Nonparametrische Bayes-Inferenz in mehrdimensionalen Item Response Modellen

Felix Naumann

München 2018

Nonparametrische Bayes-Inferenz in mehrdimensionalen Item Response Modellen

Felix Naumann

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
an der Fakultät für Psychologie und Pädagogik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Felix Naumann
aus München

München, den 25.04.2018

Erstgutachter: Prof. Dr. Markus Bühner
Zweitgutachter: Prof. Dr. Sven Hilbert
Tag der mündlichen Prüfung: 19.07.2018

Inhaltsverzeichnis

Zusammenfassung **xvii**

1 Einleitung **1**
1.1 Item Response Modelle . . . 1
1.1.1 Parametrische Item Response Modelle . . . 3
1.1.2 Nonparametrische Item Response Modelle . . . 6
1.2 Gegenstand der Arbeit . . . 8

2 Identifikation des NMIRT Modells **11**
2.1 Identifikation von statistischen Modellen . . . 11
2.1.1 Definition . . . 11
2.1.2 Beispiel . . . 12
2.2 Identifikationsbeweise . . . 13
2.2.1 Beweismethoden . . . 13
2.2.2 Beispiel . . . 14
2.3 Identifikationsbedingungen im NMIRT Modell . . . 16
2.3.1 Restriktionen . . . 16
2.3.2 Probleme für die Parameterschätzung . . . 18
2.4 Identifikationsbeweise für Reparametrisierungen des NMIRT Modells . . . 19
2.4.1 Modell ohne Beschränkung der Ladungen . . . 19
2.4.2 Modell ohne Beschränkung der ICC . . . 21
2.4.3 Modell ohne Beschränkung der Ladungen und der ICC . . . 24

3 Nonparametrische Bayes-Inferenz **25**
3.1 Parametrische Bayes-Inferenz . . . 25
3.2 Nonparametrische Bayes-Inferenz . . . 27
3.2.1 Polya Trees . . . 28
3.2.2 Dirichlet Prozess . . . 37
3.2.3 Dirichlet Prozess Mixtures . . . 39

4 Markov Chain Monte Carlo Methoden **41**
4.1 Metropolis-Hastings Algorithmus . . . 42
4.2 Gibbs Sampler . . . 45

4.3 Hamiltonian Monte Carlo 48

5 Nonparametrische Bayes-Inferenz im NMIRT Modell 51
5.1 Gibbs Sampler für ein NMIRT Modell mit Dirichlet Prozess Mixture 52
5.1.1 Reformulierung des Modells 52
5.1.2 Likelihood 54
5.1.3 Priori-Verteilungen 54
5.1.4 Hyperparameter 57
5.1.5 Posteriori-Verteilung 57
5.1.6 Vollständig bedingte Verteilungen 57
5.1.7 Gibbs Sampler 66
5.2 Gibbs Sampler für ein NMIRT Modell mit Polya Tree 67
5.2.1 Reformulierung des Modells 67
5.2.2 Likelihood 68
5.2.3 Priori-Verteilungen 68
5.2.4 Hyperparameter 69
5.2.5 Posteriori-Verteilung 69
5.2.6 Vollständig bedingte Verteilungen 69
5.2.7 Gibbs Sampler 71
5.3 HMC Sampler für ein NMIRT Modell mit Dirichlet Prozess Mixture 71
5.3.1 Reformulierung des Modells 72
5.3.2 Likelihood 73
5.3.3 Priori-Verteilungen 73
5.3.4 Hyperparameter 74
5.3.5 Posteriori-Verteilung 74
5.3.6 HMC Sampler 75
5.4 Vergleich der Algorithmen 75

6 Anwendungsbeispiel 77
6.1 TDPM Gibbs Sampler 78
6.1.1 Hyperparameter 78
6.1.2 Startwerte 78
6.1.3 Konvergenz 79
6.1.4 Schätzwerte 81
6.2 TPT Gibbs Sampler 83
6.2.1 Hyperparameter 83
6.2.2 Startwerte und Varianz der Vorschlagsverteilungen 83
6.2.3 Konvergenz 83
6.3 TDPM HMC Sampler 85
6.3.1 Hyperparameter 85
6.3.2 Startwerte und HMC Parameter 85
6.3.3 Konvergenz 85
6.3.4 Schätzwerte 87

6.4 Diskussion der Ergebnisse 89

7 Diskussion **91**
7.1 Zusammenfassung 91
7.2 Limitationen und zukünftige Forschung 91
7.2.1 Effizienz der MCMC Algorithmen 91
7.2.2 Praktische Anwendung in der psychologischen Diagnostik 92
7.2.3 Erweiterungen der MCMC Algorithmen 93
7.2.4 Allgemeinere mehrdimensionale IRT Modelle 94
7.3 Fazit 94

A Traceplots **97**
A.1 TDPM Gibbs Sampler 97
A.2 TPT Gibbs Sampler 102
A.3 TDPM HMC Sampler 106

B Verteilungsübersicht **111**

Literatur **113**

Abbildungsverzeichnis

3.1 Graphische Darstellung der Intervalle eines Polya Trees 30
3.2 Graphische Darstellung der ersten zwei Ebenen eines Polya Trees 32
3.3 Ziehungen aus einer TPT Verteilung 35
3.4 Ziehungen aus einer TPT Posteriori-Verteilung 36

6.1 Wahre ICC 78
6.2 Ausgewählte Traceplots TDPM Gibbs Sampler 80
6.3 Posteriori ICC TDPM Gibbs Sampler 82
6.4 Ausgewählte Traceplots TPT Gibbs Sampler 84
6.5 Ausgewählte Traceplots TDPM HMC Sampler 86
6.6 Posteriori ICC TDPM HMC Sampler 88

A.1 Traceplots des Varianzparameters aus dem TDPM Gibbs Sampler 97
A.2 Traceplots der ICC Parameter aus dem TDPM Gibbs Sampler 98
A.3 Traceplots der Schwierigkeitsparameter aus dem TDPM Gibbs Sampler . . 99
A.4 Traceplots der Ladungen auf Faktor 1 aus dem TDPM Gibbs Sampler . . . 100
A.5 Traceplots der Ladungen auf Faktor 2 aus dem TDPM Gibbs Sampler . . . 101
A.6 Traceplots eines ICC Parameters aus dem TPT Gibbs Sampler 102
A.7 Traceplots der Schwierigkeitsparameter aus dem TPT Gibbs Sampler . . . 103
A.8 Traceplots der Ladungen auf Faktor 1 aus dem TPT Gibbs Sampler 104
A.9 Traceplots der Ladungen auf Faktor 2 aus dem TPT Gibbs Sampler 105
A.10 Traceplots des Gammaparameters aus dem TDPM HMC Sampler 106
A.11 Traceplots der ICC Parameter aus dem TDPM HMC Sampler 107
A.12 Traceplots der Schwierigkeitsparameter aus dem TDPM HMC Sampler . . 108
A.13 Traceplots der Ladungen auf Faktor 1 aus dem TDPM HMC Sampler . . . 109
A.14 Traceplots der Ladungen auf Faktor 2 aus dem TDPM HMC Sampler . . . 110

Tabellenverzeichnis

6.1 Schätzwerte TDPM Gibbs Sampler 81
6.2 Schätzwerte TDPM HMC Sampler 87

Abkürzungsverzeichnis

DP	Dirichlet Prozess
DPM	Dirichlet Prozess Mixture
EAP	Expected-A-Posteriori Schätzer
HMC	Hamiltonian Monte Carlo
ICC	Item Characteristic Curve
IRT	Item Response Theorie
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings
NMIRT	Nonparametrisches mehrdimensionales Item Response Modell
PT	Polya Tree
TDP	Trunkierter Dirichlet Prozess
TPT	Trunkierter Polya Tree
TDPM	Trunkierte Dirichlet Prozess Mixture

Anmerkungen zur Notation

Zufallsvariablen werden in der vorliegenden Arbeit mit lateinischen Großbuchstaben bezeichnet. Ihre Realisationen sind die entsprechenden lateinischen Kleinbuchstaben. Multivariate Zufallsvariablen und ihre Realisationen sind fettgedruckt dargestellt, wobei jeweils der Kontext bestimmt, ob es sich bei ihnen um Vektoren oder Matrizen handelt.

Parameter werden mit griechischen Kleinbuchstaben bezeichnet. Wie bei den Zufallsvariablen sind Vektoren und Matrizen fettgedruckt. Um eine einheitliche Notation für frequentistische und bayesianische Modelle zu gewährleisten, wird nicht zwischen festen und zufälligen Parametern unterschieden.

Allgemeine Wahrscheinlichkeitsmaße werden mit P bezeichnet. Ihre Verteilungsfunktionen und Dichten sind jeweils als F und f notiert. Wie bei den Parametern ergibt sich jeweils aus dem Kontext, ob es sich um feste oder zufällige Größen handelt. Für die Notation konkreter Verteilungen sei auf die Verteilungsübersicht in Anhang B verwiesen.

Mengen werden mit lateinischen oder griechischen Großbuchstaben bezeichnet. Hiervon ausgenommen sind die natürlichen Zahlen $\mathbb{N}$ und die reellen Zahlen $\mathbb{R}$. Die Notation für eine allgemeine σ-Algebra über einer Menge A ist $\sigma(A)$.

Alle weiteren mathematischen Symbole werden vor ihrer Verwendung definiert, außer wenn vorausgesetzt werden kann, dass ihre Bedeutung sich aus allgemeinen Konventionen ergibt (zum Beispiel im Falle von exp als Bezeichnung für die Exponentialfunktion).

Zusammenfassung

Parametrische Item Response Modelle beschreiben die empirisch beobachteten Itemantworten von Personen in psychologischen Tests nicht immer ausreichend gut. In vielen Fällen weisen nonparametrische Modelle mit frei schätzbaren Item Characteristic Curves (ICC) eine bessere Anpassung an die Daten auf. Peress (2012) stellt in diesem Zusammenhang ein sehr flexibles Modell vor, das als nonparametrische Verallgemeinerung des faktorenanalytischen Modells für dichotome Items aufgefasst werden kann.

Gegenstand der vorliegenden Arbeit ist die Herleitung von bayesianischen Schätzverfahren für die ICC und die Item- und Personenparameter des Modells von Peress (2012). Zu diesem Zweck wird zunächst eine Reparametrisierung vorgenommen, welche die bayesianische Formulierung des Modells erleichtert. Mithilfe von Methoden aus der nonparametrischen Bayes-Inferenz werden anschließend Priori-Verteilungen für die frei schätzbare ICC festgelegt.

Hieraus resultieren drei verschiedene Modelle, für die jeweils ein Markov Chain Monte Carlo Algorithmus vorgeschlagen wird: Das erste Modell basiert auf einer Dirichlet Prozess Mixture von Normalverteilungen und erlaubt die Herleitung eines Gibbs Samplers. Im zweiten Modell wird eine Polya Tree Verteilung als Priori-Verteilung für die ICC gewählt und ein Metropolis-within-Gibbs Sampler konstruiert. Für das dritte Modell führt schließlich die Wahl einer Dirichlet Prozess Mixture von logistischen Verteilungen zu einem Hamiltonian Monte Carlo Algorithmus.

Kapitel 1

Einleitung

1.1 Item Response Modelle

Psychologische Variablen wie Intelligenz, Depression oder Extraversion haben den Nachteil, dass sie nicht direkt beobachtbar sind: Es handelt sich bei ihnen um sogenannte latente Variablen (Bühner, 2011). Trotzdem würde man in der Psychologie gerne empirische Aussagen über sie treffen. Ein Ansatz zur Lösung dieses Problems besteht darin, zunächst tatsächlich beobachtbare Variablen zu erfassen, die womöglich mit den interessierenden latenten Variablen in Beziehung stehen. Hierbei könnte es sich beispielsweise im Fall von Intelligenz um Antworten von Personen auf Items eines IQ-Tests handeln. Von den beobachtbaren Variablen versucht man in einem zweiten Schritt auf die dahinterstehenden psychologischen Variablen zu schließen. Allerdings setzt ein derartiges Vorgehen voraus, dass die genaue Beziehung zwischen den beobachtbaren und latenten Variablen bekannt ist. Modelle der Item Response Theorie (IRT) formalisieren diese Beziehung in Form von statistischen Modellgleichungen und erlauben somit die empirische Überprüfung des postulierten Zusammenhangs. Liegt ein solcher vor, kann mithilfe statistischer Inferenzmethoden auf die latenten Variablen geschlossen werden.

Da Itemantworten von Personen Zufallsschwankungen unterliegen, fasst man sie in der IRT als Zufallsvariablen auf. Unterschiedliche Itemformate schlagen sich in unterschiedlichen Wertebereichen dieser Zufallsvariablen nieder. In dieser Arbeit werden ausschließlich IRT Modelle für dichotome Items behandelt, die nur zwei Antwortalternativen wie zum Beispiel ja-nein oder richtig-falsch zulassen. Modelle für polytome und stetige Itemantworten sind nicht Gegenstand der Arbeit.

Antworten $n \in \mathbb{N}$ Personen auf $k \in \mathbb{N}$ Items, wird die Itemantwort der i-ten Person auf das j-te Item für alle $i = 1, 2, \ldots, n$ und alle $j = 1, 2, \ldots, k$ als Zufallsvariable X_{ij} auf einem Wahrscheinlichkeitsraum $(\Omega, \sigma(\Omega), P)$ aufgefasst. Die Itemantworten werden mit den Werten 0 und 1 kodiert, so dass sich für X_{ij} der Wertebereich $\{0, 1\}$ ergibt. Realisiert sich die Zufallsvariable X_{ij} in dem Wert $x_{ij} = 1$, sagt man auch, dass Person i das Item j gelöst hat. Die latente Variable wird für jede Person i durch einen Vektor $\boldsymbol{\xi}_i \in \mathbb{R}^Q$ mit $Q \in \mathbb{N}$ repräsentiert. Q entspricht hierbei der Dimension der latenten Variable.

Modelle mit festen latenten Variablen

In IRT Modellen mit festen latenten Variablen wird $\boldsymbol{\xi}_i$ für jede Person i als fester Modellparameter aufgefasst. Diese Modelle haben für dichotome Itemantworten die allgemeine Form

$$E(X_{ij}) = P(X_{ij} = 1) = F_j(\boldsymbol{\xi}_i) \tag{1.1}$$

(Peress, 2012). Die Funktion F_j in dieser Gleichung wird Item Characteristic Curve (ICC) genannt. Sie spezifiziert für jedes Item j den Zusammenhang zwischen der latenten Variable $\boldsymbol{\xi}_i$ und dem Erwartungswert der Itemantwort $E(X_{ij})$, welcher für dichotome Items der Lösungswahrscheinlichkeit $P(X_{ij} = 1)$ entspricht. Da die Itemantworten dichotom und deshalb Bernoulli-verteilt sind, ergibt sich für jede Beobachtung die folgende diskrete Dichte (van der Linden, 2016c):

$$\begin{aligned} f(x_{ij}) &= P(X_{ij} = x_{ij}) \\ &= P(X_{ij} = 1)^{x_{ij}}(1 - P(X_{ij} = 1))^{1-x_{ij}} \\ &= F_j(\boldsymbol{\xi}_i)^{x_{ij}}(1 - F_j(\boldsymbol{\xi}_i))^{1-x_{ij}} \end{aligned}$$

Die Annahme der lokalen stochastischen Unabhängigkeit der Itemantworten führt dann zur Dichte für das Antwortpattern $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top$ einer Person i (van der Linden, 2016c):

$$f(\mathbf{x}_i) = \prod_{j=1}^{k} f(x_{ij}) = \prod_{j=1}^{k} F_j(\boldsymbol{\xi}_i)^{x_{ij}}(1 - F_j(\boldsymbol{\xi}_i))^{1-x_{ij}}$$

Nimmt man zusätzlich an, dass die Personen unabhängig voneinander auf die Items antworten, ergibt sich als Dichte der gesamten Beobachtungen $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\top$

$$f(\mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x}_i) = \prod_{i=1}^{n}\prod_{j=1}^{k} F_j(\boldsymbol{\xi}_i)^{x_{ij}}(1 - F_j(\boldsymbol{\xi}_i))^{1-x_{ij}} \tag{1.2}$$

(van der Linden, 2016c). Als Funktion der Parameter betrachtet ist diese bei gegebenen Daten die Likelihood der Modelle mit festen latenten Variablen.

Modelle mit zufälligen latenten Variablen

In IRT Modellen mit zufälligen latenten Variablen wird $\boldsymbol{\xi}_i$ nicht als fester Modellparameter, sondern als Zufallsvariable aufgefasst. Dies ist zum Beispiel eine sinnvolle Annahme, falls die Personen aus einer Population zufällig in die Stichprobe gezogen werden. Modelliert wird dann der bedingte Erwartungswert beziehungsweise die bedingte Lösungswahrscheinlichkeit

$$E(X_{ij}|\boldsymbol{\xi}_i) = P(X_{ij} = 1|\boldsymbol{\xi}_i) = F_j(\boldsymbol{\xi}_i) \tag{1.3}$$

(Peress, 2012). Die auf die latente Variable bedingten Itemantworten folgen jeweils einer Bernoulli-Verteilung (van der Linden, 2016c):

$$\begin{aligned} f(x_{ij}|\boldsymbol{\xi}_i) &= P(X_{ij} = x_{ij}|\boldsymbol{\xi}_i) \\ &= P(X_{ij} = 1|\boldsymbol{\xi}_i)^{x_{ij}}(1 - P(X_{ij} = 1|\boldsymbol{\xi}_i)^{1-x_{ij}} \\ &= F_j(\boldsymbol{\xi}_i)^{x_{ij}}(1 - F_j(\boldsymbol{\xi}_i))^{1-x_{ij}} \end{aligned}$$

Auch die lokale Unabhängigkeit wird auf der Ebene der bedingten Wahrscheinlichkeiten formuliert (van der Linden, 2016c):

$$f(\mathbf{x}_i|\boldsymbol{\xi}_i) = \prod_{j=1}^{k} f(x_{ij}|\boldsymbol{\xi}_i) = \prod_{j=1}^{k} F_j(\boldsymbol{\xi}_i)^{x_{ij}}(1 - F_j(\boldsymbol{\xi}_i))^{1-x_{ij}}$$

Dies führt in Verbindung mit einer Verteilungsannahme für die latente Variable in Form einer Dichte $f(\boldsymbol{\xi}_i)$ zu folgender marginaler Dichte für ein Antwortpattern (van der Linden, 2016c):

$$\begin{aligned} f(\mathbf{x}_i) &= \int f(\mathbf{x}_i, \boldsymbol{\xi}_i) d\boldsymbol{\xi}_i \\ &= \int f(\mathbf{x}_i|\boldsymbol{\xi}_i) f(\boldsymbol{\xi}_i) d\boldsymbol{\xi}_i \\ &= \int \prod_{j=1}^{k} F_j(\boldsymbol{\xi}_i)^{x_{ij}}(1 - F_j(\boldsymbol{\xi}_i))^{1-x_{ij}} f(\boldsymbol{\xi}_i) d\boldsymbol{\xi}_i \end{aligned}$$

Die Annahme der Unabhängigkeit der Personen führt schließlich zu

$$f(\mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x}_i) = \prod_{i=1}^{n} \int \prod_{j=1}^{k} F_j(\boldsymbol{\xi}_i)^{x_{ij}}(1 - F_j(\boldsymbol{\xi}_i))^{1-x_{ij}} f(\boldsymbol{\xi}_i) d\boldsymbol{\xi}_i \tag{1.4}$$

(van der Linden, 2016c). Betrachtet man diese Funktion bei gegebenen Daten wieder als Funktion der Parameter, erhält man die Likelihood der Modelle mit zufälligen latenten Variablen.

Die Gleichungen (1.2) bzw. (1.4) gelten je nach Art der latenten Variable (Modellparameter oder Zufallsvariable) für alle IRT Modelle. Einzelne Modelle unterscheiden sich nur in ihren Annahmen bezüglich der Dimension Q der latenten Variable $\boldsymbol{\xi}_i$ und bezüglich der Restriktionen der ICC F_j. In den folgenden Abschnitten soll ein kurzer Überblick über die wichtigsten Modelle gegeben werden.

1.1.1 Parametrische Item Response Modelle

In parametrischen Item Response Modellen wird die Form der ICC F_j als bekannt vorausgesetzt (Peress, 2012). Lediglich einzelne Parameter dieser Funktion können für jedes Item j geschätzt werden.

Eindimensionale parametrische Modelle

Eindimensionale Modelle gehen davon aus, dass nur eine einzige psychologische Variable das Antwortverhalten der Personen beeinflusst und somit $\boldsymbol{\xi}_i = \xi_i \in \mathbb{R}$ ist.

Beispiele für eindimensionale parametrische Modelle sind die logistischen Modelle. Hier ist die Form der ICC durch die logistische Funktion

$$F(x) = \frac{\exp(x)}{1 + \exp(x)}$$

vorgegeben (van der Linden, 2016c). Das einfachste Modell dieser Art ist das Rasch Modell (Rasch, 1960):

$$F_j(\xi_i) = \frac{\exp(\xi_i - \sigma_j)}{1 + \exp(\xi_i - \sigma_j)} \tag{1.5}$$

Hier kann sich lediglich die durch den Paramter σ_j repräsentierte Lokation der logistischen ICC von Item zu Item unterscheiden. In dem etwas allgemeineren Birnbaum Modell (Birnbaum, 1968) können sich die Items über die Lokation hinaus in der durch einen Diskriminationsparameter δ_j bestimmten Steigung der logistischen Funktion unterscheiden:

$$F_j(\xi_i) = \frac{\exp(\delta_j(\xi_i - \sigma_j))}{1 + \exp(\delta_j(\xi_i - \sigma_j))} \tag{1.6}$$

Legt man der ICC statt der logistischen Funktion die Verteilungsfunktion der Standardnormalverteilung

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy$$

zugrunde, ergeben sich die Probit Modelle (van der Linden, 2016c). Auch hier kann zwischen Modellen mit einem oder zwei Parametern pro Item unterschieden werden: Das 1-Parameter Probit Modell mit Lokationsparameter σ_j für jedes Item hat die Form

$$F_j(\xi_i) = \Phi(\xi_i - \sigma_j). \tag{1.7}$$

Das 2-Parameter Probit Modell mit Lokationsparameter σ_j und Diskriminationsparameter δ_j besitzt die Modellgleichung

$$F_j(\xi_i) = \Phi(\delta_j(\xi_i - \sigma_j)). \tag{1.8}$$

Alle bisher beschriebenen Modelle können um einen itemspezifischen Rateparameter $\gamma_j \in [0, 1]$ ergänzt werden, der dafür sorgt, dass der linksseitige Grenzwert der ICC nicht mehr gleich Null sondern gleich γ_j ist:

$$P(X_{ij} = 1) = \gamma_j + (1 - \gamma_j)F_j(\xi_i)$$

Beispiele für Modelle dieser Art finden sich in Birnbaum (1968) und San Martín, Del Pino und Boeck (2006).

Zudem können alle Modelle als Mixture IRT Modelle formuliert werden (von Davier & Rost, 2016). In diesen wird für jede Person neben der stetigen latenten Variable ξ_i eine diskrete latente Variable mit $L \in \mathbb{N}$ Ausprägungen angenommen. Für jede Ausprägung $l = 1, 2, \ldots, L$ auf dieser latenten Klassenvariable wird dann eine eigene ICCs F_{jl} spezifiziert. Mit π_l als Wahrscheinlichkeit dafür, dass sich eine Person in der latenten Klasse l befindet, führt dies zu

$$P(X_{ij} = 1) = \sum_{l=1}^{L} \pi_l F_{jl}(\xi_i).$$

Die Methoden zur Schätzung der Parameter der eindimensionalen parametrischen Modelle unterscheiden sich je nach Modell und abhängig davon, ob man von festen oder zufälligen latenten Variablen ausgeht. Im Rasch Modell mit festen latenten Variablen werden die Parameter im frequentistischen Kontext meist mittels der Conditional Maximum Likelihood Methode geschätzt (Fischer, 1981). Die Marginal Maximum Likelihood Schätzung (Bock & Aitkin, 1981) ist die Methode der Wahl im Rasch Modell mit zufälligen latenten Variablen, im Birnbaum Modell, in den Probit Modellen und in den Modellen mit Rateparameter. Im bayesianischen Kontext schlagen Albert (1992) und Béguin und Glas (2001) jeweils Gibbs Sampler für das 2-Parameter Probit Modell mit und ohne Rateparameter vor. Zur Schätzung der Mixture Modelle kann auf den EM Algorithmus oder auf bayesianische Schätzmethoden zurückgegriffen werden (von Davier & Rost, 2016). Als Überblick über eindimensionale parametrische Modelle und die in deren Rahmen verwendeten Schätzmethoden seien van der Linden (2016a) und van der Linden (2016b) empfohlen.

Mehrdimensionale parametrische Modelle

Mehrdimensionale IRT Modelle gehen davon aus, dass die Itemantworten der Personen von mehr als einer psychologischen Variable bestimmt werden. Formal bedeutet dies, dass die latente Variable als Vektor $\boldsymbol{\xi}_i \in \mathbb{R}^Q$ mit Dimension $Q > 1$ aufgefasst wird.

Auch hier kann abhängig von der Form der ICC wieder zwischen logistischen Modellen und Probit Modellen unterschieden werden. Das einfachste logistische Modell ist das mehrdimensionale Rasch Modell (Adams, Wilson & Wang, 1997)

$$F_j(\xi_i) = \frac{\exp\left(\left(\sum_{q=1}^{Q} d_{jq}\xi_{iq}\right) - \sigma_j\right)}{1 + \exp\left(\left(\sum_{q=1}^{Q} d_{jq}\xi_{iq}\right) - \sigma_j\right)}, \tag{1.9}$$

wobei d_{jq} eine feste Indikatorvariable ist, die den Wert 1 annimmt, falls die q-te Dimension ξ_{iq} des latenten Variablenvektors $\boldsymbol{\xi}_i$ mit dem Item j zusammenhängt, und den Wert 0, falls nicht. Dabei ist d_{jq} kein Parameter des Modells, sondern muss vor der Analyse auf der Basis von inhaltlichen Erwägungen festgelegt werden.

Eine natürliche Erweiterung des Birnbaum Modells (1.6) ist das mehrdimensionale kompensatorische logistische IRT Modell. Dieses wird meist in einer etwas anderen Parametrisierung als das eindimensionale Modell formuliert. Dadurch wird die Nähe zum faktorenanalytischen Modell für stetige Itemantworten deutlich (Reckase, 2009):

$$F_j(\xi_i) = \frac{\exp(\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i)}{1 + \exp(\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i)}. \tag{1.10}$$

Hier werden im Gegensatz zum mehrdimensionalen Rasch Modell unterschiedliche, für jedes Item in einem Q-dimensionalen Vektor $\boldsymbol{\alpha}_j$ zusammengefasste Ladungen α_{jq} der Items auf die Dimensionen $q = 1, 2, \ldots, Q$ der latenten Variable zugelassen.

Für die mehrdimensionalen Probit Modelle ergeben sich jeweils die Gleichungen

$$F_j(\xi_i) = \Phi\left(\left(\sum_{q=1}^{Q} d_{jq}\xi_{iq}\right) - \sigma_j\right) \tag{1.11}$$

und

$$F_j(\xi_i) = \Phi(\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i) \tag{1.12}$$

(Bock, Gibbons & Muraki, 1988). Die Modelle (1.10) und (1.12) können als faktorenanalytische Modelle für dichotome Items aufgefasst werden.

Genau wie die eindimensionalen Modelle können auch die mehrdimensionalen parametrischen Modelle um Rateparameter (Bock et al., 1988) und um Mixturekomponenten (von Davier & Rost, 2016) erweitert werden.

Zur Schätzung der Parameter werden wie im eindimensionalen Fall Conditional und Marginal Maximum Likelihood Ansätze verwendet (Reckase, 2009). Einen Gibbs Sampler für ein bayesianisches mehrdimensionales Modell der Form (1.12) mit und ohne Rateparameter haben Béguin und Glas (2001) vorgelegt. Für einen Überblick über mehrdimensionale Modelle siehe Reckase (2009).

1.1.2 Nonparametrische Item Response Modelle

In nonparametrischen IRT Modellen fasst man die ICC F_j als weitgehend frei zu schätzenden Parameter mit einem unendlich-dimensionalen Funktionenraum als Parameterraum auf, wobei nur sehr schwache Annahmen getroffen werden (Peress, 2012). Aus inhaltlicher Sicht ist der Vorteil dieser Modelle gegenüber den parametrischen Modellen in der Lockerung der sehr restriktiven Annahme bezüglich der Form von F_j zu sehen. In vielen Fällen kann nicht davon ausgegangen werden, dass fest vorgegebene Funktionen wie die logistische Funktion oder die Verteilungsfunktion der Normalverteilung die Abhängigkeit der Itemantworten von der latenten Variable ausreichend gut beschreiben. Nonparametrische Modelle erlauben eine deutlich flexiblere Modellierung. So können sie beispielsweise auch in Situationen angewandt werden, in denen höhere Werte auf den latenten Variablen in einem Bereich des Definitionsbereichs von F_j keinen großen Einfluss auf die Lösungswahrscheinlichkeit eines Items j haben oder in denen der Zusammenhang zwischen den latenten

Variablen und der Antwortwahrscheinlichkeit weitgehend linear ist. Darüber hinaus haben nonparametrische IRT Modelle auch allein deswegen ihre Berechtigung, weil sie mit den parametrischen Modellen hinsichtlich ihrer Anpassung an die Daten verglichen werden können. Ein solcher Vergleich kann als Modelltest für das parametrische Modell interpretiert werden und Aufschluss über eine Missspezifikation der Form der ICC geben.

Eindimensionale nonparametrische Modelle

Eindimensionale nonparametrische IRT Modelle haben die Form

$$P(X_{ij} = 1) = F_j(\xi_i), \tag{1.13}$$

wobei F_j für jedes Item eine weitgehend beliebige Funktion sein kann (Douglas, 2001). Aus Identifizierungsgründen muss lediglich die Einschränkung gemacht werden, dass F_j für jedes Item j eine stetige Verteilungsfunktion, also stetig und monoton steigend mit linksseitigem Grenzwert 0 und rechtsseitigem Grenzwert 1, ist. Außerdem müssen noch einige für die Praxis weniger relevante mathematisch-technische Restriktionen bezüglich F_j getroffen werden (Douglas, 2001). Es existiert eine Vielzahl von Methoden zur Schätzung der nonparametrischen ICCs, unter anderem die klassische Mokken-Analyse (Sijtsma & Molenaar, 2016), Ansätze mit Splines (Johnson, 2007), Kernel-basierte Schätzer (Ramsay, 1991) und nonparametrische bayesianische Methoden (Duncan & MacEachern, 2008; Miyazaki & Hoshino, 2009).

Mehrdimensionale nonparametrische Modelle

Die mehrdimensionalen nonparametrischen Modelle bilden die allgemeinste Modellklasse für dichotome Itemantworten. Sie enthalten die meisten der bisher beschriebenen Modelle entweder als Spezialfälle oder können diese beliebig genau approximieren.

Ein mehrdimensionales nonparametrisches 1-Parameter Modell kann im Rahmen des von Karabatsos (2016) vorgestellten nonparametrischen Regressionsmodells formuliert werden:

$$P(X_{ij} = 1) = F_j\left(\left(\sum_{q=1}^{Q} d_{jq}\xi_{iq}\right) - \sigma_j\right), \tag{1.14}$$

wobei die festen Konstanten d_{jq} wie in (1.9) und (1.11) gewählt werden müssen und die latenten Variablen $\boldsymbol{\xi}_i$ als feste Parameter betrachtet werden. Dieses Modell ist eine Verallgemeinerung der mehrdimensionalen 1-Parameter Modelle (1.9) und (1.11) mit frei schätzbaren monotonen und stetigen Verteilungsfunktionen F_j für jedes Item j wie in (1.13). Die Schätzung der F_j und der Item- und Personenparameter ist mithilfe von nonparametrischen bayesianischen Methoden möglich (Karabatsos, 2016).

Peress (2012) schlägt eine nonparametrische Generalisierung der mehrdimensionalen kompensatorischen Modelle (1.10) und (1.12) vor:

$$P(X_{ij} = 1) = F(\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i) \tag{1.15}$$

Hierbei werden die latenten Variablen $\boldsymbol{\xi}_i$ als feste Parameter betrachtet. Dieses Modell geht von einer frei schätzbaren stetigen Verteilungsfunktion F aus, die im Gegensatz zu Modell (1.14) für jedes Item j die gleiche Form aufweist. Es sind jedoch, anders als in Modell (1.14) unterschiedliche Ladungen der Items auf die Dimensionen der latenten Variable zugelassen. Modell (1.15) hat eine Reihe günstiger Eigenschaften, die es für die Anwendung im Kontext psychologischer Tests besonders geeignet erscheinen lassen: Zunächst hat es aufgrund der nonparametrischen Form der ICC eine sehr hohe Flexibilität, so dass eine Vielzahl von Antwortprozessen modellierbar ist. Außerdem muss in psychologischen Anwendungen oft davon ausgegangen werden, dass einzelne Items von mehreren latenten Variablen unterschiedlich stark beeinflusst werden. Besonders im explorativen Kontext der Testkonstruktion ist es oft notwendig, diese Items zu identifizieren. Modell (1.15) erlaubt dies, da im Gegensatz zu (1.14) unterschliedliche Ladungen für jedes Item und jede Dimension geschätzt werden können, die den unterschiedlichen Einfluss der verschiedenen Dimensionen der latenten Variable auf die Items quantifizieren. Es kann als nonparametrisches exploratives faktorenanalytisches Modell für dichotome Itemantworten interpretiert werden. Eine detaillierte Diskussion des Modells (1.15) und der für seine Identifikation notwendigen Restriktionen findet sich im nächsten Kapitel. Es soll im folgenden als nonparametrisches mehrdimensionales IRT (NMIRT) Modell bezeichnet werden. Obwohl die ICC in diesem Modell für jedes Item durch die Kombination aus F und dem latenten Prädiktor $\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i$ festgelegt ist, wird im Folgenden auch die Funktion F für sich genommen als ICC bezeichnet.

Peress (2012) schlägt zudem ein Modell der Form

$$P(X_{ij} = 1) = F_j(\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i) \tag{1.16}$$

mit itemspezifischer Funktion F_j vor. Leider ist die Identifikation dieses Modells nur für $Q = 2$ Dimensionen sichergestellt (Peress, 2012). Daher ist es trotz der itemspezifischen ICC und der Ladungsparameter deutlich unflexibler sowohl als Modell (1.14) als auch als Modell (1.15). Es wird daher nicht weiter behandelt.

1.2 Gegenstand der Arbeit

Die Anwendung des NMIRT Modells (1.15) in der Praxis setzt voraus, dass dessen Parameter - die ICC F und die Item- und Personenparameter - aus einer Stichprobe geschätzt werden können. Tragfähige Methoden hierfür stehen bislang nicht zur Verfügung. Zur Lösung dieses Problems werden in der vorliegenden Arbeit nonparametrische bayesianische Schätzverfahren hergeleitet. In Kapitel 2 wird zunächst die Identifikation von Reparametrisierungen des NMIRT Modells bewiesen, welche die bayesianische Formulierung des Modells stark vereinfachen. Im Anschluss gibt Kapitel 3 einen Überblick über nonparametrische Verfahren der Bayes-Inferenz. Da hierbei Markov Chain Monte Carlo Algorithmen eine wichtige Rolle spielen, führt Kapitel 4 kurz in deren Grundlagen ein. In Kapitel 5 werden dann drei verschiedene nonparametrische bayesianische Formulierungen des NMIRT Modells vorgeschlagen und jeweils ein Algorithmus zur Parameterschätzung

hergeleitet. Diese Algorithmen werden in Kapitel 6 zu Illustrationszwecken auf einen simulierten Beispieldatensatz angewandt. Die Arbeit schließt mit einer Diskussion in Kapitel 7.

Kapitel 2

Identifikation des NMIRT Modells

In diesem Kapitel wird zunächst auf den Begriff der Identifikation im Rahmen von statistischen Modellen und seine Bedeutung für die Item Response Theorie eingegangen. Im Anschluss daran werden die Restriktionen diskutiert, die für den Identifikationsbeweis des NMIRT Modells von Peress (2012) notwendig sind. Im letzten Abschnitt des Kapitels wird dieser Identifikationsbeweis schließlich auf einige alternative Parametrisierungen des Modells erweitert.

2.1 Identifikation von statistischen Modellen

2.1.1 Definition

Der Begriff der Identifikation geht auf Koopmans (1949) zurück. Die hier aufgeführte Definition ist San Martín (2016) entnommen.

Sei Ω ein Ergebnisraum mit Ereignisraum $\sigma(\Omega)$ in Form einer σ-Algebra. Sei durch eine Folge von k-dimensionalen Zufallsvariablen $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ eine Stichprobe von Beobachtungen beschrieben, die in einer $n \times k$-Matrix $\mathbf{X}$ zusammengefasst sind. Der Wertebereich $\mathcal{X}$ von $\mathbf{X}$ wird mit einer geeigneten σ-Algebra $\sigma(\mathcal{X})$ versehen. Sei ferner $\mathcal{P}_{\boldsymbol{\theta}}$ eine durch einen Parametervektor $\boldsymbol{\theta} \in \Theta$ indizierte Menge von Verteilungen auf $\mathcal{X}$ mit Elementen $P_{\boldsymbol{\theta}}$. Ein statistisches Modell $(\mathcal{X}, \sigma(\mathcal{X}), \mathcal{P}_{\boldsymbol{\theta}})$ ist identifiziert, falls

$$P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}^*} \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^* \tag{2.1}$$

für alle $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta$ gilt. In dem Fall, dass jedes $P_{\boldsymbol{\theta}} \in \mathcal{P}_{\boldsymbol{\theta}}$ durch eine Likelihoodfunktion $f(\mathbf{x}|\boldsymbol{\theta})$ charakterisiert werden kann, ist diese Bedingung äquivalent zu

$$f(\mathbf{x}|\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta}^*) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^* \tag{2.2}$$

für alle $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta$ und alle $\mathbf{x} \in \mathcal{X}$.

In identifizierten Modellen können also zwei Parameterausprägungen $\boldsymbol{\theta}$ und $\boldsymbol{\theta}^*$, die zur gleichen gemeinsamen Verteilung der Zufallsvariablen der Stichprobe führen, keine unterschiedlichen Werte annehmen. Da die gemeinsame Verteilung die komplette empirische

Information des durch die Zufallsvariablen beschriebenen Zufallsexperiments enthält, ist ein eindeutiger Schluss von Empirie auf Parameterausprägungen mithilfe statistischer Inferenzverfahren in identifizierten statistischen Modellen prinzipiell möglich (San Martín, González & Tuerlinckx, 2015). In bayesianischen Modellen wird der Begriff der Identifikation etwas weiter gefasst, so dass durch die Wahl geeigneter Priori-Verteilungen auch Modelle, für deren Likelihood die Bedingung (2.2) nicht gilt, identifiziert werden können (San Martín & González, 2010). Bayesianische Modelle, die im Sinne der Bedingung (2.2) identifiziert sind, sind jedoch auch im weiteren bayesianischen Sinne identifiziert (San Martín & González, 2010).

Der Nachweis der Identifikation der Modelle ist im Rahmen der Item Response Theorie aus mehreren Gründen zentral. Der erste Grund ist rein statistischer Natur: In nicht identifizierten Modellen existiert keine konsistente Schätzfunktion für die Parameter (San Martín & Quintana, 2002). Ein zweiter Grund liegt darin, dass die Parameter in identifizierten Modellen eine empirische Interpretation erhalten (San Martín et al., 2015). Dies führt im Rahmen von IRT Modellen unter anderem dazu, dass das zunächst rein theoretische Konzept der latenten Variable eine empirische Bedeutung gewinnt. Ein dritter Grund betrifft die Invarianz der Parameter und ist besonders im Kontext der psychologischen Diagnostik von Bedeutung: Kann für ein IRT Modell gezeigt werden, dass es nicht identifiziert ist, sondern nur durch bestimmte Restriktionen identifiziert werden kann, bestimmt die Art der Nicht-Identifikation zulässige Transformationen der Personenparameter und somit in gewisser Weise das Skalenniveau (Steyer & Eid, 2001).

