- LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN -
INSTITUT FÜR STATISTIK

Modellierung der Heterogenität in Bradley-Terry-Luce Modellen

Freie wissenschaftliche Arbeit
zur Erlangung des Grades „Master of Science“ (M.Sc.),
Wintersemester 2013/14

Master-Studiengang Statistik mit wirtschafts- und
sozialwissenschaftlicher Ausrichtung

Seminar für angewandte Stochastik
Prof. Dr. Gerhard Tutz

vorgelegt von: Melanie Poppe
Abgabetermin: 20. Dezember 2013

Zusammenfassung

Subjektive Kriterien wie beispielsweise Geschmack oder Attraktivität sind nicht direkt messbar und lassen sich daher nicht auf einer Skala einordnen. Um trotzdem eine Rangordnung von Objekten aufgrund subjektiver Kriterien vorzunehmen, bieten sich Paarvergleiche an, welche eine große Rolle in der Psychologie und in angrenzenden Themengebieten spielen. In der angewandten Statistik werden Paarvergleiche häufig über Bradley-Terry-Luce Modelle angepasst. In diesen wird angenommen, dass jedes Objekt wahre Präferenzen auf einer subjektiven Skala besitzt. Für alle Versuchspersonen wird dabei die gleiche Präferenzskala vorausgesetzt und angenommen, dass sie die Reizstärken der Objekte in einem Paarvergleich gleich wahrnehmen. Da jedoch in der Realität jede Person die Reizstärken der zu vergleichenden Objekte unterschiedlich wahrnimmt, ist es sinnvoll, bei Paarvergleichen, die mehreren Personen vorgelegt werden, den Einfluss des Probanden in geeigneter Weise zu berücksichtigen. Die vorliegende Arbeit beschäftigt sich mit dem Bradley-Terry-Luce Modell und erweitert es in Bezug auf die Berücksichtigung der Heterogenität der Versuchspersonen. Durch die Einbindung eines personenspezifischen Parameters in das einfache Bradley-Terry-Luce Modell, welcher zusätzlich zu den Itemparametern geschätzt werden soll, wird eine präzisere Schätzung des Modells angestrebt. Die Anpassung dieses personenspezifischen Paarvergleichmodells erfolgt durch einen Algorithmus, in dem abwechselnd bis zur Konvergenz die Itemparameter mittels generalisierter linearer Modelle und die Personenparameter über generalisierte lineare gemischte Modelle geschätzt werden. Anhand von Simulationsstudien wird gezeigt, dass das modifizierte Bradley-Terry-Luce Modell numerisch nicht stabil ist. Die Anwendung des Modells erfolgt anhand zweier Datenbeispiele, die verdeutlichen, dass die im Modell angenommene Heterogenität der Versuchspersonen tatsächlich vorliegt. Das personenspezifische Paarvergleichsmodell, welches auf dem in dieser Arbeit vorgestellten Algorithmus zur Modellschätzung basiert, ist jedoch nicht in der Lage, diese genau zu erfassen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AMS	Arbeitsmarktservice Österreich
AU	Audiounterstütztes Lernen (Item im `trdel`-Datensatz)
BT	Bradley-Terry
BTL	Bradley-Terry-Luce
CO	Computerunterstütztes Lernen (Item im `trdel`-Datensatz)
DGP	Data Generating Process = datengenerierender Prozess
E-Mail	Elektronische Post
Ed(s).	editor(s) = Herausgeber
ed.	edition = Auflage
GL	Gedruckte Lernmittel (Item im `trdel`-Datensatz)
GLM	Generalized Linear Model = generalisiertes lineares Modell
GLMM	Generalized Linear Mixed Model = generalisiertes lineares gemischtes Modell
IIA	Independence from Irrelevant Alternatives
LCA	Luce´s Choice Axiom
ML	Maximum-Likelihood
MSE	Mean Squared Error = mittlere quadratische Abweichung
S.	Seite
s.t.	subject to = unter der Nebenbedingung / unter den Nebenbedingungen
TV	TV-unterstütztes Lernen (Item im `trdel`-Datensatz)
UV	Unterricht/Vortrag (Item im `trdel`-Datensatz)

1. Die Analyse paarweiser Vergleiche

Im Alltag werden von uns Menschen ganz selbstverständlich eine Reihe von Skalen benutzt. So achten wir beim Autofahren auf die *Geschwindigkeit* und richten uns nach ihren Begrenzungen durch Straßenschilder, bekleiden uns unterschiedlich dick je nach Anzeige der Außen*temperatur* auf dem Thermometer oder stehen zu einer bestimmten Uhr*zeit* auf und beginnen den Tag. Die gemeinsame Idee hinter diesen Messwerkzeugen ist gemäß Gediga (1998) die „Transformation von Information, die dem Menschen nicht bzw. indirekt oder nur unpräzise zur Verfügung steht, auf ein leicht ablesbares Instrument. Die Instrumente zeigen die Skalenergebnisse meist durch Zahlen [...] an."
Oftmals möchte man jedoch auch Dinge messen oder in eine Reihenfolge bringen, für die es kein objektives Messinstrument gibt. Solche subjektiven Kriterien wie beispielsweise Geschmack, Attraktivität oder Spielstärke sind nicht direkt messbar und werden als latente Variablen bezeichnet. Um trotzdem ein Ranking aufgrund einer solchen latenten Variablen vorzunehmen, bieten sich Paarvergleiche an (David, 1988).

