b}_i - n\boldsymbol{t}^{(k)'} \boldsymbol{\Sigma} \boldsymbol{t}^{(k)} \left(1 + o_p(1)\right) + o_p(1). \tag{3.1}$$

Diese Schranke ist selber durch $O_P(1)$ beschränkt und somit kann $\boldsymbol{t}^{(k)}$ höchstens von der Ordnung $O_P(n^{-1/2})$ sein, da wir sonst einen Widerspruch erhalten würden. Aus

dieser Tatsache können wir direkt die Aussage des Satzes für $k = 1$ folgern. Die Verallgemeinerung für bel. k, d.h. dass der EM Schritt nichts an der Ordnung bzw. Konsistenz verändert, zeigen Chen und Li in einem weiteren Lemma, auf welches wir an dieser Stelle nicht näher eingehen. Hier beobachtet man dann auch zusätzlich, dass $\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta} = O_P(n^{-1/6})$, was für $k = 1$ klar ist, denn wir starten die Iteration mit $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}$.

□

Neben der Aussage des vorherigen Satzes spielt die dort gezeigte obere Schranke eine entscheidende Rolle für den nun folgenden Satz über die asymptotische Verteilung der Teststatistik. Wie beim MLRT teilen wir die Kovarianzmatrix $\boldsymbol{\Sigma}$ in Blockmatrizen auf, so dass

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right), \ \boldsymbol{\Sigma}_{jk} = E\left(\boldsymbol{b}_{j1}\boldsymbol{b}_{k1}'\right) \text{ für } j,k = 1,2.$$

Wir definieren ebenfalls wie beim MLRT $\widetilde{\boldsymbol{b}}_{2i} := \boldsymbol{b}_{2i} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{b}_{1i}$. Dieser Zufallsvektor besitzt die Kovarianzmatrix $\widetilde{\boldsymbol{\Sigma}}_{22} := \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$.

Satz 3.2. *Unter den getroffenen Annahmen und der Forderung $\frac{1}{2} \in B$ gilt*

$$EM_n^{(K)} \overset{(V.)}{\to} \sup_{\boldsymbol{v}\in\mathbb{R}_{\geq}^{m_0}} \left(2\boldsymbol{v}'\boldsymbol{w} - \boldsymbol{v}'\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{v}\right) \overset{(V.)}{=} \sum_{h=0}^{m_0} a_h \chi_h^2,$$

wobei $\boldsymbol{w}$ multivariat normalverteilt ist mit Erwartungswert Null und Kovarianzmatrix $\widetilde{\boldsymbol{\Sigma}}_{22}$, die Gewichte a_h nicht negativ sind, sich zu Eins aufsummieren und abhängig von der Verteilung von $\boldsymbol{w}$ sind, χ_h^2 für $h = 1, \ldots, m_0$ eine χ^2 Verteilung mit h Freiheitsgraden beschreibt und χ_0^2 die Punktverteilung in Null darstellt.

Beweisidee. Der Beweis funktioniert sehr ähnlich wie beim MLRT, zunächst zeigt man mit Hilfe von (3.1), dass

$$\sup_{\boldsymbol{v}\in\mathbb{R}_{\geq}^{m_0}} \left(2\boldsymbol{v}' \sum_{i=1}^{n} \widetilde{\boldsymbol{b}}_{2i} - n\boldsymbol{v}'\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{v}\right) + o_P(1) \tag{3.2}$$

eine obere Schranke für $EM_n^{(K)}$ darstellt. Anschließend folgert man wieder, dass diese asymptotisch angenommen wird, d.h.

$$\left| EM_n^{(K)} - \sup_{\boldsymbol{v}\in\mathbb{R}_{\geq}^{m_0}} \left(2\boldsymbol{v}' \sum_{i=1}^{n} \widetilde{\boldsymbol{b}}_{2i} - n\boldsymbol{v}'\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{v}\right) \right| = o_P(1).$$

Hierfür benötigen wir unter anderem, dass $\frac{1}{2} \in B$ ist, denn somit wissen wir sicher, dass bereits ohne EM Iteration der Penalty asymptotisch vernachlässigbar ist. Ebenfalls wie beim MLRT konvergiert die angenommene obere Schranke in Verteilung gegen $\sup_{\boldsymbol{v}\in\mathbb{R}^{m_0}_{\geq}} \left(2\boldsymbol{v}'\boldsymbol{w} - \boldsymbol{v}'\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{v}\right)$ und wir erhalten mit dem Satz von Slutzky die Verteilungskonvergenz des Satzes. Die Gleichheit in Verteilung mit der χ^2 Mischung ist im Vergleich zum MLRT eine Verallgemeinerung und wird in [CL09] ausführlich bewiesen.

□

3.1.4 Praktische Umsetzung des Tests

Wie in Satz 3.2 erwähnt, hängen die Gewichte der χ^2 Mischung von der Kovarianzmatrix $\widetilde{\boldsymbol{\Sigma}}_{22}$ ab. Diese Matrix müssen wir also berechnen bzw. bei unbekannten Parametern schätzen. In Abschnitt 2.1.5 erklären wir dies für den MLRT. An diesem Vorgehen ändert sich fast nichts, bis auf dass wir $\boldsymbol{b}$ hier entsprechend obiger Definition berechnen bzw. schätzen müssen. Bei bekannten wahren Parametern können wir natürlich die Kovarianzen einfach über die jeweiligen Integrale (numerisch) berechnen.
Beim EM-Test gestaltet sich nun die Berechnung der Gewichte a_h in Abhängigkeit von $\widetilde{\boldsymbol{\Sigma}}_{22}$ i.A. anders als beim MLRT. Es gilt der Zusammenhang

$$a_h = P\left(\sum_{j=1}^{m_0} I\left(\hat{v}_j > 0\right) = h\right), \text{ mit } \hat{\boldsymbol{v}} = \underset{\boldsymbol{v}\in\mathbb{R}^{m_0}_{\geq}}{\operatorname{argsup}} \left(2\boldsymbol{v}'\boldsymbol{w} - \boldsymbol{v}'\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{v}\right). \tag{3.3}$$

