



LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK



Nonparametrische Bayes-Inferenz in additiven gemischten Modellen

Diplomarbeit in Statistik

Verfasser: Felix Heinzl
Betreuer: Prof. Dr. Ludwig Fahrmeir
Dr. Thomas Kneib
Abgabetermin: 09. Februar 2009

Danksagungen

Einigen Personen gebührt mein Dank für ihre Unterstützung bei der Erstellung dieser Arbeit:

- Prof. Dr. Ludwig Fahrmeir für seine hervorragende Betreuung und seine konzeptuellen Vorschläge,
- Dr. Thomas Kneib für seinen nimmermüden Einsatz und seine Hilfe in theoretischen wie praktischen Fragestellungen,
- Daniel Sabanés Bové für seine Starthilfe bei dem Programm C++,
- Manuel Eugster für seine Tipps hinsichtlich der Verlinkung der Programme R und C/C++,
- Nora Fenske für ihre Informationen bezüglich der LISA-Daten,
- Monia Mahling für ihre zuverlässige Unterstützung vor allem hinsichtlich des Softwarepakets Latex,
- Stefan Cieczynski, der für einen regen Gedankenaustausch immer zur Verfügung stand
- und meinen Eltern, die mir immer mit Verständnis und Unterstützung dienten.

Inhaltsverzeichnis

Notation	7
1 Problemstellung	9
2 Bayes-Inferenz	13
2.1 Grundmodell der statistischen Inferenz	13
2.2 Grundlagen der Bayes-Inferenz	14
2.3 Markov-Chain-Monte-Carlo-Verfahren	16
3 Regressionsmodelle	21
3.1 Das lineare Modell	21
3.2 Nonparametrische Regression	23
3.2.1 Polynom-Splines	24
3.2.2 Penalisierungsansätze	28
3.3 Das lineare gemischte Modell	31
4 Dirichlet-Prozesse	35
4.1 Definition des Dirichlet-Prozesses	36
4.2 Eigenschaften des Dirichlet-Prozesses	40
4.3 Stick-Breaking	43
4.3.1 Stick-Breaking-Repräsentation des Dirichlet-Prozesses	43
4.3.2 Stick-Breaking-Prioris	45
4.4 Pólyas Urne	47
5 MCMC-Verfahren bei Dirichlet-Prozessen	51
5.1 Gibbs-Sampling basierend auf Pólyas Urne	52
5.1.1 Algorithmus nach Escobar (1994)	52
5.1.2 Algorithmus nach West, Müller und Escobar (1994) und Bush und MacEachern (1996)	54
5.1.3 Algorithmus nach MacEachern (1994)	55
5.1.4 Algorithmus nach MacEachern und Müller (1998)	56
5.1.5 Algorithmus mit Hilfsvariablen	58
5.2 Gibbs-Sampling über Stick-Breaking	59
5.3 Zusammenfassung und Ausblick	62

6	Das additive gemischte Modell	65
6.1	Formulierung des additiven gemischten Modells	65
6.2	Inferenz im additiven gemischten Modell	67
6.3	Block-Gibbs-Sampler im additiven gemischten Modell	68
6.4	Implementierung	76
7	Analyse simulierter Daten	79
7.1	Mischverteilte Daten mit großen Unterschieden	80
7.2	Mischverteilte Daten mit geringen Unterschieden	85
7.3	Zusammenfassung der Simulationsergebnisse	89
8	Datenanalyse: LISA-Daten	91
8.1	LISA-Daten	91
8.2	DPM-Modell mit P-Spline, zufälligen Effekten 1. Grades und festen Effekten	94
8.3	DPM-Modell mit P-Spline und individuellen TP-Splines	101
8.4	DP-Modelle	103
9	Zusammenfassung	109
A	Beweise	111
A.1	Akzeptanzwahrscheinlichkeit beim Gibbs-Sampler	111
A.2	Erwartungswert des Dirichlet-Prozesses	111
A.3	3. Haupteigenschaft des Dirichlet-Prozesses	112
B	Bestimmung der vollständig bedingten Dichten	113
B.1	P-Spline	114
B.2	Lineares Modell	115
B.3	Lineares gemischtes Modell mit DPM-Priori	116
B.4	Lineares gemischtes Modell mit DP-Priori	119
	Literatur	121

Notation

Die Notation dieser Arbeit orientiert sich weitestgehend an Konventionen, wie sie in der Literatur üblich sind. Folgende Zusammenstellung dient einem Überblick über die verwendete Symbolik:

mathematisches Konstrukt	Schriftart	Beispiele
Skalar	kursiv und klein	x, y, z
Vektor	kursiv, klein und fett	$\mathbf{x}, \mathbf{y}, \mathbf{z}$
Matrix	kursiv, groß und fett	$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$
Dichtefunktion	kursiv und klein	$f(\cdot), p(\cdot)$
σ -Algebra	kalligraphische Symbole	\mathcal{A}, \mathcal{C}
Menge von Verteilungen	Fraktursatz	$\mathfrak{D}, \mathfrak{E}, \mathfrak{P}$

Eine Ausnahme in diesem Schema bilden die Borel- σ -Algebra, die mit \mathfrak{B} bezeichnet wird, und der in Abschnitt 2.1 verwendete Zufallsvektor \mathbf{Y} , der deshalb groß geschrieben ist, um ihn vom entsprechenden Datenvektor \mathbf{y} zu unterscheiden.

Des Weiteren sei erwähnt, dass ein Vektor stets als Spaltenvektor definiert ist.

Darüberhinaus werden folgende Bezeichnungen gebraucht:

\mathbb{R}	Menge der reellen Zahlen,
\mathbb{S}	Simplex von Wahrscheinlichkeiten,
$\Gamma(\cdot)$	Gammafunktion: $\Gamma(a) = \int_0^\infty x^{a-1} \exp(-x) dx$,
$B(\cdot, \cdot)$	Betafunktion: $B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$,
$I_A(\cdot)$	Indikatorfunktion für die Menge A ,
δ_x	Einpunktverteilung an der Stelle x .

An dieser Stelle sollen außerdem die in dieser Arbeit verwendeten Abkürzungen für Verteilungen zusammengestellt und benannt werden:

Verteilung	Kürzel	Parameter
Normalverteilung	$N(\mu, \sigma^2)$	Erwartungswert μ , Varianz σ^2
multivariate Normalverteilung	$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Erwartungswert $\boldsymbol{\mu}$, Kovarianzmatrix $\boldsymbol{\Sigma}$
Binomialverteilung	$B(n, \pi)$	Umfang n , Wahrscheinlichkeit π
Multinomialverteilung	$M(n, \boldsymbol{\pi})$	Umfang n , Wahrscheinlichkeitsvektor $\boldsymbol{\pi}$
Gammaverteilung	$Ga(a, b)$	Formparameter a , Rate b
inverse Gammaverteilung	$IG(a, b)$	Formparameter a , Rate b
stetige Gleichverteilung	$U(a, b)$	Intervallgrenzen a, b
Betaverteilung	$Be(a, b)$	
Dirichletverteilung	$Dir(\boldsymbol{\alpha})$	
verallg. Dirichletverteilung	$GD(\boldsymbol{a}, \boldsymbol{b})$	
Dirichlet-Prozess	$DP(\alpha)$	

Im Zusammenhang mit Verteilungen werden oft folgende Abkürzungen verwendet:

- i.i.d. unabhängig und identisch verteilt (**i**ndependent and **i**dentically **d**istributed),
- ind unabhängig (**i**ndependent).

1 Problemstellung

Ein wichtiges Anwendungsgebiet der Statistik beschäftigt sich mit der Analyse longitudinaler Daten. Solche Daten liegen vor, wenn mehrere Objekte über eine längere Zeitspanne beobachtet und hinsichtlich eines konkreten Merkmals zu bestimmten Messzeitpunkten untersucht worden sind. Die LISA-Daten, die in Abschnitt 8.1 näher beschrieben werden, stellen ein Beispiel für eine solche Längsschnittstudie dar. In dieser Studie ist u.a. die zeitliche Entwicklung des Body-Mass-Index (BMI) von Kindern in ihren ersten Lebensjahren von Interesse. Abbildung 1.1 veranschaulicht exemplarisch für zwölf Kinder die Messungen ihres BMIs an bis zu neun verschiedenen Messzeitpunkten.

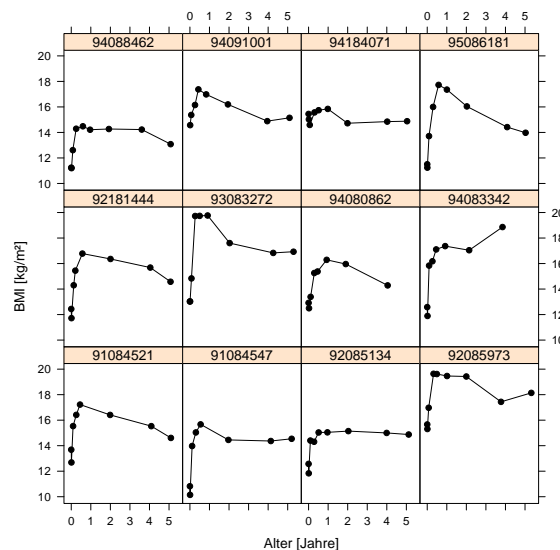


Abbildung 1.1: Individuelle BMI-Verläufe in Abhängigkeit vom Alter bei zwölf zufällig ausgewählten Kindern

Die Analyse einer solchen longitudinalen Datenstruktur sieht sich allgemein folgenden Fragen konfrontiert: Welchen generellen Zusammenhang gibt es zwischen der Zeit und dem Untersuchungsmerkmal? Gibt es bei den individuellen Verläufen bestimmte Muster? Wird der zeitliche Effekt auf das interessierende Merkmal durch andere Prädiktoren beeinflusst? Mit Hilfe eines Regressionsmodells können Antworten auf diese Fragen gefunden werden. Hierbei wird oft ein linearer Zusammenhang der Einflussgrößen auf den Response unterstellt. Im Falle der LISA-Studie wäre die Annahme eines linearen Effekts des Alters auf den BMI allerdings unzutreffend. Der BMI weist nämlich bei den meisten Kindern innerhalb des ersten Lebensjahres einen steilen Anstieg auf, wohingegen in der darauf folgenden Zeit ein leichter Rückgang zu verzeichnen ist (vgl. Abbildung 1.1). Die LISA-Daten können

daher als ein Beispiel für eine Datenstruktur angesehen werden, in der die Annahme eines linearen zeitlichen Effekts auf den Response nicht aufrecht erhalten werden kann. Eine solche Datenstruktur soll die Grundlage dieser Arbeit sein. Das Ziel wird sein, zum einen den allgemeinen nichtlinearen Effekt der Zeit auf den Response zu modellieren und zum anderen die individuellen zeitlichen Effekte zu schätzen. Gleichzeitig sollen dabei weitere Kovariablen auf einen signifikanten Effekt auf den Response untersucht werden. Dies soll im Rahmen eines additiven gemischten Modells geschehen, das für das Individuum i bei der j -ten Messung hinsichtlich der Zielvariable y folgende Struktur postuliert:

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + f(t_{ij}) + \mathbf{z}_{ij}'\mathbf{b}_i + \varepsilon_{ij}.$$

Erläuterungen zu diesem Modell werden in Kapitel 6 gegeben, wohingegen die einzelnen Bestandteile des Modells in Kapitel 3 ausführlich beschrieben werden. Diese sind wie folgt bestimmt: Zunächst wird der generelle zeitliche Effekt durch eine beliebige Funktion f beschrieben, die durch nonparametrische Inferenzmethoden geschätzt werden soll. Konkret wird hierfür ein penalisierter Spline (P-Spline) herangezogen. Der longitudinalen Struktur in den Daten wird durch ein lineares gemischtes Modell Rechnung getragen. Alle weiteren Variablen werden linear ins Modell aufgenommen.

Sämtliche im Modell zu schätzenden Parameter werden durch Methoden der Bayes-Inferenz geschätzt. Kapitel 2 liefert diesbezüglich einen kurzen Überblick. Die Anwendung bayesianischer Schätzverfahren ermöglicht im Rahmen des linearen gemischten Modells eine im Vergleich zur klassischen Sichtweise flexiblere Modellierung. Während dort typischerweise für die Verteilung der zufälligen Effekte \mathbf{b}_i eine Normalverteilung angenommen wird, erlaubt es der subjektivistische Standpunkt in der Bayes-Inferenz, jede beliebige Verteilung als Priori-Annahme hierfür zu verwenden. Noch allgemeiner kann die Verteilung der zufälligen Effekte selbst als unbekannte Größe in die Modellierung eingehen. In diesem Fall gilt es nun, für die unbekannte Verteilung eine Priori-Annahme festzulegen. Dies kann mit Hilfe einer Dirichlet-Prozess-Priori erfolgen.

Die Verwendung einer Dirichlet-Prozess-Priori bzw. des flexibleren Ansatzes einer Dirichlet-Prozess-Mischungs-Priori für die Verteilung der zufälligen Effekte steht dabei ebenso im Zentrum der Diplomarbeit wie die Kombination eines auf diese Weise gebildeten linearen gemischten Modells mit dem nonparametrischen Schätzverfahren des P-Splines. Bei Li, Lin & Müller (2007) wird von derselben Problemstellung ausgegangen – der nichtlineare zeitliche Effekt wird jedoch durch einen sog. Glättungsspline geschätzt. In diesem Sinne wird in dieser Arbeit ein hierzu alternatives und – soweit bekannt – in dieser Konstellation noch nicht verwendetes Verfahren behandelt.

Die Dirichlet-Prozesse stellen damit in ihrer Theorie und in ihrer praktischen Umsetzung einen wichtigen Bestandteil dieser Arbeit dar. Während sich Kapitel 4 mit der Theorie der Dirichlet-Prozesse befasst, werden in Kapitel 5 verschiedene Algorithmen zur Schätzung von Parametern, die aus einer unbekannten und einem Dirichlet-Prozess folgenden Verteilung stammen, erläutert und diskutiert. Die Schilderung konzentriert sich dabei im Wesentlichen auf die zwei in diesem Zusammenhang gebräuchlichen Verfahren, nämlich

die Gibbs-Sampler basierend auf Pólyas Urne und den Block-Gibbs-Sampler, der auf der sog. Stick-Breaking-Repräsentation des Dirichlet-Prozesses beruht.

Die Idee des Block-Gibbs-Samplers wird in Kapitel 6 auf das additive gemischte Modell übertragen. Hierbei wird für die Verteilung der zufälligen Effekte eine Dirichlet-Prozess-Priori bzw. eine Dirichlet-Prozess-Mischungs-Priori angenommen werden. Dieses Modell wird dann Grundlage der Analysen in den Kapiteln 7 und 8 sein.

Während sich Kapitel 7 anhand simulierter Daten in erster Linie der Fragestellung widmet, ob durch den nonparametrischen Ansatz einer unbekannten Verteilung für die zufälligen Effekte eine Verbesserung gegenüber der herkömmlichen Normalverteilungsannahme erzielt werden kann, wird in Kapitel 8 das additive gemischte Modell auf die LISA-Daten angewendet. In beiden Kapiteln liegt zudem das Augenmerk auf dem „Clustereffekt“, der mit der Verwendung eines Dirichlet-Prozesses für die Verteilung der zufälligen Effekte einhergeht. Dieser soll dazu verwendet werden, Gruppierungen hinsichtlich der Individuen zu erkennen. Bei der Analyse der LISA-Daten wird außerdem die Strategie verfolgt, das lineare gemischte Modell mit zufälligen Effekten eines bestimmten Grades auf individuelle Splines mit trunkierten Potenzen (TP-Splines) desselben Grades zu erweitern. Durch ein oder zwei zusätzliche Knoten sollen auf diese Weise die individuellen Verläufe besser erfasst werden.

2 Bayes-Inferenz

Die Grundaufgabe der statistischen Inferenz besteht darin, Schlüsse von Beobachtungen auf die zugrunde liegende Grundgesamtheit zu ziehen. Die Statistik entwickelte verschiedene, mitunter gegensätzliche Konzepte, auf welche Weise der induktive Schluss von einer Stichprobe auf die Grundgesamtheit vollzogen werden kann und wie die Unsicherheit, die diesem statistischen Schluss innewohnt, quantifiziert werden kann. Eines dieser Konzepte stellt die Bayes-Inferenz dar, die im Besonderen durch ihren subjektivistischen Standpunkt charakterisiert ist. Sie unterscheidet sich dabei klar von anderen Inferenzkonzepten wie der klassischen Inferenz und der Likelihood-Inferenz (vgl. Rüger (1999)). Bevor in Abschnitt 2.2 die Bayes-Inferenz in seinen Grundzügen dargelegt wird, soll in 2.1 kurz auf das Prinzip der statistischen Inferenz allgemein eingegangen werden. Der Abschnitt 2.3 stellt die sog. Markov-Chain-Monte-Carlo-Verfahren vor, die zur praktischen Umsetzung der Bayes-Theorie oftmals notwendig sind.

2.1 Grundmodell der statistischen Inferenz

Das Grundmodell der statistischen Inferenz gestaltet sich wie folgt: Ein reellwertiges Untersuchungsmerkmal Y folgt einer Verteilung $F(\boldsymbol{\theta})$:

$$Y \sim F(\boldsymbol{\theta}).$$

Dabei bezeichnet F einen bekannten Verteilungstyp, während $\boldsymbol{\theta}$ einen unbekannten r -dimensionalen Parametervektor mit Parameterraum $\Theta \subseteq \mathbb{R}_r$ darstellt. Ziel ist nun die Schätzung von $\boldsymbol{\theta}$ durch einen Punktschätzer $\hat{\boldsymbol{\theta}}$. Zu diesem Zwecke werden Daten y_1, \dots, y_n erhoben. Sie stellen Realisationen der Zufallsvariablen Y_1, \dots, Y_n dar, die derselben Verteilung wie Y folgen und die an dieser Stelle als unabhängig vorausgesetzt werden sollen:

$$Y_i \stackrel{i.i.d.}{\sim} F(\boldsymbol{\theta}) \quad \forall i = 1, \dots, n.$$

Der statistische Schluss entspricht nun dem Schluss von der Stichprobe $\mathbf{y} := (y_1, \dots, y_n)'$ auf den unbekannten Parameter $\boldsymbol{\theta}$, der die Verteilung der Grundgesamtheit bestimmt, aus der die y_1, \dots, y_n stammen. Der Schluss hängt also von den erhobenen Daten ab und damit auch von dem zugrunde liegenden Wahrscheinlichkeitsraum, nämlich dem des Zufallsvektors $\mathbf{Y} := (Y_1, \dots, Y_n)'$. Dieser sei durch $(\mathbb{R}_n, \mathfrak{B}_n, P(\boldsymbol{\theta}))$ bestimmt. Dabei bezeichnet \mathfrak{B}_n die n -dimensionale Borel- σ -Algebra und $P(\boldsymbol{\theta})$ die Verteilung von \mathbf{Y} . Diese ist, da sie von dem unbekannten Parameter $\boldsymbol{\theta}$ abhängt, ebenfalls nicht bekannt und lässt sich lediglich der Verteilungsfamilie $\mathfrak{P} := \{P(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ zuordnen. Diese Verteilungsfamilie

sei als dominiert vorausgesetzt, d.h. es existiert ein σ -finites Maß auf $(\mathbb{R}_n, \mathfrak{B}_n)$, das jede Verteilung $P(\boldsymbol{\theta}) \in \mathfrak{P}$ dominiert. Jede Verteilung $P(\boldsymbol{\theta})$ besitzt damit eine Dichte $p(\mathbf{y}; \boldsymbol{\theta})$ bzgl. des dominierenden Maßes.

2.2 Grundlagen der Bayes-Inferenz

Das Schätzprinzip, nach dem in der Bayes-Inferenz der statistische Schluss vollzogen wird, lässt sich durch drei Postulate beschreiben (vgl. Rüger (1999)). Das erste Bayes-Postulat folgt aus einer zur klassischen und zur Likelihood-Inferenz konträren Betrachtungsweise. Während dort der unbekannte Parameter $\boldsymbol{\theta}$ als feste, deterministische Größe aufgefasst wird, nimmt in der Bayes-Inferenz der Parameter $\boldsymbol{\theta}$ zumindest formal die Rolle einer Zufallsgröße ein. Der subjektivistische Standpunkt in der Bayes-Theorie sieht eine ganz persönliche Sichtweise des Schätzproblems vor. Die Unbekanntheit von $\boldsymbol{\theta}$ wird als persönliche Unsicherheit über $\boldsymbol{\theta}$ verstanden. Diese Unsicherheit kann von einem gewissen Vorwissen bis hin zu völliger Unwissenheit reichen. Die subjektive Herangehensweise erlaubt dem Anwender nun, das Vorwissen, das er über $\boldsymbol{\theta}$ besitzt, als a priori Annahme zu formulieren. Formal erreicht man dies über eine Verteilungsannahme bzgl. $\boldsymbol{\theta}$, der sog. Priori-Verteilung. Auf diese Weise stellt $\boldsymbol{\theta}$ formal eine Zufallsgröße dar. Gleichwohl stellt sich für einen Anhänger der Bayes-Inferenz die Frage, ob $\boldsymbol{\theta}$ objektiv betrachtet zufällig oder fest ist, nicht. Für ihn ist lediglich die rein persönliche Unsicherheit von Belang (vgl. De Finetti (1974)). Das erste Bayes-Postulat fasst dies folgendermaßen zusammen:

Erstes Bayes-Postulat
Das vor der Beobachtung vorhandene Vorwissen über den Parameter $\boldsymbol{\theta}$ wird durch eine Wahrscheinlichkeitsverteilung auf dem Parameterraum Θ , der sog. a priori Verteilung, formuliert.

In diesem Postulat steckt die schwerwiegende Annahme, dass man stets in der Lage ist, sein Vorwissen bzw. sein Nichtwissen in Form einer Verteilung auszudrücken. Diese Verteilung kann stets durch ein geeignetes σ -finites Maß dominiert werden und besitzt daher eine Dichte $p(\boldsymbol{\theta})$ bzgl. dieses Maßes. Meistens handelt es sich bei der Priori-Verteilung um eine stetige Verteilung, die durch das Lebesgue-Maß dominiert wird. Die weiteren Ausführungen in diesem Kapitel beziehen sich auf diesen Fall.

Das zweite Bayes-Postulat sieht eine Redefinition der Verteilung $P(\boldsymbol{\theta})$ vor, die angesichts der Rolle von $\boldsymbol{\theta}$ als Zufallsgröße notwendig ist:

Zweites Bayes-Postulat

Für jedes feste θ wird die Verteilung $P(\theta)$ mit der bedingten Verteilung von \mathbf{Y} gegeben θ identifiziert. Somit gilt: $p(\mathbf{y}|\theta) \equiv p(\mathbf{y}; \theta)$.

Das dritte Bayes-Postulat betrachtet die Verteilung des a posteriori Wissens $\theta|\mathbf{y}$, also des Wissens über θ nach Beobachtung der Stichprobe \mathbf{y} . Ausgehend von einer Lebesgue-stetigen Priori-Verteilung besitzt die Posteriori-Verteilung ebenfalls eine Lebesgue-Dichte. Diese lässt sich über den Satz von Bayes bestimmen:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta) p(\theta)}{\int p(\mathbf{y}|\theta) p(\theta) d\theta} = c p(\mathbf{y}|\theta) p(\theta) \propto p(\mathbf{y}|\theta) p(\theta). \quad (2.1)$$

Dabei stellt $c := [\int p(\mathbf{y}|\theta) p(\theta) d\theta]^{-1}$ eine Normierungskonstante dar. In (2.1) ist der eigentliche Kern der Bayes-Theorie enthalten: Das Wissen über θ nach Beobachtung der Stichprobe \mathbf{y} setzt sich aus zwei Komponenten zusammen. Zum einen aus dem Vorwissen, das man über θ hatte und zum anderen aus der Information, die die Stichprobe darüber liefert. Man kann also durch die Stichprobe über θ dazulernen. Konkret drückt $p(\mathbf{y}|\theta)$ die Plausibilität von θ zu der gegebenen Beobachtung \mathbf{y} aus und wird deshalb als Likelihood bezeichnet.

Das dritte Bayes-Postulat besagt nun:

Drittes Bayes-Postulat

Das nach der Beobachtung \mathbf{y} vorhandene Wissen über den Parameter θ wird durch die gemäß (2.1) bestimmte a posteriori Verteilung wiedergegeben.

Als Konsequenz aus dem dritten Bayes-Postulat gilt folgendes Korollar:

Bayes-Korollar

Ein Schluss aus einer Beobachtung darf nur von der Posteriori-Verteilung abhängen.

Dies macht die Bestimmung der Posteriori-Verteilung zu der zentralen Aufgabe in der Bayes-Inferenz. Oftmals bereitet die Integration in (2.1) große Schwierigkeiten (vgl. Abschnitt 2.3). In besonderen Fällen kann diese jedoch gänzlich vermieden werden, so dass die Bestimmung der Posteriori-Dichte kein Problem darstellt. Dies ist dann der Fall, wenn die Priori-Verteilung zur Verteilung der Stichprobe konjugiert ist. Damit ist folgendes gemeint: Angenommen die Priori-Verteilung gehört einer Familie von Verteilungen \mathfrak{D} an. Diese Verteilungsfamilie \mathfrak{D} heißt zu der Verteilungsfamilie \mathfrak{P} konjugiert, wenn für jede Priori-Verteilung aus \mathfrak{D} und jede Stichprobenverteilung aus \mathfrak{P} die Posteriori-Verteilung

ein Element von \mathfrak{D} ist. Liegt dieses vor, so steht der Verteilungstyp der Posteriori von vornherein fest und die Parameter der Verteilung können dem Produkt aus Likelihood und Priori-Dichte entnommen werden. Beispiele für solche konjugierten Paare zeigt die Tabelle 2.1.

$\mathfrak{P} = \{P(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$	\mathfrak{D}
$N(\mu, \theta = \sigma^2)$	$IG(a, b)$
$N(\theta = \mu, \sigma^2)$	$N(\mu_\theta, \sigma_\theta^2)$
$N(\boldsymbol{\theta} = \boldsymbol{\mu}, \boldsymbol{\Sigma})$	$N(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$
$B(n, \theta = \pi)$	$Be(a, b)$
$M(n, \boldsymbol{\theta} = \boldsymbol{\pi})$	$Dir(\boldsymbol{\alpha})$

Tabelle 2.1: Konjugierte Paare von Verteilungsfamilien

Basierend auf der Posteriori-Verteilung können nun Punkt- und Bereichsschätzungen angegeben werden (vgl. Rüger (1999)). Ein Punktschätzer U soll hierbei den Wert $\hat{\boldsymbol{\theta}} = U(\mathbf{y})$ von Θ angeben, für den die Posteriori am stärksten spricht. Hierfür sind die Bestimmung des Modus, des Erwartungswertes oder des Medians der Posteriori-Verteilung übliche Verfahren. Der Posteriori-Erwartungswert als Punktschätzer ist dabei wie folgt gegeben:

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{y}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \quad (2.2)$$

2.3 Markov-Chain-Monte-Carlo-Verfahren

Dem einfachen theoretischen Ansatz der Bayes-Inferenz stehen Schwierigkeiten in der praktischen Umsetzung gegenüber. In vielen Fällen ist die Posteriori-Verteilung weder analytisch noch numerisch zugänglich. So sind zur Bestimmung der Normierungskonstante c oder des Posteriori-Erwartungswertes oftmals schwierige Integrale zu berechnen, die im Falle eines hochdimensionalen Parameters in der Regel nicht lösbar sind. Wenn die Posteriori-Verteilung nicht bestimmbar ist, so kann Bayes-Inferenz jedoch mit Hilfe von Markov-Chain-Monte-Carlo- (MCMC-) Methoden durchgeführt werden. Wie bereits aus dem Namen ersichtlich wird, sind dabei zwei Konzepte von Bedeutung: Markov-Ketten und Monte-Carlo-Integration. Letzteres basiert auf der einfachen Idee, schwierige Integrationen durch die entsprechenden empirischen Analogie zu approximieren. So kann z.B. der Posteriori-Erwartungswert (vgl. (2.2)) durch das arithmetische Mittel von Zufallszahlen $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_T$ aus der Posteriori-Verteilung approximiert werden:

$$\frac{1}{T} \sum_{t=1}^T \boldsymbol{\theta}_t.$$

Von einer nicht bestimmbar Posteriore-Verteilung ausgehend ist es allerdings auch nicht möglich, direkt aus ihr Zufallszahlen zu ziehen. Um dennoch Zufallszahlen aus der Posteriori zu erhalten, müssen andere Wege beschritten werden. Eine Möglichkeit besteht darin, eine Markov-Kette zu konstruieren, die die Posteriori-Verteilung als invariante Verteilung besitzt. Nach der Konvergenzphase, der sog. „Burn In“-Phase, können die Iterationen der Markov-Kette als geringfügig abhängige Realisationen der Posteriori angesehen werden. Durch Ausdünnen kann die Abhängigkeit zwischen den Iterationen nahezu beseitigt werden. Der Metropolis-Hastings-Algorithmus liefert nun eine Methode, wie man zu gegebener Verteilung eine Markov-Kette erzeugt, die diese als invariante Verteilung besitzt. Beginnend bei einem Startwert $\theta^{(0)}$ wird sukzessive bei jeder Iteration t ein Wert θ^* vorgeschlagen und gegebenenfalls akzeptiert. Vorgeschlagen wird ein Wert, indem aus einer Vorschlagsdichte $q(\theta|\theta^{(t-1)})$ eine Zufallszahl gezogen wird. Diese hängt i.A. vom vorherigen Wert $\theta^{(t-1)}$ ab. Ein auf diese Weise vorgeschlagener Wert θ^* wird gemäß der folgenden Akzeptanzwahrscheinlichkeit als neuer Zustand $\theta^{(t)} = \theta^*$ akzeptiert:

$$\alpha(\theta^*|\theta^{(t-1)}) = \min \left\{ \frac{p(\theta^*|\mathbf{y}) q(\theta^{(t-1)}|\theta^*)}{p(\theta^{(t-1)}|\mathbf{y}) q(\theta^*|\theta^{(t-1)})}, 1 \right\}. \quad (2.3)$$

Wird θ^* nicht akzeptiert, so setzt man $\theta^{(t)} = \theta^{(t-1)}$. An (2.3) erkennt man, dass die Dichte der invarianten Verteilung nur bis auf eine Proportionalitätskonstante bekannt sein muss, da die Dichte einmal im Zähler und einmal im Nenner vorkommt. Dies ist in der Bayes-Inferenz von entscheidender Bedeutung, weil gerade die Berechnung der Normierungskonstanten c in der Posteriori-Dichte (2.1) oft nicht lösbar ist, wohingegen die Bestimmung der Posteriori-Dichte ohne Proportionalitätskonstante immer möglich ist. Somit stellt der Metropolis-Hastings-Algorithmus eine in der Bayes-Inferenz stets anwendbare Methode zur Erzeugung von Zufallszahlen aus der Posteriori-Verteilung dar. Zusammenfassend lautet er:

Metropolis-Hastings-Algorithmus:

1. Wähle einen Startwert $\theta^{(0)}$ und die Anzahl der Iterationen T . Setze $t = 1$.
2. Aufdatierungsschritt: Die Markov-Kette befinde sich im Zustand $\theta^{(t-1)}$. Ziehe einen Wert θ^* aus der Vorschlagsdichte

$$q(\theta^*|\theta^{(t-1)})$$

und akzeptiere diesen mit der gemäß (2.3) bestimmten Wahrscheinlichkeit. Andernfalls setze $\theta^{(t)} = \theta^{(t-1)}$.

3. Falls $t = T$, beende Algorithmus. Ansonsten erhöhe t um 1 und fahre fort mit 2..

Die Wahl der Vorschlagsdichte ist dabei im Grunde beliebig. Es ist jedoch darauf zu achten, dass die Vorschlagsdichte so gewählt wird, dass die Akzeptanzrate nicht zu klein

wird. Nur sehr selten angenommene Werte führen zu einer langsamen Konvergenz der Iterationen in Richtung der invarianten Verteilung. Ausreichend hohe Akzeptanzraten sind vor allem bei hochdimensionalen Parametervektoren schwierig zu erreichen. Darum wird in solchen Fällen typischerweise komponentenweise oder blockweise aufdatiert. Das Aufdatieren separater Blöcke von $\boldsymbol{\theta}$ gestaltet sich wie folgt: Zunächst wird $\boldsymbol{\theta}$ in S Blöcke $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_S$ unterteilt. Der 2. Schritt des Metropolis-Hastings-Algorithmus besteht deshalb nun aus S Stufen. Sukzessive wird in jeder Stufe $s = 1, \dots, S$ ein Wert $\boldsymbol{\theta}_s^*$ aus einer für s spezifischen Vorschlagsdichte gezogen. Zusammen mit den aktuellen Zuständen der anderen Blöcke wird damit ein Zustand $\boldsymbol{\theta}^* := (\boldsymbol{\theta}_1^{(t)}, \dots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_s^*, \boldsymbol{\theta}_{s+1}^{(t-1)}, \dots, \boldsymbol{\theta}_S^{(t-1)})'$ vorgeschlagen. Oft wird für die für s spezifische Vorschlagsdichte die vollständig bedingte Dichte verwendet. In diesem Fall spricht man vom sog. Gibbs-Sampler. Er stellt einen Spezialfall des Metropolis-Hastings-Algorithmus dar. Die vollständig bedingte Dichte ist allgemein die Dichte der bedingten Verteilung von $\boldsymbol{\theta}_s$ gegeben aller anderen Parameter $\boldsymbol{\theta}_{-s} := (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{s-1}, \boldsymbol{\theta}_{s+1}, \dots, \boldsymbol{\theta}_S)'$. Sie eignet sich deshalb als Vorschlagsdichte, da ihre Bestimmung und das Ziehen von Zufallszahlen häufig auch dann möglich ist, selbst wenn die Posteriori-Dichte nicht bestimmt werden kann. Dies lässt sich auf folgende Art und Weise veranschaulichen: Zunächst ist eine vollständig bedingte Dichte stets proportional zur Posteriori-Dichte:

$$p(\boldsymbol{\theta}_s | \boldsymbol{\theta}_{-s}, \mathbf{y}) = \frac{p(\boldsymbol{\theta} | \mathbf{y})}{p(\boldsymbol{\theta}_{-s} | \mathbf{y})} \propto p(\boldsymbol{\theta} | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}). \quad (2.4)$$

Zur Bestimmung des Kerns der vollständig bedingten Dichte sind von dem Produkt aus Likelihood und Priori-Dichte nur die $\boldsymbol{\theta}_s$ beinhaltenden Faktoren relevant. Die geringere Dimension von $\boldsymbol{\theta}_s$ im Vergleich zu $\boldsymbol{\theta}$, die beim komponentenweisen Aufdatieren sogar nur 1 beträgt, ermöglicht oftmals eine numerische oder gar analytische Bestimmung.

Darüberhinaus besitzt der Gibbs-Sampler die charakteristische Eigenschaft, dass vorgeschlagene Werte immer akzeptiert werden (Nachweis im Anhang A.1). Probleme mit zu niedrigen Akzeptanzraten sind damit von vornherein ausgeschlossen. Die Adjustierung der Iterationen $\boldsymbol{\theta}^{(t)}$ an die Daten \mathbf{y} wird beim Gibbs-Sampler ausschließlich über die Vorschlagsdichte erzielt. Ein nachträgliches Korrigieren der gezogenen Werte durch ein evtl. Nicht-Akzeptieren ist unnötig. Das macht den Gibbs-Sampler zu einem sehr effektiven Algorithmus. Er lässt sich wie folgt zusammenfassen:

Gibbs-Sampler:

1. Wähle Startwerte $\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_S^{(0)}$ und die Anzahl der Iterationen T . Setze $t = 1$.
2. Aufdatierungsschritt: Die Markov-Kette befinde sich im Zustand $\boldsymbol{\theta}^{(t-1)}$.
 - (I) Für $s = 1, \dots, S$:
 - Ziehe neuen Wert für $\boldsymbol{\theta}_s$ gemäß:

$$p(\boldsymbol{\theta}_s | \boldsymbol{\theta}_1^{(t)}, \dots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_{s+1}^{(t-1)}, \dots, \boldsymbol{\theta}_S^{(t-1)}, \mathbf{y}).$$

3. Falls $t = T$, beende Algorithmus. Ansonsten erhöhe t um 1 und fahre fort mit 2..

Für eine ausführlichere Darstellung der MCMC-Methoden sei auf Fahrmeir, Kneib & Lang (2007) verwiesen.

3 Regressionsmodelle

Die grundsätzliche Problemstellung der Regression besteht in der Untersuchung, wie diverse erklärende Variablen x_1, \dots, x_p eine Zielgröße y beeinflussen. Dabei wird das Ziel verfolgt, die Art und die Stärke des Wirkungszusammenhangs aufzudecken. Grundsätzlich geht man davon aus, dass sich die zu erklärende Variable in eine systematische und in eine stochastische Komponente zerlegen lässt. Während die systematische Komponente den zugrunde liegenden Wirkungszusammenhang beschreibt, wird durch die stochastische Komponente die zufällige Schwankung, das „Rauschen“, um diesen Effekt ausgedrückt. Es gilt nun zu untersuchen, aus welchen Einflussgrößen die systematische Komponente besteht, wie die einzelnen Effekte dabei aussehen und wie sie sich mit den anderen zusammensetzen, um so eine Zerlegung der Zielgröße in die beiden Komponenten zu erreichen, die der wahren Struktur gerecht wird. Eine der einfachsten Strukturen des Einflussterns, die dabei postuliert werden kann, stellt die Annahme eines linearen Zusammenhangs zwischen den Kovariablen und der Zielgröße dar. Ein solches lineares Modell wird im Abschnitt 3.1 erläutert, ehe in 3.2 auf den allgemeineren Ansatz der nonparametrischen Regression und in 3.3 auf das bei Longitudinal-Daten gebräuchliche lineare gemischte Modell eingegangen wird. Konzeptuell wird dabei so vorgegangen, dass die einzelnen Modelle in ihrer allgemeinen Form zunächst vorgestellt und dann in ihrer bayesianischen Umsetzung dargelegt werden. In Kapitel 6 werden diese drei Modelle zu einem additiven gemischten Modell zusammengefügt, das bei der Analyse der LISA-Daten in Kapitel 8 und in abgeschwächter Form auch bei Analyse der simulierten Daten in Kapitel 7 Verwendung findet. Die Ausführungen in diesem Kapitel orientieren sich an Fahrmeir, Kneib & Lang (2007) und Fahrmeir & Tutz (2001).

3.1 Das lineare Modell

Das lineare Modell unterstellt einen linearen Einfluss jeder Kovariablen x_r mit $r = 1, \dots, p$ auf die Zielgröße y , wobei sich die einzelnen Effekte additiv zusammensetzen. Diese systematische Komponente wird durch eine stochastische Komponente ε überlagert:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon. \quad (3.1)$$

Es wird vorausgesetzt, dass es sich bei dem Response um eine metrische und bei den Einflussgrößen um metrische oder binäre Variablen handelt. Es liegen nun Daten (y_i, \mathbf{x}_i) für $i = 1, \dots, n$ Individuen vor. Dabei entspricht $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ dem individuenspezifischen Kovariablenvektor und y_i dem Responsewert des i -ten Individuums. Da in Kapitel

6 das lineare Modell Bestandteil eines additiven gemischten Modells sein wird, soll bereits an dieser Stelle auf einen Intercept verzichtet werden. Damit gilt für die i -te Person:

$$y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i. \quad (3.2)$$

Die strukturelle Komponente ist nun durch den Prädiktor $\mathbf{x}_i' \boldsymbol{\beta}$ bestimmt, wobei $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ die Regressionskoeffizienten vektoriell zusammenfasst. Für die zufällige Komponente ε_i wird folgende Verteilungsannahme getroffen:

$$\varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2) \quad \forall i = 1, \dots, n. \quad (3.3)$$

Sie beinhaltet mehrere Aspekte: Die individuellen Abweichungen ereignen sich unabhängig voneinander und haben alle dieselbe Varianz σ^2 . In der klassischen Formulierung des linearen Modells wird i.d.R. die Normalverteilungsannahme auf die Annahme einer symmetrischen Verteilung abgeschwächt. Da jene im Rahmen der Bayes-Inferenz aber notwendig ist, wird sie an dieser Stelle gefordert.

Das individuelle lineare Modell (3.2) lässt sich über den Responsevektor $\mathbf{y} = (y_1, \dots, y_n)'$, den Vektor der Fehlervariablen $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ und der Designmatrix $\mathbf{X} = (\mathbf{x}_1', \dots, \mathbf{x}_n')'$ auch in folgende multivariate Form überführen:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Gemäß der Annahme (3.3) folgt für den Vektor der Störgrößen: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, wobei \mathbf{I}_n der n -dimensionalen Einheitsmatrix entspricht. Da der Term $\mathbf{X}\boldsymbol{\beta}$ deterministisch ist, folgt für die auf die Parameter $\boldsymbol{\beta}$ und σ^2 bedingte Verteilung von \mathbf{y} bei gegebenen Kovariablenwerten:

$$\mathbf{y} | \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n). \quad (3.4)$$

Die Designmatrix \mathbf{X} wird dabei als nichtstochastisch betrachtet, um ein Bedingen auf die konkreten Kovariablenwerte zu vermeiden und damit die Notation einfacher zu halten. Nichtsdestotrotz gilt die in (3.4) formulierte Verteilungsaussage für \mathbf{y} nur unter der Bedingung der vorliegenden Kovariablenstruktur.

In der Praxis bedeuten diese Annahmen, dass das lineare Modell dann sinnvoll anwendbar ist, wenn die Zielgröße y stetig und – gegeben die Kovariablenwerte – approximativ normalverteilt ist. Darüberhinaus ist stets der postulierte lineare Zusammenhang z.B. über eine Analyse der Residuen kritisch zu hinterfragen.

Die Aufgabe der statistischen Inferenz besteht nun darin, zu gegebenen Daten \mathbf{y} und \mathbf{X} den Parametervektor $\boldsymbol{\beta}$ und die Residuenvarianz σ^2 zu schätzen. Während das Beobachtungsmodell durch (3.4) gegeben ist, gilt es nun aus bayesianischer Perspektive eine a priori Annahme für $\boldsymbol{\beta}$ und σ^2 zu formulieren. Um die Eigenschaften der Konjugiertheit zweier Verteilungsfamilien ausnutzen zu können, bietet es sich an, folgende allgemeine Priori-Struktur zu verwenden:

$$\begin{aligned}\sigma^2 &\sim IG(a_\varepsilon, b_\varepsilon), \\ \boldsymbol{\beta} &\sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta).\end{aligned}$$

Auf diese Weise ist unter Unabhängigkeitsannahme eine gemeinsame Priori-Dichte $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}) p(\sigma^2)$ bestimmt. Es handelt sich dabei um die sog. Normal-inverse Gamma-verteilung, die zu obigem Beobachtungsmodell (3.4) konjugiert ist. Die Posteriori-Dichte lässt sich demnach analytisch berechnen:

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2)$$

Selbst bei einer solchen Festlegung der Verteilungstypen für die Priori besteht über die Hyperparameter a_ε , b_ε , $\boldsymbol{\mu}_\beta$ und $\boldsymbol{\Sigma}_\beta$ genug Freiraum für eine individuelle Formulierung des a priori Wissens. Darüber hinaus können die Hyperparameter selbst wiederum modelliert werden, wodurch man den subjektiven Einfluss auf die Schätzergebnisse gering hält und ein Maximum an Flexibilität erreicht.

