Analytische Eigenschaften von Mischungen elliptischer Verteilungen und deren Anwendung in der Clusteranalyse

Diplomarbeit
am Fachbereich Mathematik und Informatik
der Philipps-Universität Marburg

vorgelegt von
Grigory Alexandrovich

Januar 2011
Betreuer: Prof. Dr. H. Holzmann

Inhaltsverzeichnis

Einleitung

In vorliegender Arbeit untersuchen wir analytische Eigenschaften von Mischungen elliptischer Verteilungen und deren Anwendung in der Clusteranalyse.
In letzter Zeit ist das Aufkommen von Daten in allen Bereichen des menschlichen Handelns gestiegen - im Finanz- und Versicherungssektor werden große Datenmengen produziert, bei der Genexpressionsanalyse enstehen große Datensätze, Internet- und Telefonnutzungsdaten kummulieren sich in kürzester Zeit zu riesigen Datensätzen. Die große Herausforderung dabei ist nützliche Informationen aus dieser Datenflut zu extrahieren. Diese Informationen können Aussagen über die Art des datenerzeugenden Prozesses sein und gewisse Prognosen erlauben.

Die Clusteranalyse beschäftigt sich mit Identifizierung von homogenen Gruppen in Daten. Ein Ansatz, den wir in vorliegender Arbeit verfolgen, ist das probabilistische Clustering, wo Mischungen von Verteilungen zum Einsatz kommen.
Gauss-Mischungen bieten eine flexible Möglichkeit zweistufige Zufallsvorgänge zu modellieren. Die erste Stufe des Vorganges besteht aus der Auswahl einer Gruppe. Die zweite Stufe besteht aus der Generierung von Daten nach einer gruppenspezifischen Gesetzmäßigkeit. Die gruppenspezifischen Verteilungen sind im Gauss-Fall die Normalverteilungen mit gruppenspezifischen Erwartungswerten und Kovarianzmatrizen.
Dieser Ansatz wird erweitert, indem man als guppenspezifische Verteilungen Mischungen zulässt.

Trotz der hohen Fexibilität von Gauss-Mischungen, kann es in manchen Fällen besser sein andere Verteilungen zu nehmen, z.B. die t-Verteilung, die extremen Werten höhere Wahrscheinlichkeiten zuordnet als die Normalverteilung. In dieser Arbeit wird eine breite Klasse von Verteilungen behandelt - die elliptischen Verteilungen. Die Familie der elliptischen Verteilungen ist eine Verallgemeinerung der Normalverteilung und besteht aus Verteilungen, deren Dichten elliptische Höhenlinien haben.

Im 1. Kapitel definieren wir die Familie der elliptischen Verteilungen und betrachten einige ihrer Eigenschaften.
Im 2. Kapitel stellen wir die sog. **Ridgeline** vor und leiten eine Reihe nützlicher Ergebnisse mit ihrer Hilfe her. Das Kapitel besteht aus einer ausführlichen Ausarbeitung und Verallgemeinerung der Ergebnisse von Ray und Lindsay aus dem Paper „The topography of multivariate normal mixtures“, 2005 und einigen neuen Aussagen über die Anzahl von Moden in elliptischen Mischungen. Die Anzahl der Moden einer Mischung spielt insofern eine interessante Rolle, als dass wir Moden mit Subpopulationen/Clustern assoziieren und deren Anzahl gleich, kleiner oder auch größer als die der Mischungskomponenten sein kann.
Im 3. Kapitel diskutieren wir einen Clustering-Ansatz, der auf der Theorie aus dem 2. Kapitel und einem Vorschlag von C. Hennig aus dem Paper Methods for merging Gaussian mixture components“, 2009 für das Zusammenlegen von Komponeten aufbaut. Es wird eine Implementierung dieses Algorithmus diskutiert und es werden Ergebnisse von Simulationen vorgestellt.
Im 4. Kapitel konstruieren wir einen Test auf Unimodalität einer Gauss-Mischung aus zwei Komponenten, deren Kovarianzmatrizen gleich sind. Dabei greifen wir wieder auf die Theorie aus dem 2. Kapitel zurück.

Für Simulationen, Plots und Rechnungen wurde die Sprache R benutzt (http://cran.r-project.org/).

Kapitel 1

Elliptische Verteilungen

In diesem Kapitel diskutieren wir die Familie der elliptischen Verteilungen. Elliptische Verteilungen sind eine Verallgemeinerung der mehrdimensionalen Normalverteilung. In Folgenden Kapiteln betrachten wir Mischungen elliptischer Verteilungen.

1.1. Gruppen und Invarianz

In diesem Abschnitt geben wir einen sehr kurzen Einschub in das Thema Gruppen und Invarianz. Diese Begriffe brauchen wir später für einige Aussagen.

Definition 1.1.1. Sei $(G, \circ)$ eine nichtleere Menge mit binärer Verknüpfung $\circ$. Es gelte

1. $g_1, g_2 \in G \Rightarrow g_1 \circ g_2 \in G$
2. $(g_1 \circ g_2) \circ g_3 = g_1 \circ (g_2 \circ g_3)\ \forall g_1, g_2, g_3 \in G$
3. Für alle $g \in G$ existiert ein $g^{-1} \in G$, so dass gilt $g \circ g^{-1} = g^{-1} \circ g = e$, wobei e das neutrale Element ist: $g \circ e = e \circ g = g$.

Wir nennen $(G, \circ)$ eine *Gruppe*.

Im Folgenden interessieren wir uns für die Gruppe der orthogonalen Abbildungen auf $\mathbb{R}^D$, $\mathcal{O}(D)$.

Definition 1.1.2. Sei G eine Gruppe von Abbildungen von $\mathcal{X}$ nach $\mathcal{X}$. Zwei Elemente x_1, x_2 aus $\mathcal{X}$ heißen *äquivalent unter* G, falls ein $g \in G$ existiert mit $gx_1 = x_2$.
Wir schreiben:

$$x \sim y\,(mod\, G).$$

Offensichtlich hat die Äquivalenzrelation folgende Eigenschaften:

- $x \sim x \,(mod\, G)$
- $x \sim y \,(mod\, G) \Rightarrow y \sim x \,(mod\, G)$
- $x \sim y \,(mod\, G)$ und $y \sim z \,(mod\, G) \Rightarrow x \sim z \,(mod\, G)$.

Definition 1.1.3. Sei G eine Gruppe von Abbildungen von $\mathcal{X}$ nach $\mathcal{X}$. Eine Funktion f von $\mathcal{X}$ nach $\mathcal{X}$ heißt *invariant unter* G, falls

$$f(gx) = f(x) \;\; \forall x \in \mathcal{X}, g \in G.$$

Eine Funktion $f(x)$ von $\mathcal{X}$ nach $\mathcal{X}$ heißt *maximal invariant unter* G, falls f invariant unter G ist und falls $f(x_1) = f(x_2)$ impliziert $x_1 \sim x_2 \,(mod\, G)$.

Satz 1.1.4. *Sei $f : \mathcal{X} \to \mathcal{X}$ maximal invariant unter G. Eine Funktion $h : \mathcal{X} \to \mathcal{X}$ ist genau dann invariant unter G, falls h eine Funktion von f ist.*

Beweis:

„⇐“ Sei h eine Funktion von f, also $h(x) = j(f(x))$ für eine Funktion $j \Rightarrow h(gx) = j(f(gx)) = j(f(x)) = h(x)$ für alle $x \in \mathcal{X}$ und $g \in G$, also ist h invariant unter G.

„⇒“ Sei h invariant unter G, also $h(g(x)) = h(x)$ für alle $x \in \mathcal{X}$ und $g \in G$. Da f maximal invariant unter G ist, ist f injektiv auf der Menge der Repräsentanten der Äquivalenzrelation, d.h. es gibt eine Umkehrabbildung $f^{-1} : \mathcal{X} \to [\mathcal{X}]$, wobei $[\mathcal{X}]$ die Menge der Repräsentanten ist. Somit gilt $h(x) \equiv const \; \forall x \in f^{-1}(f(x_0))$ für alle $x_0 \in \mathcal{X}$, also hängt h nur durch f von x ab.

□

Beispiel 1.1.5. *Sei $G = \mathcal{O}(D)$, die Gruppe der orthogonalen Matrizen der Dimension D und $\mathcal{X} = \mathbb{R}^D$. Die Funktion $x \mapsto x^\mathsf{T} x$ ist maximal invariant unter G.*

Tatsächlich gilt für $x \in \mathbb{R}^D$: $(Ox)^\mathsf{T}(Ox) = x^\mathsf{T} O^\mathsf{T} O x = x^\mathsf{T} x$ für alle $O \in \mathcal{O}(D)$.
Auf der anderen Seite folgt aus $x^\mathsf{T} x = y^\mathsf{T} y$, daß x und y beide auf der Sphäre mit Radius $||x||$ liegen, also existiert ein $O \in \mathcal{O}(D)$ mit $x = Oy$.
Aus dem Satz oben folgt, dass jede G-invariante Funktion, die Form $f(x^\mathsf{T} x)$ hat.

1.2. Elliptische Verteilungen

Definition 1.2.1 (Sphärische Verteilungen)**.** Sei X ein Zufallsvektor der Dimension D. Wir sagen X habe eine *sphärische Verteilung* falls

$$OX \stackrel{d}{=} X \quad \forall O \in \mathcal{O}(D)$$

Satz 1.2.2. *Sei X ein $D \times 1$ Zufallsvektor. X ist genau dann sphärisch verteilt, wenn seine charakteristische Funktion $\psi(t)$ einer der folgenden äquivalenten Bedingungen genügt:*

1. $\psi(O^\mathsf{T} t) = \psi(t) \quad \forall O \in \mathcal{O}(D)$
2. *Es gibt eine Funktion $\phi : \mathbb{R} \to \mathbb{R}$, so dass $\psi(t) = \phi(t^\mathsf{T} t)$.*

Beweis:

1. Sei X sphärisch verteilt, dann $OX \stackrel{d}{=} X \quad \forall O \in \mathcal{O}(D)$. Die charakteristischen Funktionen der beiden Zufallsvektoren sind gleich. Also folgt

$$\mathbb{E}(e^{it^\mathsf{T} OX}) = \mathbb{E}(e^{i(O^\mathsf{T} t)^\mathsf{T} X}) = \psi(O^\mathsf{T} t) = \mathbb{E}(e^{it^\mathsf{T} X}) = \psi(t)$$

2. Die Funktion ψ ist invariant unter $\mathcal{O}(D)$, also ist sie eine Funktion der unter $\mathcal{O}(D)$ maximal invarianten Funktion $t^\mathsf{T} t$, d.h. $\psi(t) = \phi(t^\mathsf{T} t)$.

3. Sei $\psi(t)$ charakteristische Funktion von X und es gelte $\psi(t) = \psi(Ot)$ für alle orthogonale Matrizen O. Also gilt $X \stackrel{d}{=} OX \quad \forall O \in \mathcal{O}(D)$.

□

Satz 1.2.3. *Sei X ein $D \times 1$ Zufallsvektor. Folgende Aussagen sind äquivalent:*

1. $X \stackrel{d}{=} OX$*, für alle* $O \in \mathcal{O}(D)$
2. *Die charakteristische Funktion von X hat die Form $\phi(t^\mathsf{T} t)$ für eine skalare Funktion ϕ.*
3. *X hat die stochastische Darstellung $X \stackrel{d}{=} rU^{(D)}$, wobei $U^{(D)}$ ein auf der Einheitssphäre S_D gleichverteilter Zufallsvektor ist und $r \geq 0$ eine von $U^{(D)}$ unabhängige Zufallsvariable.*

4. Für alle $a \in \mathbb{R}^D$ gilt:

$$a^\mathsf{T} X \stackrel{d}{=} ||a|| X_1,$$

wobei X_1 die 1. Komponente von X ist.

Ein Beweis findet sich in [7].

Definition 1.2.4 (Elliptische Verteilungen)**.** *Ein $D \times 1$ Zufallsvektor X heißt elliptisch verteilt mit Parametern $\mu \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{D \times D}$, falls es ein sphärisch verteiltes Y gibt mit*

$$X \stackrel{d}{=} \mu + A^\mathsf{T} Y, \tag{1.1}$$

wobei $A^\mathsf{T} A = \Sigma$.

Folgende Eigenschaften können leicht überprüft werden:

1. Die charakteristische Funktion von X hat die Form $\psi(t) = e^{it^\mathsf{T}\mu}\phi(t^\mathsf{T}\Sigma t)$, für eine skalare Funktion ϕ.
2. X hat eine stochastische Darstellung $X \stackrel{d}{=} \mu + rA^\mathsf{T}U^{(D)}$, wobei $A^\mathsf{T}A = \Sigma$.

Bisher haben wir die Verteilungsfamilien definiert ohne Dichten zu erwähnen; sphärisch- bzw. elliptisch-verteilte Zufallsgrößen müßen auch gar keine Dichten besitzen. In vorliegender Arbeit betrachten wir jedoch nur Zufallsvektoren, die Dichten besitzen und führen unsere Argumentation hauptsächlich über die Dichten.

Satz 1.2.5 (Dichten sphärischer Verteilungen)**.** *Sei X ein sphärisch-verteilter Zufallsvektor und f seine Dichte. Dann gilt:*

$$f(x) = \varphi(x^\mathsf{T} x) \tag{1.2}$$

für eine skalare Funktion φ. Wir nennen φ (Dichte-)Generatorfunktion.

Beweis: Sei $\psi(t)$ die charakteristische Funktion von X. Der Satz 1.2.2 besagt, daß es eine skalare Funktion ϕ gibt mit $\phi(t^\mathsf{T}t) = \psi(t)$. Die Inversionsformel für die Dichte von X lautet:

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}^D} e^{-it^\mathsf{T}x}\phi(t)dt = \frac{1}{2\pi} \int_{\mathbb{R}^D} e^{-it^\mathsf{T}x}\psi(t^\mathsf{T}t)dt.$$

Also gilt für ein $O \in \mathcal{O}(D)$:

$$f(Ox) = \frac{1}{2\pi} \int_{\mathbb{R}^D} e^{-it^\mathsf{T} Ox} \psi(t^\mathsf{T} t) dt.$$

Wir substituieren $s = Ot$:

$$f(Ox) = \frac{1}{2\pi} \int_{\mathbb{R}^D} e^{-it^\mathsf{T} O^\mathsf{T} Ox} \psi(t^\mathsf{T} O^\mathsf{T} Ot)|O| dt = \frac{1}{2\pi} \int_{\mathbb{R}^D} e^{-it^\mathsf{T} x} \psi(t^\mathsf{T} t) dt = f(x).$$

Also ist f $\mathcal{O}(D)$-invariant und damit eine Funktion der unter $\mathcal{O}(D)$ maximal-invarianten Funktion $x \mapsto x^\mathsf{T} x$, d.h. $f(x) = \varphi(x^\mathsf{T} x)$ für ein passendes φ.

□

Korollar 1.2.6 (Dichten elliptischer Verteilungen)**.** *Sei $X : \Omega \to \mathbb{R}^D$ sphärisch-verteilt mit Generatorfunktion φ, $\mu \in \mathbb{R}^D$, $A \in \mathbb{R}^{D \times D}$ sei regulär. Sei $Y = \mu + A^\mathsf{T} X$ elliptisch verteilt. Dann hat Y folgende Dichte:*

$$f(y) = |\Sigma|^{-\frac{1}{2}} \varphi((y - \mu)^\mathsf{T} \Sigma^{-1} (y - \mu)), \tag{1.3}$$

wobei $\Sigma = A^\mathsf{T} A$.

Beweis: Diese Aussage folgt aus dem Transformationssatz für Dichten.

□

Es folgen zwei wichtige Definitionen der multivariaten Normal- bzw. t-Verteilung, deren Mischungen im Folgenden betrachtet werden.

Definition 1.2.7 (Multivariate Normalverteilung)**.** Sei X ein elliptisch verteilter $D \times 1$ Zufallsvektor mit Parametern μ, Σ und Generatorfunktion

$$\varphi_{gauss}(x) = \frac{1}{\sqrt{2\pi}^D} e^{-\frac{1}{2}x} \tag{1.4}$$

Wir sagen X sei *normalverteilt mit Parametern μ und Σ.*

Bemerkung 1.2.8. *Ein normalverteilter Zufallsvektor hat offensichtlich die Dichte*

$$\phi(x) = \frac{1}{\sqrt{2\pi}^D \sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\mathsf{T} \Sigma^{-1} (x-\mu)}. \tag{1.5}$$

Weiterhin gilt:

$$\mathbb{E}(X) = \mu$$

$$Cov(X) = \Sigma.$$

Ein Beweis dieser Aussage kann z.B. in [16] nachgelesen werden.

Definition 1.2.9 (Multivariate t-Verteilung)**.** Sei X ein elliptisch verteilter $D \times 1$ Zufallsvektor mit Parametern μ, Σ und Generatorfunktion

$$\varphi_t(x) = \frac{\Gamma(\frac{n+D}{2})}{\Gamma(\frac{n}{2})\sqrt{n^D}\sqrt{\pi^D}}\sqrt{(1+\frac{1}{n}x)^{n+D}}^{-1} = \frac{\Gamma(\frac{n+D}{2})}{\Gamma(\frac{n}{2})\sqrt{n^D}\sqrt{\pi^D}}(1+\frac{1}{n}x)^{-\frac{n+D}{2}} \tag{1.6}$$

Wir sagen X sei *t-verteilt mit Parametern μ, Σ und Freiheitsgrad n.*

Bemerkung 1.2.10. *Ein t-verteilter Zufallsvektor hat offensichtlich die Dichte*

$$f(x) = \frac{\Gamma(\frac{n+D}{2})}{\Gamma(\frac{n}{2})\sqrt{n^D}\sqrt{\pi^D}\sqrt{\Sigma}}(1+\frac{1}{n}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu))^{-\frac{n+D}{2}}. \tag{1.7}$$

Bemerkung zur Notation:
Im Folgenden werden wir die Generatorfunktionen ohne die von x unabhängigen multiplikativen Faktoren angeben:

$$\varphi_{gauss}(x) \equiv e^{-\frac{1}{2}x}$$

bzw.

$$\varphi_t(x) \equiv (1+\frac{1}{n}x)^{-\frac{n+D}{2}},$$

und die Dichte als

$$f(x) = k\varphi((x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu))$$

notieren, wobei k die positive Konstante ist, in der der multiplikative Vorfaktor der Generatorfunktion und $|\Sigma|^{-\frac{1}{2}}$ zusammengefasst sind.

1.3. Definition von Mischungen elliptischer Verteilungen

In vorliegender Arbeit beschäftigen wir uns mit Mischungen elliptischer Verteilungen, mit denen man unterschiedliche Verteilungen flexibel modellieren kann.

Definition 1.3.1 (Mischungen elliptischer Verteilungen)**.** Sei $C := (C_1, \ldots, C_K)$ ein $M(1, \pi)$-Zufallsvektor (multinomialverteilt mit einmaligem Ziehen und Parametervektor π), $\mathbb{P}(C = e_i) = \pi_i$, für $1 \leq i \leq K$ und $\sum_{i=1}^{K} \pi_i = 1$, wobei e_i der i'te Einheitsvektor im $\mathbb{R}^K$ ist und $X_1, \ldots, X_K$ elliptisch verteilte $D \times 1$ Zufallsvektoren mit Dichten

$$f_i(x) = k_i \varphi((x - \mu_i)^\mathsf{T} \Sigma_i^{-1} (x - \mu_i)), \tag{1.8}$$

wobei k_i eine positive Konstante ist (siehe Bemerkung zur Notation oben), $\mu_i \in \mathbb{R}^D$, $\Sigma_i \in \mathbb{R}^{D \times D}$, symmetrisch und positiv definit, $i = 1, \ldots, K$.

Der Zufallsvektor

$$X_{mix} = \sum_{i=1}^{K} C_i X_i \tag{1.9}$$

heißt *Mischung elliptischer Verteilungen*, und besitzt die Dichte

$$g(x) = \sum_{i=1}^{K} \pi_i k_i \varphi_i((x - \mu_i)^\mathsf{T} \Sigma_i^{-1} (x - \mu_i)). \tag{1.10}$$

Wir fordern, dass φ monoton fallend ist. Diese Forderung wird an einigen Stellen wichtig sein und ist in den beiden Fällen Gauss-Verteilung und t-Verteilung erfüllt.

Kapitel 2

Moden von Mischungen elliptischer Verteilungen

In diesem Kapitel untersuchen wir die Topografie (Anzahl und Lage von Moden) von Mischungen multivariater elliptischer Verteilungen. Wir verallgemeinern zunächst einige Ergebnisse aus [15] auf elliptische Verteilungen und untersuchen danach speziell die Normalverteilung und die entsprechenden Ergebnisse aus [15]. Wir gehen auch auf einige Eigenschaften der t-Verteilung ein.

2.1. Mischungen elliptischer Verteilungen

2.1.1. Ridgeline

Wir betrachten folgende Mischungsmodelle:

$$g(x;\mu_1,\ldots,\mu_K,\Sigma_1,\ldots,\Sigma_K,\pi_1,\ldots,\pi_{K-1},\vartheta_1,\ldots,\vartheta_K)=\sum_{i=1}^{K}\pi_i\cdot k_i\cdot\varphi_i((x-\mu_i)^{\mathsf{T}}\Sigma_i^{-1}(x-\mu_i)).$$

Dabei gilt: $k_i \in \mathbb{R}^+$, eine positive Konstante, die von $|\det(\Sigma_i)|^{-\frac{1}{2}}$ abhängt, $x \in \mathbb{R}^D$, $\mu_i \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{D\times D}$, für ein D aus $\mathbb{N}$.
Für die Gewichte gilt: $\pi_i \in [0,1]$ für $i = 1 \ldots K-1$, $\sum_{i=1}^{K-1}\pi_i \in [0,1]$. Das Gewicht der letzten Komponente ist durch die vorherigen Gewichte eindeutig bestimmt: $\pi_K := 1 - \sum_{i=1}^{K-1}\pi_i$.
Der Parameter ϑ_i ist ein zusätzlicher Parameter der Dichte $k_i\varphi_i((x-\mu_i)^{\mathsf{T}}\Sigma_i^{-1}(x-\mu_i))$, wie z.B. der Freiheitsgrad n_i im Falle der t-Verteilung. Gauss-Dichten haben dagegen keine weiteren Parameter außer μ_i und Σ_i. φ_i ist die Generatorfunktion der i'ten elliptischen Komponente, die von weiteren Parametern abhängen kann. Im Weiteren werden wir manchmal die Notation $\theta := (\mu_1,\ldots,\mu_K,\Sigma_1,\ldots,\Sigma_K,\pi_1,\ldots,\pi_{K-1},\vartheta_1,\ldots,\vartheta_K)$ verwenden.

Definition 2.1.1. Die K-1 dimensionale Teilmenge des $\mathbb{R}^K$

$$\mathcal{S}_K := \left\{ \alpha \in \mathbb{R}^K : \alpha_i \in [0,1], \sum_{i=1}^{K} \alpha_i = 1 \right\}$$

heißt *Einheitssimplex*.

Nun kommen wir zum wichtigsten Konstrukt dieses Kapitels:

Definition 2.1.2. Die Funktion $x^* : \mathcal{S}_K \to \mathbb{R}^D$,

$$x^*(\alpha) = \left[\alpha_1 \Sigma_1^{-1} + ... + \alpha_K \Sigma_K^{-1}\right]^{-1} [\alpha_1 \Sigma_1^{-1} \mu_1 + ... + \alpha_K \Sigma_K^{-1} \mu_K].$$

heißt *Ridgeline-Funktion*.

Bemerkung: x^* hängt von $K-1$ reellen Argumenten ab: $\alpha_1, ..., \alpha_{K-1}$. Für das letzte Element α_K gilt stets: $\alpha_K = 1 - \alpha_1 - ... - \alpha_{K-1}$, somit ist α_K keine unabhängige Variable.

Das Bild des Einheitssimplex unter dieser Abbildung, $\mathcal{M} := x^*(\mathcal{S}_K)$ nennen wir *Ridgeline-Fläche*. Wie wir gleich sehen werden, liegt $g(\mathcal{M})$ auf dem „Gebirge“ der Mischungsdichte, daher auch der Name.

Im dem wichtigen Spezialfall von Zwei-Komponenten-Mischung K=2, stellen wir die Ridgeline-Fläche etwas abweichend dar:

$$x^*(\alpha) = \left[(1-\alpha)\Sigma_1^{-1} + \alpha \Sigma_2^{-1}\right]^{-1} [(1-\alpha)\Sigma_1^{-1} \mu_1 + \alpha_2 \Sigma_2^{-1} \mu_2]. \tag{2.1}$$

Das hat den Vorteil, dass die Ridgeline-Kurve von Stelle des Gipfels der 1.Komponente zur Stelle des Gipfels der 2.Komponente führt.

Beispiel 2.1.3. *Unten sind zwei Gauss-Mischungen aus je 2 Komponenten mit jeweiligen Ridgelines (grüne Kruven) abegbildet.*

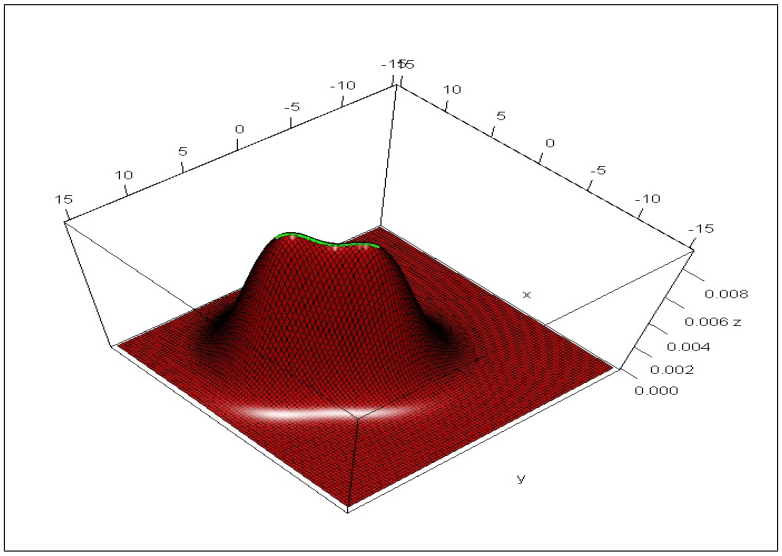


Abb. 2.1: Gauss-Mischung und Ridgeline 1

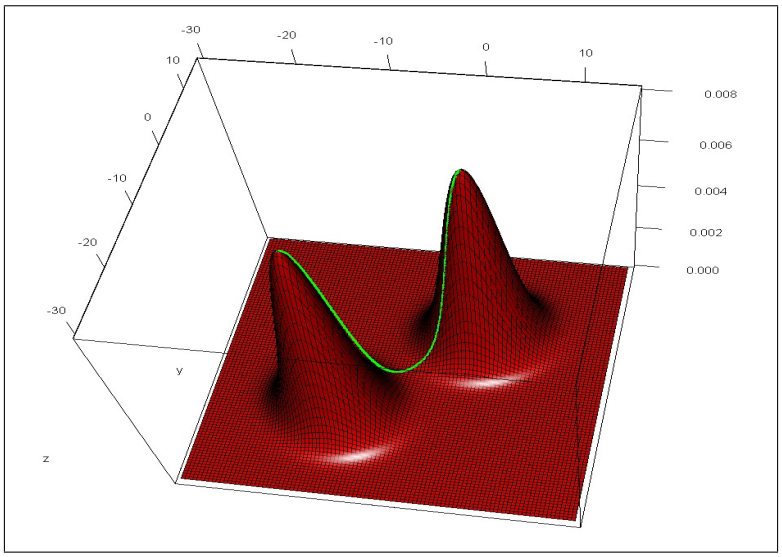


Abb. 2.2: Gauss-Mischung und Ridgeline 2

Satz 2.1.4. *Sei $g(x;\theta)$ eine K-Komponenten-Mischung mit streng monoton fallenden Generatorfunktionen φ_i wie in (1.10). Dann liegen alle kritischen Punkte von g in $\mathcal{M}$.*

Beweis: Wir benutzen hier die abkürzende Schreibweise

$$\delta(x,\mu_i) = \delta_i(x,\mu_i) := (x-\mu_i)^{\mathsf{T}}\Sigma_i^{-1}(x-\mu_i). \tag{2.2}$$

Die Gleichung für die kritischen Punkte lautet $\nabla g(x;\theta) = 0$

$$\Leftrightarrow \sum_{i=1}^{K} 2\pi_i \cdot k_i \cdot \varphi_i'(\delta(x,\mu_i)) \cdot \Sigma_i^{-1}(x-\mu_i) = 0. \tag{2.3}$$

Da nach Voraussetzung die Generatorfunktionen φ_i monoton sind, haben φ_i' überall dasselbe Vorzeichen und sind nicht alle gleichzeitig gleich Null. Also ist $\sum_{j=1}^{K} \pi_j \cdot k_j \cdot \varphi_j'(\delta(x,\mu_j)) \neq 0$ und man kann

$$\alpha_i := \frac{\pi_i \cdot k_i \cdot \varphi_i'(\delta(x,\mu_i))}{\sum_{j=1}^{K} \pi_j \cdot k_j \cdot \varphi_j'(\delta(x,\mu_j))}$$

definieren. Offensichtlich gilt $0 \leq \alpha_i \leq 1$ für alle i. Es gilt:

$$\begin{aligned}(2.3) &\Leftrightarrow \sum_{i=1}^{K} \alpha_i \cdot \Sigma_i^{-1}(x-\mu_i) = 0 \\ &\Leftrightarrow \sum_{i=1}^{K} \alpha_i \Sigma_i^{-1} x = \sum_{i=1}^{K} \alpha_i \Sigma_i^{-1} \mu_i\end{aligned}$$

und das ist äquivalent zu

$$x = \left[\sum_{i=1}^{K} \alpha_i \Sigma_i^{-1}\right]^{-1} \sum_{i=1}^{K} \alpha_i \Sigma_i^{-1} \mu_i.$$

Die letzte Gleichung definiert gerade die Ridgeline-Funktion. Es gibt also zu jedem kritischen Punkt x ein $\alpha \in \mathcal{S}_K$, so daß $x^*(\alpha) = x$. □

Korollar 2.1.5. *Falls $\Sigma_i = \Sigma$, enthält die konvexe Hülle von $\mu_1, ..., \mu_K$ alle kritischen Punkte von g.*

Beweis: Im Fall $\Sigma_i = \Sigma$ ist $\mathcal{M}$ gleich der konvexen Hülle der μ_i's. □

2.1.2. Dichte entlang der Ridgeline-Fläche

In diesem Abschnitt untersuchen wir die Eigenschaften der Abbildung

$$h(\alpha) := g(x^*(\alpha); \theta), \tag{2.4}$$

also das Verhalten der Mischungsdichte entlang der Ridgeline-Fläche. Wir werden sehen, dass die Moden und Antimoden der Funktion h zwangsläufig mit den Moden und Sattelpunkten der Dichte g korrespondieren.