Für $m_0 = 2$ können wir nun genau wie beim MLRT die Gewichte a_h berechnen. Diese ergeben sich über den Korrelationskoefizienten von $\widetilde{\boldsymbol{\Sigma}}_{22}$. Für $m_0 = 3$ leiten Chen und Li ebenfalls eine Berechnungsvorschrift her. Es gilt $a_0 + a_2 = a_1 + a_3 = 0.5$ und

$$a_0 = \left(2\pi - \arccos\omega_{12} - \arccos\omega_{13} - \arccos\omega_{23}\right)/4\pi,$$

wobei $(\omega_{ij})_{i,j=1,\ldots,m_0}$ die zu $\widetilde{\boldsymbol{\Sigma}}_{22}$ gehörige Korrelationsmatrix ist, sowie

$$a_1 = \left(3\pi - \arccos\omega_{12:3} - \arccos\omega_{13:2} - \arccos\omega_{23:1}\right)/4\pi,$$

mit

$$\omega_{ij:k} = \frac{\omega_{ij} - \omega_{ik}\omega jk}{\sqrt{(1-\omega_{ik}^2)(1-\omega_{jk}^2)}}.$$

Möchten wir für ein $m_0 > 3$ testen, so liegen keine direkten Berechnungsformeln für die Gewichte vor. Wir können aber in einer Simulation bei berechnetem bzw. geschätztem

$\widetilde{\boldsymbol{\Sigma}}_{22}$ die Werte a_h aufgrund von Gleichung (3.3) schätzen. Dazu simulieren wir N mal den Zufallsvektor W_j aus einer $\mathcal{N}(0, \widetilde{\boldsymbol{\Sigma}}_{22})$ Verteilung, berechnen den Wert

$$\hat{\boldsymbol{V}}_j = \underset{\boldsymbol{v}\in\mathbb{R}^{m_0}_{\geq}}{\operatorname{argsup}} \left(2\boldsymbol{v}'W_j - \boldsymbol{v}'\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{v}\right)$$

und schätzen hiermit die benötigten Wahrscheinlichkeiten.

3.1.5 Modifikation der Teststatistik

Wie wir bisher gesehen haben, besitzt die Teststatistik die obere Schranke aus dem Beweis zu Satz 3.2 und nimmt diese asymptotisch an. Für eine Anwendung wünschen wir uns natürlich, dass die Teststatistik in Verteilung diese Schranke möglichst schnell annimmt, um nicht für kleine Stichprobengrößen deutlich konservativ zu testen. Betrachten wir die Teststatistik ohne EM Iteration

$$\begin{aligned} EM_n^{(1)} &= \max_{\boldsymbol{\beta}\in B^{m_0}} \left\{M_n^{(1)}(\boldsymbol{\beta})\right\} = \max_{\boldsymbol{\beta}\in B^{m_0}} \left\{2\left(pl_n\left(\Psi^{(K+1)}(\boldsymbol{\beta})\right) - l_n\left(\widehat{\Psi}_0\right)\right)\right\} \\ &= \max_{\boldsymbol{\beta}\in B^{m_0}} \left\{2\left(l_n\left(\Psi^{(1)}(\boldsymbol{\beta})\right) + \boldsymbol{p}(\boldsymbol{\beta}) - l_n\left(\widehat{\Psi}_0\right)\right)\right\}, \end{aligned}$$

so sehen wir, dass diese durch den nicht positiven Penaltyterm eher verkleinert wird (außer bei $\beta = (0.5, \ldots, 0.5)$, denn hier ist der Penalty Null). Führen wir den Test durch, so wählt die Optimierung meistens den Schätzer mit $\beta = (0.5, \ldots, 0.5)$, da hier der Penalty gleich Null ist und die Verbesserung des Maximums durch eine andere Wahl von β meist geringer ausfällt, als der Vorteil den Penalty gleich Null zu wählen. Somit verkleinern wir den Alternativraum deutlich, was den Test konservativer macht. Wir könnten diesem Effekt entgegensteuern, indem wir den Parameter der Penaltyfunktion anpassen bzw. sehr klein wählen.
Wir sehen zusätzlich eine andere Möglichkeit, um die Testlevels für kleinere Stichproben unter der Nullhypothese besser einzuhalten. Analog zum zweiten Schritt schätzen wir im eingeschränkten Alternativraum $\Omega_{2m_0}(\boldsymbol{\beta})$ für $\boldsymbol{\beta} \in B^{m_0} = \{0.1, 0.3, 0.5\}^{m_0}$ den Maximum Likelihoodschätzer ohne Penaltyfunktion. Wir berechnen also

$$\Psi_{wP}^{(1)}(\boldsymbol{\beta}) := \underset{\Psi\in\Omega_{2m_0}(\boldsymbol{\beta})}{\operatorname{argmax}}\, l_n(\Psi),$$

wobei wP für without penalty steht. Das Resultat der Konsistenz der Parameter lässt sich natürlich auch auf diesen Schätzer übertragen, da wir die Null aus B ausschließen und so der Fall der Konvergenz eines Gewichtes gegen Null nicht möglich ist. Ferner ist das wesentliche Argument für die Gültigkeit der oberen Schranke von $M_n^{(1)}(\boldsymbol{\beta}) =$