Die Annahmen des bayesianischen linearen Modells lassen sich wie folgt zusammenfassen:

Bayesianisches lineares Modell

1. *Beobachtungsmodell:*

$$\mathbf{y} | \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

2. *Priori-Verteilungen:*

$$\begin{aligned}\sigma^2 &\sim IG(a_\varepsilon, b_\varepsilon), \\ \boldsymbol{\beta} &\sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta).\end{aligned}$$

3.2 Nonparametrische Regression

Oftmals ist die Annahme eines linearen Einflusses der Kovariablen x_1, \dots, x_p auf den Response y wie in (3.1) den Daten nicht angemessen. In solchen Fällen ist ein Modell, das bis auf die Additivität keine Struktur festlegt, adäquater. Ein solches additives Modell lautet:

$$y = f_1(x_1) + \dots + f_p(x_p) + \varepsilon.$$

Dabei werden mit f_r für $r = 1, \dots, p$ beliebige, für die jeweiligen Kovariablen x_r spezifische Funktionen bezeichnet. Im Folgenden werde der Einfachheit halber nur von einer

Kovariable x ausgegangen, die wie der Response als metrisch vorausgesetzt wird. Das Modell lautet nun bei n erhobenen Daten (y_i, x_i) für jedes Individuum $i = 1, \dots, n$:

$$y_i = f(x_i) + \varepsilon_i.$$

Die Funktion f wird dabei zunächst nicht näher spezifiziert. Insbesondere unterliegt sie keiner parametrischen Struktur, so dass diese Form der Regression „nonparametrisch“ genannt wird. Der Schätzung der Funktion f liegen zwei Ziele zugrunde: Einerseits soll eine Schätzung \hat{f} so weit wie möglich die Daten erfassen. Im Gegensatz zum linearen Modell, wo die Funktion f durch ihre lineare Struktur restringiert ist, kann man hier aufgrund der Beliebigkeit von f problemlos eine Schätzung \hat{f} finden, die durch alle Punkte (y_i, x_i) für $i = 1, \dots, n$ geht. Man denke hierfür z.B. an ein Polynom n -ten Grades. Ein solches „Overfitting“ spiegelt aber nur die Daten, aber nicht die dahinter stehende Struktur wider und ist daher als Regressionsansatz untauglich. Deshalb muss andererseits bei der Schätzung darauf geachtet werden, dass \hat{f} nicht zu nahe an den Daten liegt, also nicht zu rau ist. Es existieren nun verschiedene Konzepte, wie diese beiden Ziele miteinander in Einklang zu bringen sind. Solche Ansätze sind z.B. Polynom-Splines, Penalisierungsansätze oder lokale Glättungsverfahren. Im Folgenden wird eine konkrete Methode, nämlich die der penalisierten Splines, der sog. P-Splines, genau ausformuliert. Diese stellt zum einen ein Penalisierungsverfahren dar und bezieht sich zum anderen auf die Polynom-Splines. Sie stellt in diesem Sinne eine Kombination der beiden Verfahren dar. Während diese beiden im weiteren Verlauf erläutert werden, wird für andere Verfahren auf die Literatur (vgl. Fahrmeir, Kneib & Lang (2007)) verwiesen.

3.2.1 Polynom-Splines

Die naheliegendste Idee für eine Verallgemeinerung des linearen Modells, also eines Polynoms 1. Grades, wäre die Verwendung höhergradiger Polynome. Hier zeigt sich aber, dass die Modellierung mit einem Polynom für den ganzen Bildbereich der Einflussgröße nicht flexibel genug ist. Während ein solches globales Polynom bei niedrigen Graden oft nicht in der Lage ist, lokale Besonderheiten in den Daten ausreichend zu erfassen, besteht bei höherem Grad eine generelle Tendenz zum Overfitting. Ein Ansatz ist nun, in sämtlichen Teilen des Bildbereichs $[a; b]$ lokale Polynome zu schätzen und diese zu einer globalen Schätzung zu aggregieren. Knotenpunkte $a = \kappa_1 < \dots < b = \kappa_m$ dienen dabei der Festlegung der Regionen. Damit bei der Aggregation aber keine Sprünge oder Knicke auftreten, sind zusätzliche Glattheitsbedingungen erforderlich. Beim Konzept der Polynom-Splines werden diese durch die Annahme formuliert, dass die Funktion f einem sog. Polynom-Spline entspricht. Dieser ist wie folgt definiert:

Definition: Polynom-Spline

Eine Funktion $f : [a, b] \rightarrow \mathbb{R}$ heißt Polynom-Spline vom Grad $l \geq 0$ zu den Knoten $a = \kappa_1 < \dots < b = \kappa_m$, wenn sie die folgenden Bedingungen erfüllt:

1. $f(x)$ ist auf $[\kappa_j, \kappa_{j+1})$ ein Polynom vom Grad l für $j = 1, \dots, m - 1$.
2. $f(x)$ ist $(l - 1)$ -mal stetig differenzierbar.

Durch die Restriktion auf Polynom-Splines ist die Funktion f nun nicht mehr völlig beliebig. In der Tat werden durch die Glattheitsbedingungen der Polynom-Splines dem zweiten der beiden Ziele für die Schätzung von f Rechnung getragen. Darüberhinaus gilt nun der folgende, ganz zentrale Sachverhalt: Die Polynom-Splines bilden einen Vektorraum der Ordnung $d = m + l - 1$. Jedes Element aus dem Vektorraum der Polynom-Splines lässt sich daher als Linearkombination von d Basisfunktionen darstellen (vgl. Hämmerlin & Hoffmann (1994)):

$$f(x_i) = \sum_{j=1}^d \gamma_j B_j(x_i). \quad (3.5)$$

Die Basisfunktionen B_1, \dots, B_d geben allgemein Aufschluss darüber, welche Bereiche zur Schätzung von $f(x_i)$ ausschlaggebend sind und in welchem Maße. Es existieren mehrere Typen von Basisfunktionen, die zur Modellierung herangezogen werden können. Die Anpassung der Daten wird dann durch die Schätzung der Parameter $\gamma_1, \dots, \gamma_d$ erreicht. Damit liegt wieder ein parametrisches Modell vor, das in völliger Analogie zum linearen Modell geschätzt werden kann. In diesem Sinne ist die Bezeichnung „nonparametrisch“ irreführend und eigentlich nicht korrekt; sie ist aber dennoch für diesen Sachverhalt gebräuchlich.

Im konkreten Anwendungsfall sind nun die Anzahl der Knoten, ihre Platzierung und die Wahl des Basisfunktionentyps festzulegen. Während Letzteres für die Schätzergebnisse weniger entscheidend ist, spielt die Anzahl der Knotenpunkte eine große Rolle. Sie darf für eine angemessene Modellierung der Daten nicht zu gering sein, wohingegen bei einer zu hohen Anzahl die Gefahr eines Overfittings besteht. Auch die Platzierung der Knoten ist nicht unerheblich. Hier sind in der Praxis drei Verfahren üblich: äquidistante, quantilbasierte oder anhand eines Streudiagramms subjektiv festgelegte Knoten. Äquidistante Knoten sind am leichtesten zu handhaben, so dass sie Gegenstand der weiteren Ausführungen sein werden. Auch bei der Wahl der Basisfunktionen stehen mehrere Alternativen zur Auswahl. Die bekanntesten Basen sind die Basis der trunkierten Potenzen (Truncated Power Series Basis, TP-Basis) und die B-Spline-Basis (Basic Spline Basis). Die TP-Basis greift die Idee eines globalen Polynoms wieder auf und ergänzt sie um lokale Komponenten, um auch lokal gute Schätzungen zu gewährleisten. Grundlage ist folgendes für $i = 1, \dots, n$ formuliertes Regressionsmodell:

$$y_i = \gamma_1 + \gamma_2 x_i + \dots + \gamma_{l+1} x_i^l + \gamma_{l+2} (x_i - \kappa_2)_+^l + \dots + \gamma_{l+m-1} (x_i - \kappa_{m-1})_+^l + \varepsilon_i \quad (3.6)$$

mit

$$(x_i - \kappa_j)_+^l = \begin{cases} (x_i - \kappa_j)^l & x_i \geq \kappa_j, \\ 0 & \text{sonst.} \end{cases}$$

Der zum höchsten Grad des Polynoms gehörige Term wird also mehrmals aufgeführt – allerdings in einer jedes Mal unterschiedlich trunkierten Form. Auf diese Weise werden in der Tat in jedem Intervall $[\kappa_j, \kappa_{j+1})$ lokale Polynome geschätzt, die „glatt“ zusammengesetzt werden. Das Modell (3.6) lässt sich in die Form von (3.5) bringen, wenn die einzelnen Basisfunktionen folgende Gestalt haben:

Truncated Power Series Basis

Die d Basisfunktionen zum Grad l lauten:

$$B_1(x) = 1, B_2(x) = x, \dots, B_{l+1}(x) = x^l, \\ B_{l+2}(x) = (x - \kappa_2)_+^l, \dots, B_d(x) = (x - \kappa_{m-1})_+^l.$$

Man erkennt an der TP-Basis, dass mit Ausnahme von $B_1(x)$ sämtliche Basisfunktionen nach oben nicht beschränkt sind. Dies bringt bei großen x -Werten numerische Probleme mit sich. Eine in dieser Hinsicht vorzuziehende Basis liefern die B-Splines, die außerdem die Grundlage der in Abschnitt 3.2.2 ausgeführten P-Splines darstellen. Die Basisfunktionen der B-Splines sind nach oben beschränkt und sind daher numerisch stabiler. Sie unterliegen folgender rekursiven Definition:

B-Spline-Basis

Die j -te Basisfunktion zum Grad l lautet $\forall j = 1, \dots, d$:

$$B_j^0(x) = I_{[\kappa_j, \kappa_{j+1})}(x) = \begin{cases} 1 & \kappa_j \leq x < \kappa_{j+1}, \\ 0 & \text{sonst,} \end{cases} \\ B_j^l(x) = \frac{x - \kappa_j}{\kappa_{j+l} - \kappa_j} B_j^{l-1}(x) + \frac{\kappa_{j+l+1} - x}{\kappa_{j+l+1} - \kappa_{j+1}} B_{j+1}^{l-1}(x). \quad (3.7)$$

Abbildung 3.1 zeigt die Gestalt einzelner B-Spline-Basisfunktionen für verschiedene Grade bei äquidistanten Knoten. Dort lässt sich auch erkennen, dass für jede Basisfunktion $l + 2$ Knoten benötigt werden. Um die rekursive Definition in (3.7) zu ermöglichen, ist es deshalb

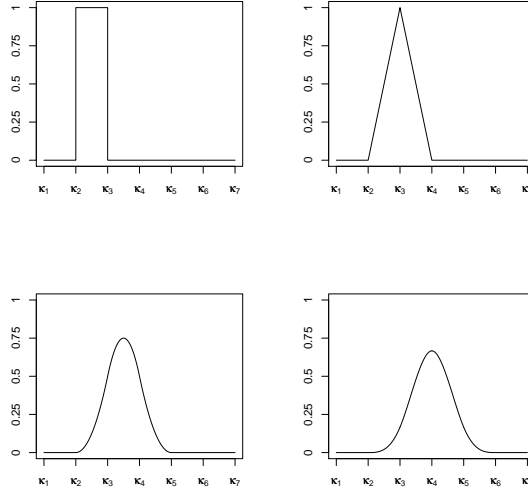


Abbildung 3.1: B-Spline-Basisfunktionen vom Grad $l = 0, 1, 2, 3$ bei äquidistanten Knoten

erforderlich, links und rechts des Definitionsbereichs $[a; b]$ jeweils l Knoten zu ergänzen. Abbildung 3.2 veranschaulicht dies. Für die Eigenschaften der B-Splines sei auf Fahrmeir, Kneib & Lang (2007) verwiesen.

Wie technisch die Bestimmung der TP-Basis, der B-Spline-Basis oder die Polynom-Splines an sich auf den ersten Blick auch erscheinen mögen, letztendlich liegt ein Modell vor, das sich formal von dem linearen Modell nicht unterscheidet:

$$\mathbf{y} = \mathbf{B}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

Dabei entspricht $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)'$ dem Koeffizientenvektor und \mathbf{B} stellt die auf folgende Weise bestimmte Designmatrix dar:

$$\mathbf{B} := \begin{pmatrix} B_1^l(x_1) & \dots & B_d^l(x_1) \\ \vdots & & \vdots \\ B_1^l(x_n) & \dots & B_d^l(x_n) \end{pmatrix}.$$

Auch hier wird für die Fehlervariablen die Annahme (3.3) getroffen. Somit gilt:

$$\mathbf{y}|\boldsymbol{\gamma}, \sigma^2 \sim N(\mathbf{B}\boldsymbol{\gamma}, \sigma^2 \mathbf{I}_n).$$

Zur bayesianischen Schätzung dieses Modells kann nun völlig analog zum bayesianischen linearen Modell in Abschnitt 3.1 vorgegangen werden.

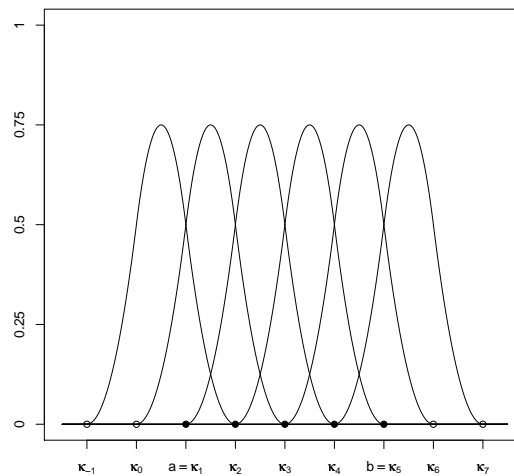


Abbildung 3.2: Knotenerweiterung für B-Spline-Basisfunktionen bei $m = 5$, $l = 2$ und äquidistanten Knoten

3.2.2 Penalisierungsansätze

Die Penalisierungsansätze beruhen direkt auf den am Beginn von Abschnitt 3.2 geäußerten Zielen: Passe auf der einen Seite die Schätzung \hat{f} möglichst gut an die Daten an und bestrafe auf der anderen Seite die Rauheit von \hat{f} . Meist wird dieses Konzept über das Prinzip der kleinsten Quadrate realisiert. Dabei wird das folgende penalisierte KQ-Kriterium minimiert:

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda J(f).$$

Dabei stellt $J(f)$ einen Strafterm dar, der die Rauheit der Funktion bestraft, während $\lambda \geq 0$ einem reellwertigen Parameter entspricht, der die gewünschte Glattheit steuert. Je höher der Glättungsparameter λ , desto mehr findet der Strafterm in der Schätzung Berücksichtigung und desto glatter wird \hat{f} . Da die zweite Ableitung einer Funktion allgemein ihre Krümmung beschreibt, wird häufig für den Strafterm $J(f) = \int (f''(u))^2 du$ verwendet. In diesem Fall spricht man von Glättungssplines. Daneben gibt es noch ein anderes, oft verwendetes Verfahren, nämlich das der P-Splines. Sie basieren auf dem Konzept der B-Splines, d.h. die Funktion f bzw. der Koeffizientenvektor γ wird gemäß Abschnitt 3.2.1 geschätzt. Darüberhinaus wird γ aber zusätzlich penalisiert. In diesem Sinne stellen die P-Splines eine Kombination von B-Splines und Penalisierungsverfahren dar. Die Idee bei der Bestrafung ist folgende: Aufeinander folgende B-Spline-Basisfunktionen B_{j-1} und B_j besitzen die Eigenschaft, dass sie sich in ihrer Wirkung auf die Schätzung ausgleichen, falls die dazugehörigen Koeffizienten γ_{j-1} und γ_j gleich groß sind. Die Penalisierung des Abstandes zweier benachbarter Koeffizienten führt daher zu einer Glättung der Funktion.

Zur Konstruktion des Strafterms dienen daher die Differenzen $\gamma_j - \gamma_{j-1}$ bzw. Differenzen höherer Ordnung $\Delta^k \gamma_j$. Diese sind auf folgende, rekursive Weise definiert:

$$\begin{aligned}\Delta^1 \gamma_j &= \gamma_j - \gamma_{j-1}, \\ \Delta^2 \gamma_j &= \Delta^1 \gamma_j - \Delta^1 \gamma_{j-1} = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2}, \\ &\vdots \\ \Delta^k \gamma_j &= \Delta^{k-1} \gamma_j - \Delta^{k-1} \gamma_{j-1}.\end{aligned}$$

Der Strafterm ist nun durch $J(f) = \sum_{j=k+1}^d (\Delta^k \gamma_j)^2 = \boldsymbol{\gamma}' \mathbf{K}_k \boldsymbol{\gamma}$ mit $\mathbf{K}_k = \mathbf{D}_k' \mathbf{D}_k$ bestimmt. Dabei gilt z.B.:

$$\mathbf{D}_1 := \begin{pmatrix} -1 & 1 & & 0 \\ & -1 & 1 & \\ & & \ddots & \ddots \\ 0 & & & -1 & 1 \end{pmatrix}, \quad \mathbf{D}_2 := \begin{pmatrix} 1 & -2 & 1 & & 0 \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ 0 & & & 1 & -2 & 1 \end{pmatrix}.$$

Mit der durch (3.5) bestimmten Schätzung von f lautet das penalisierte KQ-Kriterium nun:

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^d \gamma_j B_j(x_i) \right)^2 + \lambda \boldsymbol{\gamma}' \mathbf{K}_k \boldsymbol{\gamma}.$$

Das bisher geschilderte Schätzverfahren nach dem Prinzip der kleinsten Quadrate ist nicht der Bayes-Inferenz sondern der klassischen Inferenz zuzuordnen. Die P-Splines können aber auch bayesianisch umgesetzt werden. Hierfür stellt obige klassische Variante eine gute Hinführung dar. Die zentrale Frage bei einer solchen bayesianischen Interpretation ist, wie die Penalisierung für $\boldsymbol{\gamma}$ Berücksichtigung findet. Es gilt nun, die durch die Differenzen bestimmte Bestrafung in die Priori-Annahme einfließen zu lassen. Hierzu dienen sog. Irrfahrten (Random Walks) k -ter Ordnung, die das stochastische Analogon zu Differenzen k -ter Ordnung darstellen. Diese Analogie soll nun am Beispiel einer Irrfahrt der Ordnung 1 veranschaulicht werden, ehe sie auf beliebige Ordnungen verallgemeinert wird. Eine sog. (Gauß)-Irrfahrt der Ordnung 1 ist definiert als:

$$\gamma_j = \gamma_{j-1} + u_j, \quad u_j \stackrel{i.i.d.}{\sim} N(0, \tau^2) \quad \forall j = 2, \dots, d.$$

Man erkennt hier, dass $\gamma_j - \gamma_{j-1}$ eine um 0 symmetrische Verteilung besitzt. Dies kommt der Forderung gleich, dass γ_j und γ_{j-1} nicht in zu großem Maße voneinander abweichen sollen, so wie sie in der klassischen Form der P-Splines getroffen wurde. Für die bedingte Verteilung $\gamma_j | \gamma_{j-1}, \dots, \gamma_1$ gilt nun bei bekannter Varianz τ^2 :

$$\gamma_j | \gamma_{j-1}, \dots, \gamma_1 = \gamma_j | \gamma_{j-1} \sim N(\gamma_{j-1}, \tau^2) \quad \forall j = 2, \dots, d \quad (3.8)$$

Die durch das Gleichheitszeichen in (3.8) beschriebene Markov-Eigenschaft ist eine charakteristische Eigenschaft der Irrfahrt, die in die Bestimmung einer gemeinsamen Priori-Verteilung für γ eingeht. Hierfür muss noch eine Verteilungsannahme für den Startwert γ_1 getroffen werden. In der Regel wird dafür eine nichtinformative Priori-Verteilung verwendet:

$$p(\gamma_1) \propto \text{const.}$$

Somit gilt für die gemeinsame Verteilung von γ :

$$\begin{aligned} p(\gamma) &= p(\gamma_1) \prod_{j=2}^d p(\gamma_j | \gamma_{j-1}) = \prod_{j=2}^d \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(\gamma_j - \gamma_{j-1})^2\right) = \\ &= \frac{1}{(2\pi\tau^2)^{(d-1)/2}} \exp\left(-\frac{1}{2\tau^2} \sum_{j=2}^d (\gamma_j - \gamma_{j-1})^2\right) = \frac{1}{(2\pi\tau^2)^{(d-1)/2}} \exp\left(-\frac{1}{2\tau^2} \gamma' \mathbf{K}_1 \gamma\right). \end{aligned}$$

Fasst man die Varianz τ^2 als zu schätzenden Parameter und damit als Zufallsvariable auf, so gilt entsprechend bei 1.Ordnung:

$$p(\gamma | \tau^2) = \frac{1}{(2\pi\tau^2)^{(d-1)/2}} \exp\left(-\frac{1}{2\tau^2} \gamma' \mathbf{K}_1 \gamma\right).$$

Berechnungen für beliebige Ordnungen k lassen sich ganz analog durchführen:

$$p(\gamma | \tau^2) = \frac{1}{(2\pi\tau^2)^{(d-k)/2}} \exp\left(-\frac{1}{2\tau^2} \gamma' \mathbf{K}_k \gamma\right).$$

Obwohl die Form dieser Dichte auf den ersten Blick stark an eine multivariate Normalverteilung erinnert, so ist dies nur mit Abstrichen richtig. Für die Matrix \mathbf{K}_k gilt nämlich: $\text{rg}(\mathbf{K}_k) = d - k < d$, d.h. \mathbf{K}_k ist singulär. Somit existiert keine Inverse \mathbf{K}_k^{-1} , wie es zur Bestimmung der Kovarianzmatrix einer multivariaten Normalverteilung notwendig wäre. Man spricht hier von einer singulären Normalverteilung. Diese ist eine teilweise uneigentliche, d.h. nicht normierbare Verteilung.

Der Parameter τ^2 steuert, wie stark sich Koeffizienten γ_j und γ_{j-1} unterscheiden. Je kleiner τ^2 , desto geringer ist jene Abweichung und desto glatter ist die geschätzte Funktion \hat{f} . In diesem Sinne stellt τ^2 einen inversen Glättungsparameter dar. Es gilt sogar: $\lambda = \sigma^2/\tau^2$. In seiner Rolle als Varianzparameter bietet sich als a priori Verteilung für τ^2 ebenso eine inverse Gammaverteilung an wie für σ^2 . Mit entsprechenden Unabhängigkeitsannahmen hinsichtlich der gemeinsamen Priori kann nun die Posteriori-Verteilung bestimmt werden:

$$p(\gamma, \tau^2, \sigma^2 | \mathbf{y}) \propto p(\mathbf{y} | \gamma, \sigma^2) p(\gamma | \tau^2) p(\tau^2) p(\sigma^2).$$

Hierbei handelt es sich wie schon beim bayesianischen linearen Modell in Abschnitt 3.1 um die Normal-inverse Gammaverteilung. Ein solches Modell mit bayesianischen P-Splines

wird Grundlage der Analysen in Kapitel 8 sein. Die Annahmen seien nochmal zusammengefasst:

Bayesianische P-Splines

1. *Beobachtungsmodell:*

$$\mathbf{y}|\boldsymbol{\gamma}, \sigma^2 \sim N(\mathbf{B}\boldsymbol{\gamma}, \sigma^2 \mathbf{I}_n).$$

2. *Priori-Verteilungen:*

$$\begin{aligned} p(\gamma_j) &\propto \text{const} & \forall j = 1, \dots, k, \\ \gamma_j|\gamma_{j-1}, \tau^2 &\sim N(\gamma_{j-1}, \tau^2) & \forall j = k+1, \dots, d, \\ \tau^2 &\sim IG(a_\gamma, b_\gamma), \\ \sigma^2 &\sim IG(a_\varepsilon, b_\varepsilon). \end{aligned}$$

3.3 Das lineare gemischte Modell

Das lineare gemischte Modell erweitert das lineare Modell (3.2) dahingehend, dass neben dem generellen, den alle Individuen betreffenden Einfluss von Kovariablen x_1, \dots, x_p auf die Zielgröße y auch individuenspezifische Effekte betrachtet werden. Ein solches gemischtes Modell findet vor allem in der Analyse longitudinaler Daten Anwendung. Bei longitudinalen Daten werden mehrere Individuen innerhalb einer bestimmten Zeitspanne untersucht. Für jedes Individuum i mit $i = 1, \dots, n$ liegen n_i Messungen vor, die zu bestimmten Zeitpunkten t_{ij} mit $j = 1, \dots, n_i$ erhoben wurden. Die Kovariablen, für die individuenspezifische Effekte untersucht werden sollen, werden im Folgenden mit z_1, \dots, z_q bezeichnet. Dabei handelt es sich meist um eine Teilmenge von x_1, \dots, x_p , d.h. bei manchen Einflussgrößen wird der generelle und der individuelle Effekt analysiert und bei manchen nur der generelle. Die Datenstruktur ist also durch $(y_{i1}, \dots, y_{in_i}, \mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}, \mathbf{z}_{i1}, \dots, \mathbf{z}_{in_i})$ für $i = 1, \dots, n$ gegeben. Das lineare gemischte Modell setzt nun die einzelnen als linear angenommenen Effekte additiv zusammen:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i + \varepsilon_{ij}.$$

Die Vektoren $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})'$ und $\mathbf{z}_{ij} = (z_{ij1}, \dots, z_{ijq})'$ stehen dabei für die Kovariablenwerte der Person i zum Zeitpunkt j . Sie variieren i.A. mit der Zeit, werden aber oft als zeitkonstant angenommen. Der Vektor $\mathbf{b}_i = (b_{i1}, \dots, b_{iq})'$ symbolisiert die individuellen Koeffizienten für das i -te Individuum. Für die zufällige Komponente ε_{ij} wird nun dem linearen Modell entsprechend (vgl. (3.3)) folgende Verteilungsannahme getroffen:

$$\varepsilon_{ij} \stackrel{i.i.d.}{\sim} N(0, \sigma^2) \quad \forall i = 1, \dots, n; \forall j = 1, \dots, n_i. \quad (3.9)$$

Mit den individuellen Designmatrizen $\mathbf{X}_i = (\mathbf{x}'_{i1}, \dots, \mathbf{x}'_{in_i})'$ und $\mathbf{Z}_i = (\mathbf{z}'_{i1}, \dots, \mathbf{z}'_{in_i})'$ bzw. dem Responsevektor $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$ und dem Fehlervariablenvektor $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})'$ lässt sich das lineare gemischte Modell auch in Matrixnotation schreiben:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i.$$

Die Annahme (3.9) lässt sich ebenfalls in eine kompakte vektorielle Form bringen:

$$\boldsymbol{\varepsilon}_i \stackrel{ind}{\sim} N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i}) \quad \forall i = 1, \dots, n. \quad (3.10)$$

Gelegentlich wird hierbei auch die allgemeinere Annahme $\boldsymbol{\varepsilon}_i \stackrel{ind}{\sim} N(\mathbf{0}, \Sigma_{\varepsilon,i})$ verwendet. Im Folgenden soll aber stets die Annahme (3.10) getroffen werden. Sie bedeutet, dass die intrapersonelle Korrelation gegeben \mathbf{b}_i gleich 0 ist (vgl. (3.12)). Damit wird der Tatsache, dass gewisse Datenpunkte zu demselben Individuum gehören, nur über einen gemeinsamen Vektor \mathbf{b}_i Rechnung getragen.

Der den generellen Effekt beschreibende Vektor $\boldsymbol{\beta}$ wird in diesem Zusammenhang als fester Effekt bezeichnet, während \mathbf{b}_i zufälliger Effekt genannt wird. Dies hat seinen Ursprung in einer Sichtweise der klassischen Inferenz. Dort werden zu schätzende Parameter wie z.B. $\boldsymbol{\beta}$ grundsätzlich als feste, aber unbekannte Parameter verstanden. In gemischten Modellen wird \mathbf{b}_i nun nicht als Parameter, sondern als Zufallsgröße angesehen, für die eine für alle Individuen identische Verteilung angenommen wird. Dies ist sinnvoll, da andernfalls sehr viele, mitunter für die Gewährleistung der Schätzbarkeit des Modells zu viele Parameter zu schätzen wären. Typischerweise wird hierfür eine Normalverteilung herangezogen:

$$\mathbf{b}_i \stackrel{i.i.d.}{\sim} N(\mathbf{0}, \Sigma_b) \quad \forall i = 1, \dots, n. \quad (3.11)$$

Der Erwartungswert 0 rührt von der Vorstellung her, dass die zufälligen Effekte die individuellen Abweichungen von dem generellen Effekt darstellen. Hierzu müssen freilich für alle zufälligen Effekte auch die dazugehörigen festen Effekte im Modell enthalten sein. Die Normalverteilungsannahme wird in der klassischen Inferenz oft getroffen, weil damit – zusammen mit der Annahme (3.10) und der Unabhängigkeitsannahme von \mathbf{b}_i und $\boldsymbol{\varepsilon}_i$ – auch eine Normalverteilung für den marginalen, individuellen Response folgt:

$$\mathbf{y}_i \sim N(\mathbf{X}_i \boldsymbol{\beta}, \sigma^2 \mathbf{I}_{n_i}) \quad \forall i = 1, \dots, n.$$

In der Bayes-Inferenz, bei der sowohl die festen als auch die zufälligen Effekte als Zufallsgrößen aufgefasst werden, ist anstatt der marginalen Betrachtungsweise stets die bedingte Verteilung von \mathbf{y}_i gegeben alle die Verteilung bestimmenden Parameter von Interesse. Offensichtlich ist hierfür der Verteilungstyp der Prioris für $\boldsymbol{\beta}$, \mathbf{b}_i und σ^2 unerheblich. Für die bedingte Verteilung gilt damit stets aus der Annahme (3.10) folgend die Normalverteilung:

$$\mathbf{y}_i | \boldsymbol{\beta}, \mathbf{b}_i, \sigma^2 \sim N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i, \sigma^2 \mathbf{I}_{n_i}) \quad \forall i = 1, \dots, n. \quad (3.12)$$

Man ist daher in der Bayes-Inferenz grundsätzlich völlig frei, was die Wahl der (Priori)-Verteilung für die zufälligen Effekte betrifft, die im Folgenden mit F bezeichnet wird, und kann sich daher von der in der klassischen Inferenz üblichen Normalverteilungsannahme lösen. Ihre Adäquatheit ist nämlich in bestimmten Datenkonstellationen kritisch zu hinterfragen. Die Symmetrie und Unimodalität, die damit angenommen wird, können in der Praxis durchaus nicht gegeben sein. Zudem impliziert die Normalverteilungsannahme schmale Enden, so dass nicht all zu sehr voneinander abweichende Effekte unterstellt werden. Dunson (2008) führt an, dass Ausreißer unter dieser Annahme das Populationsmittel über die Maßen beeinflussen, während ihre zufälligen Effekte in die Richtung des Populationsmittels verzerrt würden. Eine konkrete parametrische Verteilungsannahme für die zufälligen Effekte ist darüber hinaus grundsätzlich problematisch, da sie sich als latente Variablen einer direkten Erhebung entziehen und daher eine für sie angenommene Verteilung nur schwer auf ihre Korrektheit überprüft werden kann. Ihre Verteilungsannahme hat jedoch Einfluss auf die Schätzungen der zufälligen Effekte. Eine falsche Verteilungsannahme kann deshalb zu Verzerrungen führen (vgl. Kleinman & Ibrahim (1998)).

Eine flexiblere Modellierung ist möglich, wenn man die Unsicherheit über die Verteilung der zufälligen Effekte ins Modell einbezieht. Die Verteilung F übernimmt hierbei die Rolle eines „Parameters“. Aus bayesianischer Perspektive gilt es nun, ein a priori Wissen für F zu formulieren. Für eine flexible Modellierung dessen ist eine reichhaltige Klasse von Priori-Verteilungen erforderlich. Die sog. Dirichlet-Prozess-Prioris stellen so eine Klasse für die Menge der Wahrscheinlichkeitsmaße dar. Die Dirichlet-Prozesse werden in Kapitel 4 ausführlich behandelt, ehe in Kapitel 5 verschiedene Algorithmen für Modelle mit Dirichlet-Prozessen diskutiert werden. An dieser Stelle sollen jedoch schon vorab zwei Möglichkeiten vorgestellt werden, wie eine Dirichlet-Prozess-Priori, i.Z. $DP(\alpha)$, mit der Verteilung der zufälligen Effekte verknüpft werden kann. Im ersten Fall wird, genauso wie oben ausgeführt, eine Dirichlet-Prozess-Priori für F angenommen:

Dirichlet-Prozess-Modell:

$$\begin{aligned} \mathbf{b}_i | F &\stackrel{i.i.d.}{\sim} F & \forall i = 1, \dots, n, \\ F &\sim DP(\alpha). \end{aligned}$$

Im zweiten Fall wird die Verteilung der zufälligen Effekte F in zwei Stufen modelliert. Die erste befasst sich mit der bedingten Verteilung der \mathbf{b}_i gegeben die Parameter $\boldsymbol{\theta}_i$, die im Folgenden mit $F(\boldsymbol{\theta}_i)$ bezeichnet wird. Der Verteilungstyp von $F(\boldsymbol{\theta}_i)$ sei bekannt. In der zweiten Stufe werde für die Verteilung der Parameter $\boldsymbol{\theta}_i$ – sie soll G heißen – eine Dirichlet-Prozess-Priori angenommen. Auf diese Weise resultiert für die marginale Verteilung der zufälligen Effekte F eine Mischverteilung. Das Modell wird daher Dirichlet-Prozess-Mischungs-Modell genannt.

Dirichlet-Prozess-Mischungs-Modell:

$$\begin{aligned}
\mathbf{b}_i | \boldsymbol{\theta}_i &\stackrel{\text{ind}}{\sim} F(\boldsymbol{\theta}_i) & \forall i = 1, \dots, n, \\
\boldsymbol{\theta}_i | G &\stackrel{\text{i.i.d.}}{\sim} G & \forall i = 1, \dots, n, \\
G &\sim DP(\alpha).
\end{aligned}$$

Die Beliebigkeit der Verteilung F stellt es dem Anwender nun frei, ob er zu den zufälligen Effekten auch die dazugehörigen festen Effekte im Modell mit aufführen will oder nicht. Schließlich ist der Erwartungswert von F nicht mehr auf 0 festgelegt wie in (3.11). Das DPM-Modell lässt sich nun für den Fall, dass nur zufällige Effekte betrachtet werden, wie folgt kompakt darstellen:

Lineares gemischtes Modell mit DPM-Priori

1. Beobachtungsmodell:

$$\mathbf{y}_i | \mathbf{b}_i, \sigma^2 \stackrel{\text{ind}}{\sim} N(\mathbf{Z}_i \mathbf{b}_i, \sigma^2 \mathbf{I}_{n_i}) \quad \forall i = 1, \dots, n.$$

2. Priori-Verteilungen:

$$\begin{aligned}
\mathbf{b}_i | \boldsymbol{\theta}_i &\stackrel{\text{ind}}{\sim} F(\boldsymbol{\theta}_i) & \forall i = 1, \dots, n, \\
\boldsymbol{\theta}_i | G &\stackrel{\text{i.i.d.}}{\sim} G & \forall i = 1, \dots, n, \\
G &\sim DP(\alpha), \\
\sigma^2 &\sim IG(a_\varepsilon, b_\varepsilon).
\end{aligned}$$

An dieser Stelle sei bereits erwähnt, dass die Posteriori-Verteilung für alle im Modell enthaltenen Parameter nicht in geschlossener Form darstellbar ist. Die Tatsache, dass der unbekannte Parameter G eine Verteilung und kein reellwertiger Parameter ist, wirft bereits bei der Formulierung der Priori-Verteilung viele Fragen auf, die tiefgründige Schwierigkeiten erkennen lassen: Wie kann man die Dichte $p(\boldsymbol{\theta}_i | G)$ formulieren, wenn die zugrunde liegende Verteilung G nicht feststeht? Wie lässt sich überhaupt eine Priori-Dichte für eine Verteilung angeben? Kapitel 4 liefert hierfür grundlegende Einsichten. Nichtsdestotrotz lässt sich angesichts dieser Problematik ein wenig überraschend festhalten: Durch MCMC-Verfahren können DP- bzw. DPM-Modelle geschätzt werden (vgl. Kapitel 5).

4 Dirichlet-Prozesse

Das folgende Kapitel widmet sich der Theorie der Dirichlet-Prozesse. Diese gehen zurück auf Ferguson (1973) und kommen in der Bayes-Inferenz dann zum Einsatz, wenn eine Priori-Annahme für eine unbekannte Verteilung gesucht wird. So werden sie, wie in Abschnitt 3.3 vorgestellt, z.B. bei gemischten Modellen verwendet, wenn man sich hinsichtlich der Verteilung der zufälligen Effekte nicht festlegen will. Sie kommen aber auch bei der nonparametrischen Dichteschätzung zur Anwendung. Da die Inferenz in gemischten Modellen durch die longitudinale Struktur und zahlreiche andere Parameter eine komplexere Struktur als die Dichteschätzung aufweist, soll die Idee der Dirichlet-Prozesse im Folgenden am Beispiel der Dichteschätzung in ihren Grundzügen dargelegt werden, ehe sie in Kapitel 6 auf gemischte Modelle übertragen wird.

Die Ausgangslage bei der Dichteschätzung sei nun wie folgt gegeben: Die erhobenen Daten y_1, \dots, y_n folgen einer unbekannten Verteilung F , die eine Dichte f bzgl. eines die Verteilung dominierenden Maßes besitzt:

$$y_i|F \stackrel{i.i.d.}{\sim} F \quad \forall i = 1, \dots, n. \quad (4.1)$$

Ziel ist nun die Schätzung der Dichte f , die über ihre Verteilung F identifiziert wird. Wählt man nun für F eine Dirichlet-Prozess-Priori, $F \sim DP(\alpha)$, so handelt es sich bei F – dies sei an dieser Stelle bereits vorweggenommen – um eine diskrete Verteilung. Dieser Fakt ist in Zusammenhang mit der Dichteschätzung hinderlich. Ein anderer Weg, die Verteilung F zu schätzen, kann durch die Annahme erfolgen, dass sich F als Mischung von Verteilungen $F(\theta_i)$ interpretieren lässt. Es sind nun die o.B.d.A. als univariat angenommenen Parameter θ_i mit $i = 1, \dots, n$, die einer gänzlich unbekannten Verteilung namens G folgen:

$$y_i|\theta_i \stackrel{ind}{\sim} F(\theta_i) \quad \forall i = 1, \dots, n, \quad (4.2)$$

$$\theta_i|G \stackrel{i.i.d.}{\sim} G \quad \forall i = 1, \dots, n. \quad (4.3)$$

Der Verteilungstyp der bedingten Verteilungen $F(\theta_i)$ werde dabei auf eine bestimmte stetige Verteilung festgelegt. Damit ist auch die marginale Verteilung F stetig. Ihre Dichte ist gemäß des Mischungsansatzes wie folgt bestimmt:

$$f(y_i) = \int f(y_i|\theta_i)dG(\theta_i).$$

Durch die Schätzung der Parameter $\theta_1, \dots, \theta_n$ wird nun automatisch die Verteilung F bzw. die Dichte f geschätzt. Die folgenden theoretischen Überlegungen konzentrieren sich

daher auf die unbekannten Parameter $\theta_1, \dots, \theta_n$ und die unbekannte Verteilung G . Kapitel 5 bietet dann einen Überblick, welche Inferenzmethoden hierfür zur Verfügung stehen.

Der Parameterraum der $\theta_1, \dots, \theta_n$ werde im Folgenden mit Θ bezeichnet. G entspricht nun einem Wahrscheinlichkeitsmaß auf dem Messraum (Θ, \mathcal{A}) , wobei \mathcal{A} eine zu Θ geeignete σ -Algebra darstellt. In seiner Rolle als unbekannter „Parameter“ ist G aus bayesianischer Perspektive selbst wiederum eine Zufallsgröße – ein zufälliges Wahrscheinlichkeitsmaß – und kann Werte innerhalb von \mathfrak{G} , der Menge aller Wahrscheinlichkeitsmaße auf (Θ, \mathcal{A}) , annehmen. Zusammen mit der σ -Algebra \mathcal{C} bildet $(\mathfrak{G}, \mathcal{C})$ einen Messraum. \mathcal{C} sei hierbei als die von den Mengen $\{G : G(A) < r\}$ mit $A \in \Theta$ und $r \in [0, 1]$ erzeugte σ -Algebra definiert (vgl. Sethuraman (1994)). Über ein Wahrscheinlichkeitsmaß ν auf $(\mathfrak{G}, \mathcal{C})$ lässt sich nun im bayesianischen Sinne ein a priori Wissen über G formulieren. Das a posteriori Wissen nach Beobachtung der $\theta_1, \dots, \theta_n$ wird durch das Wahrscheinlichkeitsmaß ν^θ beschrieben.

In Anlehnung an Ferguson (1973) sind an die Priori-Verteilung ν zwei Forderungen zu stellen:

1. Der Träger von G sollte möglichst groß sein.
2. Bayes-Inferenz sollte entweder über eine analytisch zugängliche Posteriori-Verteilung oder über MCMC-Verfahren durchführbar sein.

Die Dirichlet-Prozess-Prioris erfüllen beide Forderungen.

4.1 Definition des Dirichlet-Prozesses

Die Dirichlet-Prozesse basieren auf der Dirichlet-Verteilung. Daher soll zunächst diese definiert werden und ihre Eigenschaften, die in Zusammenhang mit den Dirichlet-Prozessen wichtig sind, hervorgehoben werden. Die Dirichlet-Verteilung trifft eine Verteilungsaussage für einen Vektor von Wahrscheinlichkeiten $\boldsymbol{\pi} = (\pi_1, \dots, \pi_m)'$ aus dem $(m - 1)$ -Simplex $\mathbb{S} = \{\boldsymbol{\pi} : 0 \leq \pi_j \leq 1 \ \forall j = 1, \dots, m; \sum_{j=1}^m \pi_j = 1\}$:

Definition: Dirichlet-Verteilung

Ein Zufallsvektor $\boldsymbol{\pi}$ ist genau dann Dirichlet-verteilt mit dem Parametervektor $(\alpha_1, \dots, \alpha_m)'$ für $\alpha_j > 0$, $\boldsymbol{\pi} \sim \text{Dir}(\alpha_1, \dots, \alpha_m)$, wenn die Verteilung folgende Dichte bzgl. des $(m - 1)$ -dimensionalen Lebesgue-Maßes besitzt:

$$f(\boldsymbol{\pi}) = \frac{\Gamma(\sum_{j=1}^m \alpha_j)}{\prod_{j=1}^m \Gamma(\alpha_j)} \prod_{j=1}^m \pi_j^{\alpha_j-1} I_{\mathbb{S}}(\boldsymbol{\pi}).$$

Dabei gilt $\pi_m = 1 - \sum_{j=1}^{m-1} \pi_j$.

Die Dirichlet-Verteilung besitzt für die Dirichlet-Prozesse mitunter sehr wichtige Eigenschaften, die im Folgenden ohne Nachweis aufgeführt werden:

1. Für den Spezialfall $m = 2$ gilt für $\pi := \pi_1$:

$$\pi \sim Be(\alpha_1, \alpha_2).$$

2. Aggregationseigenschaft:

$$(\pi_1, \dots, \pi_j + \pi_{j+1}, \dots, \pi_m)' \sim Dir(\alpha_1, \dots, \alpha_j + \alpha_{j+1}, \dots, \alpha_m).$$

3. Erwartungswert:

$$E(\pi_j) = \frac{\alpha_j}{\sum_{l=1}^m \alpha_l} \quad \forall j = 1, \dots, m.$$

4. Die Familie der Multinomialverteilungen ist zur Familie der Dirichlet-Verteilungen konjugiert.

Die ersten beiden Eigenschaften liefern eine Möglichkeit, wie man Realisationen für π aus der Dirichlet-Verteilung simulieren kann: Durch die Aggregationseigenschaft kann eine m -dimensionale Dirichlet-Verteilung stets zu einer 2-dimensionalen zusammengefasst und damit auf eine Betaverteilung gebracht werden, aus der leicht Zufallszahlen gezogen werden können. So können die Wahrscheinlichkeiten π_1, \dots, π_m in folgender Weise sukzessive simuliert werden:

Simulation der Dirichlet-Verteilung über Betaverteilungen:

1. Ziehe π_1 gemäß: $\pi_1 \sim Be(\alpha_1, \sum_{l=2}^m \alpha_l)$.
2. Für $j = 2, \dots, m - 1$:
 - (I) Ziehe V_j gemäß: $V_j \sim Be(\alpha_j, \sum_{l=j+1}^m \alpha_l)$.
 - (II) Setze $\pi_j = (1 - \sum_{l=1}^{j-1} \pi_l) V_j$.
3. Setze $\pi_m = 1 - \sum_{l=1}^{m-1} \pi_l$.