2.1.2 Beispiel

Die Beziehung zwischen Identifikation und zulässigen Transformationen sei am Beispiel des Rasch Modells (1.5) mit festen latenten Variablen illustriert. Hierbei handelt es sich um die Abwandlung eines von San Martín (2016) in einer anderen Parametrisierung angeführten Beispiels. Sei $c \in \mathbb{R}$ eine beliebige Konstante ungleich 0 und seien $\xi_i^* = \xi_i + c$ und $\sigma_j^* = \sigma_j + c$ für alle Personen i und alle Items j. Dann ist

$$\begin{aligned} P(X_{ij} = 1) &= \frac{\exp(\xi_i - \sigma_j)}{1 + \exp(\xi_i - \sigma_j)} \\ &= \frac{\exp(\xi_i + c - c - \sigma_j)}{1 + \exp(\xi_i + c - c - \sigma_j)} \\ &= \frac{\exp((\xi_i + c) - (\sigma_j + c))}{1 + \exp((\xi_i + c) - (\sigma_j + c))} \\ &= \frac{\exp(\xi_i^* - \sigma_j^*)}{1 + \exp(\xi_i^* - \sigma_j^*)} \end{aligned}$$

für alle Personen i und alle Items j. Das heißt, sowohl $\boldsymbol{\theta} = (\xi_1, \xi_2, \ldots, \xi_n, \sigma_1, \sigma_2, \ldots, \sigma_k)$ als auch $\boldsymbol{\theta}^* = (\xi_1^*, \xi_2^*, \ldots, \xi_n^*, \sigma_1^*, \sigma_2^*, \ldots, \sigma_k^*)$ führen zu den gleichen Lösungswahrscheinlichkeiten und somit wegen (1.2) zur gleichen Likelihood und sind daher empirisch nicht

unterscheidbar. Das Rasch Modell ist ohne zusätzliche Restriktionen somit nicht identifiziert. Die Möglichkeit der Verschiebung der Parameter um eine Konstante wird gelegentlich als Differenzskalenniveau interpretiert (Steyer & Eid, 2001).

2.2 Identifikationsbeweise

2.2.1 Beweismethoden

Während Beweise für die Nichtidentifikation statistischer Modelle häufig einfach zu erbringen sind, gestalten sich Beweise für die Identifikation in der Regel etwas schwieriger. Es gibt mehrere grundlegende Strategien:

Eine zunächst naheliegende Methode besteht darin, direkt von den Parametern und der Likelihoodfunktion des Modells auszugehen und die in der Definition als Bedingung formulierte Implikation (2.2) direkt zu bewiesen. Dies erweist sich in den meisten Fällen jedoch als sehr schwierig.

In manchen Situationen reicht ein Beweis für die Existenz einer konsistenten Schätzfunktion für die Parameter des Modells aus, um die Identifikation sicherzustellen. Auch diese Beweise sind häufig sehr aufwendig. Für Details siehe San Martín und Quintana (2002).

Die oft einfachste Variante besteht darin, die Parameter eines Modells als Funktion von Stichprobenmomenten (Erwartungswerte, Varianzen, Kovarianzen etc.) der zugrundeliegenden Zufallsvariablen darzustellen, da diese, als Parameter betrachtet, stets identifiziert sind (San Martín, 2016): Sei $g(\mathbf{X}, P_{\boldsymbol{\theta}})$ ein beliebiges Stichprobenmoment unter der Verteilung $P_{\boldsymbol{\theta}}$. Gleiche Verteilungen $P_{\boldsymbol{\theta}}$ und $P_{\boldsymbol{\theta}^*}$ implizieren gleiche Stichprobenmomente:

$$P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}^*} \Rightarrow g(\mathbf{X}, P_{\boldsymbol{\theta}}) = g(\mathbf{X}, P_{\boldsymbol{\theta}^*}) \tag{2.3}$$

Können nun die Parameter des Modells als Funktion der Stichprobenmomente dargestellt werden, also $\boldsymbol{\theta} = h(g(\mathbf{X}, P_{\boldsymbol{\theta}}))$, gilt mit (2.3)

$$\begin{aligned} P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}^*} &\Rightarrow g(\mathbf{X}, P_{\boldsymbol{\theta}}) = g(\mathbf{X}, P_{\boldsymbol{\theta}^*}) \\ &\Rightarrow h(g(\mathbf{X}, P_{\boldsymbol{\theta}})) = h(g(\mathbf{X}, P_{\boldsymbol{\theta}^*})) \\ &\Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^* \end{aligned}$$

und das Modell ist identifiziert.

Liegt für eine Parametrisierung eines Modells durch die Parameter $\boldsymbol{\theta}$ bereits ein Identifikationsbeweis vor, kann die Identifikation der Parameter $\boldsymbol{\eta}$ einer anderen Parametrisierung des Modells gezeigt werden, indem man die Parameter der alternativen Parametrisierung als Funktion der Parameter des bereits identifizierten Modells darstellt (San Martín, 2016). Man zeigt also, dass eine Funktion h existiert, so dass $\boldsymbol{\eta} = h(\boldsymbol{\theta})$ für jedes $\boldsymbol{\eta}$ gilt. Zudem weist man nach, dass unter dieser Reparametrisierung $P_{\boldsymbol{\theta}} = P_{\boldsymbol{\eta}}$ und $P_{\boldsymbol{\theta}^*} = P_{\boldsymbol{\eta}^*}$ gilt. In diesem Fall folgt nämlich

$$P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}^*} \Leftrightarrow P_{\boldsymbol{\eta}} = P_{\boldsymbol{\eta}^*}. \tag{2.4}$$

Sind die Parameter $\boldsymbol{\theta}$ identifiziert, gilt per Definition

$$P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}^*} \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^* \tag{2.5}$$

und es folgt aus (2.4) und (2.5)

$$\begin{aligned} P_{\boldsymbol{\eta}} = P_{\boldsymbol{\eta}^*} &\Rightarrow P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}^*} \\ &\Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^* \\ &\Rightarrow h(\boldsymbol{\theta}) = h(\boldsymbol{\theta}^*) \\ &\Rightarrow \boldsymbol{\eta} = \boldsymbol{\eta}^* \end{aligned}$$

für jedes $\boldsymbol{\eta}$ und das reparametrisierte Modell ist ebenfalls identifiziert.

2.2.2 Beispiel

Die letzten beiden Beweismethoden seien nun am Beispiel des Rasch Modells mit festen latenten Variablen dargestellt. Wie in Abschnitt 2.1.2 gezeigt, ist dieses nicht identifiziert. Es kann jedoch durch das Setzen von Restriktionen identifiziert werden. Eine solche Restriktion ist die Summennormierung für die Itemparameter:

$$\sum_{j=1}^{k} \sigma_j = 0 \tag{2.6}$$

Der folgende Beispielbeweis ist an den Beweis in San Martín (2016) angelehnt, geht jedoch von einer anderen Parametrisierung und einer anderen Normierung aus. Durch Umstellen der Modellgleichung des Rasch Modells folgt

$$\ln\left(\frac{P(X_{ij}=1)}{1-P(X_{ij}=1)}\right) = \xi_i - \sigma_j$$

und somit zusammen mit (2.6)

$$\begin{aligned} \frac{1}{k}\sum_{j=1}^{k} \ln\left(\frac{P(X_{ij}=1)}{1-P(X_{ij}=1)}\right) &= \frac{1}{k}\sum_{j=1}^{k}(\xi_i - \sigma_j) \\ &= \frac{1}{k}\sum_{j=1}^{k}\xi_i - \frac{1}{k}\underbrace{\sum_{j=1}^{k}\sigma_j}_{=0} \\ &= \xi_i. \end{aligned}$$

Damit ist

$$\xi_i = \frac{1}{k}\sum_{j=1}^{k} \ln\left(\frac{P(X_{ij}=1)}{1-P(X_{ij}=1)}\right) = \frac{1}{k}\sum_{j=1}^{k} \ln\left(\frac{E(X_{ij})}{1-E(X_{ij})}\right) \tag{2.7}$$

für alle Personen i. Die Personenparameter sind jeweils Funktionen der Erwartungwerte (also der ersten Momente) der Zufallsvariablen X_{ij} und daher identifiziert. Wie man in (2.7) sehen kann, erhalten die Personenparameter ξ_i durch die Summennormierung zudem eine empirische Interpretation: Der Personenparameter einer Person i entspricht unter der Summennormierung dem Mittelwert der log-Odds der Person über alle Items j des Tests. Ergebnis (2.7) kann nun für den Beweis der Identifizierung der Itemparameter verwendet werden:

$$\begin{aligned}\ln\left(\frac{P(X_{ij}=1)}{1-P(X_{ij}=1)}\right) &= \ln\left(\frac{E(X_{ij})}{1-E(X_{ij})}\right)\\ &= \xi_i - \sigma_j\\ &= \frac{1}{k}\sum_{j=1}^{k}\ln\left(\frac{E(X_{ij})}{1-E(X_{ij})}\right) - \sigma_j\end{aligned}$$

und deswegen

$$\sigma_j = \frac{1}{k}\sum_{j=1}^{k}\ln\left(\frac{E(X_{ij})}{1-E(X_{ij})}\right) - \ln\left(\frac{E(X_{ij})}{1-E(X_{ij})}\right)$$

für alle Items j und eine beliebig gewählte Person i. Damit sind alle Paramter des Rasch Modells identifiziert.

Auch andere Restriktion können zur Identifizierung des Modells führen. Beispielsweise könnte das Modell alternativ durch die Fixierung des Schwierigkeitsparameters des ersten Items auf Null identifiziert werden (San Martín et al., 2015). Damit würde sich jedoch auch die empirische Bedeutung der Parameter ändern.

Als Beispiel für die vierte Beweismethode sei nun ausgehend von dem Identifikationsbeweis des Rasch Modells (1.5) mit Summennormierung ein Beweis für die Identifikation des Rasch Modells in der multiplikativen Parametrisierung

$$P(X_{ij}=1) = \frac{\mu_i\delta_j}{1+\mu_i\delta_j} \tag{2.8}$$

mit Personenparametern μ_i und Itemparametern δ_j aufgeführt. Ein alternativer Beweis für das Modell in dieser Parametrisierung findet sich in San Martín (2016). Setze zunächst $\mu_i = \exp(\xi_i)$ für alle Personen i und $\delta_j = \exp(-\sigma_j)$ für alle Items j. Dies führt ausgehend von der ursprünglichen Parametrisierung (1.5) zu

$$P(X_{ij}=1) = \frac{\exp(\xi_i-\sigma_j)}{1+\exp(\xi_i-\sigma_j)} = \frac{\exp(\xi_i)\exp(-\sigma_j)}{1+\exp(\xi_i)\exp(-\sigma_j)} = \frac{\mu_i\delta_j}{1+\mu_i\delta_j}.$$

Damit führen $\mu_i = \exp(\xi_i)$ und $\delta_j = \exp(-\sigma_j)$ zur multipilikativen Parametrisierung (2.8) und zur gleichen Likelihood wie die Parameter ξ_i und σ_j. Die neuen Parameter können mit $\mu_i = \exp(\xi_i)$ und $\delta_j = \exp(-\sigma_j)$ als Funktionen der identifizierten Parameter ξ_i und σ_j der alten Parametrisierung dargestellt werden. Somit ist das Rasch Modell auch in der

multiplikativen Parametrisierung (2.8) identifiziert. Aus der Summennormierung (2.6) der alten Itemparameter folgt als Restriktion für die neuen Itemparameter

$$\begin{aligned} & \sum_{j=1}^{k} \sigma_j = 0 \\ \Leftrightarrow & \sum_{j=1}^{k} -\ln(\delta_j) = 0 \\ \Leftrightarrow & \sum_{j=1}^{k} \ln(\delta_j) = 0 \\ \Leftrightarrow & \ln\left(\prod_{j=1}^{k} \delta_j\right) = 0 \\ \Leftrightarrow & \prod_{j=1}^{k} \delta_j = \exp(0) = 1. \end{aligned}$$

Es ergibt sich also eine Produktnormierung.

2.3 Identifikationsbedingungen im NMIRT Modell

Für nonparametrische Modelle sind die Identifikationsbeweise deutlich aufwendiger, da der Parameteraum der ICC F ein unendlich-dimensionaler Funktionenraum ist. Einen Beweis für das NMIRT Modell

$$P(X_{ij} = 1) = F(\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i)$$

hat Peress (2012) vorgelegt. Für den Identifikationsbeweis sind einige Restriktionen nötig, die im folgenden Abschnitt dargestellt und im Hinblick auf ihre Bedeutung für die praktische Anwendung diskutiert werden sollen.

2.3.1 Restriktionen

Zur Identifizierung des NMIRT Modells müssen folgende Restriktionen gesetzt werden (Peress, 2012):

Restriktion 1. *F ist stetig differenzierbar und streng monoton steigend mit $\lim_{x\to-\infty} F(x) = 0$ und $\lim_{x\to\infty} F(x) = 1$. Zudem ist $F(0) = 0.5$ und $F(1) = p$, wobei $0.5 < p < 1$ ein beliebiger fester Wert ist.*

Die Annahme der stetigen Differenzierbarkeit der ICC ist eine technische Bedingung, die für den Identifikationsbeweis benötigt wird. Sollte die wahre ICC in einem Anwendungsfall tatsächlich Unstetigkeitsstellen (zum Beispiel in Form von Sprüngen) aufweisen, könnte sie

trotzdem durch eine stetig differenzierbare ICC approximiert werden, falls diese ansonsten frei geschätzt werden kann. Die stetige ICC hätte an der Sprungstelle beispielsweise einfach eine besonders starke Steigung.

Auch die Beschränkung auf eine ICC mit linksseitigem Grenzwert von 0 und einem rechtsseitigen Grenzwert von 1 ist nur auf den ersten Blick besonders restriktiv. Ausgeschlossen sind durch diese Restriktion unter anderem Modelle mit Rateparametern. Da die ICC ansonsten jedoch eine beliebige Form annehmen kann, ist eine Approximation von Modellen mit Rateparamter dadurch möglich, dass die ICC sich dem linken Grenzwert 0 erst in einem Bereich des Personenparameterraums annähert, in dem eine vernachlässigbar geringe Anzahl an Personen liegt. Da das Modell jedoch annimmt, dass die ICC über die Items konstant bleibt, kann es nicht in Situationen angewandt werden, in denen davon ausgegangen werden muss, dass die Ratewahrscheinlichkeiten über die Items hinweg variieren. Dies wäre bespielsweise in Leistungstests mit Multiple-Choice-Items gegeben, die eine unterschiedliche Anzahl an Distraktoren aufweisen.

Die Annahme der Monotonie der ICC bedeutet nicht, dass das Modell nur latente Variablen zulässt, die die Lösungswahrscheinlichkeiten der Items positiv beeinflussen. Der Zusammenhang lässt sich durch negative Ladungen umdrehen. Siehe hierfür auch die Diskussion zu den Restriktionen der Ladungen im nächsten Abschnitt. Die Monotonie der ICC schließt allerdings beispielsweise Modelle aus, in denen umgekehrt U-förmige ICCs spezifizert werden.

Die Quantilrestriktionen $F(0) = 0.5$ und $F(1) = p$ sind aus inhaltlicher Sicht unproblematisch, da sie lediglich eine Normierung der ICC darstellen. Verschiebungen und Streckungen der ICC könnten sonst jeweils durch Verschiebungen und Streckungen des latenten Prädiktors $\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i$ ausgeglichen werden (Peress, 2012). Im Hinblick auf die Parameterschätzung stellen diese Restriktionen jedoch ein Problem dar (siehe Abschnitt 2.3.2).

Restriktion 2. $\alpha_{jq} > 0$ *für alle Items* j *und alle Dimensionen* q *und* $\boldsymbol{\alpha}_j \neq \mathbf{0}$ *für alle Items* j*.*

Die Nichtnegativität der Ladungsparameter stellt aus inhaltlicher Sicht keine Einschränkung dar, da sie nach der Schätzung der Parameter durch Rotation der Ladungsmatrix aufgehoben werden kann. Für die Parameterschätzung ist sie jedoch vor allem aus bayesianischer Sicht problematisch (siehe Abschnitt 2.3.2).

Die Restriktion $\boldsymbol{\alpha}_j \neq \mathbf{0}$ besagt lediglich, dass der Test kein Item enthält, welches mit keiner der im Modell spezifizierten latenten Variablen zusammenhängt (Peress, 2012). Dies kann in den meisten praktischen Anwendungen als gegeben vorausgesetzt werden. Zudem kann in Fällen, in denen diese Annahme verletzt ist, ein Modell mit sehr geringen Ladungen auf dem betroffenen Item als Approximation dienen.

Restriktion 3. $\boldsymbol{\xi}_1 = (1, 0, \ldots, 0)^\top$, $\boldsymbol{\xi}_2 = (0, 1, 0, \ldots, 0)^\top$, $\ldots$, $\boldsymbol{\xi}_Q = (0, \ldots, 0, 1)^\top$ *und* $\boldsymbol{\xi}_{Q+1} = (0, \ldots, 0)^\top$.

Dies bedeutet, dass die Personenparameter der ersten $Q+1$ Personen in der Stichprobe fixiert sind. Insgesamt werden folglich $Q(Q+1)$ Personenparameterrestriktionen vorge-

nommen. Diese Restriktionen implizieren nicht, dass die Personenparameter der ersten $Q+1$ Personen bekannt sein müssen. Sie sorgen lediglich dafür, dass keine Rotationen und Translationen der Ladungs- und Personenparameter möglich sind (Peress, 2012) und entsprechen vergleichbaren Restriktionen in faktorenanalytischen Modellen. Nach der Schätzung der Ladungen können herkömmliche Rotationsmethoden auf die Ladungsmatrix und die Personenparameter angewandt werden, die die Restriktion aufheben.

Restriktion 4. *Es liegen abzählbar unendlich viele Personen und Items vor. Die Itemparameter liegen dicht in* $\mathbb{R} \times \mathbb{R}_+^Q$ *und die Personenparameter dicht in* $\mathbb{R}^Q$*.*

Da unendlich viele Personen und Items vorliegen müssen, ist die Identifikation des Modells nur asymptotisch gewährleistet. Aus rein statistischer Sicht ist dies zunächst unproblematisch, da die asymptotische Identifikation für Konsistenzbeweise ausreicht (Peress, 2012). Es ist jedoch in endlichen Stichproben nicht gewährleistet, dass es keine anderen Parameterkonfigurationen innerhalb desselben Modells gibt, die die Daten genauso gut beschreiben. Dies ist eine generelle Eigenschaft nonparametrischer IRT Modelle, die auch für eindimensionale Modelle gilt (Douglas, 2001; Peress, 2012). Die fehlende Identifikation in endlichen Stichproben ist in gewisser Weise der Preis für die größere Flexibilität dieser Modelle. Zu beachten ist zudem, dass sich die Asymptotik nicht nur auf die Anzahl der Personen in der Stichprobe, sondern auch auf die Anzahl der Items bezieht.

Der zweite Teil der Restriktion ist eine technische Bedingung für den Identifikationsbeweis. Formal bedeutet sie im Falle der Personenparameter, dass es für jede (beliebig kleine) reelle Zahl $\epsilon > 0$ und jeden Punkt $\mathbf{r}$ im Parameterraum $\mathbb{R}^Q$ eine Person i gibt, für deren Personenparameter $||\boldsymbol{\xi}_i - \mathbf{r}|| < \epsilon$ gilt, wobei $||\cdot||$ die L2-Norm ist (Peress, 2012). Die Personen in der Stichprobe decken also insofern den gesamten Parameterraum ab, als für jeden Punkt im Parameterraum eine Person in der Stichprobe gefunden werden kann, deren Personenparameter beliebig nahe an diesem Punkt liegt. Für die Itemparameter gilt dies analog. Für die praktische Anwendung bedeutend ist, dass sowohl die Item- als auch die Personenparameter sehr viele unterschiedliche Werte aufweisen sollten. Items und Personen sollten daher möglichst heterogen gewählt werden (Peress, 2012). Wenn ohnehin unendlich viele Items und Personen angenommen werden müssen, stellt diese Restriktion nur eine minimale Verschärfung dar (Peress, 2012).

2.3.2 Probleme für die Parameterschätzung

Problematisch für die Parameterschätzung sind vor allem die Quantilrestriktionen der ICC F und die Nichtnegativität der Ladungen $\boldsymbol{\alpha}_j$. Die Restriktionen der Quantile der ICC führen dazu, dass diese nicht komplett frei geschätzt werden kann, was die Möglichkeiten der nonparametrischen Inferenz stark einschränkt. Die Nichtnegativität der Ladungen $\boldsymbol{\alpha}_j$ impliziert, dass im Rahmen der Bayes-Inferenz keine multivariaten Normalverteilungen als Priori-Verteilungen für $\boldsymbol{\alpha}_j$ gewählt werden können. Wie sich in Kapitel 5 zeigen wird, ist dies jedoch eine Voraussetzung für die Implementierung eines effizienten Gibbs Samplers zur Schätzung der Parameter des Modells.

Angesichts dieser Probleme besteht ein wichtiger Teil der vorliegenden Arbeit darin, ausgehend von dem Identifikationsbeweis von Peress (2012) identifizierte Modelle mit alternativen Restriktionen herzuleiten, die für die Parameterschätzung weniger problematisch sind. Dies ist Gegenstand des folgenden Abschnitts.

2.4 Identifikationsbeweise für Reparametrisierungen des NMIRT Modells

2.4.1 Modell ohne Beschränkung der Ladungen

Zunächst soll die Beschränkung der Ladungsparameter $\boldsymbol{\alpha}_j$ des ursprünglichen Modells von Peress (2012) auf die nicht-negativen reellen Zahlen mithilfe einer Reparametrisierung der Item- und Personenparameter aufgehoben werden.

Reparametrisierung der Itemparameter

Sei hierfür

$$\alpha^*_{jq} := \xi_{(Q+2)q}\,\alpha_{jq}$$

für alle Items j und alle Dimensionen q der latenten Variable, wobei ohne Beschränkung der Allgemeinheit $\xi_{(Q+2)q} \neq 0$ für alle q angenommen sei. Das heißt, alle Ladungen der alten Parametrisierung werden mit dem Personenparameter der $(Q+2)$-ten Person auf der jeweiligen latenten Variable multipliziert. In Matrixnotation kann diese Reparametrisierung mit

$$\boldsymbol{\alpha}^*_j := \mathbf{M}\,\boldsymbol{\alpha}_j \tag{2.9}$$

für alle Items j kompakter geschrieben werden, wobei

$$\mathbf{M} = diag(\boldsymbol{\xi}_{Q+2}) = \begin{pmatrix} \xi_{(Q+2)1} & \cdots & \cdots & 0 \\ \vdots & \xi_{(Q+2)2} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \xi_{(Q+2)Q} \end{pmatrix} \tag{2.10}$$

ist. Die Reparametrisierung kann somit als Rotation der Ladungsmatrix des Modells interpretiert werden.

Reparametrisierung der Personenparameter

Da $\mathbf{M}$ in (2.10) eine Diagonalmatrix ist, ist sie invertierbar und die Inverse durch

$$\mathbf{M}^{-1} = \begin{pmatrix} \frac{1}{\xi_{(Q+2)1}} & \cdots & \cdots & 0 \\ \vdots & \frac{1}{\xi_{(Q+2)2}} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \frac{1}{\xi_{(Q+2)Q}} \end{pmatrix}$$

gegeben. Außerdem gilt $\mathbf{M}^\top = \mathbf{M}$ und $\mathbf{M}^{-\top} = \mathbf{M}^{-1}$. Sei nun

$$\boldsymbol{\xi}_i^* := \mathbf{M}^{-1}\,\boldsymbol{\xi}_i \tag{2.11}$$

für alle Personen i. Dies bedeutet, dass alle Personenparameter der alten Parametrisierung durch den Personenparameter der $(Q+2)$-ten Person auf der jeweiligen latenten Variable geteilt werden:

$$\xi_{jq}^* := \frac{\xi_{jq}}{\xi_{(Q+2)q}} \tag{2.12}$$

Eigenschaften des reparametrisierten Modells

Die reparametrisierten Personenparameter $\boldsymbol{\xi}_i^*$ unterliegen anderen Restriktionen als die ursprünglichen Personenparameter $\boldsymbol{\xi}_i$. Wegen (2.12) gilt nun $\boldsymbol{\xi}_1^* = (\xi_{11}^*, 0, \ldots, 0)^\top$, $\boldsymbol{\xi}_2^* = (0, \xi_{22}^*, 0, \ldots, 0)^\top, \ldots, \boldsymbol{\xi}_Q^* = (0, \ldots, 0, \xi_{QQ}^*)^\top$, $\boldsymbol{\xi}_{Q+1}^* = (0, \ldots, 0)^\top$ und $\boldsymbol{\xi}_{Q+2}^* = (1, \ldots, 1)^\top$ für die ersten $(Q+2)$ Personen. In den ersten Q Personenparametervektoren wird somit jeweils eine Restriktion gelockert. Die Fixierung des $(Q+1)$-ten Personenparametervektors auf den Nullvektor bleibt erhalten. Für den $(Q+2)$-ten Personenparametervektor ergeben sich in der neuen Parametrisierung durch die Fixierung auf den Einheitsvektor Q neue Restriktionen. Die Anzahl der restringierten Personenparameter in der neuen Parametrisierung ist somit genau wie in der alten Parametrisierung $Q(Q-1)+2Q = Q^2 - Q + 2Q = Q(Q+1)$.

Die Eigenschaften der nicht restringierten Personenparameter bleiben durch die Reparametrisierung erhalten. Insbesondere gilt immer noch $\boldsymbol{\xi}_i^* \in \mathbb{R}^Q$, da die Personenparameter in (2.11) lediglich durch eine Konstante geteilt werden. Da die Reparametrisierung zudem eine stetige bijektive Abbildung ist und somit topologische Eigenschaften erhält (Waldmann, 2014), liegen die neuen Personenparameter immer noch dicht in $\mathbb{R}^Q$.

Die Schwierigkeitsparameter sind von der Reparametrisierung nicht betroffen, weisen also gleiche Eigenschaften wie im urspünglichen Modell auf.

Die Ladungsparameter werden durch (2.9) von den nicht-negativen reellen Zahlen $\mathbb{R}_+^Q$ in die kompletten reellen Zahlen $\mathbb{R}^Q$ abgebildet. Da die Reparametrisierung auch in diesem Fall stetig und bijektiv ist, liegen die neuen Parameter dicht in den kompletten reellen Zahlen (Waldmann, 2014). Aus $\boldsymbol{\alpha}_j \neq 0$ folgt, dass auch $\boldsymbol{\alpha}_j^* \neq 0$ ist.

Die Eigenschaften und Restriktionen der ICC F bleiben unter der Reparametrisierung erhalten, da diese von (2.9) und (2.11) nicht betroffen ist.

Auch die Bedingung, dass unendlich viele Personen und Items vorliegen müssen, gilt nach wie vor.

Die Lösungswahrscheinlichkeiten und die lineare Struktur des latenten Prädiktors bleiben erhalten:

$$\begin{aligned} P(X_{ij} = 1) &= F(\beta_j + \boldsymbol{\alpha}_j^{*\top}\,\boldsymbol{\xi}_i^*) \\ &= F(\beta_j + (\mathbf{M}\,\boldsymbol{\alpha}_j)^\top\,\mathbf{M}^{-1}\,\boldsymbol{\xi}_i) \\ &= F(\beta_j + \boldsymbol{\alpha}_j^\top\,\mathbf{M}\,\mathbf{M}^{-1}\,\boldsymbol{\xi}_i) \\ &= F(\beta_j + \boldsymbol{\alpha}_j^\top\,\boldsymbol{\xi}_i) \end{aligned}$$

Da die Lösungswahrscheinlichkeiten unverändert bleiben und die neuen Parameter Funktionen der alten Parameter sind, ist das Modell in der neuen Parametrisierung identifiziert. Zusammenfassend ergibt sich durch die Reparametrisierung folgendes Modell:

Modell 1. *Sei i ein Personenindex und k ein Itemindex, $\boldsymbol{\xi_i} \in \mathbb{R}^Q$ der Personenparameter der Person i, $Q \in \mathbb{N}$ gleich der Dimension der latenten Variable, $\beta_j \in \mathbb{R}$ der Schwierigkeitsparameter des Items j, $\boldsymbol{\alpha_j} \in \mathbb{R}_+^Q$ der Ladungsparameter des Items j und F eine frei schätzbare Item Characteristic Curve. Seien X_{ij} unabhängige dichotome Zufallsvariablen, die jeweils für die Antwort der Person i auf das Item j stehen. Unter Modell 1 gilt für die Lösungswahrscheinlichkeiten*

$$P(X_{ij} = 1) = F(\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i)$$

mit folgenden Restriktionen:

Restriktion 1. *F ist stetig differenzierbar und streng monoton steigend mit $\lim_{x \to -\infty} F(x) = 0$ und $\lim_{x \to \infty} F(x) = 1$. Zudem ist $F(0) = 0.5$ und $F(1) = p$, wobei $0.5 < p < 1$ ein beliebiger fester Wert ist.*

Restriktion 2. *$\boldsymbol{\alpha}_j \neq 0$ für alle Items j.*

Restriktion 3. *$\boldsymbol{\xi}_1 = (\xi_{11}, 0, \ldots, 0)^\top$, $\boldsymbol{\xi}_2 = (0, \xi_{22}, 0, \ldots, 0)^\top$, $\ldots$, $\boldsymbol{\xi}_Q = (0, \ldots, 0, \xi_{QQ})^\top$, $\boldsymbol{\xi}_{Q+1} = (0, \ldots, 0)^\top$ und $\boldsymbol{\xi}_{Q+2} = (1, \ldots, 1)^\top$*

Restriktion 4. *Es liegen abzählbar unendlich viele Personen und Items vor. Die Itemparameter liegen dicht in $\mathbb{R}^{Q+1}$ und die Personenparameter dicht in $\mathbb{R}^Q$.*

2.4.2 Modell ohne Beschränkung der ICC

Die für die Parameterschätzung ungünstigen Restriktionen der Quantile der ICC können durch Reparametrisierung in Restriktionen bezüglich der Itemparameter übertragen werden.

Reparametrisierung der ICC

Ausgangspunkt für die zweite Reparametrisierung ist wieder das NMIRT Modell mit den von Peress (2012) vorgeschlagenen Restriktionen. Sei nun

$$F^*(x) := F(\beta_1 + \alpha_{11} x) \tag{2.13}$$

wobei β_1 der Schwierigkeitsparameter des ersten Items und α_{11} die Ladung des ersten Items auf die erste latente Variable ist.

Reparametrisierung der Itemparameter

Um die Lösungswahrscheinlichkeiten zu erhalten, müssen die Itemparameter ebenfalls reparametrisiert werden. Sei hierfür

$$\boldsymbol{\alpha}_j^* = \frac{\boldsymbol{\alpha}_j}{\alpha_{11}} \tag{2.14}$$

und

$$\beta_j^* = \frac{\beta_j - \beta_1}{\alpha_{11}} \tag{2.15}$$

für alle Items j.

Eigenschaften des reparametrisierten Modells

Aufgrund der Quantilrestriktionen in der ursprünglichen ICC F folgt

$$F^*\left(-\frac{\beta_1}{\alpha_{11}}\right) = F\left(\beta_1 + \alpha_{11}\left(-\frac{\beta_1}{\alpha_{11}}\right)\right) = F(\beta_1 - \beta_1) = F(0) = 0.5 \tag{2.16}$$

und

$$F^*\left(\frac{1}{\alpha_{11}} - \frac{\beta_1}{\alpha_{11}}\right) = F\left(\beta_1 + \alpha_{11}\left(\frac{1}{\alpha_{11}} - \frac{\beta_1}{\alpha_{11}}\right)\right) = F(\beta_1 + 1 - \beta_1) = F(1) = p. \tag{2.17}$$

F^* ist als Komposition einer linearen (und somit invertierbaren) Funktion mit der invertierbaren Funktion F ebenfalls invertierbar. Aus (2.16) und (2.17) ergibt sich daher

$$F^{*-1}(0.5) = -\frac{\beta_1}{\alpha_{11}} \in \mathbb{R}$$

und

$$F^{*-1}(p) = \frac{1}{\alpha_{11}} - \frac{\beta_1}{\alpha_{11}} \in \mathbb{R}.$$

Die vorher restringierten Quantile sind damit in der neuen Parametrisierung freie Parameter, die Werte in den reellen Zahlen annehmen können. F^* ist also ohne Restriktionen schätzbar.

Die weiteren Eigenschaften von F übertragen sich auf F^*: Als Komposition einer (stetig differenzierbaren) linearen Funktion mit der stetig differenzierbaren Funktion F ist F^* ebenfalls stetig differenzierbar. Die Positivität von α_{11} stellt sicher, dass F^* eine Komposition zweier monotoner Funktionen und somit wiederum monoton ist. Für die links- und rechtsseitigen Grenzwerte gilt wegen $\alpha_{11} > 0$

$$\lim_{x \to -\infty} F^*(x) = \lim_{x \to -\infty} F(\beta_1 + \alpha_{11} x) = 0$$

und

$$\lim_{x \to \infty} F^*(x) = \lim_{x \to \infty} F(\beta_1 + \alpha_{11} x) = 1.$$

Für die Itemparameter ergeben sich aus (2.14) und (2.15) die neuen Restriktionen $\alpha_{11}^* = \frac{\alpha_{11}}{\alpha_{11}} = 1$ und $\beta_1^* = \frac{\beta_1 - \beta_1}{\alpha_{11}} = 0$. In der neuen Parametrisierung sind also für das erste Item der Schwierigkeitsparameter und die Ladung auf die erste latente Variable jeweils auf 0 und 1 fixiert. Die Wertebereiche der übrigen Parameter ändern sich durch (2.14) und (2.15) nicht. Da beide Abbildungen bijektiv und stetig sind, liegen die Itemparameter zudem immer noch dicht in $\mathbb{R}_+^{Q+1}$ (Waldmann, 2014). Aus $\boldsymbol{\alpha}_j \neq 0$ folgt mit (2.14), dass auch $\boldsymbol{\alpha}_j^* \neq 0$ ist. Die Personenparameter sind von der Reparametrisierung nicht betroffen. Auch für dieses Modell müssen unendlich viele Personen und Items vorliegen.

Für die Lösungswahrscheinlichkeiten gilt

$$\begin{aligned} F^*\left(\beta_j^* + \boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i\right) &= F\left(\beta_1 + \alpha_{11}\left(\beta_j^* + \boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i\right)\right) \\ &= F\left(\beta_1 + \alpha_{11}\left(\frac{\beta_j - \beta_1}{\alpha_{11}} + \frac{\boldsymbol{\alpha}_j^\top}{\alpha_{11}}\boldsymbol{\xi}_i\right)\right) \\ &= F\left(\beta_1 + \beta_j - \beta_1 + \boldsymbol{\alpha}_j^\top\boldsymbol{\xi}_i\right) \\ &= F\left(\beta_j + \boldsymbol{\alpha}_j^\top\boldsymbol{\xi}_i\right) \end{aligned}$$

für alle Items j und alle Personen i. Die Reparametrisierung erhält somit sowohl die Lösungswahrscheinlichkeiten als auch die lineare Struktur des latenten Prädiktors. Da alle Parameter des neuen Modells zudem Funktionen der Parameter des ursprünglichen identifizierten Modells sind, ist das Modell ebenfalls identifiziert. Zusammenfassend ergibt sich durch die Reparametrisierung:

Modell 2. *Sei i ein Personenindex und k ein Itemindex, $\boldsymbol{\xi_i} \in \mathbb{R}^Q$ der Personenparameter der Person i, $Q \in \mathbb{N}$ gleich der Dimension der latenten Variable, $\beta_j \in \mathbb{R}$ der Schwierigkeitsparameter des Items j, $\boldsymbol{\alpha_j} \in \mathbb{R}_+^Q$ der Ladungsparameter des Items j und F eine frei schätzbare Item Characteristic Curve. Seien X_{ij} unabhängige dichotome Zufallsvariablen, die jeweils für die Antwort der Person i auf das Item j stehen. Unter Modell 2 gilt für die Lösungswahrscheinlichkeiten*

$$P(X_{ij} = 1) = F(\beta_j + \boldsymbol{\alpha}_j^\top\,\boldsymbol{\xi}_i)$$

mit folgenden Restriktionen:

Restriktion 1. *F ist stetig differenzierbar und streng monoton steigend mit $\lim_{x\to-\infty} F(x) = 0$ und $\lim_{x\to\infty} F(x) = 1$.*

Restriktion 2. *$\alpha_{jq} > 0$ für alle Items j und alle Dimensionen q, $\boldsymbol{\alpha}_j \neq 0$ für alle Items j. $\alpha_{11} = 1$ und $\beta_1 = 0$.*

Restriktion 3. *$\boldsymbol{\xi}_1 = (1, 0, \ldots, 0)^\top$, $\boldsymbol{\xi}_2 = (0, 1, 0, \ldots, 0)^\top$, $\ldots$, $\boldsymbol{\xi}_Q = (0, \ldots, 0, 1)^\top$ und $\boldsymbol{\xi}_{Q+1} = (0, \ldots, 0)^\top$*

Restriktion 4. *Es liegen abzählbar unendlich viele Personen und Items vor. Die Itemparameter liegen dicht in $\mathbb{R} \times \mathbb{R}_+^Q$ und die Personenparameter liegen dicht in R^Q.*

2.4.3 Modell ohne Beschränkung der Ladungen und der ICC

Die im letzten Abschnitt durch (2.13), (2.14) und (2.15) definierte Reparametrisierung kann auch auf das in Abschnitt 2.4.1 hergeleitete Modell 1 ohne Ladungsbeschränkung angewendet werden, falls als erstes Item ein Item gewählt wird, für das $\alpha_{11} > 0$ gilt. Auch hier wirkt sich die Reparametrisierung lediglich auf die ICC und die Itemparameter aus. Die Auswirkungen der Reparametrisierung auf diese Parameter sind identisch mit den im letzten Abschnitt beschriebenen Auswirkungen: Die ICC verliert ihre Quantilrestriktionen, der Schwierigkeitsparameter des ersten Items wird auf 0 fixiert und die Ladung des ersten Items auf die erste latente Variable wird auf 1 gesetzt. Die Reparametrisierung verändert zudem den Wertebereich der Itemparameter nicht, so dass die Ladungsparameter auch nach der Reparametrisierung unbeschränkt sind. Die zusätzlichen eher technischen Bedingungen bleiben alle erhalten. Es resultiert das folgende Modell:

Modell 3. *Sei i ein Personenindex und k ein Itemindex, $\boldsymbol{\xi_i} \in \mathbb{R}^Q$ der Personenparameter der Person i, $Q \in \mathbb{N}$ gleich der Dimension der latenten Variable, $\beta_j \in \mathbb{R}$ der Schwierigkeitsparameter des Items j, $\boldsymbol{\alpha_j} \in \mathbb{R}^Q$ der Ladungsparameter des Items j und F eine frei schätzbare Item Characteristic Curve. Seien X_{ij} unabhängige dichotome Zufallsvariablen, die jeweils für die Antwort der Person i auf das Item j stehen. Unter Modell 3 gilt für die Lösungswahrscheinlichkeiten*

$$P(X_{ij} = 1) = F(\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i)$$

mit den Restriktionen:

Restriktion 1. *F ist stetig differenzierbar und streng monoton steigend mit $\lim_{x\to-\infty} F(x) = 0$ und $\lim_{x\to\infty} F(x) = 1$.*

Restriktion 2. *$\boldsymbol{\alpha}_j \neq 0$ für alle Items j. $\alpha_{11} = 1$ und $\beta_1 = 0$.*

Restriktion 3. *$\boldsymbol{\xi}_1 = (\xi_{11}, 0, \ldots, 0)^\top$, $\boldsymbol{\xi}_2 = (0, \xi_{22}, 0, \ldots, 0)^\top$, $\ldots$, $\boldsymbol{\xi}_Q = (0, \ldots, 0, \xi_{QQ})^\top$, $\boldsymbol{\xi}_{Q+1} = (0, \ldots, 0)^\top$ und $\boldsymbol{\xi}_{Q+2} = (1, \ldots, 1)^\top$*

Restriktion 4. *Es liegen abzählbar unendlich viele Personen und Items vor. Die Itemparameter liegen dicht in $\mathbb{R}^{Q+1}$ und die Personenparameter dicht in $\mathbb{R}^Q$.*

Kapitel 3

Nonparametrische Bayes-Inferenz

3.1 Parametrische Bayes-Inferenz

Ausgangssituation der parametrischen Bayes-Inferenz (siehe z.B. Gelman et al., 2014; Schervish, 1995) ist eine Stichprobe der Größe $n \in \mathbb{N}$ bestehend aus Daten $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, welche als Realisationen von Zufallsvariablen $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ auf einem Wahrscheinlichkeitsraum $(\Omega, \sigma(\Omega), P)$ aufgefasst werden. Diese Daten seien je nach Dimension der $\mathbf{x}_i$ in einem Vektor oder einer Matrix $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\top$ zusammengefasst. Von Interesse ist ein Parametervektor $\boldsymbol{\theta}$ mit Parameterraum Θ, der ebenfalls als Zufallsvariable auf $(\Omega, \sigma(\Omega), P)$ verstanden und durch eine Likelihoodfunktion $f(\mathbf{x}|\boldsymbol{\theta})$ zu den Daten in Beziehung gesetzt wird. In diese Likelihood gehen Annahmen über die Verteilung der Daten bei gegebenen Parametern ein. Zudem muss eine durch die Priori-Dichte $f(\boldsymbol{\theta})$ charakterisierte Priori-Verteilung festgelegt werden, in die bereits vor der Erhebung der Daten bestehendes Wissen über $\boldsymbol{\theta}$ eingeht. Parameter, von denen die Form dieser Priori-Verteilung abhängt, werden Hyperparameter genannt.