$2\left(pl_n\left(\Psi^{(1)}(\boldsymbol{\beta})\right) - l_n\left(\widehat{\Psi}_0\right)\right)$ die Konsistenz von $\Psi^{(1)}(\boldsymbol{\beta})$ und somit erhalten wir auch, dass

$$M^{(1)}_{n,wP}(\boldsymbol{\beta}) := 2\left(l_n\left(\Psi^{(1)}_{wP}(\boldsymbol{\beta})\right) - l_n\left(\widehat{\Psi}_0\right)\right)$$

durch (3.2) nach oben beschränkt wird. Da nun

$$pl_n\left(\Psi^{(1)}(\boldsymbol{\beta})\right) = l_n\left(\Psi^{(1)}(\boldsymbol{\beta})\right) + \boldsymbol{p}(\boldsymbol{\beta}) \leq l_n\left(\Psi^{(1)}(\boldsymbol{\beta})\right) \leq l_n\left(\Psi^{(1)}_{wP}(\boldsymbol{\beta})\right),$$

ist $M^{(1)}_n(\boldsymbol{\beta}) \leq M^{(1)}_{n,wP}(\boldsymbol{\beta})$ und somit gilt auch

$$EM^{(1)}_n = \max_{\boldsymbol{\beta}\in B^{m_0}}\left\{M^{(1)}_n(\boldsymbol{\beta})\right\} \leq \max_{\boldsymbol{\beta}\in B^{m_0}}\left\{M^{(1)}_{n,wP}(\boldsymbol{\beta})\right\} := EM^{(1)}_{n,wP}. \tag{3.4}$$

$M^{(1)}_{n,wP}(\boldsymbol{\beta})$ ist für ein beliebiges $\beta \in B^{m_0}$ durch (3.2) nach oben beschränkt und daher $EM^{(1)}_{n,wP}$ natürlich auch. Wie wir bereits gesehen haben, nimmt $EM^{(1)}_n$ (3.2) asymptotisch an und deshalb muss dies wegen (3.4) auch für $EM^{(1)}_{n,wP}$ gelten.
Zusammenfassend erhalten wir also mit $EM^{(1)}_{n,wP}$ eine Teststatistik, die größer gleich der ursprünglichen Teststatistik ohne EM Iteration ist und gleichzeitig die Asymptotik erfüllt. In vielen Fällen verändert die EM Iteration sehr wenig an den Parametern, so dass wir uns durch Betrachten von $EM^{(1)}_{n,wP}$ meistens auch hier besser stellen. Würden wir nun $EM^{(1)}_{n,wP}$ alleine benutzen, so würden wir mögliche Vorteile der EM Iteration (unter nicht erfüllter Nullhypothese) verlieren. Daher betrachten wir nun als Teststatistik

$$EM^{(mod)}_n := \max\left\{EM^{(1)}_{n,wP}, EM^{(K)}_n\right\}.$$

Hiermit vereinen wir die Vorteile der EM Iteration und verbessern die Anpassung der Testlevels für kleine Stichproben.

3.1.6 Simulationen bei zwei Komponenten

Generell testen wir wieder die Anzahl der Komponenten bei Mischungen aus Normalverteilungen mit festem Erwartungswert und variablen Standardabweichungen. Zunächst untersuchen wir das Verhalten der simulierten Testlevels bei zwei und anschließend bei drei Komponenten. Ferner betrachten wir hier nur $EM^{(1)}_{n,wP}$, da wir uns unter der Nullhypothese keine Vorteile von der EM Iteration erhoffen. Wir simulieren stets die Stichprobengrößen $250, 500$ und 1000. Zur Berechnung der Testlevels führen wir jeweils 10000 Wiederholungen durch.
Um neben der isolierten Analyse des Verhaltens die Ergebnisse mit denen des MLRT vergleichen zu können, wählen wir hier die gleichen Szenarien aus. Diese werden in Tabelle 3.1 noch einmal festgehalten.

	σ_1	σ_2	π_0
N_1	1	4	0.67
N_2	2	6	0.50
N_3	1	6	0.57
N_4	2	8	0.55

Tabelle 3.1: Parameter der iid normalverteilten Zweikomponentenmischungen

Die Tabellen 3.2 bis 3.4 zeigen die Ergebnisse. Wir können unsere Beobachtungen wie folgt zusammenfassen: Für steigende Stichprobengröße beobachten wir für alle Szenarien außer N_3 einzeln Konvergenz gegen die theoretischen asymptotischen Testlevels. Im dritten Szenario testen wir wie im MLRT für $n = 500$ und $n = 1000$ etwas anti konservativ. Insgesamt können wir für Stichprobengrößen ab 250 bis 500, je nach Szenario im hier betrachteten Fall $m_0 = 2$, gute Ergebnisse erzielen.
Vergleichen wir unsere Simulationsergebnisse des EM-Tests mit denen des MLRT, so sehen wir bei $n = 250$ und $n = 500$ keine strukturellen Unterschiede. Im Fall $n = 1000$ testen wir beim MLRT etwas konservativer und somit näher an den vorgegebenen Levels. Dabei ist zu beachten, dass wir bei der Simulation des EM-Tests keine Penalisierung der Gewichte vorgenommen haben. Der Grund dafür ist, dass wir beim EM-Test mit Penalisierung der Gewichte deutlich zu konservative Ergebnisse erhielten und eine Penalisierung hier nicht nötig ist, um die Gewichte von Null wegzubeschränken. Falls wir allerdings EM-Iterationen durchführen wollten, so müssten wir die penalisierte Likelihoodfunktion direkt unter der Alternative benutzen, da in den EM Iterationen diese weiter verbessert wird.