1.1. Paarvergleiche

Bei einem Paarvergleich werden zwei Objekte oder Items einer betrachteten Objekt- oder Itemmenge systematisch gegenüber gestellt. Diese „Objekte" können je nach Anwendungsbereich Reize, Situationen, Sportmannschaften, Personen usw. sein, die paarweise miteinander verglichen werden. Häufig werden die Objekte eines Paarvergleichs von (Test-) Personen bewertet. Beispielsweise können zwei akustische Reize von einer Versuchsperson danach beurteilt werden, welcher von beiden *lauter* ist oder von zwei alternativen Menüangeboten in einer Speisekarte wird das eine gegenüber dem anderen *präferiert*, usw. Bei einem solchen Paarvergleich beurteilt eine Versuchsperson also nicht *ein* Objekt auf einer vorgegebenen Skala, sie beurteilt vielmehr *zwei* Objekte oder Items danach, welches von beiden das in Frage stehende

Merkmal (z.B. Lautstärke, Präferenz usw.) in höherem Ausmaß hat. Ein „Unentschieden“ ist in dem Fall nicht zulässig (Lukas, 1991). Werden wie in diesem Fall die Versuchspersonen dazu gezwungen, ihre klare Meinung zu äußern, spricht man auch von einem *forced choice format* (Möhring and Schlütz, 2010). Durch diese Befragungstechnik kann zwar die „Tendenz zur Mitte“ beseitigt werden, es können dadurch aber auch wichtige Informationen verloren gehen (SDI-Research, 2013).

In Fällen, in denen insgesamt mehrere Objekte zu beurteilen sind, bieten Paarvergleiche den Vorteil, dass durch Reduzierung der zu beurteilenden Objekte auf zwei Alternativen die Aufgabe vereinfacht wird und somit die Befragten nicht überfordert werden (Lukas, 1991, Courcoux and Semenou, 1997). Paarvergleiche sind jedoch nur sinnvoll, wenn die Zahl der zu beurteilenden Objekte überschaubar bleibt, da das Verfahren sonst unübersichtlich wird und die Auskunftsfähigkeit und -willigkeit der Probanden unter Umständen abnimmt (Decker and Wagner, 2002, Altobelli, 2011).

Es besteht die Möglichkeit der Bewertung der zugrunde liegenden Objektmenge, so dass sich letztendlich eine Rangordnung bzw. ein Ranking der Objekte unter größtmöglicher Berücksichtigung der paarweisen Vergleichsergebnisse ergibt. Gemäß Gaul (1979) ist dies sogar die Zielsetzung der Methode der paarweisen Vergleiche. Bereits Mallows (1957) beschrieb die Verknüpfung zwischen paarweisen Vergleichen und Ranking mit: “The judge is assumed to arrive at a ranking of n objects [...] by first making all the ${}^nC_2$ comparisons between pairs *independently*, but then only accepting the results if they are consistent with a ranking of the n objects.“
Voraussetzung für das Bilden einer Rangordnung auf der Grundlage von Paarvergleichen ist jedoch die *Transitivität* der Urteile. Transitivität bedeutet, dass bei drei paarweise zu vergleichenden Objekten A, B und C gelten muss: Wenn $A \succ B$ und $B \succ C$, dann folgt daraus $A \succ C$. Inkonsistente Antworten bei Paarvergleichen und damit nicht-transitive Rangordnungen werden von Kendall and Babington Smith (1940) auch *zirkuläre Triaden* genannt. Der Name kommt von der geometrischen Repräsentation der Präferenzen in Form eines Dreiecks. Für eine Triade ABC bedeutet beispielsweise $A \longrightarrow B$ die Bevorzugung von A gegenüber B. In Abbildung 1.1 wird der Sachverhalt der Transitivität und der zirkulären Triade noch einmal veranschaulicht.
Bei der Methode des Paarvergleichs kann jedoch eine Verletzung der Transitivitätsprämisse auftreten. Kendall and Babington Smith (1940) verdeutlichen dies an einem

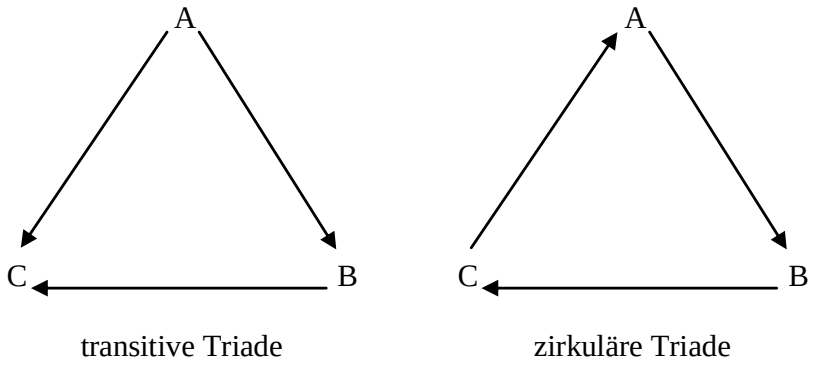

Abbildung 1.1.: Beispielhafte Darstellung von jeweils einer transitiven und zirkukären Triade ABC. $A \longrightarrow B$ stellt die Bevorzugung von A gegenüber B dar. Quelle: Eigene Darstellung.