Zunächst brauchen wir die partiellen Ableitungen: $\frac{\partial x^*}{\partial \alpha_j}(\alpha)$.

Proposition 2.1.6. *Sei $M(t)$ eine $D \times D$ Matrix mit $C^1(\mathbb{R}, \mathbb{R})$-Funktionen als Einträge, $M(t) = (f_{i,j}(t))_{i,j=1\ldots D}$. Dann gilt für die t, für die $M(t)$ regulär ist:*

$$\left[M(t)^{-1}\right]' = -M(t)^{-1}M(t)'M(t)^{-1}. \tag{2.5}$$

Beweis:

$$\begin{aligned} M(t)^{-1}M(t) = I \Rightarrow \left[M(t)^{-1}M(t)\right]' &= \left[M(t)^{-1}\right]' M(t) + M(t)^{-1}M(t)' = 0 \\ \Leftrightarrow \left[M(t)^{-1}\right]' M(t) &= -\, M(t)^{-1}M(t)' \\ \Leftrightarrow \left[M(t)^{-1}\right]' &= -\, M(t)^{-1}M(t)'M(t)^{-1}. \end{aligned}$$

□

Wir benutzen dieses Ergebnis und die Produktregel um $\frac{\partial x^*}{\partial \alpha_i}(\alpha)$ zu berechnen.
Definiere

$$S_\alpha := \left[\sum_{i=1}^{K} \alpha_i \Sigma_i^{-1}\right]. \tag{2.6}$$

Mit dieser Schreibweise erhält man

$$x^*(\alpha) = S_\alpha^{-1}\left(\sum_{i=1}^{K-1} \alpha_i \Sigma_i^{-1}\mu_i + (1 - \sum_{g=1}^{K-1} \alpha_g)\Sigma_K^{-1}\mu_K\right).$$

Korollar 2.1.7. *Im Fall $K > 2$ gilt für die partielle Ableitung von $x^*(\alpha)$ nach α_i:*

$$\frac{\partial x^*}{\partial \alpha_i}(\alpha) = S_\alpha^{-1}(v_K - v_i) \; \forall \; i = 1, ..., K-1. \tag{2.7}$$

Im Spezialfall $K = 2$ gilt:

$$\frac{\partial x^*}{\partial \alpha_i}(\alpha) = S_\alpha^{-1}(v_1 - v_2) \tag{2.8}$$

Wobei

$$v_i := \Sigma_i^{-1}(x^*(\alpha) - \mu_i). \tag{2.9}$$

Beweis:

$$d_i := \frac{\partial x^*}{\partial \alpha_i}(\alpha) = \frac{\partial S_\alpha^{-1}}{\partial \alpha_i}(\alpha) \cdot \left(\sum_{i=1}^{K-1} \alpha_i \Sigma_i^{-1}\mu_i + (1 - \sum_{g=1}^{K-1} \alpha_g)\Sigma_K^{-1}\mu_K \right) + S_\alpha^{-1}(\Sigma_i^{-1}\mu_i - \Sigma_K^{-1}\mu_K).$$

Es gilt im Fall $K > 2$:

$$\frac{\partial S_\alpha^{-1}}{\partial \alpha_i}(\alpha) \overset{(2.5)}{=} -S_\alpha^{-1} S_\alpha' S_\alpha^{-1} = -S_\alpha^{-1}(\Sigma_i^{-1} - \Sigma_K^{-1})S_\alpha^{-1}.$$

$$\begin{aligned}
\Rightarrow d_i &= -S_\alpha^{-1}(\Sigma_i^{-1} - \Sigma_K^{-1})S_\alpha^{-1} \left(\sum_{j=1}^{K-1} \alpha_j \Sigma_j^{-1}\mu_j + (1 - \sum_{g=1}^{K-1} \alpha_g)\Sigma_K^{-1}\mu_K \right) + S_\alpha^{-1}(\Sigma_i^{-1}\mu_i - \Sigma_K^{-1}\mu_K) \\
&= -S_\alpha^{-1}(\Sigma_i^{-1} - \Sigma_K^{-1})x^*(\alpha) + S_\alpha^{-1}(\Sigma_i^{-1}\mu_i - \Sigma_K^{-1}\mu_K) \\
&= -S_\alpha^{-1}\left(\Sigma_i^{-1}x^*(\alpha) - \Sigma_K x^*(\alpha) - \Sigma_i^{-1}\mu_i + \Sigma_K^{-1}\mu_K\right) \\
&= -S_\alpha^{-1}(\Sigma_i^{-1}(x^*(\alpha) - \mu_i) - \Sigma_K^{-1}(x^*(\alpha) - \mu_K)) \\
&= S_\alpha^{-1}(v_K - v_i).
\end{aligned}$$

Im Fall $K = 2$ ist α ein Skalar. Eine analoge Rechnung liefert dasselbe Ergebnis wie oben, bis auf den Faktor -1, der durch die alternative Definition der Ridgeline, bei der α und $1 - \alpha$ vertauscht wurden, zustande kommt. □

Die partiellen Ableitungen d_i, spannen den Raum aller Richtungen innerhalb der Ridgeline-Fläche auf.

Wir betrachten eine Menge spezieller Richtungen, mit der wir eine wichtige Aussage herleiten können:

$$W := \left\{ w \in \mathbb{R}^D : w^\mathsf{T} S_\alpha d_i = 0,\ \forall i = 1, ..., K-1 \right\}.$$

Das ist die Menge aller Richtungen, die zu dem von $\{ d_1, ..., d_{K-1} \}$ aufgespannten Teilraum orthogonal ist, im Sinne des von S_α erzeugten Skalarproduktes.

Satz 2.1.8. *Seien $w \in W$, $\alpha \in \mathcal{S}_K$, $g(x;\theta)$ eine elliptische Mischung mit monotonen Generatorfunktionen. Die Abbildung $(-\epsilon, \epsilon) \to \mathbb{R}$, $\delta \mapsto g(x^*(\alpha) + \delta \cdot w)$ hat das globale Maximum bei $\delta = 0$.*

Beweis: Wir betrachten Niveaulinien von

$$f_i(x) = k_i \cdot \varphi_i(\delta(x, \mu_i))$$

(der i'ten Mischungskomponente):

$$N_i(x^*(\alpha)) = \left\{ x \in \mathbb{R}^D : \varphi_i(\delta(x, \mu_i)) = \varphi_i(\delta(x^*(\alpha), \mu_i)) \right\}.$$

Zur Errinerung : $\delta(x, \mu_i) = (x - \mu_i)^\mathsf{T} \Sigma_i^{-1} (x - \mu_i)$. Die Niveaulinie $N_i(x^*(\alpha))$ ist Rand einer Ellipse, da $N_i(x^*(\alpha))$ Niveaulinie von $\delta(x, \mu_i)$ für ein passendes Niveau ist und Σ_i eine symmetrische positiv definite Matrix ist.

Behauptung: Der Gradient $\nabla f_i(x)$ steht orthogonal auf $N_i(x^*(\alpha))$.

Begründung: sei $n(t) : [0,1] \to N_i(x^*(\alpha))$ eine Parametrisierung von $N_i(x^*(\alpha))$. Die Funktion f_i ist konstant auf $N_i(x^*(\alpha)) \Rightarrow f_i(n(t)) \equiv const \Rightarrow f_i(n(t))' = \nabla f_i(x)^\mathsf{T} n'(t) = 0$. Somit ist $\nabla f_i(x)$ orthogonal zum Tangentialvektor an $N_i(x^*(\alpha))$.

$\nabla f_i(x)$ und v_i sind kollinear, d.h. $\nabla f_i(x) = \beta \cdot v_i$ für ein passendes skalares β. Deshalb steht auch v_i orthogonal auf der Niveaulinie $N_i(x^*(\alpha))$.

Die Niveaumenge $E_i(x^*(\alpha)) := \left\{ x \in R^D : f_i(x) \geq f_i(x^*(\alpha)) \right\} = \bigcup_{c:\ f_i(c) \geq f_i(x^*(\alpha))} N_i(c)$ ist eine Ellipse, da die Generatorfunktion φ_i monoton fallend ist.

Sei $w \in \mathbb{R}^D$ orthogonal zu v_i. Die Gerade $t \mapsto x^*(\alpha) + tw$ verläuft in der Stützhyperebene der Ellipse $E_i(x^*(\alpha))$. Eine Ellipse ist konvex, deswegen liegt die Gerade außerhalb von

$E_i(x^*(\alpha))$ und berührt sie nur im Punkt $x^*(\alpha)$.

Die Gerade liegt also in der Menge $E_i(x^*(\alpha))^c = \left\{ x \in R^D : f_i(x) < f_i(x^*(\alpha)) \right\}$. Daraus folgt, daß w eine Abstiegsrichtung von f_i ist.

Sei nun $w \in W$, das ist äquivalent zu

$$w^\mathsf{T}(v_i - v_K) = 0,\ i = 1, ..., K-1. \tag{2.10}$$

Wir zeigen, daß w orthogonal zu jedem $v_i,\ i = 1, ..., K-1$ ist.

Es gilt:

$$\begin{aligned}\sum_{i=1}^{K} \alpha_i v_i &= \sum_{i=1}^{K} \alpha_i \Sigma_i^{-1}(x^*(\alpha) - \mu_i) = \sum_{i=1}^{K} \alpha_i \Sigma_i^{-1}(S_\alpha^{-1} \sum_{j=1}^{K} \alpha_j \Sigma_j^{-1} - \mu_i) \\ &= \underbrace{\sum_{i=1}^{K} \alpha_i \Sigma_i^{-1}}_{=S_\alpha} S_\alpha^{-1} \sum_{j=1}^{K} \alpha_j \Sigma_j^{-1} \mu_j - \sum_{i=1}^{K} \alpha_i \Sigma_i^{-1} \mu_i = 0.\end{aligned}$$

$$\Rightarrow \sum_{i=1}^{K} \alpha_i (v_i - v_K) = -v_K \Rightarrow -v_K^\mathsf{T} w = \sum_{i=1}^{K} \alpha_i \underbrace{w^\mathsf{T}(v_i - v_K)}_{=0} = 0 \Rightarrow v_i^\mathsf{T} w = 0\ i = 1, ..., K.$$

Der Vektor w ist also orthogonal zu jedem $v_i,\ i = 1, ..., K$. Daraus folgt, daß w Abstiegsrichtung von jeder Mischungskomponente $f_i,\ i = 1, ..., K$ ist, also auch von Mischung g. □

Jetzt zeigen wir eine wichtige Optimalitätseigenschaft der Ridgeline, die es rechtfertigt, die Suche nach kritischen Punkten von g auf die Suche nach kritischen Punkten von h zu beschränken, was im Fall hochdimensionaler Daten $D \gg K-1$ eine enorme Dimensionsreduktion bedeutet.

Korollar 2.1.9. *i. Der Punkt α_{krit} ist genau dann ein kritischer Punkt von $h(\alpha) = g(x^*(\alpha))$, wenn $x^*(\alpha_{krit})$ ein kritischer Punkt von $g(x)$ ist. α_{krit} ist genau dann ein lokales Maximum von h, wenn $x^*(\alpha_{krit})$ ein lokales Maximum von g ist.*

ii. Falls $D > K-1$, hat die Mischungsdichte g keine lokalen Minima, sondern nur lokale Maxima und Sattelpunkte.

Beweis:

i. „⇐“ Sei $x^*(\alpha_{krit})$ ein kritischer Punkt von g

$$\Rightarrow \nabla h(\alpha)^\mathsf{T} = \overbrace{\nabla g(x^*(\alpha_{krit}))^\mathsf{T}}^{=0} J_x(\alpha_{krit}) = 0.$$

Wobei J_x die Jacobi-Matrix der Ridgeline-Funktion ist. Also ist α_{krit} kritischer Punkt von h.

„⇒“ Sei α_{krit} ein kritischer Punkt von h

$$\Rightarrow \nabla h(\alpha) = 0 \Leftrightarrow \nabla g(x^*(\alpha_{krit}))^\mathsf{T} d_j = 0 \; j = 1, ..., K-1.$$

Der $\mathbb{R}^D$ zerfällt in die direkte Summe von $span\{ d_1, ..., d_{K-1} \}$ und seinem orthogonalen Komplement W. Wenn wir zeigen könnten, daß auch $\nabla g(x^*(\alpha_{krit}))^\mathsf{T} w = 0 \; \forall w \in W$ gelte, folgte $\nabla g(x^*(\alpha_{krit})) = 0$ und wir wären fertig.

Dafür betrachten wir die Taylor-Entwicklung von g um $x^*(\alpha_{krit})$ an der Stelle $x^*(\alpha_{krit}) + \delta w$ für ein $w \in W$:

$$g(x^*(\alpha_{krit}) + \delta w) = g(x^*(\alpha_{krit})) + \delta \nabla g(x^*(\alpha_{krit}))^\mathsf{T} w + \frac{1}{2}\delta^2 w^\mathsf{T} H w + O(|\delta|^3),$$

wobei H die Hessematrix von g im Punkt $x^*(\alpha_{krit})$ ist. Aus Satz 2.1.8 folgt für alle $\delta \in (-\epsilon, \epsilon)$:

$$\delta \nabla g(x^*(\alpha_{krit}))^\mathsf{T} w + \frac{1}{2}\delta^2 w^\mathsf{T} H w + O(|\delta|^3) < 0$$

und damit

$$\nabla g(x^*(\alpha_{krit}))^\mathsf{T} w = 0.$$

Der Gradient $\nabla g(x^*(\alpha_{krit}))$ ist also sowohl zu $span\{ d_1, ..., d_{K-1} \}$ orthogonal, als auch zu $W = span\{ d_1, ..., d_{K-1} \}^\perp$, daraus folgt wiederum, dass er verschwindet.

Der Satz 2.1.8 besagt, daß die Mischung g beim Verlassen der Ridgeline lokal fällt. Daraus folgt, daß falls α_{krit} ein lokales Maximum von h ist, der zugehörige Punkt auf der Ridgeline $x^*(\alpha_{krit})$ ein lokales Maximum von g ist. Die umgekehrte Aussage ist trivial.

ii. Es gilt: $W^\perp = span(d_1, ..., d_{K-1}) \Rightarrow dim(W) \geq D - (K-1)$. Nach Voraussetzung gilt: $D > K - 1 \Rightarrow dim(W) \geq 1$. In jedem Punkt $x^*(\alpha)$ gibt es also eine Abstiegsrichtung, entlang derer die Dichte g fällt, d.h. es gibt keine lokalen Minima. □

2.1.3. Der Pi-Graph und die Krümmungsfunktion

Der Pi-Graph

Wir verwenden im Folgenden die Notation

$$x_\alpha := x^*(\alpha)$$

um den Lesefluß zu erleichtern.
Im letzten Abschnitt haben wir gezeigt, daß es reicht die kritischen Punkte von $h(\alpha) = g(x^*(\alpha))$ zu betrachten, da α genau dann ein kritischer Punkt von h ist, wenn $x^*(\alpha)$ ein kritischer Punkt von g ist.

Nun stellst sich die Frage wie man die kritischen Punkte von h findet. Ab jetzt betrachten wir nur den Speziallfall **K=2**. Sei α ein kritischer Punkt von $h(\alpha)$, also

$$h'(\alpha) = \pi[f_1(x_\alpha)]' + \bar{\pi}[f_2(x_\alpha)]' = 0.$$

Der Operator $'$ bezeichnet dabei wie üblich die Differentiation nach dem reellen Parameter α.
Das Umstellen der letzten Gleichung nach π liefert:

$$\pi = \frac{[f_2(x_\alpha)]'}{[f_2(x_\alpha)]' - [f_1(x_\alpha)]'} := \Pi(\alpha). \tag{2.11}$$

Um für eine konkrete Mischungsproportion π die kritischen Punkte der Dichte g zu finden, müssen wir die Lösungen α der „Π-Gleichung“:

$$\Pi(\alpha) = \pi \tag{2.12}$$

bestimmen. Daraus lassen sich die Punkte x_α errechnen. Die Anzahl der Lösungen der Π-Gleichung für ein festes π hängt von der Anzahl der Oszillationen (Vorzeichenwechsel der 1.Ableitung) der Funktion $\Pi(\alpha)$ für $\alpha \in [0, 1]$ ab.
Im Falle einer unimodalen Mischung mit Proportion π hat die Π-Gleichung nur eine Lösung, die mit dem Modus der Mischungsdichte korrespondiert. Eine bimodale Dichte hat drei π-kritische Punkte; der 1. und 3. in Reihenfolge des Verlaufs der Funktion $\Pi(\alpha)$ korrespondieren mit den jeweiligen Moden, der 2. mit dem Sattelpunkt.

Wir werden später oft folgendes Ergebnis brauchen, das uns die Ableitungen der Verkettung der Mahalanobis-Distanz zum Mittelwert μ_i mit der Ridgeline liefert:

Lemma 2.1.10. *Betrachte $\delta_i = \delta(x_\alpha, \mu_i) = (x_\alpha - \mu_i)^\mathsf{T}\Sigma_i^{-1}(x_\alpha - \mu_i)$. Dann gilt:*

$$\Delta_1 := \frac{\partial \delta_1}{\partial \alpha} = 2(\mu_2 - \mu_1)^\mathsf{T}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}S_\alpha^{-1}v_1 \tag{2.13}$$

$$\Delta_2 := \frac{\partial \delta_2}{\partial \alpha} = -\frac{\bar\alpha}{\alpha}\Delta_1 \tag{2.14}$$

Beweis: Anwenden der Kettenregel liefert:

$$\begin{aligned}\Delta_i &= 2(x_\alpha - \mu_i)^\mathsf{T}\Sigma_i^{-1}S_\alpha^{-1}(v_1 - v_2) \\ &= 2\left[S_\alpha^{-1}(\bar\alpha\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2 - S_\alpha^{-1}S_\alpha\mu_i\right]^\mathsf{T}\Sigma_i^{-1}S_\alpha^{-1}(v_1 - v_2) \\ &= 2\left[S_\alpha^{-1}(\bar\alpha\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2 - S_\alpha\mu_i)\right]^\mathsf{T}\Sigma_i^{-1}S_\alpha^{-1}(v_1 - v_2) \\ &= 2\left[S_\alpha^{-1}(\bar\alpha\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2 - \bar\alpha\Sigma_1^{-1}\mu_i - \alpha\Sigma_2^{-1}\mu_i)\right]^\mathsf{T}\Sigma_i^{-1}S_\alpha^{-1}(v_1 - v_2)\end{aligned} \tag{2.15}$$

Das ergibt:

$$\Delta_1 = 2\left[S_\alpha^{-1}\alpha\Sigma_2^{-1}(\mu_2 - \mu_1)\right]^\mathsf{T}\Sigma_1^{-1}S_\alpha^{-1}(v_1 - v_2) = 2\alpha(\mu_2 - \mu_1)^\mathsf{T}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}S_\alpha^{-1}(v_1 - v_2).$$

$$\Delta_2 \overset{\text{analog}}{=} 2\bar\alpha(\mu_1 - \mu_2)^\mathsf{T}\Sigma_1^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}(v_1 - v_2) = -2\bar\alpha(\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}(v_1 - v_2).$$

Hilfsaussage 1:

$$\Sigma_1^{-1}S_\alpha^{-1}\Sigma_2^{-1} = \Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}. \tag{2.16}$$

Begründung:

$\Sigma_1^{-1}S_\alpha^{-1}\Sigma_2^{-1} = [\Sigma_2 S_\alpha \Sigma_1]^{-1} = \left[\Sigma_2(\bar\alpha\Sigma_1^{-1} + \alpha\Sigma_2^{-1})\Sigma_1\right]^{-1} = (\bar\alpha\Sigma_2 + \alpha\Sigma_1)^{-1}$

$\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1} = [\Sigma_1 S_\alpha \Sigma_2]^{-1} = \left[\Sigma_1(\bar\alpha\Sigma_1^{-1} + \alpha\Sigma_2^{-1})\Sigma_2\right]^{-1} = (\bar\alpha\Sigma_2 + \alpha\Sigma_1)^{-1}.$

Hilfsaussage 2:

$$v_2 = -\frac{\bar\alpha}{\alpha}v_1 \tag{2.17}$$

Begründung: die Gleichung, die die Ridgeline-Funktion x^* definiert, lautet:

$$\bar\alpha\Sigma_1^{-1}(x^* - \mu_1) + \alpha\Sigma_2^{-1}(x^* - \mu_2) = \bar\alpha v_1 + \alpha v_2 = 0 \text{ (siehe Korollar 2.1.7).}$$

Mit den beiden obigen Hilfsaussagen und unseren Formeln für Δ_1 und Δ_2 folgt die Behauptung. □

Jetzt betrachten wir einige Eigenschaften der Funktion Π, die uns auf der Suche nach Lösungen der Π-Gleichung helfen werden:

Lemma 2.1.11. *Es gilt:*

$$\Pi(0) = 1$$
$$\Pi(1) = 0$$
$$\Pi(\alpha) \in [0, 1].$$

Beweis: An den Rändern des Intervalls gilt $x^*(0) = \mu_1$ und $x^*(1) = \mu_2$. Der Modus der i'ten elliptischen Mischungs-Komponente ist μ_i, dort verschwindet ihr Gradient ∇f_i. Also folgt: $[f_1(x_\alpha)]'\big|_{\alpha=0} = 0$ und $[f_2(x_\alpha)]'\big|_{\alpha=1} = 0$.

Bleibt zu zeigen: $\Pi(\alpha) \in [0, 1]$.
Behauptung:

$$[f_1(x_\alpha)]' \leq 0$$
$$[f_2(x_\alpha)]' \geq 0$$
$$\forall \alpha \in [0, 1].$$

Beweis der Behauptung: Es gilt $[f_i(x_\alpha)]' = [k_i\varphi_i(\delta(x_\alpha, \mu_i))]' = k_i\varphi_i'(\delta(x_\alpha, \mu_i))\Delta_i$. Die Generatorfunktionen sind monoton fallend, also ist $\varphi_i' \leq 0$. Weiterhin gilt $\Delta_1 = (x_\alpha - \mu_1)^\mathsf{T}\Sigma_1^{-1}S_\alpha^{-1}(v_1 - v_2)$, siehe (2.15). Das läßt sich mit (2.17) weiter umformen zu $\Delta_1 = \frac{1}{\alpha}(x_\alpha - \mu_i)^\mathsf{T}\Sigma_i^{-1}S_\alpha^{-1}v_1 = \frac{1}{\alpha}v_1^\mathsf{T}S_\alpha^{-1}v_1$. Die Matrix S_α^{-1} ist positiv definit. Damit folgt $\Delta_1 > 0$ und $\Delta_2 = -\frac{\bar{\alpha}}{\alpha}\Delta_1 < 0$.
Insgesamt liefert das $[f_1(x_\alpha)]' < 0$ und $[f_2(x_\alpha)]' > 0$.

Ende Beweis der Behauptung □

Der Nenner des Quotienten $\frac{[f_2(x_\alpha)]'}{[f_2(x_\alpha)]' - [f_1(x_\alpha)]'}$ ist stets größer als der Zähler, beide sind nichtnegativ, somit bleibt der Quotient im Intervall $[0, 1]$. □

Für die beiden speziellen Fälle *Gauss-Mischungen* und *t-Mischungen* existieren einfache Formeln um $\Pi(\alpha)$ numerisch zu bestimmen:

Proposition 2.1.12. Π *für Gauss-Mischungen*
Sei $g(x; \theta) = \pi\phi_1 + \bar{\pi}\phi_2$ *eine Gauss-Mischung. Dann gilt:*

$$\frac{1}{\Pi(\alpha)} = \frac{[\phi_2(x_\alpha)]' - [\phi_1(x_\alpha)]'}{[\phi_2(x_\alpha)]'} = 1 + \frac{\alpha\phi_1(x_\alpha)}{\bar{\alpha}\phi_2(x_\alpha)} \quad \forall \alpha \notin 0, 1 \tag{2.18}$$

Beweis:

$$\frac{[\phi_2(x_\alpha)]' - [\phi_1(x_\alpha)]'}{[\phi_2(x_\alpha)]'} = 1 + \frac{\alpha\phi_1(x_\alpha)}{\bar{\alpha}\phi_2(x_\alpha)} \Leftrightarrow -\frac{[\phi_1(x_\alpha)]'}{[\phi_2(x_\alpha)]'} = \frac{\alpha\phi_1(x_\alpha)}{\bar{\alpha}\phi_2(x_\alpha)}.$$

Jetzt betrachten wir die linke Seite der letzten Gleichung:

$$\frac{[\phi_1(x_\alpha)]'}{[\phi_2(x_\alpha)]'} \overset{K.R.}{=} \frac{\phi_1(x_\alpha)(x_\alpha-\mu_1)^\mathsf{T}\Sigma_1^{-1}S_\alpha^{-1}(v_1-v_2)}{\phi_2(x_\alpha)(x_\alpha-\mu_2)^\mathsf{T}\Sigma_2^{-1}S_\alpha^{-1}(v_1-v_2)} = \frac{\phi_1(x_\alpha)\Delta_1}{\phi_2(x_\alpha)\Delta_2} \overset{(2.14)}{=} -\frac{\alpha\phi_1(x_\alpha)}{\bar{\alpha}\phi_2(x_\alpha)}.$$

□

Proposition 2.1.13. Π *für t-Mischungen*

Sei $g(x;\theta) = \pi f_1 + \bar{\pi} f_2$ *eine t-Mischung. Dann gilt:*

$$\frac{1}{\Pi(\alpha)} = 1 + \frac{\alpha n_2 c_1 (n_1+D)[1+\frac{1}{n_1}\delta(x_\alpha,\mu_1)]^{-\frac{n_1+D+2}{2}}}{\bar{\alpha} n_1 c_2 (n_2+D)[1+\frac{1}{n_2}\delta(x_\alpha,\mu_2)]^{-\frac{n_2+D+2}{2}}} \quad \forall \alpha \notin 0,1 \tag{2.19}$$

Dabei bedeuten für $i = 1, 2$*:*

$$c_i = \frac{\Gamma(\frac{n_i+D}{2})}{\Gamma(\frac{n_i}{2})\sqrt{n_i^D}\sqrt{\pi^D}\sqrt{|\Sigma_i|}}$$
$$\delta(x_\alpha,\mu_i) = (x_\alpha-\mu_i)^\mathsf{T}\Sigma_i^{-1}(x_\alpha-\mu_i).$$

Beweis: Die Generatorfunktion der i'ten t-Komponente ist

$$\varphi_i(t) = \sqrt{(1+\frac{1}{n_i}t)^{n_i+D}}^{-1} = (1+\frac{1}{n_i}t)^{-\frac{n_i+D}{2}}.$$

Die Dichte der i'ten t-Komponente lautet somit

$$f_i(x) = c_i\sqrt{(1+\frac{1}{n_i}(x-\mu_i)^\mathsf{T}\Sigma_i^{-1}(x-\mu_i))^{n_i+D}}^{-1} = c_i(1+\frac{1}{n_i}(x_\alpha-\mu_i)^\mathsf{T}\Sigma_i^{-1}(x_\alpha-\mu_i))^{-\frac{n_i+D}{2}}.$$

Differentiation von $(f_i \circ x^*)(\alpha)$ nach α liefert:

$$[f_i(x_\alpha)]' = -\frac{c_i(n_i+D)}{n_i}(1+\frac{1}{n_i}\delta(x_\alpha,\mu_i))^{-\frac{n_i+D+2}{2}} \cdot \underbrace{(x_\alpha-\mu_i)^\mathsf{T}\Sigma_i^{-1}\dot{x}_\alpha}_{=\Delta_i}.$$

Also folgt

$$\frac{1}{\Pi(\alpha)} = 1 - \frac{c_1\varphi_1'(x_\alpha)\Delta_1}{c_2\varphi_2'(x_\alpha)\Delta_2} \overset{(2.14)}{=} 1 + \frac{\alpha c_1\varphi_1'(x_\alpha)}{\bar{\alpha}c_2\varphi_2'(x_\alpha)}$$

□

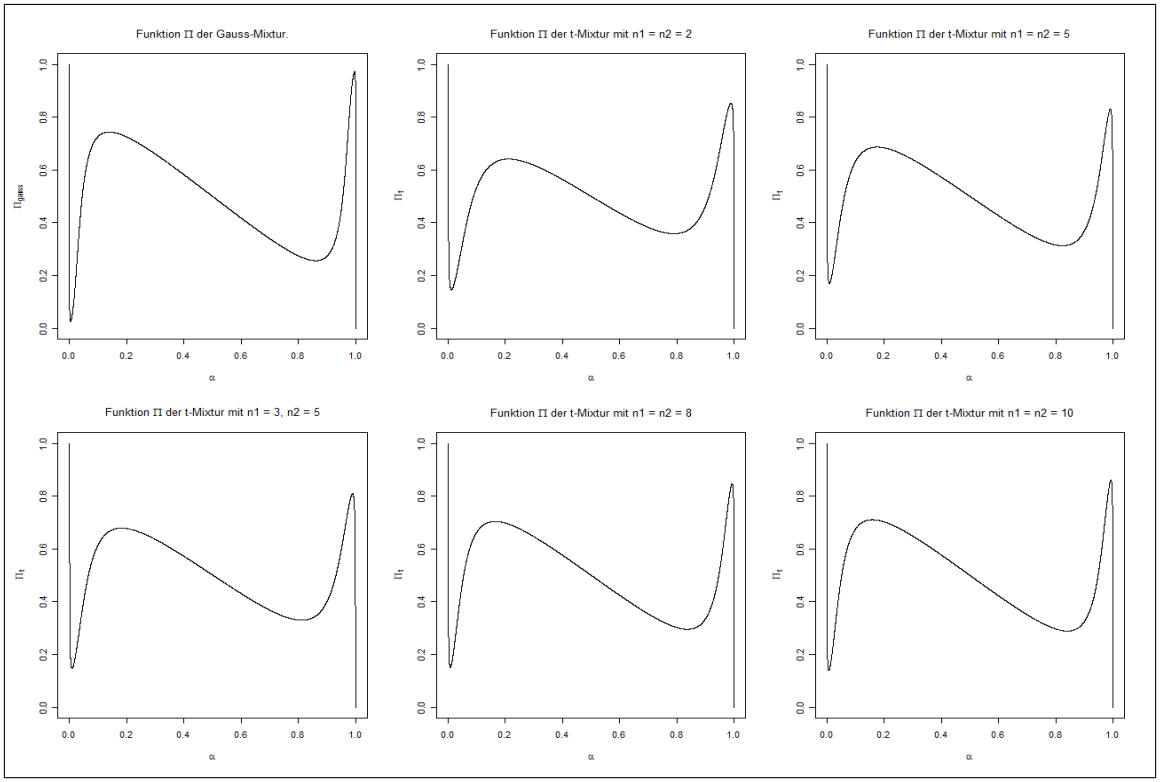
Bemerkung: Im allgemeinen Fall einer elliptischen Mischung mit Generator-Funktionen φ_1, φ_2 gilt für $\alpha \notin 0, 1$:

$$\frac{1}{\Pi(\alpha)} = 1 + \frac{\alpha c_1\varphi_1'(\delta(x_\alpha, \mu_1))}{\bar{\alpha}c_2\varphi_2'(\delta(x_\alpha, \mu_2))}.$$

Wobei c_i die Normierungskonstante der i'ten elliptischen Komponente ist. Diese Formel ergibt sich auf dieselbe Weise wie die beiden obigen mit Hilfe des Lemma 2.1.10.