Level	0.1	0.05	0.025
N_1	0.087	0.042	0.019
N_2	0.068	0.032	0.014
N_3	0.101	0.050	0.026
N_4	0.081	0.040	0.021

Tabelle 3.2: Simulierte Testlevels bei $n = 250$ (EM-Test, zwei Komponenten, iid)

Level	0.1	0.05	0.025
N_1	0.092	0.047	0.024
N_2	0.080	0.040	0.020
N_3	0.108	0.058	0.030
N_4	0.096	0.044	0.022

Tabelle 3.3: Simulierte Testlevels bei $n = 500$ (EM-Test, zwei Komponenten, iid)

Level	0.1	0.05	0.025
N_1	0.108	0.054	0.025
N_2	0.088	0.042	0.021
N_3	0.118	0.065	0.034
N_4	0.104	0.053	0.027

Tabelle 3.4: Simulierte Testlevels bei $n = 1000$ (EM-Test, zwei Komponenten, iid)

3.1.7 Simulationen bei drei Komponenten

Wir betrachten nun den EM-Test bei Dreikomponentenmischungen. Die support points der hier betrachteten vier Szenarien sind in Tabelle 3.5 aufgelistet. Für die Gewichte

	σ_1	σ_2	σ_3
N_5	1	5	20
N_6	1	10	30
N_7	0.5	3	9
N_8	0.7	4	12

Tabelle 3.5: Support points der normalverteilten Driekomponentenmischungen

legen wir für N_5, N_6 $(0.37, 0.25, 0.38)$ und für N_7, N_8 $(0.47, 0.29, 0.24)$ zu Grunde. Die Ergebnisse finden sich in den Tabellen 3.6 bis 3.8.
Wir sehen, dass wir bei $n = 250$ in allen vier Szenarien deutlich konservativ testen. Anscheinend ist die Stichprobengröße von 250 bei drei Komponenten nicht groß genug, um mit der asymptotischen Verteilung zu testen. Bei $n = 500$ sehen die Ergebnisse bereits deutlich besser aus. Wir testen zwar auch hier noch konservativ, jedoch wird die Abweichung geringer. Schließlich bei $n = 1000$ haben sich die Levels weiter denen der asymptotischen Verteilung angenähert, so dass wir insgesamt von Konvergenz der betrachteten Quantile ausgehen können.

Level	0.1	0.05	0.025
N_5	0.059	0.026	0.012
N_6	0.074	0.033	0.014
N_7	0.052	0.023	0.010
N_8	0.058	0.024	0.011

Tabelle 3.6: Simulierte Testlevels bei $n = 250$ (EM-Test, drei Komponenten, iid)

Level	0.1	0.05	0.025
N_5	0.080	0.035	0.015
N_6	0.097	0.046	0.020
N_7	0.085	0.039	0.018
N_8	0.080	0.038	0.016

Tabelle 3.7: Simulierte Testlevels bei $n = 500$ (EM-Test, drei Komponenten, iid)

Level	0.1	0.05	0.025
N_5	..	..	..
N_6	..	..	..
N_7	..	..	..
N_8	..	..	..

Tabelle 3.8: Simulierte Testlevels bei $n = 1000$ (EM-Test, drei Komponenten, iid)

3.2 Erweiterung für Hidden Markov Modelle

Wir möchten nun den EM-Test auf Hidden Markov Modelle übertragen. Dazu gehen wir davon aus, dass $X_1, \ldots, X_n$ die beobachtbaren Variablen eines HMM mit ergodischer Markov Kette sind. Unter der Nullhypothese nehmen wir weiter an, dass die Markov Kette m_0 Zustände besitzt und somit die Marginalverteilung der beobachtbaren Variablen einem Mischungsmodell mit m_0 Komponenten folgt.
Als Komponentendichten betrachten wir nur solche, die den Annahmen des EM-Tests genügen. Möglich wären z.B. Poisson-, Binomial- (mit fester Versuchsanzahl) oder Normalverteilungen (mit entweder festem Erwartungswert oder fester Standardabweichung). Konkret betrachten wir unter der Nullhypothese von zwei und drei Komponenten Normalverteilungen mit festem Erwartungswert.

3.2.1 Simulationen bei zwei Komponenten

Um eine Vergleichbarkeit zum iid Fall und MLRT zu gewährleisten, simulieren wir gemäß derselben Szenarien wie im vorherigen Abschnitt (vgl. Tabelle 3.1). Für die Übergangsmatrizen wählen wir Γ_1 bis Γ_4, die bei den Simulationen vom MLRT (vgl. Abschnitt 2.2.4) definiert wurden. Die Ergebnisse finden sich in den Tabellen 3.9 bis 3.11.
Wie schon beim EM-Test im iid-Fall, können wir auch hier bereits ab 250 bis 500 gute Ergebnisse erzielen und in allen Szenarien außer N_3 konvergieren die Levels. Es ist nicht

Level	0.1	0.05	0.025
N_1	0.083	0.041	0.020
N_2	0.070	0.035	0.017
N_3	0.104	0.056	0.029
N_4	0.086	0.040	0.020

Tabelle 3.9: Simulierte Testlevels bei $n = 250$ (EM-Test, zwei Komponenten, HMM)