Beispiel der subjektiven Bewertung der Intelligenz von Personen. Hierbei ist es möglich, dass der Beurteiler Person A für intelligenter hält als Person B, B als C und C als A, sofern dem Beurteiler die zu bewertenden individuellen Paare nacheinander präsentiert werden. Die Wahrscheinlichkeit einer solchen inkonsistenten Bewertung ist womöglich sogar erhöht, wenn anstelle von Intelligenz subjektive Kriterien wie der Musikgeschmack, die Attraktivität von Filmstars oder die Empfindung eines Parfüms betrachtet werden. Ist die Bedingung der Transitivität also nicht erfüllt, kann keine eindeutige Rangfolge aufgestellt werden (Dörsam, 2007).

Ein Beispiel für ein Paarvergleichssystem anhand von Zigarettenmarken bietet Tabelle 1.1. Die $\binom{m}{2}$ Präferenzurteile einer jeden Testperson können anschließend individuell in tabellarischer Form wie in Tabelle 1.2 dargestellt werden. Dabei erfolgt eine reihenweise Zuordnung von Rangzahlen entsprechend der Häufigkeiten, mit denen ein Objekt jeweils allen anderen vorgezogen wurde. Existieren inkonsistente Urteile und damit zirkuläre Triaden, kommen in der individuellen Paarvergleichsmatrix gleiche Reihensummen vor (Decker and Wagner, 2002). Der Eintrag einer 1 in Reihe X und Spalte Y bedeutet $X \succ Y$ und wird von einer zugehörigen 0 in Zeile Y und Spalte X begleitet. Die Diagonale wird logischerweise für Einträge gesperrt. Die Ergebnisse in Tabelle 1.2 bedeuten, dass die Testperson beispielsweise entschieden hat: *HB* $\succ$ *Lord*, *Camel* $\succ$ *HB* oder *Marlboro* $\succ$ *Reval*.
Ein Paarvergleich wird *vollständig* genannt, wenn jedes Objekt mit jedem anderen Objekt verglichen wird und somit von einer Versuchsperson $\binom{m}{2}$ Paarvergleichsur-

„Bitte markieren Sie bei den folgenden Paaren von Zigarettenmarken durch Ankreuzen jeweils jene Alternative, die Sie im direkten Vergleich bevorzugen.“				
☐	*HB*	-	*Camel*	☐
☐	*HB*	-	*Lord*	☐
☐	*Camel*	-	*Lord*	☐
☐	*HB*	-	*West*	☐
☐	*Camel*	-	*West*	☐
☐	*Lord*	-	*West*	☐
☐	*HB*	-	*Reval*	☐
☐	*Camel*	-	*Reval*	☐
☐	*Lord*	-	*Reval*	☐
☐	*West*	-	*Reval*	☐
☐	*HB*	-	*Marlboro*	☐
☐	*Camel*	-	*Marlboro*	☐
☐	*Lord*	-	*Marlboro*	☐
☐	*West*	-	*Marlboro*	☐
☐	*Reval*	-	*Marlboro*	☐

Tabelle 1.1.: Exemplarischer Aufbau einer Paarvergleichsskala anhand von sechs Zigarettenmarken. Quelle: In Anlehnung an Decker and Wagner (2002, S. 239).

teile abgegeben werden (Bortz and Döring, 2009). Im Folgenden wird grundsätzlich von vollständigen Paarvergleichen ausgegangen.

	HB	*Camel*	*Lord*	*West*	*Marlboro*	*Reval*	Summe	Rang
HB	–	0	1	1	1	1	4	2
Camel	1	–	1	1	1	1	5	1
Lord	0	0	–	0	0	0	0	6
West	0	0	1	–	0	0	1	5
Marlboro	0	0	1	1	–	1	3	3
Reval	0	0	1	1	0	–	2	4

Tabelle 1.2.: Individuelle Paarvergleichsmatrix anhand von sechs Zigarettenmarken. Der Eintrag einer 1 in Reihe X und Spalte Y bedeutet $X \succ Y$. Quelle: In Anlehnung an Decker and Wagner (2002, S. 240).

1.2. Skalierung und Datentheorie

Werden Objekte durch den direkten Vergleich miteinander in eine Rangfolge gebracht, spricht man gemäß Malhotra (2010) von einem Verfahren *komparativer* bzw. *vergleichender Skalierung*. Eine solche Skalierung erlaubt nur ordinale Aussagen und wird daher auch als nichtmetrische Skalierung bezeichnet. Das häufigste Verfahren im Rahmen komparativer Skalierung ist die Methode des Paarvergleichs. *Nichtkomparative* (oder auch metrische) *Skalierung* bedeutet im Gegensatz dazu, dass jedes Objekt unabhängig von den anderen in der Objektmenge skaliert wird (z.B. die Beurteilung des Geschmacks alternativer Fruchsäfte auf einer Skala von 1 („schmeckt überhaupt nicht“) bis 5 („schmeckt sehr gut“)). Altobelli (2011) hat dieses Konzept in ihrem Buch „Marktforschung: Methoden - Anwendungen - Praxisbeispiele“ übernommen.