Beispiel 2.1.14. *Wir vergleichen jetzt die* Π*-Funktionen zweier Mischungen; einer Gauss- und einer t-Mischung mit gleichen Parametern* $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0.05 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.05 & 0 \\ 0 & 1 \end{pmatrix}$,

und verschiedenen Freiheitsgraden der t-Mischung-Komponenten n_1, n_2.



Wie man sieht, sind die Π*-Funktionen der beiden Verteilungen ähnlich. Die Dichten der Gauss-Mischung und der t-Mischung mit* $n_1 = n_2 = 10$ *sehen beihnahe identisch aus. Interessanter ist ein Vegleich der Gauss-Mischung mit der der t-Mischung mit* $n_1 = 3$, $n_2 = 5$. *Am Verlauf der* Π*-Funktion dieser t-Mischung sieht man einen kleineren Bereich für* π *in dem die Gleichung* $\Pi(\alpha) = \pi$ *fünf Lösungen hat. So sehen die verschiedenen Mischungen im Vergleich aus:*

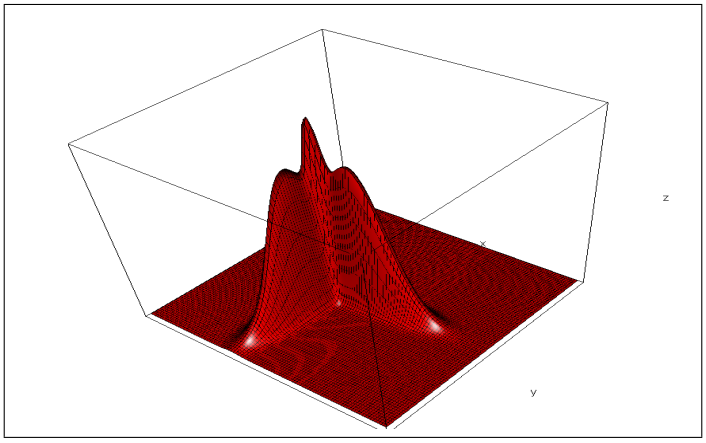


Abb. 2.3: Gauss-Mischung mit Mischungsproportion $\pi = 0.5$

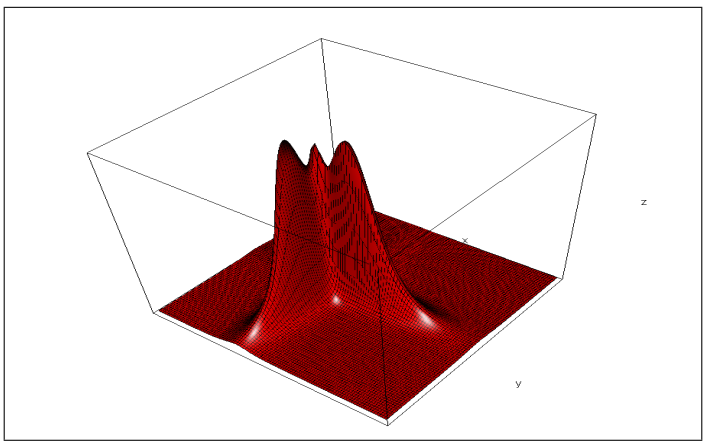


Abb. 2.4: t-Mischung mit Mischungsproportion $\pi = 0.5$

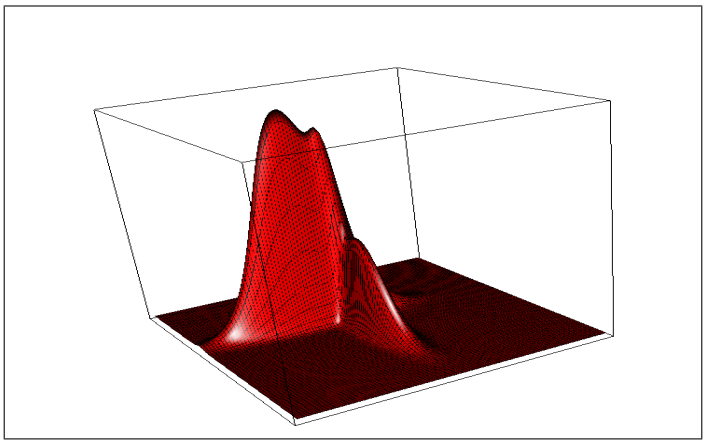


Abb. 2.5: Gauss-Mischung mit Mischungsproportion $\pi = 0.3$

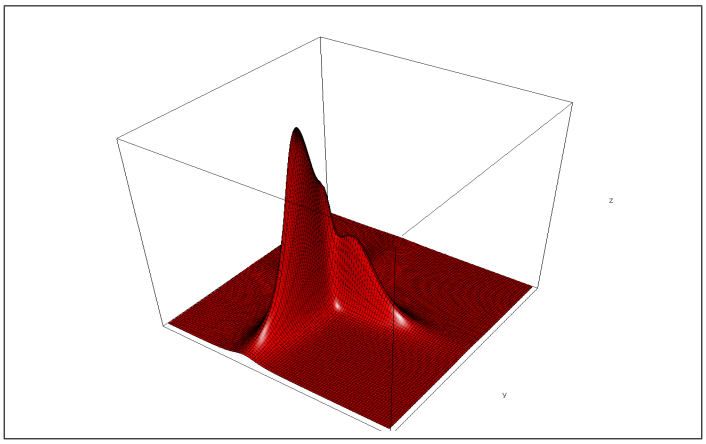


Abb. 2.6: t-Mischung mit Mischungsproportion $\pi = 0.3$

Mit obigen Formeln kann man die Funktion Π numerisch günstig ausrechnen und grafisch darstellen. Aber um allgemeine Aussagen über die Topografie von Mischungen herzuleiten, müssten wir den „Verlauf“ der Funktion Π analysieren. Dazu betrachten wir die sogenannte Krümmungsfunktion.

Die Krümmungsfunktion

Ab jetzt verwenden wir neben $'$ bzw. $''$ für die Ableitungen von Funktionen nach α, auch den Ableitungspunkt: $\dot{x}_\alpha$ bzw. $\ddot{x}_\alpha$. Beide Schreibweisen sind gleichberechtigt.

Um eine Aussage über das Oszillationsverhalten der Funktion Π zu machen, betrachten wir deren 1. Ableitung:

$$\Pi'(\alpha) = -\frac{[f_2(x_\alpha)]''[f_1(x_\alpha)]' - [f_1(x_\alpha)]''[f_2(x_\alpha)]'}{([f_2(x_\alpha)]' - [f_1(x_\alpha)]')^2} \tag{2.20}$$

Folgender Korollar macht eine Aussage über den Zusammenhang zwischen der Anzahl der Nullstellen der 1. Ableitung der Funktion Π und der Anzahl der Moden der Mischungsdichte g.

Korollar 2.1.15. *Sei $\pi \in [0,1]$ und die Funktion $\Pi'(\alpha)$ habe N Nullstellen, dann gilt:*

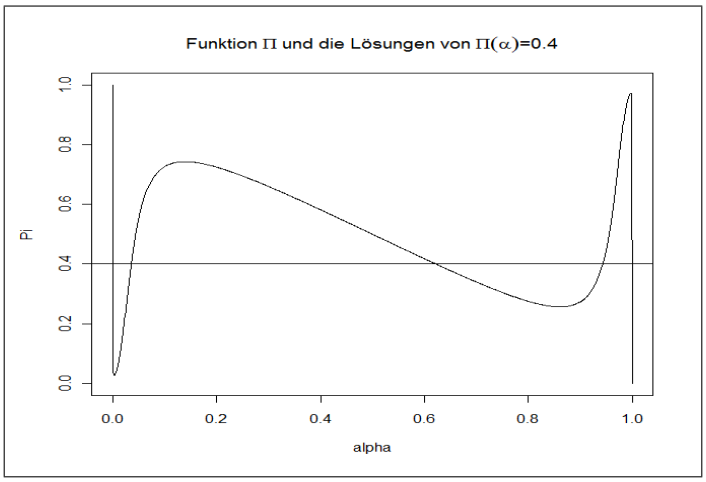
1. *N ist gerade.*
2. *Die Gleichung $\Pi(\alpha) = \pi$ hat maximal $N+1$ Lösungen $\alpha_1, ..., \alpha_{N+1}$, wobei die Indizierung in aufsteigender Reihenfolge gilt: $\alpha_{i+1} > \alpha_i$.*
3. *Für die zugehörigen Lösungen gilt: $x^*(\alpha_1)$ ist ein Modus von g, $x^*(\alpha_2)$ ist ein Sattelpunkt oder ein Minimum von g. Es gibt höchstens $1 + \frac{N}{2}$ Moden.*

Beweis:

1. Wir wissen bereits, daß $\Pi(0) = 1$, $\Pi(1) = 0$ und $\Pi(\alpha) \in [0,1]$ gilt. Daraus folgt unmittelbar $\Pi'(0) \leq 0$ und $\Pi'(1) \leq 0$. Die Ableitung Π' hat an den Rändern des Intervalls $[0,1]$ dasgleiche Vorzeichen, daraus folgt, dass Π' eine gerade Anzahl von Nullstellen im Intervall [0,1] haben muß, damit erhalten wir die 1. Aussage.
2. Die Gleichung $\Pi(\alpha) = \pi$ hat mindestens eine Lösung, da eine Mischungsdichte immer mindestens einen Modus hat. Mit jeder Oszillation der Funktion Π, d.h. mit jeder Nullstelle der Funktion Π' kommt höchstens eine neue Lösung der Gleichung $\Pi(\alpha) = \pi$ hinzu. Insgesamt kann es maximal $N+1$ Lösungen geben.
3. Wir haben bereits gezeigt: $[f_1(x_\alpha)]' < 0$ und $[f_2(x_\alpha)]' > 0$ $\forall \alpha \in [0,1]$ (vgl. Beweis von Lemma 2.1.11). $h'(\alpha) = \pi[f_1(x_\alpha)]' + \bar{\pi}[f_2(x_\alpha)]' \Rightarrow h'(0) = \bar{\pi}[f_2(x_\alpha)]' > 0 \Rightarrow$ die Mischungsdichte g steigt auf der Ridgeline für $\alpha \in (0, \alpha_1) \Rightarrow \alpha_1$ ist ein Modus. Im Extremfall wechseln sich die Moden mit den Sattelpuknten bzw. Minima ab. Erinnerung: Minima kann g nur im Fall $D = 1$ haben (siehe Korollar 2.1.9 ,ii). Im Extremfall erhalten wir: α_1 ist ein Modus, α_2 ist ein Sattelpunkt/Minimum,...,α_N ist ein Sattelpunkt/Miminum, α_{N+1} ist ein Modus. Insgesamt kann es also bei einer festen Mischungsproportion π maximal $1 + \frac{N}{2}$ Moden geben.

□

Folgendes Bild verleiht dem obigen Beweis geometrische Anschauung:



Die erste Lösung, ist die Stelle wo sich die horizontale Gerade und der Graph von Π das erste mal schneiden, sie korresponiert mit einem Modus, die zweite Lösung korrespondiert mit einem Sattelpunkt usw..

Unser Ziel ist es nun die Anzahl der Nullstellen der Funktion Π' zu untersuchen. Jede Funktion, die denselben Zähler wie Π' und einen positiven Nenner hat, bezeichnen wir als *Krümmungsfunktion* κ. Oft ist es hilfreich sich die Krümmungsfunktion anzuschauen und nicht Π', weil κ analytisch zugänglicher ist.

Im Falle eine Gaussmischung wählen wir:

$$\kappa_{gauss}(\alpha) = \frac{[\phi_2(x_\alpha)]''}{\phi_2(x_\alpha)} \frac{[\phi_1(x_\alpha)]'}{\phi_1(x_\alpha)} - \frac{[\phi_1(x_\alpha)]''}{\phi_1(x_\alpha)} \frac{[\phi_2(x_\alpha)]'}{\phi_2(x_\alpha)} \tag{2.21}$$

Im Falle der t-Mischung arbeiten wir direkt mit Π'.

Satz 2.1.16. *Sei $g(x;\theta)$ eine Mischung aus zwei Gauß-Komponenten. Dann gilt*

$$\kappa_{gauss}(\alpha) = [p(\alpha)]^2[1 - \alpha\bar{\alpha}p(\alpha)],$$

wobei:

$$p(\alpha) = (\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}(\mu_2 - \mu_1). \tag{2.22}$$

Beweis:
Zunächst einige Vorbereitungen (siehe dazu Lemma 2.1.10):

Es gilt:

1. $v_2 = -\frac{\bar{\alpha}}{\alpha}v_1$.
2. $\dot{x}_\alpha = S_\alpha^{-1}(v_1 - v_2) = \frac{1}{\alpha}S_\alpha^{-1}v_1$.
3. $\Delta_i = (x_\alpha - \mu_i)^\mathsf{T}\Sigma_i^{-1}\dot{x}_\alpha = \frac{1}{\alpha}(x_\alpha - \mu_i)^\mathsf{T}\Sigma_i^{-1}S_\alpha^{-1}v_1$
4. $\dot{\phi}_i(x_\alpha) = \phi_i(x_\alpha)\Delta_i$
5. $\frac{\dot{\phi}_i(x_\alpha)}{\phi_i(x_\alpha)} = -v_i^\mathsf{T}\dot{x}_\alpha$.

Wir definieren

$$l_i := log(\phi_i(x_\alpha)),$$

daraus folgt:

$$\dot{l}_i = \frac{\dot{\phi}_i(x_\alpha)}{\phi_i(x_\alpha)}$$

und

$$\ddot{l}_i = \frac{\ddot{\phi}_i(x_\alpha)\phi(x_\alpha)}{\phi_i(x_\alpha)^2} - \left[\frac{\dot{\phi}_i(x_\alpha)}{\phi_i(x_\alpha)}\right]^2$$

was äquivalent ist zu

$$\frac{\ddot{\phi}_i(x_\alpha)}{\phi(x_\alpha)} = \ddot{l}_i + \dot{l}_i^2.$$

Wir rechnen diese Größen aus.

$$\dot{l}_1 = -v_1^\mathsf{T}\dot{x}_\alpha$$
$$\dot{l}_2 = -v_2^\mathsf{T}\dot{x}_\alpha = \frac{\bar{\alpha}}{\alpha}v_1^\mathsf{T}\dot{x}_\alpha$$

Was leicht einzusehen ist. Die zweiten Ableitungen erfodern etwas mehr Rechenaufwand:

$$\begin{aligned}
\ddot{\phi}_1(x_\alpha) &= -\dot{\phi}_1(x_\alpha) v_1^\mathsf{T} \dot{x}_\alpha - \phi_1(x_\alpha)[\dot{v}_1^\mathsf{T} \dot{x}_\alpha + v_1^\mathsf{T} \ddot{x}_\alpha] \\
&= \phi_1(x_\alpha)[v_1^\mathsf{T} \dot{x}_\alpha]^2 - \phi_1(x_\alpha)[\dot{v}_1^\mathsf{T} \dot{x}_\alpha + v_1^\mathsf{T} \ddot{x}_\alpha] \\
&\Rightarrow \frac{\ddot{\phi}_1(x_\alpha)}{\phi_1(x_\alpha)} = [v_1^\mathsf{T} \dot{x}_\alpha]^2 - [\dot{v}_1^\mathsf{T} \dot{x}_\alpha + v_1^\mathsf{T} \ddot{x}_\alpha]
\end{aligned}$$

$$\Rightarrow \ddot{l}_1 = -\dot{v}_1^\mathsf{T} \dot{x}_\alpha - v_1^\mathsf{T} \ddot{x}_\alpha$$

Für $\ddot{l}_2$ verwenden wir

$$\left[\frac{\dot{\phi}_2(x_\alpha)}{\phi_2(x_\alpha)}\right]^2 = [v_2^\mathsf{T} \dot{x}_\alpha]^2 = \left[\frac{\bar{\alpha}}{\alpha} v_1^\mathsf{T} x_\alpha\right]^2.$$

und eine analoge Rechnung wie im Fall $\ddot{l}_1$ liefert

$$\ddot{l}_2 = \frac{\bar{\alpha}}{\alpha} \ddot{x}_\alpha^\mathsf{T} v_1 + \frac{\bar{\alpha}}{\alpha} \dot{x}_\alpha^\mathsf{T} \dot{v}_1 - \frac{1}{\alpha^2} \dot{x}_\alpha^\mathsf{T} v_1$$

Einsetzen in die Krümmungsfunktion:

$$\begin{aligned}
\kappa(\alpha) &= \left[\ddot{l}_2 + (\dot{l}_2)^2\right] \dot{l}_1 - \left[\ddot{l}_1 + (\dot{l}_1)^2\right] \dot{l}_2 \\
&= \left[\frac{\bar{\alpha}}{\alpha} \ddot{x}_\alpha^\mathsf{T} v_1 + \frac{\bar{\alpha}}{\alpha} \dot{x}_\alpha^\mathsf{T} \dot{v}_1 - \frac{1}{\alpha^2} \ddot{x}_\alpha^\mathsf{T} v_1 + (\frac{\bar{\alpha}}{\alpha} v_1^\mathsf{T} \dot{x}_\alpha)^2\right] (-v_1^\mathsf{T} \dot{x}_\alpha) \\
&- \left[-\ddot{x}_\alpha^\mathsf{T} v_1 - \dot{x}_\alpha^\mathsf{T} \dot{v}_1 + (v_1^\mathsf{T} \dot{x}_\alpha)^2\right] (\frac{\bar{\alpha}}{\alpha} v_1^\mathsf{T} \dot{x}_\alpha) \\
&= \frac{1}{\alpha^2} (\dot{x}_\alpha^\mathsf{T} v_1)^2 - [\left(\frac{\bar{\alpha}}{\alpha}\right)^2 + \frac{\bar{\alpha}}{\alpha}] (\dot{x}_\alpha^\mathsf{T} v_1)^3 \\
&= \left(\frac{\dot{x}_\alpha^\mathsf{T} v_1}{\alpha}\right)^2 \left(1 - \frac{\bar{\alpha}\alpha}{\alpha} \dot{x}_\alpha^\mathsf{T} v_1\right)
\end{aligned}$$

Weiterhin gilt:

$$\dot{x}_\alpha = \frac{1}{\alpha} S_\alpha^{-1} v_1$$

$\Rightarrow$

$$\frac{1}{\alpha} \dot{x}_\alpha^\mathsf{T} v_1 = \frac{1}{\alpha^2} v_1^\mathsf{T} S_\alpha^{-1} v_1 = \frac{1}{\alpha^2} (x_\alpha - \mu_1)^\mathsf{T} \Sigma_1^{-1} S_\alpha^{-1} \Sigma_1^{-1} (x_\alpha - \mu_1)$$

$$\begin{aligned}
x_\alpha - \mu_1 &= S_\alpha^{-1} [\bar{\alpha} \Sigma_1^{-1} \mu_1 + \alpha \Sigma_2^{-1} \mu_2] - S_\alpha^{-1} S_\alpha \mu_1 \\
&= S_\alpha^{-1} [\bar{\alpha} \Sigma_1^{-1} \mu_1 + \alpha \Sigma_2^{-1} \mu_2 - \bar{\alpha} \Sigma_1^{-1} \mu_1 - \alpha \Sigma_2^{-1} \mu_1] \\
&= \alpha S_\alpha^{-1} \Sigma_2^{-1} (\mu_2 - \mu_1)
\end{aligned}$$

$\Rightarrow$

$$\frac{1}{\alpha}\dot{x}_\alpha^\mathsf{T} v_1 = (\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}(\mu_2 - \mu_1)$$

□

2.2. Schranken für Modenanzahl

In diesem Abschnitt diskutieren wir einige Aussagen über die Anzahl der Moden von elliptischen Mischungen. Zunächst sei hier ein Ergebnis von M. Carreira-Periñán und C. Williams vorgestellt, das eine Schranke für spezielle Gauss-Mischungen aus **K** Komponenten liefert.

Satz 2.2.1. *Sei* $g(x;\mu_1,\ldots,\mu_K,\Sigma_1,\ldots,\Sigma_K,\pi_1,\ldots,\pi_{K-1}) = \sum_{i=1}^{K}\pi_i{\cdot}k_i{\cdot}\varphi_i((x-\mu_i)^\mathsf{T}\Sigma_i^{-1}(x-\mu_i))$ *eine Gauss-Mischung im* $\mathbb{R}^D$*. Die Mischung hat maximal K Moden, falls eine der folgenden Bedingungen erfüllt ist:*

1. $D = 1$ *(eindimensionale Mischung)*
2. $D \geq 1$ *und die Kovarianzmatrizen sind alle gleich:* $\Sigma_i = \Sigma$ *(homoskedastische Mischung)*
3. $D \geq 1$ *und die Kovarianzmatrizen sind vielfache der Einheitsmatrix:* $\Sigma_i = \sigma_i I$ *(isotropische Mischung)*

Der Beweis kann in [3] nachgelesen werden[1].

□

Obiger Satz liefert ein Schranke für Mischungen aus beliebig vielen Komponenten, aber mit Restriktionen an D bzw. an die Kovarianzstruktur der Mischung (Restriktionen an Kovarianzmatrizen werden wir in diesem Kapitel später nochmal aufgreifen). Jetzt beschränken wir uns wieder nur auf den Fall $K = 2$, machen dafür aber keine Restriktionen an die Parameter.

[1] Bewiesen wird nur die erste Aussage. Die Aussagen 2 und 3 sind Vermutungen, die später widerlegt wurden.

Wir verwenden zunächst die Form der Krümmungsfunktion aus Satz 2.1.16 um eine Schranke für die Modenanzahl einer allgemeinen Gauss-Mischung aus zwei Komponenten herzuleiten, danach geben wir eine Schranke für t-Mischungen an. Anschließend gehen wir auf Parametrisierungen und Restriktionen von Kovarianzmatrizen ein, die in [9] vorgeschlagen werden und leiten für spezielle Gauss-Mischungen eine weitere verschärfte Schranke her. Der Zusammenhang zwischen Π' bzw. der Krümmungsfunktion und der Modenanzahl von g ist im Korollar 2.1.15 beschrieben.

Im Folgenden bezeichne $\mathbb{R}[X]_k$ den Vektorraum der Polynome vom Grad kleiner oder gleich k und $\mathbb{R}[X]_k^{N\times M}$ eine $N \times M$-Matrix mit Einträgen aus $\mathbb{R}[X]_k$.

Wir benutzen folgendes Lemma:

Lemma 2.2.2. *Sei $A \in \mathbb{R}[X]_k^{D\times D}$ invertierbar $\Rightarrow$ für die Elemente der Inversen $A^{-1} = (r_{i,j})_{1\leq i,j\leq D}$ gilt:*

$$r_{i,j} = \frac{\zeta_{i,j}}{\nu} \tag{2.23}$$

und

$$\zeta_{i,j} \in \mathbb{R}[X]_{Dk-k}$$

$$\nu \in \mathbb{R}[X]_{Dk}.$$

Beweis: Zunächst errinern wir uns an die Leibnizformel für Determinante und stellen fest, daß für ein $A \in \mathbb{R}[X]_k^{D\times D}$

$$det(A) \in \mathbb{R}[X]_{Dk}$$

gilt. Jetzt verwenden wir die Cramerformel für die Inverse:

$$A^{-1} = \frac{1}{det(A)} A^{adj},$$

wobei A^{adj} die adjunkte Matrix von A ist, deren Elemente signierte Determinanten von Teilmatrizen von A sind, die durch Streichen von je einer Zeile und Spalte aus A entstehen. Es gilt:

$$A \in \mathbb{R}[X]_k^{D\times D} \Rightarrow A^{adj} \in \mathbb{R}[X]_k^{D-1\times D-1}$$

und

$$det(A^{adj}) \in \mathbb{R}[X]_{Dk-k}.$$

Die Aussage folgt mit $\nu = det(A)$ und $\zeta_{i,j} = A^{adj}_{i,j}$. □

Satz 2.2.3 (Schranke für Gauss-Mischungen)**.** *Sei* $g(x; \mu_1, \mu_2, \Sigma_1, \Sigma_2, \pi) = \pi\phi(x, \mu_1, \Sigma_1) + (1-\pi)\phi(x, \mu_2, \Sigma_2)$ *eine Gauss-Mischung. Dann hat* g *maximal*

$$1 + \frac{1}{2}(3D - 3D \mod 2)$$

Moden.

Beweis:
Idee: Wir zeigen, dass die Krümmungsfunktion κ_{gauss} (2.21) maximal $3D - 3D$ mod 2 Nullstellen im Interval $[0, 1]$ haben kann, daraus folgt, dass $\Pi(\alpha)$ maximal $3D - 3D$ mod 2 Oszillationen haben kann.
Behauptung: $\kappa(\alpha)$ hat maximal $E := 3D - 3D$ mod 2 Nullstellen im Intervall $[0, 1]$.
Beweis der Behauptung: Nach (2.1.16) gilt :

$$\kappa_{gauss}(\alpha) = p(\alpha)^2[1 - \alpha\bar{\alpha}p(\alpha)]$$

mit

$$p(\alpha) = (\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}(\mu_2 - \mu_1).$$

Aufgrund der positiven Definitheit der Matrizen Σ_1, Σ_2 und folglich ihrer Inversen und S_α^{-1}, ist $p(\alpha)$ stets positiv (da $\mu_1 \neq \mu_2$). Somit sind die Nullstellen von $\kappa(\alpha)$ identisch mit den Nullstellen von

$$q(\alpha) := 1 - \alpha\bar{\alpha}(\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}(\mu_2 - \mu_1). \tag{2.24}$$

Wegen $q(0) = q(1) = 1$ hat q und damit auch κ eine gerade Anzahl von Nullstellen im Intervall $[0, 1]$. Diese Aussage haben wir im allgemeinen Fall im Korollar 2.1.15 bewiesen.

Nun wenden wir das Lemma 2.2.2 auf S_α an: $S_\alpha \in \mathbb{R}[X]_1 \overset{(2.23)}{\Rightarrow}$ die Elemente $r_{i,j}$ von S_α^{-1} haben die Form

$$r_{i,j} = \frac{\zeta_{i,j}}{\nu}$$

mit

$$\zeta_{i,j} \in \mathbb{R}[X]_{D-1}$$

$$\nu \in \mathbb{R}[X]_D.$$

und folglich gilt für das Produkt

$$M(\alpha) := \Sigma_1^{-1} S_\alpha^{-1} \Sigma_2^{-1} S_\alpha^{-1} \Sigma_2^{-1} S_\alpha^{-1} \Sigma_1^{-1} = \frac{1}{\nu(\alpha)^3} \Sigma_1^{-1} \, Z \, \Sigma_2^{-1} \, Z \, \Sigma_2^{-1} \, Z \, \Sigma_1^{-1}$$

Mit $Z := (\zeta_{i,j}(\alpha))_{1 \le i,j \le D}$.