Level	0.1	0.05	0.025
N_1	0.097	0.050	0.024
N_2	0.074	0.038	0.020
N_3	0.110	0.055	0.029
N_4	0.102	0.051	0.024

Tabelle 3.10: Simulierte Testlevels bei $n = 500$ (EM-Test, zwei Komponenten, HMM)

Level	0.1	0.05	0.025
N_1	0.103	0.053	0.026
N_2	0.085	0.041	0.020
N_3	0.120	0.065	0.034
N_4	0.102	0.051	0.024

Tabelle 3.11: Simulierte Testlevels bei $n = 1000$ (EM-Test, zwei Komponenten, HMM)

verwunderlich, dass bei N_3 für $n = 1000$ noch keine deutliche Übereinstimmung zwischen den Levels vorliegt, da wir dies im iid-Fall beim EM-Test und beim MLRT (sowohl im iid- als auch beim HMM-Fall) ebenfalls beobachteten.
Vergleichen wir unsere Simulation des EM-Tests und des MLRTs im HMM-Fall bei $m_0 = 2$, so beobachten wir keine Unterschiede. Die Simulation spricht also dafür, dass der EM-Test bei zwei Komponenten auch im HMM-Fall funktioniert.

3.2.2 Simulationen bei drei Komponenten

Wir betrachten nun wie sich der EM-Test beim Zugrundelegen eines HMM im Fall von drei Komponenten verhält. Die support points der hier betrachteten vier Szenarien sind dieselben wie im iid-Fall und in Tabelle 3.5 aufgelistet.
Als Übergangsmatrizen benutzen wir für N_5, N_6 die Matrix Γ_5 und bei N_7, N_8 wählen wir Γ_6, wobei

$$\Gamma_5 := \begin{pmatrix} 0.9 & 0.05 & 0.05 \\ 0 & 0.85 & 0.15 \\ 0.1 & 0.05 & 0.85 \end{pmatrix} \text{ und } \Gamma_6 := \begin{pmatrix} 0.80 & 0.10 & 0.10 \\ 0.2 & 0.8 & 0 \\ 0.15 & 0.05 & 0.8 \end{pmatrix}.$$

Die Tabellen 3.12 bis 3.14 zeigen die Ergebnisse unserer Simulation.
Auch hier erhalten wir mit steigender Stichprobengröße eine immer besser werdende Übereinstimmung zwischen simulierten und vorgegebenen Levels. Vergleichen wir die Ergebnisse mit dem EM-Test im iid-Fall, so sehen wir keine Unterschiede. Unsere Simulationen sprechen also dafür, dass sich der EM-Test auch für $m_0 = 3$ anwenden lässt.
Insgesamt deuten unsere Simulationen darauf hin, dass sich der EM-Test wie der MLRT ebenfalls auf die von uns betrachteten Hidden Markov Modelle anwenden lässt. Im nächsten Abschnitt betrachten wir als Anwendung des HMM-Falls Renditen der Siemens AG.

Level	0.1	0.05	0.025
N_5	0.062	0.026	0.011
N_6	0.064	0.031	0.013
N_7	0.061	0.025	0.011
N_8	0.058	0.026	0.013

Tabelle 3.12: Simulierte Testlevels bei $n = 250$ (EM-Test, drei Komponenten, HMM)

Level	0.1	0.05	0.025
N_5	0.080	0.035	0.016
N_5	0.095	0.043	0.020
N_5	0.081	0.037	0.018
N_5	0.077	0.036	0.017

Tabelle 3.13: Simulierte Testlevels bei $n = 500$ (EM-Test, drei Komponenten, HMM)

Level	0.1	0.05	0.025
N_5	0.108	0.051	0.024
N_5	0.110	0.052	0.024
N_5	0.096	0.046	0.024
N_5	0.095	0.044	0.021

Tabelle 3.14: Simulierte Testlevels bei $n = 1000$ (EM-Test, drei Komponenten, HMM)

3.2.3 Anwendung des EM-Tests auf Siemensrenditen

In diesem Abschnitt betrachten wir die Tagesrenditen der Siemens AG der letzten vier Jahre in Prozent. Abbildung 3.1 veranschaulicht diesen Datensatz.

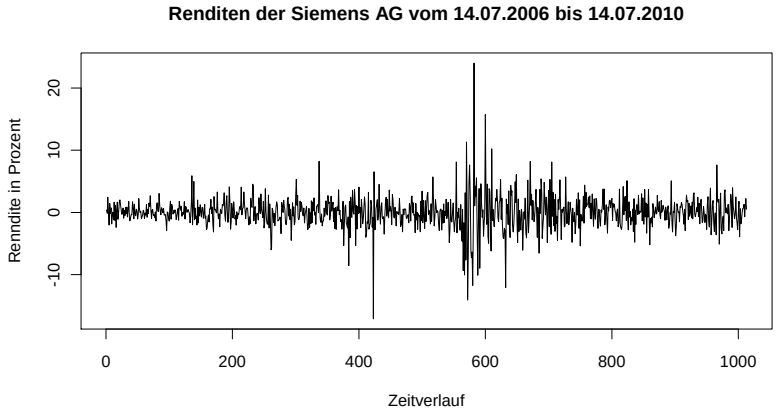
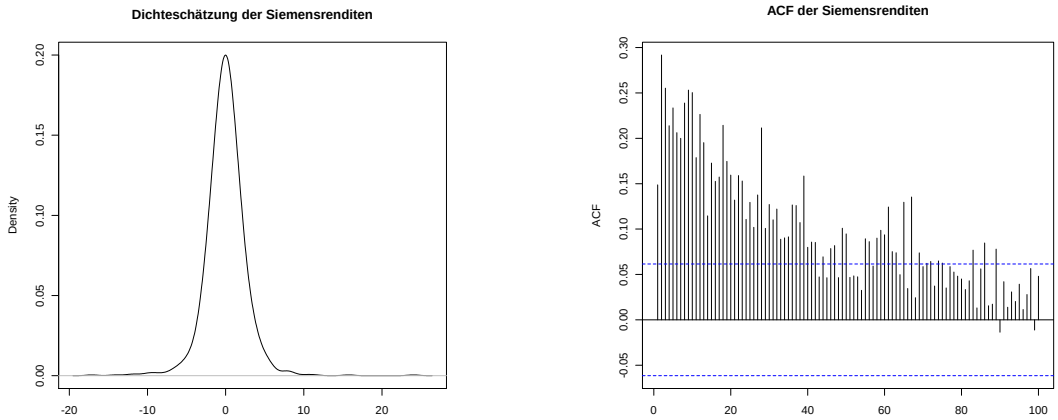