Bereits Coombs (1964) widmete Daten, die durch den Vergleich von Reizen oder Objekten entstehen, in seiner Monographie „A Theory of Data“ eine eigene Klasse in seiner vier-Quadranten-Datentheorie. Die Begriffe Reiz, Item und Objekt werden in Bezug auf Paarvergleiche häufig synonym verwendet. Coombs (1964) geht von einem Modell aus, das sowohl Personen als auch Reize als Punkte in einem gemeinsamen Raum darstellt. Neben anderen Autoren haben sich in der Folgezeit Roskam (1968), Coombs et al. (1970), Roskam (1983) und Gediga (1998) an der Coombs´schen Datentheorie orientiert.

In der Datenklassifikation nach Coombs wird zwischen folgenden vier Klassen unterschieden, Tabelle 1.3 fasst die Coombs´sche Datentheorie noch einmal zusammen (Coombs, 1964, Roskam, 1983, Gediga, 1998):

QI *Präferenzwahldaten* (preferential choice data): Ziel ist die gleichzeitige Skalierung von Reizen und Personen, indem die Abstände der Reize von Idealpunkten der Personen die Ergebnisse festlegen.
Beispiel: *Verglichen mit einem Standardgrauton, welcher dieser beiden Grautöne ist das typischere Grau?*
Hier wird angenommen, dass es für die Personen einen Idealpunkt gibt, der den Reiz mehr oder weniger gut treffen kann. Der Reiz wird umso geringer bzw. niedriger eingestuft, je weiter er vom Idealpunkt entfernt ist.

QII *Einzelreizdaten* (single stimulus data): Ziel ist die gleichzeitige Skalierung von Reizen und Personen durch direkten Vergleich von Reizen und Personen.
Beispiel: *Entweder löst die Person eine Aufgabe eines Leistungstests (dann dominiert die Person über die Aufgabe) oder die Person löst die Aufgabe des Leistungstests nicht (dann wird die Person von der Aufgabe dominiert).*

QIII *Reizvergleichsdaten* (stimulus comparison data): Dieser Datentyp liegt vor, wenn eine Person zwei oder mehr Reize in eine Reihenfolge bringt, wobei sich die Reize auf einer Dimension unterscheiden sollen. Die Skalierung der Reize erfolgt nach Dominanz. Typische Beispiele sind Paarvergleiche bezüglich Helligkeit, Lautstärke oder Tonhöhe. Das Risiko betreffend kann man auch fragen: „Ist Alternative A mit einem größeren Risiko verbunden als Alternative B?“

QIV *Ähnlichkeitsdaten* (similarities data): Zum einen kann man eine Person danach fragen, welcher Reiz aus einer Menge von Reizen einem anderen vorgegebenen Reiz an ähnlichsten ist. Ebenso kann man die Personen auch die Reize aus der Menge hinsichtlich ihrer Ähnlichkeit mit einem Bezugsreiz in eine Reihenfolge bringen lassen.
Beispiel: *Gegeben sei eine Reihe von Situationen. Geben Sie zu jeder Situation an, welche andere Situation am ähnlichsten und/oder unähnlichsten zu der gewählten Situation bezüglich Stressbelastung ist.*
Ziel ist die Beschreibung der Reize zumindest in Klassen von Reizen.

Vergleiche von	Dominanz von Paaren	Nähe von Paaren
Personen und Reizen Einsatz z.B. bei:	QII: Einzelreizdaten Leistungstests	QI: Präferenzwahldaten Persönlichkeitstests
Reizen Einsatz z.B. bei:	QIII: Reizvergleichsdaten Risikovergleich	QIV: Ähnlichkeitsdaten Klassenbildungen

Tabelle 1.3.: Datenklassifikation nach Coombs (1964) mit Beispielen von Einsatzgebieten der Vergleiche von Reizen bzw. von Personen und Reizen. Quelle: In Anlehnung an Gediga (1998, S. 21).

Die Durchführung der Methode des Paarvergleichs ist mit einem hohen Aufwand verbunden (siehe bereits Kendall and Babington Smith, 1940). Für n Personen, die unabhängig voneinander agieren und m Objekte miteinander vergleichen sollen, beträgt die Anzahl an Paarvergleichen $n \cdot \binom{m}{2} = n \cdot \frac{m(m-1)}{2}$. Bei 10 zu vergleichenden

Objekten und 20 Personen sind dies bereits 900 Paarvergleiche. Zur statistischen Modellierung muss daher auf Methoden zurückgegriffen werden, die sich für diese spezielle Datensituation eignen.