Die höchste auftretende α-Potenz in ν^3 ist kleiner gleich $3D$. Die Einträge der Matrix $\Sigma_1^{-1} \, Z \, \Sigma_2^{-1} \, Z \, \Sigma_2^{-1} \, Z \, \Sigma_1^{-1}$ sind Polynome vom Grad kleiner gleich $3D - 3$. Insgesamt ergibt das, dass die Größe

$$\frac{\varsigma(\alpha)}{\nu(\alpha)^3} := (\mu_2 - \mu_1)^\mathsf{T} M(\alpha)(\mu_2 - \mu_1)$$

eine rationale Funktion ist, mit $grad(\varsigma) \le 3D - 3$ und $grad(\nu^3) \le 3D$.
Damit lautet die Gleichung $q(\alpha) = 0$:

$$\begin{aligned} 1 - \alpha\bar{\alpha}\frac{\varsigma(\alpha)}{\nu(\alpha)^3} &= 0 \\ \Leftrightarrow \frac{\nu(\alpha)^3 - \alpha\bar{\alpha}\varsigma(\alpha)}{\nu(\alpha)^3} &= 0 \\ \Leftrightarrow \nu(\alpha)^3 - \alpha\bar{\alpha}\varsigma(\alpha) &= 0. \end{aligned}$$

Der letzte Äquivalenzpfeil ist berechtigt, da $\nu(\alpha)$ Determinante einer invertierbaren Matrix ist und deswegen $\neq 0$ für alle α.

Der Grad des Polynoms $\nu(\alpha)^3 - \alpha\bar{\alpha}\varsigma(\alpha)$ ist kleiner gleich $3D$ und der Fundamentalsatz der Algebra besagt, dass dieses Polynom maximal $3D$ Nullstellen haben kann. Da wir zusätzlich wissen, dass q im Intervall $[0, 1]$ eine gerade Anzahl von Nullstellen haben muß, folgt: q hat maximal $3D - 3D$ mod 2 Nullstellen im Intervall $[0, 1]$.

Ende Beweis der Behauptung □

Wir wissen jetzt, dass $\Pi(\alpha)$ maximal $E = 3D - 3D$ mod 2 Oszillationen hat. Daraus folgt, daß die Gleichung $\Pi(\alpha) = \pi$ für ein festes π, maximal $E + 1$ Lösungen $\alpha_1, ..., \alpha_{E+1}$ haben kann. Aus dem Verlauf der Funktion Π folgt: α_1 liefert einen Modus, α_2 liefert einen Sattelpunkt,...,α_E liefert einen Sattelpunkt, α_{E+1} liefert einen Modus (siehe auch

Korollar 2.1.15). Insgesamt gibt es also $1 + \frac{E}{2}$ Moden. □

Beispiel 2.2.4. *Die Schranke für die Modenanzahl wächst also linear mit der Dimension D mit Proportionskonstante $\frac{3}{2}$.*

Für verschiedene Dimensionen D ergeben sich folgende Schranken für die Modenanzahl der Gauss-Mischung:

$D = 1 \Rightarrow$ *Schranke = 2*
$D = 2 \Rightarrow$ *Schranke = 4*
$D = 3 \Rightarrow$ *Schranke = 5*
$D = 4 \Rightarrow$ *Schranke = 7.*

Die Schranke ist für D=1 und D=3 scharf, es gibt Beispiele, wo sie angenommen wird. Offensichtlich ist $2, 4, 5, 7, 8, 10, ...$ die Folge aller natürlichen Zahlen, die nicht durch 3 teilbar sind, in aufsteigender Reihenfolge.

Bemerkung 2.2.5. *Diese Schranke läßt sich auch direkt mit der 1.Ableitung der Funktion Π herleiten:*
Es gilt

$$\Pi(\alpha) = \left[1 + \frac{\alpha c_1 \varphi_1(\delta_1)}{\bar{\alpha} c_2 \varphi_2(\delta_2)}\right]^{-1},$$

wobei

$$\begin{aligned} \varphi_i(t) &= e^{-\frac{1}{2}t} \\ \delta_i &= (x_\alpha - \mu_i)^\mathsf{T} \Sigma_i^{-1} (x_\alpha - \mu_i) \\ c_i &= \frac{1}{|\Sigma_i|} \end{aligned}$$

siehe oben.

Differentiation nach α liefert:

$$\Pi'(\alpha) = -\left[1 + \frac{\alpha c_1 \varphi_1(\delta_1)}{\bar{\alpha} c_2 \varphi_2(\delta_2)}\right]^{-2} \left[\frac{c_1 \varphi_1(\delta_1)}{\bar{\alpha}^2 c_2 \varphi_2(\delta_2)} + \frac{\alpha c_1 [\Delta_1 \varphi_1'(\delta_1) \varphi_2(\delta_2) - \Delta_2 \varphi_2'(\delta_2) \varphi_1(\delta_1)]}{\bar{\alpha} c_2 \varphi_2(\delta_2)^2}\right]$$

Wir interessieren uns für die Nullstellen des Terms. Der Faktor $\left[1+\frac{\alpha c_1\varphi_1(\delta_1)}{\bar{\alpha}c_2\varphi_2(\delta_2)}\right]^{-2}$ *ist echt positiv, da sowohl die Konstanten* c_i*, als auch die Generatorfunktionen* φ_i *echt positiv sind. Deshalb brauchen wir diesen Faktor nicht weiter zu betrachten und untersuchen den Faktor*

$$Z(\alpha) := \frac{c_1\varphi_1(\delta_1)}{c_2\bar{\alpha}^2\varphi_2(\delta_2)} + \frac{\alpha c_1[\Delta_1\varphi_1'(\delta_1)\varphi_2(\delta_2) - \Delta_2\varphi_2'(\delta_2)\varphi_1(\delta_1)]}{\bar{\alpha}c_2\varphi_2(\delta_2)^2}$$

auf die Nullstellen.
Es gilt: $\varphi' = -\frac{1}{2}\varphi$*. Einsetzen in Z liefert:*

$$Z(\alpha) = \frac{c_1\varphi_1(\delta_1)}{c_2\bar{\alpha}^2\varphi_2(\delta_2)} + \frac{\alpha c_1[(-\frac{\Delta_1}{2})\varphi_1(\delta_1)\varphi_2(\delta_2) + \frac{\Delta_2}{2}\varphi_2(\delta_2)\varphi_1(\delta_1)]}{\bar{\alpha}c_2\varphi_2(\delta_2)^2}$$

$$\overset{\Delta 2=-\frac{\bar{\alpha}}{\alpha}\Delta_1}{=} \frac{c_1\varphi_1(\delta_1)}{c_2\bar{\alpha}^2\varphi_2(\delta_2)} - \frac{\Delta_1}{2}\frac{\alpha c_1\varphi_1(\delta_1)\varphi_2(\delta_2)(1+\frac{\bar{\alpha}}{\alpha})}{c_2\bar{\alpha}\varphi_2(\delta_2)^2}.$$

Es folgt $Z(\alpha) = 0$

$\Leftrightarrow$

$$c_1\varphi_1(\delta_1)\varphi_2(\delta_2) - \frac{\Delta_1}{2}\bar{\alpha}\alpha c_1\varphi_1(\delta_1)\varphi_2(\delta_2)\left(1+\frac{\bar{\alpha}}{\alpha}\right) = 0$$

$\Leftrightarrow$

$$\varphi_1(\delta_1)\varphi_2(\delta_2)c_1(1 - \frac{\Delta_1}{2}\bar{\alpha}\alpha\frac{1}{\alpha}) = 0$$

$\Leftrightarrow$

$$1 - \frac{\Delta_1}{2}\bar{\alpha} = 0. \tag{2.25}$$

Nun gilt

$$\begin{aligned}\Delta_1 &= 2(\mu_2-\mu_1)^\mathsf{T}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}S_\alpha^{-1}v_1, \text{ (siehe Lemma 2.1.10)}\\ &= 2(\mu_2-\mu_1)^\mathsf{T}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}S_\alpha^{-1}\Sigma_1^{-1}S_\alpha^{-1}\left[\bar{\alpha}\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2 - S_\alpha\mu_1\right]\\ &= 2\alpha(\mu_2-\mu_1)^\mathsf{T}\Sigma_2^{-1}S_\alpha^{-1}\Sigma_1^{-1}S_\alpha^{-1}\Sigma_1^{-1}S_\alpha^{-1}\Sigma_2^{-1}(\mu_2-\mu_1).\end{aligned} \tag{2.26}$$

Wir wissen bereits $S_\alpha^{-1} = \frac{1}{\nu}(r_{i,j})_{i,j}$, *mit* $\nu \in \mathbb{R}[X]_D$, $r_{i,j} \in \mathbb{R}[X]_{D-1}$. *Also gilt*

$$\frac{\Delta_1}{2} = \frac{\varsigma_\Delta}{\nu_\Delta}, \tag{2.27}$$

wobei $\varsigma_\Delta \in \mathbb{R}[X]_{3D-2}$ *und* $\nu_\Delta \in \mathbb{R}[X]_{3D}$.
Also

$$(2.25) \Leftrightarrow \nu_\Delta - \bar{\alpha}\varsigma_\Delta = 0.$$

Der Grad des Polynoms $\nu_\Delta - \bar{\alpha}\xi_\Delta$ *ist kleiner gleich* $3D$ *und dieselben Überlegungen wie im Satz 2.2.3 führen uns auf das Ergebnis.*

Ende der Bemerkung.

Im Falle der t-Verteilung betrachten wir direkt die 1.Ableitung von Π um eine Schranke für die Modenanzahl zu erhalten.

Satz 2.2.6 (Schranke für t-Mischungen)**.** *Sei* $g(x; \mu_1, \mu_2, \Sigma_1, \Sigma_2, \pi) = \pi f_1(x, \mu_1, \Sigma_1) + (1-\pi) f_2(x, \mu_2, \Sigma_2)$ *eine t-Mischung. Dann hat* g *maximal*

$$1 + \frac{7D + 2 - 7D \bmod 2}{2}.$$

Moden.

Beweis: In der obigen Bemerkung haben wir gezeigt, dass die Nullstellen von Π' dieselben sind wie die von Z mit

$$Z(\alpha) = \frac{c_1\varphi_1(\delta_1)}{c_2\bar{\alpha}^2\varphi_1(\delta_1)} + \frac{\alpha c_1[\Delta_1\varphi_1'(\delta_1)\varphi_2(\delta_2) - \Delta_2\varphi_2'(\delta_2)\varphi_1(\delta_1)]}{\bar{\alpha}c_2\varphi_2(\delta_2)^2}.$$

Für die Generatorfunktionen der t-Verteilung gilt:

$$\varphi_i'(t) = -\frac{n_i + D}{2}(1 + \frac{1}{n_i}t)^{-1}\varphi_i(t),\ i = 1, 2.$$

Eisetzen in Z liefert:

$$
\begin{aligned}
&Z(\alpha)\\
&= \frac{c_1\varphi_1(\delta_1)}{c_2\bar{\alpha}^2\varphi_2(\delta_2)} + \frac{\alpha c_1\left(\Delta_1[-\frac{n_1+D}{2}(1+\frac{1}{n_1}\delta_1)^{-1}\varphi_1(\delta_1)]\varphi_2(\delta_2) + \Delta_2[\frac{n_2+D}{2}(1+\frac{1}{n_2}\delta_2)^{-1}\varphi_2(\delta_2)]\varphi_1(\delta_1)\right)}{\bar{\alpha}c_2\varphi_2(\delta_2)^2}\\
&\overset{\Delta 2=-\frac{\bar{\alpha}}{\alpha}\Delta_1}{=} \frac{c_1\varphi_1(\delta_1)}{c_2\bar{\alpha}^2\varphi_2(\delta_2)} - \frac{\alpha c_1\Delta_1\varphi_1(\delta_1)\varphi_2(\delta_2)\left(\frac{n_1+D}{2}(1+\frac{1}{n_1}\delta_1)^{-1} + \frac{\bar{\alpha}}{\alpha}\frac{n_2+D}{2}(1+\frac{1}{n_2}\delta_2)^{-1}\right)}{\bar{\alpha}c_2\varphi_2(\delta_2)^2}
\end{aligned}
$$

Somit ist $Z(\alpha) = 0$ äquivalent zu

$$c_1\varphi_1(\delta_1)\varphi_2(\delta_2) - \bar{\alpha}\alpha c_1\Delta_1\varphi_1(\delta_1)\varphi_2(\delta_2)\left(\frac{n_1+D}{2}(1+\frac{1}{n_1}\delta_1)^{-1} + \frac{\bar{\alpha}}{\alpha}\frac{n_2+D}{2}(1+\frac{1}{n_2}\delta_2)^{-1}\right) = 0.$$

das ist äquivalent zu

$$1 - \bar{\alpha}\alpha\Delta_1\left(\frac{n_1+D}{2}(1+\frac{1}{n_1}\delta_1)^{-1} + \frac{\bar{\alpha}}{\alpha}\frac{n_2+D}{2}(1+\frac{1}{n_2}\delta_2)^{-1}\right) = 0.$$

Wir bringen den Ausdruck auf gemeinsamen Nenner und Setzen den Zähler gleich 0:

$$(1+\frac{1}{n_1}\delta_1)(1+\frac{1}{n_2}\delta_2) - \bar{\alpha}\alpha\Delta_1\left(\frac{n_1+D}{2}(1+\frac{1}{n_2}\delta_2) + \frac{\bar{\alpha}}{\alpha}\frac{n_2+D}{2}(1+\frac{1}{n_1}\delta_1)\right) = 0 \quad (2.28)$$

Zur Erinnerung:

$$\delta_i = (x_\alpha - \mu_i)^\mathsf{T}\Sigma_i^{-1}(x_\alpha - \mu_i)$$

und

$$x_\alpha = S_\alpha^{-1}(\bar{\alpha}\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2).$$

Die Betrachtung der auftretenden α-Potenzen liefert

$$x^*(\alpha) = \frac{\varsigma^*}{\nu^*},$$

mit

$$\varsigma^*,\ \nu^* \in R[X]_D.$$

Somit gilt

$$\delta_i = \frac{\varsigma_i}{\nu_i},$$

mit

$$\varsigma_i,\ \nu_i \in R[X]_{2D}.$$

Bereits in der vorangehenden Bemerkung haben wir gezeigt:

$$\Delta_1 = \frac{\varsigma_\Delta}{\nu_\Delta},$$

mit

$$\varsigma_\Delta \in \mathbb{R}[X]_{3D-2} \text{ und } \nu_\Delta \in \mathbb{R}[X]_{3D}.$$

Wir bringen die gebrochen-rationale Funktion auf der linken Seite von (2.28) auf gemeinsamen Nenner und erhalten im Zähler ein Polynom, dessen Grad nicht größer ist als der von

$$\alpha^2 \nu_1 \nu_2 \nu_\Delta.$$

Der Grad ist somit maximal $7D+2$. Die Gleichung (2.28) hat also maximal $7D+2-(7D \mod 2)$ Lösungen im Intervall $[0,1]$.
Dieselbe Überlegungskette, wie im Satz 2.2.3, liefert uns folgende Schranke für die Modenanzahl einer t-Mischung aus zwei Komponenten:

$$1 + \frac{7D + 2 - 7D \; mod \; 2}{2}.$$

□

Beispiel 2.2.7. *Die Schranke für die Modenanzahl wächst also linear mit der Dimension D mit Proportionskonstante $\frac{7}{2}$.*

Für verschiedene Dimensionen D ergeben sich folgende Schranken für die Modenanzahl der t-Mischung:

$D = 1 \Rightarrow$ *Schranke = 5*
$D = 2 \Rightarrow$ *Schranke = 9*
$D = 3 \Rightarrow$ *Schranke = 12*
$D = 4 \Rightarrow$ *Schranke = 16.*

Wie weit diese Schranken von tatsächlichen Maximalwerten liegen ist uns nicht bekannt.

Im Fall von gleichen Kovarianzmatrizen läßt sich die Schranke etwas verbessern.

Korollar 2.2.8. *Sei $g(x; \mu_1, \mu_2, \Sigma_1, \Sigma_2, \pi)$ eine t-Mischung aus 2 Komponenten und es gelte $\Sigma_1 = \Sigma_2$. Dann hat g maximal*

$$2D + 2$$

Moden.

Beweis:
In diesem Fall gilt $S_\alpha^{-1} = \Sigma_1 = \Sigma_2$ und somit

$$\Delta_1 = 2\alpha(\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2 - \mu_1)$$

(siehe 2.26).

Einsetzen in (2.28) liefert folgende Gleichung für Nullstellen von Π':

$$(1+\frac{1}{n_1}\delta_1)(1+\frac{1}{n_2}\delta_2)-2\bar{\alpha}\alpha^2(\mu_2-\mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2-\mu_1)\left(\frac{n_1+D}{2}(1+\frac{1}{n_2}\delta_2)+\frac{\bar{\alpha}}{\alpha}\frac{n_2+D}{2}(1+\frac{1}{n_1}\delta_1)\right)=0$$

Wir bringen den Ausdruck auf gemeinsamen Nenner, setzen den Zähler gleich 0 und erhalten eine polynomielle Gleichung vom Grad kleiner gleich $grad(\alpha^3\nu_1\nu_2)$. Wobei $grad(\nu_i) \leq 2D$ für $i = 1, 2$ (siehe Beweis von Satz 2.2.6). Wegen $grad(\alpha^3\nu_1\nu_2) \leq 3 + 4D$ hat Π' maximal $4D + 2$ Nulstellen im Intervall $[0, 1]$ und somit maximal $1 + \frac{4D+2}{2} = 2D + 2$ Moden.

□

2.3. Restriktionen an Kovarianzmatrizen in Mischungsmodellen

Im nächsten Kapitel werden wir uns mit modellbasierter Clusteranalyse beschäftigen. In diesem Zusammenhang tauchen oft Restriktionen an Kovarianzmatrizen einzelner Mischungskomponenten auf. Durch diese Restriktionen verringert man die Anzahl der unbekannten Parameter, die mit dem EM-Algorithmus geschätzt werden müßen. Das R-Packet **Mclust** stellt Möglichkeiten zur Verfügung derartige Restriktionen festzulegen. (siehe nächstes Kapitel).
Daten, die von Mischungen elliptischer Verteilungen produziert werden, sind charakteresiert durch ellipsoide Cluster die um die Zentren μ_i konzentriert sind, mit steigender Dichte in der Nähe der Mittelpunkte μ_i.
Die geometrischen Eigenschaften der Cluster wie Form, Ausrichtung und Volumen werden durch die Kovarianzmatrizen festgelegt, die Kovarianzmatrizen werden wiederum durch $\frac{D(D+1)}{2}$ reelle Zahlen festgelegt, die im Grunde frei gewählt werden können solange Σ_i positiv definit bleibt.

Häufig beschränkt man sich auf die Fälle $\Sigma_i = \lambda I$, was zu sphärischen Clustern mit gleichen Radien führt, oder $\Sigma_i = \Sigma$, was zu gleichen Clustern beliebiger Form führt. Im ersten Fall benötigt man einen reellen Parameter um die Kovarianz-Struktur der Mischung festzulegen im zweiten sind es $\frac{D(D+1)}{2}$ reelle Parameter und im allgemeinen Fall einer elliptischen Mischung mit K Komponenten sind es $\frac{KD(D+1)}{2}$ reelle Parameter.

Eine flexible Form Kovarianzmatrizen zu parametrisieren ist die Diagonalisierung:

$$\Sigma_i = \lambda_i D_i A_i D_i^{\mathsf{T}}, \tag{2.29}$$

(siehe [9]).
Dabei ist D_i eine orthogonale Matrix deren Spalten Eigenvektoren von Σ_i sind, A_i eine Diagonalmatrix, deren Elemente die durch den größten Eigenwert geteilte Eigenwerte von Σ_i sind λ_i der größte Eigenwert.
Mit dieser Darstellung lassen sich Restriktionen an einzelne geometrische Eigenschaften der Cluster definieren - λ_i steuert das Volumen der Komponente i, D_i die Ausrichtung, A_i die Form.
Wenn beipielsweise der größte Eigenwert von Σ_i wesentlich größer ist als die anderen Eigenwerte, dann wird sich der i'te Cluster entlang der Hauptachse im $\mathbb{R}^D$ erstrecken. Wenn dagegen zwei große Eigenwerte von Σ_i die anderen dominieren und, wird der i'te Cluster im Wesentlichen in dem durch die beiden zugehörigen Eigenvektoren aufgespannten Unterraum liegen. Falls der größte und der kleinste Eigenwert von Σ_i ungefähr den gleichen Betrag haben, ergibt sich ein sphärischer Cluster.

Mit diesem Ansatz lassen sich verschiedene Arten von Restriktionen definieren:

- $\Sigma_i = \lambda I$
- $\Sigma_i = \lambda U A U^{\mathsf{T}}$
- $\Sigma_i = \lambda_i I$
- $\Sigma_i = \lambda_i A_i$
- $\Sigma_i = \lambda_i U_i A U_i^{\mathsf{T}}$
- $\Sigma_i = \lambda_i U A_i U^{\mathsf{T}}$
- usw.

Im nächsten Abschnitt diskutieren wir Bimodalitätsbedingungen für Mischungen, für die $\Sigma_i = \lambda U A U^\mathsf{T}$ bzw. $\Sigma_i = \lambda_i U A U^\mathsf{T}$ gilt.

2.4. Bimodalitätsbedingungen für spezielle Gauss-Mischungen

In diesem Abschnitt untersuchen wir einige Aussagen über Modalität von speziellen Gauss-Mischungen, die Ray und Lindsay in [15] vorgestellt haben.

Korollar 2.4.1. *Sei $g(x;\theta)$ eine Mischung aus zwei Gauß-Komponenten mit gleichen Kovarianzmatrizen $\Sigma_1 = \Sigma_2$. Dann gilt (siehe (2.22) und (2.24)):*

$$q(\alpha) = 1 - \alpha(1-\alpha)(\mu_2 - \mu_1)^\mathsf{T}\Sigma^{-1}(\mu_2 - \mu_1),$$

$q(\alpha)$ und folglich $\kappa(\alpha)$ sind quadratische Polynome und besitzen zwei Nullstellen falls

$$(\mu_2 - \mu_1)^\mathsf{T}\Sigma^{-1}(\mu_2 - \mu_1) > 4$$

und die Mischung g ist dann bimodal, falls die Mischungsproportion $\pi \in (\pi_1, \pi_2)$, wobei

$$\frac{1}{\pi_i} = 1 + \frac{\alpha_i \phi_1(\alpha_i)}{\bar{\alpha}_i \phi_2(\alpha_i)}$$

und α_i sind die beiden Nullstellen von $q(\alpha)$ im Intervall $[0,1]$.

Beweis: die Aussage ergibt sich durch Ausrechnen der Krümmungsfunktion unter Benutzung von (2.22) und $\Sigma_1 = \Sigma_2$. □

Lemma 2.4.2. *Sei $g(x;\theta)$ eine Mischung aus zwei Gauß-Komponenten mit proportionallen Kovarianzmatrizen $\Sigma_2 = \sigma^2\Sigma_1$, für ein $\sigma^2 > 0$. Dann hat $\kappa(\alpha)$ dieselben Nullstellen in $[0,1]$ wie das Polynom*

$$q_1(\alpha) := (\sigma^2(1-\alpha) + \alpha)^3 - \alpha(1-\alpha)\mu^2\sigma^2$$

wobei $\mu^2 := (\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2 - \mu_1)$.

Beweis:

$\Sigma_2 = \sigma^2\Sigma_1 \Rightarrow S_\alpha = \Sigma_1^{-1}((1-\alpha) + \frac{\alpha}{\sigma^2})$. Daraus ergibt sich

$$
\begin{aligned}
q(\alpha) &= 1 - \alpha(1-\alpha)\frac{(\mu_2-\mu_1)^\mathsf{T}\Sigma_1^{-1}\Sigma_1\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}\Sigma_1\Sigma_1^{-1}(\mu_2-\mu_1)}{(1-\alpha)+\frac{\alpha}{\sigma^2})^3} \\
&= 1 - \alpha(1-\alpha)\frac{(\mu_2-\mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2-\mu_1)}{\sigma^4(1-\alpha+\frac{\alpha}{\sigma^2})^3} \\
&= 1 - \alpha(1-\alpha)\frac{\mu^2\sigma^2}{(\sigma^2(1-\alpha)+\frac{\alpha}{\sigma^2})^3} \\
&= \frac{1}{(\sigma^2(1-\alpha)+\frac{\alpha}{\sigma^2})^3} q_1(\alpha)
\end{aligned}
$$

Das Polynom $(\sigma^2(1-\alpha)+\frac{\alpha}{\sigma^2})^3$ hat im Intervall $[0,1]$ keine Nullstellen, somit sind die Nullstellen von $q(\alpha)$ identisch mit denen von $q_1(\alpha)$. □

Für den nächsten Korollar brauchen wir eine Definition und ein Hilfslemma, das hier ohne Beweis gegeben wird:

Definition 2.4.3. Sei

$$f(x) = ax^3 + bx^2 + cx + d,$$

mit $a \neq 0$ ein Polynom 3. Grades. und

$$D := 18abcd - 4b^3d + b^2c^2 - 4ac^3 - 27a^2d^2.$$

D heißt *Diskriminante* von f.

Lemma 2.4.4. *Sei*

$$f(x) = ax^3 + bx^2 + cx + d,$$

mit $a \neq 0$ ein Polynom 3. Grades. Sei D die Diskriminante von f. Dann gilt

- *Falls $D < 0$ besitzt f eine reelle und zwei komplexe Nullstellen.*
- *Falls $D > 0$ besitzt f drei verschiedene reelle Nullstellen.*
- *Falls $D = 0$ bestizt f mehrfache reelle Nulstellen (max. 2)*

Korollar 2.4.5. *Sei $g(x;\theta)$ eine Mischung aus zwei Gauß-Komponenten mit proportionalen Kovarianzmatrizen $\Sigma_2 = \sigma^2\Sigma_1$.*

(a) Die Dichte g ist unimodal für jede Mischungsproportion π falls

$$(\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2 - \mu_1) \leq \frac{2(1-\sigma^2+\sigma^4)^{\frac{3}{2}} - (2\sigma^6 - 3\sigma^4 - 3\sigma^2 + 2)}{\sigma^2}$$

(b) Falls die Parameter obige Bedingung nicht erfüllen, ist g bimodal genau dann, wenn $\pi \in (\pi_1, \pi_2)$, wobei

$$\frac{1}{\pi_i} = 1 + \frac{\alpha_i \phi_1(\alpha_i)}{\bar{\alpha}_i \phi_2(\alpha_i)}$$

und α_i sind die beiden Nullstellen von $q(\alpha)$ in $[0,1]$ sind.

Beweis:
Wir zeigen zuerst den Teil (a). Mit Hilfe von Lemma 2.4.2 folgt, dass die Nullstellen von $\kappa(\alpha)$ dieselben sind wie die des Polynoms $q_1(\alpha) = (\sigma^2(1-\alpha)+\alpha)^3 - \alpha(1-\alpha)\mu^2\sigma^2$. Das Polynom $q_1(\alpha)$ hat drei reelle Nullstellen, falls die Diskriminante von q_1 positiv ist (siehe Lemma oben). Das ist äquivalent zu:

$$\begin{aligned} & s(\mu) = \mu^4\sigma^2 - \mu^2(-4\sigma^6 + 6\sigma^4 + 6\sigma^2 - 4) - 27\sigma^2(\sigma^2-1)^2 \geq 0 \\ \Leftrightarrow\ & s(\mu) = \mu^4\sigma^2 + 2\mu^2(\sigma^2-2)(\sigma^2+1)(2\sigma^2-1) - 27\sigma^2(\sigma^2-1)^2 \geq 0. \end{aligned}$$

$s(\mu)$ ist quadratisch in μ und nur eine der beiden Nullstellen ist positiv. Diese Nullstelle lautet:

$$\mu_0^2 = \frac{2(1-\sigma^2+\sigma^4)^{\frac{3}{2}} - (2\sigma^6 - 3\sigma^4 - 3\sigma^2 + 2)}{\sigma^2}$$

Weiterhin gilt $s(0) \leq 0$. Deshalb gilt $s(\mu) \geq 0$ für $\mu^2 \geq \mu_0^2$. Das Anwenden von Satz 2.1.16 und Lemma 2.4.2 auf den Fall proportionaler Kovarianzmatrizen liefert, daß g genau dann unimodal ist, falls

$$(\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2 - \mu_1) \leq \frac{2(1-\sigma^2+\sigma^4)^{\frac{3}{2}} - (2\sigma^6 - 3\sigma^4 - 3\sigma^2 + 2)}{\sigma^2}.$$

Jetzt beweisen wir Teil (b). Die Dichte g ist genau dann bimodal für ein festes π, wenn die Gleichung $\Pi(\alpha) = \pi$ drei Lösungen hat, und das ist wiederum dann der Fall, wenn π im „Oszillationsgebiet“ der Funktion Π-liegt, also dem offenen Intervall, was durch die beiden Extremwerte von Π eingeschloßen wird. Diese Überlegung und die Formel 2.18

liefern die Aussage. □

Kapitel 3

Clusteranalyse mit Mischungsmodellen

In diesem Kapitel beschäftigen wir uns mit der Clusteranalyse auf Basis von Mischungsmodellen. Mit Clusteranalyse oder Unsupervised Learning wird der Vorgang des Entdeckens zusammenhängender Gruppen (Cluster) in Daten bezeichnet, ohne jegliche explizite Vorkenntnis über die Gruppenzugehörigkeit einzelner Punkte - im Gegensatz zu Diskriminanzanalyse oder Supervised Learning, wo eine Gruppeneinteilung bereits besteht und man neue Datenpunkte in diese Einteilung aufnehmen möchte.
Zwischen den Elementen eines Clusters soll größtmögliche Ähnlichkeit und zwischen verschiedenen Clustern größtmögliche Heterogenität bestehen. Dieser Ansatz erfordert die Definition eines Ähnlichkeits- bzw. eines Distanzmaßes zwischen Punkten. Die Wahl eines solchen Maßes ist entscheidend für den Ausgang der Clusteranalyse. In unserem Fall sind zwei Datenpunkte ähnlich zueinander, wenn ihre maximalen posteriori-Wahrscheinlichkeiten im gleichen Cluster angenommen werden (s. später).
Das Interesse an Clusteranalyse ist in letzten Jahren dank neuer Anwendungsfelder gestiegen. Diese erstrecken sich über Warenkorbanalyse in Kaufverhaltensdatenbanken, automatisierte Zeichenerkennung, Genexpressionsanalysen usw..
Die meisten Cluster-Algorithmen sind heuristischer Art und basieren auf menschlicher Intuition. Folgende Klassen von Clusteralgorithmen sind weit verbreitet:

- Hierarchische agglomerative Verfahren
- Iterative Partitionierungsverfahren
- Selbstorganisierende Merkmalskarten

Hierarchische agglomerative Verfahren fangen mit einer Clusterung an, bei der jeder Punkt ein Cluster ist und legen in jedem Iterationsschritt zwei Cluster zusammen, bis keine Verbesserung einer Zielfunktion mehr möglich ist. Häufige Zielfunktionen sind

kürzeste Distanz zwischen zwei Clusterpunkten (Single-Linkage) oder die Zunahme der Summe der quadrierten Abstände zum Mittelpunkt innerhalb eines Clusters (Ward).