Die statistische Inferenz basiert unter diesen Voraussetzungen auf der bedingten Verteilung der Parameter in $\boldsymbol{\theta}$ gegeben die Daten $\mathbf{x}$. Diese Verteilung wird Posteriori-Verteilung genannt. Ihre Dichtefunktion - die Posteriori-Dichte - lässt sich mithilfe des Satzes von Bayes herleiten:

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}, \boldsymbol{\theta})}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\mathbf{x})} \tag{3.1}$$

Für einen Beweis siehe beispielsweise Schervish (1995). Der Zähler

$$f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \tag{3.2}$$

ist das Produkt aus Likelihood und Priori-Dichte. Er bestimmt die Form der Posteriori-Dichte (Gelman et al., 2014). Der Nenner

$$f(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\Theta} f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{3.3}$$

ist lediglich eine Normierungskonstante, die dafür sorgt, dass das Integral der Posteriori-Dichte über dem Parameterraum gleich Eins ist, und somit die Posteriori-Dichte eine Wahrscheinlichkeitsdichte ist (Gelman et al., 2014).

Auf der Basis der Posteriori-Verteilung können dann Punkt- und Intervallschätzer und Hypothesentests für die Parameter konstruiert werden, sowie für gegebene Verlustfunktionen Entscheidungsregeln hergeleitet werden (Gelman et al., 2014). Beispielsweise kann als Punktschätzer für $\boldsymbol{\theta}$ der Erwartungswert von $\boldsymbol{\theta}$ unter der Posteriori-Verteilung

$$E(\boldsymbol{\theta}|\mathbf{x}) = \int_{\Theta} \boldsymbol{\theta} f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$$

berechnet werden. Dieser wird Expected-A-Posteriori-Schätzer (EAP) genannt (Gelman et al., 2014).

Seien zum Beispiel (Schervish, 1995) die Beobachtungen $X_1, X_2, \ldots, X_n$ unabhängig Bernoulli-verteilt mit Parameter π. Die Verteilung jeder einzelnen Zufallsvariable X_i, $i = 1, 2, \ldots, n$, hat somit die Dichte

$$f(x_i|\pi) = Ber(x_i; \pi) = \pi^{x_i}(1-\pi)^{1-x_i}. \tag{3.4}$$

Aus der Unabhängigkeitsannahme und (3.4) folgt dann für die Likelihood

$$\begin{aligned} f(\mathbf{x}|\pi) &= \prod_{i=1}^{n} f(x_i|\pi) \\ &= \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} \\ &= \prod_{i=1}^{n} \pi^{x_i} \prod_{i=1}^{n} (1-\pi)^{1-x_i} \\ &= \pi^{s}(1-\pi)^{n-s} \end{aligned} \tag{3.5}$$

mit $s := \sum_{i=1}^{n} x_i$. Als Priori-Verteilung für π muss eine auf dem Intervall $[0, 1]$ definierte Verteilung gewählt werden. Hier bietet sich die Betaverteilung an, welche durch die Dichte

$$f(\pi) = Beta(\pi; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \tag{3.6}$$

mit $B(\alpha, \beta)$ als Betafunktion charakterisiert ist. Die Hyperparameter α und β bestimmen die konkrete Form der Betaverteilung und müssen auf Basis des Vorwissens über π gewählt werden. Für $\alpha = \beta = 1$ ergibt sich z.B. eine Gleichverteilung auf $[0, 1]$.

Bei der Bestimmung der Posteriori-Dichte ist es häufig zielführend, diese zunächst nur bis auf eine multiplikative Konstante zu berechnen (Gelman et al., 2014). Dies bedeutet erstens, dass die Normierungskonstante (3.3) vernachlässigt wird und zweitens, dass im

Zähler (3.2) alle multiplikativen Terme ignoriert werden, die nicht von Elementen des Parametervektors abhängen. Für das Beispiel ergibt sich aus (3.5) und (3.6):

$$\begin{aligned} f(\pi|\mathbf{x}) &= \frac{f(\mathbf{x}|\pi)f(\pi)}{f(\mathbf{x})} \\ &\propto f(\mathbf{x}|\pi)f(\pi) \\ &= \pi^s(1-\pi)^{n-s}\frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &\propto \pi^s(1-\pi)^{n-s}\pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &= \pi^{s+\alpha-1}(1-\pi)^{n-s+\beta-1} \end{aligned}$$

Mit $\alpha^* := s + \alpha$ und $\beta^* := n - s + \beta$ ist also

$$f(\pi|\mathbf{x}) \propto \pi^{\alpha^*-1}(1-\pi)^{\beta^*-1}.$$

Diese Funktion ist proportional zur Dichte einer Betaverteilung mit Parametern α^* und β^*. Die Normierungskonstante ist somit $B(\alpha^*, \beta^*)$ und die komplette Posteriori-Dichte ist

$$f(\pi|\mathbf{x}) = \frac{1}{B(\alpha^*,\beta^*)}\pi^{\alpha^*-1}(1-\pi)^{\beta^*-1}.$$

In Fällen, in denen die Posteriori-Verteilung der gleichen Verteilungsfamilie wie die Priori-Verteilung angehört, spricht man auch davon, dass die Priori-Verteilung zur Likelihood konjugiert ist (Gelman et al., 2014). Der Erwartungswert einer betaverteilten Größe mit Parametern α^* und β^* ist $\alpha^*/(\alpha^* + \beta^*)$. Somit ist der EAP für π durch

$$E(\pi|\mathbf{x}) = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{s+\alpha}{s+\alpha+n-s+\beta} = \frac{s+\alpha}{\alpha+\beta+n}$$

gegeben (Gelman et al., 2014).

3.2 Nonparametrische Bayes-Inferenz

In der parametrischen Bayes-Inferenz gehen Annahmen bezüglich der auf die Parameter $\boldsymbol{\theta}$ bedingten Verteilung der Beobachtungen $\mathbf{x}$ in die Likelihoodfunktion $f(\mathbf{x}|\boldsymbol{\theta})$ ein. Sind diese Annahmen nicht erfüllt, ist das Modell fehlspezifiziert und die Inferenz unter Umständen verzerrt (Müller, Quintana, Jara & Hanson, 2015). Im Rahmen der nonparametrischen Bayes-Inferenz kann dieses Problem gelöst werden, indem Priori-Verteilungen nicht auf dem Parameterraum einer vorgegebenen Verteilungsfamilie, sondern direkt auf einem sehr großen Raum von Wahrscheinlichkeitsmaßen spezifiziert werden (Müller et al., 2015). Ist zum Beispiel in dem parametrischen Modell

$$\begin{aligned} X_i|\mu,\sigma^2 &\overset{iid}{\sim} N(\mu,\sigma^2) \\ \mu &\sim \text{Priori-Verteilung für } \mu \\ \sigma^2 &\sim \text{Priori-Verteilung für } \sigma^2 \end{aligned}$$

die Normalverteilungsannahme für die Zufallsvariablen X_i verletzt, könnte man ein nonparametrisches bayesianisches Modell

$$\begin{aligned} X_i|P &\overset{iid}{\sim} P \\ P &\sim \text{Priori-Verteilung für } P \end{aligned}$$

formulieren, in dem P Werte in einem sehr großen Raum $\mathcal{P}$ von Wahrscheinlichkeitsmaßen annehmen kann, der nicht nur die Menge aller Normalverteilungen enthält. In diesem Kontext stellt sich die Frage, wie Wahrscheinlichkeitsverteilungen auf einer Menge von Wahrscheinlichkeitsverteilungen konstruiert werden können. Zunächst sind einige Begriffe zu definieren: Sei $(\Omega, \sigma(\Omega), \mathbf{P})$ ein Wahrscheinlichkeitsraum, der für das gesamte Kapitel zugrundegelegt wird. Sei $\mathcal{B}$ die Borel σ-Algebra auf $\mathbb{R}$, $\mathcal{P}$ eine Menge von Wahrscheinlichkeitsmaßen auf $(\mathbb{R}, \mathcal{B})$ und $\sigma(\mathcal{P})$ eine geeignete σ-Algebra auf $\mathcal{P}$. Für die Existenz einer solchen siehe beispielsweise Schervish (1995). Eine messbare Funktion $P : \Omega \to \mathcal{P}$ heißt dann zufälliges Wahrscheinlichkeitsmaß (Schervish, 1995). P kann als Zufallsvariable verstanden werden, deren Realisationen Wahrscheinlichkeitsmaße auf $(\mathbb{R}, \mathcal{B})$ sind. Für jede Menge $A \in \mathcal{B}$ ist dann $P(A)$ eine Zufallsvariable mit dem Interval $[0, 1]$ als Wertebereich, deren Verteilung durch die Verteilung auf $\mathcal{P}$ bestimmt wird (Schervish, 1995).

3.2.1 Polya Trees

Die Polya Tree (PT) Verteilung induziert eine Verteilung auf einer Menge $\mathcal{P}$ von Wahrscheinlichkeitsmaßen durch eine Verteilung auf einer Menge von (unendlich feinen) Histogrammen (Müller et al., 2015). Eine Definition der allgemeinen PT Verteilung findet sich zum Beispiel in Schervish (1995). In der vorliegenden Arbeit wird die PT Verteilung in einer bestimmten Parametrisierung eingeführt (Hanson, 2006; Müller et al., 2015).

Sei P_0 eine parametrische Basisverteilung auf $(\mathbb{R}, \mathcal{B})$ mit Dichte f_0, Verteilungsfunktion F_0 und Quantilsfunktion F_0^{-1}. Für jedes $l \in \mathbb{N}$, und jedes $y = 1, 2, 3, \ldots, 2^l$ sei $e_l(y) = \epsilon_1\epsilon_2 \cdots \epsilon_l$ die l-bit Binärrepräsentation von $y - 1$. So ist zum Beispiel $e_3(3)$ die 3-bit Binärrepräsentation von $3 - 1 = 2$ und somit $e_3(3) = \epsilon_1\epsilon_2\epsilon_3 = 010$. Außerdem werden für jedes l die folgenden Intervalle definiert (Hanson, 2006):

$$I_{e_l(y)} =]F_0^{-1}((y-1)2^{-l}), F_0^{-1}(y2^{-l})] \tag{3.7}$$

für $y = 1, 2, 3, \ldots, 2^l - 1$ und

$$I_{e_l(y)} =]F_0^{-1}((y-1)2^{-l}), F_0^{-1}(1)[\tag{3.8}$$

für $y = 2^l$.

Sei die Basisverteilung P_0 zum Beispiel die Standardnormalverteilung und bezeichne Φ^{-1} deren Quantilsfunktion, so dass $F_0^{-1} = \Phi^{-1}$. Für $l = 1$ ist dann $2^l = 2^1 = 2$ und somit $y = 1, 2$. Für diese beiden y-Werte ist $e_1(1) = 0$ und $e_1(2) = 1$. Die zu diesen Binärindizes

gehörigen Intervalle sind

$$\begin{aligned} I_0 &= I_{e_1(1)} \\ &=]F_0^{-1}((1-1)2^{-1}), F_0^{-1}(1 \cdot 2^{-1})] \\ &=]\Phi^{-1}(0), \Phi^{-1}(0.5)] \\ &=]-\infty, 0] \end{aligned}$$

und

$$\begin{aligned} I_1 &= I_{e_1(2)} \\ &=]F_0^{-1}((2-1)2^{-1}), F_0^{-1}(1)[\\ &=]\Phi^{-1}(0.5), \Phi^{-1}(1)[\\ &=]0, \infty[. \end{aligned}$$

Für $l = 2$ ist $2^l = 2^2 = 4$ und $y = 1, 2, 3, 4$. Die Binärindizes sind dementsprechend $e_2(1) = 00$, $e_2(2) = 01$, $e_2(3) = 10$ und $e_2(4) = 11$. Als Intervalle ergeben sich

$$\begin{aligned} I_{00} &= I_{e_2(1)} \\ &=]F_0^{-1}((1-1)2^{-2}), F_0^{-1}(1 \cdot 2^{-2})] \\ &=]\Phi^{-1}(0), \Phi^{-1}(0.25)] \\ &\approx]-\infty, -0.674] \end{aligned}$$

und

$$\begin{aligned} I_{01} &= I_{e_2(2)} \\ &=]F_0^{-1}((2-1)2^{-2}), F_0^{-1}(2 \cdot 2^{-2})] \\ &=]\Phi^{-1}(0.25), \Phi^{-1}(0.5)[\\ &\approx]-0.674, 0] \end{aligned}$$

sowie

$$\begin{aligned} I_{10} &= I_{e_2(3)} \\ &=]F_0^{-1}((3-1)2^{-2}), F_0^{-1}(3 \cdot 2^{-2})] \\ &=]\Phi^{-1}(0.5), \Phi^{-1}(0.75)] \\ &\approx]0, 0.674] \end{aligned}$$

und

$$\begin{aligned} I_{11} &= I_{e_2(4)} \\ &=]F_0^{-1}((4-1)2^{-2}), F_0^{-1}(1)[\\ &=]\Phi^{-1}(0.75), \Phi^{-1}(1)[\\ &\approx]0.674, \infty[. \end{aligned}$$

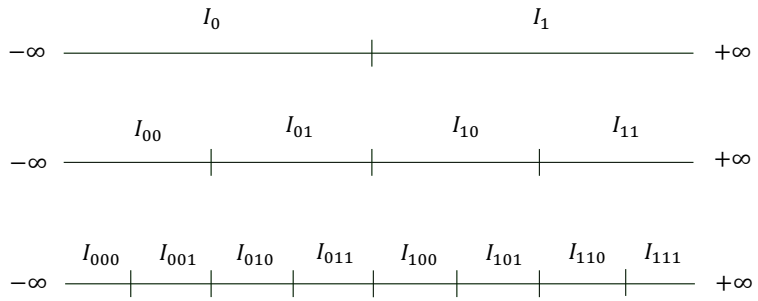

Abbildung 3.1: Graphische Darstellung der Intervalle der ersten zwei Ebenen eines Polya Trees

Allgemein stellen die Intervalle für jedes l eine Partition von $\mathbb{R}$ dar. Die Intervallgrenzen entsprechen hierbei für jedes l bestimmten Quantilen der Basisverteilung P_0. Mit wachsendem l werden diese Partitionen immer feiner: Auf der ersten Ebene ($l = 1$) gibt es nur zwei Intervalle, wobei die Grenze der Median der Basisverteilung ist. Die zweite Ebene ($l = 2$) enthält vier Intervalle, deren Grenzen die Quartile der Basisverteilung sind, etc. Zudem weisen die Intervalle per Definition eine Baumstruktur auf. Zu jedem Intervall auf Ebene l gibt es zwei aneinandergrenzende Intervalle auf Ebene $l + 1$, deren Vereinigung diesem Intervall entspricht (Hanson, 2006):

$$I_{e_l(y)} = I_{\epsilon_1\epsilon_2\cdots\epsilon_l} = I_{\epsilon_1\epsilon_2\cdots\epsilon_l 0} \cup I_{\epsilon_1\epsilon_2\cdots\epsilon_l 1}$$

Auf jeder Ebene $l+1$ wird also jedes Intervall der Ebene l in zwei kleinere Intervalle geteilt. In Abbildung 3.1 ist diese Baumstruktur für die ersten drei Ebenen dargestellt.

Im obigen Beispiel ist

$$I_0 =]-\infty, 0] = \ I_{00} \cup I_{01} \approx]-\infty, -0.674] \cup]-0.674, 0]$$

und

$$I_1 =]0, \infty[= \ I_{10} \cup I_{11} \approx]0, 0.674] \cup]0.674, \infty[$$

Die durch (3.7) und (3.8) definierte Menge von Intervallen $I_{e_l(y)}$ sei mit Π bezeichnet. Jedem Intervall $I_{e_l(y)}$ in Π sei zudem eine nichtnegative Zufallsvariable $H_{e_l(y)}$ und ein Parameter $\lambda_{e_l(y)} \in \mathbb{R}_+$ zugeordnet (Hanson, 2006). Die Zufallsvariablen $H_{e_l(y)}$ seien in einem Vektor $\mathbf{h}$ zusammengefasst.

Unter Verwendung dieser Notation kann nun die Polya Tree Verteilung definiert werden (Hanson, 2006; Müller et al., 2015): Sei P ein zufälliges Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B})$. P folgt einer Polya Tree Verteilung mit Basisverteilung P_0 und Parametern $\lambda_{e_l(y)}$, kurz $P \sim PT(\lambda_{e_l(y)}, P_0)$, falls zu jedem Intervall in Π eine Zufallsvariable $H_{e_l(y)}$ existiert, so dass alle $H_{e_l(y)0}$ unabhängig sind mit

$$H_0 \sim Beta(\lambda_0, \lambda_1) \tag{3.9}$$
$$H_1 = 1 - H_0 \tag{3.10}$$

auf der ersten Ebene und

$$H_{e_l(y)0} \sim Beta(\lambda_{e_l(y)0}, \lambda_{e_l(y)1}) \tag{3.11}$$
$$H_{e_l(y)1} = 1 - H_{e_l(y)0} \tag{3.12}$$

auf den tieferen Ebenen und falls für jedes Intervall

$$P(I_{\epsilon_1\epsilon_2\cdots\epsilon_l}) = \prod_{i=1}^{l} H_{\epsilon_1\epsilon_2\cdots\epsilon_i} \tag{3.13}$$

gilt.

Diese Definition sei am Beispiel der ersten zwei Ebenen eines PT mit der Standardnormalverteilung als Basisverteilung veranschaulicht: Die Intervalle der ersten Ebene sind $I_0 =]-\infty, 0]$ und $I_1 =]0, \infty[$ und die Intervalle auf der zweiten Ebene (gerundet) $I_{00} =]-\infty, -0.674]$, $I_{01} =]-0.674, 0]$, $I_{10} =]0, 0.674]$ und $I_{11} =]0.674, \infty[$ (siehe oben). Jedem dieser Intervalle ist ein Parameter $\lambda_{e_l(y)}$ und eine Zufallsvariable $H_{e_l(y)}$ zugeordnet. Die Parameter bestimmen die Verteilung der zu ihnen gehörigen Zufallsvariablen: H_0 wäre im Beispiel betaverteilt mit Parametern λ_0 und λ_1 und H_1 würde sich durch $1 - H_0$ ergeben. Analog würde H_{00} wegen (3.11) einer Betaverteilung mit Parametern λ_{00} und λ_{01} folgen und H_{01} wäre $1 - H_{00}$. Schließlich wäre H_{10} betaverteilt mit Parametern λ_{10} und λ_{11} und H_{11} würde sich aus $1 - H_{10}$ ergeben. Hat sich dann beispielsweise die Zufallsvariable H_0 auf der ersten Ebene des Baumes in dem Wert $h_0 = 0.4$ realisiert, ist $h_1 = 1 - h_0 = 0.6$. Falls sich ferner auf der zweiten Ebene die Zufallsvariablen H_{00} und H_{10} in $h_{00} = 0.3$ und $h_{10} = 0.5$ realisiert haben, ist $h_{01} = 1 - h_{00} = 0.7$ und $h_{11} = 1 - h_{10} = 0.5$. Die Wahrscheinlichkeiten aller Intervalle der ersten beiden Ebenen unter der Realisation des zufälligen Wahrscheinlichkeitsmaßes P können dann mithilfe von (3.13) berechnet werden. Diese Formel besagt, dass die Wahrscheinlichkeit für jedes Intervall durch das Produkt der ihm zugeordneten Variable $H_{e_l(y)}$ und der Variablen aller Intervalle auf den höheren Ebenen, die das Intervall enthalten, berechnet werden kann. Auf der ersten Ebene des Baumes bedeutet dies $P(I_0) = h_0 = 0.4$ und $P(I_1) = h_1 = 0.6$. Für die Intervalle der zweiten Ebene ist $P(I_{00}) = h_0 \cdot h_{00} = 0.4 \cdot 0.3 = 0.12$, $P(I_{01}) = h_0 \cdot h_{01} = 0.4 \cdot 0.7 = 0.28$, $P(I_{10}) = h_1 \cdot h_{10} = 0.6 \cdot 0.5 = 0.3$ und $P(I_{11}) = h_1 \cdot h_{11} = 0.6 \cdot 0.5 = 0.3$. In Abbildung 3.2 ist dieses Beispiel graphisch dargestellt.

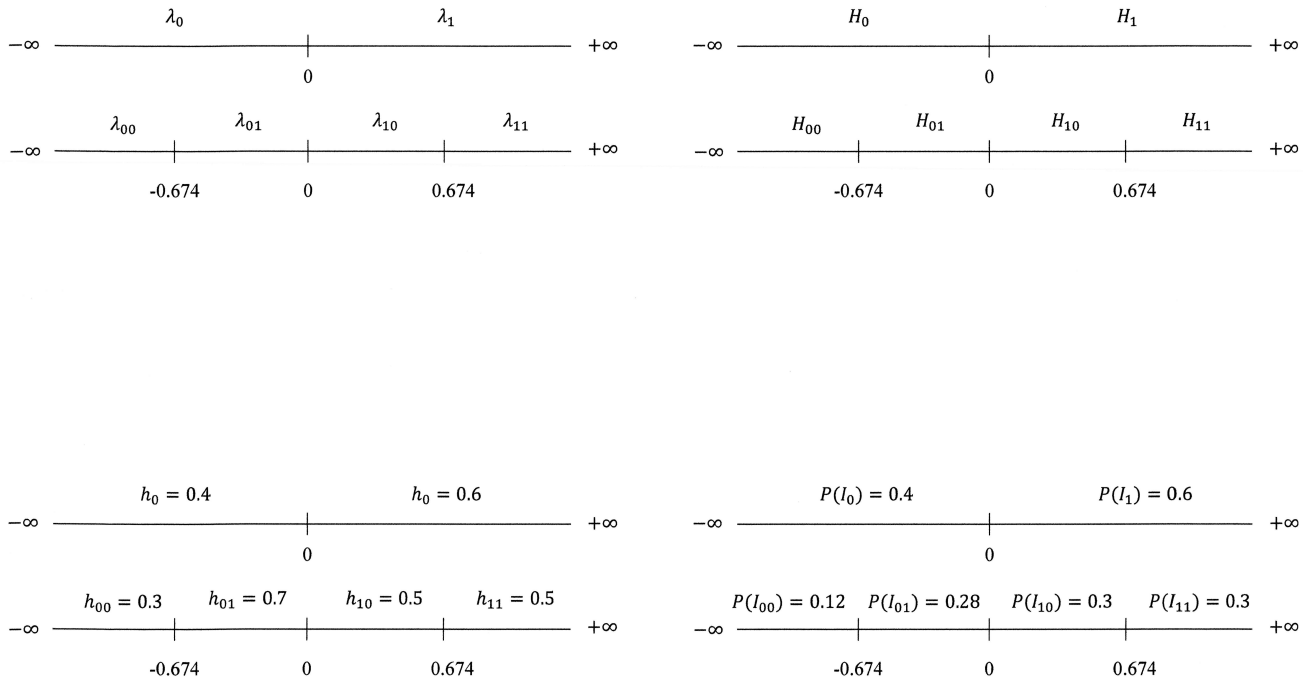

Abbildung 3.2: Graphische Darstellung der ersten zwei Ebenen eines Polya Trees. Links oben sind die den Intervallen zugeordneten Parameter veranschaulicht. Rechts oben befinden sich die den Intervallen zugeordneten Zufallsvariablen, deren Realisationen in der Graphik links unten zu finden sind. Rechts unten sind schließlich die Wahrscheinlichkeiten der Intervalle unter der Realisation des zufälligen Wahrscheinlichkeitsmaßes zu sehen.

Auf jeder Ebene l definiert die Realisation eines zufälligen Wahrscheinlichkeitsmaßes mit PT Verteilung über die Realisationen der Zufallsvariablen $H_{e_l(y)}$ in $\mathbf{h}$ ein Histogramm auf den Intervallen $I_{e_l(y)}$ dieser Ebene. Da $l \in \mathbb{N}$ und der Baum somit unendlich tief ist und je nach Realisation von $\mathbf{h}$ auf jeder Ebene ein zufälliges Histogramm resultiert, definiert der PT eine Verteilung auf einer Menge von unendlich feinen, zufälligen Histogrammen (Müller et al., 2015).

Die Wahl der Parameter $\lambda_{e_l(y)}$ hat unter anderem Auswirkungen auf die Eigenschaften der Realisationen von P. Da es in der Praxis sehr aufwendig wäre, jeden der abzählbar unendlich vielen Parameter $\lambda_{e_l(y)}$ einzeln festzulegen, wird von Hanson (2006) die Parametrisierung

$$\lambda_{e_l(y)} = \lambda l^2 \tag{3.14}$$

mit $\lambda \in \mathbb{R}_+$ vorgeschlagen. Für diesen Fall kann gezeigt werden, dass die Realisationen von P (fast sicher) stetige Verteilungen sind (Hanson, 2006). Zudem ist nur noch ein Parameter λ festzulegen, der eine intuitive Interpretation besitzt (siehe unten). Ein PT mit dieser Parametrisierung sei mit $PT(\lambda, P_0)$ bezeichnet. An dieser Stelle sei auch erwähnt, dass durch die Wahl bestimmter Werte für $\lambda_{e_l(y)}$ Quantilrestriktionen auf die Realisationen des PT gelegt werden können (Hanson, 2006).

In der Parametrisierung (3.14) folgt für alle Intervalle $I_{e_l(y)}$ in Π:

$$E(P(I_{e_l(y)})) = P_0(I_{e_l(y)}) \tag{3.15}$$

(Hanson, 2006). Die Basisverteilung des PT kann also in gewisser Weise als Erwartungswert des zufälligen Wahrscheinlichkeitsmaßes interpretiert werden. Ebenso gilt unter (3.14) für alle Intervalle $I_{e_l(y)}$ in Π:

$$Var(P(I_{e_l(y)})) = 4^{-l}\left[\prod_{i=1}^{l}\frac{2\lambda i^2 + 2}{2\lambda i^2 + 1} - 1\right]$$

(Hanson, 2006). Für $\lambda \to \infty$ konvergiert diese Varianz gegen 0 und die Realisationen des PT geht wegen (3.15) gegen die Basisverteilung P_0 (Hanson, 2006). λ kann somit als Präzisionsparameter interpretiert werden, der das Ausmaß des Vertrauens darin bestimmt, dass Realisationen nahe an der Basisverteilung liegen.

In der Praxis wird die Polya Tree Verteilung häufig auf einer vorgegebenen Ebene $l = L$ trunkiert (Hanson, 2006). Man spricht dann von einer endlichen oder trunkierten Polya Tree Verteilung (TPT) und schreibt je nach Parametrisierung $TPT(\lambda_{e_l(y)}, P_0)$ oder $TPT(\lambda, P_0)$. Der TPT hat nunmehr endlich viele Parameter und kann somit als (sehr flexibles) parametrisches Modell interpretiert werden (Hanson, 2006).

In diesem Fall muss noch festgelegt werden, dass sich die Realisationen des TPT auf Teilmengen der Intervalle der untersten Ebene L wie die Basisverteilung verhalten. Legt man fest, dass die Realisationen auf diesen Teilmengen der Basisverteilung folgen, führt dies unter anderem dazu, dass die Dichte einer Realisation einer Polya Tree Verteilung die folgende Form aufweist:

$$f_{\mathbf{h}}(x) = 2^L p_{\mathbf{h}}(i(x)) f_0(x), \tag{3.16}$$

wobei $p_{\mathbf{h}}(y)$ die bei gegebenem $\mathbf{h}$ über (3.13) berechenbare Wahrscheinlichkeit des y-ten Intervalls auf der untersten Ebene L des trunkierten TPT bezeichnet. Die Funktion $i(x)$ gibt an, im wievielten Intervall auf der untersten Ebene des TPD x liegt (Hanson, 2006). Sind also Realisationen der Zufallsvariablen $\mathbf{h}$ gegeben, legen diese über die Dichte (3.16) die Form der Realisation des TPT fest. Aus (3.16) ergibt sich durch Integration die Verteilungsfunktion einer Realisation der Polya Tree Verteilung als

$$F_{\mathbf{h}}(x) = \left[\sum_{y=1}^{i(x)-1} p_{\mathbf{h}}(y)\right] + p_{\mathbf{h}}(i(x)) \left[2^L F_0(x) - i(x) + 1\right] \tag{3.17}$$

(Hanson, 2006). Durch Invertierung von $F_{\mathbf{h}}$ kann zudem die Quantilsfunktion einer Realisation des TPT angegeben werden:

$$F_{\mathbf{h}}^{-1}(q) = F_0^{-1}\left[\frac{q - \sum_{y=1}^{c} p_{\mathbf{h}}(y) + c p_{\mathbf{h}}(c)}{2^L p_{\mathbf{h}}(c)}\right], \tag{3.18}$$

wobei c diejenige natürliche Zahl ist, für die

$$\sum_{y=1}^{c-1} p_{\mathbf{h}}(y) < q \leq \sum_{y=1}^{c} p_{\mathbf{h}}(y)$$

gilt (Hanson, 2006).

In Abbildung 3.3 sind die Verteilungsfunktionen von 30 zufälligen Realisationen von P basierend auf einem auf Ebene $L = 10$ trunkierten TPT mit Standardnormalverteilung als Basisverteilung einmal mit $\lambda = 1$ und einmal mit $\lambda = 100$ dargestellt. Hier ist deutlich zu sehen, dass die Realisationen des TPT mit hohem λ sehr nah an der Verteilungsfunktion der Basisverteilung liegen.

Ein für die Inferenz sehr wichtiges Resultat ist die Konjugiertheit des (T)PT unter iid Beobachtungen: Seien $X_1, X_2, \ldots, X_n$ Zufallsvariablen mit $X_i|P \overset{iid}{\sim} P$, $i = 1, 2, \ldots, n$, und $P \sim (T)PT(\lambda_{e_l(y)}, P_0)$. Dann ist

$$P|x_1, x_2, \ldots, x_n \sim (T)PT(\lambda^*_{e_l(y)}, P_0) \tag{3.19}$$

mit

$$\lambda^*_{e_l(y)} = \lambda_{e_l(y)} + \sum_{i=1}^{n} I(x_i \in I_{e_l(y)}) \tag{3.20}$$

für alle $\lambda_{e_l(y)}$, wobei $I(\cdot)$ die Indikatorfunktion bezeichnet (Hanson, 2006). Die Posteriori-Verteilung von P ist also wieder ein (T)PT, dessen neue Parameter $\lambda^*_{e_l(y)}$ jeweils den alten Parametern $\lambda_{e_l(y)}$ plus der Anzahl der Beobachtungen x_i, die in das zu ihnen gehörige Intervall $I_{e_l(y)}$ fallen, entsprechen. In Abbildung 3.4 sind die Verteilungsfunktionen von 100 Ziehungen aus einem Posteriori TPT bedingt auf 500 unabhängige Beobachtungen aus einer Mischung von zwei Normalverteilungen dargestellt.

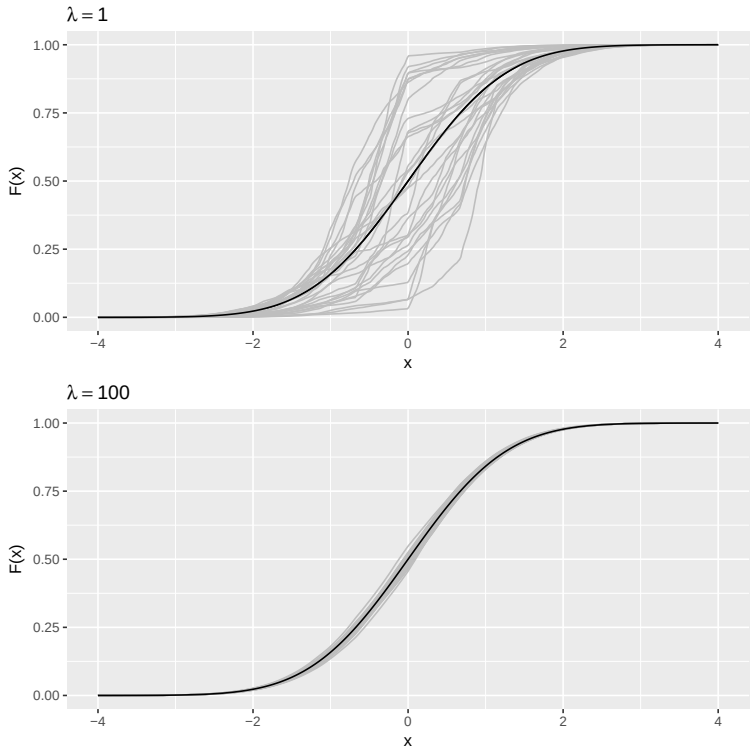

Abbildung 3.3: Ziehungen aus einer TPT Verteilung. In der oberen Graphik wurde $\lambda = 1$ gewählt, während in der unteren $\lambda = 100$ gesetzt wurde. Die schwarze Funktion die Verteilungsfunktion der Basisverteilung, welche in beiden Fällen der Verteilungsfunktion der Standardnormalverteilung entspricht.

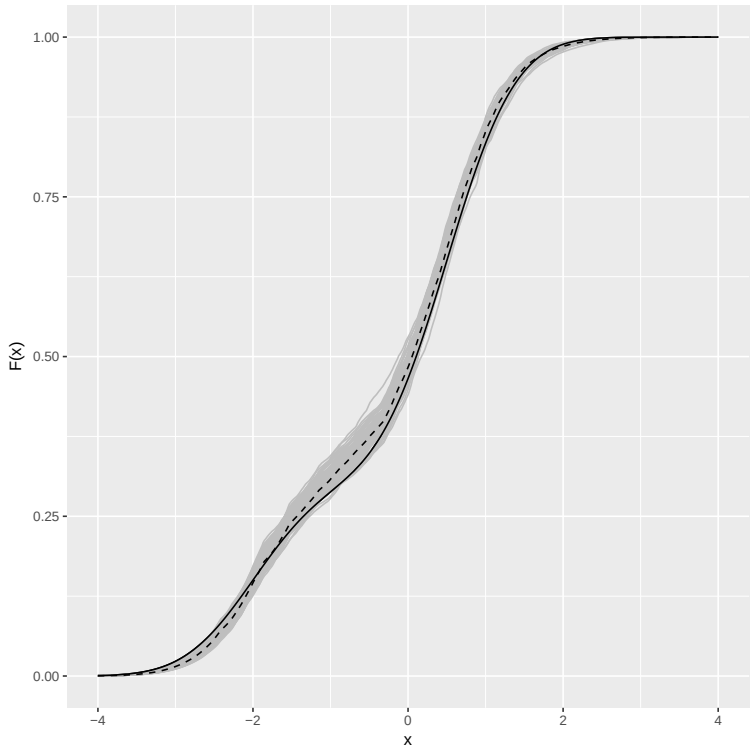

Abbildung 3.4: Ziehungen aus einer TPT Posteriori-Verteilung. Die durchgezogene schwarze Funktion ist die wahre Verteilungsfunktion. Die gestrichelte schwarze Funktion ist die punktweise Erwartungswertfunktion der Verteilungsfunktion unter ihrer Posteriori-Verteilung. Die grauen Funktionen sind 100 Ziehungen aus der Posteriori-Verteilung der Verteilungsfunktion

3.2.2 Dirichlet Prozess

Der Dirichlet Prozess (DP; Ferguson, 1973) kann als Spezialfall der PT Verteilung mit

$$\lambda_{e_l(y)} = \lambda_{e_l(y)0} + \lambda_{e_l(y)1}$$

aufgefasst werden (Ferguson, 1974). Er erlaubt eine Charakterisierung durch zwei Parameter: Einen Parameter $\lambda \in \mathbb{R}_+$ und eine Basisverteilung P_0 (Ferguson, 1973; Müller et al., 2015).

Alternativ kann der DP auch wie folgt definiert werden: Sei P ein zufälliges Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathcal{B})$. P folgt einem Dirichlet Prozess mit Parameter λ und Basisverteilung P_0, kurz $P \sim DP(\lambda, P_0)$, falls für jede endliche Partition $A_1, A_2, \ldots, A_k$ von $\mathbb{R}$ mit $A_j \in \mathcal{B}$ für alle $i = 1, 2, \ldots, k$

$$(P(A_1), P(A_2), \ldots, P(A_k)) \sim Diri(\lambda P_0(A_1), \lambda P_0(A_2), \ldots, \lambda P_0(A_k)) \tag{3.21}$$

gilt (Ferguson, 1973). Die Eigenschaften der Dirichlet-Verteilung sorgen hierbei dafür, dass jede Realisation von P ein Wahrscheinlichkeitsmaß ist.

Auf der Basis dieser Definition kann der DP kann auch sehr einfach auf mehrdimensionale Räume $(\mathbb{R}^Q, \mathcal{B}^Q)$ und mehrdimensionale zufällige Wahrscheinlichkeitsmaße verallgemeinert werden (Ferguson, 1973). Im Rahmen dieser Arbeit ist jedoch der eindimensionale DP ausreichend.

Aus (3.21) und den Eigenschaften der Dirichletverteilung ergibt sich für alle Ereignisse $A \in \mathcal{B}$:

$$E(P(A)) = P_0(A)$$

und

$$Var(P(A)) = \frac{P_0(A)(1 - P_0(A))}{\lambda + 1}$$

(Müller et al., 2015). Die Basisverteilung P_0 bestimmt also wie beim PT die erwartete Wahrscheinlichkeit für alle $A \in \mathcal{B}$ unter einer Realisation des zufälligen Wahrscheinlichkeitsmaßes P. Je größer der Parameter λ ist, desto kleiner fällt die Varianz um diesen Erwartungswert aus. Somit kann λ als Präzisionsparameter interpretiert werden (Müller et al., 2015).

Der DP erlaubt eine sogenannte Stick-Breaking Repräsentation (Ishwaran & James, 2001). Das heißt, dass unter dem DP für jedes $A \in \mathcal{B}$ fast sicher

$$P(A) = \sum_{l=1}^{\infty} \pi_l \, Dirac_{\delta_l}(A) \tag{3.22}$$

gilt, wobei $Dirac_{\delta_l}$ das Dirac-Maß

$$Dirac_{\delta_l}(A) = \begin{cases} 1 & \delta_l \in A \\ 0 & \delta_l \notin A \end{cases}$$

im Punkt δ_l bezeichnet,

$$\delta_l \overset{iid}{\sim} P_0 \tag{3.23}$$

und für die Gewichte

$$\pi_l = B_l \prod_{i=1}^{l-1}(1 - B_i)$$

mit

$$B_l \overset{iid}{\sim} Beta(1, \lambda) \tag{3.24}$$

und

$$\sum_{l=1}^{\infty} \pi_l = 1 \tag{3.25}$$

gilt. Dies ist eine für die Inferenz sehr günstige Eigenschaft, da sie die Simulation von Realisationen aus dem DP erlaubt (Ishwaran & James, 2001). Die Existenz einer solchen Darstellung impliziert allerdings auch, dass eine Realisation eines DP fast sicher eine diskrete Wahrscheinlichkeitsverteilung ist, da (3.22) bedeutet, dass jedem Punkt δ_l eine positive Wahrscheinlichkeit π_t zugewiesen wird und die abzählbar unendliche Menge der Punkte δ_l wegen (3.25) eine Wahrscheinlichkeit von 1 hat (Ishwaran & James, 2001).

In der Praxis wird der DP meist in einer trunkierten Version verwendet (Ishwaran & James, 2001; Ishwaran & Zarepour, 2000). Hierbei geht man von der Stick-Breaking Darstellung aus und setzt für ein festes $L \in \mathbb{N}$

$$P(A) = \sum_{l=1}^{L} \pi_l \, Dirac_{\delta_l}(A) \tag{3.26}$$

statt (3.22) und

$$\pi_l = B_l \prod_{i=1}^{l-1}(1 - B_i)$$

für $l = 1, 2, \ldots, L-1$, sowie

$$\pi_L = 1 - \sum_{l=1}^{L-1} \pi_l$$

statt (3.25). Die hieraus resultierende Verteilung des zufälligen Wahrscheinlichkeitsmaßes P wird trunkierter Dirichlet Prozess (TDP) genannt und sei mit $TDP(\lambda, P_0)$ bezeichnet.