Zur Skalierung von Reizen aufgrund von Daten aus Paarvergleichssituationen sind gemäß Tack (1983) vor allem zwei Modellarten verbreitet. In der ersten Modellart werden Reize durch Zufallsvariablen repräsentiert. Jede Reizdarbietung führt zu Realisierungen dieser Variablen und die sich ergebende Reaktion wird durch diese Realisierungen bestimmt. Das prominenteste Beispiel dieser Modellklasse ist Thurstones „Law of Comparative Judgement“ (Thurstone, 1927). Im anderen Fall werden Reize durch geeignete Werte repräsentiert, wobei diese Werte für jedes Reizpaar die Reaktionswahrscheinlichkeiten bestimmen. Das aus diesem Bereich bekannteste ist das *Bradley-Terry-Luce Modell*, kurz *BTL-Modell* (Bradley and Terry, 1952, Luce, 1959). Das BTL-Modell kann als Beispiel eines statistischen Verfahrens in der Coombs´schen Datentheorie für den dritten Quadranten herangezogen werden (vgl. Tabelle 1.3). Von Lukas (1991) wird es als eine außerdordentlich gut fundierte Theorie für die Auswertung von Paarvergleichsdaten angesehen, Tutz (1989), S. 11, bezeichnet es als „das am häufigsten angewandte Paarvergleichssystem“, genauso wie Chan (2011).

Die vorliegende Arbeit beschäftigt sich mit dem Bradley-Terry-Luce Modell und erweitert es in Bezug auf die Berücksichtigung der Heterogenität der Versuchspersonen. In Kapitel 2 wird daher zunächst das Bradley-Terry-Luce Modell eingeführt. Da das generalisierte lineare Modell (Nelder and Wedderburn, 1972) eine Alternative für die Modellierung von Paarvergleichen darstellt, wird in Kapitel 3 die Struktur des Paarvergleichs in seinen Kontext übertragen. Der Schwerpunkt dieser Arbeit liegt auf Kapitel 4. Mit der Erweiterung des Bradley-Terry-Luce Modells durch einen Personenparameter soll die individuelle Wahrnehmung der Objekt-Reizstärken für die einzelnen Versuchteilnehmer in das Modell aufgenommen werden und damit die Heterogenität der Personen berücksichtigt und das einfache Bradley-Terry-Luce Modell verbessert werden. Die Präsentation der zugehörigen Simulationsergebnisse befindet sich in Abschnitt 4.5. Auf zwei Beispieldatensätze werden anschließend in den Kapiteln 5 und 6 das Bradley-Terry-Luce Modell und seine personenspezifische Erweiterung angewendet und miteinander verglichen. Abschließend wird in Kapitel 7 eine Zusammenfassung der Ergebnisse, eine Schlussfolgerung für die Anwendung

des modifizierten BTL-Modells sowie ein Ausblick auf eine weitere Analysemöglichkeit gegeben.
Die Simulationen und Schätzungen der im Folgenden vorgestellten Modelle sowie die Kreation der meisten Abbildungen wurden mit der Statistiksoftware R Development Core Team (2013), Version 3.0.1, durchgeführt.

2. Das Bradley-Terry-Luce Modell

Die ersten Ansätze für Paarvergleichsmodelle gehen bereits auf Zermelo (1929) zurück, der ein Maximum-Likelihood-Verfahren zur Schätzung der Spielstärke von Schachspielern entwickelte. Wiederentdeckt wurde dieses Modell von Bradley and Terry (1952). Die Bezeichnung BTL-Modell bezieht sich auf sie ebenso wie auf Luce (1959) wegen der engen Beziehung zu seinem Wahl-Axiom. Insbesondere durch Theorien der Wahlentscheidungen und durch Fragestellungen bei der Skalierung von Reizen wurde die Entwicklung des paarweisen Vergleichs angeregt. Die Bibliographie von Davidson and Farquhar (1976) beinhaltet mehr als 350 Quellenhinweise zu dem Thema der paarweisen Vergleiche. Ein Überblick über die Entwicklungen wird in Bradley (1976) gegeben, Cattelan (2012) bietet dazu ein Update bezüglich der aktuelleren Entwicklung.
Paarvergleiche spielen eine große Rolle in der Psychologie und in angrenzenden Themengebieten. Darüberhinaus finden sich jedoch auch andere Anwendungsgebiete: Über die Anwendbarkeit des BTL-Modells bei Untersuchungen der Wahrnehmung von Schmerz schreiben Matthews and Morris (1995) und über Experimente hinsichtlich des Geschmacks von Wein und Sekt berichten Lukas (1991) und Oberfeld et al. (2009). Ein Anwendungsbeispiel im Marketingbereich zeigt Gaul (1979) auf und Dittrich et al. (2006) beschreibt einen log-linearen Ansatz des BTL-Modells im Bereich der Politik.

2.1. Datenstruktur

Das Bradley-Terry-Luce Modell, kurz BTL-Modell, ist ein Skalierungsmodell, das Präferenzurteile beschreibt (Gediga, 1998). Ausgangspunkt ist die Präsentation einer Grundmenge $A = \{a_1, a_2, \ldots, a_m\}$ beliebiger Objekte (Produkte, Situationen usw.) gegenüber einer Anzahl $i = 1, \ldots, n$ an Versuchspersonen bzw. Beurteilern in einem Paarvergleichsexperiment. Jede dieser n Personen hat nun die Aufgabe, sich für ein Item der ihr präsentierten Itempaare zu entscheiden. Das Ergebnis des

Paarvergleichs ist eine Relation $\succ \subseteq A \times A$, wobei $a_r \succ a_s$ jeweils bedeutet, dass a_r gegenüber a_s präferiert wird (Lukas, 1991). Diese Präferenzen ergeben die Beobachtungen, mit denen die Skalierungsparameter oder auch *Itemparameter* geschätzt werden sollen. Für alle Versuchspersonen wird hierbei die gleiche Präferenzskala vorausgesetzt, so dass die Beobachtungen als unabhängige Ergebnisse derselben Verteilung behandelt werden können (Frisenfeldt Tuesen, 2007).