Iterative Partitionierungsverfahren fangen mit einer Startclusterung an und verschieben iterativ einzelne Punkte zwischen den Clustern bis keine Verbesserung einer Zielfunktion mehr möglich ist. Ein bekannter Repräsentant dieser Verfahren ist der k-Means-Algorithmus.

Selbstorganisierende Merkmalskarten führen die Clusterung auf Basis künstlicher neuronaler Netzte durch.

Obwohl Clusteranalyse ein intensives Forschungsgebiet ist, gibt es eine Reihe offener Fragen die nicht präzise beantwortet werden können:

- Wie viele Cluster gibt es in den Daten?
- Wie werden Ausreißer behandelt?
- Wie sind die statistischen Eigenschaften der Algorithmen?

Früh wurde erkannt, daß man Clusteranalyse mit Wahrscheinlichkeitsmodellen betreiben kann. Diesen Ansatz wählen auch wir in vorliegender Arbeit. Wir legen den Daten ein endliches Mischungsmodell zugrunde:

$$X_{mix} = C^{\mathsf{T}} X = \sum_{i=1}^{K} C_i X_i.$$

Wir interpretieren die Größen folgendermassen: X_{mix} ist der Zufallsvektor, der alle Daten produziert hat, $\pi_i = \mathbb{P}(C = e_i)$, wobei e_i den $i'ten$ Einheitsvektor im $\mathbb{R}^K$ bezeichne, und $X_i = X_{mix}|C = e_i$. Der Zufallsvektor C wählt zufällig, mit Wahrscheinlichkeiten $\pi_1, ..., \pi_K$ Komponenten $1, \ldots, K$ aus und X_{mix} bedingt auf $C = e_i$ ist wie X_i verteilt. Für X_i wählen wir zunächst ein elliptisches Modell aus, z.B. die Gauss-Verteilung oder die t-Verteilung und erhalten nach dem sog. Mergen (Zusammenfassen) einiger Teilkomponenten eine Mischung X_i^*. Somit erhalten für am Ende unserer Analyse eine Mischung von Mischungen.

In unserem Modell lautet die Dichte von X_{mix}

$$g(x; \mu_1, ..., \mu_K, \Sigma_1, ..., \Sigma_K, \pi_1, ..., \pi_{K-1}, \vartheta_1, \ldots, \vartheta_K) = \sum_{i=1}^{K} \pi_i \cdot k_i \cdot \varphi_i((x-\mu_i)^\mathsf{T} \Sigma_i^{-1}(x-\mu_i)). \tag{3.1}$$

Der Parameter ϑ_i ist ein zusätzlicher Parameter der Dichte $k_i\varphi_i(\delta_i)$, wie z.B. der Freiheitsgrad n_i im Falle der t-Verteilung. Gauss-Dichten haben dagegen keine weiteren Parameter außer μ_i und Σ_i.
Es konnte gezeigt werden, daß manche heuristische Clusteralgorithmen approximative Lösungen der Mischungsmodellansätze liefern, k-Means ist beispielsweise äquivalent zur approximativen Maximierung der Klassifikations-Likelihood in Gauss-Mischungsmodellen mit $\Sigma_i = \sigma^2 I$ (s. auch [9]). In endlichen Mischungsmodellen entspricht zunächst jede Komponente einem Cluster. Zunächst, weil wir schlecht separierte Kompontenten (z.B. falls die Mischung zweier Komponenten unimodal oder sehr schwach bimodal ist) zusammenlegen werden und als eine Komponente behandeln werden.

3.1. Modellbasierte Clusteranalyse

Am Anfang der Analyse liegt uns eine Stichprobe $y_1, ..., y_N$ vor, die von X_{mix} produziert wurde und die wir in sinnvolle zusammenhängende Gruppen aufteilen wollen. Unsere Strategie besteht aus mehreren Schritten:

1. Schätzen der Modellparameter der Mischung:
   1.1 Suche nach geeigneten Startwerten für den EM-Algorithmus,
   1.2 ML-Schätzung der Parameter mit dem EM-Algorithmus mit Startwerten aus der Vorstufe.
2. Auswahl der Komponentenanzahl mit BIC.
3. Mergen schlecht separierter Komponenten.
4. Einteilung der Daten in Cluster nach maximalen a-posteriori-Wahrscheinlichkeiten.

Die einzelnen Schritte wollen wir im Folgenden besprechen.

3.1.1. Schätzten der Modellparameter

In dieser Phase der Clusteranalyse wollen wir den Parameter

$$\theta = (\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \pi_1, \ldots, \pi_{K-1}, \vartheta_1, \ldots, \vartheta_K)$$

der Mischungsdichte g schätzen. Dazu nehmen wir den Maximum-Likelihood-Schätzer, d.h.

$$\widehat{\theta} = \underset{\theta \in \Theta}{argmax}\ l_{mix}([\theta; y_1, .., y_N]), \tag{3.2}$$

wobei

$$l_{mix}([\theta; y_1, .., y_N]) = ln(\prod_{i=1}^{N} \sum_{j=1}^{K} \pi_j f_j(y_i; \theta_j)) \tag{3.3}$$

die Log-Likelihood der Stichporbe ist.

Mit $\theta_i := (\mu_i, \Sigma_i) \in \mathbb{R}^D \times \mathbb{R}^{D \times D}$ im Gauss-Fall bzw. $\theta_i := (n_i, \mu_i, \Sigma_i) \in \mathbb{R} \times \mathbb{R}^D \times \mathbb{R}^{D \times D}$ im t-Fall. Die Komponentenanzahl ist zunächst fest und wird mithilfe des BIC (siehe später) angepasst.
Wir lösen dieses Problem mit dem sogenannten Expectation-Maximization-Algorithmus, obgleich auch andere sinnvolle Alternativen, wie newtonartige Verfahren existieren. Der EM-Algorithmus ist ein iteratives Verfahren, das Startpunkte $\mu_i^{(0)}, \Sigma_i^{(0)}, \vartheta_i^{(0)}, \pi_i^{(0)}$ für $1 \leq i \leq K$ als Input benötigt. Um diese Startpunkte zu erhalten, führen wir als Vorstufe eine k-means Clusterung durch.

k-means

Der k-means Algorithmus ist ein iteratives Partitionierungsverfahren. Diesem Algorithmus liegt der quadrierte euklidische Abstand als Distanzmaß zu Grunde:

$$d(x_i, x_j) = \sum_{k=1}^{D} (x_{ik} - x_{jk})^2.$$

Der Algorithmus startet mit einer anfänglichen Wahl der K Zentroide (K wird vorgegeben) und ordnet in jedem Schritt jeden Datenpunkt dem ihm am nächsten liegenden Zentroid zu. Anschließend werden die Zentroide angepasst. In jedem Schritt erhält man also eine Clusterung aus Voronoi-Zellen mit Zentroiden als Zellenmittelpunkte.

k-Means Algorithmus

1. Input: Start-Zentroide $m_1, \ldots, m_K$
2. Für alle i: ordne den Datenpunkt x_i dem Zentroid $m_{C(i)}$ mit $C(i) := \underset{j}{argmin}||x_i - m_j||_2^2$ zu.
3. Für alle j: update Zentroide $m_j = \frac{1}{\#\{k|C(k)=j\}} \sum_{i:C(i)=j} x_i$.
4. Iteriere 2. und 3. solange, bis ein Konvergenz- oder Abbruchkriterium erfüllt ist.

Die anfängliche Wahl der Zentroide ist entscheidend für den Ausgang des k-Means-Algorithmus und somit auch für den nachgelagerten EM-Algorithmus, trotzdem erfolgt diese Wahl heuristisch oder zufällig, was zu den größten Nachteilen dieses Verfahrens zählt. Nachdem wir eine k-Means Clusterung erhalten haben, berechnen wir die Startwerte für den nachfolgenden EM-Algorithmus:

$$\mu_i^{(0)} := m_i,\ 1 \leq i \leq K \tag{3.4}$$

$$\Sigma_i^{(0)} := \frac{1}{\#C_i - 1} \sum_{j:C(j)=i} (y_j - m_j)(y_j - m_j)^\mathsf{T},\ 1 \leq i \leq K \tag{3.5}$$

$$\pi_i^{(0)} := \frac{\#C_i}{N},\ 1 \leq i \leq K, \tag{3.6}$$

wobei hier $C_i := \{k|C(k) = i\}$.
Für die Freiheitsgrade kann man entsprechende ML-Schätzungen in jeder Gruppe nehmen. (s. [14])

EM-Algorithmus

Um den EM-Algorithmus für die ML-Schätzung der Parameter im Mischungsmodell anwenden zu können, müssen wir unser Modell um sogenannte versteckte Variablen erweitern (engl. data augmentation). Zunächst aber beschreiben wir den EM-Algorithmus allgemein und beziehen ihn danach auf unser Problem.

Ausgangspunkt des EM-Algorithmus ist ein Zuvallsvektor $T = (X, Z)$ mit $X \in \mathbb{R}^{n_x}$ und $Z \in \mathbb{R}^{n_z}$ und ein Verteilungsmodell $f_T(t; \theta)$ mit unbekanntem θ. Von T liegt eine Teil-Stichprobe vor: $(x_1, \ldots, x_N)$, die zugehörigen Werte $(z_1, \ldots, z_N)$ konnten nicht beobachtet werden und heißen deshalb *versteckte Variablen*, $(x_1, \ldots, x_N)$ nennen wir *beobachtete Variablen*.

Ziel ist die ML-Schätzung von θ. Eine Optimierung der Gesamt-Log-Likelihood-Funktion $l_T(\theta; x, z)$ nach θ ist nicht möglich, da die Werte $z_1, \ldots, z_N$ nicht vorhanden sind.
Der EM-Algorithmus löst dieses Problem durch iterative Vorgehensweise, wo anstatt $l_T(\theta; x, z)$ der bedingte Erwartungswert der Gesamt-Log-Likelihood-Funktion bezüglich $f_{T_{\theta^{(j)}}}$ für eine Schätzung $\theta^{(j)}$

$$Q(\theta, \theta^{(j)}) := \mathbb{E}[l_T(\theta; X, Z)|X = x, \theta^{(j)}] \tag{3.7}$$

berechnet wird und im zweiten Schritt als Funktion von θ maximiert wird.

EM-Algorithmus

1. Input: $\theta^{(0)}$, setze $j := 0$.
2. *(Expectation-Schritt)* Berechne $Q(\theta, \theta^{(j)}) = \mathbb{E}[l_T(\theta; X, Z)|X = x, \theta^{(j)}]$.
3. *(Maximization-Schritt)* Berechne $\theta^{(j+1)} = \underset{\theta \in \Theta}{argmax}\, Q(\theta, \theta^{(j)})$; setze $j = j + 1$.
4. Iteriere 2. und 3. solange, bis ein Konvergenz- oder Abbruchkriterium erfüllt ist.

Eine ausführlichere Untersuchung dieses Algorithmus kann man in [6] nachlesen. Wir zeigen jetzt nur eine wichtige Eigenschaft des EM-Algorithmus, nähmlich die Monotonie.

Lemma 3.1.1. *Seien $\theta^{(j)}$ und $\theta^{(j+1)}$ zwei aufeinanderfolgende Iterierte des EM-Algorithmus, $l_X(\theta; x)$ die Log-Likelihood-Funktion der beobachteten Daten. Dann gilt*

$$l_X(\theta^{(j+1)}; x) \geq l_X(\theta^{(j)}; x) \tag{3.8}$$

Beweis: Im Folgenden bezeichne $Pr(x; \theta)$ die Dichte von X an der Stelle x falls X stetig ist und die Wahrscheinlichkeit $\mathbb{P}(X = x; \theta)$ falls X diskret ist. Sei $\theta \in \Theta, \theta^{(j)}$ eine Iterierte des EM-Algorithmus, es gilt:

$$\begin{aligned} Pr(x; \theta) &= \frac{Pr(x, z; \theta)}{Pr(z|x; \theta)} \\ \Leftrightarrow l(\theta; x) &= l(\theta; x, z) - l(\theta; z|x). \end{aligned}$$

Wir bilden auf beiden Seiten den Erwartungswert bezüglich $Pr(x, z|x; \theta^{(j)})$:

$$\begin{aligned} l(\theta; x) &= \mathbb{E}[l(\theta; X, Z)|x; \theta^{(j)}] - \mathbb{E}[l(\theta; Z|X)|x; \theta^{(j)}] \\ &= Q(\theta, \theta^{(j)}) - R(\theta, \theta^{(j)}), \end{aligned}$$

wobei wir in der letzten Gleichung $R(\theta, \theta^{(j)}) := \mathbb{E}[l(\theta; Z|X)|x; \theta^{(j)}]$ definiert haben. Auf der linken Seite ändert sich nichts, da allgemein für eine Zufallsvariable X die Gleichung $\mathbb{E}(X|\sigma(X)) = X$ gilt.
Sei $\theta^{(j+1)} = \underset{\theta\in\Theta}{argmax}\ Q(\theta, \theta^{(j)})$, wir erhalten

$$l_X(\theta^{(j+1)}; x) - l_X(\theta^{(j)}; x) = [Q(\theta^{(j+1)}, \theta^{(j)}) - Q(\theta^{(j)}, \theta^{(j)})] - [R(\theta^{(j+1)}, \theta^{(j)}) - R(\theta^{(j)}, \theta^{(j)})]. \tag{3.9}$$

Es ist klar, dass $Q(\theta^{(j+1)}, \theta^{(j)}) \geq Q(\theta^{(j)}, \theta^{(j)})$ gilt.

Behauptung: $R(\theta^{(j+1)}, \theta^{(j)}) \leq R(\theta^{(j)}, \theta^{(j)})$.

Beweis der Behauptung:

$$\begin{aligned}
R(\theta, \theta^{(j)}) &= \mathbb{E}[l(\theta; Z|X)|x; \theta^{(j)}] = \int ln(Pr(z|x;\theta))Pr(z|x;\theta^{(j)})dxdz \\
&= \int ln[Pr(z|x;\theta)\frac{Pr(z|x;\theta^{(j)})}{Pr(z|x;\theta^{(j)})}]Pr(z|x;\theta^{(j)})dxdz \\
&= \int ln[\frac{Pr(z|x;\theta)}{Pr(z|x;\theta^{(j)})}]Pr(z|x;\theta^{(j)})dxdz + \int ln[Pr(z,x;\theta^{(j)})]Pr(z|x;\theta^{(j)})dxdz \\
&= \mathbb{E}\left[ln\,\frac{Pr(Z|X;\theta)}{Pr(Z|X;\theta^{(j)})}|x;\theta^{(j)}\right] + const.
\end{aligned}$$

Also gilt

$$\underset{\theta\in\Theta}{argmax}\ R(\theta, \theta^{(j)}) = \underset{\theta\in\Theta}{argmax}\ \mathbb{E}\left[ln\frac{Pr(Z|X;\theta)}{Pr(Z|X;\theta^{(j)})}|x;\theta^{(j)}\right].$$

Die Jensen-Ungleichung liefert:

$$\mathbb{E}\left[ln\frac{Pr(Z|X;\theta)}{Pr(Z|X;\theta^{(j)})}|x;\theta^{(j)}\right] \leq ln\,\mathbb{E}\left[\frac{Pr(Z|X;\theta)}{Pr(Z|X;\theta^{(j)})}|x;\theta^{(j)}\right] = ln(1) = 0.$$

Für $\theta = \theta^{(j)}$ gilt: $\mathbb{E}\left[ln\frac{Pr(Z|X;\theta)}{Pr(Z|X;\theta^{(j)})}; \theta^{(j)}\right] = ln(1) = 0$. Also maximiert $\theta^{(j)}$ $R(\theta, \theta^{(j)})$.

Ende Beweis der Behauptung. □

Insgesamt erhalten wir $R(\theta^{(j+1)}, \theta^{(j)}) \leq R(\theta^{(j)}, \theta^{(j)})$ und $l_X(\theta^{(j+1)}; x) - l_X(\theta^{(j)}; x) = [Q(\theta^{(j+1)}, \theta^{(j)}) - Q(\theta^{(j)}, \theta^{(j)})] - [R(\theta^{(j+1)}, \theta^{(j)}) - R(\theta^{(j)}, \theta^{(j)})] \geq 0$.

□

Jetzt wissen wir, wie der EM-Algorithmus im Allgemeinen funktioniert. Um die Parameter unserer Mischungsdichte (3.1) zu schätzen, müssen wir unser Modell um versteckte Variablen erweitern.
Die beobachteten Daten sind in unserem Fall die Stichprobe von X_{mix}:

$$(y_1, \ldots, y_N).$$

Die vollständigen Daten sind die Stichprobe von $T = (X_{mix}, Z)$:

$$([y_1, z_1], \ldots, [y_N, z_N]).$$

Die vestreckten Daten sind somit die Werte von Z:

$$(z_1, \ldots, z_N).$$

Dabei gilt

$$z_i = (z_{i1}, \ldots, z_{iK})$$

und

$$z_{ik} = \begin{cases} 1 & x_i \text{ gehört zur Komponente k} \\ 0 & \text{sonst.} \end{cases}$$

e_i bezeichne den i'ten Einheitsvektor im $\mathbb{R}^K$, dann erhalten wir

$$f_{X_{mix}|Z=e_i}(x) = c_i \varphi_i((x-\mu_i)^\mathsf{T} \Sigma_i^{-1}(x-\mu_i)).$$

Wir nehmen an $Z_1, \ldots, Z_N$ seien iid $M(1, \pi_1, \ldots, \pi_K)$ (multinomial mit einmaligem Ziehen aus K Gruppen) verteilt, somit erhalten wir für die bedingte Dichte $f_{X_{mix}|Z=z_i}(y_i) = \prod_{j=1}^{K} f_i(y_i;\theta_i)^{z_{ij}}$. Die Gesamt-Log-Likelihood-Funktion lautet dann

$$l_{ges}(\theta, \pi; y, z) = \sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij} ln(\pi_j f_j(y_i;\theta)), \tag{3.10}$$

wobei bedeuten $\theta = (\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \vartheta_1, \ldots, \vartheta_K)$ und $\pi = (\pi_1, \ldots, \pi_K)$.
Für den *Exepctation-* oder *E-Schritt* ergibt sich dann die allgemeine Formel:

$$\hat{z}_{ij} = \frac{\hat{\pi}_j f_j(y_i;\hat{\theta}_j)}{\sum_{k=1}^{K} \hat{\pi}_k f_k(y_i;\hat{\theta}_k)} \tag{3.11}$$

Einsetzten dieser Werte für die fehlenden z_{ik} in die Gesamt-Log-Likelihood liefert den bedingten Erwartungswert $Q((\theta, \pi), (\hat{\theta}, \hat{\pi}))$.

Falls X_{mix} bzw. g eine Gauss-Mischung ist, ergeben sich für den M-Schritt folgende Update-Formeln:

$$\hat{\pi}_j = \frac{m_j}{N}, \quad 1 \leq j \leq K \tag{3.12}$$

$$\hat{\mu}_j = \frac{\sum_{i=1}^{N} \hat{z}_{ij} y_i}{m_j}, \quad 1 \leq j \leq K \tag{3.13}$$

$$m_j := \sum_{i=1}^{N} \hat{z}_{ij}, \quad 1 \leq j \leq K \tag{3.14}$$

Die Formeln für die Kovarianzmatrizen hängen von deren Parametrisierung ab.

Die Nachteile des EM-Algorithmus sind seine langsame Konvergenz, und die Eigenschaft abzustürzen wenn eine der Kovarianzmatrizen beinahe singulär ist. Außerdem liefert er ungenaue Ergbenisse falls einer der Cluster nur wenige Datenpunkte enthält oder falls die Datenpunkte entlang eines linearen Unterraums niedriger Dimension konzentriert sind (Stichwort Kondition der Kovarianzmatrizen).
Die Leistung des EM-Algorithmus hängt entscheidend von der Wahl der Startpunkte ab, da er bei ungünstigen Startwerten in Sattelpunkten festhängen kann. Bei guten Startwerten liefert er gute Ergebnisse. Diese Eigenschaft haben auch alle newtonartigen Verfahren. Ein Vorteil des EM-Algorithmus sind auch bekannte geschlossene Formeln für die E- bzw. M-Schritte in manchen Modellen (z.B. in Gauss-Modellen), was den Rechenaufwand deutlich reduziert.

3.1.2. Wahl der Komponentenanzahl

Bisher sind wir von einer gegebenen Komponentenanzahl K unseres Modells $g(x;\theta) = \sum_{i=1}^{K} \pi_i f_i(x;\theta)$ ausgegangen. Diese Größe ist aber i.A. unbekannt und ist optimal zu wählen.
Für diesen Zweck benutzen wir das Schwarz Information Criterion (SIC) aka Bayesian Information Criterion (BIC). Die Idee besteht grob aus folgenden Überlegungen: seien $M_1 \dots M_l$ mögliche Modelle mit a-priori-Wahrscheinlichkeiten $Pr(M_1) \dots Pr(M_l)$ (meist alle gleich) aus denen wir das optimale auswählen möchten. Nach dem Satz von Bayes gilt für die a-posteriori-Wahrscheinlichkeiten gegeben Daten D

$$Pr(M_k|D) = Pr(D|M_k)\frac{Pr(M_k)}{Pr(D)} \propto Pr(D|M_k)Pr(M_k), \tag{3.15}$$

wobei das Symbol $\propto$ „proportional zu“ bedeutet. Falls das Modell unbekannte Parameter enthält, lassen sie sich nach dem Gesetz der totalen Wahrscheinlickeit rausintegrieren:

$$Pr(D|M_k) = \int Pr(D|\theta_k, M_k)Pr(\theta_k|M_k)d\theta_k, \tag{3.16}$$

wobei $Pr(\theta_k|M_k)$ die a-priori-Verteilung des Parametervektors θ_k des Modells M_k ist. Die Größe $Pr(D|M_k)$ wird *integrierte Likelihood des Modells* M_k genannt. Man entscheidet sich für das Modell mit der höchsten a-posteriori-Wahrscheinlichkeit $Pr(M_k|D)$. Falls alle a-priori-Wahrscheinlichkeiten $Pr(M_k)$ gleich sind, ist das äquivalent zur Wahl des Modells mit der höchsten integrierten Likelihood $Pr(D|M_k)$. Im Fall $l = 2$ (Entscheidung zwischen zwei Modellen), berechnet man den sog. *Bayes-Faktor*

$$B_{12} := \frac{Pr(D|M_1)}{Pr(D|M_2)}.$$

Ein Faktor $B_{12} > 100$ spricht stark für das Modell M_1.
Dieser Ansatz birgt eine Schwierigkeit - die Auswertung des Integrals (3.16). Unter gewissen Annahmen läßt sich aber folgende Approximation zeigen:

$$2lnPr(D|M_k) \approx 2lnPr(D|\hat{\theta}_k) - \nu_k(log(N)) := BIC_k, \tag{3.17}$$

wobei θ_k die ML-Schätzung von θ und ν_k die Anzahl der zu schätzenden Parameter im Modell M_k ist. Offensichtlich bestraft BIC komplexe Modelle mit vielen Parametern umso mehr, je größer die Datenmenge ist. Für weitere Ausführungen zu BIC siehe z.B. [C.Fraley, A.Raftery].

3.1.3. Zusammenlegen von Komponenten

Unser Ziel ist die Einteilung der Daten in sinnvolle zusammenhängende Gruppen. Der naheliegende Ansatz jede Mischungskomponente, der von EM/BIC berechneten Mischung mit einem Cluster zu assoziieren, führt oft zu zu vielen Clustern. C.Henning hat in [12] einen Datensatz vorgestellt, der von EM/BIC mit einer Gaussmischung aus 9 Komponenten gefittet wurde, obwohl es nur 4 sinnvolle Cluster in dem Datensatz gab. Die Ursache dafür ist, dass die gewählte elliptische Verteilungsfamilie (z.B. die Normalverteilung) nicht felxibel genug ist um einen Cluster mit einer Komponente zu beschreiben. Beispielsweise können t-verteilte Daten nur mit mehreren Gauss-Komponenten gefittet werden, da sie mehr extreme Werte enthalten als die normalverteilten (s. auch Abschnitt Impelemtierung und Simulationen). Außerdem sind Cluster vorstellbar, die keine ellipsoide Form haben,

sie können auch nicht mit einer Komponente modelliert werden. Um diese Probleme in Griff zu kriegen, legen wir einzelne Mischungskomponenten zusammen.
Der nächste Schritt nach Anpassung einer Mischungsdichte g an die Daten ist also das Zusammenlegen (Mergen) von Komponenten zu Clustern. Um das tun, müßen wir Kriterien finden wann eine Teilmenge der Daten einen Cluster bildet. C. Hennig erwähnt folgende Ansätze:

- modalitätsbasierter Ansatz
- dichtebasierte Ansatz

Der *modalitätsbasierte Ansatz* unterscheidet Cluster nach Moden der Mischungsdichte. Punkte, die sich in der Nähe eines Modus befinden, gehören zusammen in einen Cluster. Die einzelnen Cluster werden durch Gefälle (gaps) in der Mischungsdichte getrennt. Ein Cluster hat somit immer einen Modus.
Der *dichtebasierte Ansatz* unterscheidet Cluster nach der Punktdichte in der Datenmenge. Regionen mit hoher Datendichte werden von Regionen mit niedriger Datendichte unterschieden. Der Übergang von Regionen mit hoher Datendichte zu denen mit niedriger Datendichte trennt einzelne Cluster.

Beide Ansätze sind berechtigt und die Wahl hängt von der Anwendungssituation ab, siehe [Hennig]. Wir untersuchen eine Merging-Strategie mit dem modalitätsbasierten Ansatz und greifen dafür zurück auf die Theorie aus Kapitel 2. Dort haben wir gezeigt, dass alle kritischen Punkte der Mischungsdichte g auf der Ridgeline $x^*(\mathcal{S}_K)$ liegen, wobei

$$x^*(\alpha) = \left[\alpha_1\Sigma_1^{-1} + ... + \alpha_K\Sigma_K^{-1}\right]^{-1}\left[\alpha_1\Sigma_1^{-1}\mu_1 + ... + \alpha_K\Sigma_K^{-1}\mu_K\right]$$

und $\mathcal{S}_K$ der Einheitssimplex im $\mathbb{R}^K$ ist. Die Idee des nachfolgenden Algorithmus ist es die Cluster, deren Mischung unimodal oder schlecht separiert multimodal ist, zusammen zu fassen und als einen Cluster anzusehen. Für diese Methode brauchen folgende Größe:

Definition 3.1.2. Sei g eine elliptische Mischung aus zwei Komponenten. Sei x_{min} das Minimum der Mischung g und x_{mod} der zweitgrößte Modus von g, dann heißt

$$RQ := \begin{cases} 1 & g \text{ ist unimodal} \\ \frac{g(x_{min})}{g(x_{mod})} & \text{sonst} \end{cases} \tag{3.18}$$

Ridgeline-Quotient der Mischung g.

Der Ridgeline-Quotient (RQ) kann als ein Maß für Unimodalität einer elliptischen Mischung verstanden werden - je größer RQ, desto „unimodaler“ ist die Mischung. Es ist nicht schwer einzusehen, dass $RQ \leq 1$ ist. Falls die Mischung g mehr als zwei Moden besitzt, kann es vorkommen, dass der 2. größte Modus und das kleinste lokale Minimum nicht aufeinander folgen. Das stellt aber kein Problem dar, da der Ridgeline-Quotient in diesem Fall richtigerweise auf die Nichtunimodalität deutet.
Es folgt ein generischer Merging-Algorithmus, den wir im nächsten Abschnitt präzisieren:

Ridgeline-Quotient Methode

1. Wähle eine Strafkonstante $0 < r^* < 1$ aus.
2. Beginne mit einzelnen Komponenten der Mischungsdichte g als Cluster.
3. Berechne den Ridgeline-Quotient RQ einer 2-Komponenten Mischung mit gemittelten Mittelwerten und Kovarianzmatrizen der Cluster (zu Beginn einzelne Komponenten), für alle Paare von Clustern.
4. Falls $RQ < r^*$ für alle Paare von Clustern, behalte die aktuelle Clusterung und beende den Algorithmus, sonst lege das Cluster-Paar mit maximalem RQ zusammen. Gehe zu Schritt 3.

Der Parameter r^* kann als ein Schwellenwert zwischen Unimodalität und Multimodalität angesehen werden. Je größer er gewählt wird, desto kleiner muß das Gefälle zwischen zwei Moden ausfallen um die Komponenten zusammen zu legen.

Nach dem Mergen erhalten wir eine Mischung von Mischungen f_i^*:

$$g(x) = \sum_{i=1}^{K} \pi_i f_i(x) = \sum_{j=1}^{S} f_j^*(x), \tag{3.19}$$

mit $f_i^* = \sum_{k=1}^{K_i} \pi_{i_k} f_{i_k}(x)$, $S \leq K_1 + \ldots + K_S = K$ und $\{1, \ldots, K\} = \bigcup_{i=1}^{S} \{j_1, \ldots, j_{K_i}\}$.