Der DP ist unter iid Beobachtungen konjugiert (Ferguson, 1973): Seien $X_1, X_2, \ldots, X_n$ Zufallsvariablen mit $X_i|P \overset{iid}{\sim} P$, $i = 1, 2, \ldots, n$, und $P \sim DP(\lambda, P_0)$. Dann ist

$$P|x_1, x_2, \ldots, x_n \sim DP(\lambda^*, P_0^*) \tag{3.27}$$

mit

$$\lambda^* = \lambda + n \tag{3.28}$$

und

$$P_0^* = \frac{\lambda}{\lambda + n} P_0 + \frac{n}{\lambda + n} \left(\frac{1}{n} \sum_{i=1}^{n} Dirac_{x_i} \right) \tag{3.29}$$

Die Posteriori-Verteilung von P ist also wieder ein DP, dessen Präzision um den Wert n zugenommen hat und dessen Basisverteilung als gewichtete Summe der Priori-Basisverteilung P_0 und der empirischen Verteilung der Beobachtungen $x_1, x_2, \ldots, x_n$ aufgefasst werden kann.

3.2.3 Dirichlet Prozess Mixtures

Die Dirichlet Prozess Mixture (DPM) Verteilung ist eine hierarchische Erweiterung des DP (Escobar & West, 1995): Sei D ein zufälliges Wahrscheinlichkeitsmaß und seien X_1, X_2, …, X_n Zufallsvariablen mit $X_i|D \overset{iid}{\sim} D$, $i = 1, 2, \ldots, n$. Sei zudem $G(\boldsymbol{\theta})$ eine durch einen Parametervektor $\boldsymbol{\theta}$ parametrisierte Verteilung, P_0 eine Verteilung auf dem Parameterraum von $\boldsymbol{\theta}$ und P ein zufälliges Wahrscheinlichkeitsmaß auf dem Parameterraum von $\boldsymbol{\theta}$. D folgt einer DPM Verteilung mit Parametern $G(\boldsymbol{\theta})$, P_0 und λ, kurz $D \sim TDPM(G(\boldsymbol{\theta}), P_0, \lambda)$ falls

$$\begin{aligned} X_i|\boldsymbol{\theta}_i &\overset{ind}{\sim} G(\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i|P &\overset{iid}{\sim} P \\ P &\sim DP(\lambda, P_0) \end{aligned} \tag{3.30}$$

(Ishwaran & Zarepour, 2000). Da die Realisationen des DP in (3.30) fast sicher diskrete Wahrscheinlichkeitsmaße sind und somit die $\boldsymbol{\theta}_i$ eine diskrete Verteilung aufweisen, folgt aus dieser hierarchischen Definition der DPM, dass die Realisationen von D diskrete Mischungen der parametrischen Verteilung $G(\boldsymbol{\theta})$ sind: Sei $F_G(x; \boldsymbol{\theta})$ die Verteilungsfunktion von $G(\boldsymbol{\theta})$. Dann gilt für Verteilungsfunktion F einer Realisation der DPM

$$F(x) = \sum_{l=1}^{\infty} \pi_l F_G(x; \boldsymbol{\theta}_l), \tag{3.31}$$

wobei $\boldsymbol{\theta}_l \overset{iid}{\sim} P_0$ und π_l die Gewichte des DP aus (3.30) in seiner Stick-Breaking Darstellung sind (Müller et al., 2015). Für stetige Verteilungen $G(\boldsymbol{\theta})$ sind die Realisationen der DPM also diskrete Mischungen von stetigen Verteilungen und somit selbst stetige Verteilungen.

Falls $G(\boldsymbol{\theta})$ zudem eine Dichte $f_G(x; \boldsymbol{\theta})$ besitzt, hat die Dichte f einer Realisation der DPM die Form

$$f(x) = \sum_{l=1}^{\infty} \pi_l f_G(x; \boldsymbol{\theta}_l) \tag{3.32}$$

(Müller et al., 2015).

Ersetzt man den DP in (3.30) durch einen TDP, wird die hieraus resultierende Verteilung trunkierte Dirichlet Prozess Mixture (TDPM) Verteilung genannt (Ishwaran & Zarepour, 2000). Das Trunkierungslevel L bestimmt dann die Anzahl der Mischkomponenten in (3.31) und (3.32).

Zwei wichtige Spezialfälle der (T)DPM sind Dirichlet Prozess Mean Mixtures und Dirichlet Mean Variance Mixtures von Normalverteilungen: Ein zufälliges Wahrscheinlichkeitsmaß D folgt einer Dirichlet Prozess Mean Mixture von Normalverteilungen (Ishwaran & Zarepour, 2000), kurz $D \sim (T)DPM(N(\delta_i, \sigma^2), P_0, \lambda)$, falls $X_i|D \overset{iid}{\sim} D$, $i = 1, 2, \ldots, n$ und

$$\begin{aligned} X_i|\delta_i, \sigma^2 &\overset{ind}{\sim} N(\delta_i, \sigma^2) \\ \delta_i|P &\overset{iid}{\sim} P \\ P &\sim (T)DP(\lambda, P_0). \end{aligned} \tag{3.33}$$

Ein zufälliges Wahrscheinlichkeitsmaß D folgt einer Dirichlet Prozess Mean Variance Mixture von Normalverteilungen (Ishwaran & James, 2002), falls $X_i|D \overset{iid}{\sim} D$, $i = 1, 2, \ldots, n$ und

$$\begin{aligned} X_i|\delta_i, \sigma_i^2 &\overset{ind}{\sim} N(\delta_i, \sigma_i^2) \\ (\delta_i, \sigma_i^2)|P &\overset{iid}{\sim} P \\ P &\sim (T)DP(\lambda, P_0). \end{aligned} \tag{3.34}$$

Diese beiden Spezialfälle haben sich für die nonparametrische Bayes-Inferenz als sehr vorteilhaft erwiesen, da sie die Flexibilität des (T)DP mit den günstigen mathematischen Eigenschaften der Normalverteilungen kombinieren.

Kapitel 4

Markov Chain Monte Carlo Methoden

In komplexen bayesianischen Modellen können der EAP $E(\boldsymbol{\theta}|\mathbf{x})$ und andere Größen, die auf der Posteriori-Verteilung von $\boldsymbol{\theta}$ basieren, nicht analytisch berechnet werden. Ein häufiger Grund hierfür liegt darin, dass die Normierungskonstante $f(\mathbf{x})$ der Posteriori-Dichte nicht berechnet werden kann.

In diesen Fällen kann auf Monte Carlo Simulationsmethoden zurückgegriffen werden (Robert & Casella, 2004). Das Grundprinzip dieser Methoden besteht in der Erzeugung einer Folge von computergenerierten (gegebenenfalls vektorwertigen) Zufallszahlen $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\ldots$, $\boldsymbol{\theta}_T$, für die ein Gesetz der großen Zahlen

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\theta}_t = E(\boldsymbol{\theta}|\mathbf{x}) \tag{4.1}$$

gilt (Robert & Casella, 2004). Dieses rechtfertigt dann die Approximation des EAP durch den Mittelwert der so erzeugten Folge von Zufallszahlen.

Klassische Monte Carlo Algorithmen generieren unabhängige Zufallszahlen aus der Posteriori-Verteilung. Dies ist für nahezu alle Posteriori-Dichten möglich, die bis auf eine Normierungskonstante bekannt sind. Beispiele für Simulationsmethoden dieser Art sind Importance Sampling oder der Accept-Reject Algorithmus (Robert & Casella, 2004). Für unabhängige Zufallsvariablen mit identischer Verteilung gilt das starke Gesetz der großen Zahlen und der Erwartungswert der Parameter unter der Posteriori-Verteilung kann durch den Mittelwert der Folge von Zufallszahlen approximiert werden. Leider sind diese Methoden in hochdimensionalen Parameterräumen sehr ineffizient (Robert & Casella, 2004).

Als effizienter haben sich in hochdimensionalen Problemen sogenannte Markov Chain Monte Carlo (MCMC) Algorithmen erwiesen. Diese erzeugen im Gegensatz zu klassischen Monte Carlo Algorithmen keine unabhängigen Zufallszahlen, sondern eine Markov Kette, deren stationäre Verteilung die Posteriori-Verteilung ist (Robert & Casella, 2004).

Eine Folge $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T$ von Zufallsvariablen heißt Markovkette, falls für jedes $t = 2, 3, \ldots, T$

$$f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-2}, \ldots, \boldsymbol{\theta}_1) = f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$$

gilt (Robert & Casella, 2004). Die Elemente einer Markovkette weisen somit eine bestimmte Form der Abhängigkeit auf: Die auf den aktuellen Zustand $\boldsymbol{\theta}_t$ der Kette zum Zeitpunkt t und auf die vorangegangenen Zustände $\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-2}, \ldots, \boldsymbol{\theta}_1$ bedingte Verteilung des Zustands $\boldsymbol{\theta}_{t+1}$ der Kette zum Zeitpunkt $t+1$ hängt nur von dem aktuellen Zustand $\boldsymbol{\theta}_t$ ab. Die marginale Verteilung des ersten Elements $\boldsymbol{\theta}_1$ der Markovkette wird Startverteilung genannt. Zusammen mit den bedingten Dichten $f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ legt sie die gemeinsame Verteilung aller Elemente $\boldsymbol{\theta}_t$ der Markovkette fest (Robert & Casella, 2004).

Eine durch die bedingten Dichten $f(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$ definierte Markovkette besitzt eine stationäre Verteilung P, falls aus $\boldsymbol{\theta}_1 \sim P$ folgt, dass $\boldsymbol{\theta}_t \sim P$ für alle $t = 2, 3, \ldots, T$ gilt (Robert & Casella, 2004). Wenn das erste Element der Markovkette aus der stationären Verteilung gezogen wird, entsprechen die marginalen Verteilungen aller weiteren Elemente der Markovkette ebenfalls der stationären Verteilung. Die Realisation der kompletten Markovkette ist in diesem Fall eine Folge von (abhängigen) Ziehungen aus der stationären Verteilung P. Nicht alle Markovketten besitzen eine stationäre Verteilung (Robert & Casella, 2004).

Falls für eine Markovkette $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T$ eine stationäre Verteilung P existiert und eindeutig ist, kann unter bestimmten zusätzlichen Bedingungen für beliebige Startverteilungen ein starkes Gesetz der großen Zahlen hergeleitet werden:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{n} \boldsymbol{\theta}_t = E_P(\boldsymbol{\theta}) \tag{4.2}$$

Voraussetzungen für die Existenz und Eindeutigkeit einer stationären Verteilung und zusätzliche Bedingungen für die Geltung von (4.2) finden sich in Robert und Casella (2004).

Für die Bayes-Inferenz ist dieses Ergebnis sehr bedeutend: Falls durch einen Algorithmus eine Markovkette mit der Posteriori-Verteilung als stationärer Verteilung simuliert werden kann, für die (4.2) gilt, kann der EAP durch den Mittelwert der Realisation dieser Markovkette approximiert werden (Robert & Casella, 2004). Je geringer die Autokorrelation der Kette ist, desto genauer ist die Approximation. Einige für die vorliegende Arbeit wichtige Algorithmen zur Konstruktion einer solchen Markovkette werden in den folgenden Unterkapiteln vorgestellt.

4.1 Metropolis-Hastings Algorithmus

Voraussetzung für die Anwendung des Metropolis-Hastings (MH) Algorithmus ist zunächst die Wahl einer bedingten Vorschlagsverteilung mit Dichte $g(\boldsymbol{\theta}_t^*|\boldsymbol{\theta}_{t-1})$, aus der Zufallszahlen generiert werden können. Zudem muss die Dichte $f(\boldsymbol{\theta}|\mathbf{x})$ der interessierenden Posteriori-Verteilung bis auf eine multiplikative Normierungskonstante berechenbar sein und ihr Träger muss eine Teilmenge des Trägers aller bedingten Vorschlagsverteilungen sein (Robert & Casella, 2004).

Der MH Algorithmus besteht dann aus den folgenden Schritten (Robert & Casella, 2004):

- Ziehe einen Startwert $\boldsymbol{\theta}_0$ aus einer Startverteilung. Dieser Startwert kann auch deterministisch festgesetzt werden.

- Iterativ für jedes $t = 1, 2, \ldots, T$:
  - Ziehe $\boldsymbol{\theta}_t^*$ aus der bedingten Vorschlagsverteilung mit Dichte $g(\boldsymbol{\theta}_t^*|\boldsymbol{\theta}_{t-1})$.
  - Setze $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^*$ mit Wahrscheinlichkeit

$$\rho = min\left\{\frac{f(\boldsymbol{\theta}_t^*|\mathbf{x})}{f(\boldsymbol{\theta}_{t-1}|\mathbf{x})}\frac{g(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t^*)}{g(\boldsymbol{\theta}_t^*|\boldsymbol{\theta}_{t-1})}, 1\right\} \tag{4.3}$$

    und $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$ sonst.

In jeder Iteration des Algorithmus wird somit durch Ziehen aus der bedingten Vorschlagsverteilung ein neuer Zustand $\boldsymbol{\theta}_t^*$ vorgeschlagen. Dieser vorgeschlagene Wert wird mit der in (4.3) berechneten Wahrscheinlichkeit als neuer Zustand der Kette akzeptiert. Falls der vorgeschlagene neue Zustand abgelehnt wird, ist $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$, und die Kette verweilt im Zustand $\boldsymbol{\theta}_{t-1}$.

Im Spezialfall einer symmetrischen Vorschlagsverteilung mit $g(\boldsymbol{\theta}_t^*|\boldsymbol{\theta}_{t-1}) = g(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t^*)$ für alle $\boldsymbol{\theta}_t^*$ und $\boldsymbol{\theta}_{t-1}$ vereinfacht sich die Akzeptanzwahrscheinlichkeit in (4.3) zu

$$\rho = min\left\{\frac{f(\boldsymbol{\theta}_t^*|\mathbf{x})}{f(\boldsymbol{\theta}_{t-1}|\mathbf{x})}, 1\right\} \tag{4.4}$$

(Robert & Casella, 2004). In diesem Fall wird der Algorithmus auch Metropolis Algorithmus genannt. Der Akzeptanzschritt hat hier eine einfache intuitive Interpretation: Ist die Posteriori-Dichte des vorgeschlagenen Werts höher als die des aktuellen Werts, wird der vorgeschlagene Wert mit Wahrscheinlichkeit 1 als neuer Wert der Kette akzeptiert. Ist dies nicht der Fall, entspricht die Akzeptanzwahrscheinlichkeit dem Faktor, um den die Posteriori-Dichte des vorgeschlagenen Werts niedriger als die des aktuellen Werts ist. Die Kette wandert somit bevorzugt in Regionen des Parameterraums, die eine hohe Posteriori-Dichte aufweisen.

Ein Beweis dafür, dass eine durch den MH Algorithmus erzeugte Folge von Zufallszahlen eine Markovkette mit der Posteriori-Verteilung als stationärer Verteilung ist, für die (4.2) gilt, findet sich in Robert und Casella (2004).

Die Posteriori-Dichte geht nur über den Term

$$\frac{f(\boldsymbol{\theta}_t^*|\mathbf{x})}{f(\boldsymbol{\theta}_{t-1}|\mathbf{x})}$$

in der Akzeptanzwahrscheinlichkeit ρ in (4.3) ein. Ist die Posteriori-Dichte nur bis auf eine multiplikative Konstante $c \in \mathbb{R}$ bekannt und somit statt der kompletten Posteriori-Dichte $f(\boldsymbol{\theta}|\mathbf{x})$ nur $f^*(\boldsymbol{\theta}|\mathbf{x})$ mit $f(\boldsymbol{\theta}|\mathbf{x}) = c\, f^*(\boldsymbol{\theta}|\mathbf{x})$ berechenbar, kann der MH Algorithmus wegen

$$\frac{f(\boldsymbol{\theta}_t^*|\mathbf{x})}{f(\boldsymbol{\theta}_{t-1}|\mathbf{x})} = \frac{c\, f^*(\boldsymbol{\theta}_t^*|\mathbf{x})}{c\, f^*(\boldsymbol{\theta}_{t-1}|\mathbf{x})} = \frac{f^*(\boldsymbol{\theta}_t^*|\mathbf{x})}{f^*(\boldsymbol{\theta}_{t-1}|\mathbf{x})} \tag{4.5}$$

trotzdem angewandt werden (Robert & Casella, 2004).

Dies ist von entscheidender Bedeutung für die Bayes-Inferenz, wie das folgende Beispiel veranschaulicht: Seien $X_1, X_2, \ldots, X_n$ unabhängige Ziehungen aus einer Normalverteilung mit unbekanntem Erwartungswert μ und unbekannter Varianz σ^2. Als priori-Verteilung für μ wird eine Normalverteilung mit Parametern μ_{priori} und σ^2_{priori} und als priori-Verteilung für σ^2 unabhängig von μ eine inverse Gammaverteilung mit Parametern α_{priori} und β_{priori} angenommen. Nach dem Satz von Bayes gilt:

$$\begin{aligned} f(\mu, \sigma^2|\mathbf{x}) &= \frac{f(\mathbf{x}|\mu, \sigma^2) f(\mu) f(\sigma^2)}{f(\mathbf{x})} \\ &= \frac{\prod_{i=1}^{n} N(x_i; \mu, \sigma^2) N(\mu; \mu_{priori}, \sigma^2_{priori}) IG(\sigma^2; \alpha_{priori}, \beta_{priori})}{f(\mathbf{x})} \end{aligned}$$

Der Zähler dieser Posteriori-Dichte kann zwar für jedes μ und jedes σ^2 berechnet werden, hat aber keine bekannte Form. Die Normierungskonstante $f(\mathbf{x})$ im Nenner ist nicht analytisch berechenbar. Da $f(\mathbf{x})$ jedoch wegen (4.5) vernachlässigt werden kann, bedarf es auch keiner Berechnung:

$$\begin{aligned} \frac{f(\mu_t^*, \sigma^{2*}|\mathbf{x})}{f(\mu_{t-1}, \sigma^2_{t-1}|\mathbf{x})} &= \frac{f(\mathbf{x}|\mu_t^*, \sigma_t^{2*}) f(\mu_t^*) f(\sigma_t^{2*}) / f(\mathbf{x})}{f(\mathbf{x}|\mu_{t-1}, \sigma^2_{t-1}) f(\mu_{t-1}) f(\sigma^2_{t-1}) / f(\mathbf{x})} \\ &= \frac{f(\mathbf{x}|\mu_t^*, \sigma_t^{2*}) f(\mu_t^*) f(\sigma_t^{2*})}{f(\mathbf{x}|\mu_{t-1}, \sigma^2_{t-1}) f(\mu_{t-1}) f(\sigma^2_{t-1})} \end{aligned}$$

Als bedingte Vorschlagsverteilung für (μ^*, σ^{2*}) kann in jeder Iteration t eine multivariate Normalverteilung mit dem aktuellen Zustand als Erwartungswert und Einheitskovarianzmatrix gewählt werden:

$$g(\mu^*, \sigma^{2*}|\mu_{t-1}, \sigma^2_{t-1}) = MVN(\mu^*, \sigma^{2*}; \mu_{t-1}, \sigma^2_{t-1}, \mathbf{I})$$

Aufgrund der Symmetrieeigenschaft der multivariaten Normalverteilung gilt dann

$$g(\mu^*, \sigma^{2*}|\mu_{t-1}, \sigma^2_{t-1}) = g(\mu_{t-1}, \sigma^2_{t-1}|\mu^*, \sigma^{2*})$$

und es ergibt sich

$$\rho = min \left\{ \frac{f(\mathbf{x}|\mu_t^*, \sigma_t^{2*}) f(\mu_t^*) f(\sigma_t^{2*})}{f(\mathbf{x}|\mu_{t-1}, \sigma^2_{t-1}) f(\mu_{t-1}) f(\sigma^2_{t-1})}, 1 \right\}$$

als Akzeptanzwahrscheinlichkeit in jedem Schritt.

Ein großer Nachteil des MH Algorithmus ist, dass Informationen über die Form der Posteriori-Verteilung nur in die Berechnung der Akzeptanzwahrscheinlichkeit und nicht in die Wahl der Vorschlagsverteilung eingehen. Die Effizienz des Algorithmus hängt jedoch stark von den Eigenschaften der Letzteren ab (Robert & Casella, 2004): Besitzt die Vorschlagsverteilung zu große Varianzen, werden vor allem in hochdimensionalen Parameterräumen durchgehend Werte mit sehr geringen Posteriori-Dichten vorgeschlagen, was zu sehr geringen Akzeptanzwahrscheinlichkeiten führt. Im Extremfall bleibt die Kette für

den gesamten Lauf des Algorithmus im gleichen Zustand hängen. Zu kleine Varianzen der Vorschlagsverteilung führen wiederum dazu, dass die Kette in nur sehr kleinen Schritten durch den Parameterraum wandert und langsam konvergiert. Aus diesem Grund sollten, wenn möglich, MCMC Algorithmen bevorzugt werden, die weitere Informationen über die Struktur der Posteriori-Verteilung miteinbeziehen.

Für eine weiterführende Diskussion des MH Algorithmus sei auf Robert und Casella (2004) verwiesen.

4.2 Gibbs Sampler

Sei der Parametervektor $\boldsymbol{\theta}$ in einzelne Komponenten $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B$ aufgeteilt, so dass $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B)^\top$ ist. Hierbei können die einzelnen Elemente $\boldsymbol{\theta}_b$ ihrerseits wieder vektorwertig sein. Für eine gegebene Posteriori-Dichte $f(\boldsymbol{\theta}|\mathbf{x})$ heißen die Funktionen

$$\begin{array}{c} f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_B, \mathbf{x}) \\ f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_B, \mathbf{x}) \\ \vdots \\ f(\boldsymbol{\theta}_B|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{B-1}, \mathbf{x}) \end{array} \tag{4.6}$$

vollständig bedingte Dichten (Robert & Casella, 2004). Es handelt sich bei ihnen um die Dichten der einzelnen Komponenten $\boldsymbol{\theta}_b$ gegeben die Daten und die anderen Komponenten. Um die Notation übersichtlicher zu gestalten, schreibt man oft auch $f(\boldsymbol{\theta}_b|\cdot)$ für die vollstandig bedingte Dichte einer Komponente $\boldsymbol{\theta}_b$.

Die Anwendung des Gibbs Samplers setzt voraus, dass eine Aufteilung von $\boldsymbol{\theta}$ gefunden werden kann, so dass alle vollständig bedingten Dichten eine bekannte Form aufweisen und aus den durch sie definierten Verteilungen Zufallszahlen gezogen werden können. Dies ist nicht für alle Posteriori-Verteilungen möglich und muss für jedes Modell einzeln nachgewiesen werden (Robert & Casella, 2004).

Für die vollständig bedingten Dichten gilt mit dem Satz von Bayes

$$\begin{aligned} & f(\boldsymbol{\theta}_b|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{b-1}, \boldsymbol{\theta}_{b+1}, \ldots, \boldsymbol{\theta}_B, \mathbf{x}) \\ = \; & \frac{f(\boldsymbol{\theta}_b, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{b-1}, \boldsymbol{\theta}_{b+1}, \ldots, \boldsymbol{\theta}_B, \mathbf{x})}{f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{b-1}, \boldsymbol{\theta}_{b+1}, \ldots, \boldsymbol{\theta}_B, \mathbf{x})} \\ \propto \; & f(\mathbf{x}|\boldsymbol{\theta}_b, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{b-1}, \boldsymbol{\theta}_{b+1}, \ldots, \boldsymbol{\theta}_B) f(\boldsymbol{\theta}_b, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{b-1}, \boldsymbol{\theta}_{b+1}, \ldots, \boldsymbol{\theta}_B) \\ = \; & f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \\ = \; & f(\mathbf{x}, \boldsymbol{\theta}) \end{aligned}$$

(Robert & Casella, 2004). Sie sind also für jedes $\boldsymbol{\theta}_b$ proportional zum Zähler der Posteriori-Dichte des kompletten Modells, wobei zusätzlich alle multiplikativen Terme vernachlässigt werden können, die nicht von $\boldsymbol{\theta}_b$ abhängen. Dieses Ergebnis ist bei der Herleitung der

vollständig bedingten Dichten hilfreich. Zudem liefert es Anhaltspunkte dafür, wie die Priori-Verteilungen der Parameter gewählt werden können, damit die vollständig bedingten Dichten eine bekannte Form aufweisen.

Die Schritte des Gibbs Samplers sehen wie folgt aus (Robert & Casella, 2004):

- Ziehe einen Startwert $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{20}, \ldots, \boldsymbol{\theta}_{B0})^\top$ aus einer Startverteilung. Dieser Startwert kann auch deterministisch festgesetzt werden.
- Iterativ für jedes $t = 1, 2, \ldots, T$:
  - Ziehe $\boldsymbol{\theta}_{1t}$ aus der Verteilung mit Dichte $f(\boldsymbol{\theta}_{1t}|\boldsymbol{\theta}_{2(t-1)}, \boldsymbol{\theta}_{3(t-1)}, \ldots, \boldsymbol{\theta}_{B(t-1)}, \mathbf{x})$.
  - Ziehe $\boldsymbol{\theta}_{2t}$ aus der Verteilung mit Dichte $f(\boldsymbol{\theta}_{2t}|\boldsymbol{\theta}_{1t}, \boldsymbol{\theta}_{3(t-1)}, \ldots, \boldsymbol{\theta}_{B(t-1)}, \mathbf{x})$.
  - ...
  - Ziehe $\boldsymbol{\theta}_{Bt}$ aus der Verteilung mit Dichte $f(\boldsymbol{\theta}_{Bt}|\boldsymbol{\theta}_{1t}, \boldsymbol{\theta}_{2t}, \ldots, \boldsymbol{\theta}_{(B-1)t}, \mathbf{x})$.

In jeder Iteration wird also jedes Element des Parametervektors aus seiner vollständig bedingten Verteilung gegeben die aktuellen anderen Parameter und die Daten gezogen.

Der Gibbs Sampler kann als spezieller MH Algorithmus aufgefasst werden und erzeugt daher eine Markovkette $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T$ mit der Posteriori-Verteilung als stationärer Verteilung, für die ein starkes Gesetz der großen Zahlen (4.2) gilt (Robert & Casella, 2004). Die implizite Vorschlagsverteilung ergibt sich hierbei aus den vollständig bedingten Dichten und führt dazu, dass jeder Vorschlag angenommen wird (Robert & Casella, 2004).

Im Fall des Beispiels aus Abschnitt 4.1 ist $\boldsymbol{\theta} = (\mu, \sigma^2)$. Die einzig mögliche Aufteilung des Parametervektors ist somit $\boldsymbol{\theta}_1 = \mu$ und $\boldsymbol{\theta}_2 = \sigma^2$. Für die vollständig bedingte Dichte von μ gilt:

$$\begin{aligned} f(\mu|\sigma^2, \mathbf{x}) &\propto f(\mathbf{x}|\mu, \sigma^2) f(\mu) f(\sigma^2) \\ &\propto f(\mathbf{x}|\mu, \sigma^2) f(\mu) \\ &= \prod_{i=1}^{n} N(x_i; \mu, \sigma^2) N(\mu; \mu_{priori}, \sigma^2_{priori}) \end{aligned}$$

Dieser letzte Ausdruck entspricht bei gegebenem σ^2 dem Zähler der Posteriori-Dichte von μ im Fall von n unabhängigen Ziehungen X_i aus einer Normalverteilung mit unbekanntem Erwartungswert μ und bekannter Varianz σ^2 bei normalverteilter Priori-Verteilung für μ mit Hyperparametern μ_{priori} und σ^2_{priori}. Für dieses Modell ist die Posteriori-Verteilung wieder eine Normalverteilung mit Erwartungswert

$$\mu_{post} = \frac{\frac{1}{\sigma^2} \sum_{i=1}^{n} y_i + \frac{1}{\sigma^2_{priori}} \mu_{priori}}{\frac{n}{\sigma^2} + \frac{1}{\sigma^2_{priori}}}$$

und Varianz

$$\sigma^2_{post} = \left(\frac{1}{\sigma^2_{priori}} + \frac{n}{\sigma^2} \right)^{-1}$$

(Gelman et al., 2014). Damit ist auch die vollständig bedingte Verteilung von μ eine Normalverteilung mit Erwartungswert μ_{post} und Varianz σ^2_{post}. Aus dieser können leicht Zufallszahlen gezogen werden. Die vollständig bedingte Dichte von σ^2 ist

$$\begin{aligned} f(\sigma^2|\mu, \mathbf{x}) &\propto f(\mathbf{x}|\mu, \sigma^2) f(\mu) f(\sigma^2) \\ &\propto f(\mathbf{x}|\mu, \sigma^2) f(\sigma^2) \\ &= \prod_{i=1}^{n} N(x_i; \mu, \sigma^2) IG(\sigma^2; \alpha_{priori}, \beta_{priori}). \end{aligned}$$

Auch diese hat eine bekannte Form: Sie entspricht der Posteriori-Dichte von σ^2 im Fall von n unabhängigen Ziehungen X_i aus einer Normalverteilung mit bekanntem Erwartungswert μ und unbekannter Varianz σ^2 bei inverser Gammaverteilung als Priori-Verteilung für σ^2 mit Hyperparametern α_{priori} und β_{priori}. Die Posteriori-Verteilung ist in diesem Fall wieder eine inverse Gammaverteilung mit Parametern

$$\alpha_{post} = \frac{n}{2} + \alpha_{priori}$$

und

$$\beta_{post} = \beta_{priori} + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2$$

(Gelman et al., 2014). Da auch aus dieser Verteilung leicht Zufallszahlen gezogen werden können, sind beide vollständig bedingten Verteilungen bekannt und der Implementation eines Gibbs Samplers steht nichts mehr im Wege.

Im Gegensatz zum allgemeinen MH Algorithmus berücksichtigt der Gibbs Sampler die Struktur der Posteriori-Verteilung in Form der vollständig bedingten Dichten. Dies führt dazu, dass neue Zustände immer akzeptiert werden und die Kette nicht steckenbleiben kann. Eine zu langsame Exploration des Parameterraums ist aber immer noch möglich und tritt vor allem bei stark korrelierten Parametern in hochdimensionalen Parameterräumen auf (Robert & Casella, 2004).

Auch wenn einige der vollständig bedingten Dichten nur bis auf eine multiplikative Konstante bekannt sind und nicht direkt aus ihnen gezogen werden kann, ist die Konstruktion eines Gibbs Samplers möglich (Robert & Casella, 2004): Für die Teile des Parametervektors mit unbekannten vollständig bedingten Verteilungen wird dann in jeder Iteration des Gibbs Samplers ein einzelner Metropolisschritt mit einer beliebigen Vorschlagsverteilung durchgeführt. Die Akzeptanzwahrscheinlichkeit wird hierbei auf der Basis der entsprechenden bis auf eine multiplikative Konstante bekannten vollständig bedingten Dichte berechnet. Aus den bekannten vollständig bedingten Verteilungen der übrigen Parameter kann normal gezogen werden. Gibbs Sampler dieser Art werden auch Metropolis-within-Gibbs Sampler genannt (Robert & Casella, 2004). Ihre Effizienz hängt stark von der Effizienz des Metropolisschritts ab (Robert & Casella, 2004).

4.3 Hamiltonian Monte Carlo

Der Hamiltonian Monte Carlo (HMC) Algorithmus ist eine Variante des MH Algorithmus mit einer Vorschlagsverteilung, die die geometrische Struktur der Posteriori-Dichte berücksichtigt (Neal, 2011). Dadurch ermöglicht er große Sprünge im Parameterraum, die trotzdem mit einer hohen Wahrscheinlichkeit akzeptiert werden, was in vielen Fällen zu Markovketten mit niedriger Autokorrelation führt. Die Anwendung des HMC Algorithmus setzt voraus, dass die ersten partiellen Ableitungen der logarithmierten Posteriori-Dichte bezüglich aller Parameter in $\boldsymbol{\theta}$ analytisch bestimmbar sind. Zudem müssen alle Parameter in $\boldsymbol{\theta}$ reparametrisiert werden können, so dass sie in den kompletten reellen Zahlen liegen (Neal, 2011).

Bezeichne $\nabla_{\boldsymbol{\theta}} \ln f(\boldsymbol{\theta}_t|\mathbf{x})$ den Gradienten der logarithmierten Posteriori-Dichte im aktuellen Zustand $\boldsymbol{\theta}_t$ der Markovkette, also den Vektor der ersten partiellen Ableitungen der logarithmierten Posteriori-Dichte, ausgewertet im Punkt $\boldsymbol{\theta}_t$. Sei zudem ein reeller Vektor $\mathbf{m}$ mit sogenannten Momentumvariablen definiert, der die gleiche Dimension wie der Parametervektor besitzt.

Ein Leapfrog-Schritt ausgehend von den Punkten $\boldsymbol{\theta}_t$ und $\mathbf{m}_t$ mit Schrittweite $\epsilon \in \mathbb{R}_+$ ist definiert durch (Hoffman & Gelman, 2014):

$$\begin{aligned}
\mathbf{m}_{t+\frac{\epsilon}{2}} &= \mathbf{m}_t + (\frac{\epsilon}{2})\nabla_{\boldsymbol{\theta}} \ln f(\boldsymbol{\theta}_t|\mathbf{x}) \\
\boldsymbol{\theta}_{t+\epsilon} &= \boldsymbol{\theta}_t + \epsilon \mathbf{m}_{t+\frac{\epsilon}{2}} \\
\mathbf{m}_{t+\epsilon} &= \mathbf{m}_{t+\frac{\epsilon}{2}} + (\frac{\epsilon}{2})\nabla_{\boldsymbol{\theta}} \ln f(\boldsymbol{\theta}_{t+\epsilon}|\mathbf{x})
\end{aligned}$$

Der HMC Algorithmus hat dann die folgende Form (Hoffman & Gelman, 2014):

- Ziehe Startwerte $\boldsymbol{\theta}_0$ und $\mathbf{m}_0$ aus einer Startverteilung. Diese Startwerte können auch deterministisch festgesetzt werden.

- Iterativ für jedes $t = 1, 2, \ldots, T$:

  - Ziehe alle Momentumvariablen in $\mathbf{m}_t$ unabhängig aus Standardnormalverteilungen.
  - Führe L sequentielle Leapfrog-Schritte ausgehend von $\boldsymbol{\theta}_{t-1}$ und $\mathbf{m}_t$ durch und verwende die hieraus resultierenden Werte als Vorschlagswerte $\boldsymbol{\theta}_t^*$ und $\mathbf{m}_t^*$.
  - Setze $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^*$ mit Wahrscheinlickeit

$$\rho = min\left\{\frac{\exp(\ln f(\boldsymbol{\theta}_t^*|\mathbf{x}) - \frac{1}{2}\mathbf{m}_t^{*\top}\mathbf{m}_t^*)}{\exp(\ln f(\boldsymbol{\theta}_{t-1}|\mathbf{x}) - \frac{1}{2}\mathbf{m}_t^{\top}\mathbf{m}_t)}, 1\right\}$$

    und $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$ sonst.

Die L Leapfrog-Schritte in jeder Iteration stellen eine diskrete Approximation der Lösung der partiellen Differentialgleichungen

$$\begin{aligned} \frac{\partial \boldsymbol{\theta}(s)}{\partial s} &= \mathbf{m}(s) \\ \frac{\partial \mathbf{m}(s)}{\partial s} &= \nabla_{\boldsymbol{\theta}} \ln f(\boldsymbol{\theta}(s)|\mathbf{x}) \end{aligned} \tag{4.7}$$

dar, wobei $\boldsymbol{\theta}(s)$ und $\mathbf{m}(s)$ Funktionen sind, die jeweils die Position von $\boldsymbol{\theta}$ und $\mathbf{m}$ zum Zeitpunkt s angeben (Neal, 2011). Für zweidimensionale Parametervektoren haben diese Differentialgleichungen eine intuitive physikalische Interpretation (Neal, 2011): Sie beschreiben die (reibungslose) Bewegung eines Objekts auf der Oberfläche der negativen logarithmierten Posteriori-Dichte. Der HMC Algorithmus kann in diesem Fall folgendermaßen interpretiert werden: In jeder Iteration t wird ein sich an dem aktuellen Punkt $\boldsymbol{\theta}_t$ der Kette im Parameterraum befindliches Objekt durch Ziehen der Momentumvariablen in eine zufällige Richtung mit einer zufälligen Geschwindigkeit angestoßen. Daraufhin wird die Bewegung des Objekts in diese Richtung für einige Zeit mithilfe einer diskreten Approximation der Lösungen von (4.7) simuliert. Dabei wandert das Objekt bevorzugt bergab, also in Richtung von Regionen mit hoher Posteriori-Dichte. Da es jedoch gleichzeitig auch an Geschwindigkeit bzw. an Momentum zunimmt, bleibt es nicht in diesen Regionen hängen, sondern kann sich wieder aus ihnen herausbewegen. Der aus der simulierten Bewegung resultierende Punkt wird als neuer Parameter $\boldsymbol{\theta}^*_{t+1}$ vorgeschlagen und mit Wahrscheinlichkeit ρ akzeptiert.

Der HMC Algorithmus erzeugt eine Markovkette $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T$ mit der Posteriori-Verteilung als stationärer Verteilung, für die ein starkes Gesetz der großen Zahlen (4.2) gilt (Neal, 2011). Als Hyperparameter des Algorithmus müssen die Schrittweite ϵ und Anzahl L der Leapfrog-Schritte festgelegt werden. Aktuelle Erweiterungen des Algorithmus (Hoffman & Gelman, 2014) legen diese Parameter in sinnvoller Art und Weise adaptiv fest.

Kapitel 5

Nonparametrische Bayes-Inferenz im NMIRT Modell

Ziel dieses Kapitels ist die Herleitung von MCMC Schätzverfahren für das NMIRT Modell von Peress (2012). Dieses in den Kapiteln 1 und 2 näher diskutierte Modell hat die Form

$$P(X_{ij} = 1) = F(\beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i) \tag{5.1}$$

mit festen Q-dimensionalen latenten Variablen $\boldsymbol{\xi}_i$ für jede Person $i = 1, 2, \ldots, n$, skalaren Schwierigkeitsparametern β_j und Q-dimensionalen Ladungsparametern $\boldsymbol{\alpha}_j$ für jedes Item $j = 1, 2, \ldots, k$ und einer allen Items gemeinsamen nonparametrischen ICC F. Die Parameter β_j werden in einem k-dimensionalen Vektor $\boldsymbol{\beta}$ zusammengefasst, die Ladungen in einer $k \times Q$-Matrix $\boldsymbol{\alpha}$ und die Personenparameter in einer $n \times Q$-Matrix $\boldsymbol{\xi}$. Zum Zweck der einfacheren Lesbarkeit sei außerdem die Notation $\eta_{ij} = \beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i$ für den latenten Prädiktor eingeführt, wobei die Parameter η_{ij} zudem in einer $n \times k$-Matrix $\boldsymbol{\eta}$ zusammengefasst seien. Die Itemantworten X_{ij} werden als unabhängig angenommen.

Die Parameterschätzung im NMIRT Modell stellt ein bislang ungelöstes Problem dar. Peress (2012) selbst skizziert einen Likelihood-basierten Ansatz, der die nonparametrische ICC F mithilfe von linearen Splines modelliert. Allerdings werden hier die Restriktionen des ursprünglichen Modells insbesondere bezüglich der ICC F nicht eingehalten, so dass das der Schätzung zugrundeliegende Modell nicht identifiziert ist. Doch auch wenn der Ansatz mithilfe der in Kapitel 2 vorgeschlagenen Reparametrisierungen modifiziert werden könnte, blieben noch einige Probleme bestehen: Erstens liefert die von Peress (2012) beschriebene Methode lediglich Punktschätzer. Eine Quantifizierung der Schätzunsicherheit ist in ihrem Rahmen somit nicht möglich. Zweitens bleibt unklar, wie effizient der von Peress (2012) vorgeschlagene numerische Optimierungsalgorithmus zur Bestimmung des Maximums der Likelihoodfunktion in der Praxis wäre. In hochdimensionalen Parameterräumen bleiben Algorithmen dieser Art sehr häufig in lokalen Maxima hängen (Robert & Casella, 2004).

In anderen nonparametrischen IRT Modellen hat sich die Parameterschätzung mithilfe nonparametrischer bayesianischer Methoden als vielversprechend erwiesen. So wählen Duncan und MacEachern (2008) für die Schätzung des eindimensionalen nonparametrischen Modells (1.13) einen auf einem DP basierenden Ansatz. Miyazaki und Hoshino (2009)

spezifizieren für dasselbe Modell eine aus einer TDPM abgeleitete Priori-Verteilung für die ICC. Karabatsos (2016) stellt ein mehrdimensionales nonparametrisches Modell der Form (1.14) vor, dem eine Verallgemeinerung der DPM zugrunde liegt. Ein entscheidender Vorteil dieser bayesianischen Ansätze besteht darin, dass über die Varianz der marginalen Posteriori-Verteilungen eine Quantifizierung der Schätzunsicherheit bezüglich der Item- und Personenparameter und der ICC möglich ist. Dies ist im Rahmen frequentistischer Methoden deutlich aufwendiger und in vielen Fällen sogar unmöglich. Dem nonparametrischen bayesianischen Ansatz soll daher auch hier gefolgt werden.