Abbildung 3.1: Renditen der Siemens AG

Zu Beginn standardisieren wir die Daten durch Subtraktion des Erwartungswertes. Abbildung 3.2 zeigt die empirisch geschätzte Dichte der Daten und die empirische Autokorrelationsfunktion (ACF) vom Betrag der Renditen. Da wir eine unimodale Dichte beobachten, liegt es nahe sie über eine Mischung aus Normalverteilungen mit festem Erwartungswert und variabler Standardabweichung zu modellieren.



(a) Dichteschätzung

(b) empirische ACF

Abbildung 3.2: Dichteschätzung und empirische ACF der Siemensrenditen

Stellt man sich die Frage, ob einzelne Renditen unabhängig voneinander sind, so kann

man die ACF vom Betrag der Aktienrenditen betrachten, denn wären sie unabhängig voneinander, so dürften wir hier keine Autokorrelationen beobachten. In Abbildung 3.2 (b) können wir allerdings starke Autokorrelationen erkennen und somit kann keine Unabhängigkeit vorliegen. Eine direkte Anwendung des EM-Tests wäre also nicht sinnvoll. Legen wir nun ein Hidden Markov Modell zu Grunde, so können wir eine Abhängigkeit des Umweltzustandes, unter der die Rendite entstanden ist, modellieren.
Wir führen zuerst den EM-Test für $m_0 = 1$ durch, d.h wir testen, ob die Daten bereits durch eine Normalverteilung mit Erwartungswert Null und bel. Standardabweichung beschrieben werden können. Unter der Hypothese schätzen wir hier eine Standardabweichung von 2.59. Im Gegensatz dazu ergibt sich unter der Alternative eine Mischung mit den Parametern $\sigma_1 = 1.77, \sigma_2 = 6.25$ und $p = 0.9$. Der Wert der Teststatistik beträgt in diesem Fall 285.82, was einen P-Wert von ca. Null zur Folge hat. Wir können also eindeutig $H_0 : m_0 = 1$ verwerfen und bestätigen die visuelle Intuition, dass die vorliegende Dichte zu spitz für eine einfache Normalverteilung ist.
Als nächstes betrachten wir die Nullhypothese von zwei Komponenten. Unter der Hypothese schätzen wir hier ein Mischungsmodell mit den Parametern $\sigma_1 = 1.77, \sigma_2 = 6.30$ und $p = 0.902$. Die Tatsache, dass die Schätzung unter der Alternative im ersten Test und die Schätzung unter der Hypothese im zweiten Test (beides mal passen wir eine Zweikomponentenmischung an) nicht genau übereinstimmen, ist kein numerischer Fehler. Es liegt daran, dass wir im ersten Test nur die Werte $0.1, 0.3$ und 0.5 für das Gewicht des kleineren support points zulassen und im zweiten Test das Gewicht unter der Hypothese komplett mit anpassen. Die Alternative im zweiten Test ist eine Vierkomponentenmischung. Wir schätzen hier die support points $\sigma_1 = 1.32, \sigma_2 = 2.33, \sigma_3 = 4.90$ und $\sigma_4 = 9.42$. Für die Gewichte ergibt sich nach Berücksichtigung, dass die optimale Aufteilung durch $\boldsymbol{\beta} = (0.5, 0.3)$ gegeben ist: $p_1 = 0.46, p_2 = 0.46, p_3 = 0.06$ und $p_4 = 0.02$. Die Teststatistik nimmt ferner den Wert 12.27 an. Um hierfür einen P-Wert angeben zu können, schätzen wir zunächst die Gewichte der χ^2-Mischung der asymptotischen Verteilung der Teststatistik. Diese ergeben sich mittels den geschätzten Parametern unter der Hypothese zu $(0.32, 0.5, 0.18)$. Wir können nun den P-Wert als die Wahrscheinlichkeit dieser Verteilung für eine größeren Wert als 12.27 berechnen und erhalten als Ergebnis $6.28 \cdot 10^{-4}$. Somit können wir auch die Hypothese von $H_0 : m_0 = 2$ bei einem Signifikanzniveau von 5% deutlich verwerfen.
Bis an diese Stelle hätten wir auch mit dem MLRT testen können. Um nun aber auch den Test für $H_0 : m_0 = 3$ durchführen zu können, benötigen wir zwingend den EM-Test. Unter der Hypothese schätzen wir hier eine Dreikomponentenmischung mit Parametern

$\sigma_1 = 1.43, \sigma_2 = 2.74$ und $\sigma_3 = 8.28$ sowie Gewichten $p_1 = 0.61, p_2 = 0.35$ und $p_3 = 0.04$. Die Schätzung unter der Alternative ist gegeben durch die support points $\boldsymbol{\sigma} = (0.53, 1.50, 1.83, 2.62, 4.32, 8.62)$ und Gewichte $\boldsymbol{p} = (0.05, 0.46, 0.13, 0.29, 0.035, 0.035)$. Der Wert der Teststatistik lautet 0.8717 und liefert (nach Berechnung der Gewichte der χ^2-Mischung gemäß der Schätzung unter der Hypothese) einen P-Wert von 0.28. Wir können also bei einem 5%-Signifikanzniveau die Hypothese von drei Komponenten nicht verwerfen und behalten sie daher bei.