Sei $Y_{(r,s)}$ eine binäre Zufallsvariable, welche die Präferenz zwischen zwei Objekten beschreibt, so dass gilt:

$$Y_{(r,s)} = \begin{cases} 1, & \text{falls die Versuchsperson Objekt } a_r \text{ gegenüber Objekt } a_s \text{ bevorzugt,} \\ 0, & \text{falls die Versuchsperson Objekt } a_s \text{ gegenüber Objekt } a_r \text{ bevorzugt} \end{cases}$$

für alle $r, s = 1, \ldots, m$.
Sei $p_{rs} := P(a_r \succ a_s) = P(Y_{(r,s)} = 1)$ zudem die Wahrscheinlichkeit, dass ein Proband Objekt a_r gegenüber Objekt a_s bevorzugt. Die Verteilung von $Y_{(r,s)}$ ist dann gegeben durch

$$Y_{(r,s)} \quad \sim \quad \mathrm{B}(1, p_{rs}) \qquad \forall\, r, s = 1, \ldots, m.$$

2.2. Das Bradley-Terry Modell

Einen Ansatz zur Quantifizierung von Paarvergleichen entwickelte bereits Guttman (1946). Er stand jedoch vor dem Problem numerische Werte für die einzelnen Objekte einer Objektmenge zu bestimmen, welche die besten Vergleiche in Bezug auf eine bestimmte Eigenschaft repräsentieren sollten.
Bradley and Terry (1952) präsentieren als einfache Lösung zu Guttmans Problem der Quantifizierung eine verallgemeinerte Form des Binomialmodells und formulieren sowohl ein mathematisches Modell als auch Maximum-Likelihood Schätzer für die Bewertung der Präferenzen der einzelnen Objekte.

Ausgangspunkt sei ein Paarvergleichsexperiment, bei dem m Objekte der Menge A (siehe Abschnitt 2.1) miteinander verglichen werden sollen. Es wird angenommen, dass jedes dieser Objekte wahre Präferenzen $\pi(a_1), \ldots, \pi(a_m)$ auf einer subjektiven Skala besitzt. Zudem gelten die Bedingungen $\pi(a_r) \geq 0$ und $\sum_{r=1}^{m} \pi(a_r) = 1$, wobei

die letztgenannte Gründen der Identifizierbarkeit dient. Unter der Annahme eines Paarvergleichs zwischen Objekt a_r und Objekt a_s ist die Wahrscheinlichkeit p_{rs} für die Bevorzugung von Item a_r gegenüber Item a_s definiert durch

$$p_{rs} := P(a_r \succ a_s) = \frac{\pi(a_r)}{\pi(a_r) + \pi(a_s)}$$

oder kürzer

$$p_{rs} = \frac{\pi_r}{\pi_r + \pi_s}. \tag{2.1}$$

Gleichung (2.1) lässt sich umformen zu

$$p_{rs} = \frac{1}{1 + \frac{\pi_s}{\pi_r}}$$

und man sieht unmittelbar, dass p_{rs} nur abhängt vom Verhältnis der beiden Skalenwerte π_r und π_s. Je größer π_r ist bei konstantem π_s (bzw. je kleiner π_s ist bei konstantem π_r), desto größer ist auch p_{rs}. Sind beide Skalenwerte gleich groß, also $\pi_r = \pi_s$, dann ist $p_{rs} = p_{sr} = 0.5$.
Aus Gleichung (2.1) folgt außerdem für alle a_r, a_s:

$$p_{rs} + p_{sr} = 1,$$

das heißt, es gibt nur zwei Alternativen: Entscheidung für a_r oder Entscheidung für a_s. Ein „Unentschieden“ als Ergebnis ist in diesem Modell nicht zulässig.
Die Skalenwerte π_r sind ein Maß für die zu beurteilende Eigenschaft. In den unter Abschnitt 1.1 aufgeführten Beispielen also für die Lautstärke zweier akustischer Reize, die Attraktivität der Menüangebote bzw. die Spielstärke zweier Fußballmanschaften. Es lassen sich allerdings nicht für beliebige Paarvergleichsdaten Skalenwerte finden, die Gleichung (2.1) erfüllen. Dafür ist die sogenannte Multiplikationsbedingung

$$\frac{p_{rs}}{p_{sr}} \cdot \frac{p_{st}}{p_{ts}} = \frac{p_{rt}}{p_{tr}} \tag{2.2}$$