Diese Simulationsidee für Dirichlet-Verteilungen wird in Abschnitt 4.3 auf Dirichlet-Prozesse übertragen werden. Abgesehen davon existiert noch ein weiteres leicht zugängliches Simulationsprinzip basierend auf Gammaverteilungen:

Simulation der Dirichlet-Verteilung über Gammaverteilungen:

1. Ziehe Z_j gemäß: $Z_j \stackrel{ind}{\sim} Ga(\alpha_j, 1)$ für $\forall j = 1, \dots, m$.
2. Setze $\pi_j = \frac{Z_j}{\sum_{l=1}^m Z_l}$ für $\forall j = 1, \dots, m$.

Die folgenden Betrachtungen rücken die Frage nach der Priori-Verteilung für die unbekannte Verteilung G wieder in den Vordergrund. Wäre der Wertebereich Θ eine endliche Menge $\Theta = \{\theta_1, \dots, \theta_m\}$, so ließe sich die Dirichlet-Verteilung als Priori-Verteilung für die unbekannte Verteilung G verwenden: Über die Parameter α_j für $j = 1, \dots, m$ kann man sein persönliches Vorwissen über die Größe der einzelnen Wahrscheinlichkeiten $\pi_j = P(\theta_j)$ ausdrücken. Hält man beispielsweise tendenziell größere Werten von Θ für wahrscheinlicher als kleinere Werte, so bietet es an, deren Parameter α_j in der Dirichlet-Verteilung entsprechend höher zu wählen. Abbildung 4.1 zeigt z.B. eine Realisation einer Dirichlet-Verteilung $Dir(1, 2, 3, 4, 5, 6, 7, 8)$.

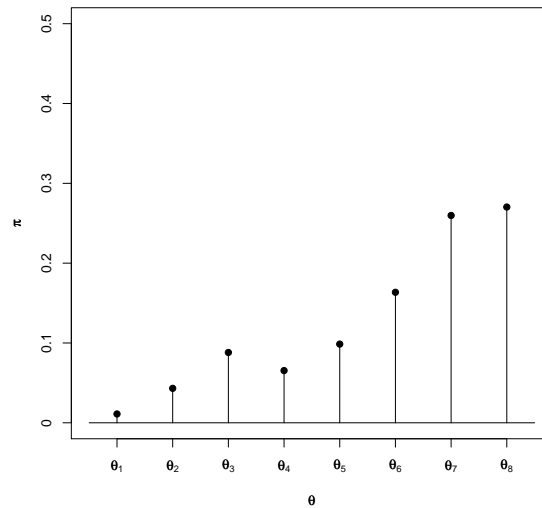


Abbildung 4.1: Realisation der Dirichlet-Verteilung $Dir(1, 2, 3, 4, 5, 6, 7, 8)$

Ein Dirichlet-Prozess stellt nun auf der einen Seite eine Erweiterung der Dirichlet-Verteilung dar, indem Θ einen beliebigen messbaren Raum bilden kann. Auf der anderen Seite basiert ein Dirichlet-Prozess in dem Sinne auf der Dirichlet-Verteilung, dass der beliebige messbare Raum Θ durch die Vergrößerung in eine endliche messbare Partition $\{A_1, \dots, A_m\}$, also in messbare und disjunkte Mengen A_j mit $\bigcup_{j=1}^m A_j = \Theta$, wieder auf den endlichen Fall reduziert wird.

Die Definition des Dirichlet-Prozesses, die auf Ferguson (1973) zurückgeht, lautet nun wie folgt:

Definition: Dirichlet-Prozess

Gegeben sei ein beliebiger messbarer Raum Θ mit der zugehörigen σ -Algebra \mathcal{A} . Sei α ein finites, von 0 verschiedenes Maß auf dem Messraum (Θ, \mathcal{A}) .

Dann gilt:

G ist genau dann ein durch die Mengen A_j indizierter Dirichlet-Prozess auf (Θ, \mathcal{A}) mit Parameter α , $G \sim DP(\alpha)$, wenn für jede endliche messbare Partition $\{A_1, \dots, A_m\}$ von Θ gilt: Der Zufallsvektor $(G(A_1), \dots, G(A_m))'$ besitzt eine Dirichlet-Verteilung mit Parameter $(\alpha(A_1), \dots, \alpha(A_m))'$.

Ein Dirichlet-Prozess auf (Θ, \mathcal{A}) bildet folglich einen stochastischen Prozess mit Parameterraum \mathcal{A} und Zustandsraum $[0, 1]$.

Ein Dirichlet-Prozess wird also durch das Maß α bestimmt: Für jedes feste α gibt es einen Dirichlet-Prozess. Die Zufälligkeit dieses stochastischen Prozesses liegt in der Zufälligkeit der Dirichlet-Verteilung mit Parameter $(\alpha(A_1), \dots, \alpha(A_m))'$. Das einmalige Ziehen aus dieser Verteilung zu gegebener Partition und gegebenem α liefert eine Realisation des Dirichlet-Prozesses.

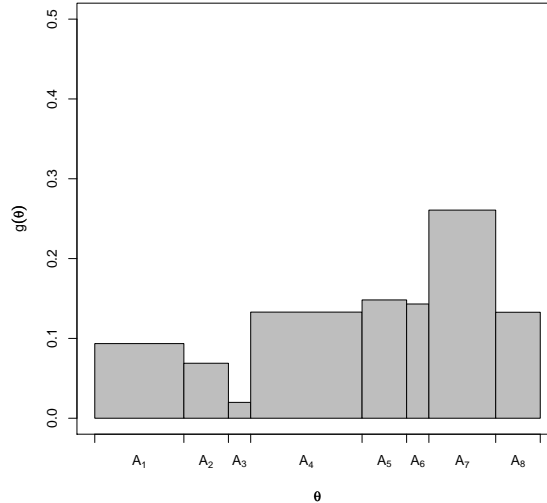


Abbildung 4.2: Realisation von $G \sim DP(\alpha)$, wobei α jeder Menge A_j den Wert 1 zuweist

Beispiel:

Der Parameterraum Θ werde in eine Partition $\{A_1, \dots, A_8\}$ unterteilt. Das Maß α sei so bestimmt, dass es jeder dieser als messbar vorausgesetzten Mengen A_j für $j = 1, \dots, 8$ unabhängig ihrer Größe den Wert 1 zuweist. Die Dirichlet-Verteilung in der Definition des Dirichlet-Prozesses lautet demnach: $Dir(1, 1, 1, 1, 1, 1, 1, 1)$. Eine Realisation von G bzw. der entsprechenden Dichte g lässt sich der Abbildung 4.2 entnehmen.

4.2 Eigenschaften des Dirichlet-Prozesses

Der folgende Abschnitt dient dem Zweck, den Dirichlet-Prozess anhand seiner wichtigsten Eigenschaften zu veranschaulichen (vgl. Sethuraman (1994)). Die Haupteigenschaft knüpft an die ursprüngliche Fragestellung an: Auf welche Weise kann ein a priori Wissen ν über eine unbekannte Verteilung G formuliert werden? Wie bereits mehrfach angedeutet kann ein Dirichlet-Prozess hierfür als Priori-Verteilung fungieren. Dies ist möglich, da folgende Grundvoraussetzung erfüllt ist:

1. Haupteigenschaft des Dirichlet-Prozesses

Der Dirichlet-Prozess $DP(\alpha)$ ist ein Wahrscheinlichkeitsmaß auf $(\mathfrak{G}, \mathcal{C})$.

Ein Nachweis dieser Aussage erfolgt in Abschnitt 4.3.

Konkret wird das Vorwissen durch das Maß α bzw. durch den bei gegebener Partition festgelegten Parametervektor $(\alpha(A_1), \dots, \alpha(A_m))'$ formuliert. Bei dem Maß α handelt es sich gemäß Definition um ein finites Maß, d.h. $\alpha(\Theta) < \infty$. Es ist also auf eine reelle Zahl $\alpha_0 := \alpha(\Theta)$ normiert. Daher lässt sich das Maß α auch stets als Produkt der Zahl α_0 und des entsprechenden, auf 1 normierten Wahrscheinlichkeitsmaßes $G_0 = \alpha/\alpha(\Theta)$ darstellen: $\alpha = \alpha_0 G_0$. Zwischen G und G_0 besteht nun folgender im Anhang A.2 nachgewiesener Zusammenhang:

$$E(G) = G_0.$$

Dieser Sachverhalt sagt aus, dass sich die Verteilung G im Mittel um G_0 konzentriert. Dies ermöglicht nun eine inhaltlich schön interpretierbare Priori-Annahme: G_0 entspricht der Vermutung, die der Anwender für G ansetzt, und α_0 einem Parameter, der ausdrückt, in welchem Maße der Anwender seiner Vermutung vertraut. Anders formuliert: α_0 steuert, wie stark sich G um G_0 konzentriert. Man bezeichnet die Parameter daher auch wie folgt:

G_0	Basisverteilung
$\alpha_0 > 0$	Präzisionsparameter

Abbildung 4.3 illustriert Realisationen von G mit $G \sim DP(\alpha_0 G_0)$ für unterschiedliche Präzisionsparameter.

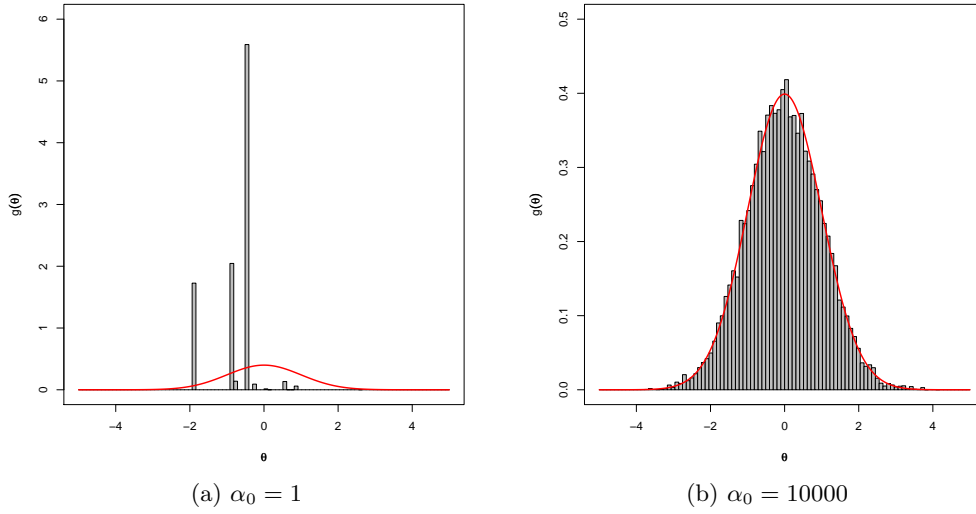


Abbildung 4.3: Realisation von $G \sim DP(\alpha_0 G_0)$ mit $G_0 = N(0; 1)$, deren Dichte in roter Farbe dargestellt ist

Mit der Datengrundlage und dem Mischungsansatz gemäß (4.2) und (4.3) ergibt sich nun, wenn man als Priori-Annahme für G einen Dirichlet-Prozess annimmt, das bereits in Abschnitt 3.3 vorgestellte Dirichlet-Prozess-Mischungs-Modell:

Dirichlet-Prozess-Mischungs-Modell (DPM-Modell):

$$\begin{aligned}
 y_i | \theta_i &\stackrel{i.i.d.}{\sim} F(\theta_i) & \forall i = 1, \dots, n, \\
 \theta_i | G &\stackrel{i.i.d.}{\sim} G & \forall i = 1, \dots, n, \\
 G &\sim DP(\alpha_0 G_0).
 \end{aligned}$$

Verzichtet man wie in (4.1) auf den Mischungsansatz und wählt für F den Dirichlet-Prozess als Priori-Verteilung, so liegt ein Dirichlet-Prozess-Modell vor:

Dirichlet-Prozess-Modell (DP-Modell):

$$\begin{aligned}
 y_i | F &\stackrel{i.i.d.}{\sim} F & \forall i = 1, \dots, n, \\
 F &\sim DP(\alpha_0 G_0).
 \end{aligned}$$

Die Dirichlet-Prozesse sind des Weiteren durch folgende charakteristische Eigenschaft geprägt:

2. Haupteigenschaft des Dirichlet-Prozesses

Für die Menge der diskreten Verteilungen $\mathfrak{G}_{\text{diskret}} \subset \mathfrak{G}$ gilt: $DP(\alpha_0 G_0)(\mathfrak{G}_{\text{diskret}}) = 1$, d.h. jede Realisation von G ist eine diskrete Verteilung.

Ein Nachweis dieser Aussage erfolgt ebenfalls in Abschnitt 4.3.

Die diskrete Natur der Verteilung G führt bezüglich $\theta_i | G \stackrel{i.i.d.}{\sim} G$ für $i = 1, \dots, n$ dazu, dass verschiedene Individuen $i' \neq i''$ dasselbe $\theta_{i'} = \theta_{i''}$ annehmen können und somit einem gemeinsamen Cluster angehören. Diese Cluster-Eigenschaft ist ein zentrales Merkmal der Dirichlet-Prozesse. Sie lässt sich wie folgt konkretisieren: Allgemein liegen bei n Individuen $k \leq n$ Cluster vor. Seien nun c_1, \dots, c_n Variablen, die die Clusterzugehörigkeit jedes Individuums angeben, und seien ϕ_1, \dots, ϕ_k die Clusterlokationen mit $\theta_i = \phi_{c_i}$. Die Information der $\theta_1, \dots, \theta_n$ ist folglich mit der Information von c_1, \dots, c_n und ϕ_1, \dots, ϕ_k äquivalent. Dieser Sachverhalt wird sich bei der Konstruktion von Algorithmen als nützlich erweisen.

Die dritte Eigenschaft stellt einen Zusammenhang zwischen der Priori-Verteilung für G namens ν und der Posteriori-Verteilung ν^θ bei gegebenen Daten $\theta_1, \dots, \theta_n$ her.

3. Haupteigenschaft des Dirichlet-Prozesses

Die Familie der Verteilungen $DP(\alpha_0 G_0)$ stellt eine zu $\theta_i | G \stackrel{iid}{\sim} G$, $i = 1, \dots, n$ konjugierte Verteilungsfamilie dar. Konkret gilt:

Gegeben sei ein Verteilungsmodell

$$\theta_i | G \stackrel{i.i.d.}{\sim} G \quad \forall i = 1, \dots, n$$

mit der Priori-Annahme

$$G \sim DP(\alpha_0 G_0).$$

Dann gilt für die Posteriori-Verteilung:

$$G | \theta_1, \dots, \theta_n \sim DP \left(\alpha_0 + n, \frac{1}{n + \alpha_0} \sum_{i=1}^n \delta_{\theta_i} + \frac{\alpha_0}{n + \alpha_0} G_0 \right).$$

Ein Nachweis dieser Aussage erfolgt im Anhang A.3. Die Posteriori-Verteilung ist demnach ein gewichtetes Mittel der empirischen Verteilung der $\{\theta_1, \dots, \theta_n\}$ und der Priori-Verteilung G_0 . Deren Gewicht $\frac{\alpha_0}{n + \alpha_0}$ konvergiert für $n \rightarrow \infty$ gegen 0, so dass der Einfluss der Priori-Annahme für steigenden Stichprobenumfang immer kleiner wird.

4.3 Stick-Breaking

4.3.1 Stick-Breaking-Repräsentation des Dirichlet-Prozesses

Ebenso wie ein Dirichlet-verteilter Vektor von Wahrscheinlichkeiten simuliert werden kann (vgl. Abschnitt 4.1), gibt es auch eine – zunächst theoretische – Möglichkeit eine Verteilung zu konstruieren, deren Verteilung einem Dirichlet-Prozess entspricht. Diese sog. „Stick-Breaking“-Konstruktion sieht gemäß Sethuraman (1994) folgende Darstellung für ein Wahrscheinlichkeitsmaß $G \sim DP(\alpha_0 G_0)$ vor:

$$G = \sum_{h=1}^{\infty} \pi_h \delta_{\phi_h}, \quad (4.4)$$

- $\phi_h \stackrel{i.i.d.}{\sim} G_0 \quad \forall h \in \mathbb{N},$
- $V_h \stackrel{i.i.d.}{\sim} Be(1, \alpha_0) \quad \forall h \in \mathbb{N},$
- $\pi_h := V_h \prod_{l < h} (1 - V_l) \quad \forall h \in \mathbb{N}.$

Dabei ist $\phi = (\phi_1, \phi_2, \dots)'$ von $V = (V_1, V_2, \dots)'$ und damit auch von $\pi = (\pi_1, \pi_2, \dots)'$ unabhängig.

Das Wahrscheinlichkeitsmaß G lässt sich also als eine unendliche Mischung von Einpunkt-Verteilungen darstellen. Sethuraman (1994) liefert den Nachweis, dass die Verteilung von G tatsächlich dem Wahrscheinlichkeitsmaß $DP(\alpha_0 G_0)$ entspricht. Der Name dieses Konstruktionsprinzips folgt aus der rekursiven Definition der Gewichte. Wegen

$$\prod_{h=1}^N (1 - V_h) = \prod_{h=1}^{N-1} (1 - V_h) - \pi_N = \dots = (1 - V_1) - \sum_{h=2}^N \pi_h = 1 - \sum_{h=1}^N \pi_h \quad (4.5)$$

erkennt man u.a.:

$$\begin{aligned} \pi_1 &= V_1, \\ \pi_2 &= V_2(1 - \pi_1), \\ \pi_3 &= V_3(1 - \pi_1 - \pi_2), \\ &\vdots \end{aligned}$$

Von einem „Stab“ der Länge 1 wird also sukzessive immer wieder ein Stück „weggebrochen“ und entfernt. Hierbei fällt auf, dass die Gewichte in ganz ähnlicher Weise wie die Wahrscheinlichkeiten bei der Dirichlet-Verteilung simuliert werden. Abbildung 4.4 visualisiert exemplarisch die Stick-Breaking-Prozedur.

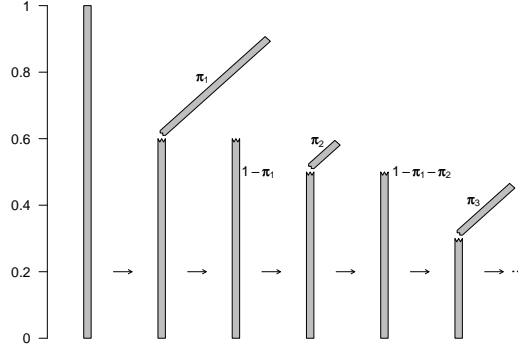


Abbildung 4.4: Visualisierung der Stick-Breaking-Prozedur

Die Vorteile der Stick-Breaking-Repräsentation sind durchaus vielschichtig. Zunächst können aus ihr die Haupteigenschaften 1 und 2 des Dirichlet-Prozesses leicht abgeleitet werden: Aus der Umformung (4.5) folgt, dass $\sum_{h=1}^N \pi_h = 1 - \prod_{h=1}^N (1 - V_h) \rightarrow 1$ für $N \rightarrow \infty$ mit Wahrscheinlichkeit 1. Daher ist G als eine gewichtete Summe von Wahrscheinlichkeitsmaßen, deren Gewichte sich zu 1 addieren, selbst wiederum ein Wahrscheinlichkeitsmaß: $G \in \mathfrak{G}$. Genauer gesagt ist G eine messbare Abbildung von (Θ, \mathcal{A}) nach $(\mathfrak{G}, \mathcal{C})$, womit die erste Haupteigenschaft bewiesen ist. Die zweite Haupteigenschaft folgt aus der Abzählbarkeit der Summe.

Darüber hinaus offeriert die Stick-Breaking-Konstruktion eine Möglichkeit ein durch einen Dirichlet-Prozess bestimmtes Wahrscheinlichkeitsmaß G zu simulieren. Die unendliche Summation, wie es (4.4) vorsieht, stellt dabei für die Praxis ein Problem dar. Muliere & Tardella (1998) schlagen deshalb eine gestutzte Form der Stick-Breaking-Konstruktion vor:

$$G_N = \sum_{h=1}^N \pi_h \delta_{\phi_h}, \quad (4.6)$$

- $\phi_h \stackrel{i.i.d.}{\sim} G_0 \quad \forall h \in \{1, \dots, N\},$
- $V_h \stackrel{i.i.d.}{\sim} Be(1, \alpha_0) \quad \forall h \in \{1, \dots, N-1\},$
 $V_N := 1,$
- $\pi_h := V_h \prod_{l < h} (1 - V_l) \quad \forall h \in \{1, \dots, N\}.$

Auch hier ist $\phi = (\phi_1, \dots, \phi_N)'$ unabhängig von $\mathbf{V} = (V_1, \dots, V_N)'$ bzw. $\pi = (\pi_1, \dots, \pi_N)'$.

Es gibt jedoch auch Verfahren, die die Trunkierung vermeiden (vgl. Walker (2007) und Papaspiliopoulos & Roberts (2008)). Nichtsdestotrotz zeichnet sich die gestutzte Form der Stick-Breaking-Konstruktion durch ihre Einfachheit und durch folgende Rechtfertigung aus: Ishwaran & James (2001) weisen nämlich nach, dass selbst bei großen Stichprobenumfängen bereits eine Trunkierung von $N = 150$ zu einer guten Approximation von G führt. Dies liegt daran, dass die Gewichte π_h mit steigendem h stochastisch kleiner werden. So gilt:

$$\begin{aligned} E \left(\sum_{h=N+1}^{\infty} \pi_h \right) &= E \left(1 - \sum_{h=1}^N \pi_h \right) = E \left(\prod_{h=1}^N (1 - V_h) \right) = \prod_{h=1}^N E(1 - V_h) = \\ &= \prod_{h=1}^N (1 - E(V_h)) = \prod_{h=1}^N \left(\frac{\alpha}{\alpha + 1} \right) = \left(\frac{\alpha}{\alpha + 1} \right)^N \xrightarrow{N \rightarrow \infty} 0. \end{aligned}$$

4.3.2 Stick-Breaking-Prioris

In der Darstellung eines Dirichlet-Prozesses als gewichtete Summe von Einpunktverteilungen liegt auch eine Möglichkeit zur Verallgemeinerung des Dirichlet-Prozesses, die auf Ishwaran & James (2001) zurück geht: Ein zufälliges Wahrscheinlichkeitsmaß G werde als Stick-Breaking-Priori, $G \sim P(\mathbf{a}, \mathbf{b})$, bezeichnet, wenn es folgende Form aufweist:

$$G = \sum_{h=1}^{\infty} \pi_h \delta_{\phi_h},$$

- $\phi_h \stackrel{i.i.d.}{\sim} G_0 \quad \forall h \in \mathbb{N}$,
- $V_h \stackrel{ind}{\sim} Be(a_h, b_h) \quad \forall h \in \mathbb{N}$,
- $\pi_h := V_h \prod_{l < h} (1 - V_l) \quad \forall h \in \mathbb{N}$.

Dabei ist $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots)'$ von $\mathbf{V} = (V_1, V_2, \dots)'$ und damit auch von $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)'$ unabhängig. Für $\mathbf{a} = (a_1, a_2, \dots)'$ und $\mathbf{b} = (b_1, b_2, \dots)'$ gilt: $a_h, b_h > 0 \quad \forall h \in \mathbb{N}$.

Stick-Breaking-Prioris umfassen u.a. folgende Spezialfälle:

1. Dirichlet-Prozess:

$$\begin{aligned} a_h &= 1 \quad \forall h \in \mathbb{N}, \\ b_h &= \alpha_0 \quad \forall h \in \mathbb{N}. \end{aligned}$$

2. Beta-Zwei-Parameter-Prozess:

$$\begin{aligned} a_h &= a \quad \forall h \in \mathbb{N}, \\ b_h &= b \quad \forall h \in \mathbb{N}. \end{aligned}$$

3. Pitman-Yor-Prozess (Zwei-Parameter Poisson-Dirichlet-Prozess):

$$\begin{aligned} a_h &= 1 - a & \forall h \in \mathbb{N}, \\ b_h &= b + h a & \forall h \in \mathbb{N}, \end{aligned}$$

mit $0 \leq a < 1$ und $b > -a$.

Analog zum trunkeierten Dirichlet-Prozess lautet die gestutzte Form der Stick-Breaking-Priori, $G_N \sim P_N(\mathbf{a}, \mathbf{b})$:

$$G_N = \sum_{h=1}^N \pi_h \delta_{\phi_h}, \quad (4.7)$$

- $\phi_h \stackrel{i.i.d.}{\sim} G_0 \quad \forall h \in \{1, \dots, N\},$
- $V_h \stackrel{ind}{\sim} Be(a_h, b_h) \quad \forall h \in \{1, \dots, N-1\},$
 $V_N := 1,$
- $\pi_h := V_h \prod_{l < h} (1 - V_l) \quad \forall h \in \{1, \dots, N\}.$

Auch hier ist $\boldsymbol{\phi} = (\phi_1, \dots, \phi_N)'$ unabhängig von $\mathbf{V} = (V_1, \dots, V_N)'$ bzw. $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)'$.

Allgemein führt die Trunkierung der Stick-Breaking-Priori zu endlich dimensionalen und damit gut handhabbaren Parametervektoren $\boldsymbol{\phi} = (\phi_1, \dots, \phi_N)'$, $\mathbf{V} = (V_1, \dots, V_N)'$ und $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)'$. Für den gemäß (4.7) konstruierten Vektor $\boldsymbol{\pi}$ führt die Endlichkeit außerdem dazu, dass $\boldsymbol{\pi}$ der verallgemeinerten Dirichlet-Verteilung folgt:

Definition: verallgemeinerte Dirichlet-Verteilung

Gegeben seien Parameter $\mathbf{a} = (a_1, \dots, a_{N-1})'$ und $\mathbf{b} = (b_1, \dots, b_{N-1})'$ mit $a_h, b_h > 0 \quad \forall h = 1, \dots, N-1$. Der Zufallsvektor $\boldsymbol{\pi}$ ist genau dann verallgemeinert Dirichlet-verteilt, $\boldsymbol{\pi} \sim GD(\mathbf{a}, \mathbf{b})$, wenn die Verteilung folgende Dichte bzgl. des $(N-1)$ -dimensionalen Lebesgue-Maßes besitzt:

$$f(\boldsymbol{\pi}) = \frac{\pi_N^{b_{N-1}-1}}{\prod_{h=1}^{N-1} B(a_h, b_h)} \prod_{h=1}^{N-1} \left[\pi_h^{a_h-1} \left(\sum_{l=h}^N \pi_l \right)^{b_{h-1}-(a_h+b_h)} \right] I_{\mathbb{S}}(\boldsymbol{\pi}).$$

Dabei gilt $\pi_N = 1 - \sum_{h=1}^{N-1} \pi_h$.

Der Vorteil dieses Zusammenhangs liegt vor allem darin, dass er – ungeachtet der unübersichtlichen Form der Dichte – eine kompakte Schreibweise der Verteilung von $\boldsymbol{\pi}$ ermöglicht. Im Spezialfall eines Dirichlet-Prozesses lautet diese $\boldsymbol{\pi} \sim GD(\mathbf{1}, \mathbf{1}_{\alpha_0})$ mit $\mathbf{1} = (1, \dots, 1)'$.

Nach diesem Exkurs über Stick-Breaking-Prioris werden im Folgenden lediglich die Dirichlet-Prozesse betrachtet werden.

4.4 Pólyas Urne

Ein Dirichlet-Prozess lässt sich auch als Grenzverteilung eines erweiterten Pólya-Urnen-Modells betrachten (vgl. Blackwell & MacQueen (1973)). Pólyas Urnenmodell liegt die Idee zugrunde, dass in mehreren Schritten aus einer Urne mit endlich vielen, farbigen Kugeln gezogen wird. In jedem Schritt wird eine Kugel gezogen und anschließend diese Kugel und eine weitere Kugel derselben Farbe in die Urne gelegt. Dieses Schema lässt sich dahingehend erweitern, dass die Urne ein stetiges Farbspektrum mit überabzählbar vielen Kugeln enthält.

Formal lässt sich diese Prozedur durch eine Pólya-Folge beschreiben. Die Menge der Kugeln sei hierbei mit Θ bezeichnet.

Definition: Pólya-Folge

Sei Θ ein polnischer Raum, d.h. ein vollständiger, separabler metrischer Raum. Das Maß $\alpha = \alpha_0 G_0$ auf (Θ, \mathcal{A}) beschreibt die Anfangsverteilung auf Θ . Die Folge von Zufallsvariablen $(\theta_n)_{n \in \mathbb{N}}$ mit $\theta_n \in \Theta$ heißt genau dann Pólya-Folge mit Parameter α bzw. $\alpha_0 G_0$, wenn für jede Menge $A \in \mathcal{A}$ gilt:

$$\begin{aligned} \text{(a)} \quad & P(\theta_1 \in A) = \frac{\alpha(A)}{\alpha(\Theta)} = G_0(A), \\ \text{(b)} \quad & P(\theta_{n+1} \in A | \theta_1, \dots, \theta_n) = \frac{\alpha(A) + \sum_{i=1}^n \delta_{\theta_i}(A)}{\alpha(\Theta) + \sum_{i=1}^n \delta_{\theta_i}(\Theta)} = \frac{\alpha_0 G_0(A) + \sum_{i=1}^n \delta_{\theta_i}(A)}{\alpha_0 + n}. \end{aligned}$$

In der Definition der Pólya-Folge lässt sich Pólyas Urnenmodell, also der Spezialfall einer Urne mit endlichen vielen Kugeln, wiederfinden, wenn man für das Maß α das Zählmaß ansetzt.

Blackwell & MacQueen (1973) brachten Pólya-Folge und Dirichlet-Prozess miteinander in Verbindung. Dieser im weiteren Verlauf aufgeführte Zusammenhang erfährt dadurch eine Einschränkung, dass man bei einer Pólya-Folge auf einen polnischen Raum beschränkt ist, während die Definition des Dirichlet-Prozesses einen beliebigen messbaren Raum zulässt. Gleichwohl ist diese Einschränkung eher theoretischer Natur und in der Praxis ohne Bedeutung. Der Zusammenhang lautet nun wie folgt:

Eine Pólya-Folge mit Parameter $\alpha_0 G_0$ besitzt für $n \rightarrow \infty$ eine (diskrete) stationäre Verteilung G , für die gilt:

1. $G \sim DP(\alpha_0 G_0)$,
2. $\theta_n | G \stackrel{i.i.d.}{\sim} G \quad \forall n \in \mathbb{N}$.

Ebenfalls lässt sich zeigen, dass von 1. und 2. ausgehend eine Pólya-Folge mit Parameter $\alpha_0 G_0$ resultiert: Da G_0 das a priori Wissen über θ beschreibt, folgt (a) unmittelbar. Der Nachweis von (b) lässt sich durch folgende Marginalisierung erreichen:

$$\begin{aligned}
 P(\theta_{n+1} \in A | \theta_1, \dots, \theta_n) &= \int P(\theta_{n+1} \in A, G | \theta_1, \dots, \theta_n) dG \\
 &= \int P(\theta_{n+1} \in A | G, \theta_1, \dots, \theta_n) f(G | \theta_1, \dots, \theta_n) dG \\
 &= \int P(\theta_{n+1} \in A | G) f(G | \theta_1, \dots, \theta_n) dG \\
 &= \int G(A) f(G | \theta_1, \dots, \theta_n) dG \\
 &= E(G(A) | \theta_1, \dots, \theta_n) \\
 &= \frac{1}{n + \alpha_0} \left(\sum_{i=1}^n \delta_{\theta_i}(A) + \alpha_0 G_0(A) \right).
 \end{aligned}$$

Hier wird die Bedeutung des Zusammenhangs zwischen Dirichlet-Prozess und Pólyas Urne deutlich: Über die Urnendarstellung lässt sich die prädiktive Verteilung gewinnen:

$$\theta_{n+1} | \theta_1, \dots, \theta_n \sim \frac{1}{n + \alpha_0} \sum_{i=1}^n \delta_{\theta_i} + \frac{\alpha_0}{n + \alpha_0} G_0. \quad (4.8)$$

Dies beinhaltet eine Möglichkeit, wie man Realisierungen aus G erhalten kann. Die Simulationsidee für θ_{n+1} bei gegebenen $\theta_1, \dots, \theta_n$ sieht also so aus, dass mit Wahrscheinlichkeit $\frac{n}{n + \alpha_0}$ aus der empirischen Verteilung der $\{\theta_1, \dots, \theta_n\}$ und mit Wahrscheinlichkeit $\frac{\alpha_0}{n + \alpha_0}$ aus G_0 gezogen wird. Das „Herausmarginalisieren“ von G hat also folgende Konsequenz: Es können Realisationen aus aber nicht von G simuliert werden. Dieser Sachverhalt wird für die Konstruktion von Gibbs-Sampling-Algorithmen, wie sie in Abschnitt 5.1 erläutert werden, von zentraler Bedeutung sein.

Aufgrund der Clustereigenschaft des Dirichlet-Prozesses können bei den einzelnen $\theta_1, \dots, \theta_n$ durchaus dieselben Werte auftreten (vgl. Abschnitt 4.2). Die prädiktive Verteilung kann daher auch über die Clusterlokationen ϕ_1, \dots, ϕ_k mit $\theta_i = \phi_{c_i}$ ausgedrückt werden:

$$\phi_{c_{n+1}} | \phi_1, \dots, \phi_k \sim \frac{1}{n + \alpha_0} \sum_{c=1}^k n_c \delta_{\phi_c} + \frac{\alpha_0}{n + \alpha_0} G_0. \quad (4.9)$$

Dabei bezeichnet n_c die Häufigkeit der Elemente in Cluster c . Dieses Simulationsprinzip nennt man auch „Chinese-Restaurant“-Prozess. Der Name erklärt sich, wenn man in (4.9) die Clusterlokationen mit Tischen eines Restaurants assoziiert. An jedem der Tische können beliebig viele Personen Platz nehmen. Die in chinesischen Restaurants oft vorzufindenden runden Tischen dienen hierfür zur Veranschaulichung. Der erste Gast des Abends findet ein leeres Lokal vor und wählt zufällig einen Tisch aus. Dieser sei mit ϕ_1 bezeichnet. Wenn nun zu einem fortgeschrittenen Zeitpunkt k Tische des Restaurants durch n Gäste besetzt sind, steht der $(n+1)$ -te Gast vor der Wahl, sich zu anderen an einen Tisch mit dazu zu setzen oder einen freien Tisch zu wählen. Gemäß (4.9) wählt er den Tisch c mit Wahrscheinlichkeit $\frac{n_c}{n+\alpha_0}$ und den nächstgelegenen freien Tisch $k+1$ mit $\frac{\alpha_0}{n+\alpha_0}$ (vgl. Abbildung 4.5). Stellt man sich nun ganz analog zum erweiterten Pólya-Urnenmodell ein Restaurant mit überabzählbar vielen Tischen vor, dann entspricht der Chinese-Restaurant-Prozess einer Pólya-Folge mit der Grenzverteilung $G \sim DP(\alpha_0 G_0)$. Das Verwenden des Clusterprinzips, wie es der Chinese-Restaurant-Prozess vollzieht, kann die Effizienz von Algorithmen zur Schätzung von $\theta_1, \dots, \theta_n$ deutlich erhöhen (vgl. Abschnitt 5.1).

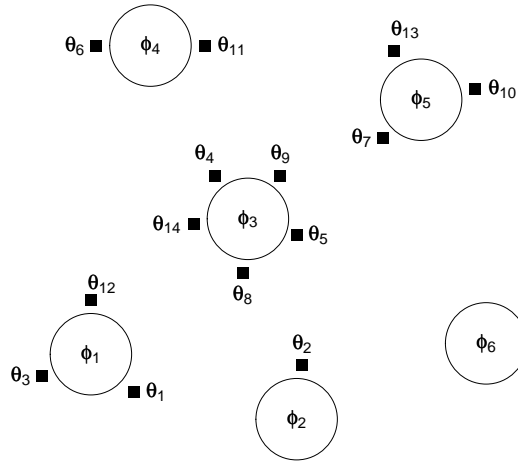


Abbildung 4.5: Veranschaulichung des „Chinese-Restaurant“-Prozesses

5 MCMC-Verfahren bei Dirichlet-Prozessen

Während in Kapitel 4 der theoretische Hintergrund der Dirichlet-Prozesse dargelegt wurde, widmet sich dieses Kapitel der Schätzung eines DPM-Modells, wie es dort vorgestellt wurde.

$$\begin{aligned} y_i | \theta_i &\stackrel{\text{ind}}{\sim} F(\theta_i) & \forall i = 1, \dots, n, \\ \theta_i | G &\stackrel{\text{i.i.d.}}{\sim} G & \forall i = 1, \dots, n, \\ G &\sim DP(\alpha_0 G_0). \end{aligned} \tag{5.1}$$

Dabei werden die Basisverteilung G_0 , der Verteilungstyp von $F(\theta_i)$ sowie die Daten $\mathbf{y} = (y_1, \dots, y_n)'$ als gegeben betrachtet. Auch der Präzisionsparameter α_0 sei zunächst bekannt. Wenn man nun analog zu Kapitel 4 die Dichte f bzw. die entsprechende Verteilung F schätzen will, ist es erforderlich, die Parameter $\theta_1, \dots, \theta_n$ zu bestimmen. Deren Schätzung ist das erklärte Ziel dieses Kapitels. Sie werden als reellwertig angenommen, so dass der dem Dirichlet-Prozess zugrunde liegende Messraum durch $(\Theta, \mathcal{A}) = (\mathbb{R}, \mathfrak{B})$ bestimmt ist. Da \mathbb{R} ein polnischer Raum ist, kann der Zusammenhang des Dirichlet-Prozesses zur Pólya-Folge hergestellt werden (vgl. Abschnitt 4.4).

Wie bereits in Abschnitt 3.3 angedeutet, ist man bei der Schätzung von $\theta_1, \dots, \theta_n$ mit einem grundlegenden Problem konfrontiert: Wie ist die unbekannte Verteilung G zu handhaben? Schließlich handelt es sich bei G nicht um einen reellwertigen Parameter, sondern um eine Verteilung. Es haben sich zwei Wege entwickelt, wie man diesem Schätzproblem begegnen kann. Der eine Weg basiert auf Pólyas Urne. Dort entledigt man sich durch ein „Herausmarginalisieren“ von G des Problems einer unbekannten Verteilung. Dieses Konzept wird daher auch als marginale Methode bezeichnet (vgl. Abschnitt 5.1). Der zweite Weg greift die Stick-Breaking-Repräsentation wieder auf. In ihrer gestutzten Form wird die Verteilung G_N , die für hinreichend große N als gute Approximation von G aufgefasst werden kann, in eine parametrische Darstellung gebracht. Es liegen damit ausschließlich reellwertige, zwar entsprechend hochdimensionale Parameter vor, die durch übliche bayesianische Inferenzmethoden geschätzt werden können. Da hierbei G nicht herausmarginalisiert, sondern die bedingte Struktur beibehalten wird, heißt dieser Weg auch die bedingte Methode (vgl. Abschnitt 5.2). Beide Wege verwenden, da die Posteriori-Verteilungen nicht in geschlossener Form darstellbar sind, MCMC-Verfahren. Konkret handelt es sich bei den in den Abschnitten 5.1 und 5.2 vorgestellten Algorithmen stets um Gibbs-Sampler. In Abschnitt 5.3 sollen schließlich deren Charakteristika gegenüber gestellt und mit Erweiterungsmöglichkeiten ergänzt werden.

5.1 Gibbs-Sampling basierend auf Pólyas Urne

Der folgende Abschnitt beschreibt und vergleicht fünf Algorithmen, die innerhalb des marginalen Konzepts eine zentrale Rolle spielen. Die Ausführungen orientieren sich dabei an der guten Zusammenstellung in Neal (2000) und an der jeweiligen Primärliteratur.

5.1.1 Algorithmus nach Escobar (1994)

Escobar (1994) entwickelte den ersten MCMC-Algorithmus für Dirichlet-Prozesse. Diese Methode stellt ein Gibbs-Sampling-Verfahren dar, bei dem die $\theta_1, \dots, \theta_n$ komponentenweise aufdatiert werden. Die Vorschlagsdichte für θ_i ist gemäß (2.4) durch folgende vollständig bedingte Dichte bestimmt:

$$p(\theta_i | \theta_{-i}, \mathbf{y}) \propto f(y_i | \theta_i) p(\theta_i | \theta_{-i}).$$

Während die Likelihood $f(y_i | \theta_i)$ als Funktion von θ_i von vornherein gegeben ist, kann die Verteilung von $\theta_i | \theta_{-i}$ über die Gleichung (4.8) hergeleitet werden: Da wegen $\theta_i | G \stackrel{i.i.d.}{\sim} G$ die θ_i vertauschbar sind, darf die Beobachtung i , für die ein Aufdatieren erfolgen soll, stets als die „letzte“ der n Beobachtungen aufgefasst werden:

$$\theta_i | \theta_{-i} \sim \frac{1}{n-1+\alpha_0} \sum_{j \neq i} \delta_{\theta_j} + \frac{\alpha_0}{n-1+\alpha_0} G_0. \quad (5.2)$$

Dabei gilt $\theta_{-i} = \{\theta_1, \dots, \theta_n\} \setminus \theta_i$.

Man bezieht sich hier also auf die Verbindung des Dirichlet-Prozesses zu Pólyas Urne und der daraus resultierenden prädiktiven Verteilung (4.8). Zu deren Herleitung wurde über die unbekannte Verteilung G integriert. Durch dieses „Herausmarginalisieren“ von G wurde das bei der Modellschätzung auftretende Problem, dass mit $\theta_1, \dots, \theta_n$ Werte geschätzt werden sollen, die einer unbekannten Verteilung folgen, auf elegante Weise gelöst.

Die Struktur von Gleichung (5.2) lässt sich auf folgende Weise abstrahieren:

$$\theta_i | \theta_{-i} \sim \sum_{j \neq i} q_{ij} \delta_{\theta_j} + r_i Q \quad (5.3)$$

mit

$$\begin{aligned} Q &= G_0, \\ q_{ij} &= \frac{1}{n-1+\alpha_0}, \\ r_i &= \frac{\alpha_0}{n-1+\alpha_0}. \end{aligned}$$

Dabei entspricht r_i der Wahrscheinlichkeit, dass θ_i einen Wert annimmt, der sich von denen aller anderen Individuen unterscheidet. In diesem Fall kann θ_i jeden Wert des Parameterraums Θ annehmen. Die Wahrscheinlichkeit hierfür lautet zunächst nach dem Satz von der totalen Wahrscheinlichkeit $\int dG_0(\theta_i)$, wird jedoch durch die aus der Pólya-Urnen-Repräsentation resultierenden Gewichtung zu: $r_i = \frac{\alpha_0}{n-1+\alpha_0} \int dG_0(\theta_i) = \frac{\alpha_0}{n-1+\alpha_0}$. Unter q_{ij}

verstehen man die Wahrscheinlichkeit, dass die Person i denselben θ -Wert wie die Person j annimmt. Diese ist hier unabhängig von j .

Unter Berücksichtigung der beobachteten Daten bleibt die Struktur erhalten. Die Gewichte ändern sich dahingehend, dass für q_{ij} die Plausibilität von θ_j unter der Beobachtung y_i einbezogen werden muss ebenso wie für r_i die Likelihood $f(y_i|\theta_i)$ in den Satz der totalen Wahrscheinlichkeit eingeht. In der Verteilung Q wird dem durch die Beobachtung gewonnenen Wissen Rechnung getragen. Der Algorithmus lautet demnach:

Algorithmus 1:

Die Markov-Kette befinde sich im Zustand $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)'$.