3.1.4. Einteilung der Daten in die Cluster

Wir haben ein passendes Modell für unsere Daten konstruiert und wollen jetzt die Stichprobe $y_1, \ldots, y_N$ in Cluster einteilen. Als Cluster i definierten wir die Datenpunkte, die die posteriori-Wahrscheinlichkeit $Pr(C = e_i | X_{mix} = y)$ maximieren. Formal:

$$Cluster_i := \left\{ y \in \{y_1, \ldots, y_N\} \mid \frac{f_i^*(y)}{\sum_{j=1}^{S} f_j^*(y)} \geq \frac{f_k^*(y)}{\sum_{j=1}^{S} f_j^*(y)} \; \forall k \in \{1, \ldots, K\} \right\} \tag{3.20}$$

$$= \{y \in \{y_1, \ldots, y_N\} \mid f_i^*(y) \geq f_k^*(y) \; \forall k \in \{1, \ldots, K\}\}$$

Für die Parameter der Dichten f_i^* nehmen wir die ML-Schätzungen des EM-Algorithmus, d.h.

$$f_i^*(y) = \pi_1 f_{i_1}(y; \hat{\theta}_{i_1}) + \ldots + \pi_{K_i} f_{i_{K_i}}(y; \hat{\theta}_{i_{K_i}}).$$

3.2. Implementierung und Simulation

3.2.1. Details zur Implementierung

Der im vorherigen Abschnitt vorgesellte Ansatz wurde für Gauss-Mischungen implementiert. Als Programmiersprache haben wir R (s. http://www.r-project.org/) gewählt. Die Schätzung der Mischungsdichte g wird mit dem Paket **Mclust** von Chris Fraley und Adrian Raftery realisiert. Das Paket beinhaltet die Funktion `Mclust()` in der der EM-Algorithmus in Kombination mit BIC implemetiert ist. Für ausführlichere Informationen siehe [10]. Der Ridgeline-Merging-Algorithmus wurde in der Funktion `ridgeline_merge()` neu implemetiert. Die Funktion `ridgeline_merge()` erwartet als Input das von Mclust berechnete Objekt, den Schwellenparameter r^* und die Anzahl der Durchläufe der Optimierungsroutine bei der Suche nach Extrema auf der Ridgeline N. Als Ergebnis gibt die Funktion 4 Objekte zurück:

1. Vektor mit neuen Datenlables, in dem zu jedem Datenpunkt sein Cluster angegeben ist.

2. Liste der gemergten Cluster

3. Matrix mit neuen a-posteriori-Wahrscheinlichkeiten $A = (a_{i,j})_{i,j}$ mit $a_{i,j} = \mathbb{P}(Z = e_j | x_i) = \frac{f_j^*(x_i)}{\sum_{k=1}^{S} f_k^*(x_i)}$

4. Ridgeline-Quotienten für jedes gemergte Cluster-Paar

Die Berechnung des Ridgeline-Quotienten RQ ist ein kritischer Schritt des Ridgeline-Merging-Algoritmus, den wir jetzt kurz erläutern.
Wir müssen zum einen x_{min} - das kleinste lokale Minimum der Mischungsdichte und zum anderen alle Moden für das x_{mod} finden, da wir wissen müssen welcher Modus der zweitgrößte ist.
Das stellt ein Problem dar, da die bekannten Optimierungsroutinen die auf Gradientenabstieg oder Newtonverfahren basieren nur lokale Extrema finden. Unser Ansatz zum Auffinden aller Maxima bzw. Minima besteht in 3 Schritten:

1. Wähle Anzahl von Suchvorgängen N.

2. Ziehe eine Stichprobe vom Umfang N aus $U(0,1)$

3. Führe N-Mal Maximierungs- /Minimierungsroutine für $g(x^*(\alpha))$ mit Startpunkten aus Schritt 2 aus und speichere die Ergebnisse in einer Liste.

Wir hoffen, daß falls N groß genug gewählt wurde, so alle Extrema gefunden werden können. Die Größe N soll von der Dimension D abhängen, da in höheren Dimensionen größere Anzahlen von Moden möglich sind (s. Kapitel 2, Schrankensätze). Im Fall $D = 2$ haben Rechnungen mit N zwischen 10 und 15 gute Ergebnisse geliefert. Ein möglicher Ansatz für die Wahl von N ist $N = \frac{15}{2}D$ da die bekannte Schranke für die Modenanzahl einer Gauss-Mischung proportional zu D mit dem Faktor $\frac{3}{2}$ wächst (s. Satz 2.2.3). Als Optimierungsroutine in Schritt 3 haben wir die Funktion `nlminb()` aus dem R Paket `stats` gewählt, das zur Standart-Distribution von R gehört.

3.2.2. Simulationen

Wir haben den Gesamtalgorithmus (Mclust + Merging) an drei simulierten und einem realen Datensatz getestet. Die Ergebnisse wollen wir jetzt diskutieren. Um eine Mischung

$$g(x;\mu_1,\ldots,\mu_K,\Sigma_1,\ldots,\Sigma_K,\pi_1,\ldots,\pi_{K-1},\vartheta_1,\ldots,\vartheta_K) = \sum_{i=1}^{K} \pi_i \cdot k_i \cdot \varphi_i((x-\mu_i)^{\mathsf{T}}\Sigma_i^{-1}(x-\mu_i)).$$

zu simulieren, gingen wir in zwei Schritten vor:

1. Generiere einen Wert aus $M(1,(\pi_1,\ldots,\pi_K))$ (Multinomialverteilung).

2. Sei e_i der Wert aus Schritt 1. Generiere einen Wert aus $k_i\varphi_i((x-\mu_i)^\mathsf{T}\Sigma_i^{-1}(x-\mu_i))$.

Diese Vorgehensweise soll so oft wiederholt werden, wieviele Punkte man generieren möchte.

Experiment 1: Simulation einer Gauss-Mischung aus 5 Komponenten

In unserem 1. Experiment simulierten wir wiederholt eine Gauss-Mischung mit folgenden Parametern:

$K = 5$, $D = 2$, $\pi_1 = \pi_4 = \pi_5 = 0.2$, $\pi_2 = 0.25$, $\pi_3 = 0.15$,

$$\mu_1 = \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1.5 \\ 5 \end{pmatrix}, \mu_3 = \begin{pmatrix} 2 \\ 4.5 \end{pmatrix}, \mu_4 = \begin{pmatrix} 4.1 \\ 1 \end{pmatrix}, \mu_5 = \begin{pmatrix} 5 \\ 1 \end{pmatrix},$$

$$\Sigma_1 = \begin{pmatrix} 0.3 & 0.05 \\ 0.05 & 0.3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix},$$

$$\Sigma_4 = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix} = \Sigma_5$$

Auf der folgenden Abbildung ist die Mischungsdichte zu sehen:

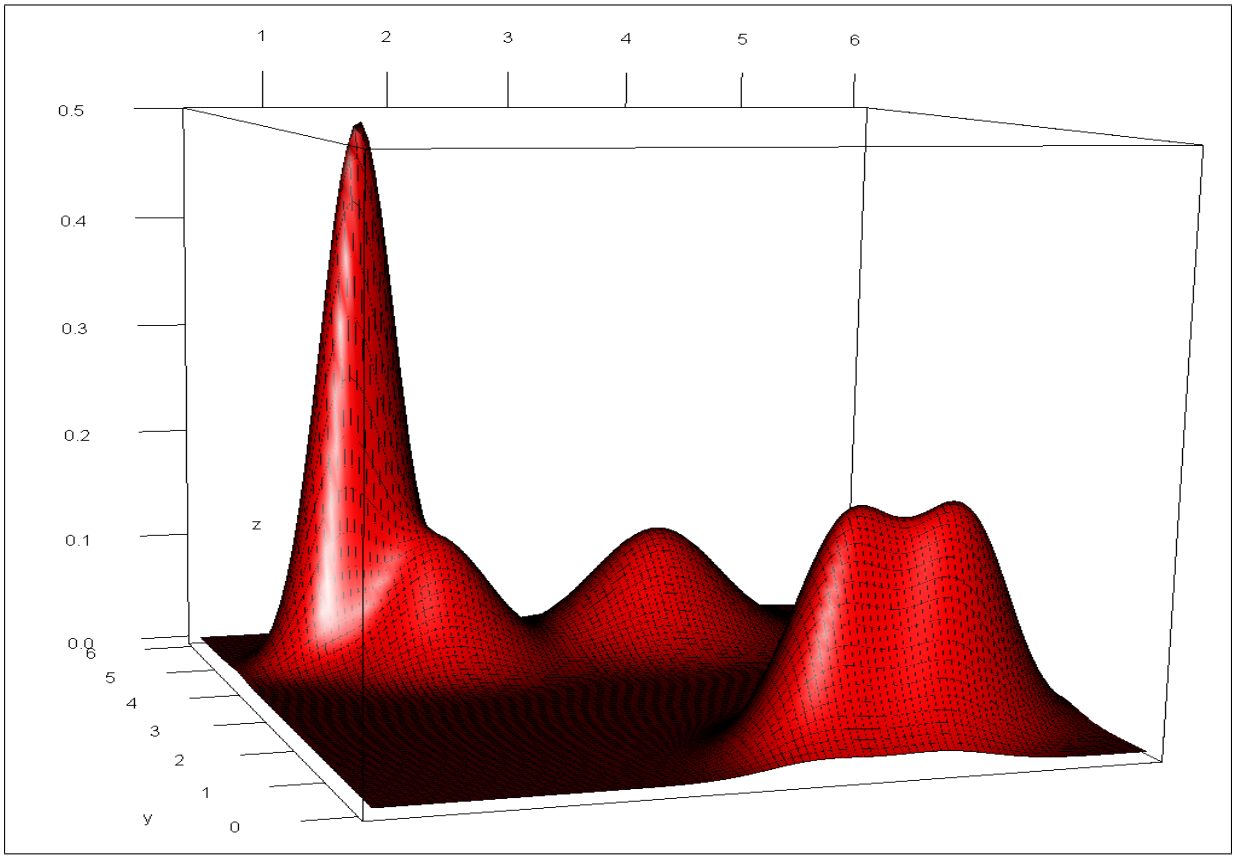


Abb. 3.1: Dichte der Gauss-Mischung

Man kann deutlich erkennen, daß die Komponenten 2 und 3 (hinten links) und 5 und 6 (vorne rechts) schwach getrennt sind, wobei 2 und 3 unimodal und 5 und 6 schwach bimodal sind. Aus Sicht unseres modalitätsbasierten Ansatzes sollten diese Paare zusammengelegt werden.

Wir generierten mit obigem Simulationsverfahren 600 Punkte aus dieser Mischung und wendeten darauf anschließend den Mclust-Algorithmus ab. Diesen Vorgang wiederholten wir 4 Mal und betrachteten die sich ergebenen Clusterungen und die jeweiligen Ridgeline-Quotienten.

Mclust reagierte empfindlich auf die Änderungen der Daten und fittete die Daten in 2 Fällen mit 4 Gauss-Komponenten und in den anderen 2 Fällen mit 5 bzw. 6 Gauss-Komponenten (s. Abb. 3.2). Die Farben und die Nummerierungen der Datenpunkte in den folgenden Abbildungen haben keine besondere Bedeutung und dienen lediglich zur Unterscheidung von verschiedenen Cluster, wobei die Nummer des Datenpunktes gleich der Nummer des Clusters ist, dem der Punkt zugeordnet wurde.

Wir reichten die Ergebnisse von `Mclust()` an die Funktion `ridgeline_merge()` mit Parametern $r^* = 0.7$ und $N = 15$. Der Merging-Algorithmus lieferte in jedem Fall jeweils 3 Cluster, die gut voneinander gertrennt waren. Das Verfahren erkannte zuverlässig die Unimodalität bzw. die schwache Bimodalität der jeweiligen Teilmischungen und legte sie zusammen.

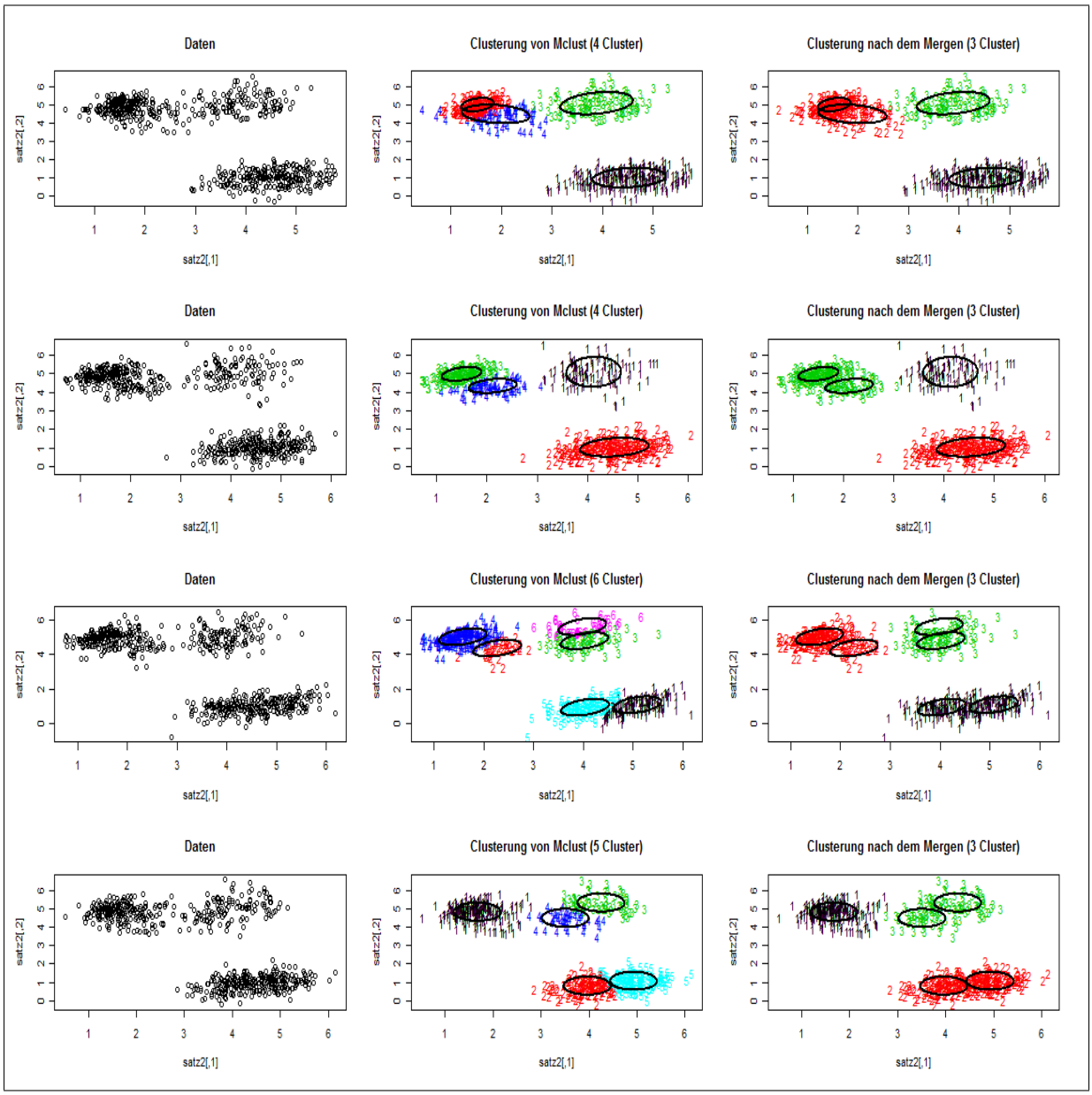


Abb. 3.2: Simulationen einer Gauss-Mischung, v.l.n.r.: simulierte Daten, Clusterungen von Mclust, Clusterungen nach dem Mergen. Die Ellipsen deuten die 50%-Konfidenzgebiete der jeweiligen Gauss-Komponenten an.

In der unteren Tabelle ist zu jedem gemergten Cluster-Paar angeben welcher Ridgeline-Quotient sich ergeben hat. Für die Lables der Cluster gilt: nach jedem Zusammenlegen von zwei Clustern erhält der neue Cluster die kleinere beiden Nummern, d.h. beim Mergen von $Cluster_i$ mit $Cluster_j$ bekommt der neue Cluster die Nummer $min(i, j)$.

Zur Errinnerung: RQ = 1 bedeutet Unimodalität.

Simulation/Iteration	Zusammengelegte Cluster	Ridgeline-Quotient
Simulation 1		
1	(2,4)	1
Simulation 2		
1	(3,4)	0.823
Simulation 3		
1	(2,4)	1
2	(3,6)	1
3	(1,5)	0.840
Simulation 4		
1	(2,5)	0.973
2	(3,4)	0.900

Tabelle 3.1: Experiment 1. Ergebnisse des Merging-Algorithmus

Experiment 2: Simulation einer t-Mischung aus 3 Komponenten

In unserem nächsten Versuch betrachteten wir eine t-Mischung mit folgenden Parametern:

$K = 3$, $D = 2$, $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$, $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 15 \\ 15 \end{pmatrix}$, $\mu_1 = \begin{pmatrix} 20 \\ 0 \end{pmatrix}$,

$\Sigma_1 = \begin{pmatrix} 1 & 0.05 \\ 0.05 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_3 = \Sigma_1$.

So sieht die Mischungsdichte aus:

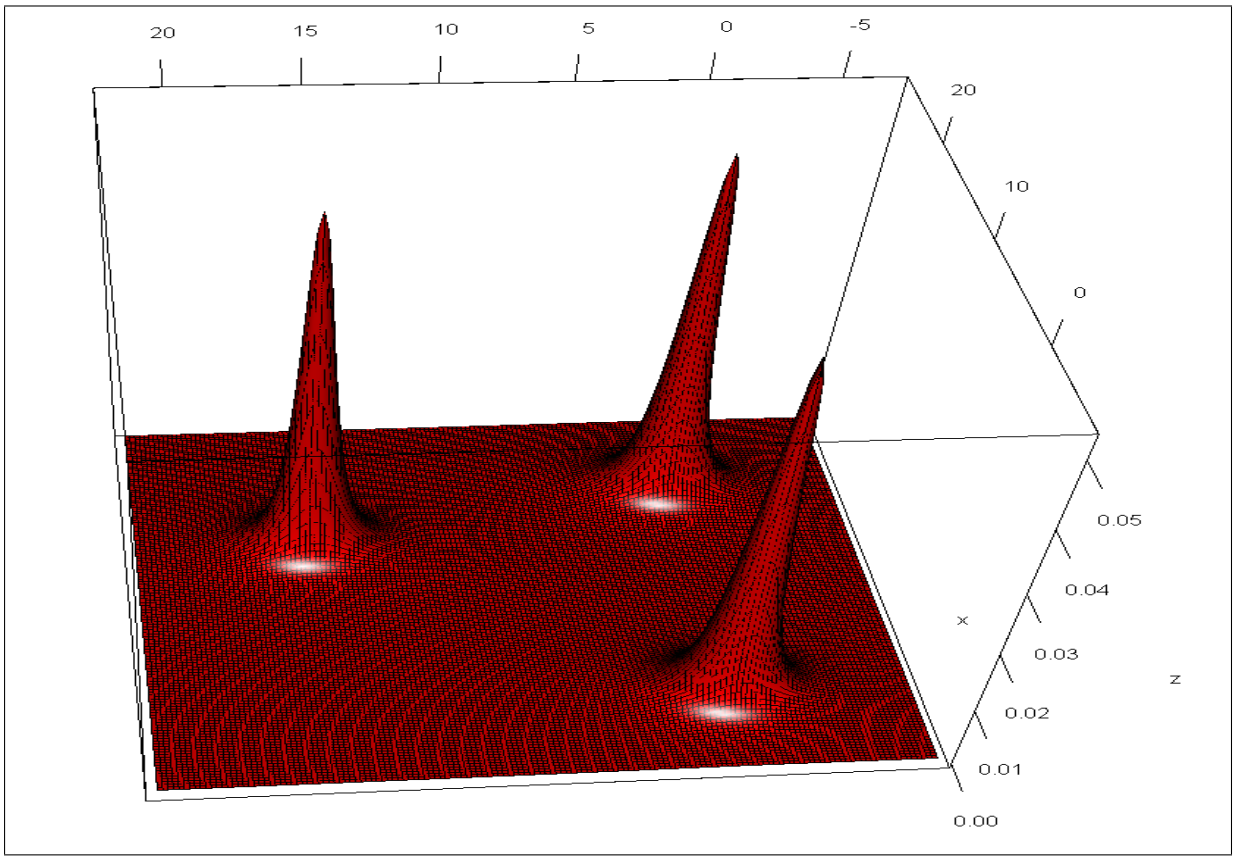


Abb. 3.3: Dichte der t-Mischung

Wir generierten 600 Punkte aus der Mischung und wendeten darauf den Mclust-Algorithmus an. Mclust lieferte eine Mischung aus 7 Gauss-Komponenten. Er modellierte einzelne t-Komponenten mit je zwei Gauss-Komponenten, eine für die um den Mittelpunkt gehäuften Werte und eine für die weiter außerhalb liegenden Werte. Die ganz extremen Werte aller t-Komponenten wurden duch eine Gauss-Komponente mit großer Varianz gefittet. Wieder erhalten wir zu viele Cluster.

Die von Mclust erzeugte Clusterung verarbeiteten wir weiter in der Funktion `ridgeline_merge()` mit den gleichen Parametern wie im 1. Experiment. Das Ergebnis ist in der folgenden Abbildung zusammengefasst. Die Funktion vereinigte die Gauss-Komponenten mit denen einzelne t-Komponenten modelliert wurden, so daß pro t-Komponente je ein Cluster entstand. Die extremen Werte, die von Mclust mit einer breiten Gauss-Komponente modelliert wurden, gehörten nun zu einem der drei Cluster. Diese Clusterung stimmte eher mit dem den Daten zugrunde liegenden Modell überein. Die Ridgeline-Quotienten der gemergten Paare betrugen in jedem Schritt 1. Das blieb

auch bei mehrfacher Wiederholung der Simulation so, wobei Mclust bei unterschiedlichen Simulationen wieder unterschiedlich viele Gauss-Komponenten generierte, abhängig von den Daten.

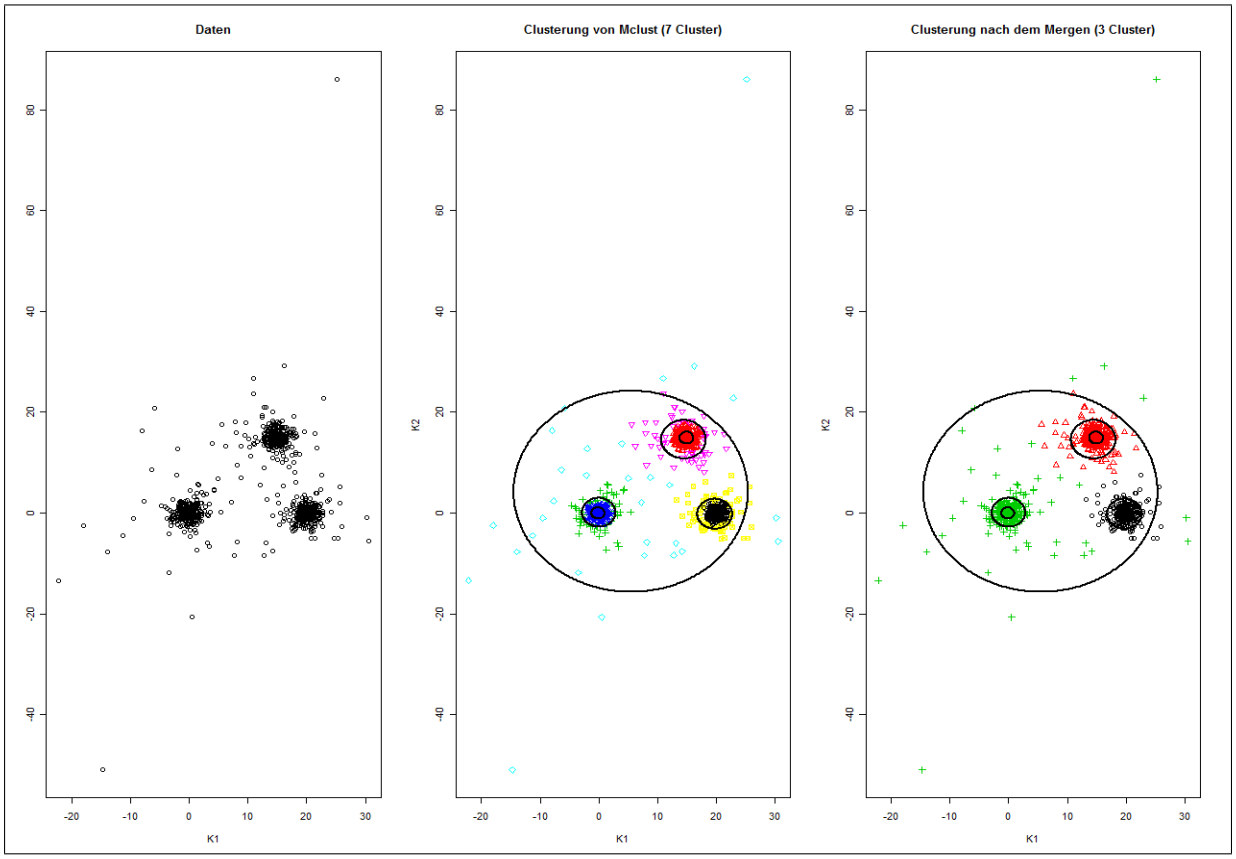


Abb. 3.4: Simulation einer t-Mischung, v.l.n.r.: simulierte Daten, Clusterung von Mclust, Clusterung nach dem Mergen. Die Ellipsen deuten die 50%-Konfidenzgebiete der jeweiligen Gauss-Komponenten an.

Experiment 3: Schiefe Normalverteilung

In diesem Experiment simulierten wir eine schiefe Normalverteilung. Ein D-dimensionaler Zufallsvektor X ist schief-normalverteilt, falls er folgende Dichte besitzt:

$$f(x; \Sigma, a) = 2\phi(x, \Sigma)\Phi(x^{\mathsf{T}}a),$$

wobei $\phi(x, \Sigma)$ Dichte einer D-dimensionalen Normalverteilung mit Erwartungswert 0 und Kovarianzmatrix Σ ist, Φ Verteilungsfunktion einer $N(0, 1)$ Zufallsvariable und $a \in \mathbb{R}^D$. Der Parameter a wird Schiefe-Parameter genannt. Für $a = 0$ erhalten wir die Standardnormalverteilung im $\mathbb{R}^D$. Mehr Informationen über die schiefe Normalverteilung gibt es z.B. bei A. Azzalini, A. Capitanio [1].

Wir betrachteten schiefe Normalverteilungen mit steigenden Schiefeparametern. Aus der Definition der schiefen Normalverteilung wird klar, dass Punkte x für die $x^\mathsf{T}a$ eine große negative Zahl ist, wenig Masse besitzen. Das sind Punkte die einen großen Winkel zu a haben. Je größer der Betrag von a, desto weniger Masse hat die Dichte bei x. Wir führten 4 Simulationen einer schiefen Normalverteilung durch mit folgenden Parametern:

$$D = 2,\ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ a \in \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 25 \\ 25 \end{pmatrix}, \begin{pmatrix} 125 \\ 125 \end{pmatrix} \right\}.$$

Der Schiefeparameter wurde also nach jeder Simulation mit 5 multipliziert.
Wir generierten jedesmal 600 Punkte aus der schiefen Normalverteilung mit der Funktion `rmsn()` aus dem Paket **SN** von Adelchi Azzalini und clusterten sie mit Mclust. Anschließend übergaben wir das Ergebnis von Mclust der Funktion `ridgeline_merge()`.
Uns interessierte ob die von einer einzigen schiefen Normalverteilung produzierten Daten von unseren Clustering-Algorithmen als homogen erkannt werden. Mclust brauchte mehrere Gauss-Komponenten um diese Daten zu fitten, außer im Fall $a = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, wo er nur eine Gauss-Komponente brauchte. Der Mergingalgorithmus erkannte richtigerweise immer nur einen Cluster, allerdings nur mit einem recht niedrigem Schwellenparameter $r^* = 0.3$.

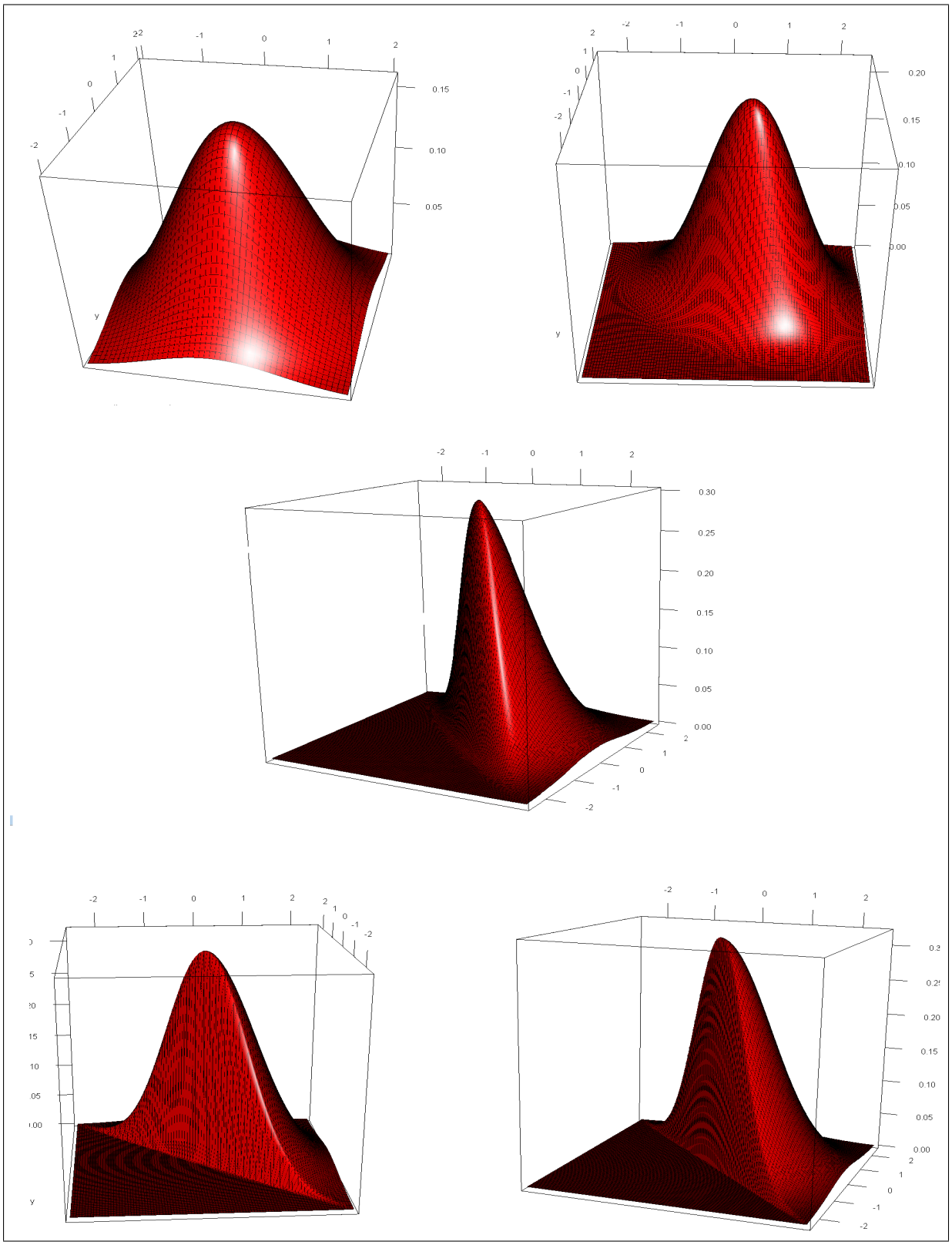


Abb. 3.5: Dichten der schiefen Normal-Verteilungen, von links nach rechts steigt der Schiefeparameter.