Ausgangspunkt sind die in Kapitel 2 hergeleiteten Reparametrisierungen des NMIRT Modells. Für deren Parameter werden mithilfe der in Kapitel 3 vorgestellten zufälligen Wahrscheinlichkeitsmaße Priori-Verteilungen formuliert. Zur Schätzung wird dann auf die in Kapitel 4 beschriebenen MCMC Methoden zurückgegriffen.

5.1 Gibbs Sampler für ein NMIRT Modell mit Dirichlet Prozess Mixture

Der in diesem Abschnitt hergeleitete Gibbs Sampler basiert auf der in Modell 3 vorgeschlagenen Reparametriesierung des NMIRT Modells. In dieser Parametrisierung ist die ICC F ohne Quantilrestriktionen schätzbar. Dafür ist der Schwierigkeitsparameter des ersten Items β_1 auf 0 und die Ladung des ersten Items auf den ersten Faktor α_{11} auf 1 gesetzt. Zudem sind folgende Restriktionen der Personenparameter nötig: $\boldsymbol{\xi}_1 = (\xi_{11}, 0, \ldots, 0)^\top$, $\boldsymbol{\xi}_2 = (0, \xi_{22}, 0, \ldots, 0)^\top$, $\ldots$, $\boldsymbol{\xi}_Q = (0, \ldots, 0, \xi_{QQ})^\top$, $\boldsymbol{\xi}_{Q+1} = (0, \ldots, 0)^\top$ und $\boldsymbol{\xi}_{Q+2} = (1, \ldots, 1)^\top$. Da die Ladungsparameter in dieser Parametrisierung nicht positiv sein müssen, können alle Item- und Personenparameter Werte in den kompletten reellen Zahlen annehmen. Dies wird sich als Voraussetzung dafür erweisen, dass die vollständig bedingten Dichten aller Parameter eine bekannte Form besitzen.

5.1.1 Reformulierung des Modells

Da die ICC F in Modell 3 monoton und stetig mit linksseitigem Grenzwert 0 sowie rechtsseitigem Grenzwert 1 und somit eine stetige Verteilungsfunktion ist, kann das Modell mithilfe von in einer $n \times k$ Matrix $\mathbf{z}$ zusammengefassten stetigen latenten Hilfsvariablen Z_{ij} als latentes Schwellenmodell reformuliert werden (Albert, 1992):

Sei P_F das durch die Verteilungsfunktion F definierte Wahrscheinlichkeitsmaß und

$$Z_{ij} = -\eta_{ij} + \epsilon_{ij} \tag{5.2}$$

mit

$$\epsilon_{ij} \overset{i.i.d.}{\sim} P_F \tag{5.3}$$

und

$$X_{ij} = I(Z_{ij} \leq 0). \tag{5.4}$$

Diese Formulierung ist äquivalent zu (5.1), da

$$\begin{aligned} P(X_{ij} = 1) &= P(I(Z_{ij} \leq 0) = 1) \\ &= P(Z_{ij} \leq 0) \\ &= P(-\eta_{ij} + \epsilon_{ij} \leq 0) \\ &= P(\epsilon_{ij} \leq \eta_{ij}) \\ &= F(\eta_{ij}) \end{aligned}$$

und die Itemantworten X_{ij} wegen der Unabhängigkeit der ϵ_{ij} ebenfalls unabhängig sind.

Wird in (5.3) zum Beispiel eine Standardnormalverteilung für P_F festgesetzt, ergibt sich das mehrdimensionale parametrische Probit Modell (1.12). Mit der logistischen Verteilung resultiert das mehrdimensionale logistische Modell (1.10). Soll nun die ICC frei geschätzt werden, muss die Verteilung P_F als Parameter betrachtet werden und im Rahmen der nonparametrischen Bayes-Inferenz eine Priori-Verteilung erhalten. Wird als Priori-Verteilung für P_F ähnlich wie in Miyazaki und Hoshino (2009) und Karabatsos (2016) eine trunkierte Dirichlet Prozess Mean Mixture von Normalverteilungen mit Varianzparameter σ^2 angenommen, also

$$P_F \sim TDPM(N(\delta_{ij}, \sigma^2), P_0, \lambda) \tag{5.5}$$

folgt mit (5.3) und (3.33)

$$\epsilon_{ij}|\delta_{ij}, \sigma^2 \overset{ind}{\sim} N(\delta_{ij}, \sigma^2)$$

$$\delta_{ij}|P \overset{i.i.d.}{\sim} P$$

$$P \sim TDP(\lambda, P_0)$$

und deshalb wegen (5.2)

$$Z_{ij}|\eta_{ij}, \delta_{ij}, \sigma^2 \overset{ind}{\sim} N(\delta_{ij} - \eta_{ij}, \sigma^2). \tag{5.6}$$

Da δ_{ij} als Verteilung die Realisation eines auf Ebene L trunkierten TDP hat, gibt es aufgrund der Stick-Breaking Repräsentation (3.26) des TDP bei gegebenem P eine endliche Menge von L Ausprägungen, die δ_{ij} annehmen kann. Diese seien in einem Vektor $\boldsymbol{\delta}$ mit Elementen δ_l, $l = 1, 2, \ldots, L$, zusammengefasst. Die zu den δ_l gehörenden Mischverteilungsgewichte π_l, $l = 1, 2, \ldots, L$ aus (3.26) seien in dem Vektor $\boldsymbol{\pi}$ zusammengefasst.

Die Reformulierung des Modells induziert in Verbindung mit der Stick-Breaking Repräsentation (3.26) des TDP eine Priori-Verteilung für die ICC F. Eine Realisation der ICC hat wegen (3.31) die Form

$$F(x) = \sum_{l=1}^{L} \pi_l \Phi(x; \delta_l, \sigma^2), \tag{5.7}$$

wobei $\Phi(x; \delta_l, \sigma^2)$ die Verteilungsfunktion einer Normalverteilung mit Erwartungswert δ_l und Varianz σ^2 ist.

Ishwaran und Zarepour (2000) sowie Miyazaki und Hoshino (2009) folgend seien zudem die in einer Matrix $\mathbf{c}$ zusammengefassten diskreten Clustervariablen $c_{ij} \in \{1, 2, \ldots, L\}$ mit der Eigenschaft

$$c_{ij} = l \Leftrightarrow \delta_{ij} = \delta_l \tag{5.8}$$

definiert. Diese Clustervariablen geben für jede Beobachtung an, an welcher Position in $\boldsymbol{\delta}$ sich das zu dieser Beobachtung gehörige δ_{ij} befindet. Man kann dann äquivalent zu (5.6) schreiben:

$$Z_{ij}|\eta_{ij}, \delta, c_{ij}, \sigma^2 \overset{ind}{\sim} N(\delta_{c_{ij}} - \eta_{ij}, \sigma^2) \tag{5.9}$$

Die Parameter des in dieser Art und Weise reformulierten Modells sind die zur TDPM gehörigen Parameter $\boldsymbol{\pi}$, $\boldsymbol{\delta}$ und σ^2, die Itemparameter $\boldsymbol{\beta}$ und $\boldsymbol{\alpha}$, die Personenparameter $\boldsymbol{\xi}$, sowie die Hilfsparameter $\mathbf{z}$ und $\mathbf{c}$.

5.1.2 Likelihood

Die Likelihood hängt wegen (5.4) nur von den latenten Hilfsvariablen Z_{ij} ab und hat die Form

$$\begin{aligned} f(\mathbf{x}|\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) &= f(\mathbf{x}|\mathbf{z}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{k} P(X_{ij} = x_{ij}|Z_{ij} = z_{ij}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{k} [I(x_{ij} = 1)I(z_{ij} \leq 0) + I(x_{ij} = 0)I(z_{ij} > 0)] \end{aligned}$$

(Albert, 1992). Falls $z_{ij} \leq 0$ ist, nimmt die Wahrscheinlichkeit dafür, dass Person i Item j löst, den Wert 1 an. Ist hingegen $z_{ij} > 0$, so ist die Wahrscheinlichkeit gleich 0.

5.1.3 Priori-Verteilungen

Die Dichte $f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi})$ der Priori-Verteilung der Parameter des Modells kann in mehreren Schritten hierarchisch konstruiert werden.

Die latenten Hilfsvariablen Z_{ij} sind wegen (5.9) bedingt auf die anderen Parameter unabhängig von $\boldsymbol{\pi}$. Die gemeinsame Priori-Dichte der Parameter kann deshalb faktorisiert werden:

$$\begin{aligned} f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) &= f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) \\ &= f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) \end{aligned}$$

Aufgrund von (5.9) gilt für die latenten Hilfsvariablen

$$f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) = \prod_{i=1}^{n} \prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2).$$

Für die Personenparameter $\boldsymbol{\xi}$, die Itemparameter $\boldsymbol{\beta}$ und $\boldsymbol{\alpha}$ und die zur ICC gehörigen Parameter σ^2, $\boldsymbol{\delta}$, $\mathbf{c}$ und $\boldsymbol{\pi}$ werden unabhängige Priori-Verteilungen gewählt:

$$f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) = f(\boldsymbol{\xi}) f(\boldsymbol{\beta}, \boldsymbol{\alpha}) f(\sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi})$$

Die Parameter $\boldsymbol{\xi}_i$ der einzelnen Personen können a priori als unabhängig angenommen werden:

$$f(\boldsymbol{\xi}) = \prod_{i=1}^{n} f(\boldsymbol{\xi}_i)$$

In den restringierten Personenparametern der Personen 1 bis Q ist jeweils nur der i-te Parameter frei. Für diesen kann jeweils eine Normalverteilung mit Erwartungswert 0 und Varianz σ_ξ^2 gewählt werden, also

$$f(\boldsymbol{\xi}_i) = f(\xi_{ii}) = N(\xi_{ii}; 0, \sigma_\xi^2)$$

für $i = 1, 2, \ldots, Q$. Die Personenparameter der Personen $Q+1$ und $Q+2$ sind vollständig fixiert. Für diese beiden Personen bedarf es daher keiner Priori-Verteilung. Für die unrestringierten Personenparameter der Personen $Q+3$ bis n können die folgenden Q-dimensionalen Normalverteilungen angenommen werden:

$$f(\boldsymbol{\xi}_i) = MVN(\boldsymbol{\xi}_i; \mathbf{0}, \sigma_\xi^2 \mathbf{I}_Q)$$

für $i = Q+3, Q+4, \ldots, n$, wobei $\mathbf{I}_Q$ die Q-dimensionale Einheitsmatrix ist.

Die Itemparameter sind ebenfalls unabhängig:

$$f(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{j=1}^{k} f(\beta_j, \boldsymbol{\alpha}_j)$$

Aufgrund der in Modell 3 vorgeschlagenen Reparametrisierung sind die Ladungsparameter unbeschränkt und es können für die Vektoren $(\beta_j \boldsymbol{\alpha}_j) := (\beta_j, \alpha_{j1}, \alpha_{j2}, \ldots \alpha_{jQ})^\top$ der unrestringierten Items $j = 2, 3, \ldots, k$ multivariate $(Q+1)$-dimensionale Normalverteilungen mit Erwartungswertvektor $\mathbf{0}$ und Varianzen $\sigma_{\beta\alpha}^2$ angenommen werden:

$$f(\beta_j, \boldsymbol{\alpha}_j) = MVN((\beta_j \boldsymbol{\alpha}_i); \mathbf{0}, \sigma_{\beta\alpha}^2 \mathbf{I}_{Q+1})$$

für $j = 2, 3, \ldots, k$. Das erste Item hat wegen der Fixierung des Schwierigkeitsparameters und der ersten Ladung nur $Q-1$ Parameter. Für diese wird wie für die unrestringierten Items eine multivariate Normalverteilung mit gleichen Varianzen $\sigma_{\beta\alpha}^2$ für die Komponenten gewählt:

$$f(\beta_1, \boldsymbol{\alpha}_1) = f(\boldsymbol{\alpha}_1) = MVN(\boldsymbol{\alpha}_1; \mathbf{0}, \sigma_{\beta\alpha}^2 \mathbf{I}_{Q-1})$$

Die Priori-Verteilungen für die ICC-Parameter σ^2, $\boldsymbol{\delta}$, $\mathbf{c}$ und $\boldsymbol{\pi}$ ergeben sich aus der Definition des TDPM und der Stick-Breaking Repräsentation (3.26) des TDP (Ishwaran

& Zarepour, 2000). Zunächst sind σ^2 und δ unabhängig voneinander und unabhängig von $\mathbf{c}$ und $\boldsymbol{\pi}$:

$$f(\sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) = f(\sigma^2) f(\boldsymbol{\delta}) f(\mathbf{c}, \boldsymbol{\pi})$$

Für σ^2 kann aus Gründen der Konjugiertheit eine inverse Gammaverteilung mit Shape-Parameter α_{σ^2} und Rate-Parameter β_{σ^2} gewählt werden (Ishwaran & Zarepour, 2000):

$$f(\sigma^2) = IG(\sigma^2; \alpha_{\sigma^2}, \beta_{\sigma^2})$$

Die Elemente δ_l von $\boldsymbol{\delta}$ sind wegen (3.23) unabhängig und folgen jeweils der Basisverteilung des TDP. Für diese Basisverteilung wird eine Normalverteilung mit Erwartungswert 0 und Varianz σ_δ^2 gewählt:

$$f(\boldsymbol{\delta}) = \prod_{l=1}^{L} f(\delta_l) = \prod_{l=1}^{L} N(\delta_l; 0, \sigma_\delta^2)$$

Die Priori-Verteilung für $\mathbf{c}$ und $\boldsymbol{\pi}$ ist hierarchisch definiert:

$$f(\mathbf{c}, \boldsymbol{\pi}) = f(\mathbf{c}|\boldsymbol{\pi}) f(\boldsymbol{\pi}),$$

wobei die Indikatorvariablen c_{ij} bei gegebenem $\boldsymbol{\pi}$ unabhängig sind. Aufgrund von (5.8) und der Stick-Breaking Darstellung (3.26) folgen sie jeweils einer Multinomialverteilung mit Parametervektor $\boldsymbol{\pi}$ (Ishwaran & Zarepour, 2000):

$$f(\mathbf{c}|\boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{j=1}^{k} f(c_{ij}|\boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{j=1}^{k} M(c_{ij}; \boldsymbol{\pi})$$

Die Priori-Verteilung von $\boldsymbol{\pi}$ ergibt sich ebenfalls aus der Stick-Breaking Repräsentation (3.26) des TDP (Ishwaran & Zarepour, 2000):

$$\pi_l = B_l \prod_{i=1}^{l-1} (1 - B_i)$$

für $l = 1, 2, \ldots, L-1$ mit $B_l \overset{iid}{\sim} Beta(1, \lambda)$ und λ als Präzisionsparameter des TDP und

$$\pi_L = 1 - \sum_{l=1}^{L-1} \pi_l.$$

Aus dieser Definition der π_l lässt sich die Dichte $f(\boldsymbol{\pi})$ herleiten. Es handelt sich um die bekannte Dichte einer generalisierten Dirichletverteilung (Ishwaran & James, 2001).

Zusammenfassend hat die gemeinsame Priori-Dichte die Form

$$f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) = f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}, \boldsymbol{\alpha}) f(\sigma^2) f(\boldsymbol{\delta}) f(\mathbf{c}|\boldsymbol{\pi}) f(\boldsymbol{\pi}).$$

5.1.4 Hyperparameter

Als Hyperparameter des Modells müssen festgelegt werden:

- die Varianz σ_ξ^2 der Priori-Verteilungen der Personenparameter,
- die Varianz $\sigma_{\beta\alpha}^2$ der Priori-Verteilungen der Itemparameter,
- der Präzisionsparameter λ des TDP,
- die Varianz σ_δ^2 der Basisverteilung des TDP und
- die Parameter α_{σ^2} und β_{σ^2} der inversen Gamma-Priori-Verteilung für σ^2.

Schließlich muss noch ein Wert für die Trunkierungskonstante L des TDP festgelegt werden.

5.1.5 Posteriori-Verteilung

Die Dichte der Posteriori-Verteilung ist nach dem Satz von Bayes proportional zur gemeinsamen Dichte der Daten und der Parameter beziehungsweise zum Produkt der Likelihood und der gemeinsamen Priori-Dichte:

$$\begin{aligned} f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}|\mathbf{x}) &\propto f(\mathbf{x}, \mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) \\ &\propto f(\mathbf{x}|\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) \\ &= f(\mathbf{x}|\mathbf{z}) f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\pi}) \\ &= f(\mathbf{x}|\mathbf{z}) f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}, \boldsymbol{\alpha}) f(\sigma^2) f(\boldsymbol{\delta}) f(\mathbf{c}|\boldsymbol{\pi}) f(\boldsymbol{\pi}) \end{aligned}$$

5.1.6 Vollständig bedingte Verteilungen

Mit der Reformulierung des Modells in Abschnitt 5.1.1 und den in Abschnitt 5.1.3 gewählten Priori-Verteilungen können die vollständig bedingten Verteilungen für alle Parameter hergeleitet werden. Der hier geführte Beweis ergibt sich aus einer Kombination von Resultaten aus Ishwaran und Zarepour (2000) mit Ideen aus Béguin und Glas (2001). Von Béguin und Glas (2001) wurde insbesondere der Ansatz übernommen, die vollständig bedingten Dichten der Item- und Personenparameter auf Posteriori-Dichten von linearen Modellen zurückzuführen.

Latente Hilfsvariablen

Die vollständig bedingte Dichte der latenten Hilfsvariablen $\mathbf{z}$ ist proportional zur Posteriori-Dichte und somit

$$\begin{aligned}
f(\mathbf{z}|\cdot) &\propto f(\mathbf{x}|\mathbf{z})f(\mathbf{z}|\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\boldsymbol{\delta},\mathbf{c})f(\boldsymbol{\xi})f(\boldsymbol{\beta},\boldsymbol{\alpha})f(\sigma^2)f(\boldsymbol{\delta})f(\mathbf{c}|\boldsymbol{\pi})f(\boldsymbol{\pi}) \\
&\propto f(\mathbf{x}|\mathbf{z})f(\mathbf{z}|\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\boldsymbol{\delta},\mathbf{c}) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{k}[I(x_{ij}=1)I(z_{ij}\leq 0)+I(x_{ij}=0)I(z_{ij}>0)]\prod_{i=1}^{n}\prod_{j=1}^{k}N(z_{ij};\delta_{c_{ij}}-\eta_{ij},\sigma^2) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{k}\begin{cases}N(z_{ij};\delta_{c_{ij}}-\eta_{ij},\sigma^2)I(z_{ij}\leq 0) & \text{falls } x_{ij}=1\\ N(z_{ij};\delta_{c_{ij}}-\eta_{ij},\sigma^2)I(z_{ij}>0) & \text{falls } x_{ij}=0\end{cases}
\end{aligned}$$

Damit kann $Z_{ij}|\cdot$ für jede Beobachtung unabhängig aus einer trunkierten Normalverteilung gezogen werden: Ist $x_{ij}=1$ wird $Z_{ij}|\cdot$ aus einer rechtsseitig an der Stelle $z_{ij}=0$ trunkierten Normalverteilung mit Erwartungswert $\delta_{c_{ij}}-\eta_{ij}$ und Varianz σ^2 gezogen. Ist $x_{ij}=0$, wird $Z_{ij}|\cdot$ aus einer linksseitig an der Stelle $z_{ij}=0$ trunkierten Normalverteilung mit Erwartungswert $\delta_{c_{ij}}-\eta_{ij}$ und Varianz σ^2 gezogen.

Personenparameter

Für die Personenparameter $\boldsymbol{\xi}$ ergibt sich:

$$\begin{aligned}
f(\boldsymbol{\xi}|\cdot) &\propto f(\mathbf{x}|\mathbf{z})f(\mathbf{z}|\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\boldsymbol{\delta},\mathbf{c})f(\boldsymbol{\xi})f(\boldsymbol{\beta},\boldsymbol{\alpha})f(\sigma^2)f(\boldsymbol{\delta})f(\mathbf{c}|\boldsymbol{\pi})f(\boldsymbol{\pi}) \\
&\propto f(\mathbf{z}|\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\boldsymbol{\delta},\mathbf{c})f(\boldsymbol{\xi}) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{k}N(z_{ij};\delta_{c_{ij}}-\eta_{ij},\sigma^2)\prod_{i=1}^{n}f(\boldsymbol{\xi}_i) \\
&= \prod_{i=1}^{n}\left[\prod_{j=1}^{k}N(z_{ij};\delta_{c_{ij}}-\eta_{ij},\sigma^2)f(\boldsymbol{\xi}_i)\right]
\end{aligned}$$

Da die vollständig bedingte Dichte faktorisierbar ist, sind die Personenparameter $\boldsymbol{\xi}_i$ bedingt auf die Daten und alle übrigen Parameter des Modells unabhängig.

Für die unrestringierten Personenparameter $i=Q+3,Q+4,\ldots,n$ gilt:

$$\begin{aligned}
f(\boldsymbol{\xi}_i|\cdot) &\propto \prod_{j=1}^{k}N(z_{ij};\delta_{c_{ij}}-\eta_{ij},\sigma^2)MVN(\boldsymbol{\xi}_i;\mathbf{0},\sigma_\xi^2\mathbf{I}_Q) \\
&= \prod_{j=1}^{k}N(z_{ij};\delta_{c_{ij}}-\beta_j-\boldsymbol{\alpha}_j^\top\boldsymbol{\xi}_i,\sigma^2)MVN(\boldsymbol{\xi}_i;\mathbf{0},\sigma_\xi^2\mathbf{I}_Q) \\
&= \prod_{j=1}^{k}N(z_{ij}-\delta_{c_{ij}}+\beta_j;-\boldsymbol{\alpha}_j^\top\boldsymbol{\xi}_i,\sigma^2)MVN(\boldsymbol{\xi}_i;\mathbf{0},\sigma_\xi^2\mathbf{I}_Q),
\end{aligned}$$

wobei im letzten Schritt die Eigenschaft $N(x; \mu + c, \sigma^2) = N(x - c; \mu, \sigma^2)$ der Dichte der Normalverteilung ausgenutzt wurde. Die resultierende Funktion

$$\prod_{j=1}^{k} N(z_{ij} - \delta_{c_{ij}} + \beta_j; -\boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i, \sigma^2) MVN(\boldsymbol{\xi}_i; \mathbf{0}, \sigma_\xi^2 \mathbf{I}_Q)$$

entspricht dem Zähler der Dichte der Posteriori-Verteilung der Regressionsgewichte $\boldsymbol{\xi}_i$ eines bayesianischen linearen Modells mit k unabhängigen Beobachtungen der normalverteilten Kriteriumsvariable $z_{ij} - \delta_{c_{ij}} + \beta_j$ und der Prädiktorvariablen $-\boldsymbol{\alpha}_j$ bei bekannter Varianz σ^2 und multivariater Normalverteilung als Priori-Verteilung für $\boldsymbol{\xi}_i$. Die $k \times Q$ Design-Matrix dieses linearen Modells hat die folgende Form:

$$\mathbf{X}_\xi := \begin{pmatrix} -\alpha_{11} & -\alpha_{12} & \cdots & -\alpha_{1Q} \\ -\alpha_{21} & -\alpha_{22} & \cdots & -\alpha_{2Q} \\ \vdots & \vdots & \vdots & \vdots \\ -\alpha_{k1} & -\alpha_{k2} & \cdots & -\alpha_{kQ} \end{pmatrix}$$

Sie ist für alle Personen $i = Q + 3, Q + 4, \ldots, n$ identisch, da in ihr nur die Itemparameter aus $\boldsymbol{\alpha}_j$ vorkommen.

Der k-dimensionale Kriteriumsvektor ist für jede Person $i = Q + 3, Q + 4, \ldots, n$

$$\mathbf{y}_{\xi i} := \begin{pmatrix} z_{i1} - \delta_{c_{i1}} + \beta_1 \\ z_{i2} - \delta_{c_{i2}} + \beta_2 \\ \vdots \\ z_{ik} - \delta_{c_{ik}} + \beta_k \end{pmatrix}.$$

Dieser unterscheidet sich von Person zu Person. Damit kann das lineare Modell für alle $i = Q + 3, Q + 4, \ldots, n$ wie folgt geschrieben werden:

$$\begin{aligned} \mathbf{y}_{\xi i} &\sim MVN(\mathbf{X}_\xi \boldsymbol{\xi}_i, \sigma^2 \mathbf{I}_k) \\ \boldsymbol{\xi}_i &\sim MVN(\mathbf{0}, \sigma_\xi^2 \mathbf{I}_Q) \end{aligned}$$

Die Posteriori-Verteilung von $\boldsymbol{\xi}_i$ in diesem linearen Modell ist eine Normalverteilung mit Erwartungswertvektor

$$\boldsymbol{\mu}_{\xi i} = \left(\mathbf{X}_\xi^\top \mathbf{X}_\xi + \frac{\sigma^2}{\sigma_\xi^2} \mathbf{I}_Q \right)^{-1} \mathbf{X}_\xi^\top \mathbf{y}_{\xi i}$$

und Kovarianzmatrix

$$\boldsymbol{\Sigma}_\xi = \sigma^2 \left(\mathbf{X}_\xi^\top \mathbf{X}_\xi + \frac{\sigma^2}{\sigma_\xi^2} \mathbf{I}_Q \right)^{-1}$$

(Gelman et al., 2014). Daher kann $\boldsymbol{\xi}_i | \cdot$ für jedes $i = Q + 3, Q + 4, \ldots, n$ aus einer Normalverteilung mit Erwartungswertvektor $\boldsymbol{\mu}_{\xi i}$ und Kovarianzmatrix $\boldsymbol{\Sigma}_\xi$ gezogen werden.

Für die restringierten Personenparameter $i = 1, 2, \ldots, Q$ ist

$$\begin{aligned} f(\boldsymbol{\xi}_i|\cdot) &\propto \prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) N(\xi_{ii}; 0, \sigma_\xi^2) \\ &= \prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \beta_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i, \sigma^2) N(\xi_{ii}; 0, \sigma_\xi^2) \\ &= \prod_{j=1}^{k} N(z_{ij} - \delta_{c_{ij}} + \beta_j; -\alpha_{ji}\xi_{ii}, \sigma^2) N(\xi_{ii}; 0, \sigma_\xi^2). \end{aligned}$$

Der letzte Term ist proportional zu der Posteriori-Dichte eines Regressionsgewichts ξ_{ii} in einem bayesianischen linearen Modell mit k unabhängigen Beobachtungen einer normalverteilten Kriteriumsvariable $z_{ij} - \delta_{c_{ij}} + \beta_j$ und einer Prädiktorvariable $-\alpha_{ji}$ mit normalverteilter Priori-Verteilung für das Regressionsgewicht ξ_{ii} bei bekannter Varianz σ^2 der Beobachtungen.

Die $k \times 1$ Designmatrix dieses Modells ist

$$\mathbf{X}_{\xi_{restr}i} := \begin{pmatrix} -\alpha_{1i} \\ -\alpha_{2i} \\ \vdots \\ -\alpha_{ki} \end{pmatrix}$$

und der k-dimensionale Kriteriumsvektor ist

$$\mathbf{y}_{\xi_{restr}i} := \begin{pmatrix} z_{i1} - \delta_{c_{i1}} + \beta_1 \\ z_{i2} - \delta_{c_{i2}} + \beta_2 \\ \vdots \\ z_{ik} - \delta_{c_{ik}} + \beta_k \end{pmatrix}$$

für jede Person $i = 1, 2, \ldots, Q$. Das vollständige Modell ist für jede Person durch

$$\begin{aligned} \mathbf{y}_{\xi_{restr}i} &\sim MVN(\mathbf{X}_{\xi_{restr}i}\xi_{ii}, \sigma^2 \mathbf{I}_k) \\ \xi_{ii} &\sim N(0, \sigma_\xi^2) \end{aligned}$$

gegeben. Die Posteriori-Verteilung von ξ_{ii} ist in diesem Fall wieder eine Normalverteilung mit Erwartungswert

$$\mu_{\xi_{restr}i} = \left(\mathbf{X}_{\xi_{restr}i}^\top \mathbf{X}_{\xi_{restr}i} + \frac{\sigma^2}{\sigma_\xi^2} \mathbf{I}_1 \right)^{-1} \mathbf{X}_{\xi_{restr}i}^\top \mathbf{y}_{\xi_{restr}i} = \frac{\mathbf{X}_{\xi_{restr}i}^\top \mathbf{y}_{\xi_{restr}i}}{\mathbf{X}_{\xi_{restr}i}^\top \mathbf{X}_{\xi_{restr}i} + \frac{\sigma^2}{\sigma_\xi^2}}$$

und Varianz

$$\sigma_{\xi_{restr}i}^2 = \sigma^2 \left(\mathbf{X}_{\xi_{restr}i}^\top \mathbf{X}_{\xi_{restr}i} + \frac{\sigma^2}{\sigma_\xi^2} \mathbf{I}_1 \right)^{-1} = \frac{\sigma^2}{\mathbf{X}_{\xi_{restr}i}^\top \mathbf{X}_{\xi_{restr}i} + \frac{\sigma^2}{\sigma_\xi^2}}$$

(Gelman et al., 2014). Die Parameter $\xi_{ii}|\cdot$ der Personen $i = 1, 2, \ldots, Q$ können daher jeweils aus einer Normalverteilung mit Erwartungswert $\mu_{\xi_{restr}i}$ und Varianz $\sigma^2_{\xi_{restr}i}$ gezogen werden.

Die Parameter der Personen $Q + 1$ und $Q + 2$ sind komplett fixiert und müssen daher nicht gezogen werden.

Itemparameter

Für die Itemparameter $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ ergibt sich:

$$\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{\alpha}|\cdot) &\propto f(\mathbf{x}|\mathbf{z}) f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}, \boldsymbol{\alpha}) f(\sigma^2) f(\boldsymbol{\delta}) f(\mathbf{c}|\boldsymbol{\pi}) f(\boldsymbol{\pi}) \\
&\propto f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) f(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) \prod_{j=1}^{k} f(\beta_j, \boldsymbol{\alpha}_j) \\
&= \prod_{j=1}^{k} \left[\prod_{i=1}^{n} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) f(\beta_j, \boldsymbol{\alpha}_j) \right]
\end{aligned}$$

Somit sind die Itemparameter $(\beta_j, \boldsymbol{\alpha}_j)$ bedingt auf die Daten und alle übrigen Parameter des Modells unabhängig.

Für alle unrestringierten Items $j = 2, 3, \ldots, k$ gilt:

$$\begin{aligned}
f(\beta_j, \boldsymbol{\alpha}_j|\cdot) &\propto \prod_{i=1}^{n} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) MVN((\beta_j \boldsymbol{\alpha}_j); \mathbf{0}, \sigma^2_{\beta\alpha} \mathbf{I}_{Q+1}) \\
&= \prod_{i=1}^{n} N(z_{ij}; \delta_{c_{ij}} - \beta_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i, \sigma^2) MVN((\beta_j \boldsymbol{\alpha}_j); \mathbf{0}, \sigma^2_{\beta\alpha} \mathbf{I}_{Q+1}) \\
&= \prod_{i=1}^{n} N(z_{ij} - \delta_{c_{ij}}; -\beta_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i, \sigma^2) MVN((\beta_j \boldsymbol{\alpha}_j); \mathbf{0}, \sigma^2_{\beta\alpha} \mathbf{I}_{Q+1})
\end{aligned}$$

Diese vollständig bedingte Dichte ist proportional zu der Posteriori-Dichte eines bayesianischen linearen Modells mit $n \times (Q + 1)$ Designmatrix

$$\mathbf{X}_{\alpha\beta} := \begin{pmatrix} -1 & -\xi_{11} & -\xi_{12} & \cdots & -\xi_{1Q} \\ -1 & -\xi_{21} & -\xi_{22} & \cdots & -\xi_{2Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -\xi_{n1} & -\xi_{n2} & \cdots & -\xi_{nQ} \end{pmatrix}$$

und n-dimensionalem Kriteriumsvektor

$$\mathbf{y}_{\alpha\beta j} := \begin{pmatrix} z_{1j} - \delta_{c_{1j}} \\ z_{2j} - \delta_{c_{2j}} \\ \vdots \\ z_{nj} - \delta_{c_{nj}} \end{pmatrix}.$$

Dieses lineare Modell kann wie folgt formuliert werden:

$$\begin{aligned} \mathbf{y}_{\alpha\beta j} &\sim MVN(\mathbf{X}_{\alpha\beta}(\beta_j \boldsymbol{\alpha}_j), \sigma^2 \mathbf{I}_n) \\ (\beta_j \boldsymbol{\alpha}_j) &\sim MVN(\mathbf{0}, \sigma^2_{\beta\alpha} \mathbf{I}_{Q+1}) \end{aligned}$$

Die Posteriori-Verteilung von $(\beta_j \boldsymbol{\alpha}_j)$ in diesem Modell ist eine multivariate Normalverteilung mit Erwartungswertvektor

$$\boldsymbol{\mu}_{\alpha\beta j} = \left(\mathbf{X}^\top_{\alpha\beta} \mathbf{X}_{\alpha\beta} + \frac{\sigma^2}{\sigma^2_{\alpha\beta}} \mathbf{I}_{Q+1} \right)^{-1} \mathbf{X}^\top_{\alpha\beta} \mathbf{y}_{\alpha\beta j}$$

und Kovarianzmatrix

$$\boldsymbol{\Sigma}_{\alpha\beta} = \sigma^2 \left(\mathbf{X}^\top_{\alpha\beta} \mathbf{X}_{\alpha\beta} + \frac{\sigma^2}{\sigma^2_{\alpha\beta}} \mathbf{I}_{Q+1} \right)^{-1}$$

(Gelman et al., 2014). Die Parameter $(\beta_j \boldsymbol{\alpha}_j)|\cdot$ können daher für alle Items außer dem ersten jeweils unabhängig aus dieser Verteilung gezogen werden.

Für das erste Item ergibt sich aufgrund der Restriktionen:

$$\begin{aligned} f(\beta_1, \boldsymbol{\alpha}_1|\cdot) &= f(\boldsymbol{\alpha}_1|\cdot) \\ &\propto \prod_{i=1}^{n} N(z_{i1}; \delta_{c_{i1}} - \eta_{i1}, \sigma^2) MVN(\boldsymbol{\alpha}_1; \mathbf{0}, \sigma^2_{\beta\alpha} \mathbf{I}_{Q-1}) \\ &= \prod_{i=1}^{n} N(z_{i1}; \delta_{c_{i1}} - \beta_1 - \boldsymbol{\alpha}_1^\top \boldsymbol{\xi}_i, \sigma^2) MVN(\boldsymbol{\alpha}_1; \mathbf{0}, \sigma^2_{\beta\alpha} \mathbf{I}_{Q-1}) \\ &= \prod_{i=1}^{n} N(z_{i1}; \delta_{c_{i1}} - 0 - \alpha_{11}\xi_{i1} - \sum_{q=2}^{Q} \alpha_{1q}\xi_{iq}, \sigma^2) MVN(\boldsymbol{\alpha}_1; \mathbf{0}, \sigma^2_{\beta\alpha} \mathbf{I}_{Q-1}) \\ &= \prod_{i=1}^{n} N(z_{i1}; \delta_{c_{i1}} - \xi_{i1} - \sum_{q=2}^{Q} \alpha_{1q}\xi_{iq}, \sigma^2) MVN(\boldsymbol{\alpha}_1; \mathbf{0}, \sigma^2_{\beta\alpha} \mathbf{I}_{Q-1}) \\ &= \prod_{i=1}^{n} N(z_{i1} - \delta_{c_{i1}} + \xi_{i1}; -\sum_{q=2}^{Q} \alpha_{1q}\xi_{iq}, \sigma^2) MVN(\boldsymbol{\alpha}_1; \mathbf{0}, \sigma^2_{\beta\alpha} \mathbf{I}_{Q-1}) \end{aligned}$$

Diese Funktion ist wieder proportional zur Posteriori-Dichte eines linearen Modells mit Designmatrix

$$\mathbf{X}_{\alpha_1} := \begin{pmatrix} -\xi_{12} & -\xi_{13} & \cdots & -\xi_{1Q} \\ -\xi_{22} & -\xi_{23} & \cdots & -\xi_{2Q} \\ \vdots & \vdots & \vdots & \vdots \\ -\xi_{n2} & -\xi_{n3} & \cdots & -\xi_{nQ} \end{pmatrix}$$

und Kriteriumsvektor

$$\mathbf{y}_{\alpha_1} := \begin{pmatrix} z_{11} - \delta_{c_{11}} + \xi_{11} \\ z_{21} - \delta_{c_{21}} + \xi_{21} \\ \vdots \\ z_{n1} - \delta_{c_{n1}} + \xi_{n1} \end{pmatrix},$$

so dass

$$\begin{aligned} \mathbf{y}_{\alpha_1} &\sim MVN(\mathbf{X}_{\alpha_1}\boldsymbol{\alpha}_1, \sigma^2\mathbf{I}_n) \\ \boldsymbol{\alpha}_1 &\sim MVN(\mathbf{0}, \sigma^2_{\beta\alpha}\mathbf{I}_{Q-1}). \end{aligned}$$

Die Posteriori-Verteilung von $\boldsymbol{\alpha}_1$ in diesem Modell ist eine multivariate Normalverteilung mit Erwartungswertvektor

$$\boldsymbol{\mu}_{\alpha_1} = \left(\mathbf{X}^\top_{\alpha_1}\mathbf{X}_{\alpha_1} + \frac{\sigma^2}{\sigma^2_{\alpha\beta}}\mathbf{I}_{Q-1}\right)^{-1}\mathbf{X}^\top_{\alpha_1}\mathbf{y}_{\alpha_1}$$

und Kovarianzmatrix

$$\boldsymbol{\Sigma}_{\alpha_1} = \sigma^2\left(\mathbf{X}^\top_{\alpha_1}\mathbf{X}_{\alpha_1} + \frac{\sigma^2}{\sigma^2_{\alpha\beta}}\mathbf{I}_{Q-1}\right)^{-1}$$

(Gelman et al., 2014). Die vollständig bedingte Verteilung der Itemparameter des ersten Items entspricht somit dieser Verteilung.

Varianzparameter der TDPM

Für den Varianzparameter σ^2 der TDPM gilt Ishwaran und Zarepour (2000) folgend:

$$\begin{aligned} f(\sigma^2|\cdot) &\propto f(\mathbf{x}|\mathbf{z})f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c})f(\boldsymbol{\xi})f(\boldsymbol{\beta}, \boldsymbol{\alpha})f(\sigma^2)f(\boldsymbol{\delta})f(\mathbf{c}|\boldsymbol{\pi})f(\boldsymbol{\pi}) \\ &\propto f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c})f(\sigma^2) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2)IG(\sigma^2; \alpha_{\sigma^2}, \beta_{\sigma^2}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \beta_j - \boldsymbol{\alpha}_j^\top\boldsymbol{\xi}_i, \sigma^2)IG(\sigma^2; \alpha_{\sigma^2}, \beta_{\sigma^2}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{k} N(z_{ij} - \delta_{c_{ij}} + \beta_j + \boldsymbol{\alpha}_j^\top\boldsymbol{\xi}_i; 0, \sigma^2)IG(\sigma^2; \alpha_{\sigma^2}, \beta_{\sigma^2}) \end{aligned}$$

Dies entspricht dem Zähler einer Posteriori-Dichte im Fall von nk unabhängigen, normalverteilten Beobachtungen $y_{\sigma^2 ij} := z_{ij} - \delta_{c_{ij}} + \beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i$ mit bekanntem Erwartungswert 0 und unbekannter Varianz σ^2, für die eine inverse Gammaverteilung mit den Parametern α_{σ^2} und β_{σ^2} angenommen wird. Da die inverse Gammaverteilung in diesem Fall konjugiert ist, folgt $\sigma^2|\cdot$ wieder einer inversen Gammaverteilung mit Parametern

$$\alpha_{\sigma^2 post} = \frac{nk}{2} + \alpha_{\sigma^2}$$

und

$$\beta_{\sigma^2 post} = \beta_{\sigma^2} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} y_{\sigma^2 ij}^2$$

(Gelman et al., 2014). $\sigma^2|\cdot$ kann daher aus einer inversen Gammaverteilung mit Parametern $\alpha_{\sigma^2 post}$ und $\beta_{\sigma^2 post}$ gezogen werden.