(I) Für $i = 1, \dots, n$:

- Entferne θ_i .
- Ziehe neuen Wert für θ_i gemäß:

$$\theta_i | \theta_{-i}, y_i \sim \sum_{j \neq i} q_{ij} \delta_{\theta_j} + r_i Q$$

mit

$$\begin{aligned} Q &= H_i, \\ q_{ij} &= b \frac{1}{n-1+\alpha_0} f(y_i|\theta_j), \\ r_i &= b \frac{\alpha_0}{n-1+\alpha_0} \int f(y_i|\theta_i) dG_0(\theta_i). \end{aligned}$$

Dabei ist H_i die a posteriori mit der Likelihood $f(y_i|\theta_i)$ und der Priori-Verteilung G_0 , während b hier sowie im Folgenden eine Proportionalitätskonstante darstellt, so dass gilt: $\sum_{j \neq i} q_{ij} + r_i = 1$. Auch wenn sich der Faktor $\frac{1}{n-1+\alpha_0}$ in die Proportionalitätskonstante hineinziehen ließe, so soll er zur besseren Vergleichbarkeit mit anderen Algorithmen weiterhin aufgeführt werden.

Der Algorithmus erfordert zum einen die Berechnung des Integrals $\int f(y_i|\theta_i) dG_0(\theta_i)$ und zum anderen das Ziehen von Zufallszahlen aus H_i . Im Falle der Konjugiertheit der Likelihood $f(y_i|\theta_i)$ zur a priori G_0 kann das Integral analytisch bestimmt werden und das Ziehen von Zufallszahlen aus H_i stellt kein Problem dar. Ohne Konjugiertheit sind zur Berechnung des Integrals meist rechenintensive, numerische Verfahren notwendig (vgl. MacEachern & Müller (1998)). Auch das Ziehen von Zufallszahlen aus H_i ist dann i.d.R. schwierig.

Der Algorithmus wird in Escobar (1994) mit $F(\theta_i) = N(\theta_i, 1)$ und in Escobar & West (1995) mit $F(\theta_i = (\mu_i, \sigma_i^2)) = N(\mu_i, \sigma_i^2)$ verwendet. Er konvergiert allerdings sehr langsam zur Posteriori-Verteilung, da ein separates Aufdatieren für die Individuen, wie es der Algorithmus vollzieht, wegen der Clustereigenschaft der θ_i ineffizient ist (vgl. Neal (2000)).

Bilden einige Individuen einen Cluster, geht beim Aufdatieren eines Individuums aus diesem Cluster die Tendenz dahin, dass wieder der Wert dieses Clusters angenommen wird. Änderungen von Clusterlokationen ereignen sich selten.

5.1.2 Algorithmus nach West, Müller und Escobar (1994) und Bush und MacEachern (1996)

Eine wesentlich höhere Konvergenzgeschwindigkeit lässt sich erzielen, wenn für alle Individuen eines Clusters gleichzeitig ein Update erfolgt, also wenn die Stellen der Cluster statt die der Individuen aufdatiert werden. Gleichung (5.2) lautet demnach mit Summation über die Cluster statt über die Individuen:

$$\theta_i | \theta_{-i} \sim \sum_{c=1}^{k-i} q_{ic}^* \delta_{\phi_c} + r_i Q \quad (5.4)$$

mit

$$\begin{aligned} Q &= G_0, \\ q_{ic}^* &= \frac{n_{-i,c}}{n-1+\alpha_0}, \\ r_i &= \frac{\alpha_0}{n-1+\alpha_0}. \end{aligned}$$

Dabei steht $k-i$ für die Anzahl der Cluster und $n_{-i,c}$ für die Anzahl der Individuen im Cluster c – jeweils ohne Individuum i . q_{ic}^* entspricht der Wahrscheinlichkeit, dass die Person i in den Cluster c fällt, während r_i weiterhin für die Wahrscheinlichkeit steht, dass i einen neuen Cluster bildet.

Werden nun die beobachteten Daten mit einbezogen, so ändern sich die Gewichte q_{ic}^* und r_i analog zu Algorithmus 1 zu:

$$\begin{aligned} q_{ic}^* &= b \frac{n_{-i,c}}{n-1+\alpha_0} f(y_i | \phi_c), \\ r_i &= b \frac{\alpha_0}{n-1+\alpha_0} \int f(y_i | \phi_{c_i}) dG_0(\phi_{c_i}). \end{aligned}$$

Jeder Iterationsschritt besteht nun aus zwei Stufen. In der ersten Stufe wird anhand q_{ic}^* bzw. r_i bestimmt, welche Person zu welchem Cluster gehört. In der zweiten Stufe werden die Clusterlokationen erneuert.

Algorithmus 2:

Die Markov-Kette befinde sich im Zustand $\mathbf{c} = (c_1, \dots, c_n)'$ und $\boldsymbol{\phi} = (\phi_c : c \in \{c_1, \dots, c_n\})$.

(I) Für $i = 1, \dots, n$:

- Entferne c_i .
- Falls $n_{-i,c_i} = 0$, entferne ϕ_{c_i} .

- Ziehe neuen Wert für c_i gemäß:

$$P(c_i = c \mid c_{-i}, y_i, \phi) = b \frac{n_{-i,c}}{n-1+\alpha_0} f(y_i \mid \phi_c) \quad \text{falls } \exists j \neq i \text{ mit } c_j = c,$$

$$P(c_i \neq c_j, \forall j \neq i \mid c_{-i}, y_i, \phi) = b \frac{\alpha_0}{n-1+\alpha_0} \int f(y_i \mid \phi_{c_i}) dG_0(\phi_{c_i}).$$

- Falls $n_{c_i} = 1$, ziehe Wert für ϕ_{c_i} aus H_i .

(II) Für alle $c \in (c_1, \dots, c_n)$:

- Ziehe neuen Wert für ϕ_c aus H_c .

Dabei entspricht H_c der Posteriori, deren Dichte sich wie folgt darstellt:

$$p(\phi_c \mid y_i : c_i = c) \propto \left(\prod_{i: c_i = c} f(y_i \mid \phi_c) \right) g_0(\phi_c).$$

Wie im Algorithmus 1 ist auch hier die Berechnung des Integrals $\int f(y_i \mid \phi_{c_i}) dG_0(\phi_{c_i})$ und das Ziehen von Zufallszahlen aus H_i sowie zusätzlich aus H_c notwendig. Im Falle der Konjugiertheit sind all diese Berechnungen durchführbar.

Bush & MacEachern (1996) gebrauchen diese Form des Aufdatierens mit dem Unterschied, dass dort statt des Schritts (I) der Algorithmus 1 verwendet wird, wobei von den simulierten θ -Werten nur die Clusterstruktur interessiert. West, Müller & Escobar (1994) verwenden exakt den Algorithmus 2 für $F(\theta_i = (\mu_i, \Sigma_i)') = N(X_i \mu_i, \Sigma_i)$ mit einer hierzu konjugierten Normal- bzw. inversen Wishartverteilung für die Basisverteilung G_0 . Sie erweitern diesen Ansatz zudem für den nichtkonjugierten Fall. Dies macht es notwendig in Schritt (II) die clusterspezifischen μ_i und Σ_i abwechselnd und wechselseitig aufeinander bedingt aufzudatieren. Außerdem wird vorgeschlagen, das Integral durch numerische Quadratur oder durch Monte-Carlo-Integration zu berechnen. Diese Approximation führt allerdings zu leicht verfälschten Übergangswahrscheinlichkeiten. Die auf diese Weise konstruierte Markov-Kette besitzt deshalb eine nur approximativ mit der Posteriori übereinstimmende stationäre Verteilung (vgl. MacEachern & Müller (1998)).

5.1.3 Algorithmus nach MacEachern (1994)

Im Algorithmus 2 bedingen sich beide Stufen gegenseitig. Von dieser gegenseitigen Abhängigkeit kann man sich lösen, indem man in den Vorschlagswahrscheinlichkeiten von Stufe (I) über die Clusterlokationen integriert. So kann die Clusterstruktur unabhängig von den Clusterlokationen iterativ bestimmt werden. Liegt jene vor, kann bei der Schätzung der Clusterlokationen auf iterative Verfahren verzichtet werden. Unter Annahme der Konjugiertheit sind nämlich die Posteriori-Verteilungen in der zweiten Stufe analytisch zugänglich.

Algorithmus 3:

Die Markov-Kette befinde sich im Zustand $\mathbf{c} = (c_1, \dots, c_n)'$.

(I) Für $i = 1, \dots, n$:

- Entferne c_i .
- Ziehe neuen Wert für c_i gemäß:

$$P(c_i = c \mid c_{-i}, y_i) = b \frac{n_{-i,c}}{n-1+\alpha_0} \int f(y_i \mid \phi_c) dH_{-i,c}(\phi_{c_i}) \quad \text{falls } \exists j \neq i \text{ mit } c_j = c,$$

$$P(c_i \neq c_j, \forall j \neq i \mid c_{-i}, y_i) = b \frac{\alpha_0}{n-1+\alpha_0} \int f(y_i \mid \phi_{c_i}) dG_0(\phi_{c_i}).$$

Anschließend werden die Clusterlokationen gemäß H_c bestimmt.

MacEachern (1994) benutzt für $F(\theta_i) = N(\theta_i, \sigma_i^2)$ ein solches Gibbs-Sampling-Verfahren, das sich nur auf die Clusterstruktur bezieht. Dort wird zur Bestimmung der Vorschlagswahrscheinlichkeiten eine alternative Methode angegeben, die komplizierte Berechnungen nötig macht, dafür aber auf Integrationen verzichtet.

5.1.4 Algorithmus nach MacEachern und Müller (1998)

Der Algorithmus von MacEachern & Müller (1998) kommt ohne die Berechnung des Integrals aus und ist damit auch anwendbar, wenn $y_i \mid \theta_i$ nicht zur a priori G_0 konjugiert ist. Die Idee des Algorithmus setzt an folgendem Punkt an: Bei den Algorithmen 1–3 definieren sich die Cluster allein durch identische θ -Werte bei den Individuen. Für die Wahrscheinlichkeit, dass eine Person i in keinen der bestehenden k_{-i} Cluster fällt, sondern einen neuen, eigenen Cluster bildet, ist daher stets die Berechnung des Integrals $\int f(y_i \mid \phi_{c_i}) dG_0(\phi_{c_i})$ notwendig, da die Stelle dieses Clusters jeden Wert des Parameterraums Θ annehmen kann.

MacEachern & Müller (1998) hingegen ordnen die Cluster nach folgender Struktur: Die Clusterlokationen ϕ_1, \dots, ϕ_k werden durch die Lokationen potentieller Cluster $\phi_{k+1}, \dots, \phi_n$ ergänzt:

$$\underbrace{(\phi_1, \dots, \phi_k)}_{\phi_F}, \underbrace{(\phi_{k+1}, \dots, \phi_n)}_{\phi_E}.$$

Dabei steht ϕ_F für die „vollen“ Cluster und ϕ_E für die „leeren“ Cluster. In dieser Anordnung der Cluster liegt die Annahme, dass es zwischen den vollen Clustern keine „Lücke“ gibt, also dass sich zwischen zwei vollen Clustern niemals ein leerer Cluster befindet. Der Algorithmus wird daher auch als „no gaps“-Algorithmus bezeichnet.

Die Annahme hat zwei Konsequenzen:

1. Wenn durch das Entfernen der Beobachtung i eine Lücke entsteht, also wenn $n_{-i,c} = 0$ für $1 \leq c < k$, dann muss i mit Wahrscheinlichkeit 1 wieder in diesen Cluster.

2. Wenn durch das Entfernen der Beobachtung i keine Lücke entsteht, dann entspricht die Wahrscheinlichkeit, dass i einen neuen Cluster bildet, der, dass i in den $(k_{-i} + 1)$ -ten Cluster fällt. Entscheidend ist hierbei, dass man den neuen Cluster nun nicht mehr durch alle möglichen Stellen, die dieser Cluster als Lokation haben kann, identifiziert, sondern einzig durch die Tatsache, dass es sich dabei um den Cluster $(k_{-i} + 1)$ handelt.

Die Gewichte von (5.4) lauten demnach für den Fall, dass durch das Entfernen der Beobachtung i keine Lücke entstanden ist:

$$q_{ic}^* = b \frac{n_{-i,c}}{n-1+\alpha_0},$$

$$r_i = b \frac{\alpha_0}{n-1+\alpha_0} \frac{1}{k_{-i}+1}.$$

Bezieht man die beobachteten Daten ein, so lauten sie:

$$q_{ic}^* = b \frac{n_{-i,c}}{n-1+\alpha_0} f(y_i|\phi_c),$$

$$r_i = b \frac{\alpha_0}{n-1+\alpha_0} \frac{1}{k_{-i}+1} f(y_i|\phi_{k_{-i}+1}).$$

Auf diese Weise kann sich die Anzahl der Cluster k nur dann verringern, wenn $n_{c_i} = 1$ und wenn $c_i = k$. Deshalb ist es im Fall $n_{c_i} = 1$ notwendig, die Personen zu permutieren. Dies führt mit Wahrscheinlichkeit $\frac{1}{k}$ zu $c_i = k$ und mit Wahrscheinlichkeit $\frac{k-1}{k}$ zu $1 \leq c_i < k$. Zusammenfassend lautet der Algorithmus:

Algorithmus 4:

Die Markov-Kette befinde sich im Zustand $\mathbf{c} = (c_1, \dots, c_n)'$ und $\boldsymbol{\phi} = (\phi_c : c \in \{c_1, \dots, c_n\})$.

(I) Für $i = 1, \dots, n$:

- Falls $n_{c_i} > 1$, ziehe einen Wert für $\phi_{k_{-i}+1}$ aus G_0 .
- Falls $n_{c_i} = 1$, behalte entweder mit Wahrscheinlichkeit $\frac{k-1}{k}$ den aktuellen Wert von c_i bei und beende den Iterationsdurchlauf an dieser Stelle oder fahre mit Wahrscheinlichkeit $\frac{1}{k}$ in der Iterationsschleife fort.
- Entferne c_i .
- Ziehe neuen Wert für c_i gemäß:

$$P(c_i = c | c_{-i}, y_i, \phi_1, \dots, \phi_{k_{-i}+1}) = b \frac{n_{-i,c}}{n-1+\alpha_0} f(y_i|\phi_c) \quad \text{für } 1 \leq c \leq k_{-i},$$

$$P(c_i = c | c_{-i}, y_i, \phi_1, \dots, \phi_{k_{-i}+1}) = b \frac{\alpha_0}{n-1+\alpha_0} \frac{1}{k_{-i}+1} f(y_i|\phi_c) \quad \text{für } c = k_{-i}+1.$$

- Entferne die ϕ_c mit $n_c = 0$.

(II) Für alle $c \in (c_1, \dots, c_n)$:

- Ziehe neuen Wert für ϕ_c aus H_c .

Während eine eventuell schwierige Berechnung des Integrals nicht notwendig ist, kann das Ziehen von Zufallzahlen aus H_c im nichtkonjugierten Fall weiterhin problematisch sein. Darüberhinaus liegt dem Algorithmus die Tendenz zu Grunde, dass sich Änderungen in der Anzahl der Cluster zu selten ereignen. Das Gewicht r_i von Algorithmus 4 ist im Falle ohne Daten wegen des Faktors $\frac{1}{k_{-i}+1}$ niedriger als es Pólyas Urnendarstellung (5.4) vorsieht, was auch bei Einbeziehung der Daten zu tendenziell niedrigeren Wahrscheinlichkeiten für eine Erhöhung der Clusteranzahl führt. Entsprechend sind die Gewichte q_{ic}^* höher. Eine gleichbleibende Clusteranzahl wird dadurch ebenfalls wahrscheinlicher wie durch den Umstand, dass im Falle $n_{c_i} = 1$ mit Wahrscheinlichkeit $\frac{k-1}{k}$ keine Veränderung an c_i vorgenommen wird. Insgesamt macht dies den Sampling-Mechanismus ineffizient.

5.1.5 Algorithmus mit Hilfsvariablen

In Algorithmus 4 fungiert $\phi_{k_{-i}+1}$ als Hilfsvariable. Als Lokation eines zunächst leeren Clusters scheint sie verzichtbar, dient aber dem Zweck, der Beobachtung i einen Cluster zur Verfügung zu stellen, falls jene in keinen der mit anderen Beobachtungen assoziierten Cluster gelangt. Dieses Prinzip lässt sich nun dahingehend erweitern, dass nicht nur eine, sondern mehrere Hilfsvariablen stellvertretend für mehrere potentielle Cluster während jedes Iterationsschritts erzeugt werden. Die Idee eines solchen Samplings kann man allgemein wie folgt beschreiben:

Die stationäre Verteilung einer Markov-Kette sei $\pi(x)$. Ein Iterationsschritt, in dem x aufdatiert werden soll, lautet:

1. Ziehe einen Wert für y gemäß $\pi(y|x)$.
2. Datiere (x, y) mit der gemeinsamen Verteilung $\pi(x, y)$ als stationärer Verteilung auf.
3. Entferne y .

Dieses Vorgehen liefert eine Markov-Kette für x , solange $\pi(x)$ die marginale Verteilung von $\pi(x, y)$ ist.

Im Folgenden wird ein Algorithmus mit m Hilfsvariablen $\phi_{k_{-i}+1}, \dots, \phi_{k_{-i}+m}$ konstruiert. Analog zu (5.4) soll die Wahrscheinlichkeit, dass i in einen der m leeren Cluster fällt, $\frac{\alpha_0/m}{n-1+\alpha_0}$ sein. Der Algorithmus gestaltet sich folgendermaßen:

Algorithmus 5:

Die Markov-Kette befinde sich im Zustand $\mathbf{c} = (c_1, \dots, c_n)'$ und $\boldsymbol{\phi} = (\phi_c : c \in \{c_1, \dots, c_n\})$.

(I) Für $i = 1, \dots, n$:

- Falls $n_{c_i} > 1$, ziehe Werte für $\phi_{k_{-i}+1}, \dots, \phi_{k_{-i}+m}$ unabhängig aus G_0 .

- Falls $n_{c_i} = 1$, vertausche innerhalb \mathbf{c} so, dass $c_i = k_{-i} + 1$ und ziehe Werte für $\phi_{k_{-i}+2}, \dots, \phi_{k_{-i}+m}$ unabhängig aus G_0 .
- Entferne c_i .
- Ziehe neuen Wert für c_i gemäß:

$$P(c_i = c | c_{-i}, y_i, \phi_1, \dots, \phi_{k_{-i}+m}) = b \frac{n_{-i,c}}{n-1+\alpha_0} f(y_i | \phi_c) \text{ für } 1 \leq c \leq k_{-i},$$

$$P(c_i = c | c_{-i}, y_i, \phi_1, \dots, \phi_{k_{-i}+m}) = b \frac{\alpha_0/m}{n-1+\alpha_0} f(y_i | \phi_c) \text{ für } k_{-i} < c \leq k_{-i}+m.$$

- Entferne die ϕ_c mit $n_c = 0$.

(II) Für alle $c \in (c_1, \dots, c_n)$:

- Ziehe neuen Wert für ϕ_c aus H_c .

Dieser Algorithmus besitzt für $m = 1$ eine starke Ähnlichkeit zu Algorithmus 4, unterscheidet sich aber dadurch, dass die Wahrscheinlichkeit, dass i in keinen der durch die anderen Personen besetzten Cluster fällt, größer ist. Ebenso ist die Wahrscheinlichkeit für eine Reduzierung der Clusteranzahl höher.

Für $m \rightarrow \infty$ ähnelt der Algorithmus dem Algorithmus 2. Die Wahrscheinlichkeit, dass i in einen der durch die m Hilfsvariablen definierten Cluster fällt, lautet:

$$\sum_{c=1}^m b \frac{\alpha_0/m}{n-1+\alpha_0} f(y_i | \phi_c) = b \frac{\alpha_0}{n-1+\alpha_0} \frac{1}{m} \sum_{c=1}^m f(y_i | \phi_c).$$

Da die m gezogenen ϕ -Werte Zufallszahlen aus G_0 sind, entspricht die rechte Seite der Wahrscheinlichkeit $P(c_i \neq c_j, \forall j \neq i | c_{-i}, y_i, \phi)$ im Algorithmus 2 bei Monte-Carlo-Integration. Nichtsdestotrotz führt Algorithmus 5 für alle $m \in \mathbb{N}$ zur exakten stationären Verteilung, obwohl dies bei Algorithmus 2 mit Monte-Carlo-Integration nicht der Fall ist.

5.2 Gibbs-Sampling über Stick-Breaking

Mit Hilfe der Stick-Breaking-Konstruktion kann ein zufälliges Wahrscheinlichkeitsmaß G , dessen Verteilung durch einen Dirichlet-Prozess bestimmt ist, durch abzählbar unendlich viele Parameter identifiziert werden (vgl. (4.4)). Bezieht man sich gemäß der Rechtfertigung in Abschnitt 4.3.1 auf die trunkierte Form des Dirichlet-Prozesses (vgl. (4.6)) so kann die Verteilung zumindest approximativ durch endlich viele reellwertige Parameter repräsentiert werden. Man kann daher Realisationen von G_N , die dann als Realisierungen von G aufgefasst werden, erhalten, indem man die Parametervektoren ϕ_1, \dots, ϕ_N und V_1, \dots, V_{N-1} simuliert. $\phi = (\phi_1, \dots, \phi_N)'$ symbolisiert dabei die Lokationen der maximal N möglichen Cluster, während $\mathbf{V} = (V_1, \dots, V_N)'$ zur Konstruktion der Gewichte $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)'$ benötigt wird. Zufallszahlen $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)'$ sind dann wegen $\theta_i = \phi_{c_i}$ über die Lokationen ϕ und die Klassifikationsvariablen $\mathbf{c} = (c_1, \dots, c_n)'$ bestimmbar. Diese erhält man, indem man unter Verwendung der Wahrscheinlichkeiten $\boldsymbol{\pi}$

per Zufall entscheidet, welcher der theoretisch möglichen Lokationen ϕ_1, \dots, ϕ_N eine Person i für $i = 1, \dots, n$ zugeordnet wird. Es lassen sich also durch die Simulation von ϕ , π und \mathbf{c} Realisationen aus G , $\theta = (\theta_1, \dots, \theta_n)'$, bestimmen.

In demselben Sinne kann auch eine Schätzung von θ , wie sie zu Beginn dieses Kapitels als Ziel ausgegeben wurde, über die Schätzung von ϕ , π und \mathbf{c} erreicht werden. Das Modell in (5.1) lässt sich daher neu formulieren:

$$\begin{aligned} y_i | \phi, \mathbf{c} &\stackrel{\text{ind}}{\sim} F(\phi_{c_i}) & \forall i = 1, \dots, n, \\ \phi &\sim G_0^{\otimes N}, \\ \mathbf{c} | \pi &\sim K^{\otimes n}, \\ \pi &\sim GD(\mathbf{1}, \mathbf{1}\alpha_0). \end{aligned}$$

Dabei entspricht $G_0^{\otimes N} := G_0 \otimes \dots \otimes G_0$ dem N -fachen Produktmaß von G_0 und $K^{\otimes n} := K \otimes \dots \otimes K$ dem n -fachen Produktmaß von $K := \sum_{h=1}^N \pi_h \delta_h$. Weiterhin gilt: $\phi \perp \pi$.

Unter den gegebenen Voraussetzungen folgt für die Posteriori $\phi, \mathbf{c}, \pi | \mathbf{y}$:

$$p(\phi, \mathbf{c}, \pi | \mathbf{y}) \propto \left(\prod_{i=1}^n f(y_i | \phi, \mathbf{c}, \pi) \right) p(\phi, \mathbf{c}, \pi) \propto \left(\prod_{i=1}^n f(y_i | \phi, \mathbf{c}) \right) p(\phi) p(\mathbf{c} | \pi) p(\pi).$$

Die Verteilung der Posteriori ist analytisch nicht zugänglich, so dass MCMC-Verfahren zur Schätzung von ϕ , π und \mathbf{c} notwendig sind. Der sog. Block-Gibbs-Algorithmus geht nun so vor, dass die Parameter ϕ , π und \mathbf{c} blockweise über ihre jeweiligen vollständig bedingten Dichten in jedem Iterationsschritt aufdatiert werden. Im Folgenden wird hergeleitet, auf welche Weise die vollständig bedingten Dichten zu bestimmen sind.

Aufdatieren von ϕ :

$$\begin{aligned} p(\phi | \mathbf{c}, \pi, \mathbf{y}) &\propto p(\phi | \mathbf{c}, \mathbf{y}) \propto \left(\prod_{i=1}^n f(y_i | \phi, \mathbf{c}) \right) p(\phi) = \\ &= \left(\prod_{h=1}^N \prod_{i: c_i=h} f(y_i | \phi_h) \right) \left(\prod_{h=1}^N g_0(\phi_h) \right) = \prod_{h=1}^N \left(g_0(\phi_h) \prod_{i: c_i=h} f(y_i | \phi_h) \right). \end{aligned}$$

Es lassen sich daher für das Aufdatieren von ϕ_h zwei Fälle unterscheiden: Für die leeren Cluster wird als Vorschlagsdichte für ϕ_h die bekannte Dichte g_0 verwendet. Gehören dem Cluster h hingegen Individuen an, so wird ϕ_h aus der Posteriori gezogen, deren Priori-Dichte durch g_0 bestimmt ist und deren Likelihood sich aus den y -Werten der zugehörigen Individuen zusammensetzt. Hierfür ist die Konjugiertheit von F zu G_0 nützlich.

Aufdatieren von \mathbf{c} :

Die Verteilung von \mathbf{c} entspricht einer diskreten Verteilung. Die vollständig bedingten Wahrscheinlichkeiten $P(c_i = h | \phi, \pi, \mathbf{y})$ für $i = 1, \dots, n$ lassen sich wie folgt bestimmen:

$$P(c_i = h | \phi, \pi, \mathbf{y}) \propto f(y_i | \phi_h) \pi_h.$$

Aufdatieren von π :

$$p(\pi | \phi, \mathbf{c}, \mathbf{y}) \propto p(\pi | \mathbf{c}) \propto p(\mathbf{c} | \pi) p(\pi).$$

Bei der Bestimmung von $p(\mathbf{c} | \pi)$ ist zu bedenken, dass die Information von $\mathbf{c} = (c_1, \dots, c_n)'$ gleichbedeutend ist mit der von $(n_1, \dots, n_N)'$, dem Vektor der Häufigkeiten in allen möglichen Clustern. Da $(n_1, \dots, n_N)' \sim M(n, \pi)$, gilt für $p(\mathbf{c} | \pi)$:

$$\begin{aligned} p(\mathbf{c} | \pi) &= p(n_1, \dots, n_N | \pi) \propto \prod_{h=1}^N \pi_h^{n_h} = \prod_{h=1}^N \left(V_h \prod_{l=1}^{h-1} (1 - V_l) \right)^{n_h} = \\ &= \prod_{h=1}^N V_h^{n_h} \cdot \prod_{h=1}^N \prod_{l=1}^{h-1} (1 - V_l)^{n_h} = \prod_{h=1}^{N-1} V_h^{n_h} \cdot \prod_{h=1}^{N-1} (1 - V_h)^{\sum_{l=h+1}^N n_l}. \end{aligned}$$

Für $p(\pi)$ gilt:

$$p(\pi) = p(V_1, \dots, V_{N-1}) = \prod_{h=1}^{N-1} p(V_h) = \prod_{h=1}^{N-1} \frac{1}{B(1, \alpha_0)} (1 - V_h)^{\alpha_0 - 1} \propto \prod_{h=1}^{N-1} (1 - V_h)^{\alpha_0 - 1}.$$

Daraus folgt für $p(\pi | \mathbf{c})$:

$$p(\pi | \mathbf{c}) \propto \prod_{h=1}^{N-1} V_h^{n_h} (1 - V_h)^{\alpha_0 + \sum_{l=h+1}^N n_l - 1}.$$

Zusammenfassend lautet der Block-Gibbs-Sampler daher:

Algorithmus 6:

Die Markov-Kette befinde sich im Zustand $\phi = (\phi_1, \dots, \phi_N)'$, $\mathbf{c} = (c_1, \dots, c_n)'$ und $\pi = (\pi_1, \dots, \pi_N)'$.

(I) Für $h = 1, \dots, N$:

- Ziehe neuen Wert für ϕ_h gemäß:

$$\phi_h | \mathbf{c}, \mathbf{y} \sim G_0 \quad \text{falls } \nexists i : c_i = h,$$

$$\phi_h | \mathbf{c}, \mathbf{y} \sim H_h \quad \text{falls } \exists i : c_i = h.$$

(II) Für $i = 1, \dots, n$:

- Ziehe neuen Wert für c_i gemäß:

$$c_i | \phi, \pi, y_i \sim \sum_{h=1}^N c^* f(y_i | \phi_h) \pi_h \delta_h.$$

(III) Für $h = 1, \dots, N$:

- Ziehe neuen Wert für V_h (außer für $h = N$, da stets $V_N = 1$) gemäß:

$$V_h | c \sim Be(1 + n_h, \alpha_0 + \sum_{l=h+1}^N n_l).$$

- Konstruiere π_h gemäß:

$$\pi_h = V_h \prod_{l < h} (1 - V_l).$$

Dabei entspricht H_h der Posteriori, deren Dichte durch

$$p(\phi_h | y_i : c_i = h) \propto \left(\prod_{i: c_i = h} f(y_i | \phi_h) \right) g_0(\phi_h).$$

beschrieben ist, während c^* einer Proportionalitätskonstante entspricht, für die gilt:

$$\sum_{h=1}^N c^* f(y_i | \phi_h) \pi_h = 1.$$

Die Stick-Breaking-Konstruktion macht es möglich, dass zu jedem Iterationsschritt t auch eine „Ziehung“ von G_N vorliegt:

$$G_N^{(t)} = \sum_{h=1}^N \pi_h^{(t)} \delta_{\phi_h^{(t)}}.$$

Diese Ziehungen können zur Schätzung von G und von Funktionalen davon verwendet werden (vgl. Ishwaran & James (2001)).

5.3 Zusammenfassung und Ausblick

Die in den Abschnitten 5.1 und 5.2 erläuterten Algorithmen sollen nun hinsichtlich ihrer Gemeinsamkeiten und Unterschiede verglichen werden. Dabei werden auch Erweiterungsmöglichkeiten aufgezeigt. Darüber hinaus dient dieser Abschnitt einem Ausblick, welche alternativen Verfahren zur Schätzung von DPM-Modellen noch existieren.

Ein zentrales Thema bei den Inferenzkonzepten der marginalen Methode ist die notwendige Integration (vgl. Algorithmen 1–3) oder wie sie vermieden werden kann (vgl. Algorithmen 4 und 5). So umgeht Algorithmus 5 dieses Problem mit Hilfe von zusätzlichen Variablen und Algorithmus 4 durch die „no gaps“-Restriktion, wobei letzterer den Nachteil eines schlechteren Mischungsverhalten aufweist. Der Wert des ersten Algorithmus liegt vor allem darin, dass er als erster MCMC-Algorithmus für Dirichlet-Prozesse zur Entwicklung anderer Verfahren beigetragen hat. Er ist jedoch aufgrund seiner langsamen Konvergenz nicht

empfehlenswert. Algorithmus 2 und 3 sind diesbezüglich zu bevorzugen. Der Nachteil der notwendigen Integration wiegt oft gar nicht so schwer, da sie im Falle der Konjugiertheit oft analytisch durchführbar ist (vgl. Neal (2000)) und diese schon deshalb oft angenommen wird, um das Ziehen von Zufallszahlen für θ_i bzw. ϕ_c , wie es in den Algorithmen 1, 2, 4 und 5 nötig ist, zu erleichtern. Ähnliches gilt für Algorithmus 3: Dort können die Clusterlokationen im Falle der Konjugiertheit ohne MCMC-Verfahren bestimmt werden, so dass diese Annahme oft getroffen wird. Stört man sich dennoch an den Integralberechnungen, so besteht eine generelle Möglichkeit zu deren Vermeidung darin, den entsprechenden Gibbs-Sampling-Aufdatierungsschritt durch einen Metropolis-Hastings-Schritt zu ersetzen. Dies sei an Algorithmus 2 illustriert: Dort entsprechen die Wahrscheinlichkeiten, mit denen ein Individuum i einem vollen Cluster oder einem leeren Cluster zugeordnet wird, den Gewichten, die unter Berücksichtigung der Daten aus der Pólya-Urnen-Darstellung resultieren. Man kann nun stattdessen die Pólya-Urnen-Gewichte ohne Datenbezug als Vorschlag für c_i verwenden:

$$\begin{aligned} P(c_i^* = c | c_{-i}) &= \frac{n_{-i,c}}{n-1+\alpha_0} && \text{falls } \exists j \neq i \text{ mit } c_j = c, \\ P(c_i^* \neq c_j, \forall j \neq i | c_{-i}) &= \frac{\alpha_0}{n-1+\alpha_0}. \end{aligned}$$

Die Wahrscheinlichkeit, dass c_i^* als neuer Zustand akzeptiert wird, lautet dann:

$$\alpha(c_i^* | c_i) = \min \left\{ \frac{f(y_i | \phi_{c_i^*}) p(c_i^* | c_{-i}) p(c_{-i}) p(c_i | c_{-i})}{f(y_i | \phi_{c_i}) p(c_i | c_{-i}) p(c_{-i}) p(c_i^* | c_{-i})}, 1 \right\} = \min \left\{ \frac{f(y_i | \phi_{c_i^*})}{f(y_i | \phi_{c_i})}, 1 \right\}.$$

Die Anpassung von c_i an den Datenpunkt y_i erfolgt also zu 100% im Akzeptanzschritt und zu 0% durch den Vorschlag. Beim Gibbs-Sampler verhält es sich genau umgekehrt. Eine niedrige Akzeptanzrate und damit eine langsame Konvergenz können die Folge sein.

In ähnlicher Weise kann man vorgehen, wenn im nichtkonjugierten Fall das Aufdatieren von θ_i bzw. ϕ_c Schwierigkeiten bereitet. Auch da bietet es sich an, den Gibbs-Sampling-Schritt durch einen Metropolis-Hastings-Schritt zu ersetzen und sich bei der Vorschlagsdichte an der vollständig bedingten Dichte zu orientieren (vgl. Papaspiliopoulos & Roberts (2008)).

Abgesehen von der Gemeinsamkeit, dass im Block-Gibbs-Sampler aus einer Posteriori H_h gezogen wird, die der Posteriori H_c der marginalen Algorithmen 2, 4 und 5 entspricht, bestehen zwischen den Pólya-Urnen-Gibbs-Samplern und dem Block-Gibbs-Sampler kaum Ähnlichkeiten. Sie unterscheiden sich vielmehr in dem Punkt, dass im marginalen Ansatz durch das Rausintegrieren von G nur die $\theta_1, \dots, \theta_n$ geschätzt werden können, wohingegen im Block-Gibbs-Sampler auch Posteriori-Inferenz für G selbst möglich ist. So können die Ziehungen $G_N^{(t)}$ mit $t = 1, \dots, T$ zur Schätzung von G und von Funktionalen davon genutzt werden (vgl. Ishwaran & James (2001)). Des Weiteren kommt der Block-Gibbs-Sampler ohne jegliche Integration aus. Integralberechnungen wie in den Algorithmen 1–3 oder Strategien zu deren Vermeidung wie in Algorithmus 4 und 5 sind daher unnötig. Ein dritter Vorteil des Block-Gibbs-Samplers gegenüber den marginalen Ansätzen liegt in dem guten Mischungsverhalten, das aus dem blockweise Aufdatieren resultiert. Pólya-Urnen-Gibbs-Sampler tendieren dazu, diesbezüglich eine langsamere Konvergenz aufzuweisen (vgl. Ishwaran & James (2001) und Ishwaran & Zarepour (2000)). Aus diesen Gründen soll das

additive gemischte Modell, das in Kapitel 6 ausformuliert und in den Kapiteln 7 und 8 auf Daten angewendet wird, durch den Block-Gibbs-Sampler implementiert werden.

Die in dieser Arbeit verwendete Version des Block-Gibbs-Samplers bezieht sich stets auf die gestutzte Form gemäß Muliere & Tardella (1998). Es existieren mittlerweile auch andere Verfahren, die ohne Trunkierung auskommen. So schlagen beispielsweise Papaspiliopoulos & Roberts (2008) einen retrospektiven Ansatz vor. Dessen Idee basiert auf dem Block-Gibbs-Sampler und setzt an dem Aufdatierungsschritt für c_i von Algorithmus 6 an, bei dem entschieden wird, welchem Cluster eine Person i zugewiesen wird: Wenn allgemein aus einer endlichen Menge $\{1, \dots, N\}$ gemäß der Wahrscheinlichkeiten $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)'$ ein Element gezogen werden soll, so geht man typischerweise so vor, dass man bei gegebenem $\boldsymbol{\pi}$ eine Zufallszahl u_i aus der stetigen Gleichverteilung $U(0, 1)$ zieht und dann der Person i den Wert c zuordnet, der folgendes Kriterium erfüllt:

$$\sum_{h=0}^{c-1} \pi_h < u_i \leq \sum_{h=1}^c \pi_h. \quad (5.5)$$

Dabei gilt $\pi_0 = 0$. Analog sieht es der Algorithmus 6 vor, für gegebenes $\boldsymbol{\phi} = (\phi_1, \dots, \phi_N)'$ und $\boldsymbol{\pi}$ die Zufallszahl u_i zu ziehen, um dann zu prüfen, für welches Cluster das Kriterium (5.5) erfüllt ist. Ohne Trunkierung müsste man eigentlich unendlich dimensionale Vektoren $\boldsymbol{\phi}$ und $\boldsymbol{\pi}$ zugrunde legen. Papaspiliopoulos & Roberts (2008) vermeiden dies nun, indem sie zuerst die Zufallszahl u_i ziehen und dann retrospektiv so viele Paare (ϕ_h, π_h) konstruieren, wie es zur Erfüllung von (5.5) notwendig ist. Wenn die Idee auch recht einfach ist, so ist ihre Umsetzung keineswegs trivial (vgl. Walker (2007)).

Auch Walker (2007) liefert eine Methode, wie die Stick-Breaking-Konstruktion ohne Trunkierung verwendet werden kann. Dort wird die Endlichkeit der betreffenden Parametervektoren durch die Einführung einer latenten Variable erzeugt.

Sämtliche bisher geschilderten Algorithmen stellen MCMC-Verfahren dar. Neben dem marginalen und dem bedingten Ansatz gibt es zur Schätzung von DPM-Modellen in Form von sogenannten „reversible jump MCMC“-Verfahren noch eine dritte Form im Rahmen der MCMC-Methoden. Jain & Neal (2004) und Dahl (2005) entwickelten in diesem Kontext Algorithmen. Darüberhinaus existieren Verfahren, die nicht auf MCMC basieren, wie z.B. gewichtete „Chinese-restaurant“-Algorithmen (vgl. Ishwaran & Takahara (2002)), sequentielles „importance sampling“ (vgl. MacEachern, Clyde & Liu (1999)) und partielle, prädiktive Rekursion (vgl. Newton & Zhang (1999)).

6 Das additive gemischte Modell

6.1 Formulierung des additiven gemischten Modells

Dieser Abschnitt verfolgt das Ziel, die drei in Kapitel 3 hergeleiteten Modelle, nämlich das lineare Modell, die nonparametrische Regression mittels P-Spline und das lineare gemischte Modell mit Dirichlet-Prozess-Priori, zu einem Modell zusammenzuführen und damit eine formale Grundlage für die Analysen in Kapitel 7 und vor allem in Kapitel 8 zu schaffen. Dort gilt das Interesse in erster Linie der Untersuchung des Effekts der Zeit t auf den Response y . Zu diesem Zwecke liegen analog zu Abschnitt 3.3 longitudinale Daten vor. Man will nun einerseits den generellen Effekt der Zeit auf die Zielgröße als auch die individuellen, zeitlichen Effekte schätzen. Darüberhinaus sollen neben der Zeit noch andere Einflussgrößen x_1, \dots, x_p auf einen signifikanten Einfluss auf den Response untersucht werden. Für diese Analysen seien longitudinale Daten $(y_{i1}, \dots, y_{in_i}, \mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}, t_{i1}, \dots, t_{in_i}, \mathbf{z}_{i1}, \dots, \mathbf{z}_{in_i})$ für $i = 1, \dots, n$ gegeben. Dabei stellen t_{i1}, \dots, t_{in_i} die Messzeitpunkte und damit gleichzeitig die Ausprägungen der zeitlichen Variable t des i -ten Individuums dar. Das im folgenden betrachtete und für Kapitel 8 zugrunde liegende Modell sieht nun für die Person i zum Zeitpunkt j mit $i = 1, \dots, n$ und $j = 1, \dots, n_i$ folgende Struktur vor:

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + f(t_{ij}) + \mathbf{z}_{ij}'\mathbf{b}_i + \varepsilon_{ij}. \quad (6.1)$$

Dieses Modell wird additives gemischtes Modell genannt. Der Name stammt aus der additiven Zusammensetzung der Komponenten, zu denen auch das lineare gemischte Modell gehört. Der generelle Effekt der Zeit wird durch eine beliebige, nichtlineare Funktion f modelliert, so dass in der Bezeichnung „additives gemischtes Modell“ statt „lineares gemischtes Modell“ auch zum Ausdruck kommt, dass (mindestens) eine nichtlineare Komponente dem Modell angehört. Man spricht in diesem Zusammenhang auch von einem semiparametrischen Modell, da es sowohl parametrische als auch nonparametrische Elemente enthält. Für die Verteilung der Fehlervariable wird analog zu (3.9) folgende Annahme getroffen:

$$\varepsilon_{ij} \stackrel{i.i.d.}{\sim} N(0, \sigma^2) \quad \forall i = 1, \dots, n; \forall j = 1, \dots, n_i.$$

Der generelle, zeitliche Effekt soll durch einen P-Spline modelliert werden. Aufgrund der longitudinalen Struktur liegt bereits für jedes Individuum i eine *Designmatrix* vor. Diese hat in Analogie zu Abschnitt 3.2.1 folgende Gestalt, wobei die Basisfunktionen B_1, \dots, B_d durch die B-Spline-Basis bestimmt sind:

$$\mathbf{B}_i := \begin{pmatrix} B_1^l(t_{i1}) & \dots & B_d^l(t_{i1}) \\ \vdots & & \vdots \\ B_1^l(t_{in_i}) & \dots & B_d^l(t_{in_i}) \end{pmatrix}.$$

Bei der Modellierung der nichtlinearen Funktion f werden alle Datenpunkte als gleichwertig betrachtet, egal ob sie zu demselben oder zu verschiedenen Individuen gehören. Die longitudinale Struktur der Daten findet hierbei keine Berücksichtigung. Diese soll nun durch die zufällige Effekte erfasst werden. Darin kann z.B. für jede Person ein eigener Intercept (Random-Intercept) und eine individuelle Steigung (Random-Slope) enthalten sein. Man besitzt im Grad der zufälligen Effekte verschiedene Variationsmöglichkeiten. Die Designmatrix für die Person i lautet allgemein bei Grad l :

$$\mathbf{Z}_i := \begin{pmatrix} 1 & t_{i1} & \dots & t_{i1}^l \\ \vdots & \vdots & & \vdots \\ 1 & t_{in_i} & \dots & t_{in_i}^l \end{pmatrix}.$$

Die übrigen Kovariablen, die i.A. auch zeitvariierend sein können, werden linear ins Modell aufgenommen. Hierbei unterscheidet das Modell wie schon beim P-Spline nicht, ob die einzelnen Messungen zu derselben oder zu verschiedenen Personen gehören. Die individuelle Designmatrix \mathbf{X}_i ist durch $\mathbf{X}_i := (\mathbf{x}'_{i1}, \dots, \mathbf{x}'_{in_i})'$ bestimmt. Zusammenfassend liegt für jede Person i mit $i = 1, \dots, n$ folgendes Modell zugrunde:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{B}_i \boldsymbol{\gamma} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i.$$

Dieses lässt sich auch noch kompakter für alle Individuen zusammen darstellen. Dazu werden folgende Vektoren für den Response, die Störgrößen und die zufälligen Effekte definiert:

$$\mathbf{y} := \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} := \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix}, \quad \mathbf{b} := \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{pmatrix}.$$

Auch die individuellen Designmatrizen werden zu Matrizen, die alle Personen beschreiben, zusammengefasst:

$$\mathbf{X} := \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix}, \quad \mathbf{B} := \begin{pmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_n \end{pmatrix}, \quad \mathbf{Z} := \begin{pmatrix} \mathbf{Z}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{Z}_n \end{pmatrix}.$$

Somit lautet das additive gemischte Modell für alle Individuen:

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{B} \boldsymbol{\gamma} + \mathbf{Z} \mathbf{b} + \boldsymbol{\varepsilon}.$$

6.2 Inferenz im additiven gemischten Modell

Die Schätzung innerhalb des additiven gemischten Modells basiert auf den Inferenzmethoden der einzelnen Modellkomponenten. In Kapitel 5 wurden diverse Gibbs-Sampling-Verfahren erläutert, die bei DP(M)-Modellen zur Schätzungen der zufälligen Effekte herangezogen werden können. Analog dazu soll auch für das additive gemischte Modell ein Gibbs-Sampling-Verfahren herangezogen werden. Konkret wird in Abschnitt 6.3 der Block-Gibbs-Sampler für das additive gemischte Modell mit einem DP-Modell bzw. einem DPM-Modell für die zufälligen Effekte ausgeführt. Dieser Abschnitt konzentriert sich auf allgemeine Inferenzprinzipien und die generellen Schwierigkeiten, die bei mehreren Teilmodellen entstehen können.