Als letzten Schritt möchten wir nun noch das angepasste Dreikomponenten Modell interpretieren. Den größten Anteil machen die ersten beiden Zustände mit einer Gesamtwahrscheinlichkeit von 0.96 (einzelne Wahrscheinlichkeiten sind 0.61 und 0.35) und den Tagesstandardabweichungen von 1.43% und 2.74% aus. Den wesentlichen Anteil machen also die beiden Zustände mit niedriger (1.43%) und höherer (2.74%) Varianz aus. Eine gesonderte Rolle hat der dritte Zustand, der mit nur 4% Wahrscheinlichkeit eintritt und mit 8.28% eine ungleich größere Standardabweichung aufweist. Es liegt die Überlegung nahe, dass dieser Zustand extrem hoher Schwankung gerade die Finanzkrise im vierten Quartal 2008 widerspiegelt. Um dieser Vermutung nachzugehen, schätzen wir im HMM Modell den MLE, d.h. wir schätzen hier mit Hilfe der HMM Log-Likelihoodfunktion. Für diese Schätzungen benutzen wir das R-Paket *SwitchingVolatility*. Dieses Paket wurde am Lehrstuhl von Prof. Holzmann mit Hilfe von Funktionen aus [ZM09] entwickelt. Die Schätzung für die Übergangsmatrix lautet:

$$\widehat{\Gamma} = \begin{pmatrix} 0.97 & 0.03 & 0 \\ 0.04 & 0.95 & 0.01 \\ 0 & 0.09 & 0.91 \end{pmatrix}.$$

Wir können nun mit Hilfe des Viterbi Algorithmus schätzen welcher Zustand jeweils am wahrscheinlichsten ist. Abbildung 3.3 bestätigt, dass die Renditen der Siemensaktien in der Zeit der Finanzkrise (in diesem Datensatz läuft das vierte Quartal 2008 von $i = 550, \ldots, 610$) im Vergleich zum restlichen betrachteten Intervall eine wesentlich höhere Standardabweichung aufweisen. Außerdem sehen wir, dass (laut Schätzung) der dritte Zustand (bis auf zwei Ausnahmen) nur in der Finanzkrise aufgetreten ist. Wir können diesem Zustand somit als die Finanzkrise interpretieren. Ein weiterer interessanter Aspekt ist die Tatsache, dass sich die Renditen der Siemensaktie nach der Finanzkrise ca. ein Jahr im Zustand hoher Varianz befanden.