Lokationsparameter der TDPM

Für $\boldsymbol{\delta}$ ergibt sich:

$$\begin{aligned}
f(\boldsymbol{\delta}|\cdot) &\propto f(\mathbf{x}|\mathbf{z}) f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}, \boldsymbol{\alpha}) f(\sigma^2) f(\boldsymbol{\delta}) f(\mathbf{c}|\boldsymbol{\pi}) f(\boldsymbol{\pi}) \\
&\propto f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) f(\boldsymbol{\delta}) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) \prod_{l=1}^{L} N(\delta_l; \mu_\delta, \sigma_\delta^2) \\
&= \prod_{l=1}^{L} \left[\prod_{i=1}^{n} \prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) N(\delta_l; \mu_\delta, \sigma_\delta^2) \right].
\end{aligned}$$

Damit sind die $\delta_l|\cdot$ unabhängig und für jedes $l = 1, 2, \ldots, L$ ist

$$f(\delta_l|\cdot) \propto \prod_{i=1}^{n} \prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) N(\delta_l; \mu_\delta, \sigma_\delta^2). \tag{5.10}$$

Für alle $\delta_l|\cdot$ mit $c_{ij} \neq l$ für alle i und alle j gilt somit

$$f(\delta_l|\cdot) \propto N(\delta_l; \mu_\delta, \sigma_\delta^2),$$

da der Term $\prod_{i=1}^{n} \prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2)$ in (5.10) in diesem Fall nicht von δ_l abhängt und somit eine multiplikative Konstante darstellt. Diese $\delta_l|\cdot$ können also unabhängig aus der Basisverteilung des TDP gezogen werden (Ishwaran & Zarepour, 2000).

Für alle weiteren $\delta_l|\cdot$ gilt nach Ishwaran und Zarepour (2000):

$$\begin{aligned} f(\delta_l|\cdot) &\propto \prod_{ij:c_{ij}=l} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) N(\delta_l; \mu_\delta, \sigma_\delta^2) \\ &= \prod_{ij:c_{ij}=l} N(z_{ij} + \eta_{ij}; \delta_l, \sigma^2) N(\delta_l; \mu_\delta, \sigma_\delta^2) \\ &= \prod_{ij:c_{ij}=l} N(z_{ij} + \beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i; \delta_l, \sigma^2) N(\delta_l; \mu_\delta, \sigma_\delta^2) \end{aligned}$$

Diese Funktion ist für jedes dieser δ_l proportional zur Posteriori-Dichte im Fall von $n_l := |\{ij : c_{ij} = l\}|$ unabhängigen, normalverteilten Beobachtungen $y_{\delta_l ij} := z_{ij} + \beta_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\xi}_i$ mit unbekanntem Erwartungswert δ_l und bekannter Varianz σ^2, wobei eine Normalverteilung mit den Parametern μ_δ und σ_δ^2 als Priori-Verteilung für δ_l angenommen wird. Aus der Konjugiertheit der Normalverteilung in diesem Fall folgt, dass die Posteriori-Verteilung wieder eine Normalverteilung mit Erwartungswert

$$\mu_{\delta_l post} = \frac{\frac{1}{\sigma^2} \sum_{ij:c_{ij}=l} y_{\delta_l ij} + \frac{1}{\sigma_\delta^2} \mu_\delta}{\frac{n_l}{\sigma^2} + \frac{1}{\sigma_\delta^2}}$$

und Varianz

$$\sigma^2_{\delta_l post} = \left(\frac{1}{\sigma_\delta^2} + \frac{n_l}{\sigma^2} \right)^{-1}$$

ist (Gelman et al., 2014). Die übrigen $\delta_l|\cdot$ können somit aus unabhängigen Normalverteilungen jeweils mit Erwartungswert $\mu_{\delta_l post}$ und Varianz $\sigma^2_{\delta_l post}$ gezogen werden.

Clustervariablen der TDPM

Für die diskreten Clustervariablen $\mathbf{c}$ ergibt sich:

$$\begin{aligned} f(\mathbf{c}|\cdot) &\propto f(\mathbf{x}|\mathbf{z}) f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}, \boldsymbol{\alpha}) f(\sigma^2) f(\boldsymbol{\delta}) f(\mathbf{c}|\boldsymbol{\pi}) f(\boldsymbol{\pi}) \\ &\propto f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\delta}, \mathbf{c}) f(\mathbf{c}|\boldsymbol{\pi}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{k} N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) \prod_{i=1}^{n} \prod_{j=1}^{k} M(c_{ij}; \boldsymbol{\pi}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{k} \left[N(z_{ij}; \delta_{c_{ij}} - \eta_{ij}, \sigma^2) M(c_{ij}; \boldsymbol{\pi}) \right] \end{aligned}$$

Die $c_{ij}|\cdot$ sind somit für alle Beobachtungen unabhängig multinomialverteilt mit Parametern

$$\pi^*_{lij} \propto N(z_{ij}; \delta_l - \eta_{ij}, \sigma^2) \pi_l$$

(Ishwaran & Zarepour, 2000).

Mischverteilungsgewichte der TDPM

Für die Gewichte $\boldsymbol{\pi}$ ist schließlich

$$\begin{aligned} f(\boldsymbol{\pi}|\cdot) &\propto f(\mathbf{x}|\mathbf{z})f(\mathbf{z}|\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\boldsymbol{\delta},\mathbf{c})f(\boldsymbol{\xi})f(\boldsymbol{\beta},\boldsymbol{\alpha})f(\sigma^2)f(\boldsymbol{\delta})f(\mathbf{c}|\boldsymbol{\pi})f(\boldsymbol{\pi}) \\ &\propto f(\mathbf{c}|\boldsymbol{\pi})f(\boldsymbol{\pi}). \end{aligned}$$

Die Verteilung von $\boldsymbol{\pi}|\cdot$ entspricht somit der Posteriori-Verteilung von $\boldsymbol{\pi}$ gegeben die Clustervariablen $\mathbf{c}$. Diese hat aufgrund der Definition der Clustervariablen und der Konjugiertheit der generalisierten Dirichletverteilung zur Multinomialverteilung die folgende Posteriori Stick-Breaking Darstellung (Ishwaran & Zarepour, 2000):

$$\pi_l|\cdot = B_l^* \prod_{i=1}^{t-1}(1 - B_i^*)$$

für $l = 1, 2, \ldots, L-1$ mit

$$B_l^* \overset{ind}{\sim} Beta(1 + n_l, \lambda + \sum_{i=l+1}^{T} n_i)$$

wobei n_l gleich der Anzahl der c_{ij} mit $c_{ij} = l$ ist und

$$\pi_L|\cdot = 1 - \sum_{l=1}^{L-1} \pi_l|\cdot$$

Diese Darstellung kann zum Ziehen der $\boldsymbol{\pi}|\cdot$ verwendet werden.

5.1.7 Gibbs Sampler

Damit kann ein Gibbs Sampler konstruiert werden, in dem iterativ aus den vollständig bedingten Verteilungen der folgenden Parameter gezogen wird:

$$\begin{aligned} (\mathbf{z}|\cdot) \\ (\mathbf{c}|\cdot) \\ (\boldsymbol{\pi}|\cdot) \\ (\boldsymbol{\delta}|\cdot) \\ (\sigma^2|\cdot) \\ (\boldsymbol{\beta},\boldsymbol{\alpha}|\cdot) \\ (\boldsymbol{\xi}|\cdot) \end{aligned}$$

Dieser Gibbs Sampler soll als TDPM Gibbs Sampler bezeichnet werden.

5.2 Gibbs Sampler für ein NMIRT Modell mit Polya Tree

Auf Mischverteilungen basierende Modelle wie die TDPM führen in der Regel zu äußerst komplexen Posteriori-Dichten, die in der Praxis zu Konvergenzproblemen des TDPM Gibbs Samplers führen können. In diesem Fall muss auf alternative MCMC Algorithmen zurückgegriffen werden.

Der in diesem Abschnitt vorgeschlagene Gibbs Sampler vermeidet die Mischverteilungsstruktur, indem statt einer TDPM ein TPT als Priori-Verteilung für die ICC angenommen wird. Der hieraus resultierende Algorithmus stellt eine direkte Verallgemeinerung des von Hanson (2006) für ein binäres Regressionsmodell mit nonparametrischer Linkfunktion vorgeschlagenen Gibbs Samplers dar.

Ausgangspunkt ist die Parametrisierung des NMIRT in Modell 2 mit unrestringierter ICC und positiven Ladungsparametern. Der Schwierigkeitsparameter des ersten Items β_1 ist auf 0 und die Ladung des ersten Items auf den ersten Faktor α_{11} auf 1 gesetzt. Die Restriktionen für die Personenparameter sind $\boldsymbol{\xi}_1 = (1, 0, \ldots, 0)^\top$, $\boldsymbol{\xi}_2 = (0, 1, 0, \ldots, 0)^\top$, $\ldots$, $\boldsymbol{\xi}_Q = (0, \ldots, 0, 1)^\top$ und $\boldsymbol{\xi}_{Q+1} = (0, \ldots, 0)^\top$.

5.2.1 Reformulierung des Modells

Zunächst sei für jedes Item j und jede Dimension q

$$\alpha^*_{jq} := \ln(\alpha_{jq}) \tag{5.11}$$

definiert, so dass $\alpha^*_{jq} \in \mathbb{R}$ gilt. Die Struktur des latenten Prädiktors η_{ij} in Abhängigkeit von α^*_{jq} ist dann durch

$$\eta_{ij} = \beta_j + \sum_{q=1}^{Q} \exp(\alpha^*_{jq})\xi_{iq}$$

gegeben. Als neue Restriktion ergibt sich $\alpha^*_{11} = \ln(\alpha_{11}) = \ln(1) = 0$. Zudem kann das Modell wieder ähnlich wie in (5.2) - (5.4) mit latenten Hilfsvariablen Z_{ij} reformuliert werden (Hanson, 2006):

$$Z_{ij} \overset{i.i.d.}{\sim} P_F \tag{5.12}$$

mit

$$X_{ij} = I(Z_{ij} \leq \eta_{ij}). \tag{5.13}$$

Für P_F wird jedoch nun statt einer TDPM ein TPT in der Parametrisierung (3.14) als Priori-Verteilung angenommen:

$$P_F \sim TPT(\lambda, P_0) \tag{5.14}$$

Eine Realisation der ICC F hat in diesem Modell die Form der Verteilungsfunktion $F_{\mathbf{h}}$ einer Realisation des TPT in (3.17).

Die Parameter des reformulierten Modells sind die in einem Vektor $\mathbf{h}$ zusammengefassten Variablen $H_{e_l(y)}$ des TPT aus (3.9) - (3.12), welche die Realisationen des TPT festlegen, die Itemparameter $\boldsymbol{\beta}$ und $\boldsymbol{\alpha}^*$, die Personenparameter $\boldsymbol{\xi}$ und die Hilfsvariablen $\mathbf{z}$.

5.2.2 Likelihood

Die Likelihood des Modells kann auf zwei Arten formuliert werden: Unter Verwendung von (1.2) ohne die latenten Hilfsvariablen ergibt sich

$$f(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) = \prod_{i=1}^{n} \prod_{j=1}^{k} F_{\mathbf{h}}(\eta_{ij})^{x_{ij}} (1 - F_{\mathbf{h}}(\eta_{ij}))^{1-x_{ij}} \tag{5.15}$$

(Hanson, 2006), wobei die ICC $F_{\mathbf{h}}$ durch die Verteilungsfunktion der Realisation des TPT gegeben $\mathbf{h}$ aus (3.17) ist. In der Formulierung des Modells mit den latenten Hilfsvariablen $\mathbf{z}$ aus (5.12) ist die Verteilung der Daten $\mathbf{x}$ bedingt auf die die anderen Parameter wegen (5.13) unabhängig von $\mathbf{h}$ und es folgt für die Likelihood (Hanson, 2006):

$$\begin{aligned} f(\mathbf{x}|\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) &= f(\mathbf{x}|\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{k} [I(x_{ij} = 1) I(z_{ij} \leq \eta_{ij}) + I(x_{ij} = 0) I(z_{ij} > \eta_{ij})] \end{aligned} \tag{5.16}$$

5.2.3 Priori-Verteilungen

Die Verteilung von $\mathbf{h}$ ergibt sich direkt aus (3.9) - (3.12) in der Definition des TPT. Sie besitzt eine Dichte $f(\mathbf{h})$, die sich als Produkt von Dichten der Betaverteilung darstellen lässt. Die Personen- und Itemparameter werden a priori als unabhängig voneinander und von $\mathbf{h}$ angenommen, so dass

$$f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) = f(\boldsymbol{\xi}) f(\boldsymbol{\beta}) f(\boldsymbol{\alpha}^*) f(\mathbf{h})$$

gilt. Die Hilfsvariablen $\mathbf{z}$ hängen wegen (5.12) nur vom TPT und somit nur von $\mathbf{h}$ ab, wodurch sich

$$\begin{aligned} f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) &= f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h},) f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) \\ &= f(\mathbf{z}|\mathbf{h}) f(\mathbf{h}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}) f(\boldsymbol{\alpha}^*) \end{aligned}$$

ergibt.

Die auf $\mathbf{h}$ bedingte Dichte $f(\mathbf{z}|\mathbf{h})$ der Hilfsvariablen $\mathbf{z}$ kann aufgrund der bedingten Unabhängigkeit der Z_{ij} gegeben die Realisation des TPT in (5.12) faktorisiert werden:

$$f(\mathbf{z}|\mathbf{h}) = \prod_{i=1}^{n} \prod_{j=1}^{k} f(z_{ij}|\mathbf{h}) = \prod_{i=1}^{n} \prod_{j=1}^{k} f_{\mathbf{h}}(z_{ij}), \tag{5.17}$$

wobei $f_{\mathbf{h}}(z_{ij})$ die in (3.16) definierte Dichte einer Realisation des TPT ist.

Für die Personenparameter $\boldsymbol{\xi}$ und die Itemparameter $\boldsymbol{\beta}$ und $\boldsymbol{\alpha}^*$ werden unter Berücksichtigung der Restriktionen analog zu Abschnitt 5.1.3 unabhängige Normalverteilungen mit Erwartungswert 0 und Varianz σ_{ξ}^2 bzw. $\sigma_{\beta\alpha^*}^2$ gewählt. Wegen (5.11) folgt hieraus, dass die positiven Ladungsparameter in $\boldsymbol{\alpha}$ a priori lognormalverteilt sind.

5.2.4 Hyperparameter

Als Hyperparameter des Modells müssen festgelegt werden:

- die Varianz σ_ξ^2 der Priori-Verteilungen der Personenparameter,
- die Varianz $\sigma_{\beta\alpha^*}^2$ der Priori-Verteilungen der Itemparameter,
- der Präzisionsparameter λ des TPT und
- die Basisverteilung P_0 des TPT.

Außerdem muss noch ein Wert für die Trunkierungskonstante L des TPT festgesetzt werden.

5.2.5 Posteriori-Verteilung

Die Dichte der Posteriori-Verteilung ist nach dem Satz von Bayes proportional zur gemeinsamen Dichte der Daten und der Parameter bzw. zum Produkt der Likelihood und der gemeinsamen Priori-Dichte. In der Formulierung des Modells mit Hilfsvariablen führt dies zu

$$\begin{aligned} f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}|\mathbf{x}) &\propto f(\mathbf{x}, \mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) \\ &= f(\mathbf{x}|\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) \\ &= f(\mathbf{x}|\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*) f(\mathbf{z}|\mathbf{h}) f(\mathbf{h}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}) f(\boldsymbol{\alpha}^*). \end{aligned} \tag{5.18}$$

In der Formulierung ohne Hilfsvariablen ist die Posteriori-Dichte durch

$$\begin{aligned} f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}|\mathbf{x}) &\propto f(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) \\ &= f(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) \\ &= f(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) f(\mathbf{h}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}) f(\boldsymbol{\alpha}^*) \end{aligned} \tag{5.19}$$

gegeben.

5.2.6 Vollständig bedingte Verteilungen

Parameter des TPT

Die vollständig bedingte Dichte für die Parameter $\mathbf{h}$ des TPT kann anhand der Posteriori-Dichte aus (5.18) hergeleitet werden:

$$\begin{aligned} f(\mathbf{h}|\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{x}) &\propto f(\mathbf{x}|\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*) f(\mathbf{z}|\mathbf{h}) f(\mathbf{h}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}) f(\boldsymbol{\alpha}^*) \\ &\propto f(\mathbf{z}|\mathbf{h}) f(\mathbf{h}) \end{aligned}$$

Die durch sie definierte Verteilung entspricht somit der Posteriori-Verteilung von $\mathbf{h}$ gegeben die unabhängigen Beobachtungen z_{ij}. Aufgrund von (5.12) und der Konjugiertheitseingenschaft (3.19) des TPT, können die $\mathbf{h}|\cdot$ einfach aus dem auf die Beobachtungen z_{ij} bedingten Posteriori-TPT gezogen werden (Hanson, 2006). Hierfür werden zunächst die Parameter des Posteriori-TPT anhand von (3.20) berechnet. Anschließend werden die $\mathbf{h}$ gemäß (3.9) - (3.12) gezogen.

Item- und Personenparameter und Hilfsparameter

Die vollständig bedingte Dichte der übrigen Parameter kann wie folgt faktorisiert werden (Hanson, 2006):

$$f(\mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*|\mathbf{h}, \mathbf{x}) = f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}, \mathbf{x}) f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*|\mathbf{h}, \mathbf{x})$$

Um aus der durch sie definierten vollständig bedingten Verteilung zu ziehen, können also in einem ersten Schritt die Parameter $\boldsymbol{\xi}$, $\boldsymbol{\beta}$ und $\boldsymbol{\alpha}^*$ aus der durch den zweiten Faktor

$$f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*|\mathbf{h}, \mathbf{x}) \tag{5.20}$$

definierten Verteilung gezogen werden. In einem zweiten Schritt können bedingt auf diese Parameter die Hilfsvariablen $\mathbf{z}$ aus der durch den ersten Term

$$f(\mathbf{z}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}, \mathbf{x}) \tag{5.21}$$

definierten Verteilung gezogen werden.

Die Funktion in (5.20) ist identisch zur vollständig bedingten Dichte der Parameter $\boldsymbol{\xi}$, $\boldsymbol{\beta}$ und $\boldsymbol{\alpha}^*$ basierend auf der Posteriori-Dichte in (5.19). Sie ist somit proportional zu dieser und hat die Form

$$f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*|\mathbf{h}, \mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) f(\mathbf{h}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}) f(\boldsymbol{\alpha}^*) \tag{5.22}$$

$$\propto f(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \mathbf{h}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}) f(\boldsymbol{\alpha}^*). \tag{5.23}$$

Allerdings hat diese vollständig bedingte Dichte aufgrund der durch den TPT vorgegebenen komplizierten Form von $F_{\mathbf{h}}$ in der Likelihood (5.15) keine bekannte Form. Um Zufallszahlen aus ihr zu ziehen, muss also ein Metropolisschritt in den Gibbs Sampler eingebaut werden (Hanson, 2006). Hierfür werden neue Parameterwerte $\boldsymbol{\xi}_{neu}$, $\boldsymbol{\beta}_{neu}$ und $\boldsymbol{\alpha}^*_{neu}$ aus unabhängigen Normalverteilungen mit den Werten der einzelnen Komponenten aus der letzten Iteration jeweils als Erwartungswert gezogen. Die neuen Werte werden mit Wahrscheinlichkeit

$$\rho = min \left\{ \frac{f(\mathbf{x}|\boldsymbol{\xi}_{neu}, \boldsymbol{\beta}_{neu}, \boldsymbol{\alpha}^*_{neu}, \mathbf{h}) f(\boldsymbol{\xi}_{neu}) f(\boldsymbol{\beta}_{neu}) f(\boldsymbol{\alpha}^*_{neu})}{f(\mathbf{x}|\boldsymbol{\xi}_{alt}, \boldsymbol{\beta}_{alt}, \boldsymbol{\alpha}^*_{alt}, \mathbf{h}) f(\boldsymbol{\xi}_{alt}) f(\boldsymbol{\beta}_{alt}) f(\boldsymbol{\alpha}^*_{alt})}, 1 \right\} \tag{5.24}$$

akzeptiert. Werden die vorgeschlagenen Werte abgelehnt, werden die alten Werte von $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}^*$ und $\mathbf{z}$ bis zur nächsten Iteration des Gibbs Samplers beibehalten.

In dem Fall, dass die neuen Werte für $\boldsymbol{\xi}$, $\boldsymbol{\beta}$ und $\boldsymbol{\alpha}^*$ akzeptiert werden, kann $\mathbf{z}|\cdot$ aus der durch (5.21) definierten Verteilung gezogen werden. Die Dichte in (5.21) ist die vollständig bedingte Dichte von $\mathbf{z}$ basierend auf der Posteriori-Dichte (5.18). Für sie gilt daher:

$$\begin{aligned}
f(\mathbf{z}|\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha}^*,\mathbf{h},\mathbf{x}) &\propto f(\mathbf{x}|\mathbf{z},\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha}^*)f(\mathbf{z}|\mathbf{h})f(\mathbf{h})f(\boldsymbol{\xi})f(\boldsymbol{\beta})f(\boldsymbol{\alpha}^*) \\
&\propto f(\mathbf{x}|\mathbf{z},\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha}^*)f(\mathbf{z}|\mathbf{h}) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{k}\left[I(x_{ij}=1)I(z_{ij}\leq\eta_{ij})+I(x_{ij}=0)I(z_{ij}>\eta_{ij})\right]\prod_{i=1}^{n}\prod_{j=1}^{k}f_{\mathbf{h}}(z_{ij}) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{k}f_{\mathbf{h}}(z_{ij})\left[I(x_{ij}=1)I(z_{ij}\leq\eta_{ij})+I(x_{ij}=0)I(z_{ij}>\eta_{ij})\right] \\
&= \prod_{i=1}^{n}\prod_{j=1}^{k}\begin{cases} f_{\mathbf{h}}(z_{ij})I(z_{ij}\leq\eta_{ij}) & \text{falls } x_{ij}=1 \\ f_{\mathbf{h}}(z_{ij})I(z_{ij}>\eta_{ij}) & \text{falls } x_{ij}=0 \end{cases}
\end{aligned}$$

wobei sich der dritte Schritt aus (5.16) und (5.17) ergibt (Hanson, 2006). Die Hilfsvariablen sind somit bedingt auf alle anderen Parameter und die Daten unabhängig und folgen für alle Beobachtungen mit $x_{ij} = 1$ einer rechtsseitig in η_{ij} trunkierten Verteilung mit Dichte $f_{\mathbf{h}}(z_{ij})$ und für alle Beobachtungen mit $x_{ij} = 0$ einer linksseitig in η_{ij} trunkierten Verteilung mit Dichte $f_{\mathbf{h}}(z_{ij})$. Da die zu $f_{\mathbf{h}}(z_{ij})$ gehörige inverse Verteilungsfunktion eine durch (3.18) definierte Form hat, kann $\mathbf{z}|\cdot$ leicht mithilfe der Inversionsmethode gezogen werden (Hanson, 2006).

5.2.7 Gibbs Sampler

Dies definiert einen Metropolis-within-Gibbs Sampler, in dem iterativ aus den folgenden Verteilungen gezogen wird:

$$\begin{aligned}
(\mathbf{h}|\cdot) \\
(\mathbf{z},\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha}^*|\cdot)
\end{aligned}$$

Dieser Gibbs Sampler soll im Folgenden als TPT Gibbs Sampler bezeichnet werden.

5.3 HMC Sampler für ein NMIRT Modell mit Dirichlet Prozess Mixture

Das in Abschnitt 5.1 mithilfe einer TDPM formulierte Modell kann so angepasst werden, dass auf seine Posteriori-Verteilung ein HMC Algorithmus angewandt werden kann. Dies ist in den bisherigen Modellen nicht möglich, da in dem in Abschnitt 5.1 vorgestellten Modell die partiellen Ableitungen der Likelihood nicht analytisch hergeleitet werden können und in dem Modell aus Abschnitt 5.2 die Likelihood nicht differenziertbar ist. Der hieraus

resulierende HMC Sampler könnte sich in manchen Situationen als effizienter als die beiden Gibbs Sampler erweisen.

Ausgangspunkt ist die Parametrisierung in Modell 2 mit frei schätzbarer ICC und positiven Ladungsparametern, in welcher der Schwierigkeitsparameter des ersten Items β_1 auf 0 und die Ladung des ersten Items auf den ersten Faktor α_{11} auf 1 gesetzt ist. Die Restriktionen für die Personenparameter sind: $\boldsymbol{\xi}_1 = (1, 0, \ldots, 0)^\top$, $\boldsymbol{\xi}_2 = (0, 1, 0, \ldots, 0)^\top$, $\ldots$, $\boldsymbol{\xi}_Q = (0, \ldots, 0, 1)^\top$ und $\boldsymbol{\xi}_{Q+1} = (0, \ldots, 0)^\top$.

5.3.1 Reformulierung des Modells

Zunächst seien die positiven Ladungsparameter wieder wie in (5.11) für jedes Item j und jede Dimension q logarithmiert:

$$\alpha^*_{jq} := \ln(\alpha_{jq})$$

Es folgt

$$\eta_{ij} = \beta_j + \sum_{q=1}^{Q} \exp(\alpha^*_{jq})\xi_{iq}$$

und $\alpha^*_{11} = \ln(\alpha_{11}) = \ln(1) = 0$. Die hierarchische Formulierung des Modells durch (5.2) - (5.4) unter Verwendung einer TDP Mean Mixture von Normalverteilungen führte zu Realisationen der ICC mit der Form (5.7):

$$F(x) = \sum_{l=1}^{L} \pi_l \Phi(x; \delta_l, \sigma^2)$$

Anstelle der Verteilungsfunktion Φ der Normalverteilung mit Varianzparameter σ^2 kann als Verteilung der Mischkomponenten im TDPM jedoch auch die logistische Verteilungsfunktion mit allgemeinem Diskriminationsparameter $exp(\gamma)$, $\gamma \in \mathbb{R}$, gewählt werden:

$$F(x) = \sum_{l=1}^{L} \pi_l \frac{\exp(\delta_l + \exp(\gamma)x)}{1 + \exp(\delta_l + \exp(\gamma)x)} \tag{5.25}$$

Die Gewichte π_l sind hierbei wieder durch die $L-1$ unabhängigen, betaverteilten Variablen B_l aus der Stick-Breaking Darstellung des TDP determiniert:

$$\pi_l = B_l \prod_{i=1}^{l-1}(1 - B_i) \tag{5.26}$$

für $l = 1, 2, \ldots, L-1$ und

$$\pi_L = 1 - \sum_{l=1}^{L-1} \pi_l. \tag{5.27}$$

Werden diese Parameter in einem Vektor $\mathbf{b}$ zusammengefasst, resultiert ein nonparametrisches bayesianisches TDPM Modell, das von den Parametern $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}^*$, γ, $\boldsymbol{\delta}$ und $\mathbf{b}$ abhängt.

5.3.2 Likelihood

Dieses Modell besitzt wegen (1.2) und (5.25) eine Likelihood der Form

$$\begin{aligned}
f(\mathbf{x}|\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha}^*,\gamma,\boldsymbol{\delta},\mathbf{b}) =& \prod_{i=1}^{n}\prod_{j=1}^{k} F(\eta_{ij})^{x_{ij}}(1-F(\eta_{ij}))^{1-x_{ij}} \\
=& \prod_{i=1}^{n}\prod_{j=1}^{k}\left(\sum_{l=1}^{L}\pi_l \frac{exp(\delta_l+\exp(\gamma)\eta_{ij})}{1+exp(\delta_l+\exp(\gamma)\eta_{ij})}\right)^{x_{ij}} \\
& \prod_{i=1}^{n}\prod_{j=1}^{k}\left(1-\sum_{l=1}^{L}\pi_l \frac{exp(\delta_l+\exp(\gamma)\eta_{ij})}{1+exp(\delta_l+\exp(\gamma)\eta_{ij})}\right)^{1-x_{ij}} \\
=& \prod_{i=1}^{n}\prod_{j=1}^{k}\left(\sum_{l=1}^{L}\pi_l \frac{exp(\delta_l+\exp(\gamma)(\beta_j+\boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i))}{1+exp(\delta_l+\exp(\gamma)(\beta_j+\boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i))}\right)^{x_{ij}} \\
& \prod_{i=1}^{n}\prod_{j=1}^{k}\left(1-\sum_{l=1}^{L}\pi_l \frac{exp(\delta_l+\exp(\gamma)(\beta_j+\boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i))}{1+exp(\delta_l+\exp(\gamma)(\beta_j+\boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i))}\right)^{1-x_{ij}},
\end{aligned}$$

wobei die Gewichte π_l rekursiv durch (5.26) und (5.27) definiert sind. Die logarithmierte Likelihood hat die Form

$$\begin{aligned}
\ln f(\mathbf{x}|\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\alpha}^*,\gamma,\boldsymbol{\delta},\mathbf{b}) =& \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}\ln\left(\sum_{l=1}^{L}\pi_l \frac{exp(\delta_l+\exp(\gamma)(\beta_j+\boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i))}{1+exp(\delta_l+\exp(\gamma)(\beta_j+\boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i))}\right) + \\
& \sum_{i=1}^{n}\sum_{j=1}^{k}(1-x_{ij})\ln\left(1-\sum_{l=1}^{L}\pi_l \frac{exp(\delta_l+\exp(\gamma)(\beta_j+\boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i))}{1+exp(\delta_l+\exp(\gamma)(\beta_j+\boldsymbol{\alpha}_j^{*\top}\boldsymbol{\xi}_i))}\right).
\end{aligned}$$

Da in der logarithmierten Likelihood sowie in (5.26) und (5.27) nur Additionen, Multiplikationen, Divisionen und die Exponentialfunktion vorkommen, ist sie bezüglich aller Parameter differenzierbar und die partiellen Ableitungen sind unter Verwendung der Kettenregel in geschlossener Form darstellbar.

5.3.3 Priori-Verteilungen

Die Variablen in $\mathbf{b}$ folgen aufgrund von (3.24) in der Stick-Breaking Darstellung des TDP unabhängigen Betaverteilungen mit Parametern 1 und λ. Als Priori-Dichte ergibt sich somit

$$f(\mathbf{b}) = \prod_{l=1}^{L-1} Beta(b_l;1,\lambda) \propto \prod_{l=1}^{L-1} b_l^{1-1}(1-b_l)^{\lambda-1} = \prod_{l=1}^{L-1}(1-b_l)^{\lambda-1}.$$

Die logarithmierte Priori-Dichte ist also

$$\ln f(\mathbf{b}) = c + \sum_{l=1}^{L-1}(\lambda-1)\ln(1-b_l),$$

wobei $c \in \mathbb{R}$ eine Konstante ist. Die partiellen Ableitungen dieser Funktion können mithilfe der Kettenregel für alle b_l bestimmt werden.

Alle anderen Priori-Verteilungen werden unabhängig voneinander und von $\mathbf{b}$ gewählt:

$$f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \gamma, \boldsymbol{\delta}, \mathbf{b}) = f(\boldsymbol{\xi}) f(\boldsymbol{\beta}) f(\boldsymbol{\alpha}^*) f(\gamma) f(\boldsymbol{\delta}) f(\mathbf{b})$$

Für alle unrestringierten Komponenten können hierbei unabhängige Normalverteilungen mit Erwartungswert 0 angenommen werden, so dass die logarithmierten Priori-Dichten Summen von quadratischen Funktionen in den Parametern sind. Die partiellen Ableitungen können dann leicht berechnet werden.

5.3.4 Hyperparameter

Als Hyperparameter des Modells müssen festgelegt werden:

- die Varianz σ_ξ^2 der Priori-Verteilungen der Personenparameter,
- die Varianz $\sigma_{\beta\alpha^*}^2$ der Priori-Verteilungen der Itemparameter,
- die Varianz σ_γ^2 des Parameters γ, der den allgemeinen Diskriminationsparameter der logistischen Mischkomponenten bestimmt,
- die Varianz σ_δ^2 der Lokationsparameter δ_l der logistischen Mischkomponenten und
- der Parameter λ des TDP.

Schließlich muss noch ein Wert für die Anzahl L der Mischkomponenten festgelegt werden.

5.3.5 Posteriori-Verteilung

Für die Dichte der Posteriori-Verteilung gilt

$$\begin{aligned} f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \gamma, \boldsymbol{\delta}, \mathbf{b} | \mathbf{x}) &\propto f(\mathbf{x} | \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \gamma, \boldsymbol{\delta}, \mathbf{b}) f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \gamma, \boldsymbol{\delta}, \mathbf{b}) \\ &= f(\mathbf{x} | \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \gamma, \boldsymbol{\delta}, \mathbf{b}) f(\boldsymbol{\xi}) f(\boldsymbol{\beta}) f(\boldsymbol{\alpha}^*) f(\gamma) f(\boldsymbol{\delta}) f(\mathbf{b}). \end{aligned}$$

Damit folgt für die logarithmierte Posteriori-Dichte

$$\begin{aligned} &\ln f(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \gamma, \boldsymbol{\delta}, \mathbf{b} | \mathbf{x}) \\ &= \ln f(\mathbf{x} | \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\alpha}^*, \gamma, \boldsymbol{\delta}, \mathbf{b}) \\ &+ \ln f(\boldsymbol{\xi}) + \ln f(\boldsymbol{\beta}) + \ln f(\boldsymbol{\alpha}^*) + \ln f(\gamma) + \ln f(\boldsymbol{\delta}) + \ln f(\mathbf{b}) + c, \end{aligned}$$

wobei $c \in \mathbb{R}$ eine Konstante ist.

5.3.6 HMC Sampler

Da sowohl die logarithmierte Likelihood als auch die logarithmierten Priori-Dichten bezüglich aller Parameter differenzierbar sind, gilt dies auch für ihre Summe. Somit kann der Gradient der logarithmierten Posteriori-Dichte berechnet und der in Abschnitt 4.3 beschriebene HMC Algorithmus angewandt werden. Die Berechnung des Gradienten und die Reparametrisierung der auf das Intervall $[0, 1]$ beschränkten betaverteilten Parameter in $\mathbf{b}$ kann automatisch mithilfe geeigneter Software zur automatischen Differenzierung erfolgen. Der hieraus resultierende Algorithmus soll als TDPM HMC Sampler bezeichnet werden.

5.4 Vergleich der Algorithmen

Alle drei in diesem Kapitel vorgeschlagenen MCMC Algorithmen können zur Schätzung der Parameter des NMIRT Modells eingesetzt werden. Sie basieren lediglich auf unterschiedlichen Priori-Verteilungen für die ICC und unterschiedlichen Parametrisierungen. Für die praktische Anwendung ist es von Vorteil, wenn mehrere MCMC Algorithmen zur Schätzung des gleichen Modells zur Verfügung stehen, da nicht davon ausgegangen werden kann, dass jeder Algorithmus in jedem Datensatz konvergiert.

Im TDPM Gibbs Sampler sind alle vollständig bedingten Verteilungen bekannt. Dies führt in der Regel zu einer effektiven Exploration des Parameterraums, da die Markovkette in jeder Iteration ihren Zustand ändert. Als problematisch könnte sich aber die durch die TDPM bedingte Mischverteilungsstruktur der Posteriori-Verteilung erweisen. Der TPT Gibbs Sampler vermeidet diese Mischverteilungsstruktur durch die Wahl eines TPT als Priori-Verteilung. Allerdings haben hier nicht alle vollständig bedingten Verteilungen eine bekannte Form, so dass ein Metropolisschritt für die Item- und Personenparameter nötig wird. Dies kann dazu führen, dass die Markovkette in bestimmten Regionen des Parameterraums hängen bleibt. Der TDPM HMC Sampler nutzt die im Gradienten der logarithmierten Posteriori-Dichte enthaltene Information zur Konstruktion effizienter Vorschlagsverteilungen mit hohen Akzeptanzwahrscheinlichkeiten. Die durch ihn generierten Markovketten dürften in den meisten Fällen eine geringere Autokorrelation als die Ketten des TDPM Gibbs Samplers aufweisen. Wie bei diesem kann jedoch die Mischverteilungsstruktur der Posteriori-Verteilung zu Konvergenzproblemen führen. Da die Evaluation des Gradienten für hochdimensionale Parametervektoren zudem sehr viel Zeit kostet, ist eine hohe Laufzeit zu erwarten.

Kapitel 6

Anwendungsbeispiel

Um ihre praktische Anwendbarkeit zu untersuchen, wurden die drei in Kapitel 5 vorgeschlagenen Schätzalgorithmen auf einen simulierten Datensatz angewandt. Hierfür wurden Antworten von 500 Personen auf zehn Items aus einem NMIRT Modell mit $Q = 2$ Faktoren zufällig in R (R Core Team, 2018) generiert. Als ICC diente ähnlich wie in Miyazaki und Hoshino (2009) eine Mischung aus den Verteilungsfunktionen zweier Normalverteilungen mit Erwartungswerten von jeweils 0.5 und -2, Mischverteilungsgewichten von jeweils 0.7 und 0.3 und einer Standardabweichung von 0.7 für beide Komponenten:

$$F(x) = 0.7\Phi(x; 0.5, 0.7^2) + 0.3\Phi(x; -2, 0.7^2)$$

Die ICC ist in Abbildung 6.1 dargestellt. Die Schwierigkeitsparameter β_j der zehn Items wurden in gleichen Abständen von -2 bis 2 aufsteigend gewählt, wobei β_1 den Restriktionen in den Modellen entsprechend auf Null gesetzt wurde. Die Ladungsparameter in $\boldsymbol{\alpha}_j$ wurden für beide Faktoren zunächst aus einer Gleichverteilung auf dem Intervall $[0.2, 1.5]$ gezogen und anschließend für den ersten Faktor absteigend und den zweiten Faktor aufsteigend sortiert. Die Ladung α_{11} des ersten Items auf den ersten Faktor wurde auf 1 fixiert. Die Personenparameter in $\boldsymbol{\xi}_i$ wurden für jede Person aus unabhängigen Standardnormalverteilungen gezogen.

Die beiden Gibbs Sampler wurden in R unter Verwendung der Pakete FastGP (Gopalan & Bornn, 2016), invgamma (Kahle & Stamey, 2017) und truncnorm (Mersmann, Trautmann, Steuer & Bornkamp, 2014) geschrieben. Die Visualisierung der Ergebnisse erfolgte mithilfe des Pakets ggplot2 (Wickham, 2016), das auch für die Darstellungen in Kapitel 3 verwendet wurde. Das logistische TDPM Modell wurde in Stan (Stan Development Team, 2017) geschrieben, so dass auf die hier implementierte automatische Differenzierung der logarithmierten Posteriori-Dichte und den automatischen HMC Algorithmus mit Tuning der MCMC Parameter (Hoffman & Gelman, 2014) zurückgegriffen werden konnte. Als Interface diente das R Paket rstan (Stan Development Team, 2018). Die Laufzeit der Algorithmen betrug auf einem Macbook Pro (13 Zoll, Mitte 2012, 2.9 GHz, 8 GB Arbeitsspeicher) jeweils ungefähr zehn Stunden.

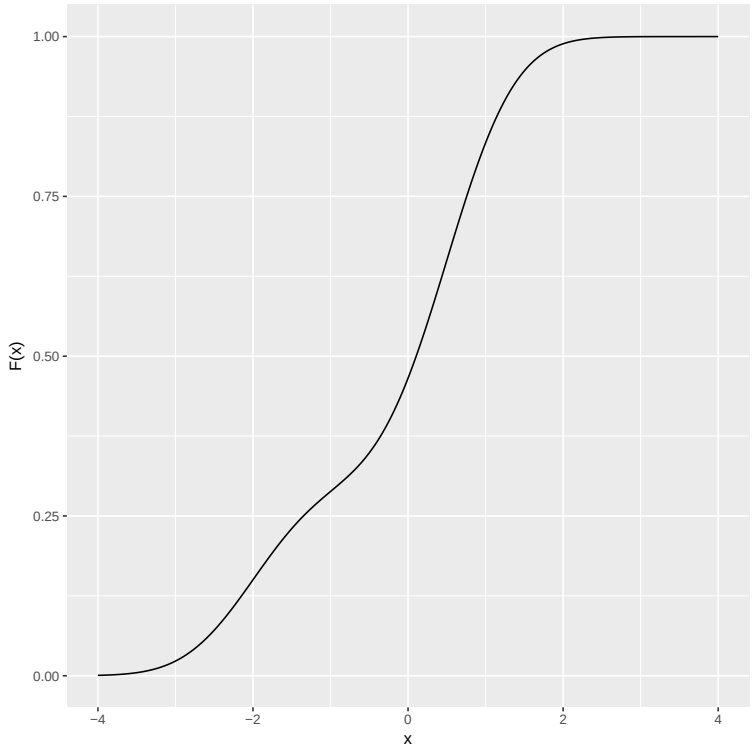

Abbildung 6.1: Wahre ICC

6.1 TDPM Gibbs Sampler

6.1.1 Hyperparameter

Der Präzisionsparameter wurde auf $\lambda = 10$ gesetzt, da dieser Wert in Testläufen zu einer relativ schnellen Konvergenz der Markovketten führte. In einer Stichprobe mit $nk = 5000$ Beobachtungen führt $\lambda = 10$ zudem noch immer zu einer wenig informativen Priori-Verteilung für die ICC. Für den TDP erwies sich ein Trunkierungslevel von $L = 5$ als ausreichend, da die Schätzwerte für die Gewichte weiterer Mischungskomponenten in Testläufen gegen Null gingen. Als Basisverteilung des TDP diente die Standardnormalverteilung. Die punktweise Priori-Erwartungswertfunktion der ICC entsprach somit der ICC der Probit Modelle. Die Parameter der inversen Gammaverteilung für die Varianz σ^2 der Mischkomponenten wurden auf die Werte $\alpha_{\sigma^2} = 3$ und $\beta_{\sigma^2} = 1$ gesetzt. Für alle freien Item- und Personenparameter wurde wie in Béguin und Glas (2001) eine Standardnormalveteilung als Priori-Verteilung gewählt. Die Anzahl der Faktoren wurde als bekannt vorausgesetzt.