Beim Gibbs-Sampling werden beim Aufdatieren eines Parameters die vollständig bedingten Dichten verwendet und damit alle anderen Parameter berücksichtigt (vgl. (2.4)). Beim Aufdatieren eines der Regressionskoeffizienten γ , β und \mathbf{b} müssen daher die Zustände der anderen Parameter durch einen Arbeitsresponse eingebunden werden. Die Reihenfolge, in der die Koeffizienten aufdatiert werden, ist dabei im Grunde beliebig – es ist nur darauf zu achten, dass von den anderen Parametern die jeweils aktuellen Zustände verwendet werden. Liegt die Reihenfolge γ , β und \mathbf{b} vor, so lauten die entsprechenden Arbeitsresponses:

$$\text{Arbeitsresponse beim Aufdatieren von } \gamma^{(t+1)}: \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\beta^{(t)} - \mathbf{Z}\mathbf{b}^{(t)},$$

$$\text{Arbeitsresponse beim Aufdatieren von } \beta^{(t+1)}: \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{B}\gamma^{(t+1)} - \mathbf{Z}\mathbf{b}^{(t)},$$

$$\text{Arbeitsresponse beim Aufdatieren von } \mathbf{b}^{(t+1)}: \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\beta^{(t+1)} - \mathbf{B}\gamma^{(t+1)}.$$

Es gibt aber noch einen weiteren Umstand, den man beim Aufdatieren der Regressionskoeffizienten beachten muss: Der zeitliche Effekt wird sowohl durch den P-Spline als auch durch die zufälligen Effekte modelliert. Diese Kombination erweist sich beim Gibbs-Sampling als durchaus diffizil. Die „Aufgabenbereiche“ des Splines und des linearen gemischten Modells überlagern sich nämlich in gewisser Weise. Werden z.B. zufällige Effekte 1. Grades verwendet, so werden innerhalb des linearen gemischten Modells sämtliche individuellen linearen Effekte – und damit implizit auch der generelle lineare Effekt – der Zeit auf den Response geschätzt. Aber auch im P-Spline, der den generellen zeitlichen Effekt modellieren soll, steckt ein gewisser linearer Anteil. Dies kann dazu führen, dass der generelle, lineare Effekt der Zeit während des Algorithmus das eine Mal eher vom Spline und das andere Mal eher vom linearen gemischten Modell erfasst wird. In der Folge treten in den Samplingpfaden von γ und \mathbf{b} Auf- und Abbewegungen bzw. regelrechte Sprünge auf. Eine Konvergenz der Markov-Ketten stellt sich nicht ein. Darauf basierende Schätzer sind somit unbrauchbar. Es ist daher notwendig eine Restriktion einzubauen, die eine saubere Trennung der Aufgabenbereiche der beiden Modellkomponenten gewährleistet. Hier sind zwei Möglichkeiten denkbar. Die eine sieht vor, die Trennung hinsichtlich des Grades des zeitlichen Effekts zu vollziehen. Wenn z.B. zufällige Effekte 1. Grades im Modell enthalten sind, so soll das lineare gemischte Modell den linearen Anteil des zeitlichen Effekts zu 100% erfassen – sowohl den generellen als auch die individuellen. Der P-Spline hingegen soll nur den generellen zeitlichen Effekt modellieren, der über den 1. Grad hinausgeht. Dies erfordert

beim Aufdatieren der P-Spline-Basiskoeffizienten eine Restriktion, die schwer umsetzbar ist. Im Gegensatz zur TP-Basis, bei der die einzelnen Basiskoeffizienten direkt mit dem entsprechenden Grad des zeitlichen Effekts in Verbindung gebracht werden kann, ist dies bei der B-Spline-Basis nur schwer ersichtlich. Die andere Restriktion ist dagegen leichter zugänglich: Dort wird die Trennung hinsichtlich des generellen und des individuellen Effekts vollzogen. Der Spline soll ganz im Sinne früherer Überlegungen den generellen Effekt des Alters beschreiben. Die zufälligen, individuellen Effekte verstehen sich lediglich als Abweichungen davon. Um dies sicher zu stellen, ist bei jeder Iteration eine Zentrierung der zufälligen Effekte erforderlich. Hierbei muss auf zwei Dinge geachtet werden: Erstens darf der abgezogene Mittelwert bei der Schätzung der Fehlervarianz nicht außer Acht gelassen werden. Zweitens ist die Reihenfolge, in der die Modellkomponenten „feste Effekte“, „zufällige Effekte“ und „P-Spline“ aufdatiert werden, nun nicht mehr beliebig. Das Aufdatieren der Basiskoeffizienten des Splines soll dem der zufälligen Effekte folgen, damit diese den generellen linearen Effekt in korrekter Weise aufnehmen können. Im DP-Modell unterscheidet sich das Vorgehen von dem im DPM-Modell geringfügig. Da dort die zufälligen Effekte in zwei Schritten aufdatiert werden, nämlich zuerst in der Bestimmung der Clusterlokationen und dann in der Entscheidung, welches Individuum zu welchem Cluster gehört, bietet es sich an, die Zentrierung gleich für die Clusterlokationen durchzuführen. Von den beiden Ansätzen wird im weiteren Verlauf der zweite angewandt werden.

Würde man die Kovariable „Zeit“ nicht nur durch einen P-Spline und im Rahmen der zufälligen Effekte modellieren, sondern auch als festen Effekt ins Modell aufnehmen, träte eine weitere Schwierigkeit bei der Trennung der Aufgabenbereiche hinzu. Dies ist aber nicht notwendig. Schließlich wird der generelle zeitliche Effekt bereits durch den P-Spline beschrieben. Deshalb soll dieser Fall im Folgenden nicht berücksichtigt werden.

6.3 Block-Gibbs-Sampler im additiven gemischten Modell

Der folgende Abschnitt setzt nun die allgemeinen Gedanken zur Inferenz aus dem vorherigen Abschnitt in die Tat um. Für das Aufdatieren der zufälligen Effekte wird ein Block-Gibbs-Sampler gemäß Abschnitt 5.2 verwendet. Auch alle anderen Parameter werden über ihre vollständig bedingten Dichten aufdatiert. Der Algorithmus wird nun sowohl für den Fall eines DP- als auch eines DPM-Modells ausgeführt. Folgende Auflistung dient hierbei als Überblick über sämtliche in den Modellen vorkommenden Anzahlen:

- n Anzahl der Individuen,
- n_i Anzahl der Messungen bei Person i ,
- n_d Anzahl aller Messungen: $n_d = \sum_{i=1}^n n_i$,
- p Anzahl der festen Effekte,
- d Anzahl der Basisfunktionen,
- q Anzahl der zufälligen Effekte.

Die Annahmen der Modelle basieren exakt auf den Ausführungen zu den einzelnen Modellkomponenten in Kapitel 3. Sie werden lediglich dahingehend noch erweitert, dass Hyper-

parameter wie $\boldsymbol{\mu}_\beta$, $\boldsymbol{\Sigma}_\beta$, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, α_0 und im Falle des DPM-Modells $\boldsymbol{\Sigma}_b$ ebenfalls modelliert werden, um das Modell flexibler zu gestalten. Die einzelnen Priori-Annahmen orientieren sich zum einen an dem Prinzip der Konjugiertheit und zum anderen an der Annahme, dass es innerhalb der Parametervektoren keine Korrelationen gibt, d.h. sämtliche Kovarianzmatrizen besitzen Diagonalgestalt. Am Ende dieses Abschnitts werden alternative Modellannahmen diskutiert.

Für den Fall, dass für die zufälligen Effekte ein DPM-Modell verwendet wird, werden nun für das entsprechende additive gemischte Modell folgende Annahmen getroffen:

Additives gemischtes Modell bei DPM-Priori (6.2)

1. *Beobachtungsmodell:*

$$\mathbf{y}_i | \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}_i, \sigma^2 \stackrel{\text{ind}}{\sim} N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{B}_i \boldsymbol{\gamma} + \mathbf{Z}_i \mathbf{b}_i, \sigma^2 \mathbf{I}_{n_i}) \quad \forall i = 1, \dots, n.$$

2. *Priori-Verteilungen:*

$$\begin{aligned} \sigma^2 &\sim IG(a_\varepsilon, b_\varepsilon), \\ \boldsymbol{\beta} | \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta &\sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \quad \text{mit } \boldsymbol{\Sigma}_\beta = \text{diag}(\sigma_{\beta_1}^2, \dots, \sigma_{\beta_p}^2), \\ \boldsymbol{\mu}_\beta &\sim N(\mathbf{m}_\beta, \mathbf{S}_\beta) \quad \text{mit } \mathbf{S}_\beta = \text{diag}(s_{\beta_1}^2, \dots, s_{\beta_p}^2), \\ \sigma_{\beta_r}^2 &\sim IG(a_\beta, b_\beta) \quad \forall r = 1, \dots, p, \\ p(\gamma_j) &\propto \text{const} \quad \forall j = 1, \dots, k, \\ \gamma_j | \gamma_{j-1}, \tau^2 &\sim N(\gamma_{j-1}, \tau^2) \quad \forall j = k+1, \dots, d, \\ \tau^2 &\sim IG(a_\gamma, b_\gamma), \\ \mathbf{b}_i | \boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b &\sim N(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b) \quad \boldsymbol{\Sigma}_b = \text{diag}(\sigma_{b_1}^2, \dots, \sigma_{b_q}^2), \\ \sigma_{b_r}^2 &\sim IG(a_b, b_b) \quad \forall r = 1, \dots, q, \\ \boldsymbol{\theta}_i | G &\stackrel{i.i.d.}{\sim} G \quad \forall i = 1, \dots, n, \\ G &\sim DP(\alpha_0 G_0), \\ G_0 &= N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad \text{mit } \boldsymbol{\Sigma}_0 = \text{diag}(\sigma_{0_1}^2, \dots, \sigma_{0_q}^2), \\ \boldsymbol{\mu}_0 &\sim N(\mathbf{m}_0, \mathbf{S}_0) \quad \text{mit } \mathbf{S}_0 = \text{diag}(s_{0_1}^2, \dots, s_{0_q}^2), \\ \sigma_{0_r}^2 &\sim IG(a_0, b_0) \quad \forall r = 1, \dots, q, \\ \alpha_0 &\sim Ga(a_\alpha, b_\alpha). \end{aligned}$$

Auf diesen Annahmen basierend soll nun ein Gibbs-Sampler dargelegt werden. Die Herleitung sämtlicher vollständig bedingten Dichten befindet sich im Anhang B. Der Block-Gibbs-Sampler für das additive gemischte Modell mit DPM-Modell lautet nun:

Block-Gibbs-Algorithmus bei DPM-Priori:

Die Markov-Kette befinde sich im Zustand $\gamma, \tau^2, \beta, \mu_\beta, \Sigma_\beta, \mathbf{b}, \Sigma_b, \phi, \mathbf{c}, \pi, \alpha_0, \mu_0, \Sigma_0$ und σ^2 .

(I) Aufdatieren der zum P-Spline gehörigen Parameter:

- ▷ Bilde Arbeitsresponse $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}$.
- ▷ Ziehe neuen Wert für γ gemäß:

$$\begin{aligned} \gamma | \tau^2, \beta, \mathbf{b}, \mathbf{y}, \sigma^2 &\sim N(\mu_\gamma^*, \Sigma_\gamma^*), \\ \mu_\gamma^* &= \left(\frac{1}{\tau^2} \mathbf{K} + \frac{1}{\sigma^2} \mathbf{B}' \mathbf{B} \right)^{-1} \frac{1}{\sigma^2} \mathbf{B}' \tilde{\mathbf{y}}, \\ \Sigma_\gamma^* &= \left(\frac{1}{\tau^2} \mathbf{K} + \frac{1}{\sigma^2} \mathbf{B}' \mathbf{B} \right)^{-1}. \end{aligned}$$

- ▷ Ziehe neuen Wert für τ^2 gemäß:

$$\tau^2 | \gamma \sim IG(a_\gamma + 0.5 \, \text{rg}(\mathbf{K}), b_\gamma + 0.5 \, \gamma' \mathbf{K} \gamma).$$

(II) Aufdatieren der zu den festen Effekten gehörigen Parameter:

- ▷ Bilde Arbeitsresponse $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{B}\gamma - \mathbf{Z}\mathbf{b}$.
- ▷ Ziehe neuen Wert für β gemäß:

$$\begin{aligned} \beta | \mu_\beta, \Sigma_\beta, \gamma, \mathbf{b}, \mathbf{y}, \sigma^2 &\sim N(\mu_\beta^*, \Sigma_\beta^*), \\ \mu_\beta^* &= \left(\Sigma_\beta^{-1} + \frac{1}{\sigma^2} \mathbf{X}' \mathbf{X} \right)^{-1} \left(\Sigma_\beta^{-1} \mu_\beta + \frac{1}{\sigma^2} \mathbf{X}' \tilde{\mathbf{y}} \right), \\ \Sigma_\beta^* &= \left(\Sigma_\beta^{-1} + \frac{1}{\sigma^2} \mathbf{X}' \mathbf{X} \right)^{-1}. \end{aligned}$$

- ▷ Für $r = 1, \dots, p$:

- Ziehe neuen Wert für μ_{β_r} gemäß:

$$\mu_{\beta_r} | \sigma_{\beta_r}^2, \beta_r \sim N \left(\left(\frac{1}{\sigma_{\beta_r}^2} + \frac{1}{s_{\beta_r}^2} \right)^{-1} \left(\frac{\beta_r}{\sigma_{\beta_r}^2} + \frac{m_{\beta_r}}{s_{\beta_r}^2} \right), \left(\frac{1}{\sigma_{\beta_r}^2} + \frac{1}{s_{\beta_r}^2} \right)^{-1} \right).$$

- Ziehe neuen Wert für $\sigma_{\beta_r}^2$ gemäß:

$$\sigma_{\beta_r}^2 | \mu_{\beta_r}, \beta_r \sim IG(a_\beta + 0.5, b_\beta + 0.5(\beta_r - \mu_{\beta_r})^2).$$

(III) Aufdatieren der zu den zufälligen Effekten gehörigen Parameter:

- ▷ Bilde Arbeitsresponse $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\beta - \mathbf{B}\gamma$.
- ▷ Für $i = 1, \dots, n$:

$$\begin{aligned} \mathbf{b}_i | \theta_i, \Sigma_b, \beta, \gamma, \mathbf{y}_i, \sigma^2 &\sim N(\mu_b^*, \Sigma_b^*), \\ \mu_b^* &= \left(\Sigma_b^{-1} + \frac{1}{\sigma^2} \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} \left(\Sigma_b^{-1} \theta_i + \frac{1}{\sigma^2} \mathbf{Z}_i' \tilde{\mathbf{y}}_i \right), \\ \Sigma_b^* &= \left(\Sigma_b^{-1} + \frac{1}{\sigma^2} \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1}. \end{aligned}$$

▷ Korrekturschritt (falls P-Spline im Modell enthalten):

- Bestimme Mittelwert $\bar{\mathbf{b}}$.
- Für $i = 1, \dots, n$:
 - Ersetze \mathbf{b}_i durch $\mathbf{b}_i - \bar{\mathbf{b}}$.

▷ Für $h = 1, \dots, N$:

- Ziehe neuen Wert für ϕ_h gemäß:
 - Falls $\nexists i : c_i = h$:

$$\phi_h | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

- Falls $\exists i : c_i = h$:

◇ Für $r = 1, \dots, q$:

$$\begin{aligned} \phi_{h_r} | \sigma_{b_r}^2, \mu_{0_r}, \sigma_{0_r}^2, \mathbf{b}, \mathbf{c} &\sim N(\mu_{0_r}^*, \sigma_{0_r}^{2*}), \\ \mu_{0_r}^* &= \left(\frac{n_h}{\sigma_{b_r}^2} + \frac{1}{\sigma_{0_r}^2} \right)^{-1} \left(\frac{n_h}{\sigma_{b_r}^2} \bar{b}_{r,h} + \frac{\mu_{0_r}}{\sigma_{0_r}^2} \right), \\ \sigma_{0_r}^{2*} &= \left(\frac{n_h}{\sigma_{b_r}^2} + \frac{1}{\sigma_{0_r}^2} \right)^{-1}. \end{aligned}$$

▷ Für $i = 1, \dots, n$:

- Ziehe neuen Wert für c_i gemäß:

$$c_i | \boldsymbol{\pi}, \boldsymbol{\phi}, \mathbf{b}_i, \boldsymbol{\Sigma}_b \sim \sum_{h=1}^N c^* f(\mathbf{b}_i | \boldsymbol{\phi}_h, \boldsymbol{\Sigma}_b) \pi_h \delta_h,$$

$f \triangleq$ Dichte der multivariaten Normalverteilung,

$c^* \triangleq$ Konstante, so dass Summe der Wahrscheinlichkeiten 1 ergibt.

- Setze $\boldsymbol{\theta}_i = \boldsymbol{\phi}_{c_i}$.

▷ Für $h = 1, \dots, N$:

- Ziehe neuen Wert für V_h (außer für $h = N$, da stets $V_N = 1$) gemäß:

$$V_h | \mathbf{c} \sim \text{Be}(1 + n_h, \alpha_0 + \sum_{l=h+1}^N n_l).$$

- Konstruiere π_h gemäß:

$$\pi_h = V_h \prod_{l < h} (1 - V_l).$$

▷ Ziehe neuen Wert für α_0 gemäß:

$$\alpha_0 | \boldsymbol{\pi} \sim \text{Ga}(N - 1 + a_\alpha, b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h)).$$

▷ Für $r = 1, \dots, p$:

- Ziehe neuen Wert für μ_{0_r} gemäß:

$$\mu_{0_r} | \sigma_{0_r}^2, \boldsymbol{\theta} \sim N \left(\left(\frac{n}{\sigma_{0_r}^2} + \frac{1}{s_{0_r}^2} \right)^{-1} \left(\frac{n}{\sigma_{0_r}^2} \bar{\theta}_r + \frac{m_{0_r}}{s_{0_r}^2} \right), \left(\frac{n}{\sigma_{0_r}^2} + \frac{1}{s_{0_r}^2} \right)^{-1} \right).$$

- Ziehe neuen Wert für $\sigma_{0_r}^2$ gemäß:

$$\sigma_{0_r}^2 | \mu_{0_r}, \boldsymbol{\theta} \sim \text{IG} \left(a_0 + 0.5 n, b_0 + 0.5 \sum_{i=1}^n (\theta_{i_r} - \mu_{0_r})^2 \right).$$

- Ziehe neuen Wert für $\sigma_{b_r}^2$ gemäß:

$$\sigma_{b_r}^2 | \boldsymbol{\theta}, \mathbf{b} \sim IG(a_b + 0.5n, b_b + 0.5 \sum_{i=1}^n (b_{i_r} - \theta_{i_r})^2).$$

(IV) Aufdatieren der Fehlervarianz:

$$\sigma^2 | \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}, \mathbf{y} \sim IG(a_\varepsilon + 0.5n_d, b_\varepsilon + 0.5 \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - \mu_{ij})^2),$$

$$\mu_{ij} = (\mathbf{X}_i \boldsymbol{\beta} + \mathbf{B}_i \boldsymbol{\gamma} + \mathbf{Z}_i \bar{\mathbf{b}} + \mathbf{Z}_i \mathbf{b}_i)_j.$$

Analog zu 6.2 werden für das additive gemischte Modell, dessen zufällige Effekte durch eine DP-Priori modelliert werden, folgende Annahmen getroffen:

Additives gemischtes Modell bei DP-Priori		(6.3)
1. Beobachtungsmodell:		
$\mathbf{y}_i \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}_i, \sigma^2 \stackrel{ind}{\sim} N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{B}_i \boldsymbol{\gamma} + \mathbf{Z}_i \mathbf{b}_i, \sigma^2 \mathbf{I}_{n_i}) \quad \forall i = 1, \dots, n.$		
2. Priori-Verteilungen:		
σ^2	\sim	$IG(a_\varepsilon, b_\varepsilon),$
$\boldsymbol{\beta} \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta$	\sim	$N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \quad \text{mit } \boldsymbol{\Sigma}_\beta = \text{diag}(\sigma_{\beta_1}^2, \dots, \sigma_{\beta_p}^2),$
$\boldsymbol{\mu}_\beta$	\sim	$N(\mathbf{m}_\beta, \mathbf{S}_\beta) \quad \text{mit } \mathbf{S}_\beta = \text{diag}(s_{\beta_1}^2, \dots, s_{\beta_p}^2),$
$\sigma_{\beta_r}^2$	\sim	$IG(a_\beta, b_\beta) \quad \forall r = 1, \dots, p,$
$p(\gamma_j)$	\propto	$const \quad \forall j = 1, \dots, k,$
$\gamma_j \gamma_{j-1}, \tau^2$	\sim	$N(\gamma_{j-1}, \tau^2) \quad \forall j = k+1, \dots, d,$
τ^2	\sim	$IG(a_\gamma, b_\gamma),$
$\mathbf{b}_i G$	$\stackrel{i.i.d.}{\sim}$	$G \quad \forall i = 1, \dots, n,$
G	\sim	$DP(\alpha_0 G_0),$
G_0	$=$	$N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad \text{mit } \boldsymbol{\Sigma}_0 = \text{diag}(\sigma_{0_1}^2, \dots, \sigma_{0_q}^2),$
$\boldsymbol{\mu}_0$	\sim	$N(\mathbf{m}_0, \mathbf{S}_0) \quad \text{mit } \mathbf{S}_0 = \text{diag}(s_{0_1}^2, \dots, s_{0_q}^2),$
$\sigma_{0_r}^2$	\sim	$IG(a_0, b_0) \quad \forall r = 1, \dots, q,$
α_0	\sim	$Ga(a_\alpha, b_\alpha).$

Entsprechend lautet der Block-Gibbs-Sampler, dessen vollständig bedingte Dichten wiederum in Anhang B hergeleitet werden, für das additive gemischte Modell bei DP-Priori:

Block-Gibbs-Algorithmus bei DP-Priori:

Die Markov-Kette befinde sich im Zustand $\gamma, \tau^2, \beta, \mu_\beta, \Sigma_\beta, \phi, \mathbf{c}, \pi, \alpha_0, \mu_0, \Sigma_0$ und σ^2 .

(I) Aufdatieren der zum P-Spline gehörigen Parameter:

- ▷ Bilde Arbeitsresponse $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}$.
- ▷ Ziehe neuen Wert für γ gemäß:

$$\begin{aligned} \gamma | \tau^2, \beta, \mathbf{b}, \mathbf{y}, \sigma^2 &\sim N(\mu_\gamma^*, \Sigma_\gamma^*), \\ \mu_\gamma^* &= \left(\frac{1}{\tau^2} \mathbf{K} + \frac{1}{\sigma^2} \mathbf{B}' \mathbf{B} \right)^{-1} \frac{1}{\sigma^2} \mathbf{B}' \tilde{\mathbf{y}}, \\ \Sigma_\gamma^* &= \left(\frac{1}{\tau^2} \mathbf{K} + \frac{1}{\sigma^2} \mathbf{B}' \mathbf{B} \right)^{-1}. \end{aligned}$$

- ▷ Ziehe neuen Wert für τ^2 gemäß:

$$\tau^2 | \gamma \sim IG(a_\gamma + 0.5 \operatorname{rg}(\mathbf{K}), b_\gamma + 0.5 \gamma' \mathbf{K} \gamma).$$

(II) Aufdatieren der zu den festen Effekten gehörigen Parameter:

- ▷ Bilde Arbeitsresponse $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{B}\gamma - \mathbf{Z}\mathbf{b}$.
- ▷ Ziehe neuen Wert für β gemäß:

$$\begin{aligned} \beta | \mu_\beta, \Sigma_\beta, \gamma, \mathbf{b}, \mathbf{y}, \sigma^2 &\sim N(\mu_\beta^*, \Sigma_\beta^*), \\ \mu_\beta^* &= \left(\Sigma_\beta^{-1} + \frac{1}{\sigma^2} \mathbf{X}' \mathbf{X} \right)^{-1} \left(\Sigma_\beta^{-1} \mu_\beta + \frac{1}{\sigma^2} \mathbf{X}' \tilde{\mathbf{y}} \right), \\ \Sigma_\beta^* &= \left(\Sigma_\beta^{-1} + \frac{1}{\sigma^2} \mathbf{X}' \mathbf{X} \right)^{-1}. \end{aligned}$$

- ▷ Für $r = 1, \dots, p$:

- Ziehe neuen Wert für μ_{β_r} gemäß:

$$\mu_{\beta_r} | \sigma_{\beta_r}^2, \beta_r \sim N \left(\left(\frac{1}{\sigma_{\beta_r}^2} + \frac{1}{s_{\beta_r}^2} \right)^{-1} \left(\frac{\beta_r}{\sigma_{\beta_r}^2} + \frac{m_{\beta_r}}{s_{\beta_r}^2} \right), \left(\frac{1}{\sigma_{\beta_r}^2} + \frac{1}{s_{\beta_r}^2} \right)^{-1} \right).$$

- Ziehe neuen Wert für $\sigma_{\beta_r}^2$ gemäß:

$$\sigma_{\beta_r}^2 | \mu_{\beta_r}, \beta_r \sim IG(a_\beta + 0.5, b_\beta + 0.5(\beta_r - \mu_{\beta_r})^2).$$

(III) Aufdatieren der zu den zufälligen Effekten gehörigen Parameter:

- ▷ Bilde Arbeitsresponse $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\beta - \mathbf{B}\gamma$.
- ▷ Für $h = 1, \dots, N$:
 - Ziehe neuen Wert für ϕ_h gemäß:
 - Falls $\nexists i : c_i = h$:

$$\phi_h | \mu_0, \Sigma_0 \sim N(\mu_0, \Sigma_0).$$

- Falls $\exists i : c_i = h$:

$$\phi_h | \mu_0, \Sigma_0, \beta, \gamma, \mathbf{y}, \sigma^2 \sim N(\mu_0^*, \Sigma_0^*),$$

$$\begin{aligned} \mu_0^* &= (\Sigma_0^{-1} + \frac{1}{\sigma^2} \sum_{i:c_i=h} \mathbf{Z}_i' \mathbf{Z}_i)^{-1} (\Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} \sum_{i:c_i=h} \mathbf{Z}_i' \tilde{\mathbf{y}}_i), \\ \Sigma_0^* &= (\Sigma_0^{-1} + \frac{1}{\sigma^2} \sum_{i:c_i=h} \mathbf{Z}_i' \mathbf{Z}_i)^{-1}. \end{aligned}$$

- ▷ Korrekturschritt (falls P-Spline im Modell enthalten):

- Bestimme Mittelwert $\bar{\phi}$.
- Für $h = 1, \dots, N$:
 - Ersetze ϕ_h durch $\phi_h - \bar{\phi}$.

- ▷ Für $i = 1, \dots, n$:

- Ziehe neuen Wert für c_i gemäß:

$$c_i | \pi, \beta, \gamma, \phi, \mathbf{y}_i, \sigma^2 \sim \sum_{h=1}^N c^* f(\mathbf{y}_i | \mathbf{X}_i \beta + \mathbf{B}_i \gamma + \mathbf{Z}_i \phi_h, \sigma^2 \mathbf{I}_{n_i}) \pi_h \delta_h,$$

$f \triangleq$ Dichte der multivariaten Normalverteilung,

$c^* \triangleq$ Konstante, so dass Summe der Wahrscheinlichkeiten 1 ergibt.

- Setze $\mathbf{b}_i = \phi_{c_i}$.

- ▷ Für $h = 1, \dots, N$:

- Ziehe neuen Wert für V_h (außer für $h = N$, da stets $V_N = 1$) gemäß:

$$V_h | \mathbf{c} \sim Be(1 + n_h, \alpha_0 + \sum_{l=h+1}^N n_l).$$

- Konstruiere π_h gemäß:

$$\pi_h = V_h \prod_{l < h} (1 - V_l).$$

- ▷ Ziehe neuen Wert für α_0 gemäß:

$$\alpha_0 | \pi \sim Ga(N - 1 + a_\alpha, b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h)).$$

- ▷ Für $r = 1, \dots, p$:

- Ziehe neuen Wert für μ_{0_r} gemäß:

$$\mu_{0_r} | \sigma_{0_r}^2, \mathbf{b} \sim N \left(\left(\frac{n}{\sigma_{0_r}^2} + \frac{1}{s_{0_r}^2} \right)^{-1} \left(\frac{n}{\sigma_{0_r}^2} \bar{b}_r + \frac{m_{0_r}}{s_{0_r}^2} \right), \left(\frac{n}{\sigma_{0_r}^2} + \frac{1}{s_{0_r}^2} \right)^{-1} \right).$$

- Ziehe neuen Wert für $\sigma_{0_r}^2$ gemäß:

$$\sigma_{0_r}^2 | \mu_{0_r}, \mathbf{b} \sim IG(a_0 + 0.5n, b_0 + 0.5 \sum_{i=1}^n (b_{i_r} - \mu_{0_r})^2).$$

(IV) Aufdatieren der Fehlervarianz:

$$\sigma^2 | \beta, \gamma, \mathbf{b}, \mathbf{y} \sim IG \left(a_\varepsilon + 0.5n_d, b_\varepsilon + 0.5 \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - \mu_{ij})^2 \right),$$

$$\mu_{ij} = (\mathbf{X}_i \beta + \mathbf{B}_i \gamma + \mathbf{Z}_i \bar{\phi} + \mathbf{Z}_i \mathbf{b}_i)_j.$$

Zu den in den Modellen (6.2) und (6.3) getroffenen Annahmen gibt es freilich Alternativen. So könnte man für die Kovarianzmatrizen Σ_β und Σ_0 auch Matrizen verwenden, die keine Diagonalgestalt aufweisen. Als Priori-Verteilungen für die Kovarianzmatrizen bieten sich dann inverse Wishartverteilungen an. Bei der Priori-Verteilung für den Präzisionsparameter ist zu beachten, dass eine Gammaverteilung in Kombination mit dem Aufdatierungsschritt für π , genauer gesagt für V_1, \dots, V_{N-1} , durchaus problematisch ist. Dieser sei nochmal für $h = 1, \dots, N - 1$ aufgeführt:

$$V_h | \mathbf{c} \sim Be \left(1 + n_h, \alpha_0 + \sum_{l=h+1}^N n_l \right).$$

Die vollständig bedingte Dichte für α_0 lautet im Falle einer Gammaverteilung als Priori:

$$\alpha_0 | \pi \sim Ga \left(N - 1 + a_\alpha, b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h) \right).$$

Niedrige Werte von α_0 nahe bei 0 resultieren in V_h -Werten nahe bei 1. Diese führen wiederum zu sehr kleinen α_0 -Werten. Ist z.B. $\alpha_0 = 0.25$ und wird V_h für einen leeren Cluster aufdatiert, dem nur noch leere Cluster folgen, so lautet die Aufdatierungsdichte für V_h : $Be(1, 0.25)$. Hierbei kann es passieren, dass für ein V_h exakt 1 gezogen werden. Dies führt im nächsten Iterationsschritt wegen $\alpha_0 | \pi \sim Ga(N - 1 + a_\alpha, \infty)$ zu $\alpha_0 = 0$. Falls der letzte Cluster $h = N$ leer ist, ist für mindestens ein $h = 1, \dots, N - 1$ die vollständig bedingte Dichte durch $Be(1 + n_h, 0)$ bestimmt und damit die allgemeine Annahme $b > 0$ bei der Beta-Verteilung $Be(a, b)$ verletzt.

Dies lässt sich vermeiden, wenn niedrige α_0 -Werte durch eine diskrete Verteilung per se ausgeschlossen werden. So kann für α_0 beispielsweise ein Gitter $\Omega = \{0.5, 0.6, \dots, 100\}$ verwendet werden. Die Priori-Verteilung gestaltet sich dann wie folgt:

$$\alpha_0 \sim \sum_{\omega \in \Omega} P(\alpha_0 = \omega) \delta_\omega \quad \Omega = \{0.5, 0.6, \dots, 100\}.$$

Damit resultiert folgende vollständig bedingte Dichte:

$$\alpha_0 | \pi \sim \sum_{\omega \in \Omega} \exp((N - 1) \log(\omega) + (\omega - 1) \sum_{h=1}^{N-1} \log(1 - V_h)) P(\alpha_0 = \omega) \delta_\omega.$$

Nichtsdestotrotz soll im weiteren Verlauf die Gammaverteilung als Priori für den Präzisionsparameter verwendet werden. Möglicherweise auftretende fehlende Werte innerhalb der Ziehungen werden in Kauf genommen.

6.4 Implementierung

Die additiven gemischten Modelle mit DP- bzw. DPM-Priori, wie sie in Abschnitt 6.3 aufgeführt sind, werden Grundlage der Analysen in Kapitel 7 und 8 sein. Aus diesem Grund wurden die entsprechenden Block-Gibbs-Algorithmen im Rahmen dieser Arbeit eigens implementiert. Hierfür wurde sowohl das Programm R als auch C++ verwendet. Die Struktur der Implementierung ist dabei so konzipiert, dass die Modelle von R aus durch die Funktionen `blockDP()` bzw. `blockDPM()` aufgerufen werden, welche intern auf die C++-Funktionen `R_DPmodel()` bzw. `R_DPMmodel()` zugreifen, in denen dann die vom jeweiligen Algorithmus vorgesehenen Berechnungen durchgeführt werden. Grundsätzlich wurde beim Schreiben der Programme auf eine ausführliche Kommentierung und eine möglichst allgemeine Implementierung geachtet. Die Modelle lassen sich daher leicht auf beliebige Datensätze bei einer großen Vielfalt an Modifikationsmöglichkeiten anwenden.

Im Folgenden soll die Struktur der Implementierung näher beleuchtet werden: Durch die R-Funktionen `blockDP()` bzw. `blockDPM()` werden die jeweiligen Modellinformationen eingelesen und verarbeitet. Die Bestandteile des zu berechnenden Modells können dabei beliebig zusammengestellt werden. So kann einerseits wie in Abschnitt 6.3 vorgesehen ein additives gemischtes Modell bestehend aus einem linearen Modell, einem P-Spline und einem linearen gemischten Modell berechnet werden und andererseits auch nur eine oder zwei dieser Modellkomponenten gewählt werden. Des Weiteren werden in den R-Funktionen diverse Parameter initialisiert, die dann an die C++-Funktionen `R_DPmodel()` bzw. `R_DPMmodel()` übergeben werden. Dort werden die in Abschnitt 6.3 beschriebenen Algorithmen jeweils 1:1 umgesetzt. Hierfür ist das Rechnen mit Matrizen notwendig, das durch die Benutzung der `Newmat`-Bibliothek gewährleistet wurde. Die C++-Funktionen `R_DPmodel()` und `R_DPMmodel()` sind in der Datei `blockDP.cpp` bzw. `blockDPM.cpp` enthalten. Darin befinden sich auch diverse Hilfsfunktionen, die zum einen die Transformation von Modellgrößen von R nach C++ und umgekehrt ermöglichen und die zum anderen während des Algorithmus benötigt werden. Auf die Funktionen der Dateien `blockDP.cpp` bzw. `blockDPM.cpp` kann von R aus durch das Laden der Dateien `blockDP.dll` bzw. `blockDPM.dll` zugegriffen werden. Nach Beendigung des Algorithmus werden schließlich die Informationen über sämtliche gezogenen Werte in Matrixform ins R zurückgegeben. In der R-Funktion `BlockDP()` bzw. `BlockDPM()` werden dann lediglich die gezogenen Werte mit dem zugehörigen Variablennamen versehen. Die Ausgabe der Funktionen stellt eine Liste dar, die zum einen die beschriftete Matrix der Ziehungen und zum anderen eine Zusammenstellung diverser Modellinformationen enthält.

Zuvor geschriebene R-Funktionen erwiesen sich wegen zahlreicher verschachtelter Schleifen für praktische Zwecke als zu langsam. Durch die Umlagerung der Algorithmen in C++ konnte eine enorme Verkürzung der Rechenzeit erzielt werden: Die Rechengeschwindigkeit erhöhte sich in etwa um den Faktor 28. Dennoch sind zur Berechnung von Modellen, die auf Datensätzen mit vielen Individuen beruhen, einige Stunden notwendig.

Der Diplomarbeit ist eine CD beigelegt, die die zur Berechnung der Block-Gibbs-Algorithmen notwendigen Dateien enthält. Im Folgenden werde ein Überblick über diese Dateien gegeben:

- `blockDP.cpp` (beinhaltet u.a. die C++-Funktion `R_DPmodel()`),
- `blockDPM.cpp` (beinhaltet u.a. die C++-Funktion `R_DPMmodel()`),
- `blockDP.dll` (kompilierte Fassung von `blockDP.cpp`),
- `blockDPM.dll` (kompilierte Fassung von `blockDPM.cpp`),
- `BlockDP.r` (beinhaltet die R-Funktion `blockDP()`),
- `BlockDPM.r` (beinhaltet die R-Funktion `blockDPM()`),
- `CallingBlockDP.r` (enthält exemplarische Aufrufe von `blockDP()` sowie Auswertungsmöglichkeiten der Modelle),
- `CallingBlockDPM.r` (enthält exemplarische Aufrufe von `blockDPM()` sowie Auswertungsmöglichkeiten der Modelle),
- `Hilfsfunktionen.r` (enthält Funktionen, die zum Auswerten eines Modells benötigt werden),
- `SimMixed3.r` (beschreibt die Simulation in Kapitel 7),
- `Latexplots.r` (enthält Funktionen, die zum Erstellen von Graphiken in dieser Arbeit benötigt wurden).

7 Analyse simulierter Daten

Im Folgenden soll untersucht werden, inwieweit die Verwendung eines DP-Modells bzw. eines DPM-Modells für die Verteilung der zufälligen Effekte die Schätzergebnisse bezüglich der b_i im Vergleich zu einer traditionellen Normalverteilungsannahme verbessern kann. Das Kapitel greift damit die Diskussion in Abschnitt 3.3 wieder auf und will anhand simulierter Daten folgende These überprüfen: Im Falle nichtnormalverteilter zufälliger Effekte führt die Verwendung eines DPM-Modells zu besseren Schätzungen für die zufälligen Effekte. Als Vergleichskriterium dient bei n Individuen der geschätzte Mean Square Error (MSE):

$$\widehat{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n (b_i - \hat{b}_i)^2.$$

Für die Simulation werde der Einfachheit halber von einem linearen gemischten Modell mit folgender Struktur ausgegangen: Für den Einfluss der Kovariable t werde ein linearer Zusammenhang auf den Response y angenommen. Zudem erlaubt ein Random-Intercept individuelle Verschiebungen hinsichtlich des Niveaus der Regressionsgeraden. Das Modell lautet demnach für die Person $i = 1, \dots, n$ bei der Messung $j = 1, \dots, n_i$:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i^* + \varepsilon_{ij}. \quad (7.1)$$

Hierbei bezeichnet b_i^* den zentrierten zufälligen Effekt, während der gesamte zufällige Effekt durch $b_i = \beta_0 + b_i^*$ beschrieben ist. Für die Fehlervariable wird die übliche Annahme $\varepsilon_{ij} \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ für alle $i = 1, \dots, n$ und $j = 1, \dots, n_i$ getroffen. Als Besonderheit der Simulation wird nun für b_i^* statt einer Normalverteilung eine Mischverteilung bestehend aus drei Normalverteilungen angenommen:

$$b_i^* \stackrel{i.i.d.}{\sim} w_1 N(\mu_1, \sigma_b^2) + w_2 N(\mu_2, \sigma_b^2) + w_3 N(\mu_3, \sigma_b^2) \quad \forall i = 1, \dots, n. \quad (7.2)$$

Die Gewichte w_1 , w_2 und w_3 addieren sich hierbei zu 1. Zudem sollen die Erwartungswerte μ_1 , μ_2 und μ_3 so aufeinander abgestimmt sein, dass für die Verteilung der b_i^* ein Erwartungswert von Null vorliegt. In den folgenden beiden Abschnitten werden zwei Simulationen diskutiert, deren Einstellungen sich nur dahingehend unterscheiden, dass μ_1 , μ_2 und μ_3 einmal relativ weit entfernt voneinander liegen und das andere Mal nur geringe Abstände aufweisen.

7.1 Mischverteilte Daten mit großen Unterschieden

Zunächst wird eine Datenstruktur simuliert, die auf stark voneinander abweichenden Erwartungswerten der drei Normalverteilungen innerhalb der Mischverteilung (7.2) beruht. Konkret werden für die Simulation die Werte $\mu_1 = 4.5$, $\mu_2 = -1.5$ und $\mu_3 = -4.5$ gewählt. Sämtliche Einstellungen der Simulation können der Tabelle 7.1 entnommen werden. Die Messzeitpunkte sind dabei für alle Individuen gleich.

Daten	n	n_i	t_{i1}	t_{i2}	t_{i3}	t_{i4}	t_{i5}
	20	5	1	2	3	4	5
Fehlervarianz	σ^2						
	0.5						
feste Effekte	β_0	β_1					
	8	3					
zufälliger Effekt	σ_b^2	μ_1	μ_2	μ_3	w_1	w_2	w_3
	1	4.5	-1.5	-4.5	0.4	0.3	0.3

Tabelle 7.1: Simulationseinstellungen für große Clusterunterschiede

Für die 20 Individuen ergibt sich damit beispielsweise ein Verlauf, wie er in Abbildung 7.1 ersichtlich ist.