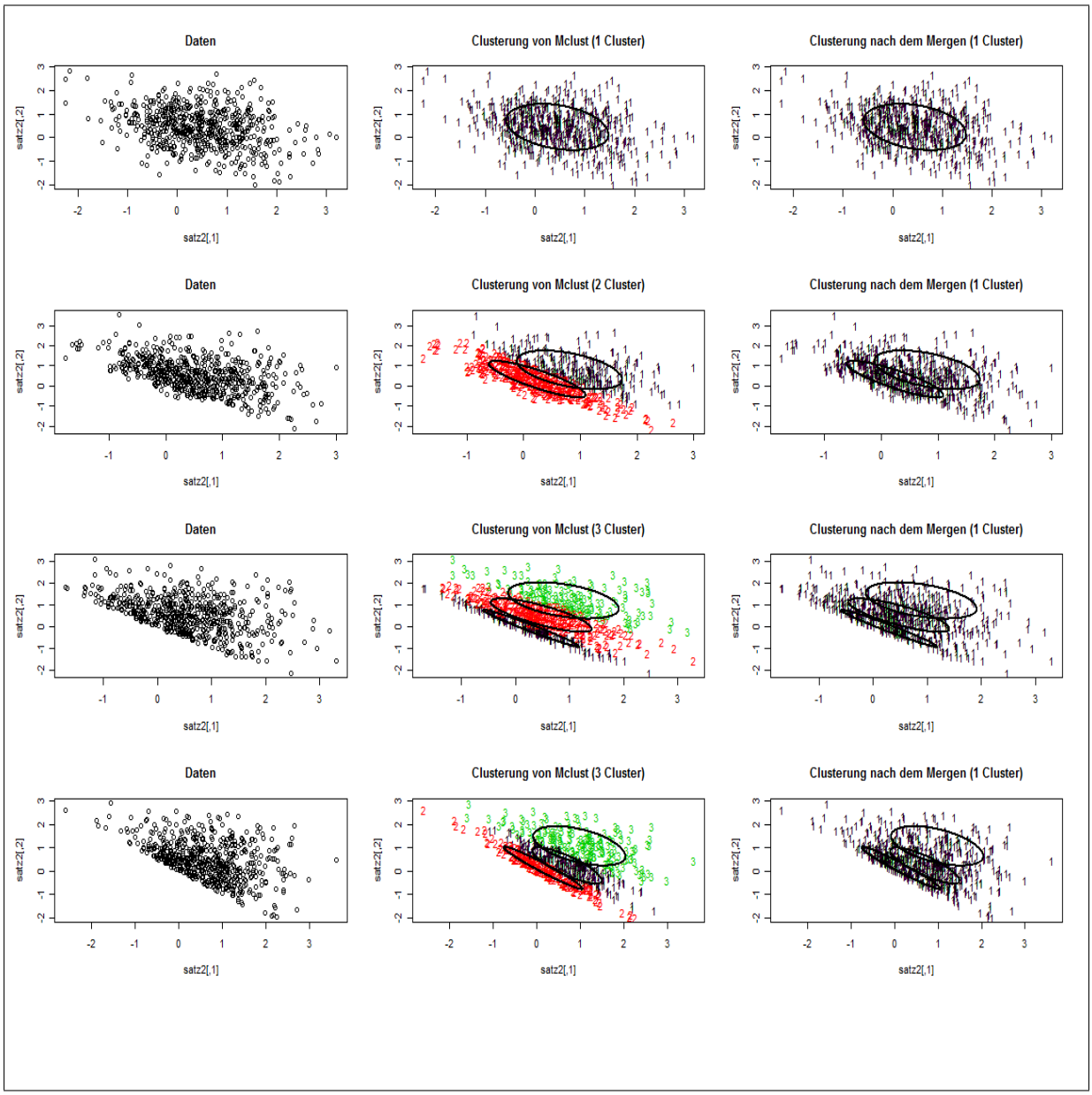


Abb. 3.6: Simulation einer schiefen NV, v.l.n.r.: simulierte Daten, Clusterung von Mclust, Clusterung nach dem Mergen. Die Ellipsen deuten die 50%-Konfidenzgebiete der jeweiligen Gauss-Komponenten an.

Simulation/Iteration	Zusammengelegte Cluster	Ridgeline-Quotient
Simulation 1		
1	keine	-
Simulation 2		
1	(1,2)	1
Simulation 3		
1	(2,3)	1
2	(1,2)	0.360
Simulation 4		
1	(1,3)	1
2	(1,2)	0.440

Tabelle 3.2: Experiment 3. Ergebnisse des Merging-Algorithmus

Wie man sieht, sind die Ridgeline-Quotienten in den Simulationen 3 bzw. 4 recht klein, das bedeutet, dass die zugehörigen Komponeten gut getrennt waren und bei einem Schwellenparameter $r^* > 0.36$ bzw.$r^* > 0.44$ nicht zusammengelegt worden wären.

Experiment 4: Realer Datensatz

Hier betrachteten wir den Datensatz **World Population Prospects: The 2008 Revision** der UN, wo für verschiedene Länder und Zeitperioden von je 5 Jahren von 1950-2005 Fruchtbarkeit und Sterblichkeit der Bevölkerung gemessen wurde. Der Datensatz beinhaltet 2024 Datenpunkte. Unser Ziel war wieder die Daten in Cluster einzuteilen.
Mclust teilte die Daten in 6 Cluster ein. Bei einer optischen Inspektion der geschätzten Dichte sind aber nur 3 Moden feststellbar, was bei unserem Modalitätsbasierten Ansatz 3 Cluster impliziert.

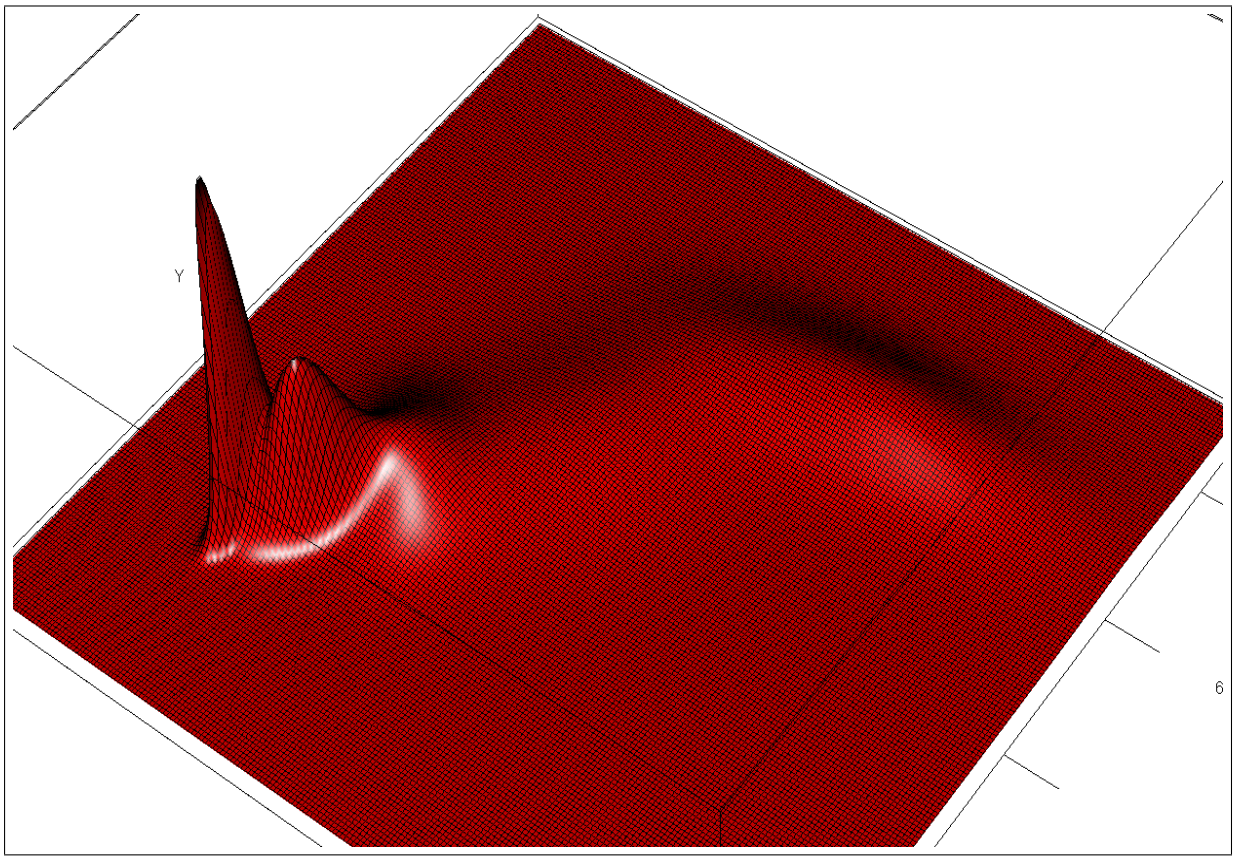


Abb. 3.7: Gauss-Mischung geschätzt von Mclust

Wir führten wieder die Funktion `ridgeline_merge()` mit $r^* = 0.7$ und $N = 15$ aus und erhielten erwartungsgemäß eine Clusterung aus 3 Clustern.

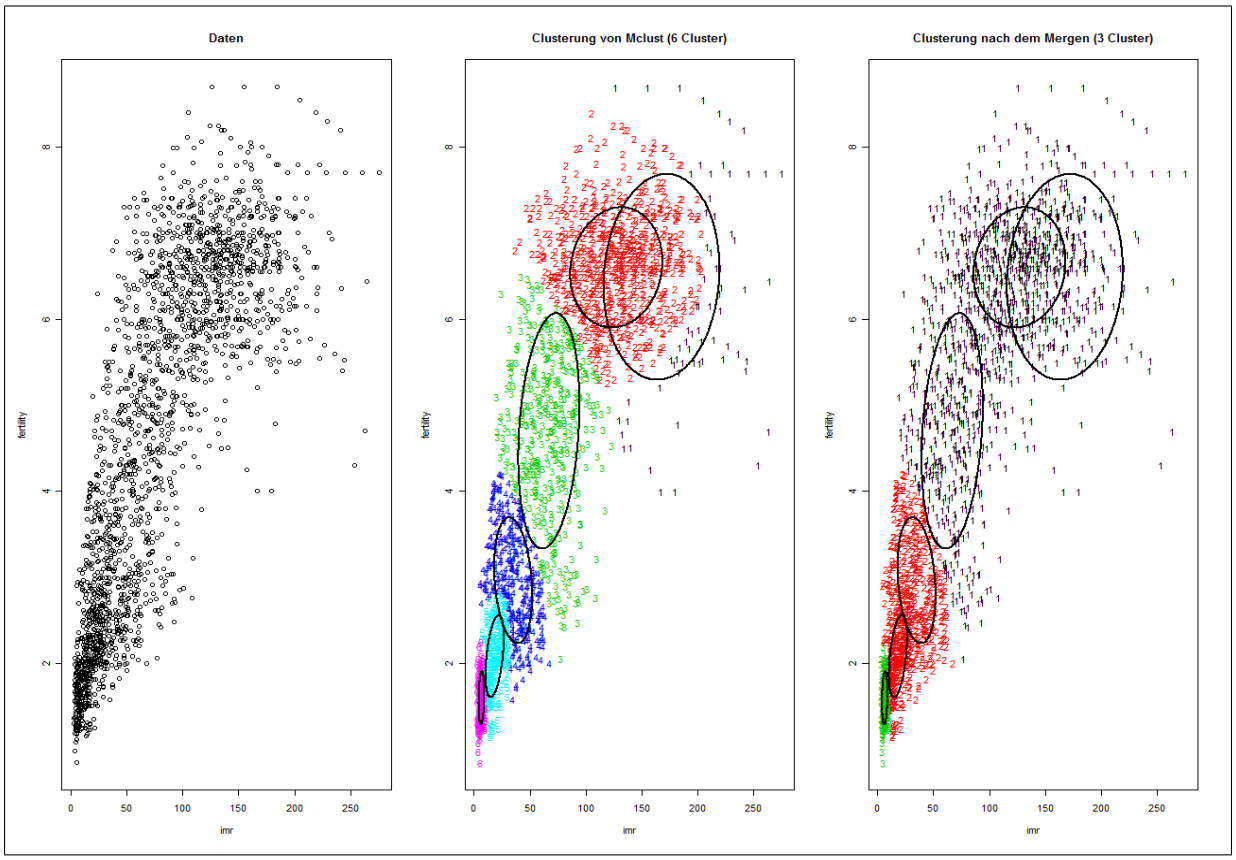


Abb. 3.8: Realer Datensatz, v.l.n.r.: simulierte Daten, Clusterung von Mclust, Clusterung nach dem Mergen. Die Ellipsen deuten die 50%-Konfidenzgebiete der jeweiligen Gauss-Komponenten an.

Iteration	**Zusammengelegte Cluster**	**Ridgeline-Quotient**
1	(1,2)	1
2	(4,5)	1
3	(1,3)	0.911

Tabelle 3.3: Experiment 4. Ergebnisse des Merging-Algorithmus

Zusammenfassend läßt sich sagen, dass die Hintereinanderausführung von Mclust und Ridgeline-Merge bessere Ergebnisse liefert als Mclust alleine. Der neue Algorithmus ist in vielen Fällen in der Lage Cluster zu erkennen, die nicht von normalverteilten Prozessen stammen. Besonders hohe Bedeutung hat die Robustheit des Merging-Ansatzes - der neue Algorithmus ist weniger datenempfindlich als Mclust. In den betrachteten Beispielen hat der Merging-Algorithmus immer die wahren Cluster gefunden, während Mclust verschiedene Clusterungen produziert hat.

Das Konstrukt Ridgeline erlaubt es uns zum einen, die Suche nach Moden und Antimoden

einer Gauss-Mischung auf eine niederdimensionale Untermannigfaltigkeit zu beschränken und zum anderen gibt es uns durch den Ridgeline-Quotient Kriterien für das Zusammenlegen von Komponenten.
Kritisch ist bei dem Verfahren die Wahl der Paramerer r^* (Schwellenparameter zwischen Unimodalität und Bimodalität) und N (Anzahl der Durchläufe der Optimierungsroutine). Falls N zu klein gewählt wird, könnte es vorkommen, daß nicht alle Moden bzw. Antimoden der Dichte gefunden werden. Dadurch könnten Cluster vereinigt werden, die in Wahrheit getrennt sind. Die Wahl von r^* sollte davon abhängen, wie stark die verschiedenen Cluster von einander getrennt sein sollen. Falls man eine stark ausgeprägte Trennung zwischen den Clustern fordert, sollte man r^* klein (≤ 0.5) wählen, falls man dagegen schon bei leichter Bimodalität von verschiedenen Clustern ausgeht, sollte man r^* groß wählen. Falls die Daten sich nicht sinnvoll durch eine Gaussmischung approximieren lassen entstehen zu viele gut getrennte Gausskomponenten pro Cluster, in diesem Fall kann auch der Merging-Algorithmus nicht die wahren Cluster identifizieren.

Kapitel 4

Ein Likelihood-Ratio Test auf Bimodalität im Zwei-Komponenten-Fall

In diesem Kapitel konstruieren wir einen Test auf Bimodalität. Unsere Ausgangssituation sind Daten einer Gaussmischung aus zwei Komponenten mit gleichen Varianzen:

$$g(x; \mu_1, \mu_2, \Sigma, \pi) = \pi\phi(x; \mu_1, \Sigma) + (1 - \pi)\phi(x; \mu_2, \Sigma).$$

Die Parameter $\mu_1, \mu_2, \Sigma, \pi$ sind unbekannt. Das Ziel ist es anhand der beobachteten Daten zu entscheiden ob die zugrunde liegende Mischung unimodal oder bimodal ist. Unsere Nullhypothese ist „g ist unimodal“. Im Kapitel 2 haben wir für den Fall proportionaler Kovarianzmatrizen $\Sigma_1 = \sigma^2\Sigma_2$ ein Kriterium angegeben wann g unimodal ist (Korollar 2.4.5). Wir geben es an dieser Stelle noch einmal für den Fall $\sigma^2 = 1$ an:

(a) Die Dichte g ist unimodal für jede Mischungsproportion π falls

$$(\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2 - \mu_1) \leq 4$$

(b) Falls die Parameter obige Bedingung nicht erfüllen, ist g bimodal genau dann, wenn $\pi \in (\pi_1, \pi_2)$, wobei

$$\frac{1}{\pi_i} = 1 + \frac{\alpha_i\phi_1(\alpha_i)}{\bar{\alpha}_i\phi_2(\alpha_i)}$$

und α_i sind die beiden Nullstellen von $q_1(\alpha) := 1 - \alpha(1 - \alpha)\mu^2$ sind und $\mu^2 = (\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2 - \mu_1)$.

Zur Erinnerung $\phi(\alpha) = \phi(x^*(\alpha))$. Der Fall $D = 1$ wurde in [13] besprochen.

4.1. Grundbegriffe der Testtheorie

4.1.1. Allgemeine Tests

Es folgt eine knappe Zusammenfassung der wichtigsten Begriffe der Testtheorie.

Definition 4.1.1. Sei $(\Omega, \mathcal{A}, \mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsräumen, $\mathcal{X} = (\mathbb{R}^D, \mathcal{B})$ ein Messraum, X eine Zufallsvariable $(\Omega, \mathcal{A}, \mathbb{P}_\theta) \to \mathcal{X}$. Sei $\Theta = \Theta_0 \cup\ \Theta_1$ die Menge der zulässigen Parameter, $\alpha \in [0, 1]$.

Wir betrachten das Entscheidungsproblem gilt $H_0 : \theta \in \Theta_0$ oder gilt $H_1 : \theta \in \Theta_1$? H_0 heißt *Nullhypothese*, H_1 heißt *Alternativhypothese*.

Eine meßbare Abbildung

$$T : \mathcal{X} \to \{0, 1\}$$

mit

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(T = 1) \leq \alpha$$

heißt *statistischer Test zum Niveau* α. Der Parameter α heißt *Signifikanzniveau des Tests* T.

Die Menge

$$K := \{x \in \mathcal{X} | T(x) = 1\}$$

heißt *Verwerfungsbereich des Tests* T.

Die Abbildung

$$\theta \mapsto \mathbb{P}_\theta(T = 1)$$

heißt *Power-Funktion des Tests* T.

Die Größe

$$\beta := \inf_{\theta\in\Theta_1} \mathbb{P}_\theta(T = 1)$$

heißt *Power des Tests* T.

Der Größen

$$\sup_{\theta \in \Theta_0} \mathbb{P}(T = 1)$$

bzw.

$$\sup_{\theta \in \Theta_1} \mathbb{P}(T = 0)$$

heißen *Fehler 1. Art* bzw. *Fehler 2. Art*.

Bei einem statistischen Test geht es also darum, anhand der beobachteten Daten zu entscheiden, ob man die Nullhypothese verwirft oder nicht. Ein Test wird typischerweise so konstruiert, dass aus den Daten x eine Teststatistik $S(x)$ berechnet wird, deren Verteilung unter H_0 bekannt ist. Falls der beobachtete Wert „zu extrem“ ist, um zufällig unter H_0 aufzutreten, wird H_0 verworfen. „Zu extrem“ kann z.B. (falls $S(X)$ reelwertig und große Werte gegen H_0 sprechen) bedeuten, dass der Wert $S(x)$ größer ist als das $(1-\alpha)$-Quantil der Verteilung von $S(X)$ unter H_0. Falls man eine große Stichprobe hat, α klein gewählt ist (≤ 0.05), und es zu einer Verwerfung von H_0 kommt, hat man eine starke Aussage und kann von Nichtgültigkeit der Nullhypothese ausgehen. Falls dagegen die Nullhypothese nicht verworfen werden kann, kann man nicht von deren Gültigkeit ausgehen.

4.1.2. Likelihood-Ratio-Tests

Jetzt diskutieren wir eine weit verbreitete Klasse von Tests - die Likelihood-Ratio-Tests. Getestet wird wieder $H_0 : \theta \in \Theta_0$ gegen $H_1 : \theta \in \Theta_1$. Die Testentscheidung wird aufgrund des Wertes des Likelihood-Quotienten getroffen:

Definition 4.1.2. Seien $\Theta = \Theta_0 \cup \Theta_1$, $\theta_i \neq \emptyset, i = 0, 1$, $\mathbf{x} = (x_1, \ldots, x_N)$ eine Stichprobe von iid Zufallsvariablen $X_1, \ldots, X_N$, X_i sei verteilt gemäß Dichte $f(x, \theta)$.

Seien

$$L(\theta) = \prod_{i=1}^{N} f(x_i; \theta),$$

$$\hat{\theta}_U := arg \sup_{\theta \in \Theta} L(\theta) \text{ unrestringierter Schätzer,} \tag{4.1}$$

$$\hat{\theta}_R := arg \sup_{\theta \in \Theta_0} L(\theta) \text{ restringierter Schätzer.} \tag{4.2}$$

Die Größe

$$\lambda_N := \frac{log[L(\hat{\theta}_R)]}{log[L(\hat{\theta}_U)]} \tag{4.3}$$

heißt *Likelihood-Quotient* der Stichprobe $\mathbf{x}$.

Es ist klar, dass falls der wahre Parameter im Inneren von Θ_0 liegt, beide Schätzer für große N identisch sind und $\lambda_N \xrightarrow{\mathcal{D}} \delta_1$, wobei δ_1 das Dirac-Maß im Punkt 1 ist.
Wenn das wahre θ am Rand von Θ_0 liegt, läßt sich die Verteilung von λ_N nicht exakt angeben.
In einigen Speziallfällen existieren Aussagen für die asymptotische Verteilung von

$$\nu_N := -2log(\lambda_N) \tag{4.4}$$

unter gewissen Regularitätsbedingungen. Es sind im Wesentlichen die Regularitätsbedingungen für die asymptotische Normalität des Maximum-Likelihood-Schätzers (s. Annahmen 1-3 unten).

So gilt z.B. im Fall $H_0 : \theta_1 = \ldots = \theta_r = 0$ und H_0 ist wahr: $\nu_N \xrightarrow{D} \chi^2_r$ (s. z.B. Ferguson).

Chernoff hat gezeigt, dass falls das wahre θ an der Grenze zwischen Θ_0 und Θ_1 liegt und Θ_0 und Θ_1 in einer Umgebung von dem wahren θ durch disjunkte Kegel C_{Θ_0} bzw. C_{Θ_1} approximierbar sind, die Verteilung von ν_N, dieselbe ist wie im Test $\theta \in C_{\Theta_0}$ gegen $\theta \in C_{\Theta_1}$ basiert auf **einer** Beobachtung von $N(\theta, J^{-1})$. Dabei gilt $J := E\nabla^2 log[f(X)]$. Daraus folgt in vielen Fällen, dass die asymptotische Verteilung von ν_N eine Mischung von χ^2-Verteilungen verschiedener Grade ist.

Wir präsentieren an dieser Stelle ein wichtiges Ergebnis aus dem Paper von H.Holzmann und S.Vollmer [13] über die asymptotische Verteilung von ν_N bei einem Test auf Bimodalität einer Mischung $g(x; \theta_1, \theta_2, \pi) = \pi f(x; \theta_1) + (1 - \pi) f(x; \theta_2)$, wobei $x \in \mathbb{R}^D$, $\theta_i \in \mathbb{R}^P$, $\pi \in [0, 1]$, die aus dem chernoff'schem Satz folgt. Der zulässige Parameterraum für die Mischungsdichte Θ habe Dimension $q \leq 2P + 1$, ($\leq$ damit z.B. der Fall gleicher oder proportionaler Kovarianzmatrizen) zugelassen wird. Es gilt $\Theta = \Theta_0 \cup \Theta_1$, wobei die Mischungsdichte g unimodal ist, falls $(\theta_1, \theta_2, \pi) \in \Theta_0$ und sonst bimodal.
Wir machen zuerst drei Annahmen über die zugrunde liegenden Verteilungen und den wahren Parameter $(\theta_1^0, \theta_2^0, \pi^0)$:

Annahme 1 Die partiellen Ableitungen von $log[g(x;\theta_1,\theta_2,\pi)]$ 3. Ordnung nach θ_1,θ_2 und π existieren in einer Umgebung $U(\theta_1^0,\theta_2^0,\pi^0)$.

Annahme 2 Für $(\theta_1,\theta_2,\pi) \in U(\theta_1^0,\theta_2^0,\pi^0)$ sind die Beträge der 1. und 2. partiellen Ableitungen von g gleichmässig beschränkt durch eine Funktion $F(x) \in L_1(\mathbb{R})$, und die Beträge der 3. partiellen Ableitungen von $log(g)$ sind gleichmässig beschänkt durch eine Funktion $H(x)$ mit $\mathbb{E}H(X_1) \leq \infty$

Annahme 3 $\mathbb{E}\nabla^2 log\, g(X_1;\theta_1,\theta_2,\pi)$ existiert in $U(\theta_1^0,\theta_2^0,\pi^0)$ und ist dort positiv definit.

Satz 4.1.3. *Der wahre Parameter $(\theta_1^0,\theta_2^0,\pi^0)$ der Mischungsdichte g liege auf dem Rand $\partial\Theta_0$, der Θ_0 und Θ_1 voneinander trennt. In einer Umgebung $U(\theta_1^0,\theta_2^0,\pi^0)$ sei der Rand $\partial\Theta_0$ eine glatte $q-1$-dimensionale Untermannigfaltigkeit im $\mathbb{R}^q$. Falls die Annahmen 1-3 erfüllt sind, gilt:*

$$\nu_N := 2\left(\sup_{(\theta_1,\theta_2,\pi)\in\Theta} log\, L(\theta_1,\theta_2,\pi) - \sup_{(\theta_1,\theta_2,\pi)\in\Theta_0} log\, L(\theta_1,\theta_2,\pi)\right) \xrightarrow{\mathcal{D}} \frac{1}{2}(\delta_0 + \chi_1^2), \qquad (4.5)$$

Die Aussage folgt aus dem Satz von Chernoff über die asymptotische Verteilung des Likelihood-Quotienten, s. [4].

4.2. Test auf Bimodalität

In diesem Abschnitt diskutieren wir einen Likelihood-Ratio-Test auf Bimodalität einer Gauss-Mischung mit gleichen Kovarianzen. Um einen solchen Test zu konstruieren, müssen wir zunächst unser Testproblem aufstellen. Die Nullhypothese H_0 lautet „g ist unimodal“ und der zugehörige Parameterraum der Nullhypothese lautet mit $\mu^2 := (\mu_2-\mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2-\mu_1)$:

$$\Theta_0 = \left\{\mu^2 \leq 4\right\} \cup \left\{\mu^2 > 4\right\} \cap \left\{\pi \notin (\pi_1,\pi_2)\right\}. \qquad (4.6)$$

Die Grenzen π_1 und π_2 ergeben sich aus $\frac{1}{\pi_i} = 1 + \frac{\alpha_i\phi_1(\alpha_i)}{\bar{\alpha}_i\phi_2(\alpha_i)}$ und $1+\mu^2\alpha_i^2-\mu^2\alpha_i = 0$. Lösen der quadratischen Gleichung liefert

$$\alpha_{1,2} = \frac{\mu^2 \mp \sqrt{\mu^4-4\mu^2}}{2\mu^2}. \qquad (4.7)$$

Damit könnnen wir π_i ausrechnen.