6.1.2 Startwerte

Als Startwert für die Mischverteilungsgewichte π_l wurde für alle fünf Komponenten der Wert 0.2 gewählt. Die Varianz σ^2 der Mischungskomponenten zu Beginn der Kette wurde

auf 1 gesetzt. Startwerte für die Clustervariablen $\mathbf{c}$ wurden für jede Beobachtung aus einer diskreten Gleichverteilung auf der Menge $\{1, 2, 3, 4, 5\}$ gezogen. Die Startwerte für die Schwierigkeitsparameter β_j wurden zufällig aus ihrer Priori-Verteilung gezogen. Für alle übrigen freien Item- und Personenparameter und für die Lokationsparameter der Mischkomponenten in $\boldsymbol{\delta}$ wurde unabhängig aus Normalverteilungen mit Erwartungswert 0 und Standardabweichung 0.001 gezogen. Die Startwerte für die latenten Hilfsvariablen in $\mathbf{z}$ wurden schließlich gemäß (5.9) für jede Beobachtung auf den Wert des sich aus den anderen Startwerten ergebenden negativen linearen Prädiktors $-\eta_{ij}$ gesetzt.

6.1.3 Konvergenz

Da die Markovketten einiger Parameter in Testläufen hohe Autokorrelationen aufwiesen, wurde die Gesamtlänge der simulierten Kette auf 500000 Iterationen festgesetzt. In Abbildung 6.2 sind die Verläufe der Markovketten - die sogenannten Traceplots - für die Parameter des zweiten Items, die Parameter der ersten Mischungskomponente der TDPM und den Varianzparameter σ^2 beispielhaft dargestellt. Die Markovketten der Gewichte π_l und der Lokationsparameter δ_l weisen das für Mischverteilungsmodelle typische Label-Switching auf (Stephens, 2000): Sie springen zwischen Permutationen der Labels l umher. Dies ist jedoch unter Voraussetzung der Konvergenz der Ketten der übrigen Parameter unproblematisch, da keine Inferenz bezüglich der einzelnen Mischungskomponenten, sondern nur bezüglich der durch sie definierten ICC gewünscht ist (Stephens, 2000). Die Traceplots für die übrigen Itemparameter und Mischkomponenten befinden sich in Anhang A.1. Obwohl die Markovketten vor allem der Ladungsparameter eine hohe Autokorrelation zeigen, kann insgesamt von der Konvergenz der Ketten ausgegangen werden: Wie in Abbildung 6.2 zu sehen ist, explorieren die Ketten den Parameterraum gleichmäßig und bleiben zu keinem Zeitpunkt in Teilregionen hängen.

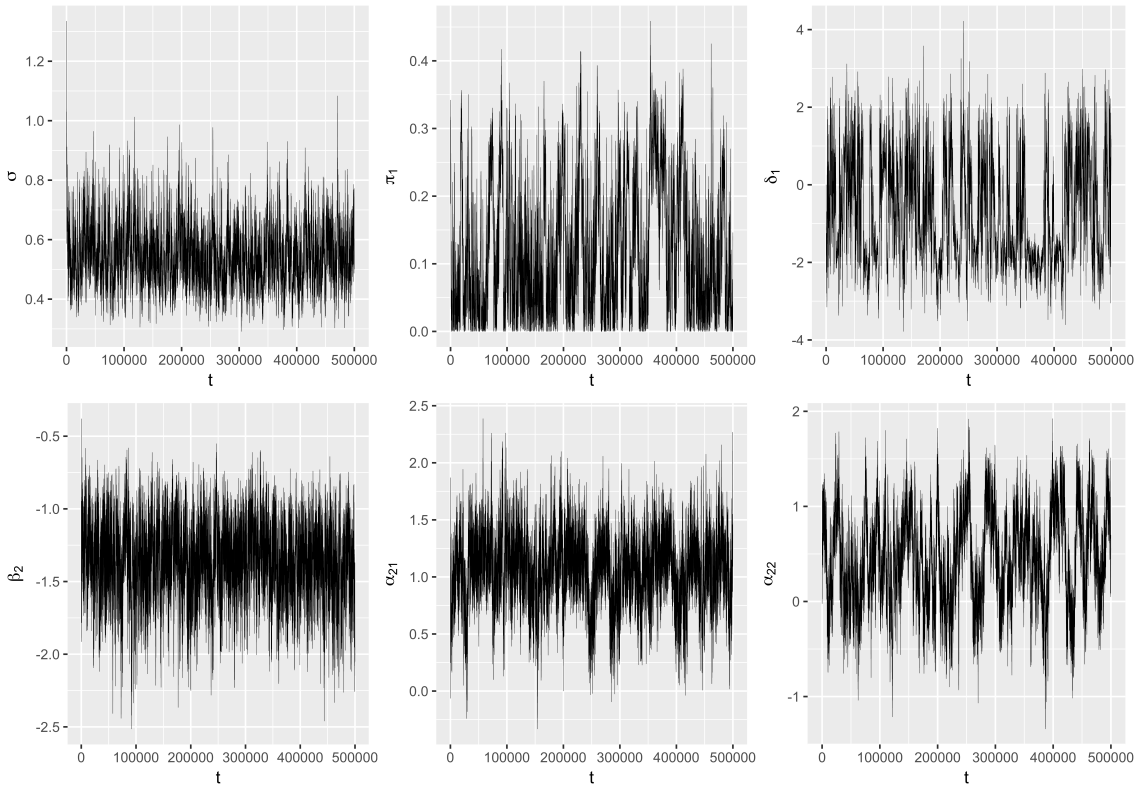

Abbildung 6.2: Ausgewählte Traceplots aus dem TDPM Gibbs Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

Tabelle 6.1: Schätzwerte für die Itemparameter aus dem TDPM Gibbs Sampler.

Item	β_j	$\hat{\beta}_j$	$SD(\hat{\beta}_j)$	α_{j1}	$\hat{\alpha}_{j1}$	$SD(\hat{\alpha}_{j1})$	α_{j2}	$\hat{\alpha}_{j2}$	$SD(\hat{\alpha}_{j2})$
1	0.00	0.00	0.00	1.00	1.00	0.00	0.24	0.14	0.47
2	-1.56	-1.36	0.24	1.16	1.06	0.30	0.44	0.46	0.47
3	-1.11	-0.95	0.20	1.11	0.94	0.34	0.53	0.53	0.43
4	-0.67	-0.61	0.17	1.00	0.98	0.31	0.58	0.52	0.43
5	-0.22	-0.01	0.13	0.98	0.71	0.28	0.69	0.42	0.33
6	0.22	0.24	0.12	0.67	0.53	0.38	0.73	0.81	0.29
7	0.67	0.74	0.14	0.45	0.26	0.47	1.07	1.09	0.29
8	1.11	1.09	0.16	0.35	0.34	0.39	1.15	0.90	0.27
9	1.56	1.68	0.28	0.28	0.42	0.59	1.27	1.32	0.38
10	2.00	1.79	0.30	0.23	0.36	0.46	1.42	1.06	0.33

Anmerkungen. Wahre und mithilfe des TDPM Gibbs Samplers geschätzte Itemparameter. β_j = wahrer Schwierigkeitsparameter des Items j, $\hat{\beta}_j$ = EAP des Schwierigkeitsparameters des Items j, $SD(\hat{\beta}_j)$ = Posteriori-Standardabweichung des Schwierigkeitsparameters des Items j, α_{j1} = wahre Ladung des j-ten Items auf Faktor 1, $\hat{\alpha}_{j1}$ = EAP der Ladung des j-ten Items auf Faktor 1, $SD(\hat{\alpha}_{j1})$ = Posteriori-Standardabweichung der Ladung des j-ten Items auf Faktor 1, α_{j2} = wahre Ladung des j-ten Items auf Faktor 2, $\hat{\alpha}_{j2}$ = EAP der Ladung des j-ten Items auf Faktor 2, $SD(\hat{\alpha}_{j2})$ = Posteriori-Standardabweichung der Ladung des j-ten Items auf Faktor 2.

6.1.4 Schätzwerte

Nach Inspektion der Traceplots wurde eine Burnin-Periode von 20000 Iterationen gewählt. Die wahren Itemparameter, die EAP-Schätzwerte und ihre Posteriori-Standardabweichung sind in Tabelle 6.1 aufgeführt. In Abbildung 6.3 sind 100 zufällige Ziehungen aus der Posteriori-Verteilung der ICC, die punktweise Posteriori-Erwartungswertfunktion und die wahre ICC dargestellt.

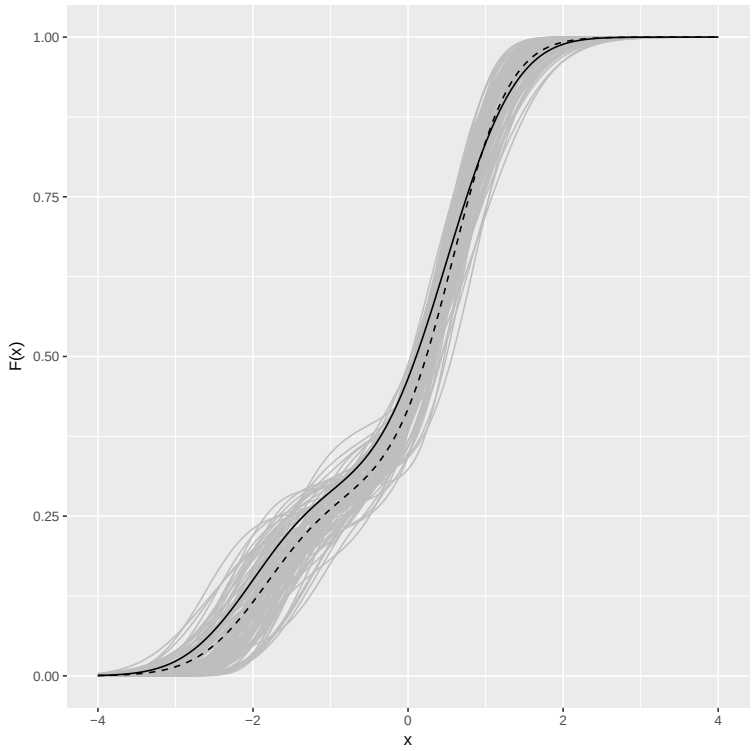

Abbildung 6.3: 100 Ziehungen aus der Posteriori-Verteilung der ICC aus dem TDPM Gibbs Sampler. Die durchgezogene schwarze Funktion ist die wahre ICC. Die gestrichelte schwarze Funktion ist die punktweise Erwartungswertfunktion der ICC unter ihrer Posteriori-Verteilung. Die grauen Funktionen sind 100 Ziehungen aus der Posteriori-Verteilung der ICC.

6.2 TPT Gibbs Sampler

6.2.1 Hyperparameter

Der TPT wurde in der Parametrisierung (3.14) mit Präzisionsparameter λ spezifiziert und entsprechend der Daumenregel von Hanson (2006) auf der Ebene $L = \log_2(nk) = \log_2(5000) \approx 12$ trunkiert. Für kleine Werte des Präzisionsparameters λ zeigte sich in Testläufen eine sehr schnelle Divergenz der Markovketten, so dass $\lambda = 10$ gesetzt wurde. Als Basisverteilung wurde eine Standardnormalverteilung gewählt. Für die Schwierigkeitsparameter β_j, die logarithmierten Ladungsparameter $\boldsymbol{\alpha}_j^*$ und alle Personenparameter $\boldsymbol{\theta}_i$ wurden unabhängige Standardnormalverteilungen als Priori-Verteilungen angenommen. Die Anzahl der Faktoren wurde als bekannt vorausgesetzt.

6.2.2 Startwerte und Varianz der Vorschlagsverteilungen

Mit zufälligen Startwerten zeigte sich in Testläufen ein starkes Divergenzverhalten des TPT Gibbs Samplers. Aus diesem Grund wurde zunächst in Stan ein parametrisches bayesianisches mehrdimensionales Modell (1.10) mit identischer Parametrisierung und identischen Priori-Verteilungen geschätzt. Die Schätzwerte aus diesem Modell wurden dann als Startwerte für die Item- und Personenparameter im TPT Gibbs Sampler verwendet. Die Startwerte für die latenten Hilfsvariablen Z_{ij} wurden jeweils auf den linearen Prädiktor η_{ij} gesetzt, der aus den Startwerten für die Item- und Personenparameter berechnet wurde. Die Startwerte für die betaverteilten Variablen des TPT wurden aus ihrer bedingten Verteilung gegeben die Startwerte der anderen Parameter gezogen. Zudem wurden die Varianzen der Vorschlagsverteilungen für die Item- und Personenparameter proportional zu den Posteriori-Varianzen der entsprechenden Parameter aus dem parametrischen Modell gewählt (Hanson, 2006). Diese Varianzen wurden schließlich mit dem Wert 0.01 multipliziert, da sich für die hieraus resultierenden Vorschlagsvarianzen in Testläufen eine Akzeptanzrate nahe des von Robert und Casella (2004) für hochdimensionale Modelle empfohlenen Werts von 0.25 für den Metropolisschritt ergab.

6.2.3 Konvergenz

Die Traceplots einiger ausgewählter (nicht logarithmierter) Ladungsparameter sind in Abbildung 6.4 dargestellt. Auch nach einer sehr hohen Anzahl von 300000 Iterationen kann nicht von der Konvergenz des TPT Gibbs Samplers ausgegangen werden. Vielmehr bleiben die Markovketten immer wieder für sehr lange Zeit in Regionen nahe Null hängen. Für die restlichen Ladungsparameter ergibt sich ein ähnliches Bild. Deren Traceplots befinden sich in Anhang A.2.

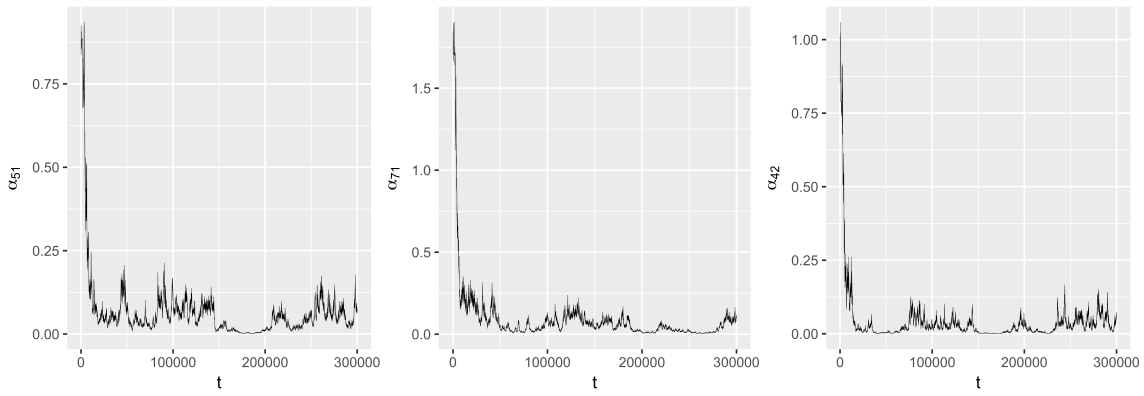

Abbildung 6.4: Ausgewählte Traceplots aus dem TPT Gibbs Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

6.3 TDPM HMC Sampler

6.3.1 Hyperparameter

Um Vergleichbarkeit mit dem TDPM Gibbs Sampler zu gewährleisten, wurden die Anzahl der Mischkomponenten auf $L = 5$ und der Präzisionsparameter des TDP auf $\lambda = 10$ gesetzt. Für den Parameter γ, welcher den allgemeinen Diskriminationsparameter $\exp(\gamma)$ bestimmt, und die Lokationsparameter δ_l der Mischkomponenten wurden unabhängige Standardnormalverteilungen als Priori-Verteilungen gewählt. Da das Label-Switching der Mischkomponenten in einigen Testläufen zu Konvergenzproblemen führte, wurde zudem eine Ordnungsrestriktion auf die Lokationsparameter der Mischkomponenten gelegt (Stephens, 2000). Die Priori-Verteilungen für die Item- und Personenparameter wurden analog zum TPT Gibbs Sampler gewählt. Die Anzahl der Faktoren wurde als bekannt vorausgesetzt.

6.3.2 Startwerte und HMC Parameter

Die Parameter des HMC Algorithmus werden in Stan automatisch und adaptiv während der Burnin-Phase getuned (Hoffman & Gelman, 2014). Auch die Startwerte werden automatisch generiert. Lediglich der die adaptive Wahl der Schrittweite ϵ bestimmende Parameter δ_{adapt} (Hoffman & Gelman, 2014) musste manuell auf den Wert $\delta_{adapt} = 0.875$ gesetzt werden, da es unter Verwendung der Standardeinstellung $\delta_{adapt} = 0.8$ zu Konvergenzproblemen kam.

6.3.3 Konvergenz

Die durch den TDPM HMC Sampler erzeugten Markovketten weisen eine relativ niedrige Autokorrelation auf, weswegen sich eine Gesamtlänge von 20000 Iterationen als ausreichend erwies. In Abbildung 6.5 sind beispielhaft die Traceplots für die nicht logarithmierten Itemparameter des zweiten Items, für den allgemeinen Diskriminationsparameter $\exp(\gamma)$ und die Parameter der ersten Mischkomponente dargestellt. Die Traceplots für die übrigen Item- und ICC-Parameter finden sich in Anhang A.3. Insgesamt kann von der Konvergenz der Markovketten bei relativ niedriger Autokorrelation ausgegangen werden.

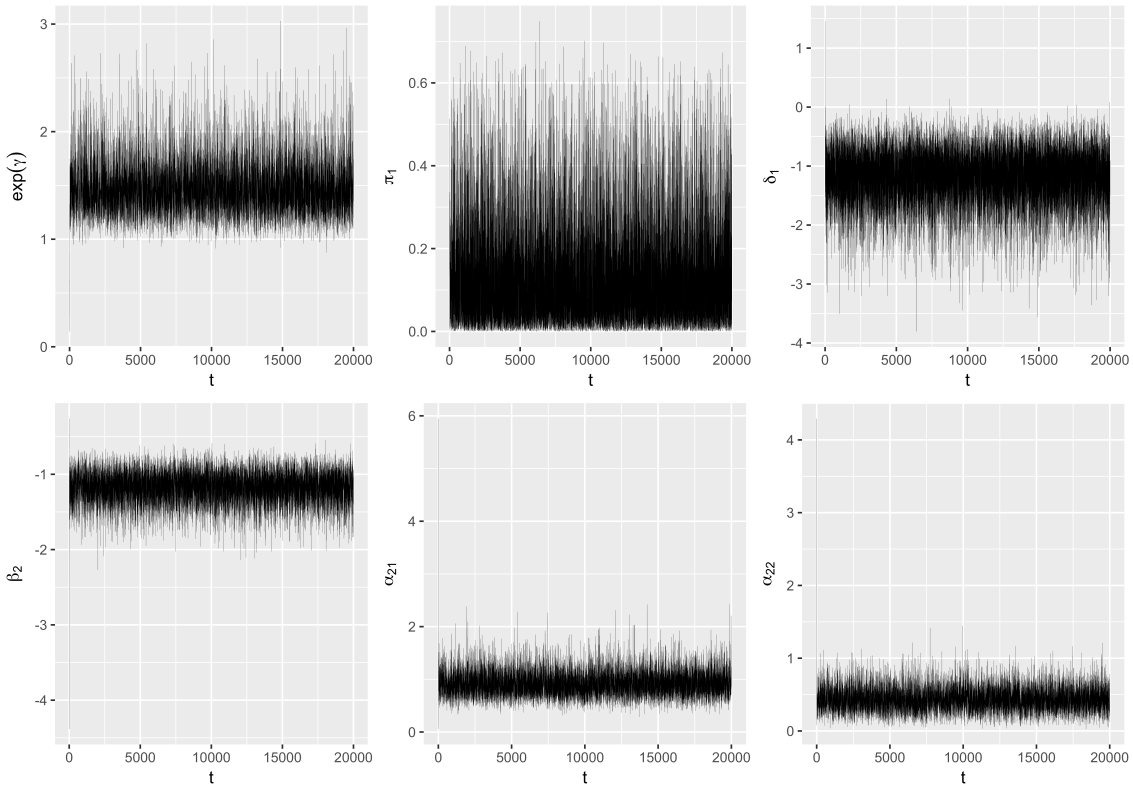

Abbildung 6.5: Ausgewählte Traceplots aus dem TDPM HMC Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

Tabelle 6.2: Schätzwerte für die Itemparameter aus dem TDPM HMC Sampler.

Item	β_j	$\hat{\beta}_j$	$SD(\hat{\beta}_j)$	α_{j1}	$\hat{\alpha}_{j1}$	$SD(\hat{\alpha}_{j1})$	α_{j2}	$\hat{\alpha}_{j2}$	$SD(\hat{\alpha}_{j2})$
1	0.00	0.00	0.00	1.00	1.00	0.00	0.24	0.26	0.12
2	-1.56	-1.18	0.20	1.16	0.92	0.23	0.44	0.43	0.16
3	-1.11	-0.86	0.17	1.11	0.87	0.23	0.53	0.58	0.17
4	-0.67	-0.60	0.15	1.00	0.99	0.26	0.58	0.51	0.16
5	-0.22	-0.07	0.11	0.98	0.78	0.20	0.69	0.54	0.15
6	0.22	0.21	0.12	0.67	0.61	0.18	0.73	0.86	0.19
7	0.67	0.81	0.18	0.45	0.32	0.15	1.07	1.33	0.29
8	1.11	1.33	0.23	0.35	0.42	0.18	1.15	1.34	0.28
9	1.56	2.02	0.33	0.28	0.56	0.24	1.27	1.74	0.35
10	2.00	2.18	0.34	0.23	0.50	0.22	1.42	1.39	0.29

Anmerkungen. Wahre und mithilfe des TDPM HMC Samplers geschätzte Itemparameter. β_j = wahrer Schwierigkeitsparameter des Items j, $\hat{\beta}_j$ = EAP des Schwierigkeitsparameters des Items j, $SD(\hat{\beta}_j)$ = Posteriori-Standardabweichung des Schwierigkeitsparameters des Items j, α_{j1} = wahre Ladung des j-ten Items auf Faktor 1, $\hat{\alpha}_{j1}$ = EAP der Ladung des j-ten Items auf Faktor 1, $SD(\hat{\alpha}_{j1})$ = Posteriori-Standardabweichung der Ladung des j-ten Items auf Faktor 1, α_{j2} = wahre Ladung des j-ten Items auf Faktor 2, $\hat{\alpha}_{j2}$ = EAP der Ladung des j-ten Items auf Faktor 2, $SD(\hat{\alpha}_{j2})$ = Posteriori-Standardabweichung der Ladung des j-ten Items auf Faktor 2.

6.3.4 Schätzwerte

Die ersten 1000 Iterationen wurden als Burnin-Phase betrachtet. Dies ist die Standardeinstellung in Stan und überdies durch die Traceplots gerechtfertigt. Die wahren Itemparameter, die EAP-Schätzwerte und die Posteriori-Standardabweichungen für die Itemparameter sind in Tabelle 6.2 zu sehen. In Abbildung 6.6 sind 100 zufällige Ziehungen aus der Posteriori-Verteilung der ICC, die punktweise Posteriori-Erwartungswertfunktion und die wahre ICC dargestellt.

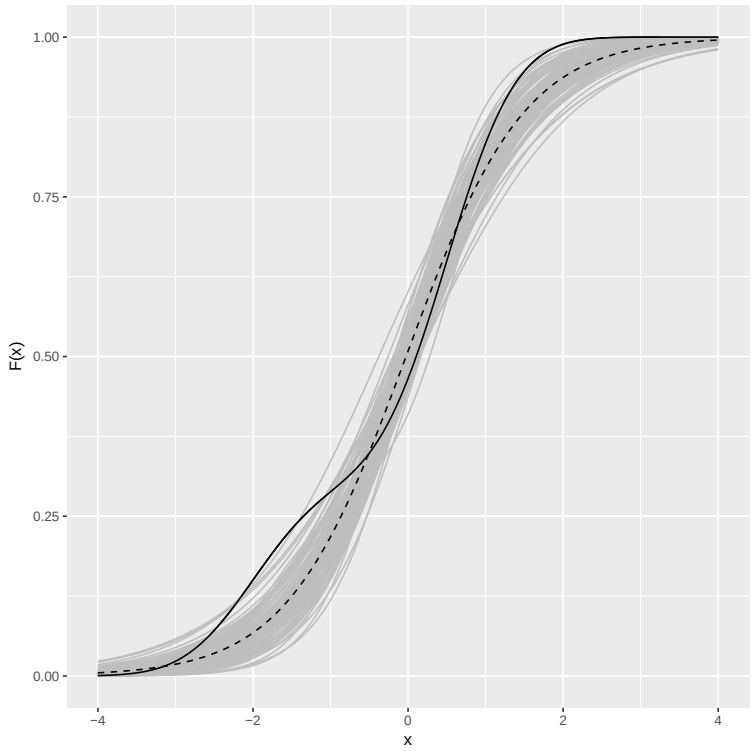

Abbildung 6.6: Ziehungen aus der Posteriori-Verteilung der ICC aus dem TDPM HMC Sampler. Die durchgezogene schwarze Funktion ist die wahre ICC. Die gestrichelte schwarze Funktion ist die punktweise Erwartungswertfunktion der ICC unter ihrer Posteriori-Verteilung. Die grauen Funktionen sind 100 Ziehungen aus der Posteriori-Verteilung der ICC.

6.4 Diskussion der Ergebnisse

Der TPT Gibbs Sampler konvergierte auch nach aufwendigem Tuning nicht, so dass die aus ihm resultierenden Schätzwerte nicht interpretierbar sind. Dieser Umstand kann nicht durch die Parametrisierung oder eine ungünstige Wahl der Priori-Verteilungen erklärt werden, da das zugrundeliegende Modell identisch mit dem des TDPM HMC Samplers ist. Die plausibelste Erklärung für die Konvergenzprobleme ist, dass ein wichtiger und hochdimensionaler Teil des Parametervektors (alle Item- und Personenparameter) im Rahmen des TPT Gibbs Samplers nur sehr ineffizient mithilfe von Metropolissschritten gezogen werden konnte. In der hier simulierten Stichprobe konnte dieser Nachteil nicht durch die im Vergleich zu den TDPM prinzipiell günstigeren Eigenschaften der TPT Verteilung ausgeglichen werden. Eventuell kann der TPT Gibbs Sampler mit noch systematischerem Tuning der Varianzen der Vorschlagsverteilung und der Komplexitätsparameter L und λ zur Konvergenz gebracht werden. Dies würde jedoch die ohnehin schon sehr hohe Rechenzeit deutlich erhöhen. Zudem stellt es einen erheblichen Nachteil dar, wenn die praktische Anwendung ein aufwendiges manuelles Tuning erfordert.

Die durch den TDPM Gibbs Sampler und den TDPM HMC Sampler erzeugten Markovketten konvergierten. Erwartungsgemäß traten beim Gibbs Sampler höhere Autokorrelation auf. Ausgeglichen wurde dies jedoch durch eine um den Faktor 25 geringere Rechendauer des Gibbs Samplers, welche die Generierung deutlich längerer Ketten in gleicher Zeit erlaubte. Die Tabellen 6.1 und 6.2 zeigen, dass die Itemparameter durch den TDPM Gibbs Sampler und den TDPM HMC Algorithmus schon bei 500 Personen relativ gut geschätzt werden konnten. Die absoluten Abweichungen in der hier simulierten Stichprobe sind vergleichbar mit denen der Schätzwerte parametrischer Modelle in Stichproben ähnlicher Größe (Béguin & Glas, 2001). Die im TDPM HMC Sampler geschätzten Posteriori-Standardabweichungen für die Ladungen waren etwas niedriger als die durch den TDPM Gibbs Sampler geschätzten Posteriori-Standardabweichungen. Eine Ursache hierfür könnte darin liegen, dass höhere Momente wie die Varianz langsamer konvergieren als der Mittelwert und die Ketten des TDPM Gibbs Samplers aufgrund ihrer hohen Autokorrelation noch etwas länger laufen müssten, um eine genaue Schätzung der Posteriori-Standardabweichungen zu ermöglichen.

Der TDPM Gibbs Sampler konnte die Form der ICC sehr gut reproduzieren. Dies ist angesichts der kleinen Stichprobe bemerkenswert. Demgegenüber ist die geschätzte ICC im TDPM HMC Sampler relativ weit von der wahren ICC entfernt. Hierbei sollte jedoch bedacht werden, dass die wahre ICC eine Mischung von Verteilungsfunktionen der Normalverteilung ist und somit innerhalb des Parameterraums des TDPM Modells liegt. Das Modell hinter dem TDPM HMC Algorithmus basiert hingegen auf einer Mischung von logistischen Verteilungsfunktionen und kann die wahre ICC im besten Fall approximieren. Allerdings kommen als Ursache auch die im Vergleich zu den Standardnormalverteilungen des TDPM Gibbs Samplers etwas weniger informativen Lognormal-Priori-Verteilungen für die Ladungsparameter in Betracht. Zudem ist nicht auszuschließen, dass die Mischverteilungsstruktur der Posteriori-Verteilung für den HMC Sampler ein größeres Problem darstellt als für den Gibbs Sampler. Möglicherweise gelangt er zu selten in Bereiche der Posteriori-Dichte, die für die ICC relevant sind.

Die hier berichteten Ergebnisse sollten jedoch auf keinen Fall überinterpretiert werden, da es sich lediglich um eine Beispielanwendung der Schätzalgorithmen und nicht um eine systematische Simulation handelt. Vor allem in komplexen Modellen kann das Konvergenzverhalten von MCMC Algorithmen von Datensatz zu Datensatz stark variieren. In der praktischen Anwendung sollten daher stets alle drei Algorithmen angewandt und das Konvergenzverhalten im Einzelfall verglichen werden. Dennoch sind die Resultate zumindest für den TDPM Gibbs Sampler und den TDPM HMC Sampler durchaus vielversprechend.

Kapitel 7

Diskussion

7.1 Zusammenfassung

Das Ziel der vorliegenden Arbeit lag in der Herleitung von Methoden zur Schätzung der Parameter des NMIRT Modells (1.15) von Peress (2012). Zu diesem Zweck wurde das Modell zunächst reparametrisiert und dann Duncan und MacEachern (2008), Miyazaki und Hoshino (2009) und Karabatsos (2016) folgend als nonparametrisches bayesianisches Modell formuliert. Hierauf aufbauend konnten drei verschiedene MCMC Algorithmen konstruiert werden: Ein Gibbs Sampler, der auf einer TDPM Priori-Verteilung für die ICC F und einer hierarchischen Reformulierung des Modells mit diskreten Hilfsvariablen basiert, ein Metropolis-within-Gibbs Sampler für ein Modell mit TPT Priori-Verteilung für F und ein HMC Algorithmus für ein Modell mit TDPM Priori-Verteilung für F mit logistischen Mischverteilungskomponenten. In einer Beipielanwendung wurden diese Algorithmen auf einen simulierten Datensatz angewandt. Hierbei traten bei dem auf dem TPT basierenden Gibbs Sampler Konvergenzprobleme auf. Der TDPM Gibbs Sampler und der TDPM HMC Sampler konvergierten und konnten die Parameterstrukur sehr gut reproduzieren. Mithilfe des TDPM Gibbs Samplers gelang darüber hinaus auch die Schätzung der Form der ICC.

7.2 Limitationen und zukünftige Forschung

7.2.1 Effizienz der MCMC Algorithmen

Hinsichtlich der Effizienz der vorgeschlagenen Algorithmen im Allgemeinen besteht noch Forschungsbedarf. So konnte die Beispielanwendung in Kapitel 6 nur einen ersten Eindruck von den praktischen Anwendungsmöglichkeiten der Algorithmen vermitteln. Wie sich die Algorithmen unter anderen datengenerierenden Prozessen und in anderen Stichproben verhalten, muss noch im Rahmen von systematischen Simulationsstudien untersucht werden. Von besonderem Interesse wird hierbei sein, unter welchen Umständen die verschiedenen Algorithmen Konvergenz- bzw. Divergenzverhalten zeigen. Ein weiterer Schwerpunkt künftiger Forschung sollte auf die Frage gelegt werden, ab welcher Stichprobengröße

von einer zufriedenstellenden Genauigkeit in der Parameterschätzung ausgegangen werden kann. Diesbezüglich wäre auch ein Vergleich mit fehlspezifizierten einfacheren Modellen wie den parametrischen mehrdimensionalen Modellen interessant. Möglicherweise können die Ladungs- und Schwierigkeitsparameter in diesen Modellen unter bestimmten Verlustfunktionen besser geschätzt werden. Sie liefern zwar aufgrund der Fehlspezifikation verzerrte Schätzer, weisen aber aufgrund der geringeren Komplexität eventuell eine höhere Stabilität im Sinne einer niedrigeren Varianz auf. Auch die Sensitivität der Parameterschätzungen gegenüber der Wahl der Hyperparameter der Priori-Verteilungen sollte in zukünftigen Studien noch untersucht werden, da sich hier schon in parametrischen Modellen große Unterschiede zeigen (Béguin & Glas, 2001).

7.2.2 Praktische Anwendung in der psychologischen Diagnostik

Die in dieser Arbeit hergeleiteten Algorithmen ermöglichen die Schätzung der Parameter des NMIRT Modells. Damit ist der wichtigste Schritt in Richtung praktischer Anwendbarkeit getan. Bevor das Modell tatsächlich in der psychologischen Diagnostik eingesetzt werden kann, sind jedoch noch einige offene Fragen zu klären.

Ein erstes Problem besteht darin, dass die Parameter des NMIRT Modells in den Parametrisierungen von Modell 2 und Modell 3 nur selten eine psychologische Bedeutung haben. Die Fixierung einzelner Personenparameter stellt zwar die Eindeutigkeit der Ladungsmatrix und der Personenparameter sicher, führt aber nicht zwangsläufig zu einer psychologisch interpretierbaren Lösung. Mit diesem Problem kann jedoch leicht umgegangen werden, da der latente Prädiktor im NMIRT Modell die gleiche lineare Struktur wie im klassischen faktorenanalytischen Modell aufweist. Um eine Lösung nahe einer psychologisch sinnvollen Einfachstruktur zu erhalten, kann daher auf herkömmliche Rotationsmethoden zurückgegriffen werden: Die Ladungsmatrix des NMIRT Modells wird zunächst unter den Restriktionen der Personenparameter geschätzt und anschließend orthogonal oder oblique rotiert.

Ein zweiter für die Anwendung wichtiger Fragenkomplex bezieht sich auf die Modellwahl. Diese ist in mehrfacher Hinsicht von Bedeutung: Da die Dimension Q der latenten Variable im NMIRT Modell kein Modellparameter, sondern ein vor der Parameterschätzung festzulegender unbekannter Wert ist, müssen in der Praxis mehrere NMIRT Modelle mit unterschiedlicher Anzahl an Faktoren miteinander verglichen werden. Darüber hinaus ist es im Anwendungskontext sinnvoll, das NMIRT Modell mit einfacheren Modellen wie beispielsweise parametrischen mehrdimensionalen Modellen hinsichtlich der Modellpassung zu vergleichen. Alle behandelten Modelle enthalten lediglich die trunkierten Versionen des DP und des PT und besitzen somit einen Parameterraum von endlicher Dimensionalität. Da zudem in jeder MCMC Iteration die logarithmierte Likelihood ausgewertet werden kann, ist ein Modellvergleich unter Verwendung von bayesianischen Informationskriterien wie beispielsweise dem Deviance Information Criterion möglich (Sinharay, 2016). Inwiefern andere bayesianische Modellwahlverfahren wie posterior predictive checks oder Bayes Faktoren (Sinharay, 2016) im Kontext der Modelle zur Anwendung kommen können, bleibt noch zu klären.

Damit das NMIRT Modell nach erfolgreicher Schätzung der Itemparameter und der nonparametrischen ICC in der psychologischen Einzelfalldiagnostik eingesetzt werden kann, sind zudem Verfahren zur Schätzung der Personenparameter nötig. Naheliegend, aber zeitaufwendig wäre es, eine neu getestete Person zur Normstichprobe hinzuzufügen und die MCMC Algorithmen mit dieser zusätzlichen Beobachtung erneut laufen zu lassen. Dadurch würden Ziehungen aus der Posteriori-Verteilung der Personenparameter der neuen Person generiert. Ein Vorteil bei diesem Vorgehen bestünde in der Berücksichtigung der Unsicherheit hinsichtlich der Itemparameter. Man könnte die Personenparameter der neuen Person auch wie in anderen Modellen auf der Basis der auf den EAP festgesetzten Itemparametern schätzen. Ein solches Vorgehen gestaltet sich zwar weniger aufwendig, ist aber stark von der Genauigkeit der Itemparameterschätzung in den Normstichproben abhängig, da die Varianz der Posteriori-Verteilung der Itemparameter ignoriert wird.

7.2.3 Erweiterungen der MCMC Algorithmen

TDPM Algorithmen

Die beiden auf der TDPM Priori-Verteilung für die ICC F basierenden Modelle wären um eine zusätzliche Priori-Verteilung für den Präzisionsparameter λ erweiterbar. Dies würde sich unter Umständen positiv auf das Konvergenzverhalten der Algorithmen auswirken, da λ in diesem Fall implizit aus den Daten geschätzt wird (Ishwaran & Zarepour, 2000).

Zudem könnte als Priori-Verteilung für die ICC F statt der hier verwendeten TDP Mean-Mixture (3.33) die allgemeinere TDP Mean-Variance-Mixture (3.34) gewählt werden (Ishwaran & James, 2002). Dies würde das Modell noch flexibler machen, allerdings auch die Herleitung der vollständig bedingten Verteilungen der Item- und Personenparameter im TDPM Gibbs Sampler erschweren.

Denkbar wären auch Priori-Verteilungen für die ICC, die auf Verallgemeinerungen der Stick-Breaking Darstellung des DP basieren. Hierbei würde man in (3.24) statt der sich aus dem DP ergebenden $Beta(1, \lambda)$ Verteilung eine allgemeine, unter Umständen auch von Kovariablen abhängige, Betaverteilung wählen (Ishwaran & James, 2001). Dies könnte sich in bestimmten Situationen als effizienter als der TDP erweisen. Priori-Verteilungen dieser Art wären außerdem im Rahmen des in dieser Arbeit vorgeschlagenen TDPM Gibbs Samplers einfach zu implementieren, da sich alternative Stick-Breaking Darstellungen lediglich auf die vollständig bedingte Verteilung der Mischverteilungsgewichte $\boldsymbol{\pi}$ auswirken würden.

TPT Algorithmus

Durch die Einführung von Priori-Verteilungen für die Parameter der Basisverteilung können TPTs zu stetigen Mischungen von TPTs erweitert werden (Hanson, 2006). Dies wirkt sich eventuell positiv auf die Konvergenz des TPT Algorithmus aus, da in diesem Fall implizit eine datengesteurte Wahl der Basisverteilung stattfinden würde. Zudem hätten die Realisationen der ICCs unter einer stetigen Mischung von TPTs eine deutlich glattere Struktur (Hanson, 2006). Allerdings würde dies im Rahmen des TPT Gibbs Samplers einen weiteren

Metropolisschritt zur Ziehung aus der vollständig bedingten Verteilung dieser Parameter nach sich ziehen.