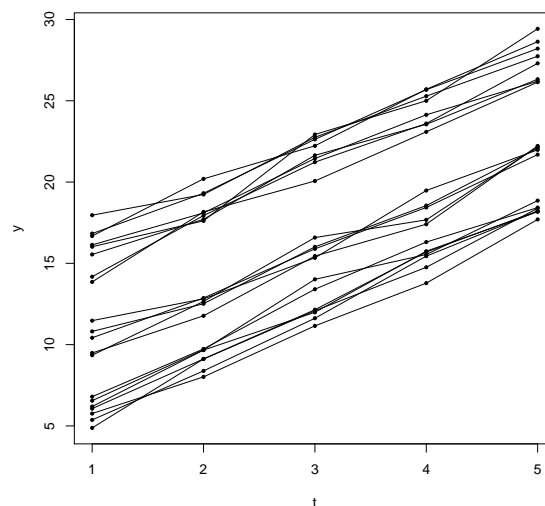


Abbildung 7.1: Simuliertes Datenbeispiel für das Modell (7.1) gemäß (7.2) und den Einstellungen in Tabelle 7.1

In der Simulationsstudie sollen nun vier verschiedene Verfahren hinsichtlich des geschätzten MSEs verglichen werden: zum einen die Maximum-Likelihood(ML)- und die restringierte Maximum-Likelihood(REML)-Schätzung, die beide normalverteilte zufällige Effekte unterstellen, und zum anderen das DP- und das DPM-Modell. Für letztere werden zudem drei unterschiedliche Einstellungen für den Präzisionsparameter α_0 verwendet: Einmal wird α_0 konstant gleich 1 gesetzt und die anderen beiden Male wird eine Gamma-Verteilung, nämlich $\alpha_0 \sim Ga(2, 2)$ bzw. $\alpha_0 \sim Ga(2, 4)$, als a priori gewählt. Diese sind durch Escobar & West (1995) bzw. Ishwaran & James (2002) motiviert. Tabelle 7.2 stellt die drei verschiedenen Modellannahmen zusammen. Werden in nachfolgenden Tabellen und Graphiken Modelle mit Indizes versehen, so beziehen sie sich hierauf.

Modell 1	Modell 2	Modell 3
$\alpha_0 = 1$	$\alpha_0 \sim Ga(2, 2)$	$\alpha_0 \sim Ga(2, 4)$

Tabelle 7.2: Die drei verschiedenen Modellannahmen für den Präzisionsparameter

Die (RE)ML-Schätzungen wurden mit Hilfe der Funktion `lme()` aus dem R-Paket `nlme` bestimmt. Für Erläuterungen zur genauen Modellspezifikation und Berechnungsweise sei auf Laird & Ware (1982) bzw. Lindstrom & Bates (1988) verwiesen.

Das DP- und DPM-Modell werden gemäß Abschnitt 6.3 formuliert und durch die dort angegebenen Algorithmen im Rahmen eines additiven gemischten Modells geschätzt. Das blockweise Aufdatieren in diesem Gibbs-Sampler macht es nämlich möglich, dass Modell-Komponenten wie hier der P-Spline ohne Probleme weggelassen werden können. Das additive gemischte Modell kann daher auf ein lineares gemischtes Modell reduziert werden. Es muss allerdings beachtet werden, dass die Algorithmen keine interne Trennung der festen und zufälligen Effekte vorsehen. Dies war im additiv gemischten Modell nicht nötig, da dort die zeitliche Variable stets nichtlinear modelliert wurde. Für die Trennung des nichtlinearen und des zufälligen Effekts wurde dann eine Korrektur angewendet (vgl. Abschnitt 6.2). Nun soll die Trennung der festen und zufälligen Effekte dahingehend erfolgen, dass die Steigung als fester und der Intercept ausschließlich als zufälliger Effekt betrachtet wird. Die Analyse der zufälligen Effekte stützt sich daher auf den „gesamten“ Intercept b_i , also auf den generellen Intercept plus die individuelle Abweichung davon. Die Berechnungen hierzu wurden mit eigens geschriebenen R- bzw. C++-Funktionen durchgeführt (vgl. Abschnitt 6.4).

Die Tabelle 7.3 enthält die Informationen über sämtliche Modellannahmen. W beschreibt die Ausdünnung bei der Markovkette und gibt konkret darüber Aufschluss, jedes wievielte Glied der Kette in die Analyse eingeht. Die Bedeutung der übrigen Kürzel kann den Modellgleichungen in 6.3 entnommen werden. Im Falle eines DP-Modells sind die Hyperparameter a_b und b_b ohne Bedeutung. In den Hyperparameter m_0 soll bereits das Vorwissen eingehen, dass der zufällige Effekt b_i den globalen Intercept beinhaltet und nicht um ihn zentriert ist. Ein vorab gerechnetes lineares Modell führte hier zur Priori-Annahme $m_0 = 8$.

MCMC	<i>Iterationen</i>	<i>Burnin</i>	<i>W</i>				
	33000	3000	30				
Fehlervarianz	a_ε	b_ε					
	0.0001	0.0001					
fester Effekt	a_β	b_β	m_β	s_β^2			
	0.005	0.005	0	100			
zufälliger Effekt	a_b	b_b	a_0	b_0	m_0	s_0^2	N
	0.005	0.005	0.005	0.005	8	4	50

Tabelle 7.3: Annahmen bzgl. des MCMC-Algorithmus und der Hyperparameter

Für die Simulation wurden 50 Datensätze simuliert. Für jeden Datensatz und jedes Verfahren wird die Quadratsumme der Differenz zwischen wahrem b_i und geschätztem \hat{b}_i berechnet und durch n geteilt. Der so resultierende geschätzte MSE dient als Gütenachweis des Verfahrens. Die Tabelle 7.4 fasst die Ergebnisse der Simulation zusammen, indem sie den Mittelwert bzw. den Median der jeweiligen geschätzten MSEs bildet.

	ML	REML	DP₁	DP₂	DP₃	DPM₁	DPM₂	DPM₃
Mittelwert	0.1166	0.1165	0.1645	0.1533	0.1637	0.1158	0.1163	0.1163
Median	0.1106	0.1105	0.1547	0.1462	0.1524	0.1103	0.1102	0.1110

Tabelle 7.4: Vergleich der Schätzgenauigkeit anhand der geschätzten MSEs

Man erkennt anhand Tabelle 7.4 zunächst, dass die drei DPM-Modelle hinsichtlich des arithmetischen Mittels die niedrigsten und damit die besten Werte aufweisen. Auch bezüglich des Medians schneiden die DPM-Modelle mit Ausnahme des Modells mit Priori-Annahme $\alpha_0 \sim Ga(2, 4)$ besser als die (RE)ML-Modelle ab. Die Unterschiede zu den ML- bzw. REML-Schätzungen, die mit Normalverteilungsannahme für die zufälligen Effekte arbeiten, sind allerdings nur minimal. Eine wesentlich schlechtere Prädiktion liegt bei den DP-Modellen vor. Offensichtlich ist eine diskrete Verteilung bei den zufälligen Effekte für eine gute Anpassung nicht flexibel genug. Dies heben auch die Boxplots der geschätzten MSEs in Abbildung 7.2 deutlich hervor. Die Annahme für den Konzentrationsparameter spielt hingegen in dieser Simulation eine untergeordnete Rolle.

Vergleicht man die (RE-)ML-Schätzungen ausschließlich mit den DPM-Modellen, so muss man feststellen, dass die bessere Schätzgüte der DPM-Modelle kaum der Rede wert ist. Dies erstaunt – schließlich liegen den simulierten Daten durch die Clusterstruktur ideale Voraussetzungen für ein DPM-Modell zugrunde. Oder anders formuliert: Durch die Multimodalität der Verteilung der zufälligen Effekte erscheint eine Normalverteilungsannahme hierfür denkbar unangebracht. Dennoch, und das ist die wesentliche Erkenntnis dieser Untersuchung, sind darauf aufbauende Modelle auch bei nicht normalverteilten zufälligen

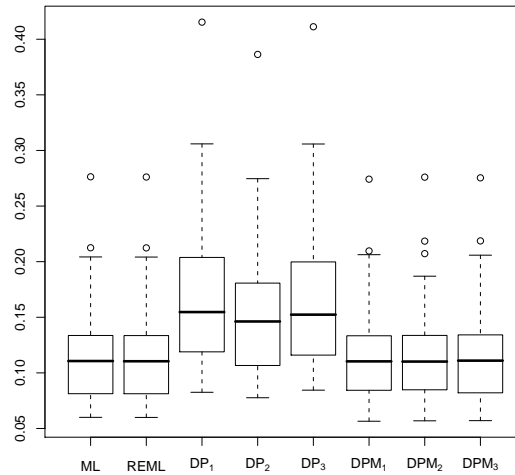


Abbildung 7.2: Boxplots zu den geschätzten MSEs

Effekten erstaunlich flexibel. Das stellt zwar nicht die Schätzgüte der DPM-Modelle an sich in Frage – ihre Motivierung erfährt jedoch eine Einschränkung.

Während bisher die Güte der einzelnen Verfahren in frequentistischer Weise beurteilt wurden, womit man streng genommen eigentlich den Rahmen der Bayes-Inferenz verlässt, sollen nun die Schätzergebnisse des DPM-Modells an einem konkreten Datenbeispiel genauer analysiert werden. Hierzu dient die Datenstruktur gemäß Abbildung 7.1. Das Interesse gilt dabei vor allem der für Dirichlet-Prozesse charakteristischen Clustereigenschaft. Im DPM-Modell ist es die Verteilung der $\theta_1, \dots, \theta_n$, deren Verteilung durch einen Dirichlet-Prozess bestimmt ist. Auf sie wirkt der Cluster-Effekt. Die θ_i mit $i = 1, \dots, n$ entsprechen den Erwartungswerten bzgl. $b_i \stackrel{\text{ind}}{\sim} N(\theta_i, \sigma_b^2)$. Es stellt sich nun die Frage, welche Clusterstruktur das Modell ausgibt. Die Antwort darauf ist mit folgender Problematik verbunden: Dadurch, dass zur Schätzung der $\theta_1, \dots, \theta_n$ MCMC-Verfahren notwendig sind, liefert das DPM-Modell lediglich zu jeder Iteration die Information, welche Individuen zu welchem Cluster gehören. Wie aber fasst man dies zu einer Gesamtaussage über die Clusterstruktur zusammen? Eine Möglichkeit bestünde darin, die in den Iterationen häufigste Clusterstruktur anzugeben. Hier tritt allgemein folgendes Problem auf: Bei großer Individuenanzahl n und großer Trunkierung N sind hierbei so viele Kombinationen möglich, dass sich kaum eine Clusterstruktur von den anderen hervorhebt. Unter Umständen liegen sogar ausschließlich verschiedene Strukturen vor. Eine andere Möglichkeit wäre, für jedes Individuenpaar zu prüfen, wie oft sie in den Iterationen demselben Cluster zugeordnet wurden, um daraus die zugrunde liegende Clusterstruktur herauszuarbeiten. Eine Zusammenstellung dieser Information für sämtliche Paare ist vor allem bei hohem n schwer auswertbar. Des Weiteren könnte man sich auch von der genauen Clustereinteilung in jedem Iterationsschritt lösen und sie lediglich als einen beim Sampling im Hintergrund ablaufenden

Mechanismus auffassen. Als Endergebnis liefert das Modell Schätzungen $\hat{\theta}_1, \dots, \hat{\theta}_n$ sowie einen Schätzer für die Anzahl der Cluster, der mit \hat{k} bezeichnet werden soll. Anhand dieser Informationen kann über externe Methoden der Clusteranalyse zu vorgegebener Clusterzahl eine optimale Zuordnung erlangt werden. Das sog. Varianzkriterium (engl. *k* means clustering) stellt ein solches Verfahren dar, das eine Partition dann als optimal ansieht, wenn die Streuung innerhalb der Cluster minimal und zwischen den Clustern maximal ist. Für eine genaue Beschreibung des Varianzkriteriums sei auf Fahrmeir, Hamerle & Tutz (1996) verwiesen. Bei diesem Vorgehen ist zu beachten, dass die vom Modell ausgegebene geschätzte Anzahl der Cluster maßgebend von dem Präzisionsparameter bestimmt wird. Ein kleines α_0 , wie es z.B. durch eine Priori-Annahme $\alpha_0 \sim Ga(2, 4)$ induziert wird, führt zu einer sehr groben Clusterstruktur, d.h. zu einer geringen Anzahl an Cluster (vgl. Abbildung 4.3). Die Clustereigenschaft des Dirichlet-Prozesses darf also nicht als ein Instrument angesehen werden, das stets zu jeder Form von Daten die wahre, zugrunde liegende Clusterstruktur ausgibt (vgl. Dunson (2008)). Die resultierende Clusterstruktur ist stets mit den zugehörigen Modellannahmen verknüpft.

Ein DPM-Modell mit den Einstellungen gemäß Tabelle 7.3 und $\alpha_0 \sim Ga(2, 4)$ liefert für die Anzahl der Cluster einen Schätzwert von 3. Dies erhält man, wenn man den Median von den Clusteranzahlen sämtlicher Iterationen bildet. Ordnet man nun über das Varianzkriterium, die Schätzungen $\hat{\theta}_1, \dots, \hat{\theta}_n$ bzw. $\hat{b}_1, \dots, \hat{b}_n$ 3 Clustern zu, ergibt sich Abbildung 7.3.

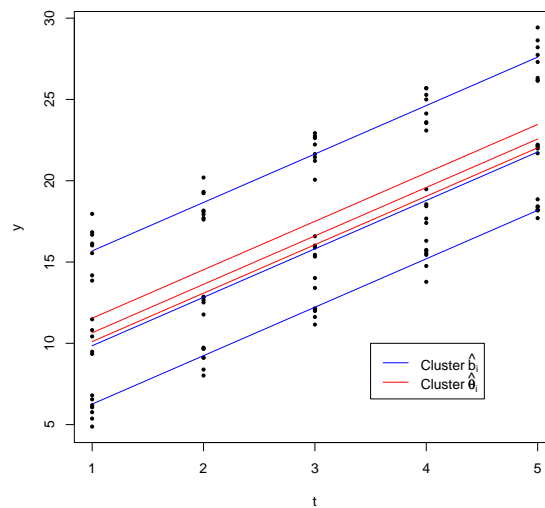


Abbildung 7.3: Clusterzuweisung der Individuen hinsichtlich $\hat{\theta}_i$ und \hat{b}_i

Es fällt dabei auf, dass zwar für die zufälligen Effekte b_i die Niveaustufen der 3 Cluster klar ersichtlich sind, jedoch wäre zu erwarten gewesen, dass dies gerade für die Erwartungswerte θ_i , für die eigentlich die Einteilung in Cluster durch den Dirichlet-Prozess erfolgt, der Fall ist. Die drei Clusterlokationen der θ_i liegen allerdings relativ nahe im Bereich des

generellen Effekts beieinander. Offenbar werden den beiden äußeren Clustern während des Algorithmus häufig Individuen zugewiesen, die der wahren Struktur zu Folge eigentlich zu einem anderen gehören. Auf diese Weise entsteht für alle drei Clusterlokationen eine Tendenz zur Mitte. Dies macht plausibel, warum die DPM-Modelle kaum eine Verbesserung gegenüber den traditionellen (RE)ML-Schätzungen vorweisen können. Wenn hinsichtlich der Erwartungswerte der drei Normalverteilungen in (7.2) keine Unterschiede vorliegen, dann entspricht das de facto einer für alle Individuen einheitlichen Normalverteilungsannahme für die zufälligen Effekte. Aus demselben Grund liefern Kerndichteschätzer für die geschätzten zufälligen Effekte bezüglich des DPM-Modells und des gemäß Maximum-Likelihood bestimmten Modells mit Normalverteilungsannahme dasselbe Ergebnis. Die geschätzten Kurven sind nahezu deckungsgleich (vgl. Abbildung 7.4).

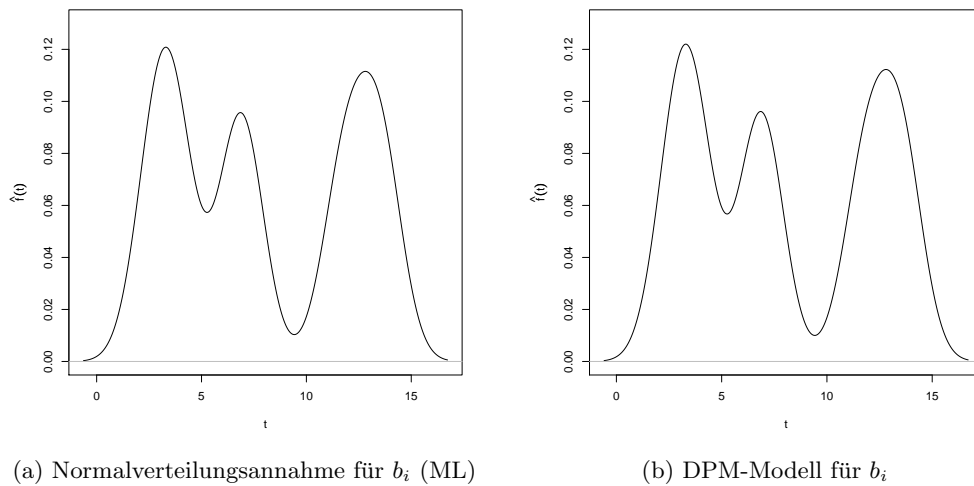


Abbildung 7.4: Kerndichteschätzer für \hat{b}_i mit Bandweite 1 und Gauß-Kern

7.2 Mischverteilte Daten mit geringen Unterschieden

Dieselben Untersuchungen wie in Abschnitt 7.1 sollen nun an simulierten Daten durchgeführt werden, deren zufällige Effekte einer Mischverteilung mit nahe beieinander liegenden Klassenmittelpunkten folgen. Die Erwartungswerte der Normalverteilungen in (7.2) werden nun mit $\mu_1 = 1.5$, $\mu_2 = -0.5$ und $\mu_3 = -1.5$ gewählt. Abgesehen davon entsprechen die Simulationseinstellungen denen in Abschnitt 7.1. Sie sind nochmal in Tabelle 7.5 zusammengefasst.

Damit resultiert zum Beispiel eine Datenstruktur wie in Abbildung 7.5. Dort sieht man, dass die Clusterstruktur in den Daten kaum noch wahrnehmbar ist.

Die zu vergleichenden Modelle sind analog zu Abschnitt 7.1 gegeben. Die Wahl der Hyperparameter und der Einstellungen des MCMC-Algorithmus werden erneut gemäß Tabelle

Daten	n	n_i	t_{i1}	t_{i2}	t_{i3}	t_{i4}	t_{i5}
	20	5	1	2	3	4	5
Fehlervarianz	σ^2						
	0.5						
feste Effekte	β_0	β_1					
	8	3					
zufälliger Effekt	σ_b^2	μ_1	μ_2	μ_3	w_1	w_2	w_3
	1	1.5	-0.5	-1.5	0.4	0.3	0.3

Tabelle 7.5: Simulationseinstellungen für kleine Clusterunterschiede

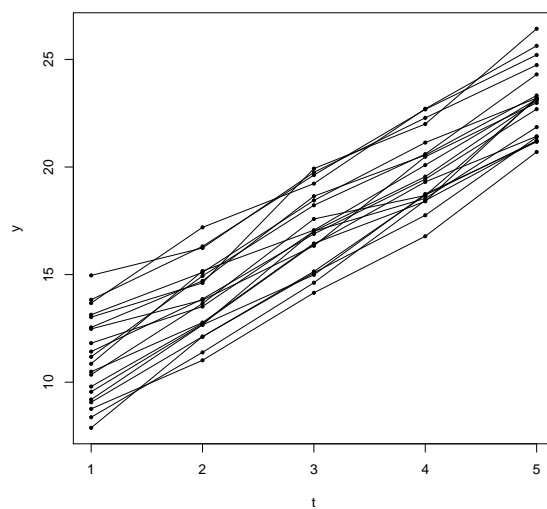


Abbildung 7.5: Simuliertes Datenbeispiel für das Modell (7.1) gemäß (7.2) und den Einstellungen in Tabelle 7.5

7.3 gewählt. Das Ergebnis der Simulation mit 50 Datensätzen ist dem der Datensituation mit weit auseinander liegenden Clustern sehr ähnlich (vgl. Tabelle 7.6 und Abbildung 7.6).

	ML	REML	DP ₁	DP ₂	DP ₃	DPM ₁	DPM ₂	DPM ₃
Mittelwert	0.1227	0.1226	0.1496	0.1409	0.1504	0.1231	0.1229	0.1232
Median	0.1249	0.1252	0.1436	0.1356	0.1427	0.1243	0.1228	0.1234

Tabelle 7.6: Vergleich der Schätzgenauigkeit anhand der geschätzten MSEs

Die DPM-Modelle weisen zwar hinsichtlich des Medians der schätzten MSEs bessere Werte auf als die Modelle mit Normalverteilungsannahme für die zufälligen Effekte, bezüglich des arithmetischen Mittels verhält es sich jedoch umgekehrt. Insgesamt sind die Differenzen der Verfahren gering. Die DP-Modelle sind hinsichtlich der Schätzungsgüte deutlich schlechter, wenn auch die Unterschiede nicht so groß sind wie in Abschnitt 7.1.

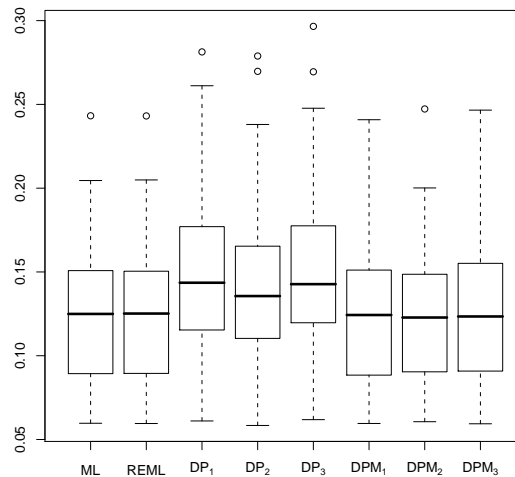


Abbildung 7.6: Boxplots zu den geschätzten MSEs

Abgesehen von einem Vergleich der Schätzungsgüte ist es nun interessant, inwieweit die drei nahe beieinander liegenden Clusterlokationen von dem DPM-Modell erfasst werden. Hierzu soll an den Daten in Abbildung 7.5 dieselben Untersuchungen vorgenommen werden wie in Abschnitt 7.1: Die $\hat{\theta}_1, \dots, \hat{\theta}_n$ Schätzungen und $\hat{b}_1, \dots, \hat{b}_n$ werden jeweils den \hat{k} vorgegebenen Clustern nach dem Varianzkriterium zugewiesen. Abbildung 7.7 zeigt das Ergebnis eines DPM-Modells mit den Einstellungen gemäß Tabelle 7.3 und $\alpha_0 \sim Ga(2, 4)$. Die Anzahl der Cluster wurde dabei korrekterweise mit 3 geschätzt.

Während die geschätzten $\hat{b}_1, \dots, \hat{b}_n$ der wahren Struktur gemäß zugeordnet werden, liegen hinsichtlich der $\hat{\theta}_1, \dots, \hat{\theta}_n$ offenbar keine Unterschiede vor, die eine differenzierte Cluster-

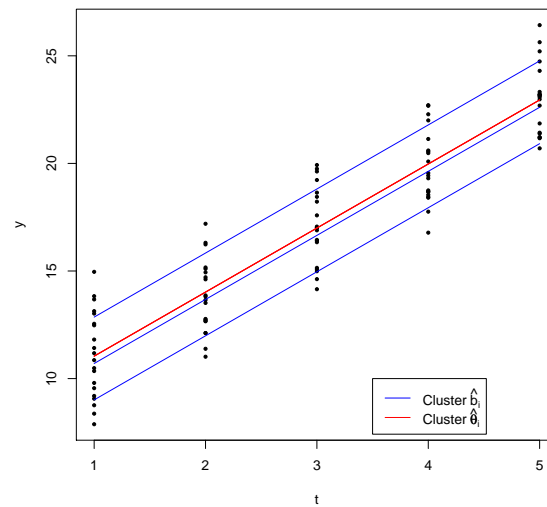
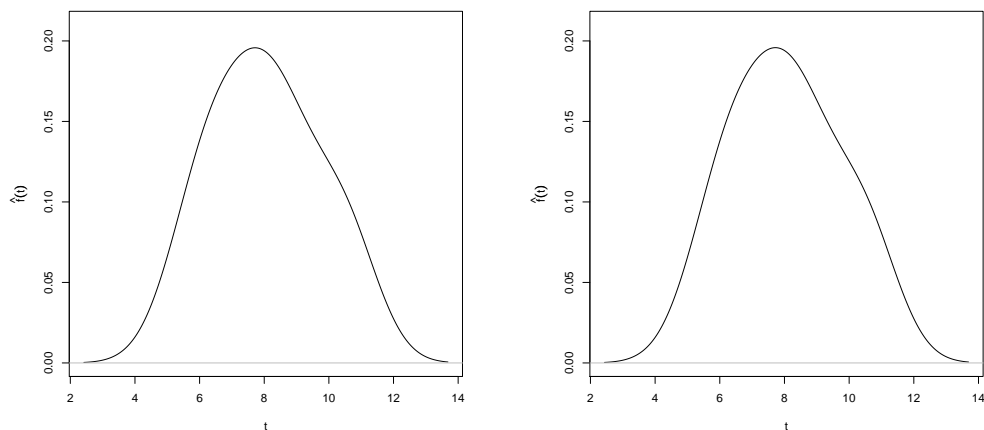


Abbildung 7.7: Clusterzuweisung der Individuen hinsichtlich $\hat{\theta}_i$ und \hat{b}_i

einteilung möglich machen würde. Waren bei der Datenstruktur in Abschnitt 7.1 hierfür noch verschiedene Stufen erkennbar (vgl. Abbildung 7.3), so sind diese nun egalisiert, was zu den nahezu äquivalenten Schätzergebnissen der DPM-Modelle und der Modelle mit Normalverteilungsannahme für die zufälligen Effekte führt. Die Kerndichteschätzer für die zufälligen Effekte sind folglich wiederum praktisch identisch (vgl. Abbildung 7.8).



(a) Normalverteilungsannahme für b_i (ML)

(b) DPM-Modell für b_i

Abbildung 7.8: Kerndichteschätzer für \hat{b}_i mit Bandweite 1 und Gauß-Kern

Die Kerndichteschätzer in Abbildung 7.8 zeigen auch, dass die Erwartungswerte der Normalverteilungen in der Mischverteilung 7.2 offensichtlich zu nahe beieinander liegen, als dass dies anhand der Daten erkannt werden könnte. Statt einer trimodalen Dichte wie in

Abbildung 7.4 zeigt sich nun eine unimodale geschätzte Dichte. Lediglich eine Wölbung in dem Bereich um $t = 10$ weist auf eine Asymmetrie der zufälligen Effekte hin.

7.3 Zusammenfassung der Simulationsergebnisse

Die Simulationsbeispiele in den Abschnitten 7.1 und 7.2 zeigen, dass DPM-Modelle und Modelle, die mit Normalverteilungsannahme für die zufälligen Effekte arbeiten, eine ähnliche Schätzgüte aufweisen. DP-Modelle lieferten bezüglich der geschätzten MSE deutlich schlechtere Resultate, was auf die diskrete Verteilung zurückzuführen ist. Ein Hinweis, warum die DPM-Modelle trotz des für die zufälligen Effekte gewählten Mischungsansatzes die Schätzgüte der Modelle mit Normalverteilungsannahme für die zufälligen Effekte nicht wesentlich überbieten konnten, geben die beiden konkreten Datenbeispiele. Sie zeigen, dass die DPM-Modelle es nicht vermochten, die drei verschiedenen Niveaustufen des Intercepts durch die Schätzungen $\hat{\theta}_1, \dots, \hat{\theta}_n$ in ausreichender Weise zu erfassen. Es stellt sich die Frage, ob durch eine geschicktere Adjustierung der Modellgrößen bessere Ergebnisse erzielt werden könnten. Ein ausführliches Studium dieses Sachverhalts liegt aber jenseits dieser Arbeit, so dass die Frage unbeantwortet bleibt. Nichtsdestotrotz kann eine wichtige Erkenntnis festgehalten werden: Die Annahme einer Normalverteilung für die zufälligen Effekte erweist sich selbst bei einer zugrunde liegenden Mischverteilung als ein gutes Schätzinstrument. Dies deckt sich mit den Ergebnissen einer umfangreichen Simulationsstudie von Li, Lin & Müller (2007). Sie stellten fest, dass im Rahmen eines additiven gemischten Modells bei bimodal verteilten zufälligen Effekte die Normalverteilungsannahme immer noch zu robusten Schätzungen aller Modellparameter führt. Effizienznachteile gegenüber einem DP-Modell im Sinne eines höheren MSEs können lediglich bei bestimmten festen Effekten festgestellt werden. Ein Erklärungsansatz fußt auf der Tatsache, dass es sich bei der Dirichlet-Prozess-Priori im bayesianischen Sinne letztlich nur um eine Priori-Annahme handelt. Bei wachsendem Stichprobenumfang sinkt ihr Einfluss und die Schätzergebnisse werden hauptsächlich durch die Daten selbst bestimmt. Hinzu kommt, dass, selbst wenn die für alle Individuen einheitliche Priori der zufälligen Effekte ungünstig gewählt ist, die Posteriori-Verteilung für b_i eine für jedes Individuum separate Anpassung gewährt.

8 Datenanalyse: LISA-Daten

In diesem Kapitel wird das additive gemischte Modell aus Kapitel 6 auf ein Datenbeispiel angewendet. Hierfür dienen die LISA-Daten, die in Abschnitt 8.1 kurz vorgestellt werden. Eine ausführlichere Beschreibung der Daten findet sich bei Fenske (2008). In den anschließenden Abschnitten werden die Ergebnisse berechneter Modelle veranschaulicht und diskutiert.

8.1 LISA-Daten

Die LISA-Studie stellt eine Längsschnittstudie dar, deren Ziel es ist, die Zusammenhänge zwischen den Lebensumständen und der Entwicklung des Immunsystems bei Kindern in Deutschland in ihren ersten Lebensjahren zu untersuchen. So steht die Abkürzung LISA für *Influences of Life-style factors on the development of the Immune System and Allergies in East and West Germany*. Die Studie lässt sich jedoch auch dazu verwenden – und so soll es im weiteren Verlauf geschehen –, die zeitliche Entwicklung des Körpergewichts der Kinder zu analysieren. Konkret dient als Untersuchungsmerkmal der Body Mass Index, der das Körpergewicht auf folgende Weise in Relation zur Körpergröße setzt:

$$\text{BMI} = \frac{\text{Körpergewicht [kg]}}{(\text{Körpergröße})^2 [\text{m}^2]}.$$

Für die LISA-Studie wurden an vier Standorten in Deutschland (München, Leipzig, Wesel und Bad Honnef) insgesamt 3097 Kinder beobachtet, die zwischen Ende 1997 und Anfang 1999 geboren wurden. Das vorliegende Datenmaterial der Studie umfasst für jedes Kind von seiner Geburt an bis zum Alter von sechs Jahren neun Messzeitpunkte, an denen es untersucht und seine Eltern mittels Fragebogen über die Lebensumstände ihres Kindes befragt wurden. Neben den Angaben zum Alter und zum Body Mass Index beinhaltet der zugrunde liegende Datensatz zeitkonstante Kovariablen. Die folgende Analyse konzentriert sich jedoch im Besonderen auf die Untersuchung des Alterseffekts auf den Body Mass Index.

Dem Problem mit fehlenden Werten wird im Folgenden durch eine Complete-Case-Analysis begegnet, d.h. die Beobachtungen zum Zeitpunkt t_{ij} finden nur dann in der Analyse Berücksichtigung, wenn dort sämtliche Angaben vorhanden sind. Fehlt also für ein Individuum i bei der j -ten Messung die Information zum Alter oder zum BMI, wird nur die entsprechende Beobachtung entfernt, fehlt hingegen der Wert einer zeitkonstanten Kovariablen, wird die Person gänzlich ausgeschlossen. Damit resultieren 2043 Individuen bei insgesamt 17316 Beobachtungen.

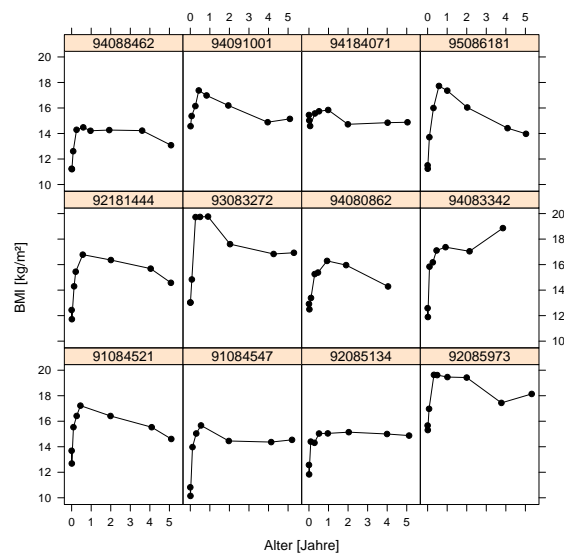


Abbildung 8.1: Individuelle BMI-Verläufe in Abhängigkeit vom Alter bei zwölf zufällig ausgewählten Kindern

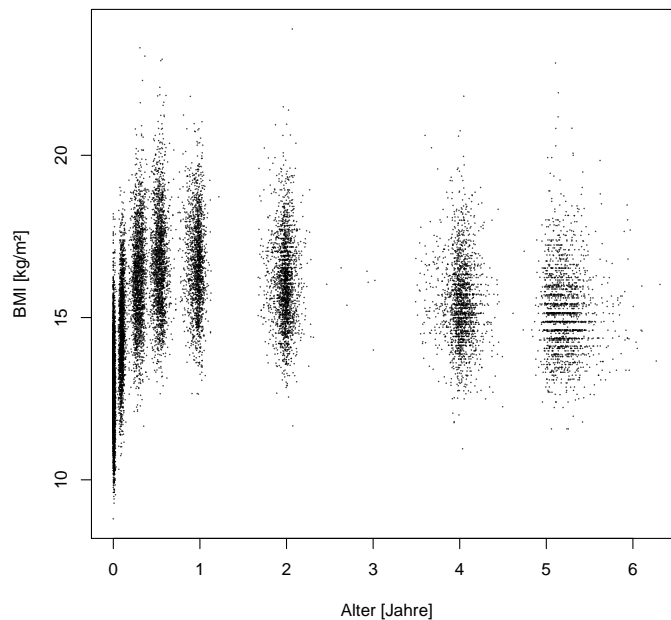


Abbildung 8.2: Streudiagramm zu den BMI-Messungen in Abhängigkeit vom Alter

Abbildung 8.1 veranschaulicht für zwölf zufällig ausgewählte Kinder die individuellen Verläufe der Zielgröße BMI hinsichtlich der Einflussgröße Alter. Das Alter ist dabei stets mit der beobachteten Zeitspanne identisch, da die erste Messung bei jedem Kind zur Geburt stattfand. Die sogenannten Traceplots zeigen bei den meisten Kindern einen steilen Anstieg des BMI in den ersten sechs Lebensmonaten und einen daraufhin folgenden leichten Rückgang. Ein solcher Eindruck lässt sich auch im Streudiagramm in Abbildung 8.2 erkennen. Dort werden für alle Messzeitpunkte die jeweiligen BMI-Werte eingetragen, wobei nicht unterschieden wird, welche Punkte zu welchem Individuum gehören.

Neben dem Alter können noch weitere Kovariablen hinsichtlich eines signifikanten Effekts auf den BMI überprüft werden. Die Tabelle 8.1 und die Tabelle 8.2 liefern eine Übersicht über die kategorialen bzw. metrischen Prädiktoren auf der Basis des reduzierten Datensatzes.

Variable	Beschreibung	Kategorien	rel. Häuf.	abs. Häuf.
sex	Geschlecht	0 = weiblich	47.2%	964
		1 = männlich	52.8%	1079
breast	Nahrung in den ersten 6 Lebensmonaten	0 = Flasche und Muttermilch	40.5%	828
		1 = nur Muttermilch	59.5%	1215
tvpc	Zeit pro Tag vor Fernseher und Computer im Alter von 4 Jahren	1 = weniger als eine Stunde	72.1%	1472
		2 = 1 bis 2 Stunden	26.6%	543
		3 = mehr als 3 Stunden	1.4%	28
mSmoke	Rauchen der Mutter in der Schwangerschaft	0 = nein	86.0%	1756
		1 = ja	14.0%	287
mEdu	höchster Schulabschluss der Mutter	1 = kein Abschluss	0.4%	9
		2 = Hauptschulabschluss	6.1%	124
		3 = Realschulabschluss	36.1%	737
		4 = Fachabitur	19.0%	389
		5 = Abitur	38.4%	784
area	Umgebung	0 = ländlich	21.5%	439
		1 = städtisch	78.5%	1604

Tabelle 8.1: Beschreibung der kategorialen Kovariablen

Variable	Beschreibung	Median	Mittelw.	Stand.
ageY	Alter (in Jahren)	0.52	1.39	1.76
wgain2y	Gewichtszuwachs bis zum Alter von 2 Jahren (in <i>kg</i>)	8.83	8.91	1.29
outdoor	im Freien verbrachte Stunden im Alter von 4 Jahren (in Stunden pro Tag)	3.50	3.45	1.15
mBMI	BMI der Mutter (in <i>kg/m²</i>)	21.72	22.58	3.74
mDiffBMI	BMI-Differenz der Mutter in der Schwangerschaft (in <i>kg/m²</i>)	4.96	5.12	1.63

Tabelle 8.2: Beschreibung der metrischen Kovariablen

8.2 DPM-Modell mit P-Spline, zufälligen Effekten 1. Grades und festen Effekten

Die folgende Regressionsanalyse befasst sich in erster Linie mit dem Effekt des Alters auf den BMI. Den Traceplots in Abbildung 8.1 und dem Streudiagramm in Abbildung 8.2 zu Folge liegt hierfür ein nichtlinearer Zusammenhang vor. Das Verwenden eines linearen Modells für die Variable Alter ist daher ausgeschlossen und es muss auf nonparametrische Inferenz-Methoden gemäß Abschnitt 3.2 zurückgegriffen werden. Diese sind allerdings mit folgendem Problem konfrontiert: Die besondere Struktur der Daten mit Messungen zu regelmäßigen Zeitpunkten resultiert in mehreren Punktwolken, wohingegen in den dazwischen liegenden Zeitspannen kaum Messungen vorliegen. Lokale Schätzungen in diesem Bereich werden daher nur durch wenige Datenpunkte gesteuert. Dies birgt die Gefahr, dass dort der zeitliche Effekt in Richtung dieser Daten verzerrt wird. Penalisierte Splines, wie z.B. die P-Splines, die die Rauheit der geschätzten Funktion bestrafen, sind wirkungsvolle Methoden, solche Ausschläge der Funktion zu unterbinden. Der generelle Effekt des Alters auf den BMI soll daher im Folgenden durch einen P-Spline modelliert werden.

Um der longitudinalen Struktur gerecht zu werden, sollen zufällige Effekte die individuellen Aspekte des Alters erfassen. Die in Abschnitt 6.2 beschriebene Problematik der „doppelten Erfassung“ des zeitlichen Effekts soll durch eine Zentrierung der zufälligen Effekte in jedem Iterationsschritt gelöst werden. Diese lassen sich damit auch stets als individuelle Abweichungen vom generellen Alterseffekt interpretieren. Den Grad der zufälligen Effekte betreffend konzentriert sich die Analyse im Wesentlichen auf zufällige Effekte 1. Grades. Dies lässt sich folgendermaßen motivieren: Betrachtet man exemplarisch die zeitlichen Verläufe des Body-Mass-Index für ein paar Individuen (vgl. Abbildung 8.1), so fällt zunächst auf, dass sie sich in erster Linie hinsichtlich des Niveaus unterscheiden. So manches Kind weist schon in den ersten Lebensmonaten einen höheren BMI als andere Kinder auf und entwickelt sich dann aber ähnlich zu den anderen Kindern - nur auf einem anderen Level.

Die Analyse sollte daher in jedem Fall einen Random-Intercept enthalten. Darüber hinaus fällt auf, dass bei der zeitlichen Entwicklung im Bereich des jeweiligen BMI-Maximums deutliche Unterschiede vorliegen. Bei manchen Kindern ist die Krümmung im ersten Lebensjahr viel deutlicher ausgeprägt als bei anderen. Durch zufällige Effekte ersten Grades sollen daher auch die individuellen Steigungen berücksichtigt werden.

Abgesehen vom Alter sollen noch weitere Kovariablen in die Regressionsanalyse eingehen und auf einen signifikanten Effekt auf den BMI überprüft werden. Hierfür wurden die Dummyvariablen „sex“, „mSmoke“ und „area“, sowie die stetige Variable „mBMI“ herangezogen. Sie werden linear ins Modell aufgenommen.

Für diese Zielstellung soll nun ein additives gemischtes Modell berechnet werden, wie es in Kapitel 6 hergeleitet wurde. Für die Verteilung der zufälligen Effekte wird ein DPM-Modell verwendet, welches sich nach Kapitel 7 als besser als das DP-Modell erwiesen hat. Die Berechnungen werden mittels der R-Funktion `BlockDPM()` (vgl. Abschnitt 6.4) durchgeführt, die das Modell (6.2) berechnet. Die Hyperparameter und MCMC-Einstellungen werden gemäß Tabelle 8.3 gewählt.

MCMC	<i>Iterationen</i>	<i>Burnin</i>	<i>W</i>						
	55000	5000	50						
Fehlervarianz	a_ε	b_ε							
	0.0001	0.0001							
feste Effekte	a_β	b_β	m_β	s_β^2					
	0.005	0.005	0	100					
P-Spline	a_γ	b_γ	m	l	k				
	0.005	0.005	12	3	2				
zufällige Effekte	a_b	b_b	a_0	b_0	a_α	b_α	m_0	s_0^2	N
	0.005	0.005	0.005	0.005	2	4	0	10	100

Tabelle 8.3: Annahmen bzgl. des MCMC-Algorithmus und der Hyperparameter

Die Bedeutung der Kürzel kann weitestgehend dem Abschnitt 6.3 entnommen werden. Hinsichtlich der P-Spline-Einstellungen sei daran erinnert, dass m für die Anzahl der Knoten (ohne Knotenerweiterung), l für den Grad der polynomialen Struktur und k für die Ordnung steht, mit der der Spline penalisiert wird (vgl. Abschnitt 3.2). Grundsätzlich wurden die Einstellungen so gewählt, dass kaum ein a priori Wissen in die Modellierung eingeht. Lediglich die für den Präzisionsparameter α_0 angenommene $Ga(2, 4)$ fällt aus diesem Schema heraus; diese wurde gewählt, um eine relativ grobe Clustereinteilung herbeizuführen.

Die Modellschätzung liefert nun folgende Resultate: Zunächst fasst Tabelle 8.4 die Information aus den Ziehungen der Fehlervarianz und der festen Effekte zusammen. Die Schätzungen $\hat{\beta}$ sind durch den Mittelwert oder durch den Median gegeben, während die

empirische Standardabweichung auf die Schätzpräzision hinweist. Zudem werde für die Schätzer jeweils ein Kreditabilitätsintervall zum Grad 0.95 angegeben, das ebenfalls die Schätzgenauigkeit beschreibt und Aufschluss darüber gibt, ob es sich um einen signifikanten Effekt handelt. Hierfür werden symmetrische Kreditabilitätsintervalle aufgeführt, die an beiden Enden der empirischen Verteilung der Ziehungen jeweils 2.5% abschneiden.

	Mittelwert	Median	emp. Stand.	2.5%-Quantil	97.5%-Quantil
σ^2	0.97791	0.97728	0.01469	0.94968	1.00745
sex	0.31300	0.31190	0.04132	0.23430	0.39693
mBMI	0.04176	0.04183	0.00569	0.03118	0.05282
mSmoke	0.08490	0.08576	0.05695	-0.02314	0.18757
area	0.00696	0.00885	0.05318	-0.09640	0.11140

Tabelle 8.4: Schätzergebnisse der Fehlervarianz und der festen Effekte

Man erkennt anhand der Tabelle 8.4, dass es sich bei den Variablen „sex“ und „mBMI“ um signifikante Einflussgrößen auf dem Niveau 5% handelt. Die jeweiligen symmetrischen 95%-Vertrauensintervalle enthalten nicht die Null, so dass jeweils die Hypothese $\beta_r = 0$ zum Niveau 5% abgelehnt werden kann. Der Body-Mass-Index ist bei Jungen durchschnittlich um ca. 0.31 höher als bei Mädchen, wenn hinsichtlich der anderen Einflussgrößen dieselben Ausprägungen vorliegen. Ebenso führt ein höherer BMI der Mutter auch zu einem höheren BMI des Kindes, sofern alle anderen Kovariablen konstant sind. Es zeigt sich jedoch, dass die Tatsache, ob die Mutter in der Schwangerschaft geraucht hat oder nicht, keinen signifikanten Einfluss auf den BMI des Kindes ausübt. Selbiges lässt sich für die Kovariable „area“ sagen: Ob ein Kind in der Stadt oder im ländlichen Raum aufwächst, hat hinsichtlich des BMIs keine wesentliche Bedeutung. Diese Aussagen beziehen sich dabei auf die hier gewählte Kovariablenstruktur und gelten unter den jeweiligen Ausprägungen der anderen Einflussgrößen.