Jetzt wollen wir den Likelihood-Quotienten berechnen. Das ist ein nicht einfaches Problem, da die Optimierung der Likelihood-Funktion über Θ_0 problematisch ist.
Die Likelihoodfunktion bei einer gegebenen Stichprobe $\mathbf{x}$ lautet

$$L(\theta_1, \theta_2, \pi) = \prod_{i=1}^{N} g(x_i, \theta_1, \theta_2, \pi) = \prod_{i=1}^{N} \pi f(x_i, \theta_1) + (1-\pi) f(x_i, \theta_2)$$

und die Log-Likelihood-Funktion lautet

$$l(\theta) = log \prod_{i=1}^{N} \pi f(x_i, \theta_1) + (1-\pi) f(x_i, \theta_2) = \sum_{i=1}^{N} log[\pi f(x_i, \theta_1) + (1-\pi) f(x_i, \theta_2)],$$

s. Formel (3.3). Im Folgenden gehen wir davon aus, dass die Likelihood-Funktion ein Maximum besitzt. Das Optimierungsproblem für den restringierten Schätzer lautet

$$\begin{aligned} &\max\ l(\theta) \\ &\text{s.t. } \theta \in \Theta_0 \end{aligned}$$

Der zulässige Bereich für dieses Optimierungsproblem Θ_0 zerfällt in 2 disjunkte Mengen

$$M_1 := \left\{ \mu_1, \mu_2, \Sigma_1, \pi \,|\, (\mu_2 - \mu_1)^{\mathsf{T}} \Sigma_1^{-1} (\mu_2 - \mu_1) \leq 4 \right\} \tag{4.8}$$

$$M_2 := \left\{ \mu_1, \mu_2, \Sigma_1, \pi \,|\, (\mu_2 - \mu_1)^{\mathsf{T}} \Sigma_1^{-1} (\mu_2 - \mu_1) > 4,\ \pi \notin (\pi_1, \pi_2) \right\}, \tag{4.9}$$

wobei $\pi_i = \pi_i(\mu_1, \mu_2, \Sigma_1)$.

Im Folgenden benutzen wir die Notation $\theta := (\mu_1, \mu_2, \Sigma, \pi)$. Wir lösen das Optimierungsproblem auf folgende Weise:
1. Berechne unrestringierten Schätzer $\hat{\theta}_U$.
2. Falls $\hat{\theta}_U \in M_1$ oder $\hat{\theta}_U \in M_2$ setze $\hat{\theta}_R = \hat{\theta}_U$.
Sonst setze $\hat{\theta}_R = arg \max_{\theta} \left\{ \max_{\theta \in \partial M_1} l(\theta), \max_{\theta \in M_2} l(\theta) \right\}$.

Wir optimieren also separat über die beiden Teilmengen. Falls der unrestringierte Schätzer nicht in M_1 liegt, können wir davon ausgehen, dass der auf M_1 beschränkte Schätzer auf dem Rand von ∂M_1 liegen wird. Deshalb bertachten wir nur den Rand.
Jetzt behandeln wir die beiden Optimierungsteilprobleme.

Problem 1

$$\begin{aligned} &\max\ l(\theta) \\ &\text{s.t. } \theta \in \partial M_1 \end{aligned}$$

$\Leftrightarrow$

$$\begin{aligned} &\max\ l(\mu_1, \mu_2, \Sigma_1, \pi) \\ &\text{s.t. } (\mu_2 - \mu_1)^\mathsf{T} \Sigma_1^{-1} (\mu_2 - \mu_1) = 4 \\ &\pi \in [0, 1] \end{aligned}$$

Dazu lösen wir eine Folge von folgenden Optimierungsproblemen:

$$\begin{aligned} \max\ l_1^k(\mu_1, \mu_2, \Sigma, \pi) &:= l(\mu_1, \mu_2, \Sigma, \pi) - [4 - (\mu_2 - \mu_1)^\mathsf{T} \Sigma_1^{-1} (\mu_2 - \mu_1)]^2 c_k \qquad (4.10) \\ \text{s.t. } \pi &\in [0, 1], \end{aligned}$$

wobei c_k eine Strafkonstante der Iteration k ist, die jedes mal, wenn die Lösung θ^* ausserhalb des zulässigen Bereiches liegt um das 10-fache vergößert wird $c_{k+1} = 10c_k$ und $c_1 = 1$. Die Likelihood-Funktion wurde also um eine Penaltyfunktion $T_1 := [4 - (\mu_2 - \mu_1)^\mathsf{T} \Sigma_1^{-1} (\mu_2 - \mu_1)]^2 c_k$ ergänzt, die dafür sorgt, dass die Lösung im zulässigen Bereich bleibt. Die Verletzung der Restriktion $(\mu_2 - \mu_1)^\mathsf{T} \Sigma_1^{-1} (\mu_2 - \mu_1) = 4$ wird also bestraft und führt zu schlechteren Funktionswerten der Zielfunktion, falls c_k groß genug ist, also mindestens so groß wie der (unbekannte) Lagrange-Multiplikator der Restriktion im Optimum, wird die Optimierungsroutine eine Lösung finden in der die Nebenbedingungen eingehalten werden. Da wir a-priori nicht wissen wie groß die Strafkonstante sein muss, gehen wir iterativ vor mit immer größeren Strafkonstanten, bis wie eine Lösung erhalten, die im zulässigen Bereich liegt.

Problem 2

$$\begin{aligned} &\max\ l(\theta) \\ &\text{s.t. } \theta \in M_2 \end{aligned}$$

Um dieses Problem zu lösen, definieren wir eine neue Zielfunktion, die nur implizit von π abhängt und auch penalisiert wird:

$$l_2^k(\mu_1, \mu_2, \Sigma) := l(\mu_1, \mu_2, \Sigma, \pi^*) - \max\left\{(\mu_2 - \mu_1)^\mathsf{T} \Sigma_1^{-1} (\mu_2 - \mu_1) - 4, 0\right\} c_k \qquad (4.11)$$

wobei $\pi^* = arg \max_{\pi \notin (\pi_1, \pi_2)} l(\mu_1, \mu_2, \Sigma, \pi)$, mit

$$\pi_1 = \left[1 + \frac{(\mu^2 - \sqrt{\mu^4 - 4\mu^2})\phi_1(x^*(\frac{\mu^2 - \sqrt{\mu^4 - 4\mu^2}}{2\mu^2}))}{(\mu^2 + \sqrt{\mu^4 - 4\mu^2})\phi_2(x^*(\frac{\mu^2 - \sqrt{\mu^4 - 4\mu^2}}{2\mu^2}))} \right]^{-1}, \tag{4.12}$$

$$\pi_2 = \left[1 + \frac{(\mu^2 + \sqrt{\mu^4 - 4\mu^2})\phi_1(x^*(\frac{\mu^2 + \sqrt{\mu^4 - 4\mu^2}}{2\mu^2}))}{(\mu^2 - \sqrt{\mu^4 - 4\mu^2})\phi_2(x^*(\frac{\mu^2 + \sqrt{\mu^4 - 4\mu^2}}{2\mu^2}))} \right]^{-1} \tag{4.13}$$

s. Formel (4.7). Als Startwerte für die Optimierungsprobleme nehmen wir die unrestringierte Lösung. Für die Maximierung der Funktionen benutzen wir die R-Funktion `nlminb()`.

Um die asymptotische Verteilung von ν_N, für $\theta \in \partial\Theta_0$ zu untersuchen, führen wir eine Simulation durch. Wir generieren 500 Punkte aus einer Gauss-Mischung mit folgenden Parametern:

$$\mu_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \mu_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix} \text{ und } \pi = 0.14$$

und berechnen die Teststatistik ν_N. Dieser Vorgang wird 1000 mal wiederholt. Die gewählte Parameterkonstellation liegt auf dem Rand des Unimodalitätsbereiches, die Wahl von π entspricht π_1 in (4.12). Am Ende erhalten wir eine Stichprobe von ν_N mit $N = 1000$.

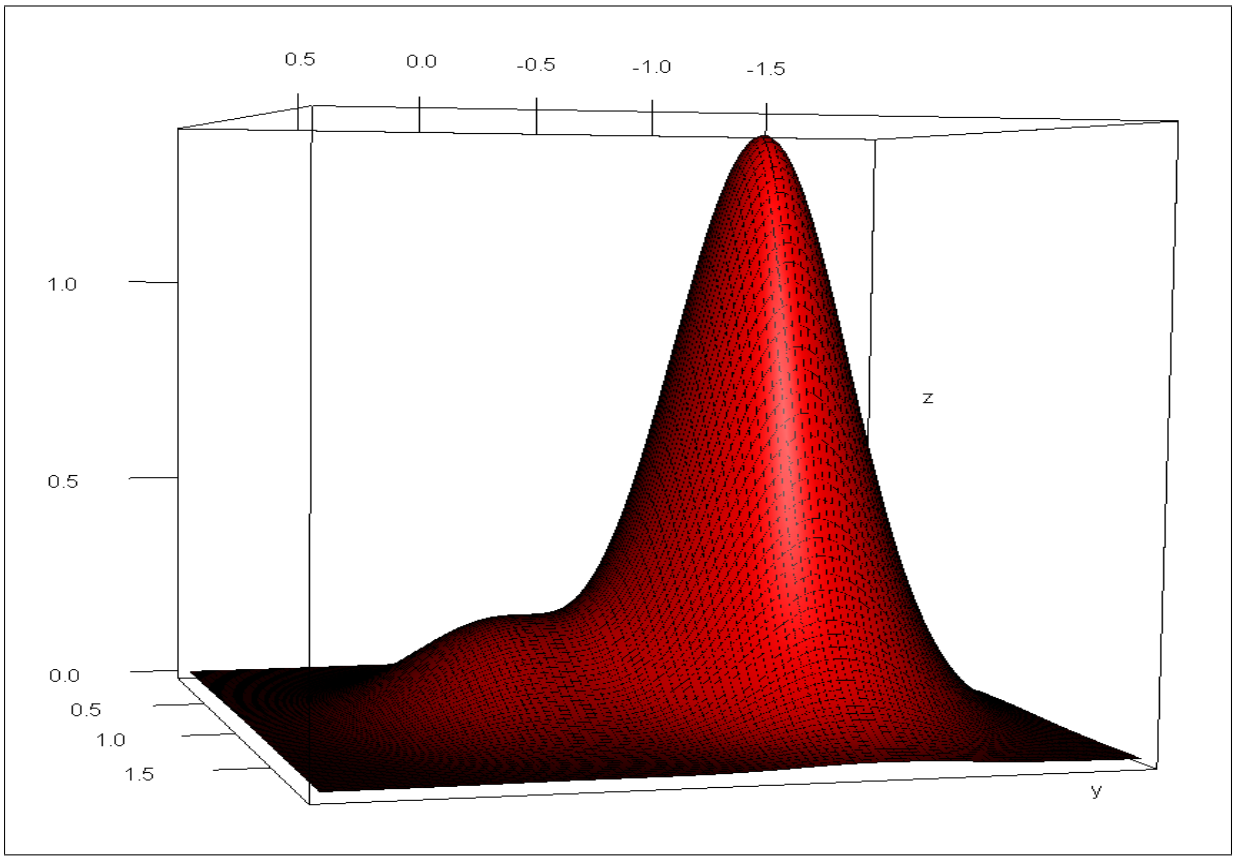


Abb. 4.1: Dichte der simulierten Gauss-Mischung auf Rand der Unimodalität

Es folgen einige Statistiken zu der erhaltenen Stichprobe:

Min	**1. Quartil**	**Median**	**Mittelwert**	**3. Quartil**	**Max**
0.000000	0.000000	0.006211	0.583200	0.634800	9.169000

Tabelle 4.1: Zusammenfassung einer Stichprobe von ν_{500}

Die simulierte Stichprobe enthält 485 Nullen. Die Dichteschätzung der positiven Werte des Log-Likelihood-Quotienten und ein QQ-Plot mit der Chi-Quadrat-Verteilung mit 1 Freiheitsgrad sind auf der folgenden Abbildung angegeben:

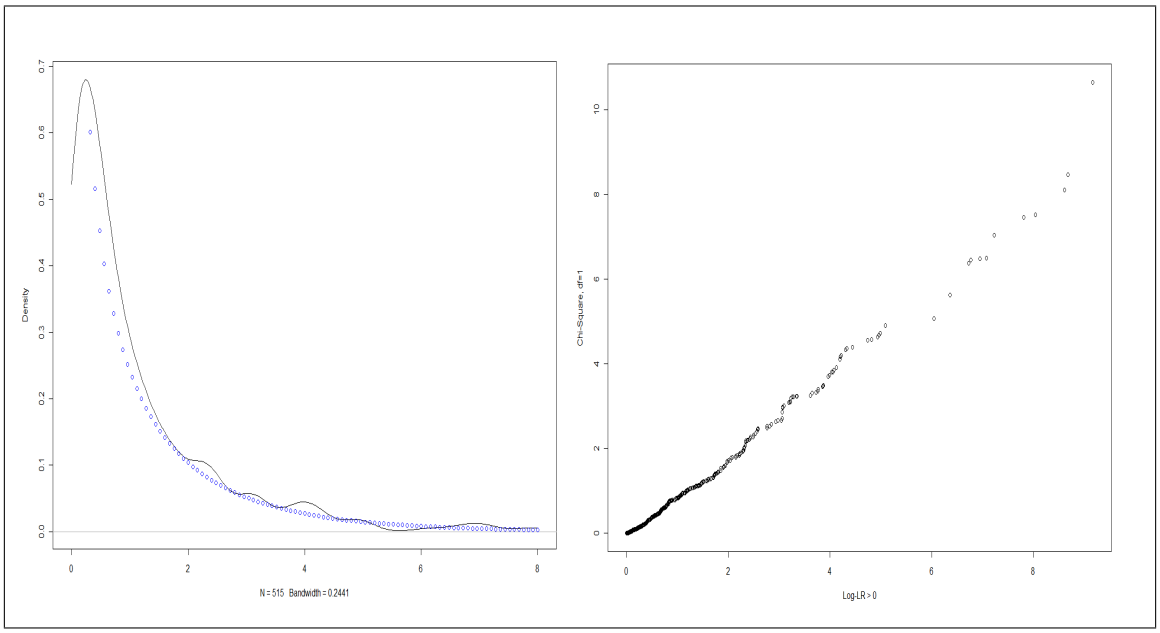


Abb. 4.2: Links: Dichteschätzung der positiven Werte von ν_{500} (schwarze Linie) und Dichte einer χ^2_1-Verteilung. Rechts: QQ-Plot mit χ^2_1-Verteilung

Die Verteilung der postiven Werte der Teststatistik läßt sich durch die χ^2_1-Verteilung gut approximieren. Knapp die Hälfte der Werte war 0 (485 von 1000). Die Annahmen für den Satz 4.1.3 scheinen in dem gewählten Punkt zu gelten. Der Rand von Θ_0 läßt sich also lokal um θ durch eine Hyperebene approximieren.

Die Annahmen gelten aber nicht überall, wie ein Blick auf die Bimodalitätsregion zeigt (s. Abbilung 4.3). Die Bimodalität hängt in unserem Spezialfall ($\Sigma_1 = \Sigma_2$) nur von $(\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2 - \mu_1)$ und π ab. Wie man aus der Abbildung erkennen kann, läßt sich die Menge um den Punkt $(\mu_2 - \mu_1)^\mathsf{T}\Sigma_1^{-1}(\mu_2 - \mu_1) = 4$ und $\pi = 0.5$ nicht durch eine Hyperebene approximieren (wegen der Spitze), deshalb gilt dort auch nicht die asymptotische Verteilung von ν_N aus dem Satz 4.1.3. Einige Statistiken aus einer Simulation von ν_{500} in dieser Ecke sind in der Tabelle 3.2.2 zusammengefasst.

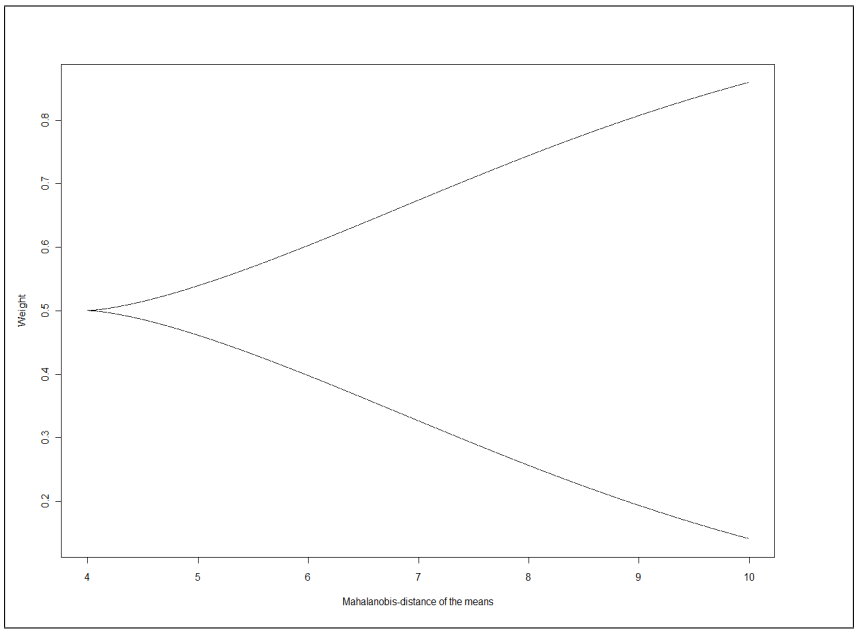


Abb. 4.3: Bimodalitätsregion der Gaussmischung

Min	**1. Quartil**	**Median**	**Mittelwert**	**3. Quartil**	**Max**
0.000000	0.000000	0.000000	0.1187	0.000000	4.0980

Tabelle 4.2: Zusammenfassung einer Stichprobe von ν_{500} in einem nichtregulären Punkt

790 Mal war der Wert der Stichprobe 0. Die empirischen Verteilungsfunktionen aus den beiden Simulationen sind auf der folgenden Abbildung angegeben.

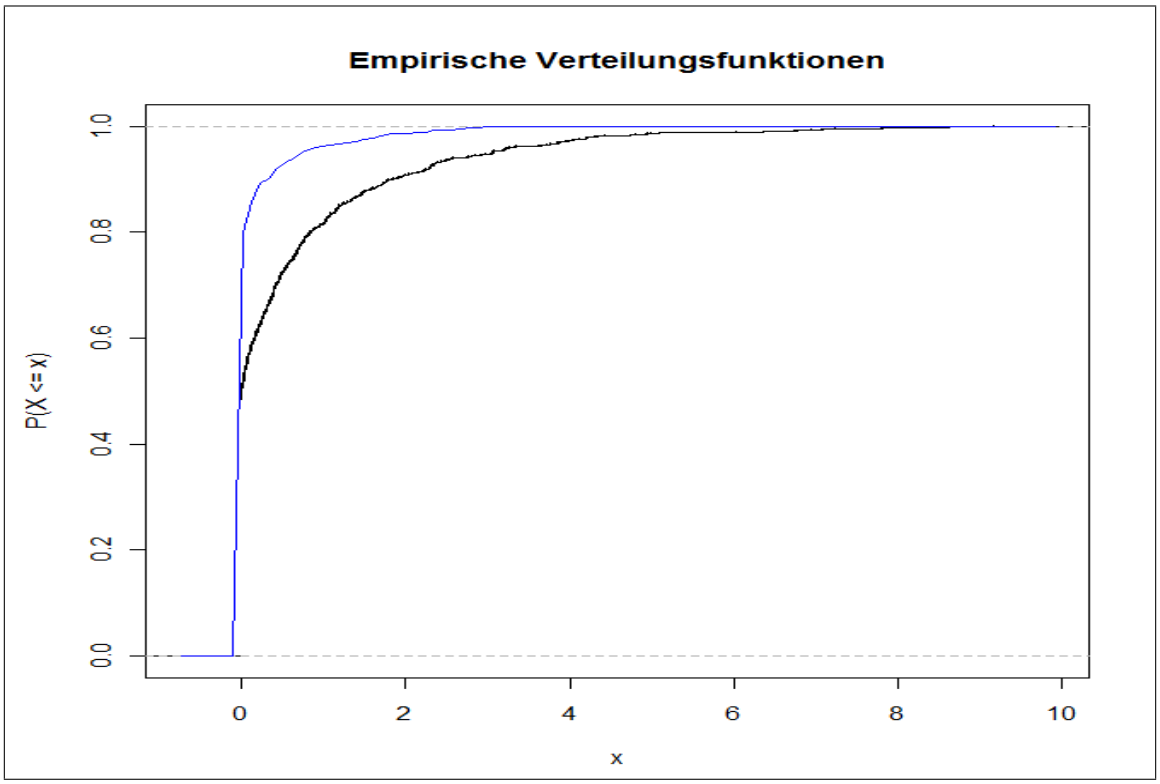


Abb. 4.4: Empirische Verteilungsfunktionen von ν_{500}: im regulären Fall (schwarz, untere Kurve) und in der Spitze des Bimodalitätsbereiches (blau, obere Kurve).

Die Verteilung der Teststatistik scheint eine Mischung aus dem Diracmaß im Punkt 0 und einer Verteilung zu sein, deren Masse näher am Ursprung liegt, als es bei χ^2_1 der Fall ist. Das Gewicht des Diracmaßes in der Mischung ist ca. 0.8. Ein Test, der ausgehend von $T := \frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1$ als Verteilung von ν_N, die Nullhypothese verwerfen würde, falls $\nu_N > q_{1-\alpha}$ ist, würde das Niveau α auch im nichtregulären Punkt für große N einhalten. Die Größe $q_{1-\alpha}$ ist das $1-\alpha$ Quantil von T.

Um die Wirkung der Restriktionen an den ML-Schätzer zu veranschaulichen, betrachten wir jetzt eine bimodale Mischung mit folgenden Parametern:

$$\mu_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix} \text{ und } \pi = 0.5.$$

Aus dieser Mischung simulieren wir 100 Punkte und berechnen $\hat{\theta}_U$ und $\hat{\theta}_R$ (siehe 4.1). Die zugehörigen Mischungen sind auf Abbildung 4.5 zu sehen. Der zugehörige Wert des Log-Likelihood-Quotioneten ν_N ist 31.013. Dieser Wert wäre unter wahrer Nullyhypothese unwahrscheinlich.

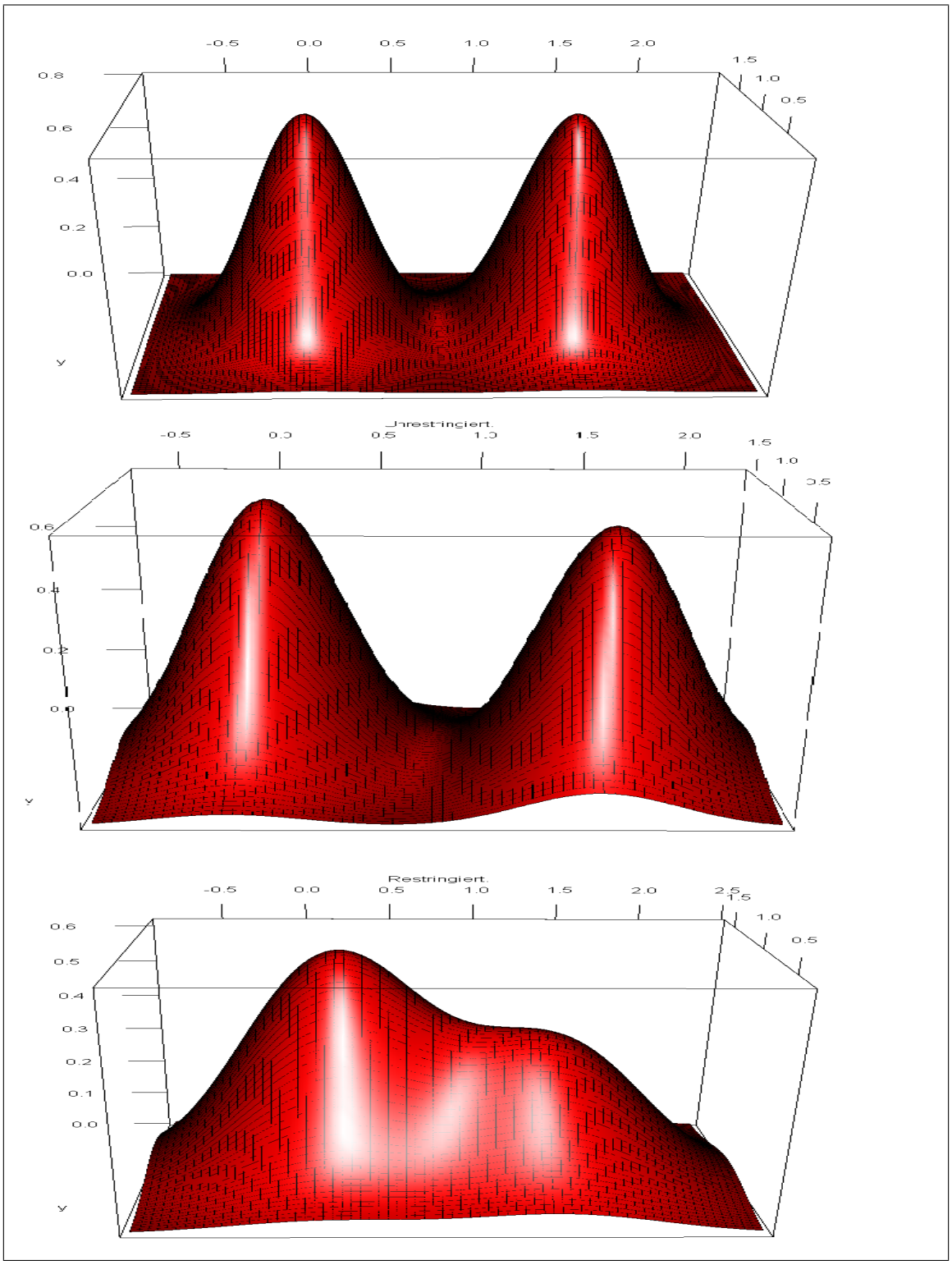


Abb. 4.5: Von oben nach unten: Originaldichte, unrestringierte Schätzung, restringierte Schätzung.

Schlusswort, offene Fragen

In dieser Arbeit wurden einige analytische Eigenschaften von Mischungen elliptischer Verteilungen untersucht, besonderes Augenmerk galt der Normal- und der t-Verteilung. Es wurde ein Clustering-Algorithmus vorgestellt, der auf Mclust aufbaut und seine Ergebnisse oft verbessern kann. Ein zentrales Konstrukt, der dieser Arbeit zugrunde liegenden Theorie, ist die Ridgeline.

Im 2. Kapitel wurden Schranken für die Modenanzahl in verschiedenen Fällen hergeleitet. Eine interessante Fragestellung in diesem Zusammenhang wäre die Verifikation ob die Schranken scharf sind oder nicht. Im letzteren Fall wäre die Herleitung einer scharfen Schranke von Interesse. Im Gauss-Fall sind die Schranken für $D = 1$ und $D = 3$ bekannterweise scharf, für andere Dimensionen sind uns keine Beispiele bekannt, wo die Schranken angenommen werden.

Weiterhin sind analoge Schranken für Mischungen aus K Komponenten für $K \geq 3$ bis auf die Speziallfälle aus Satz 2.2.1 unbekannt. Angesichts der Komplexität der Rechnungen für $K = 2$ stellt sich eine direkte Analyse der Modalitätsanzahl über die Ridgeline als kein einfaches Problem dar.

Im 4. Kapitel konstruierten wir einen Test auf Unimodalität für eine Mischung aus 2 Gauss-Komponenten mit $\Sigma_1 = \Sigma_2$. Im Korollar 2.4.5 sind Kriterien für den allgemeineren Fall $\Sigma_1 = \sigma^2\Sigma_2$ angeben, eine Erweiterung des Tests auf diesen Fall wäre interessant, ist aber mit höherem Aufwand bei der restringierten Optimierung im Rahmen der Likelihood-Ratio-Statistik verbunden.

Danksagung

An dieser Stelle möchte ich Herrn Prof. Dr. H. Holzmann für seine ausgezeichnete Betreuung danken. Ich danke weiterhin F. Ketterer und F. Schwaiger für ihre wertvollen Tipps.

Erklärung

Hiermit versichere ich, dass ich die vorgelegte Diplomarbeit selbstständig verfasst und noch nicht anderweitig zu Prüfungszwecken vorgelegt habe. Alle benutzten Quellen und Hilfsmittel sind angegeben.

Marburg, den 30.01.2011

Literaturverzeichnis

[1] Azzalini, A., Capitanio, A.: *Statistical applications of the multivariate skew-normal distribution*. Journal of the Royal Statistical Society, 61 (1999), pp. 579-602

[2] Bronstein I. N., Semendjajew K. A., Musiol G., Muehlig H.: *Taschenbuch der Mathematik*, Harri Deutsch, 2005.

[3] Carreira-Periñán M., Williams, C.,: *On the number of Modes of a Gaussian Mixture*. Lecture Notes in Computer Science, 2695 (2003), pp. 625-640

[4] Chernoff, H.: *On the distribution of the likelihood ratio*. Ann. Math. Statist. 25 (1954), pp. 573-578

[5] Davison, A.C.: *Statistical Models*, Cambridge University Press, 2009.

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

[7] Fang, K.T., Kotz, S., Ng, K.W.: *Symmetric Multivariate and Related Distributions*. Chapman & Hall, 1989.

[8] Ferguson, T.: *A Course in Large Sample Theory*. Chapman & Hall, 1996.

[9] Fraley, C., Raftery, A.E.: *Model-Based Clustering, Discriminant Analysis, and Density Estimation*. Journal of the American Statistical Association, 97 (2002), pp. 611-631

[10] Fraley, C., Raftery, A.E.: *Technical Report MCLUST Version 3 for R*, 2006.

[11] Hastie, T., Tibshirani R., Jerome F.: *The Elements of Statistical Learning*. Springer, 2001.

[12] Hennig, C.: *Methods for merging Gaussian mixture components*. Research Report no. 302 , Department of Statistical Science, UCL, 2009.

[13] Holzmann, H., Vollmer, S.: *A likelihood ratio test for bimodality on two-component mixtures - with application to regional income distribution in the EU*. AStA - Advances in Statistical Analysis, 92 (2008), pp. 57-69.

[14] Peel, D., McLachnan, G.J.: *Robust mixture modelling using the t-distribution*. Statistics and Computing, 10 (2000), pp. 339-348.

[15] Ray, S., Lindsay, B.G.: *The Topography of Multivariate Normal Mixtures*, Ann. Statist. 33, Number 5 (2005), pp. 2042-2065

[16] Shirayev, A. N.: *Probability*. Springer, 1995.