Ein großer Nachteil des TPT Gibbs Samplers ist die Notwendigkeit von Metropolisschritten bei der Ziehung aus den vollständig bedingten Verteilungen der Item- und Personenparameter. Neben der hier vorgeschlagenen Verwendung von unabhängigen Normalverteilungen mit auf der Posteriori-Verteilung im parametrischen Modell basierenden komponentenspezifischen Varianzen, wären auch andere Vorschlagsverteilungen denkbar und gegebenenfalls effizienter. So würde es sich zum Beispiel anbieten, die Item- und Personenparameter aus den vollständig bedingten Verteilungen des mehrdimensionalen parametrischen Modells zu ziehen. Liegt die wahre ICC nahe an der parametrischen Form, wäre dadurch eine Approximation der vollständig bedingten Verteilungen des TPT Modells möglich. Zudem ist bekannt, dass die Konvergenz von MCMC Algorithmen stark von der Parametrisierung abhängen kann (Robert & Casella, 2004). Eventuell könnten noch weitere Reparametrisierungen des Modells hergeleitet werden, unter denen der auf dem TPT basierende Algorithmus ein günstigeres Konvergenzverhalten zeigt.

7.2.4 Allgemeinere mehrdimensionale IRT Modelle

Alle in dieser Arbeit vorgeschlagenen Modelle und MCMC Algorithmen basieren auf Reparametrisierungen des Modells (1.15) von Peress (2012). Trotz der hohen Flexibilität dieses Modells ist eine weitere Verallgemeinerung nicht ausgeschlossen. Eine erste schon von Peress (2012) vorgeschlagene Generalisierung wäre das Modell (1.16) mit itemspezifischen nonparametrischen ICCs. Hier müsste jedoch zunächst der Identifikationsbeweis von Peress für $Q = 2$ Dimensionen auf einer beliebige Anzahl an Faktoren verallgemeinert werden.

Denkbar ist auch, dass unter bestimmten Restriktionen ein Identifikationsbeweis für additive Modelle der Form

$$P(X_{ij} = 1) = g\left(\sum_{q=1}^{Q} F_{jq}(\xi_{iq})\right)$$

mit g als inverser Linkfunktion und nonlinearen monotonen Funktionen F_{jq} für jedes Item j und jede Dimension q der latenten Variable geführt werden könnte. In diesem Fall könnten auch Modelle geschätzt werden, die nonlineare Einflüsse der einzelnen latenten Variablen zuließen und trotzdem gut interpretierbar wären.

Angesichts der bereits sehr hohen Komplexität der in dieser Arbeit vorgestellten Modelle erscheint es jedoch sinnvoller, zunächst deren Eigenschaften genauer zu untersuchen, bevor man zu noch allgemeineren Modellen übergeht.

7.3 Fazit

Item Response Modelle dienen dem Zweck, von beobachtbaren auf latente Variablen zu schließen. Hierbei bewegen sie sich im Spannungsfeld zwischen psychologischer Interpretierbarkeit und Anpassung an die empirisch beobachteten Daten. Das NMIRT Modell von

Peress (2012) stellt in diesem Zusammenhang einen vielversprechenden Kompromiss dar, da es die Interpretierbarkeit der explorativen Faktorenanalyse mit der Flexibilität der nonparametrischen IRT Modelle verbindet. Die hier vorgeschlagenen Schätzverfahren tragen dazu bei, dass das Modell auch tatsächlich in der psychologischen Diagnostik eingesetzt werden kann.

Anhang A

Traceplots

Dieser Anhang enthält die Traceplots der Itemparameter aus der Anwendung der drei MCMC Algorithmen in Kapitel 6.

A.1 TDPM Gibbs Sampler

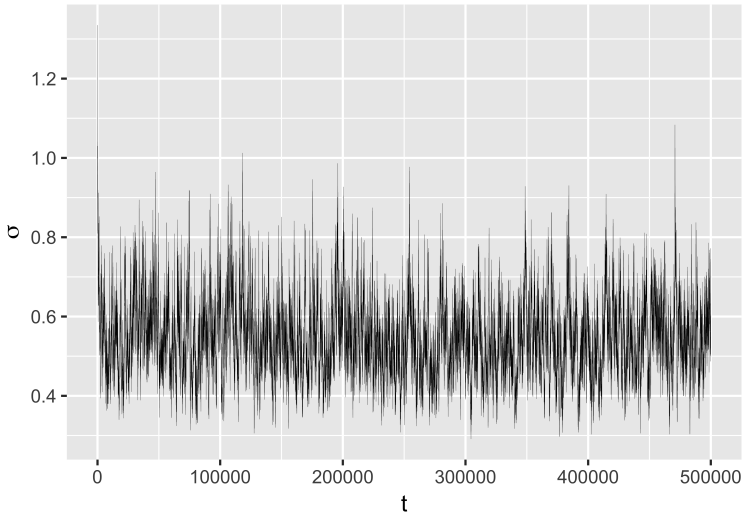

Abbildung A.1: Traceplot des Sigmaparameters aus dem TDPM Gibbs Sampler. Auf der y-Achse ist der Parameterwert in Iteration t des MCMC Algorithmus abgetragen.

Abbildung A.2: Traceplots der ICC Parameter aus dem TDPM Gibbs Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

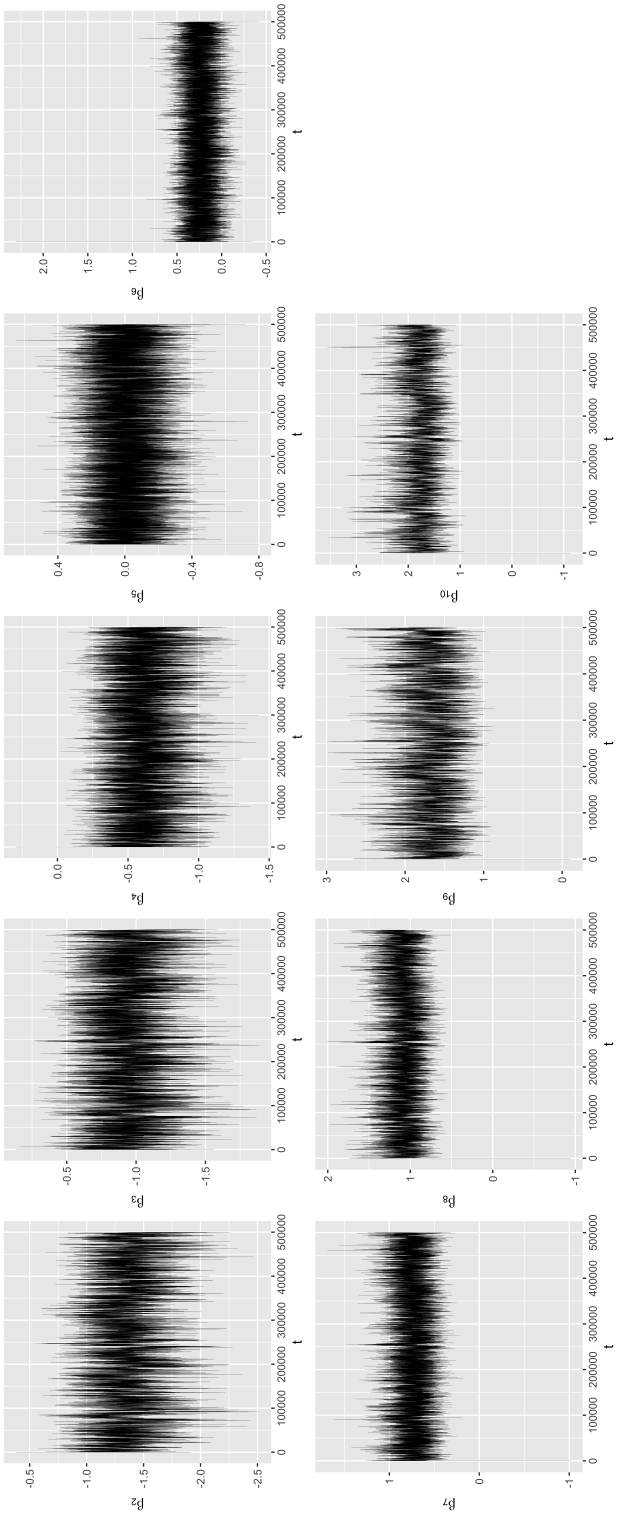

Abbildung A.3: Traceplots der Schwierigkeitsparameter aus dem TDPM Gibbs Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

Abbildung A.4: Traceplots der Ladungen auf Faktor 1 aus dem TDPM Gibbs Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

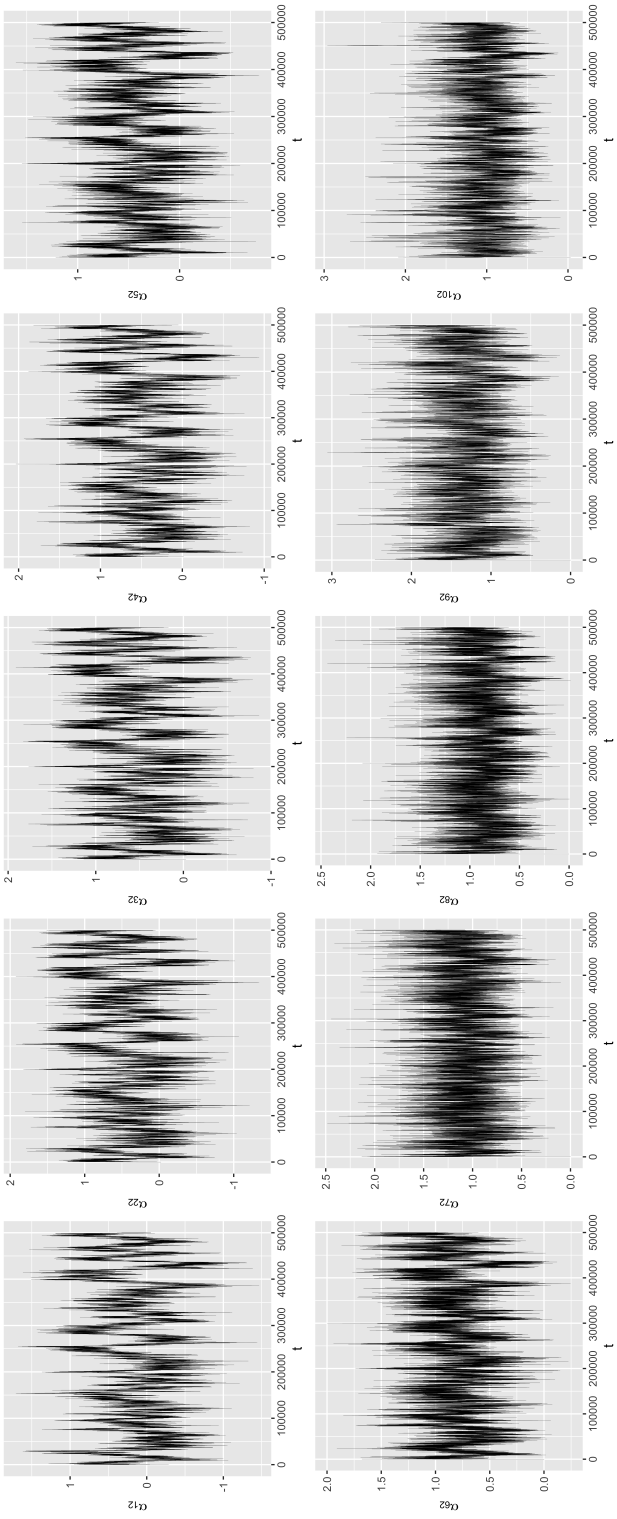

Abbildung A.5: Traceplots der Ladungen auf Faktor 2 aus dem TDPM Gibbs Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

A.2 TPT Gibbs Sampler

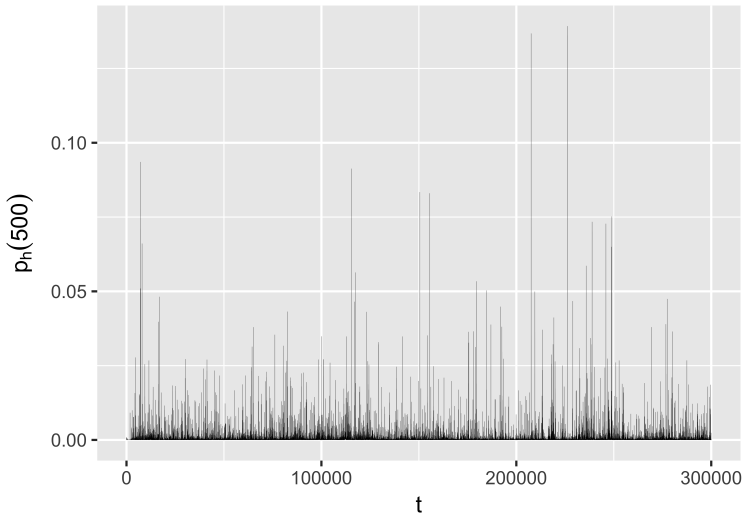

Abbildung A.6: Traceplot eines ICC Parameters aus dem TPT Gibbs Sampler. Auf der y-Achse ist der Parameterwert in Iteration t des MCMC Algorithmus abgetragen.

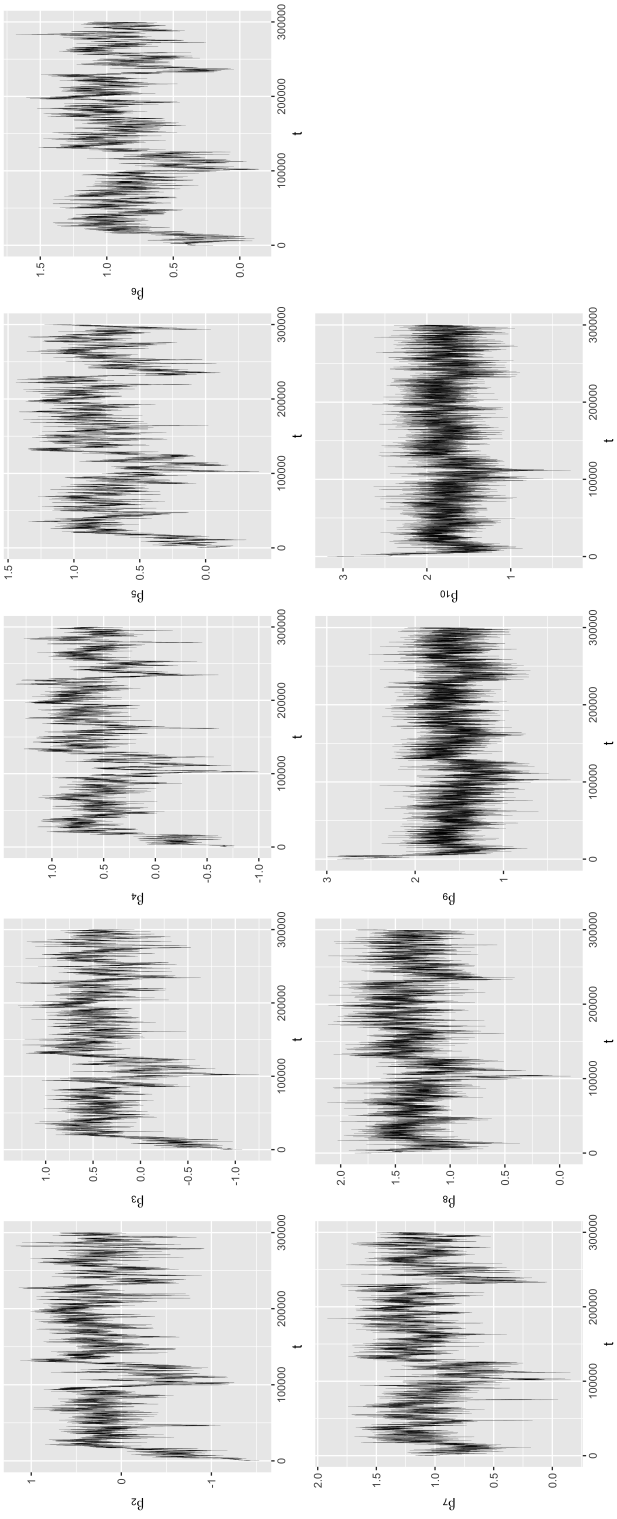

Abbildung A.7: Traceplots der Schwierigkeitsparameter aus dem TPT Gibbs Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

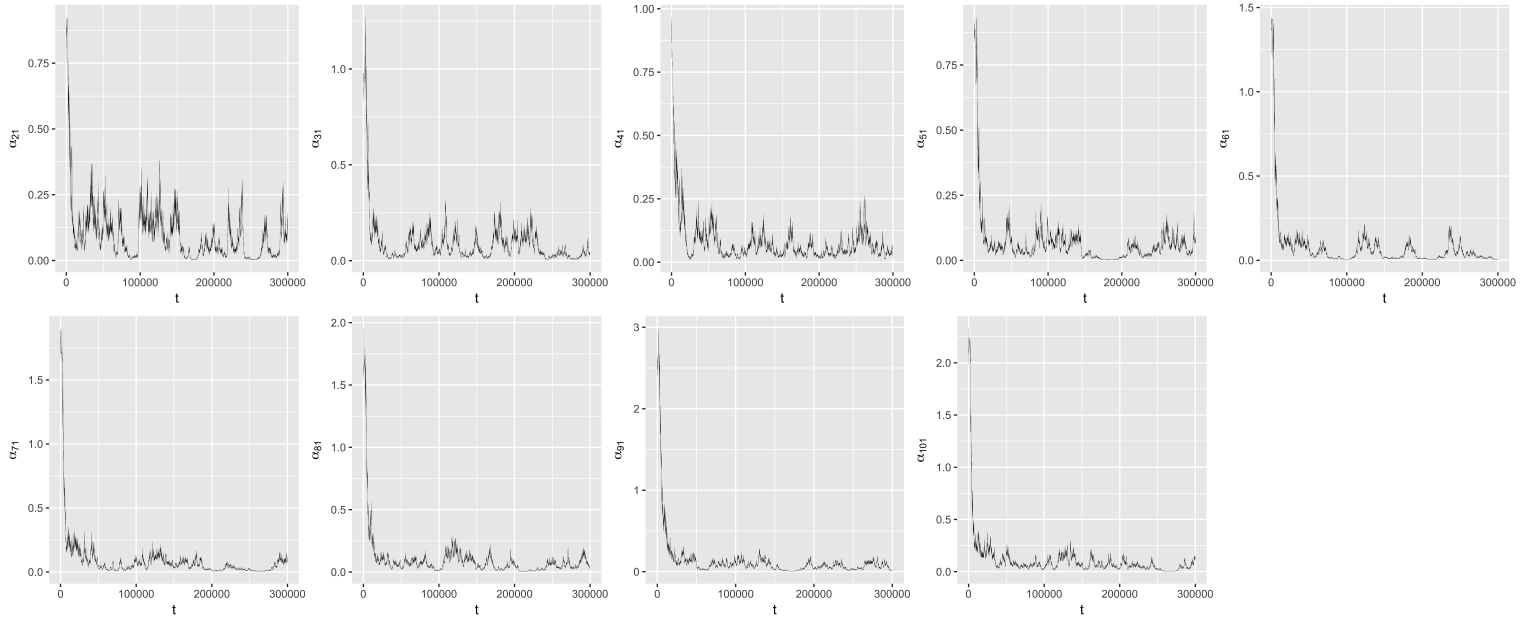

Abbildung A.8: Traceplots der Ladungen auf Faktor 1 aus dem TPT Gibbs Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

Abbildung A.9: Traceplots der Ladungen auf Faktor 2 aus dem TPT Gibbs Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

A.3 TDPM HMC Sampler

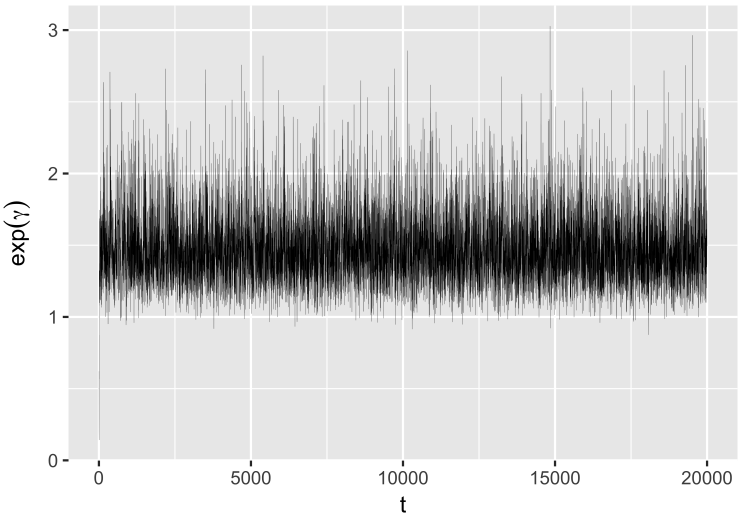

Abbildung A.10: Traceplot des Gammaparameters aus dem TDPM HMC Sampler. Auf der y-Achse ist der Parameterwert in Iteration t des MCMC Algorithmus abgetragen.

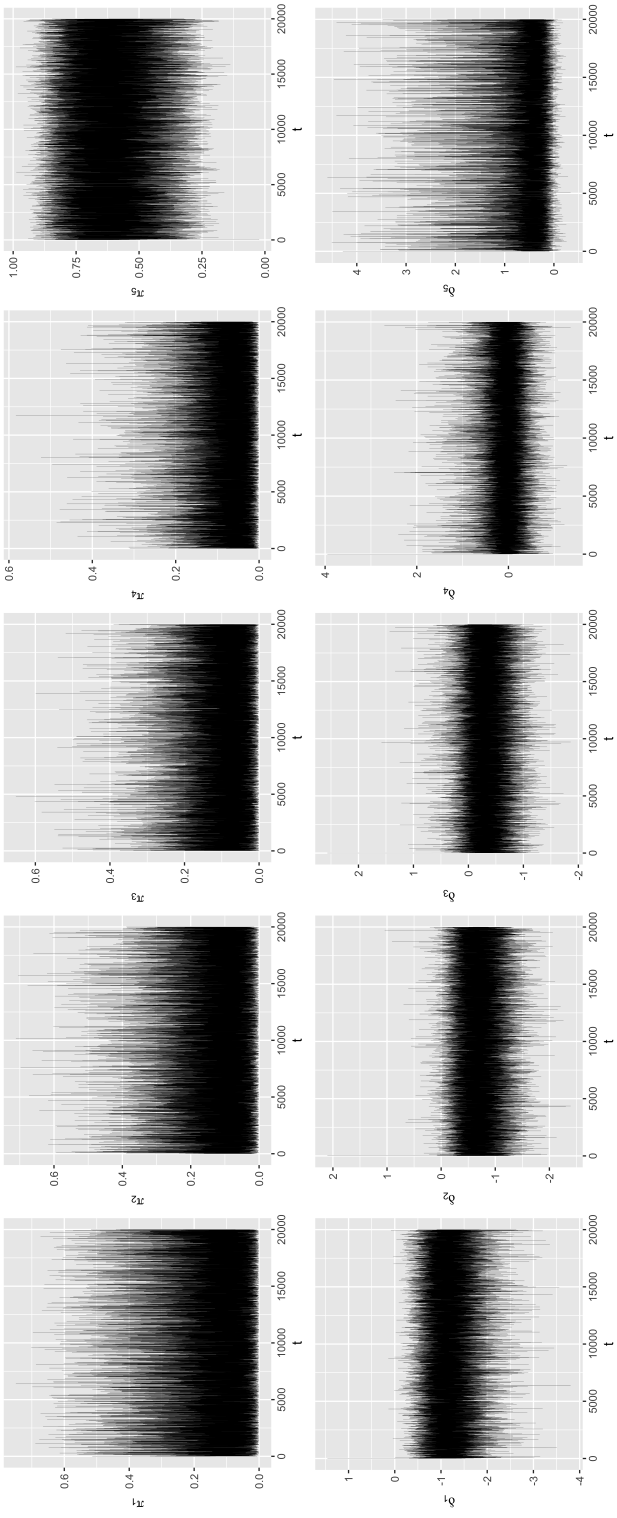

Abbildung A.11: Traceplots der ICC Parameter aus dem TDPM HMC Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

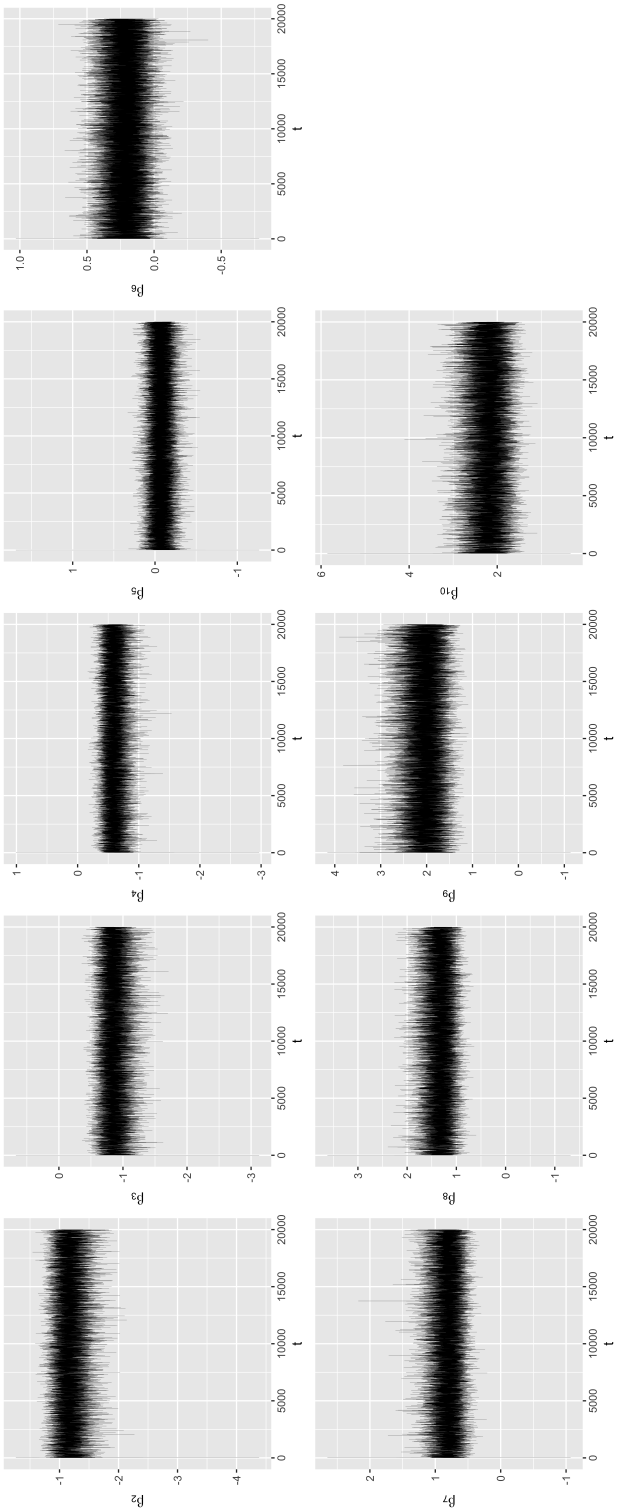

Abbildung A.12: Traceplots der Schwierigkeitsparameter aus dem TDPM HMC Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

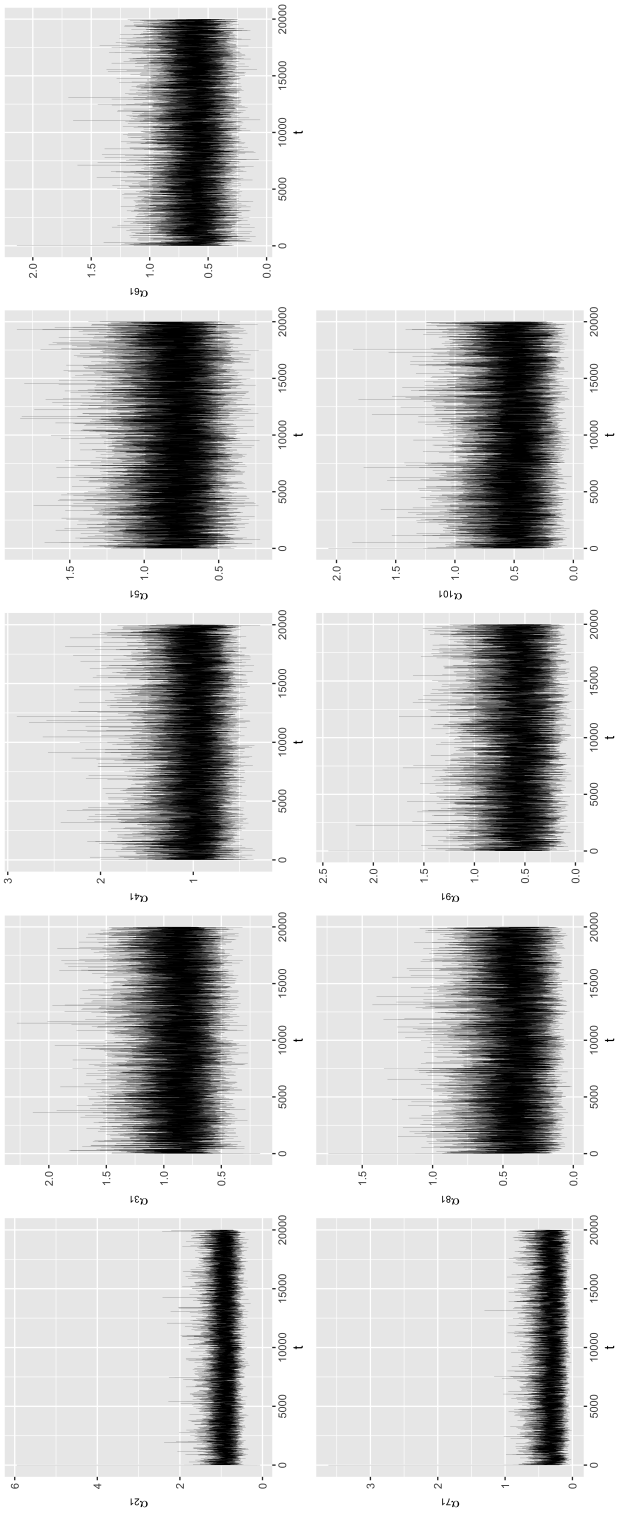

Abbildung A.13: Traceplots der Ladungen auf Faktor 1 aus dem TDPM HMC Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

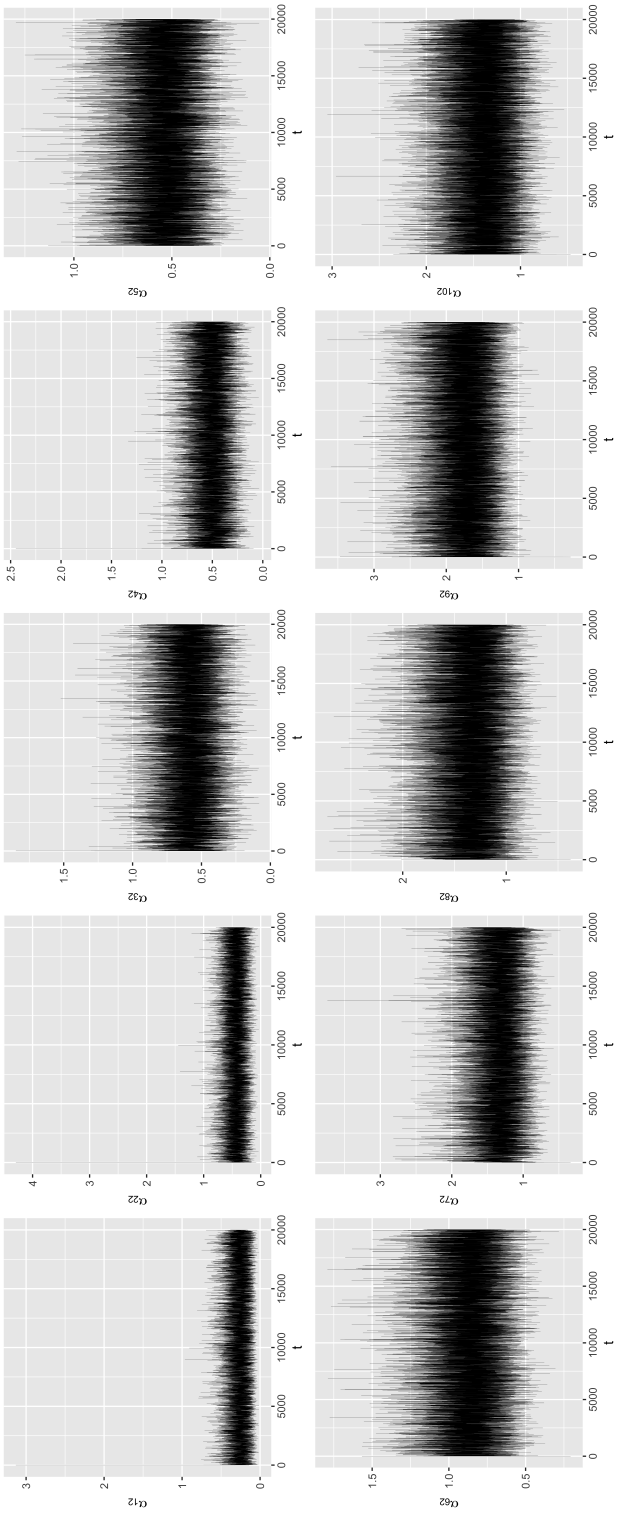

Abbildung A.14: Traceplots der Ladungen auf Faktor 2 aus dem TDPM HMC Sampler. Auf der y-Achse sind jeweils die Parameterwerte in Iteration t des MCMC Algorithmus abgetragen.

Anhang B

Verteilungsübersicht

Univariate Normalverteilung

Symbol:

$$N(\mu, \sigma^2)$$

Dichte:

$$f(x|\mu, \sigma^2) = N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

mit $x \in \mathbb{R}$ $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}_+$

Multivariate Normalverteilung

Symbol:

$$MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Dichte:

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = MVN(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

mit $\mathbf{x} \in \mathbb{R}^k$, $\boldsymbol{\mu} \in \mathbb{R}^k$, $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$ positiv definit mit Determinante $|\boldsymbol{\Sigma}|$ und $k \in \mathbb{N}$

Inverse Gammaverteilung

Symbol:

$$IG(\alpha, \beta)$$

Dichte:

$$f(x|\alpha, \beta) = IG(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$$

mit $x \in \mathbb{R}_+$, $\alpha \in \mathbb{R}_+$, $\beta \in \mathbb{R}_+$ und Γ als Gammafunktion

Bernoulli-Verteilung

Symbol:

$$Ber(\pi)$$

Dichte:

$$f(x|\pi) = Ber(x;\pi) = \pi^x(1-\pi)^{1-x}$$

mit $x \in \{0,1\}$ und $\pi \in [0,1]$

Betaverteilung

Symbol:

$$Beta(\alpha,\beta)$$

Dichte:

$$f(x|\alpha,\beta) = Beta(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

mit $x \in [0,1]$, $\alpha \in \mathbb{R}_+$, $\beta \in \mathbb{R}_+$ und B als Betafunktion

Multinomialverteilung

Symbol:

$$M(\boldsymbol{\pi})$$

Dichte:

$$f(x|\boldsymbol{\pi}) = M(x;\boldsymbol{\pi}) = \prod_{i=1}^{k}\pi_i^{I(x=i)}$$

mit $x \in \{1,2,\ldots,k\}$, $\pi_i \in [0,1]$ für alle $i = 1,2,\ldots,k$, $\sum_{i=1}^{k}\pi_i = 1$ und $k \in \mathbb{N}$

Dirichletverteilung

Symbol:

$$Diri(\alpha_1,\alpha_2,\ldots,\alpha_k)$$

Dichte:

$$\begin{aligned} f(x_1,x_2,\ldots,x_k|\alpha_1,\alpha_2,\ldots,\alpha_k) &= Diri(x_1,x_2,\ldots,x_k;\alpha_1,\alpha_2,\ldots,\alpha_k) \\ &= \frac{1}{B(\alpha_1,\alpha_2,\ldots,\alpha_k)}\prod_{i=1}^{k}x_i^{\alpha_i-1} \end{aligned}$$

mit $x_i \in [0,1]$ für alle $i = 1,2,\ldots,k$, $\sum_{i=1}^{k}x_i = 1$, $\alpha_i \in \mathbb{R}$ für alle $i = 1,2,\ldots,k$, $k \in \mathbb{N}$ und B als multivariater Betafunktion

Literatur

Adams, R. J., Wilson, M. & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*(3), 251–269.

Béguin, A. & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–562.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examineés ability. In F. M. Lord & M. R. Novick (Hrsg.), *Statistical theories of mental test scores* (S. 392–479). Reading, MA: Addison-Wesley.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bock, R. D., Gibbons, R. & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium.

Douglas, J. A. (2001). Asymptotic identifiability of nonparametric item response models. *Psychometrika*, *66*(4), 531–540.

Duncan, K. A. & MacEachern, S. N. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, *8*(1), 41–66.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*(2), 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, *2*(4), 615–629.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*(1), 59–77.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2014). *Bayesian data analysis*. New York: Chapman und Hall/CRC.

Gopalan, G. & Bornn, L. (2016). *FastGP: Efficiently using gaussian processes with Rcpp and RcppEigen*. R package version 1.2.

Hanson, T. E. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, *101*(476), 1548–1565.

Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1351–1381.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 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*(3), 508–532.

Ishwaran, H. & Zarepour, M. (2000). Markov Chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, *87*(2), 371–390.

Johnson, M. S. (2007). Modeling dichotomous item responses with free-knot splines. *Computational Statistics and Data Analysis*, *51*, 4178–4192.

Kahle, D. & Stamey, J. (2017). *invgamma: The inverse gamma distribution*. R package version 1.1.

Karabatsos, G. (2016). Bayesian nonparametric response models. In W. J. van der Linden (Hrsg.), *Handbook of item response theory: Vol. 1. Models* (S. 323–335). New York: Chapman und Hall/CRC.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica*, *17*(2), 125–144.

Mersmann, O., Trautmann, H., Steuer, D. & Bornkamp, B. (2014). *truncnorm: Truncated normal distribution*. R package version 1.0-7.

Miyazaki, K. & Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika*, *74*(3), 375–393.

Müller, P., Quintana, F. A., Jara, A. & Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones & X.-L. Meng (Hrsg.), *Handbook of Markov Chain Monte Carlo* (S. 113–162). New York: Chapman und Hall/CRC.

Peress, M. (2012). Identification of a semiparametric item response model. *Psychometrika*, *77*(2), 223–243.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611–630.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Robert, C. P. & Casella, G. (2004). *Monte Carlo statistical methods*. New York: Springer.

San Martín, E. (2016). Identification of item response theory models. In W. J. van der Linden (Hrsg.), *Handbook of item response theory: Vol. 2. Statistical tools* (S. 127–150). New York: Chapman und Hall/CRC.

San Martín, E., Del Pino, G. & Boeck, P. D. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*(3), 183–203.

San Martín, E. & González, J. (2010). Bayesian identifiability: Contributions to an inconclusive debate. *Chilean Journal of Statistics*, *1*(2), 69–91.

San Martín, E., González, J. & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*(2), 450–467.

San Martín, E. & Quintana, F. (2002). Consistency and identifiability revisited. *Brazilian Journal of Probability and Statistics*, *16*, 99–106.

Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.

Sijtsma, K. & Molenaar, I. W. (2016). Mokken models. In W. J. van der Linden (Hrsg.), *Handbook of item response theory: Vol. 1. Models* (S. 303–321). New York: Chapman und Hall/CRC.

Sinharay, S. (2016). Bayesian model fit and model comparison. In W. J. van der Linden (Hrsg.), *Handbook of item response theory: Vol. 2. Statistical tools* (S. 379–394). New York: Chapman und Hall/CRC.

Stan Development Team. (2017). *The Stan core library*.

Stan Development Team. (2018). *RStan: The R interface to Stan*. R package version 2.17.3.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*(4), 795–809.

Steyer, R. & Eid, M. (2001). *Messen und Testen*. Berlin: Springer.

van der Linden, W. J. (Hrsg.). (2016a). *Handbook of item response theory: Vol. 1. Models*. New York: Chapman und Hall/CRC.

van der Linden, W. J. (Hrsg.). (2016b). *Handbook of item response theory: Vol. 2. Statistical tools*. New York: Chapman und Hall/CRC.

van der Linden, W. J. (2016c). Unidimensional logistic response models. In W. J. van der Linden (Hrsg.), *Handbook of item response theory: Vol. 1. Models* (S. 13–30). New York: Chapman und Hall/CRC.

von Davier, M. & Rost, J. (2016). Logistic mixture-distribution response models. In W. J. van der Linden (Hrsg.), *Handbook of item response theory: Vol. 1. Models* (S. 393–406). New York: Chapman und Hall/CRC.

Waldmann, S. (2014). *Topology. An introduction*. Springer.

Wickham, H. (2016). *ggplot2: Create elegant data visualisations using the grammar of graphics*. R package version 2.2.1.