In der Abbildung 8.3 können die Samplingpfade und die Dichteschätzungen zu den festen Effekten begutachtet werden. Daran erkennt man ebenso wie in Abbildung 8.4, die die Autokorrelationsfunktionen der Ziehungen darstellt, dass sich für die Einflussgrößen „sex“ und „mSmoke“ nahezu perfekte Samples ergeben haben, während bei den Variablen „mBMI“ und „area“ trotz einer Ausdünnung vom Faktor 50 noch Autokorrelationen innerhalb der Samples vorzufinden sind. Dennoch kann in den Pfaden ein Rauschen um einen festen Wert beobachtet werden, so dass dieser als Schätzung für den entsprechenden Koeffizienten dient.

Der geschätzte zeitliche Verlauf soll durch Abbildung 8.5 veranschaulicht werden. Die Kovariablenwerte der festen Effekte werden hierbei gemittelt, so dass die Graphik ausschließlich den Effekt der Variable „ageY“ auf den BMI visualisiert. Jene zeigt diesbezüglich zum einen den allgemeinen geschätzten Effekt (rote Linie) und zum anderen die Datenpunkte und den geschätzten Effekt für 4 ausgewählte Individuen. Die Tabelle 8.5 setzt die Farben

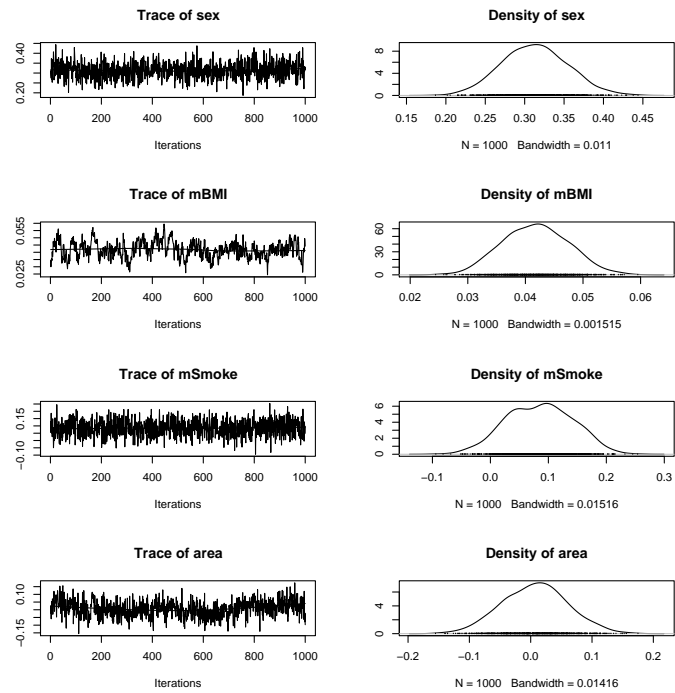


Abbildung 8.3: Samplingpfade und Dichteschätzungen zu den festen Effekten

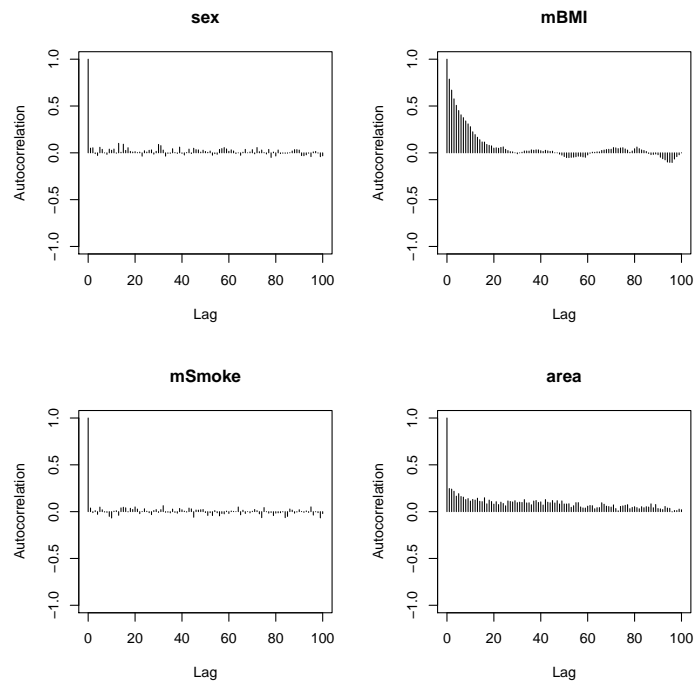


Abbildung 8.4: Autokorrelationsfunktionen zu den festen Effekten

in der Graphik mit den Personen in Verbindung. Diese können entweder durch die dem Datensatz zugehörige Variable „id“ oder durch eine Personennummer identifiziert werden. Eine Personennummer 824 sagt beispielsweise aus, dass die entsprechende Person von den $n = 2043$, nach der Variable „id“ aufsteigend geordneten Individuen die 824. ist.

Variable „id“	Personennummer	Farbe
92185191	824	hellgrün
92189214	888	orange
94182011	1786	hellblau
95089461	2037	lila

Tabelle 8.5: Identifikationsschlüssel der vier ausgewählten Individuen

Die vier Individuen in Abbildung 8.5 wurden deshalb ausgewählt, da sie relativ auffällige Verläufe aufweisen und exemplarisch für bestimmte Entwicklungsformen stehen. So weist ein Kind generell sehr niedrige BMI-Werte (id=92189214) auf, ein weiteres ist durch relativ hohe BMI-Werte mit einem sehr stark ausgeprägtem Maximum im ersten Lebensjahr charakterisiert (id=92185191), bei einem drittem Kind findet nach dem 1. Lebensjahr untypischerweise ein Anstieg des BMI statt (id=95089461) und zu guter Letzt wird ein Individuum betrachtet, dass mit seinen BMI-Werten weitestgehend im Normalbereich liegt, jedoch im Alter von 2 Jahren einen sehr hohen Wert aufweist, dessen Korrektheit angezweifelt werden kann (id=94182011). Das Augenmerk liegt nun darauf, wie das Modell auf solche extremen Ausprägungen reagiert.

Zunächst zeigt Abbildung 8.5 einen nichtlinearen Verlauf für den generellen Alterseffekt: Einem sehr steilen Anstieg in den ersten sechs Lebensmonaten folgt ein langsamer Rückgang. Insgesamt weist der P-Spline eine leicht unruhige Form auf. Dies liegt daran, dass in bestimmten Bereichen kaum Messungen vorliegen und diese dann die Schätzung in den jeweiligen Bereichen stark beeinträchtigen. Die Penalisierung des Splines wirkt diesem Phänomen zwar entgegen, eine völlig glatte Kurve konnte dadurch aber nicht erreicht werden. Hinsichtlich der ausgewählten Individuen erkennt man, wie die zufälligen Effekte 1. Grades einerseits eine individuelle Niveau-Verschiebung zulassen und andererseits auch die Steigung individuell gestalten können. Für eine genauere individuelle Anpassung, wie es z.B. für das Individuum 92185191 im Bereich des Maximums nötig wäre, ist der Grad allerdings zu niedrig.

Exemplarisch sollen in Abbildung 8.6 für die vier Individuen die Samplingpfade für den Random-Intercept bezüglich der jeweils 1000 gespeicherten Werte angegeben werden. Diese sind alles in allem ganz passabel; sie zeigen jedoch auch auf, dass gelegentlich kleinere Sprünge auftreten. Die Ursache liegt in der Clusterstruktur hinsichtlich des Erwartungswerts für die Verteilung eines zufälligen Effekts. Das Aufdatieren einer Clusterlokation richtet sich stets nach allen Individuen in diesem Cluster. So kann es passieren, dass ein Individuum i , wenn es in gewisser Weise zwischen zwei Clustern steht, eine Zeit lang mehr

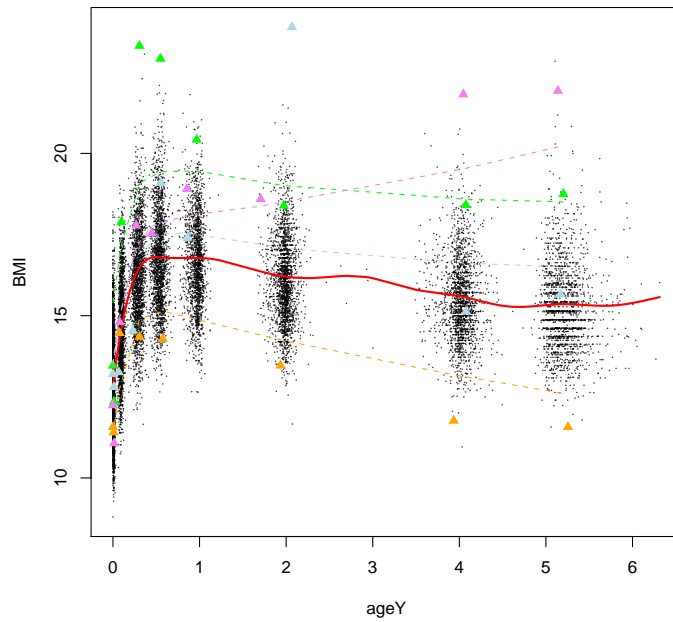


Abbildung 8.5: Geschätzter zeitlicher Effekt

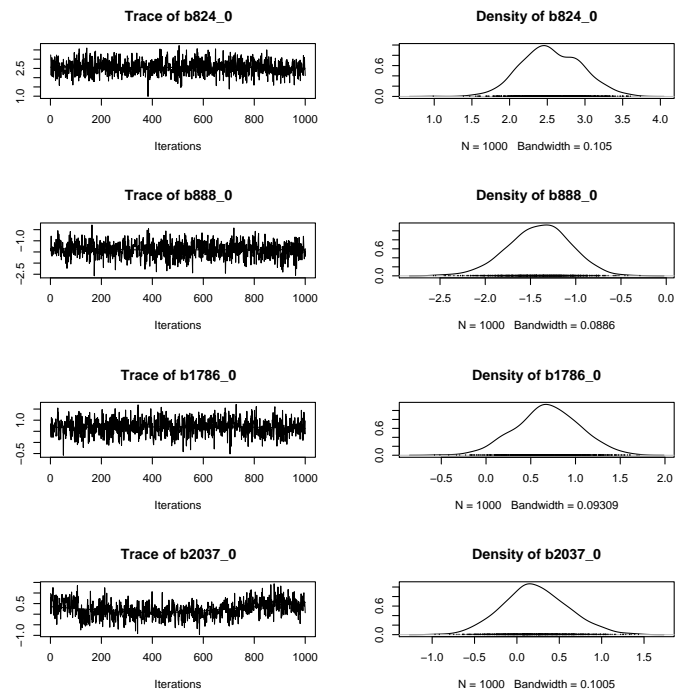


Abbildung 8.6: Samplingpfade und Dichteschätzungen des Random-Intercepts für vier ausgewählte Individuen

dem einen und dann, wenn sich die andere Clusterlokation in die eigene Nähe verschoben hat, dem anderen Cluster zugewiesen wird. Ein baldiges Zurückspringen in den alten Cluster kann ausbleiben, wenn sich dessen Lokation durch die Abwesenheit der Person i von eben dieser Person entfernt hat. Die Samplingpfade der anderen zufälligen Effekte und die der Parameter des P-Splines können mit zufriedenstellend bis tadellos beurteilt werden.

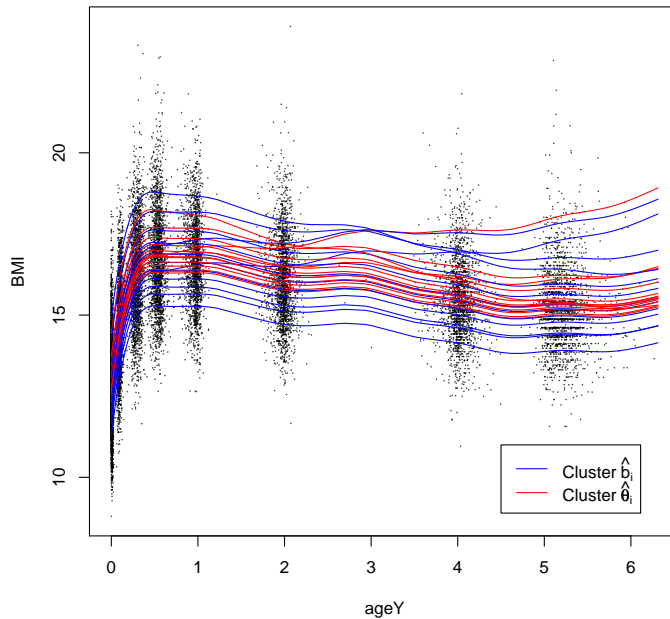


Abbildung 8.7: Clusterzuweisung der Individuen hinsichtlich $\hat{\theta}_i$ und \hat{b}_i

Schließlich werden in Analogie zu Kapitel 7 die Schätzungen $\hat{\theta}_1, \dots, \hat{\theta}_n$ bzw. $\hat{b}_1, \dots, \hat{b}_n$ in die geschätzten 13 Cluster eingeteilt. Man erkennt anhand Abbildung 8.7, dass sich die damit resultierenden Cluster sowohl hinsichtlich der $\hat{\theta}_i$ als auch der \hat{b}_i hauptsächlich im Niveau unterscheiden. Der Anstieg des BMIs in den ersten sechs Lebensmonaten fällt in den Clustern unterschiedlich steil aus. Danach aber sind die Verläufe nahezu parallel. Dies bedeutet, dass für ein ein halbes Jahr altes Kind die Entwicklung seines BMIs gut prognostiziert werden kann, indem man das BMI-Niveau, das das Kind zu diesem Zeitpunkt hat, entsprechend weiterverfolgt. Es gibt jedoch eine Gruppe von Kindern, die aus diesem Schema herausfällt. Sie haben hinsichtlich ihres BMIs kein Maximum im ersten Lebensjahr. Stattdessen steigt ihr BMI von einem zum Zeitpunkt $ageY = 0.5$ noch völlig durchschnittlichen Level in den nächsten Jahren kontinuierlich an, bis sie schließlich im Alter von 4 Jahren verglichen mit den anderen Clusterstufen zu den Kindern mit dem größten BMI und damit zu den dicksten Kindern zählen. Dieser Cluster wird auch hinsichtlich der $\hat{\theta}_i$ erfasst. Ein Indiz dafür, dass er wirklich ernst zu nehmen ist. Schließlich unterliegt den $\hat{\theta}_i$ eine Tendenz zur Mitte, der die Unterschiede zwischen den $\hat{\theta}_i$ gering werden lässt (vgl. Kapitel 7). Diese Tendenz erkennt man an den $\hat{\theta}_i$ sowie an den \hat{b}_i , wenn auch nicht in so starkem Maße. Der bei der Korrektur vollzogene Zentrierungsschritt spielt

in diesem Zusammenhang auch eine Rolle. Er legt fest, dass sich die zufälligen Effekte b_i aller Individuen insgesamt aufheben. Gäbe es nur einen einzigen Cluster, so führte das zu $\hat{\theta}_i = 0 \forall i = 1, \dots, n$. Ebenso ist es plausibel, dass auch bei paar Clustern die Unterschiede zwischen den $\hat{\theta}_i$ nicht allzu groß ausfallen werden. Umso hervorhebenswerter ist es, wenn wie in diesem Beispiel relativ deutlich Unterschiede hinsichtlich der $\hat{\theta}_i$ vorliegen.

8.3 DPM-Modell mit P-Spline und individuellen TP-Splines

Im weiteren Verlauf steht ausschließlich der zeitliche Effekt auf den BMI im Fokus. Ein Modell mit zufälligen Effekten 1. Grades wie in Abschnitt 8.2 macht es möglich, hinsichtlich des Alterseffekts individuelle Niveauverschiebungen und Steigungsänderungen in Relation zum generellen P-Spline zu modellieren. Abbildung 8.5 zeigt aber, dass zufällige Effekten 1. Grades nicht flexibel genug sind, um die unterschiedlich ausgeprägten individuellen Maxima im Bereich des ersten Lebensjahres adäquat zu modellieren. So ist beispielsweise an Individuum 92185191 erkennbar, dass der steile Anstieg des BMIs bis zu dem auf sehr hohem Niveau liegenden Maximum und der entsprechend steile Rückgang nicht ausreichend erfasst werden. In diesem Zusammenhang ist es interessant, das Modell zu individuellen TP-Splines 1. Grades zu erweitern. Hierbei sollen ein oder zwei Knoten im Bereich des ersten Lebensjahres bessere lokale Schätzungen ermöglichen. Konkret bietet es sich an, einen Knoten bei 0.5 zu wählen, da dort in etwa die meisten Kinder ihr persönliches BMI-Maximum haben. Auch ein Knoten bei 1 scheint sinnvoll zu sein, weil sich die Steigung, die sich zu diesem Zeitpunkt eingestellt hat, danach nicht mehr wesentlich ändert. Auf diese Weise können in den Intervallen $[0; 0.5]$, $[0.5; 1]$ und $[1; 6]$ jeweils unterschiedliche Steigungen modelliert werden. Allgemein lautet nun die Designmatrix für das Individuum i bei Grad $l = 1$ und zwei Knoten κ_2 und κ_3 mit $t_{i1} < t_{i2} < t_{i3} < \kappa_2 < t_{i4} < t_{i5} < \kappa_3 < t_{i6} < t_{i7} < t_{i8} < t_{i9}$:

$$\mathbf{Z}_i := \begin{pmatrix} 1 & t_{i1} & 0 & 0 \\ 1 & t_{i2} & 0 & 0 \\ 1 & t_{i3} & 0 & 0 \\ 1 & t_{i4} & t_{i4} - \kappa_2 & 0 \\ 1 & t_{i5} & t_{i5} - \kappa_2 & 0 \\ 1 & t_{i6} & t_{i6} - \kappa_2 & t_{i6} - \kappa_3 \\ 1 & t_{i7} & t_{i7} - \kappa_2 & t_{i7} - \kappa_3 \\ 1 & t_{i8} & t_{i8} - \kappa_2 & t_{i8} - \kappa_3 \\ 1 & t_{i9} & t_{i9} - \kappa_2 & t_{i9} - \kappa_3 \end{pmatrix}.$$

Der für jede Person zu schätzende Parametervektor ist durch $\mathbf{b}_i = (b_{i0}, b_{i1}, b_{i2}, b_{i3})'$ gegeben.

Zur praktischen Umsetzung dieser Idee soll nun wie in Abschnitt 8.2 ein DPM-Modell mit Einstellungen, die sich im Vergleich zu 8.3 nur durch das Weglassen der festen Effekte unterscheidet, mittels der R-Funktion `BlockDPM()` berechnet werden. Die MCMC-Einstellungen und die der Hyperparameter sind in Tabelle 8.6 zusammengefasst.

MCMC	<i>Iterationen</i>	<i>Burnin</i>	<i>W</i>						
	55000	5000	50						
Fehlervarianz	a_ε	b_ε							
	0.0001	0.0001							
P-Spline	a_γ	b_γ	m	l	k				
	0.005	0.005	12	3	2				
zufällige Effekte	a_b	b_b	a_0	b_0	a_α	b_α	m_0	s_0^2	N
	0.005	0.005	0.005	0.005	2	4	0	10	100

Tabelle 8.6: Annahmen bzgl. des MCMC-Algorithmus und der Hyperparameter

Die Samplingpfade dieses Modells zeigen allerdings für alle die Steigung betreffenden zufälligen Effekte, also für b_{i1} , b_{i2} und b_{i3} , eine inakzeptable Gestalt. Exemplarisch soll diesbezüglich \mathbf{b}_i für $i = 888$ in Abbildung 8.8 visualisiert werden. Für alle anderen Individuen ergab sich meist ein ähnliches Bild.

Man sieht in Abbildung 8.8, dass, während für b_{i0} die Samplingpfade zumindest nach der 300. gespeicherten Iteration vernünftig ausschauen, sich der lineare zeitliche Effekt zwischen den Koeffizienten b_{i1} , b_{i2} und b_{i3} verlagert. Dies ist vor allem an dem Bruchpunkt bei b_{i1} und b_{i3} in der Nähe der 100. gespeicherten Iteration ersichtlich. Man erkennt auch, dass eine längerer Burn-In oder eine größere Ausdünnung keine Abhilfe schaffen würden. Offenbar kann für diese Kombination aus generellem P-Spline und individuellen TP-Splines mit DPM-Priori hinsichtlich mancher Parameter keine Konvergenz erreicht werden. Als mögliche Ursache lässt sich anführen, dass an diesen Parametern zu viele unterschiedliche Kräfte wirken. Zum einen versucht das Modell für jedes Individuum separat die individuelle Abweichung der Steigung gegenüber dem P-Spline den Daten entsprechend auf die drei Steigungsparameter b_{i1} , b_{i2} und b_{i3} zu übertragen. Zum anderen bewirkt der Clustereffekt des Dirichlet-Prozesses eine Anbindung dieses Individuums zu anderen Personen und versucht innerhalb dieses Personenkreises eine allgemeingültige Steigung zu finden. Eine ständig wechselnde Konstellation in diesem Cluster erschwert nun eine Konvergenz sämtlicher zufälliger Effekte.

Insgesamt können die betreffenden Samples nicht als unabhängige Zufallsziehungen aus der Posteriori aufgefasst werden. Darauf aufbauende Schätzungen wären ohne Wert.

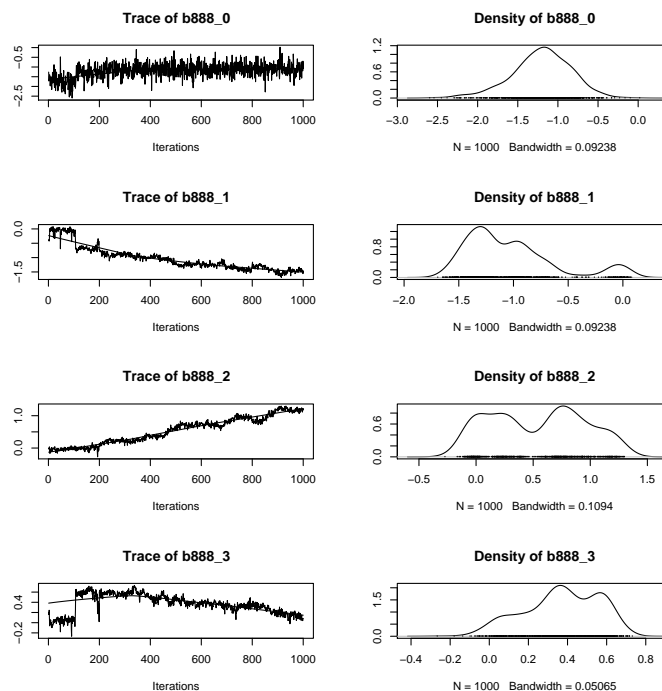


Abbildung 8.8: Samplingpfade und Dichteschätzungen zu b_i für $i = 888$

8.4 DP-Modelle

Die in Abschnitt 8.3 untersuchte Fragestellung, ob sich durch die Erweiterung von zufälligen Effekten 1. Grades auf individuelle TP-Splines 1. Grades bessere Schätzungen erzielen lassen, soll nun auch für das DP-Modell überprüft werden. Dabei ist insbesondere interessant, ob ähnliche Probleme bei den Samplingpfaden auftreten wie im Falle des DPM-Modells. Konkret wurden drei Modelle gerechnet: Allen drei ist gemein, dass sie einen P-Spline enthalten, wohingegen feste Effekte ausgeblendet werden. Sie unterscheiden sich darin, dass die zufälligen Effekte 1. Grades einmal ohne Knoten, einmal mit einem Knoten bei $ageY = 0.5$ und einmal mit zwei Knoten bei $ageY = 0.5$ und $ageY = 1$ modelliert wurden. Während das erste genauer analysiert wird, werden für letztere lediglich die Resultate angegeben. Die drei Modelle wurden mit der Funktion `BlockDP()` berechnet (vgl. Abschnitt 6.4). Die dabei gewählten Einstellungen des MCMC-Algorithmus und die Wahl der Hyperparameter für die drei DP-Modelle lassen sich in Tabelle 8.7 wiederfinden.

Im Gegensatz zu den Abschnitten 8.2 und 8.3 wird bei den DP-Modellen mit weniger Iterationen gearbeitet, da die Berechnung der DP-Modelle bei den LISA-Daten mit deutlich längeren Laufzeiten verbunden ist. Die Ursache dieses Phänomens ist wert, genauer betrachtet zu werden. Der mit Abstand aufwendigste Aufdatierungsschritt des Block-Gibbs-Samplers mit DP- bzw. DPM-Priori ist derjenige, bei dem entschieden wird, welche Person welchem Cluster zugeordnet wird. Dieser wird zudem von der Iterationsschleife umschlos-

MCMC	<i>Iterationen</i>	<i>Burnin</i>	<i>W</i>						
	33000	3000	30						
Fehlervarianz	a_ε	b_ε							
	0.0001	0.0001							
P-Spline	a_γ	b_γ	m	l	k				
	0.005	0.005	12	3	2				
zufällige Effekte	a_b	b_b	a_0	b_0	a_α	b_α	m_0	s_0^2	N
	-	-	0.005	0.005	2	4	0	10	100

Tabelle 8.7: Annahmen bzgl. des MCMC-Algorithmus und der Hyperparameter

sen, so dass drei ineinander geschachtelte Schleifen vorliegen. Wenn nun in der x -ten Iteration für die i -te Person untersucht wird, wie plausibel der Cluster h für sie ist, ist die Evaluierung einer multivariaten Normalverteilung notwendig. Diese ist im Fall des DPM-Modells mit zufälligen Effekten 1. Grades zweidimensional, da es sich um die Verteilung von $\mathbf{b}_i = (b_{i0}, b_{i1})'$ handelt. Im Falle des DP-Modells ist die multivariate Normalverteilung unabhängig von der Gestalt der zufälligen Effekte des Vektors \mathbf{y}_i , der bei den LISA-Daten in der Regel neundimensional ist. Den Funktionswert einer neundimensionalen Normalverteilung zu bestimmen, bedeutet eigentlich einen nur geringfügig größeren Rechenaufwand als für eine zweidimensionale Normalverteilung; da dies aber bei 33000 Iterationen, 2043 Individuen und 100 möglichen Clusterlokationen insgesamt 6741900000 mal gemacht wird, führt dies zu einer deutlich längeren Rechenzeit.

Das Modell mit zufälligen Effekten 1. Grades liefert alles in allem ein ähnliches Bild zu dem DPM-Modell in Abschnitt 8.3. So zeigt Abbildung 8.9, dass die geschätzten zeitlichen Effekte in etwa denen in Abbildung 8.5 entsprechen. Hinsichtlich der Clusterstruktur zeigt sich wie schon im DPM-Modell: Während sich die meisten Cluster durch verschiedene Niveaus des zeitlichen BMI-Verlaufs unterscheiden, gibt es auch eine Gruppe von Individuen deren BMI stetig ansteigt und kein lokales Maximum im ersten Lebensjahr aufweist (vgl. Abbildung 8.10). Für diese Graphik wurden wiederum die geschätzten zufälligen Effekte anhand des Varianzkriteriums der Anzahl an Clustern zugewiesen werden, die das Modell ausgegeben hat.

Den Samplingpfaden der zufälligen Effekte im DP-Modell muss hingegen besondere Aufmerksamkeit zuteil werden. Sie weisen auf ein spezifisches Problem hin: Da der Clustereffekt direkt auf die zufälligen Effekte wirkt und nicht auf den Erwartungswert des jeweiligen Effekts wie im DPM-Fall, können die Ziehungen gelegentlich eine bimodale Struktur aufweisen. Dies ist dann der Fall, wenn das sich zwischen zwei Clustern befindende \mathbf{b}_i mal dem einen und mal dem anderen Cluster zugeordnet wird. Abbildung 8.11 zeigt so ein Beispiel für b_{i0} mit $i = 2037$. Es lässt sich in so einem Fall diskutieren, inwieweit der Median oder das arithmetische Mittel der Ziehungen den zugrunde liegenden Effekt widerspiegeln. Für andere Parameter wie γ , τ^2 und σ^2 wiesen die Pfade eine tadellose Form auf.

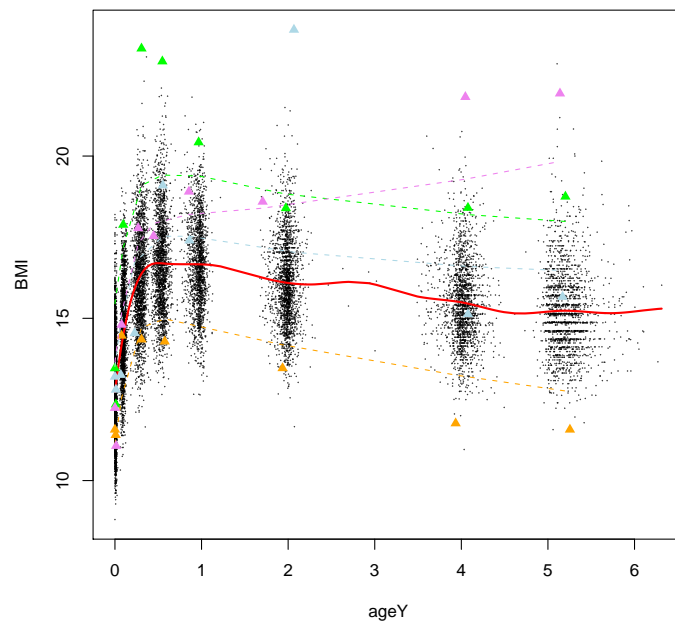
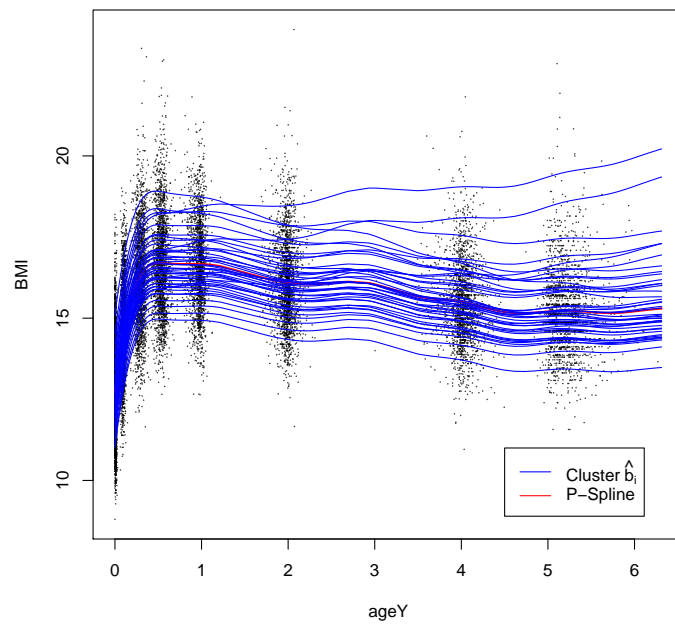


Abbildung 8.9: Geschätzter zeitlicher Effekt (DP-Modell ohne Knoten)

Abbildung 8.10: Clusterzuweisung der Individuen hinsichtlich \hat{b}_i (DP-Modell ohne Knoten)

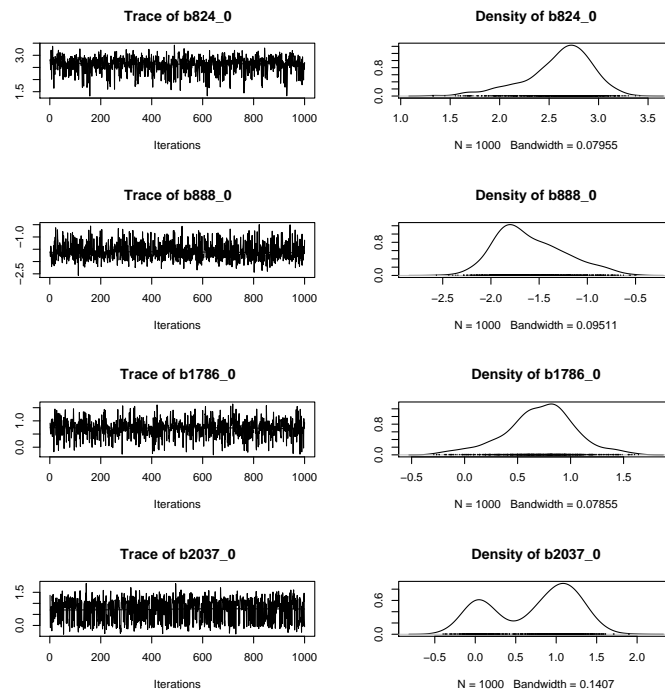


Abbildung 8.11: Samplingpfade und Dichteschätzungen des Random-Intercepts für vier ausgewählte Individuen (DP-Modell ohne Knoten)

Nun soll die eingangs gestellte Frage nach einer möglichen Verbesserung der Modellschätzung durch ein oder zwei ergänzte Knoten wieder aufgegriffen werden. Zunächst werden für das Modell mit den Knoten bei 0.5 und 1 analog zu Abschnitt 8.3 die Samplingpfade der zufälligen Effekte für $i = 888$ betrachtet, um zu vergleichen, ob die Probleme, die beim entsprechenden DPM-Modell auftraten, auch hier vorzufinden sind. Abbildung 8.12 zeigt, dass dies nicht der Fall ist. Die Pfade weisen ein zufriedenstellende Gestalt auf. Dieses Bild zeigt sich auch für alle anderen Parameter sowohl bei dem Modell mit zwei Knoten als auch bei dem mit einem Knoten. Ebenfalls lässt sich für beide Modelle festhalten, dass Probleme mit einer deutlich bimodalen Form der geschätzten Dichte der Ziehungen nicht mehr auftreten. Diese ist so gut wie immer unimodal und in den wenigen Fällen, in denen sie es nicht ist, liegen die beiden Modi sehr dicht beieinander.

Die gezogenen Werte können daher zur Schätzung der Parameter verwendet werden. Der generelle zeitliche Effekt und der der ausgewählten Individuen kann nun den Abbildungen 8.13 und 8.14 entnommen werden.

Man erkennt an den Abbildungen, dass für das Individuum 92185191 der zeitliche Verlauf des BMIs besser gefittet wird. Das erreichte Maximum des Fits ist bei beiden Modellen höher als es in Abbildung 8.9 der Fall war. Man sieht an Individuum 92185191 auch deutlich den Einfluss des zweiten Knotens bei 1: Dadurch wird ein steiler Anstieg bis $ageY = 0.5$ und ein Abstieg innerhalb $[0.5; 1]$ sichtbar, während in dem DP-Modell mit einem Knoten bei $ageY = 0.5$ diese „Spitze“ nicht annähernd so ausgeprägt ist. Den-

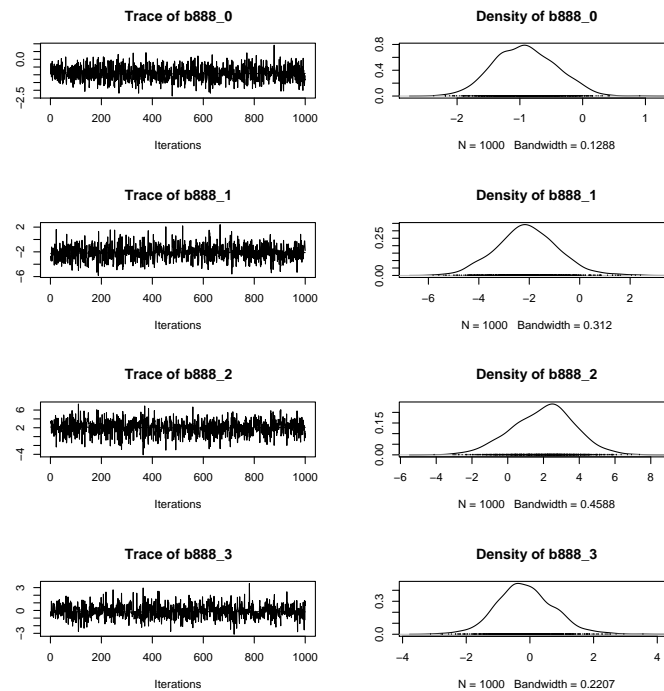


Abbildung 8.12: Samplingpfade zu b_i für $i = 888$ (DP-Modell mit Knoten bei 0.5 und 1)

noch würde man sich hierbei eine den Daten noch besser angepasste Form des zeitlichen Verlaufs wünschen; der Clustereffekt verhindert allerdings eine auf „Ausreißer“ wie Individuum 92185191 zu sehr ausgerichtete Modellierung. Dies erkennt man auch an Individuum 92189214, das deutlich unterdurchschnittliche BMI-Werte aufweist. Der entsprechende Fit passt sich nur in Maßen dieser Entwicklung an. Für das Individuum 94182011 lässt sich feststellen, dass der extrem hohe Wert des BMIs im Alter von ca. 2 Jahren den Verlauf zwischen einem Jahr und sechs Jahren deutlich beeinflusst. Während die Messungen im Alter von 4 und etwas über 5 Jahren eher für eine durchschnittliche BMI-Entwicklung sprechen, ist der Fit klar über dem P-Spline. Eine im Vergleich zu Abbildung 8.9 viel bessere Anpassung wird für das Individuum 95089461 erreicht. Der Anstieg des BMIs nach dem ersten Lebensjahr ist bei den Modellen mit einem bzw. zwei Knoten wesentlich steiler und damit den Daten angemessener.

Alles in allem stellt die Erweiterung der zufälligen Effekte auf individuelle TP-Splines einen guten Ansatz dar, wie die Schätzung der individuellen Verläufe verbessert werden kann. Hervorhebenswert ist dabei, dass dies bei DP-Modellen wesentlich besser funktioniert als bei DPM-Modellen, da bei DP-Modellen die in Abschnitt 8.3 angesprochenen Probleme bei den Samplingpfaden nicht auftraten. Dies ist bemerkenswert, wenn man bedenkt, dass sich das DP-Modell in Kapitel 7 hinsichtlich der Schätzgüte als schlechter als das DPM-Modell erwiesen hat. Offensichtlich erschwert die zusätzliche Stufe in einem DPM-Modell bei einem allgemeinen P-Spline und individuellen TP-Splines ein gutes Mischungsverhältnis für die Koeffizienten, die in ihrer Bedeutung auf den BMI eng miteinander verbunden sind.

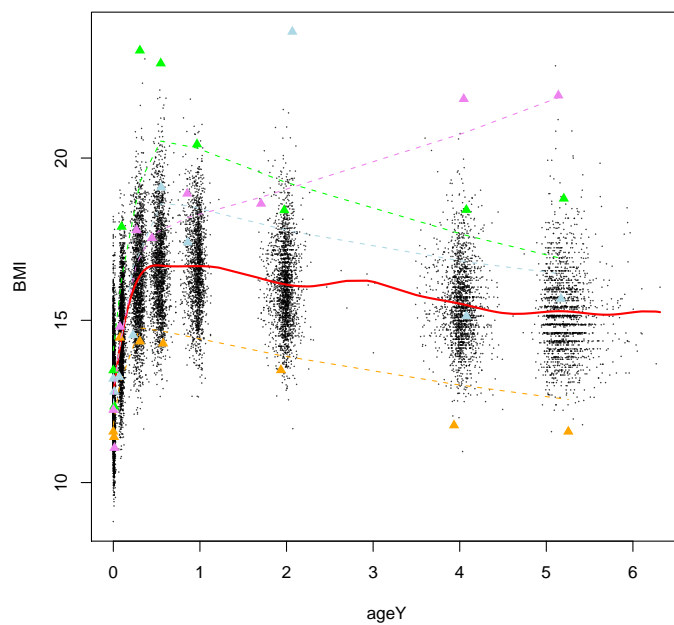


Abbildung 8.13: Geschätzter zeitlicher Effekt (DP-Modell mit einem Knoten bei 0.5)

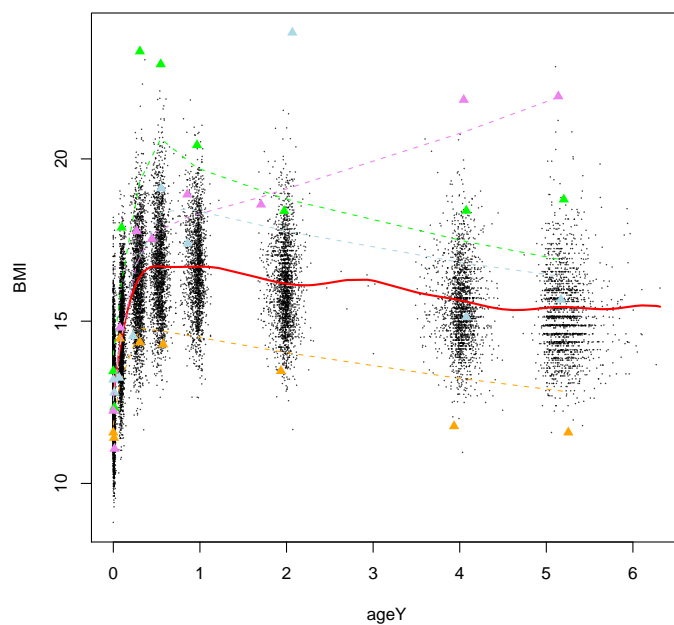


Abbildung 8.14: Geschätzter zeitlicher Effekt (DP-Modell mit Knoten bei 0.5 und 1)

9 Zusammenfassung

In dieser Diplomarbeit wurde ein additives gemischtes Modell behandelt, das sowohl den generellen zeitlichen Effekt als auch die individuellen zeitlichen Effekte auf die Zielgröße untersucht. Für Ersteres wurde ausgehend von einer nichtlinearen Datenstruktur ein P-Spline, für Zweiteres ein Dirichlet-Prozess-Modell oder ein Dirichlet-Prozess-Mischungs-Modell verwendet. Die Schätzung sämtlicher im Modell vorkommenden Parameter wurde mittels eines Block-Gibbs-Samplers vollzogen, der auf der Stick-Breaking-Repräsentation des Dirichlet-Prozesses basiert.

Grundsätzlich waren in dieser Arbeit zwei Belange von besonderem Interesse: Zum einen wurde anhand simulierter Daten untersucht, ob durch die flexiblere Modellierung der Verteilung der zufälligen Effekte bessere Schätzergebnisse die zufälligen Effekte betreffend erzielt werden können, als wenn hierfür die traditionelle Normalverteilungsannahme veranschlagt wird. Es zeigte sich, dass selbst bei mischverteilten zufälligen Effekte die Normalverteilungsannahme zu erstaunlich guten Ergebnisse führt und hinsichtlich des DPM-Modells kaum Unterschiede vorlagen. Das DP-Modell erwies sich hingegen aufgrund der implizit für die zufälligen Effekte angenommenen diskreten Verteilung als deutlich schlechter. Des Weiteren stand die Kombination eines über einen Dirichlet-Prozess modellierten linearen gemischten Modells mit einem P-Spline im Fokus. Diesbezüglich kann festgehalten werden, dass durch eine Zentrierung der zufälligen Effekte bei jedem Iterationsschritt des MCMC-Algorithmus stabile Schätzergebnisse erzielt werden können. Die Aufgabenbereiche des P-Splines und der zufälligen Effekte werden dadurch klar voneinander getrennt, selbst wenn beide Modellkomponenten den zeitlichen Effekt auf den Response beschreiben. Die zufälligen Effekte können so als die individuellen Abweichungen vom generellen, durch den P-Spline beschriebenen zeitlichen Effekt verstanden werden. Darüber hinaus kann der generelle P-Spline auch mit individuellen TP-Splines kombiniert werden. Durch zusätzliche Knoten konnten die individuellen Verläufe etwas adäquater modelliert werden. Für das DPM-Modell offenbarten sich allerdings Probleme hinsichtlich der Samplingpfade. Das DP-Modell wies diese Schwierigkeiten nicht auf und ist in diesem Fall zu bevorzugen.