für alle Objekttripel a_r, a_s und a_t notwendig und hinreichend (Suppes and Zinnes, 1963). Dabei wird vorausgesetzt, dasss keine der Wahrscheinlichkeiten p_{rs} die Werte 0 oder 1 annimmt. Existieren Skalenwerte π_r, dann sind sie eindeutig bis auf Multiplikation mit einer Konstanten und bilden damit eine Verhältnisskala. Allgemeinere Formulierungen der Eindeutigkeitseigenschaften des Bradley-Terry(-Luce) Modells

werden von Colonius (1980) diskutiert.
Es ist ebenso möglich die Wahrscheinlichkeiten p_{rs} in die sogenannten „logits“ zu transformieren gemäß der Formel

$$\text{logit}(p) := \log\left(\frac{p}{1-p}\right).$$

Äquivalent lautet damit das Bradley-Terry Modell (BT-Modell) aus Gleichung (2.1):

$$\begin{aligned}\text{logit}(p_{rs}) &= \log\left(\frac{p_{rs}}{1-p_{rs}}\right) \\ &= \log\left(\frac{\pi_r}{\pi_r+\pi_s}\cdot\frac{\pi_r+\pi_s}{\pi_s}\right) \\ &= \log\left(\frac{\pi_r}{\pi_s}\right) \\ &= \log(\pi_r) - \log(\pi_s).\end{aligned} \tag{2.3}$$

Werden also logarithmierte Skalierungsparameter $\gamma_r = \log(\pi_r)$ für alle $a_r \in A$ definiert, dann ergeben sich die logit-transformierten Wahrscheinlichkeiten p_{rs} als Linearkombination der γ_r und man erhält als äquivalente Form des BT-Modells die Darstellung

$$\text{logit}(p_{rs}) = \gamma_r - \gamma_s. \tag{2.4}$$

Alternativ gilt ebenso nach Umformulierung der Parameter ($\gamma_r = \log(\pi_r) \Leftrightarrow \pi_r = \exp(\gamma_r)$)

$$p_{rs} = \frac{\exp(\gamma_r)}{\exp(\gamma_r)+\exp(\gamma_s)} = \frac{\exp(\gamma_r-\gamma_s)}{1+\exp(\gamma_r-\gamma_s)}. \tag{2.5}$$

Gleichung (2.5) wird in der Literatur oft auch als Grundgleichung des Bradley-Terry-Modells bezeichnet (Gediga, 1998).
Häufig lässt sich auch folgende Darstellung des BT-Modells finden (z.B. Cattelan, 2012):

$$p_{rs} = F(\gamma_r - \gamma_s). \tag{2.6}$$

Hierbei bezeichnet F die kumulative Verteilungsfunktion einer um Null symmetrischen Zufallsvariablen. Wird für F die Normalverteilung angenommen, definiert Gleichung (2.6) das Thurstone-Modell (Thurstone, 1927). Liegt für F die logistische

Verteilungsfunktion vor, beschreibt (2.6) das Bradley-Terry Modell (Bradley and Terry, 1952). Modelle in der Form von Gleichung (2.6) werden von David (1988) „lineare Modelle“ genannt.

2.3. Luce´s Choice Axiom

Bei einem Paarvergleich hat die Versuchsperson die Aufgabe, sich für jeweils ein Objekt der ihr präsentierten Objektpaare zu entscheiden. Beim Paarvergleich geht es also um ein ganz allgemeines Verhalten, nämlich das der Wahl, wobei ein Individuum eine Option aus einem Set mit mehreren Alternativen auswählt.
Luce (1959) hat mit den Inhalten seiner Monographie „Individual Choice Behavior: A Theoretical Analysis“ das Bradley-Terry Modell entscheidend erweitert, was zur Namensgebung *Bradley-Terry-Luce Modell* führte.
Luce beginnt mit zwei grundlegenden Theoremen des individuellen Wahlverhaltens:

(a) es ist probabilistisch und

(b) die Wahrscheinlichkeit für die Auswahl eines Objekts aus der Menge von mehreren Alternativen steht in einem Zusammenhang mit der Wahrscheinlichkeit für die Auswahl des selben Objekts aus einer größeren Menge von Alternativen.

Das Theorem (a), welches besagt, dass Wahlentscheidungen als probabilistisch zu betrachten sind, steht im Gegensatz zu der Annahme von deterministen Entscheidungen. Zum Verständnis dieser Unterscheidung sei eine Situation betrachtet, in der eine Person ein Objekt x aus der Objektmenge T wähle. Für den Fall eines Paarvergleichs nimmt die deterministische Theorie eine binäre Relation von Präferenzen $\succ$ an, $x \succ y$, $y \succ x$ oder $x \sim y$ für alle $x, y \in T$, wobei $\sim$ für Indifferenz zwischen den beiden Objekten steht.
Eine probabilistische Entscheidungstheorie macht zusätzlich die Annahme, dass bei Entscheidungen für eine Alternative nur Wahrscheinlichkeiten betrachtet werden. Für Paarvergleiche wird daher in der probabilistischen Entscheidungstheorie eine Funktion $P(x, y)$ betrachtet, die jedes mögliche Paar von Alternativen in dem geschlossenen Intervall $[0, 1]$ abbildet. Es werden ebenso in allgemeinen Entscheidungssituationen, nicht nur bei Paarvergleichen, Wahrscheinlichkeiten $P_T(x)$ für die Wahl von x aus einer Menge T spezifiziert. Falls S eine Teilmenge von T ist ($S \subset T$), sei im Folgenden mit $P_T(S)$ die Wahrscheinlichkeit bezeichnet, dass die Wahl einer

Person auf ein Objekt der Teilmenge S fällt.
Eine probabilistische Theorie befolgt die üblichen Wahrscheinlichkeitsaxiome:

(i) Für $S \subset T$, $0 \leq P_T(S) \leq 1$.