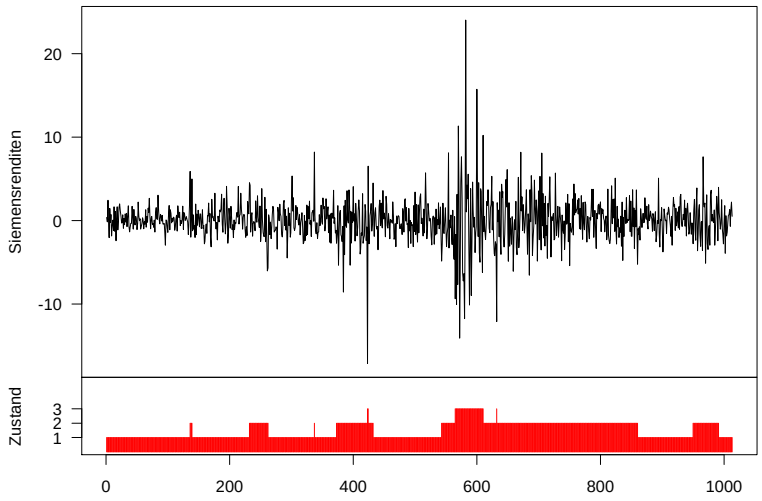


Abbildung 3.3: Renditen der Siemens AG und Schätzung der Zustände des HMM

Ausblick

Mit dem MLRT und dem EM-Test haben wir in dieser Arbeit zwei Verfahren untersucht, die für unabhängig identisch verteilte Stichproben die Anzahl der Komponenten einer Mischung testen. Weiterhin haben wir in Kapitel 2.2 eine Erweiterung des MLRT betrachtet, bei der die Stichproben die beobachtbaren Variablen eines stationären Hidden Markov Modells sind. Für diese Testsituation wurde in [DH08] nachgewiesen, dass unter der Nullhypothese die asymptotische Verteilung der Teststatistik mit der des MLRT bei unabhängigen Stichproben übereinstimmt.
Für den EM-Test wurde eine solche Verallgemeinerung bisher nicht durchgeführt. Wir haben daher für diese Arbeit den EM-Test mit normalverteilten Komponenten (fester Erwartungswert) implementiert und Simulationen durchgeführt, bei denen wir als Stichproben ebenfalls die beobachtbaren Variablen eines Hidden Markov Modells benutzten. Die Ergebnisse deuten daraufhin, dass sich auch der EM-Test auf diese Abhängigkeitsstruktur erweitern lässt.
Einen Beweis dafür haben wir bisher nicht geführt, jedoch sollten wir die Konsistenzaussagen analog zum MLRT verallgemeinern können. Der nächste Schritt wäre nun eine entsprechenden Aussage wie in Gleichung (2.43) in Kapitel 2.1.4 zu folgern, so dass der eigentliche Beweis darin bestünde zu zeigen, dass $n^{-1/2} \sum_{i=1}^{n} \widetilde{\boldsymbol{b}}_{2i}$ asymptotisch normalverteilt ist mit Kovarianzmatrix $\mathrm{COV}(\widetilde{\boldsymbol{b}}_{21})$.
Bei der Verallgemeinerung des MLRT zeigten wir die Konvergenz in Verteilung gegen die Normalverteilung mit entsprechender Kovarianzmatrix über den Zentralen Grenzwertsatz für stationäre Martingaldifferenzenfolgen. Wir werden versuchen einen ähnlichen Beweis für den EM-Test zu führen, um so die vermutete Asymptotik der Teststatistik zu zeigen.
Wir führten die Simulationen mit einer modifizierten Teststatistik durch (unter der Alternative keine Penalisierung und kein EM Schritt, vgl. Kapitel 3.1.5). Die Begründung für diese Wahl ist einerseits, dass die Penalisierung den Test bei kleinen Stichproben zu konservativ macht und andererseits haben wir so keinen Parameter für den Penaltyterm, der anzupassen ist. Diese Aspekte sind im Beweis natürlich ebenfalls zu berücksichtigen

und sollten die Herleitung eher erleichtern.
Mit dieser Teststatistik würden natürlich jegliche Vorteile einer EM Iteration verloren gehen und so die Güte des Tests evtl. verschlechtern, da wir bei nicht gültiger Nullhypothese den Maximum Likelihoodschätzer gegebenenfalls nicht exakt genug anpassen können. Die Penalisierung wird grundsätzlich eingeführt, um bei der EM Iteration (Schätzung unter der Alternative) zu verhindern, dass einzelne Gewichte der $2m_0$-Komponentenmischung gegen Null konvergieren. Um diese Tatsache zu gewährleisten, könnten wir auch direkt über alle $2m_0$ Gewichte optimieren, anstatt erst über m_0 und dann über alle in der EM Iteration, wenn wir als Restriktion $\beta_h > \varepsilon$ mit $\varepsilon > 0$ vorgeben. Wir könnten so in einer Optimierung direkt den Maximumlikelihood Schätzer unter der Alternative berechnen und müssten nicht m_0^3 Optimierungen mit anschließender EM Iteration durchführen. Eine andere Möglichkeit wäre die in Kapitel 3.1.5 angesprochene Berechnung des Maximums aus der original Teststatistik mit EM Iteration und der ohne Penalisierung und ohne EM Iteration.

Literaturverzeichnis

[Bil61] Billingsley, P.: The Lindeberg-Lévy Theorem for Martingales. In: *Proceedings of the American Mathematical Society* 12 (1961), Nr. 5, S. 788 – 792.

[CCK01] Chen, H. ; Chen, J. ; Kalbfleisch, J. D.: A Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63 (2001), Nr. 1, S. 19–29.

[CCK04] Chen, H. ; Chen, J. ; Kalbfleisch, J. D.: Testing for a Finite Mixture Model with Two Components. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 66 (2004), Nr. 1, S. 95 – 115.

[CL09] Chen, H. ; Li, P.: *Testing the Order of a Finite Mixture.* Under preparation, 2009.

[DCG99] Dacunha-Castelle, D. ; Gassiat, E.: Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. In: *The Annals of Statistics* 27 (1999), Nr. 4, S. 19 – 40.

[DH08] Dannemann, J. ; Holzmann, H.: Testing for two states in a hidden Markov model. In: *The Canadian Journal of Statistics* 36 (2008), Nr. 4, S. 505 – 520.

[Fer96] Ferguson, T. S.: *A course in large sample theory.* 1. Auflage. London, Chapman & Hall, 1996

[FS06] Frühwirth-Schnatter, S.: *Finite Mixture and Markov Switching Models.* 1. Auflage. New York, Springer, 2006.

[Kle06] Klenke, A.: *Wahrscheinlichkeitstheorie.* 1. Auflage. Berlin Heidelberg, Springer, 2006.

[Ler92a] Leroux, B.G.: Consistent Estimation of a Mixing Distribution. In: *The Annals of Statistics* 20 (1992), Nr. 3, S. 1350 – 1360.

[Ler92b] Leroux, B.G.: Maximum-likelihood estimation for hidden Markov models. In: *Stochastic Processes and their Applications* 40 (1992), Nr. 1, S. 127 – 143.

[Lin78] Lindgren, G.: Markov Regime Models for Mixed Distributions and Switching Regressions. In: *Scandinavian Journal of Statistics* 5 (1978), Nr. 2, S. 81 – 91.

[MS05] Meintrup, D. ; Schäffler, S.: *Stochastik*. 1. Auflage. Berlin Heidelberg, Springer, 2005.

[ZM09] Zucchini, W. ; MacDonald, I.L.: *Hidden Markov Models for Time Series*. 2. Auflage. Boca Raton, Chapman & Hall/CRC, 2009.

Anhang

Eidesstattliche Erklärung

Ich versichere, dass ich diese Diplomarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Marburg, den 17. August 2010

_____

Florian Schwaiger