Hinsichtlich des durch den Dirichlet-Prozess induzierten Clustereffekts fällt die Bewertung unterschiedlich aus. Zunächst darf die Clustereigenschaft des Dirichlet-Prozesses nicht als eine automatische Einteilung der Individuen in die wahre zugrunde liegende Struktur verstanden werden. Dies liegt nicht nur daran, dass die vom Modell ausgegebene Schätzung für die Anzahl der Cluster in hohem Maße von den gewählten Einstellungen hinsichtlich des Präzisionsparameters abhängt, sondern hat ihre Ursache vor allem in der Schwierigkeit, wie die Information der Clustereinteilung zu jedem Iterationsschritt des MCMC-Algorithmus zu einer allgemein gültigen Clustereinteilung gebündelt werden kann. Der Clustereffekt des Dirichlet-Prozesses ist vielmehr ein während des Algorithmus ablaufender Mechanismus,

der stets die Informationen mehrerer Personen zu einer Gruppe zusammenfasst, so dass letztendlich für den zufälligen Effekt eines bestimmten Individuums eine Schätzung resultiert, die nicht durch dieses Individuum allein, sondern auch durch seine „Nachbarschaft“ geprägt ist. Auf diese Weise können grundlegende Muster in den Daten herausgearbeitet werden.

Die ansprechende Clustereigenschaft des Dirichlet-Prozesses kann auch im Rahmen eines additiv gemischten Modells genutzt werden. In diesem Zusammenhang lässt sich festhalten – und das ist das wesentliche Fazit dieser Arbeit –, dass einer Kombination eines linearen gemischten Modells basierend auf Dirichlet-Prozess-Prioris mit einem P-Spline nichts im Wege steht und dass diese Form der Kombination bei longitudinalen Daten, die durch einen nichtlinearen zeitlichen Effekt charakterisiert sind, einen leistungsfähigen Ansatz darstellt.

A Beweise

A.1 Akzeptanzwahrscheinlichkeit beim Gibbs-Sampler

Für den Nachweis, dass die Akzeptanzwahrscheinlichkeit beim Gibbs-Sampler 1 beträgt (vgl. Abschnitt 2.3), sei folgende Situation gegeben:

Es werde aus der vollständig bedingten Dichte $p(\theta_s | \theta_{-s}, \mathbf{y})$ ein Wert θ_s^* gezogen. Zusammen mit den aktuellen Zuständen der anderen Blöcke wird ein Zustand $\theta^* := (\theta_1^{(t)}, \dots, \theta_{s-1}^{(t)}, \theta_s^*, \theta_{s+1}^{(t-1)}, \dots, \theta_S^{(t-1)})'$ vorgeschlagen. Gegenwärtig befindet sich die Markov-Kette noch im Zustand $\theta^\circ := (\theta_1^{(t)}, \dots, \theta_{s-1}^{(t)}, \theta_s^{(t-1)}, \theta_{s+1}^{(t-1)}, \dots, \theta_S^{(t-1)})'$.

Die Akzeptanzwahrscheinlichkeit lautet nun:

$$\alpha(\theta^* | \theta^\circ) = \min \left\{ \frac{p(\theta^* | \mathbf{y}) q(\theta^\circ | \theta^*)}{p(\theta^\circ | \mathbf{y}) q(\theta^* | \theta^\circ)}, 1 \right\} = \min \left\{ \frac{p(\theta_{-s}^* | \mathbf{y}) p(\theta_s^* | \theta_{-s}^*, \mathbf{y}) p(\theta_s^\circ | \theta_{-s}^\circ, \mathbf{y})}{p(\theta_{-s}^\circ | \mathbf{y}) p(\theta_s^\circ | \theta_{-s}^\circ, \mathbf{y}) p(\theta_s^* | \theta_{-s}^*, \mathbf{y})}, 1 \right\} = 1.$$

Der letzte Schritt gilt wegen $\theta_{-s}^* = \theta_{-s}^\circ$.

A.2 Erwartungswert des Dirichlet-Prozesses

Gemäß der Definition des Dirichlet-Prozesses $G \sim DP(\alpha_0 G_0)$ gilt für jede endliche messbare Partition $\{A_1, \dots, A_m\}$ von Θ :

$$(G(A_1), \dots, G(A_m))' \sim \text{Dir}(\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_m)).$$

Somit gilt für alle $j = 1, \dots, m$ (vgl. hierzu die Eigenschaften der Dirichlet-Verteilung in Abschnitt 4.1):

$$E(G(A_j)) = \frac{\alpha_0 G_0(A_j)}{\sum_{l=1}^m \alpha_0 G_0(A_l)} = G_0(A_j).$$

A.3 3. Haupteigenschaft des Dirichlet-Prozesses

Für den Nachweis der in Abschnitt 4.2 formulierten 3. Haupteigenschaft des Dirichlet-Prozesses wird die Posteriori $G|\theta_1, \dots, \theta_n$ bei folgenden Annahmen bestimmt:

$$\begin{aligned}\theta_i|G &\stackrel{i.i.d.}{\sim} G & \forall i = 1, \dots, n, \\ G &\sim DP(\alpha_0 G_0).\end{aligned}$$

Für die Priori $G \sim DP(\alpha_0 G_0)$ gilt gemäß der Definition des Dirichlet-Prozesses für jede endliche messbare Partition $\{A_1, \dots, A_m\}$ von Θ :

$$(G(A_1), \dots, G(A_m))' \sim Dir(\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_m)).$$

Ebenso lässt sich die Posteriori-Verteilung $G|\theta_1, \dots, \theta_n$ durch die Verteilung des Wahrscheinlichkeitsvektors $(G(A_1), \dots, G(A_m))'|\theta_1, \dots, \theta_n$ identifizieren. Deren Dichte lässt sich folgendermaßen berechnen:

$$\begin{aligned}p(G(A_1), \dots, G(A_m)|\theta_1, \dots, \theta_n) &\propto p(\theta_1, \dots, \theta_n|G(A_1), \dots, G(A_m)) p(G(A_1), \dots, G(A_m)) \\ &\propto \prod_{i=1}^n p(\theta_i|G(A_1), \dots, G(A_m)) \prod_{j=1}^m G(A_j)^{\alpha_0 G_0(A_j)-1} \\ &\propto \prod_{j=1}^m G(A_j)^{n_j} \prod_{j=1}^m G(A_j)^{\alpha_0 G_0(A_j)-1} \\ &= \prod_{j=1}^m G(A_j)^{n_j + \alpha_0 G_0(A_j)-1}.\end{aligned}$$

\Rightarrow Für jede endliche messbare Partition $\{A_1, \dots, A_m\}$ von Θ gilt:

$$(G(A_1), \dots, G(A_m))'|\theta_1, \dots, \theta_n \sim Dir(n_1 + \alpha_0 G_0(A_1), \dots, n_m + \alpha_0 G_0(A_m)).$$

$\Rightarrow G|\theta_1, \dots, \theta_n \sim DP(\alpha_0^* G_0^*)$, wobei sich die Parameter wie folgt bestimmen lassen:

$$\begin{aligned}\alpha_0^* &= \sum_{j=1}^m (n_j + \alpha_0 G_0(A_j)) = n + \alpha_0, \\ G_0^*(A_j) &= \frac{1}{n + \alpha_0} (n_j + \alpha_0 G_0(A_j)) = \frac{1}{n + \alpha_0} (\sum_{i=1}^n \delta_{\theta_i}(A_j) + \alpha_0 G_0(A_j)). \\ \Rightarrow G_0^* &= \frac{1}{n + \alpha_0} \sum_{i=1}^n \delta_{\theta_i} + \frac{\alpha_0}{n + \alpha_0} G_0.\end{aligned}$$

B Bestimmung der vollständig bedingten Dichten

Im Folgenden werden sämtliche vollständig bedingten Dichten (vgl. (2.4)) hergeleitet, die im Rahmen des additiven gemischten Modell verwendet wurden. Als Grundlage dienen die Modelle (6.3) bzw. (6.2). Auf den Zentrierungsschritt soll dabei verzichtet werden. Für die Bestimmung der Likelihood werde zunächst folgende notationelle Vereinfachung eingeführt:

$$\boldsymbol{\mu}_i := \mathbf{X}_i\boldsymbol{\beta} + \mathbf{B}_i\boldsymbol{\gamma} + \mathbf{Z}_i\mathbf{b}_i.$$

Damit lautet die Likelihood:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}, \sigma^2) &= \prod_{i=1}^n f(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}_i, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{0.5n_i}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}_i - \boldsymbol{\mu}_i)'(\mathbf{y}_i - \boldsymbol{\mu}_i)\right) \\ &\propto \frac{1}{(\sigma^2)^{0.5n_d}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu}_i)'(\mathbf{y}_i - \boldsymbol{\mu}_i)\right). \end{aligned}$$

Des Weiteren werde die Notation beim Aufdatieren der Parameter $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ und \mathbf{b} dadurch knapp gehalten, dass stets mit dem jeweiligen Arbeitsresponse $\tilde{\mathbf{y}}$ gearbeitet wird. Ehe in den folgenden Abschnitten die einzelnen Modellkomponenten und ihre Parameter behandelt werden, soll zunächst die vollständig bedingte Dichte der Fehlervarianz, die mit allen Modellkomponenten in Verbindung steht, hergeleitet werden.

Dichte der Priori-Verteilung für σ^2 :

$$p(\sigma^2) \propto \frac{1}{(\sigma^2)^{a_\varepsilon+1}} \exp\left(-\frac{b_\varepsilon}{\sigma^2}\right).$$

Vollständig bedingte Dichte für σ^2 :

$$\begin{aligned} p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}, \mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}, \sigma^2) p(\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{0.5n_d}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu}_i)'(\mathbf{y}_i - \boldsymbol{\mu}_i)\right) \frac{1}{(\sigma^2)^{a_\varepsilon+1}} \exp\left(-\frac{b_\varepsilon}{\sigma^2}\right) \\ &= \frac{1}{(\sigma^2)^{a_\varepsilon+0.5n_d+1}} \exp\left(-\frac{1}{\sigma^2} \left(b_\varepsilon + \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu}_i)'(\mathbf{y}_i - \boldsymbol{\mu}_i)\right)\right). \end{aligned}$$

B.1 P-Spline

Dichten der Priori-Verteilungen:

$$\begin{aligned} p(\gamma|\tau^2) &\propto \frac{1}{(\tau^2)^{0.5rg(\mathbf{K})}} \exp\left(-\frac{1}{2\tau^2}\gamma'\mathbf{K}\gamma\right), \\ p(\tau^2) &\propto \frac{1}{(\tau^2)^{a_\gamma+1}} \exp\left(-\frac{b_\gamma}{\tau^2}\right). \end{aligned}$$

Vollständig bedingte Dichten:

$$\begin{aligned} p(\gamma|\tau^2, \beta, \mathbf{b}, \mathbf{y}, \sigma^2) &\propto p(\mathbf{y}|\beta, \gamma, \mathbf{b}, \sigma^2) p(\gamma|\tau^2) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (\tilde{\mathbf{y}}_i - \mathbf{B}_i\gamma)'(\tilde{\mathbf{y}}_i - \mathbf{B}_i\gamma)\right) \exp\left(-\frac{1}{2\tau^2}\gamma'\mathbf{K}\gamma\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (\gamma' \mathbf{B}_i' \mathbf{B}_i \gamma - 2\gamma' \mathbf{B}_i' \tilde{\mathbf{y}}_i)\right) \exp\left(-\frac{1}{2\tau^2}\gamma'\mathbf{K}\gamma\right) \\ &= \exp\left(-\frac{1}{2} \left(\gamma' \left(\frac{1}{\tau^2} \mathbf{K} + \frac{1}{\sigma^2} \sum_{i=1}^n \mathbf{B}_i' \mathbf{B}_i \right) \gamma - 2\gamma' \left(\frac{1}{\sigma^2} \sum_{i=1}^n \mathbf{B}_i' \tilde{\mathbf{y}}_i \right) \right)\right) \\ &= \exp\left(-\frac{1}{2} \left(\gamma' \underbrace{\left(\frac{1}{\tau^2} \mathbf{K} + \frac{1}{\sigma^2} \mathbf{B}' \mathbf{B} \right)}_{=:\boldsymbol{\Sigma}_\gamma^{*-1}} \gamma - 2\gamma' \left(\frac{1}{\sigma^2} \mathbf{B}' \tilde{\mathbf{y}} \right) \right)\right) \\ &= \exp\left(-\frac{1}{2} \left(\gamma' \boldsymbol{\Sigma}_\gamma^{*-1} \gamma - 2\gamma' \boldsymbol{\Sigma}_\gamma^{*-1} \underbrace{\boldsymbol{\Sigma}_\gamma^* \left(\frac{1}{\sigma^2} \mathbf{B}' \tilde{\mathbf{y}} \right)}_{\boldsymbol{\mu}_\gamma^*} \right)\right). \end{aligned}$$

$$\begin{aligned} p(\tau^2|\gamma) &\propto p(\gamma|\tau^2) p(\tau^2) \\ &\propto \frac{1}{(\tau^2)^{0.5rg(\mathbf{K})}} \exp\left(-\frac{1}{2\tau^2}\gamma'\mathbf{K}\gamma\right) \frac{1}{(\tau^2)^{a_\gamma+1}} \exp\left(-\frac{b_\gamma}{\tau^2}\right) \\ &= \frac{1}{(\tau^2)^{a_\gamma+0.5rg(\mathbf{K})+1}} \exp\left(-\frac{1}{\tau^2}(b_\gamma + 0.5\gamma'\mathbf{K}\gamma)\right). \end{aligned}$$

B.2 Lineares Modell

Dichten der Priori-Verteilungen:

$$\begin{aligned}
 p(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) &\propto |\boldsymbol{\Sigma}_\beta|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)' \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right), \\
 p(\mu_{\beta_r}) &\propto \exp\left(-\frac{1}{2s_{\beta_r}^2}(\mu_{\beta_r} - m_{\beta_r})^2\right), \\
 p(\sigma_{\beta_r}^2) &\propto \frac{1}{(\sigma_{\beta_r}^2)^{a_{\beta_r}+1}} \exp\left(-\frac{b_{\beta_r}}{\sigma_{\beta_r}^2}\right).
 \end{aligned}$$

Vollständig bedingte Dichten:

$$\begin{aligned}
 p(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \boldsymbol{\gamma}, \mathbf{b}, \mathbf{y}, \sigma^2) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}, \sigma^2) p(\boldsymbol{\beta}|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \\
 &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (\tilde{\mathbf{y}}_i - \mathbf{X}_i \boldsymbol{\beta})' (\tilde{\mathbf{y}}_i - \mathbf{X}_i \boldsymbol{\beta})\right) \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)' \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)\right) \\
 &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (\boldsymbol{\beta}' \mathbf{X}_i' \mathbf{X}_i \boldsymbol{\beta} - 2\boldsymbol{\beta}' \mathbf{X}_i' \tilde{\mathbf{y}}_i)\right) \exp\left(-\frac{1}{2}(\boldsymbol{\beta}' \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\beta} - 2\boldsymbol{\beta}' \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta)\right) \\
 &= \exp\left(-\frac{1}{2} \left(\boldsymbol{\beta}' \left(\boldsymbol{\Sigma}_\beta^{-1} + \frac{1}{\sigma^2} \sum_{i=1}^n \mathbf{X}_i' \mathbf{X}_i \right) \boldsymbol{\beta} - 2\boldsymbol{\beta}' \left(\boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta + \frac{1}{\sigma^2} \sum_{i=1}^n \mathbf{X}_i' \tilde{\mathbf{y}}_i \right) \right)\right) \\
 &= \exp\left(-\frac{1}{2} \left(\boldsymbol{\beta}' \underbrace{\left(\boldsymbol{\Sigma}_\beta^{-1} + \frac{1}{\sigma^2} \mathbf{X}' \mathbf{X} \right)}_{=:\boldsymbol{\Sigma}_\beta^{*-1}} \boldsymbol{\beta} - 2\boldsymbol{\beta}' \left(\boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta + \frac{1}{\sigma^2} \mathbf{X}' \tilde{\mathbf{y}} \right) \right)\right) \\
 &= \exp\left(-\frac{1}{2} \left(\boldsymbol{\beta}' \boldsymbol{\Sigma}_\beta^{*-1} \boldsymbol{\beta} - 2\boldsymbol{\beta}' \underbrace{\boldsymbol{\Sigma}_\beta^{*-1} \boldsymbol{\Sigma}_\beta^* \left(\boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta + \frac{1}{\sigma^2} \mathbf{X}' \tilde{\mathbf{y}} \right)}_{\boldsymbol{\mu}_\beta^*} \right)\right).
 \end{aligned}$$

$$\begin{aligned}
p(\mu_{\beta_r} | \sigma_{\beta_r}^2, \beta_r) &\propto p(\beta_r | \mu_{\beta_r}, \sigma_{\beta_r}^2) p(\mu_{\beta_r}) \\
&\propto \exp\left(-\frac{1}{2\sigma_{\beta_r}^2}(\beta_r - \mu_{\beta_r})^2\right) \exp\left(-\frac{1}{2s_{\beta_r}^2}(\mu_{\beta_r} - m_{\beta_r})^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_{\beta_r}^2}\beta_r^2 - 2\frac{1}{\sigma_{\beta_r}^2}\beta_r\mu_{\beta_r} + \frac{1}{\sigma_{\beta_r}^2}\mu_{\beta_r}^2 + \frac{1}{s_{\beta_r}^2}\mu_{\beta_r}^2 - 2\frac{1}{s_{\beta_r}^2}\mu_{\beta_r}m_{\beta_r} + \frac{1}{s_{\beta_r}^2}m_{\beta_r}^2\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\mu_{\beta_r}^2\left(\frac{1}{\sigma_{\beta_r}^2} + \frac{1}{s_{\beta_r}^2}\right) - 2\mu_{\beta_r}\left(\frac{\beta_r}{\sigma_{\beta_r}^2} + \frac{m_{\beta_r}}{s_{\beta_r}^2}\right)\right)\right) \\
&= \exp\left(-\frac{1}{2\left(\frac{1}{\sigma_{\beta_r}^2} + \frac{1}{s_{\beta_r}^2}\right)^{-1}}\left(\mu_{\beta_r}^2 - 2\mu_{\beta_r}\left(\frac{1}{\sigma_{\beta_r}^2} + \frac{1}{s_{\beta_r}^2}\right)^{-1}\left(\frac{\beta_r}{\sigma_{\beta_r}^2} + \frac{m_{\beta_r}}{s_{\beta_r}^2}\right)\right)\right).
\end{aligned}$$

$$\begin{aligned}
p(\sigma_{\beta_r}^2 | \mu_{\beta_r}, \beta_r) &\propto p(\beta_r | \mu_{\beta_r}, \sigma_{\beta_r}^2) p(\sigma_{\beta_r}^2) \\
&\propto \frac{1}{(\sigma_{\beta_r}^2)^{0.5}} \exp\left(-\frac{1}{2\sigma_{\beta_r}^2}(\beta_r - \mu_{\beta_r})^2\right) \frac{1}{(\sigma_{\beta_r}^2)^{a_\beta+1}} \exp\left(-\frac{b_\beta}{\sigma_{\beta_r}^2}\right) \\
&= \frac{1}{(\sigma_{\beta_r}^2)^{a_\beta+0.5+1}} \exp\left(-\frac{1}{\sigma_{\beta_r}^2}\left(b_\beta + \frac{1}{2}(\beta_r - \mu_{\beta_r})^2\right)\right).
\end{aligned}$$

B.3 Lineares gemischtes Modell mit DPM-Priori

Dichten der Priori-Verteilungen:

$$\begin{aligned}
p(\mathbf{b}_i | \boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b) &\propto |\boldsymbol{\Sigma}_b|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\theta}_i)' \boldsymbol{\Sigma}_b^{-1}(\mathbf{b}_i - \boldsymbol{\theta}_i)\right), \\
p(\sigma_{b_r}^2) &\propto \frac{1}{(\sigma_{b_r}^2)^{a_b+1}} \exp\left(-\frac{b_b}{\sigma_{b_r}^2}\right), \\
p(\boldsymbol{\theta}_i | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &\propto |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_0)\right), \\
p(\boldsymbol{\phi}_h | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &\propto |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\phi}_h - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\phi}_h - \boldsymbol{\mu}_0)\right), \\
p(\mu_{0_r}) &\propto \exp\left(-\frac{1}{2s_{0_r}^2}(\mu_{0_r} - m_{0_r})^2\right), \\
p(\sigma_{0_r}^2) &\propto \frac{1}{(\sigma_{0_r}^2)^{a_0+1}} \exp\left(-\frac{b_0}{\sigma_{0_r}^2}\right), \\
p(\alpha_0) &\propto \alpha_0^{a_\alpha-1} \exp(-b_\alpha \alpha_0).
\end{aligned}$$

Weitere notwendige Dichten:

$$\begin{aligned}
 p(\mathbf{c}|\boldsymbol{\pi}) &= p(n_1, \dots, n_N|\boldsymbol{\pi}) \propto \prod_{h=1}^N \pi_h^{n_h} = \prod_{h=1}^N \left(V_h \prod_{l=1}^{h-1} (1 - V_l) \right)^{n_h} = \\
 &= \prod_{h=1}^N V_h^{n_h} \cdot \prod_{h=1}^N \prod_{l=1}^{h-1} (1 - V_l)^{n_h} = \prod_{h=1}^{N-1} V_h^{n_h} \cdot \prod_{h=1}^{N-1} (1 - V_h)^{\sum_{l=h+1}^N n_l}.
 \end{aligned}$$

$$\begin{aligned}
 p(\boldsymbol{\pi}|\alpha_0) &= p(V_1, \dots, V_{N-1}|\alpha_0) = \prod_{h=1}^{N-1} p(V_h) = \prod_{h=1}^{N-1} \frac{1}{B(1, \alpha_0)} (1 - V_h)^{\alpha_0 - 1} = \\
 &= \prod_{h=1}^{N-1} \frac{\Gamma(1 + \alpha_0)}{\Gamma(1)\Gamma(\alpha_0)} (1 - V_h)^{\alpha_0 - 1} = \prod_{h=1}^{N-1} \alpha_0 (1 - V_h)^{\alpha_0 - 1}.
 \end{aligned}$$

Vollständig bedingte Dichten:

$$\begin{aligned}
 p(\mathbf{b}_i|\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{y}, \sigma^2) &\propto p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}_i, \sigma^2) p(\mathbf{b}_i|\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_b) \\
 &\propto \exp\left(-\frac{1}{2\sigma^2}(\tilde{\mathbf{y}}_i - \mathbf{Z}_i \mathbf{b}_i)'(\tilde{\mathbf{y}}_i - \mathbf{Z}_i \mathbf{b}_i)\right) \exp\left(-\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\theta}_i)' \boldsymbol{\Sigma}_b^{-1}(\mathbf{b}_i - \boldsymbol{\theta}_i)\right) \\
 &\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{b}_i' \mathbf{Z}_i' \mathbf{Z}_i \mathbf{b}_i - 2\mathbf{b}_i' \mathbf{Z}_i' \tilde{\mathbf{y}}_i)\right) \exp\left(-\frac{1}{2}(\mathbf{b}_i' \boldsymbol{\Sigma}_b^{-1} \mathbf{b}_i - 2\mathbf{b}_i' \boldsymbol{\Sigma}_b^{-1} \boldsymbol{\theta}_i)\right) \\
 &= \exp\left(-\frac{1}{2}\left(\underbrace{\mathbf{b}_i' \left(\boldsymbol{\Sigma}_b^{-1} + \frac{1}{\sigma^2} \mathbf{Z}_i' \mathbf{Z}_i\right) \mathbf{b}_i}_{=: \boldsymbol{\Sigma}_b^{*-1}} - 2\mathbf{b}_i' \left(\boldsymbol{\Sigma}_b^{-1} \boldsymbol{\theta}_i + \frac{1}{\sigma^2} \mathbf{Z}_i' \tilde{\mathbf{y}}_i\right)\right)\right) \\
 &= \exp\left(-\frac{1}{2}\left(\mathbf{b}_i' \boldsymbol{\Sigma}_b^{*-1} \mathbf{b}_i - 2\mathbf{b}_i' \boldsymbol{\Sigma}_b^{*-1} \underbrace{\boldsymbol{\Sigma}_b^* \left(\boldsymbol{\Sigma}_b^{-1} \boldsymbol{\theta}_i + \frac{1}{\sigma^2} \mathbf{Z}_i' \tilde{\mathbf{y}}_i\right)}_{\boldsymbol{\mu}_b^*}\right)\right).
 \end{aligned}$$

$$\begin{aligned}
 p(\sigma_{b_r}^2|\boldsymbol{\theta}, \mathbf{b}) &\propto \left(\prod_{i=1}^n p(b_{i_r}|\theta_{i_r}, \sigma_{b_r}^2)\right) p(\sigma_{b_r}^2) \\
 &\propto \frac{1}{(\sigma_{b_r}^2)^{0.5n}} \exp\left(-\frac{1}{2\sigma_{b_r}^2} \sum_{i=1}^n (b_{i_r} - \theta_{i_r})^2\right) \frac{1}{(\sigma_{b_r}^2)^{a_b+1}} \exp\left(-\frac{b_b}{\sigma_{b_r}^2}\right) \\
 &= \frac{1}{(\sigma_{b_r}^2)^{a_b+0.5n+1}} \exp\left(-\frac{1}{\sigma_{b_r}^2} \left(b_b + \frac{1}{2} \sum_{i=1}^n (b_{i_r} - \theta_{i_r})^2\right)\right).
 \end{aligned}$$

$$\begin{aligned}
p(\mu_{0_r} | \sigma_{0_r}^2, \boldsymbol{\theta}) &\propto \left(\prod_{i=1}^n p(\theta_{i_r} | \mu_{0_r}, \sigma_{0_r}^2) \right) p(\mu_{0_r}) \\
&\propto \exp \left(-\frac{1}{2\sigma_{0_r}^2} \sum_{i=1}^n (\theta_{i_r} - \mu_{0_r})^2 \right) \exp \left(-\frac{1}{2s_{0_r}^2} (\mu_{0_r} - m_{0_r})^2 \right) \\
&\propto \exp \left(-\frac{1}{2\sigma_{0_r}^2} n(\bar{\theta}_r - \mu_{0_r})^2 \right) \exp \left(-\frac{1}{2s_{0_r}^2} (\mu_{0_r} - m_{0_r})^2 \right) \\
&= \exp \left(-\frac{1}{2} \left(\frac{n}{\sigma_{0_r}^2} \bar{\theta}_r^2 - 2\frac{n}{\sigma_{0_r}^2} \bar{\theta}_r \mu_{0_r} + \frac{n}{\sigma_{0_r}^2} \mu_{0_r}^2 + \frac{1}{s_{0_r}^2} \mu_{0_r}^2 - 2\frac{1}{s_{0_r}^2} \mu_{0_r} m_{0_r} + \frac{1}{s_{0_r}^2} m_{0_r}^2 \right) \right) \\
&\propto \exp \left(-\frac{1}{2} \left(\mu_{0_r}^2 \left(\frac{n}{\sigma_{0_r}^2} + \frac{1}{s_{0_r}^2} \right) - 2\mu_{0_r} \left(\frac{n}{\sigma_{0_r}^2} \bar{\theta}_r + \frac{m_{0_r}}{s_{0_r}^2} \right) \right) \right) \\
&= \exp \left(-\frac{1}{2 \left(\frac{n}{\sigma_{0_r}^2} + \frac{1}{s_{0_r}^2} \right)^{-1}} \left(\mu_{0_r}^2 - 2\mu_{0_r} \left(\frac{n}{\sigma_{0_r}^2} + \frac{1}{s_{0_r}^2} \right)^{-1} \left(\frac{n}{\sigma_{0_r}^2} \bar{\theta}_r + \frac{m_{0_r}}{s_{0_r}^2} \right) \right) \right).
\end{aligned}$$

$$\begin{aligned}
p(\sigma_{0_r}^2 | \mu_{0_r}, \boldsymbol{\theta}) &\propto \left(\prod_{i=1}^n p(\theta_{i_r} | \mu_{0_r}, \sigma_{0_r}^2) \right) p(\sigma_{0_r}^2) \\
&\propto \frac{1}{(\sigma_{0_r}^2)^{0.5n}} \exp \left(-\frac{1}{2\sigma_{0_r}^2} \sum_{i=1}^n (\theta_{i_r} - \mu_{0_r})^2 \right) \frac{1}{(\sigma_{0_r}^2)^{a_0+1}} \exp \left(-\frac{b_0}{\sigma_{0_r}^2} \right) \\
&= \frac{1}{(\sigma_{0_r}^2)^{a_0+0.5n+1}} \exp \left(-\frac{1}{\sigma_{0_r}^2} \left(b_0 + \frac{1}{2} \sum_{i=1}^n (\theta_{i_r} - \mu_{0_r})^2 \right) \right).
\end{aligned}$$

$$\begin{aligned}
p(\phi_{h_r} | \sigma_{b_r}^2, \mu_{0_r}, \sigma_{0_r}^2, \mathbf{b}, \mathbf{c}) &\propto \left(\prod_{i:c_i=h} p(b_{i_r} | \phi_{h_r}, \sigma_{b_r}^2) \right) p(\phi_{h_r}) \\
&\propto \exp \left(-\frac{1}{2\sigma_{b_r}^2} \sum_{i:c_i=h} (b_{i_r} - \phi_{h_r})^2 \right) \exp \left(-\frac{1}{2\sigma_{0_r}^2} (\phi_{h_r} - \mu_{0_r})^2 \right) \\
&\propto \exp \left(-\frac{1}{2\sigma_{b_r}^2} n_h (\bar{b}_{r,h} - \phi_{h_r})^2 \right) \exp \left(-\frac{1}{2\sigma_{0_r}^2} (\phi_{h_r} - \mu_{0_r})^2 \right) \\
&= \exp \left(-\frac{1}{2} \left(\frac{n_h}{\sigma_{b_r}^2} \bar{b}_{r,h}^2 - 2\frac{n_h}{\sigma_{b_r}^2} \bar{b}_{r,h} \phi_{h_r} + \frac{n_h}{\sigma_{b_r}^2} \phi_{h_r}^2 + \frac{1}{\sigma_{0_r}^2} \phi_{h_r}^2 - 2\frac{1}{\sigma_{0_r}^2} \phi_{h_r} \mu_{0_r} + \frac{1}{\sigma_{0_r}^2} \mu_{0_r}^2 \right) \right) \\
&\propto \exp \left(-\frac{1}{2} \left(\phi_{h_r}^2 \left(\frac{n_h}{\sigma_{b_r}^2} + \frac{1}{\sigma_{0_r}^2} \right) - 2\phi_{h_r} \left(\frac{n_h}{\sigma_{b_r}^2} \bar{b}_{r,h} + \frac{\mu_{0_r}}{\sigma_{0_r}^2} \right) \right) \right) \\
&= \exp \left(-\frac{1}{2 \left(\frac{n_h}{\sigma_{b_r}^2} + \frac{1}{\sigma_{0_r}^2} \right)^{-1}} \left(\phi_{h_r}^2 - 2\phi_{h_r} \left(\frac{n_h}{\sigma_{b_r}^2} + \frac{1}{\sigma_{0_r}^2} \right)^{-1} \left(\frac{n_h}{\sigma_{b_r}^2} \bar{b}_{r,h} + \frac{\mu_{0_r}}{\sigma_{0_r}^2} \right) \right) \right).
\end{aligned}$$

$$P(c_i = h | \boldsymbol{\pi}, \boldsymbol{\phi}, \mathbf{b}_i, \boldsymbol{\Sigma}_b) \propto p(\mathbf{b}_i | \boldsymbol{\phi}_h, \boldsymbol{\Sigma}_b) \pi_h \propto |\boldsymbol{\Sigma}_b|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\mathbf{b}_i - \boldsymbol{\phi}_h)' \boldsymbol{\Sigma}_b^{-1} (\mathbf{b}_i - \boldsymbol{\phi}_h) \right) \pi_h.$$

$$\begin{aligned}
 p(\boldsymbol{\pi}|\mathbf{c}, \alpha_0) &\propto p(\mathbf{c}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\alpha_0) \\
 &\propto \prod_{h=1}^{N-1} V_h^{n_h} (1 - V_h)^{\sum_{l=h+1}^N n_l} \prod_{h=1}^{N-1} (1 - V_h)^{\alpha_0 - 1} \\
 &\propto \prod_{h=1}^{N-1} V_h^{n_h} (1 - V_h)^{\alpha_0 + \sum_{l=h+1}^N n_l - 1}.
 \end{aligned}$$

$$\begin{aligned}
 p(\alpha_0|\boldsymbol{\pi}) &\propto p(\boldsymbol{\pi}|\alpha_0) p(\alpha_0) \\
 &\propto \prod_{h=1}^{N-1} \alpha_0 (1 - V_h)^{\alpha_0 - 1} \alpha_0^{a_\alpha - 1} \exp(-b_\alpha \alpha_0) \\
 &\propto \alpha_0^{N-1+a_\alpha-1} \exp\left(\log\left(\prod_{h=1}^{N-1} (1 - V_h)^{\alpha_0}\right)\right) \exp(-b_\alpha \alpha_0) \\
 &= \alpha_0^{N-1+a_\alpha-1} \exp\left(\sum_{h=1}^{N-1} \alpha_0 \log(1 - V_h) - b_\alpha \alpha_0\right) \\
 &= \alpha_0^{N-1+a_\alpha-1} \exp\left(-\alpha_0 \left(b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h)\right)\right).
 \end{aligned}$$

B.4 Lineares gemischtes Modell mit DP-Priori

Dichten der Priori-Verteilungen:

$$\begin{aligned}
 p(\boldsymbol{\phi}_h|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &\propto |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\phi}_h - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\phi}_h - \boldsymbol{\mu}_0)\right), \\
 p(\mu_{0_r}) &\propto \exp\left(-\frac{1}{2s_{0_r}^2}(\mu_{0_r} - m_{0_r})^2\right), \\
 p(\sigma_{0_r}^2) &\propto \frac{1}{(\sigma_{0_r}^2)^{a_0+1}} \exp\left(-\frac{b_0}{\sigma_{0_r}^2}\right), \\
 p(\alpha_0) &\propto \alpha_0^{a_\alpha-1} \exp(-b_\alpha \alpha_0).
 \end{aligned}$$

Vollständig bedingte Dichten:

$$\begin{aligned}
 p(\sigma_{0_r}^2|\mu_{0_r}, \mathbf{b}) &\propto \left(\prod_{i=1}^n p(b_{i_r}|\mu_{0_r}, \sigma_{0_r}^2)\right) p(\sigma_{0_r}^2) \\
 &\propto \frac{1}{(\sigma_{0_r}^2)^{0.5n}} \exp\left(-\frac{1}{2\sigma_{0_r}^2} \sum_{i=1}^n (b_{i_r} - \mu_{0_r})^2\right) \frac{1}{(\sigma_{0_r}^2)^{a_0+1}} \exp\left(-\frac{b_0}{\sigma_{0_r}^2}\right) \\
 &= \frac{1}{(\sigma_{0_r}^2)^{a_0+0.5n+1}} \exp\left(-\frac{1}{\sigma_{0_r}^2} \left(b_0 + \frac{1}{2} \sum_{i=1}^n (b_{i_r} - \mu_{0_r})^2\right)\right).
 \end{aligned}$$

$$\begin{aligned}
p(\mu_{0r} | \sigma_{0r}^2, \mathbf{b}) &\propto \left(\prod_{i=1}^n p(b_{ir} | \mu_{0r}, \sigma_{0r}^2) \right) p(\mu_{0r}) \\
&\propto \exp \left(-\frac{1}{2\sigma_{0r}^2} \sum_{i=1}^n (b_{ir} - \mu_{0r})^2 \right) \exp \left(-\frac{1}{2s_{0r}^2} (\mu_{0r} - m_{0r})^2 \right) \\
&\propto \exp \left(-\frac{1}{2\sigma_{0r}^2} n(\bar{b}_r - \mu_{0r})^2 \right) \exp \left(-\frac{1}{2s_{0r}^2} (\mu_{0r} - m_{0r})^2 \right) \\
&= \exp \left(-\frac{1}{2} \left(\frac{n}{\sigma_{0r}^2} \bar{b}_r^2 - 2\frac{n}{\sigma_{0r}^2} \bar{b}_r \mu_{0r} + \frac{n}{\sigma_{0r}^2} \mu_{0r}^2 + \frac{1}{s_{0r}^2} \mu_{0r}^2 - 2\frac{1}{s_{0r}^2} \mu_{0r} m_{0r} + \frac{1}{s_{0r}^2} m_{0r}^2 \right) \right) \\
&\propto \exp \left(-\frac{1}{2} \left(\mu_{0r}^2 \left(\frac{n}{\sigma_{0r}^2} + \frac{1}{s_{0r}^2} \right) - 2\mu_{0r} \left(\frac{n}{\sigma_{0r}^2} \bar{b}_r + \frac{m_{0r}}{s_{0r}^2} \right) \right) \right) \\
&= \exp \left(-\frac{1}{2 \left(\frac{n}{\sigma_{0r}^2} + \frac{1}{s_{0r}^2} \right)^{-1}} \left(\mu_{0r}^2 - 2\mu_{0r} \left(\frac{n}{\sigma_{0r}^2} + \frac{1}{s_{0r}^2} \right)^{-1} \left(\frac{n}{\sigma_{0r}^2} \bar{b}_r + \frac{m_{0r}}{s_{0r}^2} \right) \right) \right).
\end{aligned}$$

$$\begin{aligned}
p(\phi_h | \mu_0, \Sigma_0, \beta, \gamma, \mathbf{y}, \sigma^2) &\propto \left(\prod_{i:c_i=h} p(\mathbf{y}_i | \beta, \gamma, \phi_h, \sigma^2) \right) p(\phi_h | \mu_0, \Sigma_0) \\
&\propto \exp \left(-\frac{1}{2\sigma^2} \sum_{i:c_i=h} (\tilde{\mathbf{y}}_i - \mathbf{Z}_i \phi_h)' (\tilde{\mathbf{y}}_i - \mathbf{Z}_i \phi_h) \right) \exp \left(-\frac{1}{2} (\phi_h - \mu_0)' \Sigma_0^{-1} (\phi_h - \mu_0) \right) \\
&\propto \exp \left(-\frac{1}{2\sigma^2} \sum_{i:c_i=h} (\phi_h' \mathbf{Z}_i' \mathbf{Z}_i \phi_h - 2\phi_h' \mathbf{Z}_i' \tilde{\mathbf{y}}_i) \right) \exp \left(-\frac{1}{2} (\phi_h' \Sigma_0^{-1} \phi_h - 2\phi_h' \Sigma_0^{-1} \mu_0) \right) \\
&= \exp \left(-\frac{1}{2} \left(\underbrace{\phi_h' \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} \sum_{i:c_i=h} \mathbf{Z}_i' \mathbf{Z}_i \right) \phi_h}_{=:\Sigma_0^{*-1}} - 2\phi_h' \left(\Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} \sum_{i:c_i=h} \mathbf{Z}_i' \tilde{\mathbf{y}}_i \right) \right) \right) \\
&= \exp \left(-\frac{1}{2} \left(\phi_h' \Sigma_0^{*-1} \phi_h - 2\phi_h' \Sigma_0^{*-1} \underbrace{\Sigma_0^* \left(\Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} \sum_{i:c_i=h} \mathbf{Z}_i' \tilde{\mathbf{y}}_i \right)}_{\mu_0^*} \right) \right).
\end{aligned}$$

$$\begin{aligned}
P(c_i = h | \boldsymbol{\pi}, \beta, \gamma, \phi, \mathbf{y}_i, \sigma^2) &\propto p(\mathbf{y}_i | \beta, \gamma, \phi_h, \sigma^2) \pi_h \\
&\propto \frac{1}{(\sigma^2)^{0.5n_i}} \exp \left(-\frac{1}{2\sigma^2} (\mathbf{y}_i - \mathbf{X}_i \beta - \mathbf{B}_i \gamma - \mathbf{Z}_i \phi_h)' (\mathbf{y}_i - \mathbf{X}_i \beta - \mathbf{B}_i \gamma - \mathbf{Z}_i \phi_h) \right) \pi_h.
\end{aligned}$$

$p(\boldsymbol{\pi} | \mathbf{c}, \alpha_0)$ und $p(\alpha_0 | \boldsymbol{\pi})$: analog zu Abschnitt B.3.

Literaturverzeichnis

- Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, **1**: 353–355.
- Bush, C. A. & MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**: 275–285.
- Dahl, D. B. (2005). Sequentially-allocated merge-split sampler for conjugate and non-conjugate Dirichlet process mixture models. *Technical Report*, Texas A&M University, Department of Statistics.
- De Finetti, B. (1974). *Theory of probability*, John Wiley and Sons, New York.
- Dunson, D. B. (2008). Nonparametric Bayes applications to biostatistics. *Department of Statistical Science*, Duke University. Version: 2008.
- Escobar, M. D. (1994). Estimating normal means with a Dirichlet Process Prior. *Journal of the American Statistical Association*, **89**: 268–277.
- Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**: 577–588.
- Fahrmeir, L., Hamerle, A. & Tutz, G. (1996). *Multivariate statistische Verfahren (2. Auflage)*, De Gruyter, Berlin.
- Fahrmeir, L. & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models (2. Auflage)*, Springer, New York.
- Fahrmeir, L., Kneib, T. & Lang, S. (2007). *Regression: Modelle, Methoden und Anwendungen*, Springer, Berlin.
- Fenske, N. (2008). *Flexible Longitudinaldaten-Regression mit Anwendungen auf Adipositas*, Diplomarbeit, Institut für Statistik, Ludwig-Maximilians-Universität München.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**: 209–230.
- Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**: 615–629.
- Hämmerlin, G. & Hoffmann, K.-H. (1994). *Numerische Mathematik (4. Auflage)*, Springer, Berlin.

- Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**: 161–173.
- Ishwaran, H. & James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, **11**: 508–532.
- Ishwaran, H. & Takahara, G. (2002). Independent and identically distributed Monte Carlo algorithms for semiparametric linear mixed models. *Journal of the American Statistical Association*, **97**: 1154–1166.
- Ishwaran, H. & Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, **87**: 371–390.
- Jain, S. & Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, **13**: 158–182.
- Kleinman, K. P. & Ibrahim, J. G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics*, **54**: 921–938.
- Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**: 963–974.
- Li, Y., Lin, X. & Müller, P. (2007). Bayesian inference in semiparametric mixed models for longitudinal data. *Department of Biostatistics Working Paper Series*, UT MD Anderson Cancer Center.
- Lindstrom, M. J. & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**: 1014–1022.
- MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation*, **23**: 727–741.
- MacEachern, S. N. & Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**: 223–238.
- MacEachern, S. N., Clyde, M. & Liu, J. S. (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *Canadian Journal of Statistics*, **27**: 251–267.
- Muliere, P. & Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics*, **26**: 283–297.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**: 249–265.
- Newton, M. A. & Zhang, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, **86**: 15–26.

- Papaspiliopoulos, O. & Roberts, G. (2008). Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika*, **95**: 169–186.
- Rüger, B. (1999). *Test- und Schätztheorie. Band I: Grundlagen*, R. Oldenburg Verlag, München.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**: 639–650.
- Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, **36**: 45–54.
- West, M., Müller, P. & Escobar, M. D. (1994). Hierarchical priors and mixture models with application in regression and density estimation, *in* P. R. Freeman & A. F. M. Smith (eds), *Aspects of uncertainty*, Wiley, New York.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Diplomarbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 09. Februar 2009

Felix Heinzl