(ii) $P_T(\emptyset) = 0,\ P_T(T) = 1$.

(iii) Falls $R, S \subset T$ und $R \cap S = \emptyset$, dann $P_T(R \cup S) = P_T(R) + P_T(S)$.

Durch diese Wahrscheinlichkeitsaxiome werden alle Wahrscheinlichkeitsmaße P_T beschränkt, aber es wird keine Verbindung zwischen verschiedenen Maßgrößen angenommen. Es lässt sich aber vermuten, dass zumindest für das Verhalten bei Entscheidungen die unterschiedlichen Wahrscheinlichkeitsmaße nicht vollständig voneinander unabhängig sind. Beispielhaft formuliert besteht sicherlich ein Zusammenhang zwischen der Wahrscheinlichkeit, dass eine beliebige Person ein Fischgericht zum Abendessen in einem Restaurant aus der Speisekarte wählt, und der Wahrscheinlichkeit, dass dieselbe Person dasselbe Gericht zum Mittagessen aus dem angebotenen Menü wählt (Pleskac, 2012).
Diese Beobachtung motiviert das Theorem (b): Die Wahl eines Items aus einer kleineren Menge von Alternativen steht in einem Verhältnis zur Wahl desselben Items aus einer größeren Menge von Wahlmöglichkeiten. Luce (1959) hat dies wie folgt in dem sogenannten *Luce´s Choice Axiom (LCA)* in zwei Teilen formalisiert. Zum Zweck einer einfacheren Notation stehe die Wahrscheinlichkeit $P(x, y)$ bei einer Menge $S = \{x, y\}$ für das Vorziehen von Alternative x gegenüber Alternative y im Folgenden für $P_{\{x,y\}}(x)$, wenn $x \neq y$.

Sei T eine abgeschlossene Menge, so dass P_S für jedes $S \subset T$ definiert ist.

1. Falls $P(x, y) \neq 0, 1$ für alle $x, y \in T$, dann gilt für $R \subset S \subset T$

$$P_T(R) = P_S(R)P_T(S). \tag{1}$$

2. Falls $P(x, y) = 0$ für einige $x, y \in T$, dann gilt für jedes $S \subset T$

$$P_T(S) = P_{T-\{x\}}(S - \{x\}). \tag{2}$$

Die Notation $S - \{x\}$ steht dabei für $S \setminus \{x\}$, in Worten: „S ohne x“. Zur grafischen Verdeutlichung von Luce´s Wahlaxiom dient Abbildung 2.1.

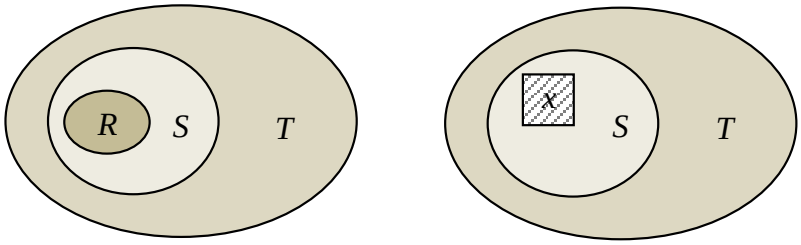

Abbildung 2.1.: Grafische Verdeutlichung von Luce´s Choice Axiom. Teil 1 der Definition ist links, Teil 2 ist rechts dargestellt. Quelle: In Anlehnung an Cohen (2004).

Rückwärts vorgehend ist Teil (2) des Wahlaxioms mehr oder weniger eine „Ordnungs-“ oder „Aufräumannahme“ (Pleskac, 2012). Bei Paarvergleichen erlaubt dieser Teil die Entfernung von niemals gewählten Objekten aus der Teilmenge S, ohne dass dies eine Auswirkung auf die Wahlwahrscheinlichkeiten hätte. Wird beispielsweise bei Paarvergleichen zwischen Gerichten einer Speisekarte mit Fisch oder Huhn niemals das Fischgericht gewählt, dann kann bei einer Objektmenge bestehend aus Fisch-, Huhn- und Rindgerichten das Fischgericht sorglos aus der Objektmenge entfernt werden, so dass die Objektmenge auf nur Huhn oder Rind reduziert wird.
Teil (1) des Axioms lässt sich für $P_T(S) > 0$ auch als bedingte Wahrscheinlichkeit schreiben:

$$P_T(R \,|\, S) = P_S(R) = \frac{P_T(R)}{P_T(S)}.$$

Als konkretes Beispiel dafür sei angenommen, dass T eine Reihe von Hauptgerichten auf einer Speisekarte darstelle, S sei eine echte Teilmenge von T, welche ein Fischgericht enthält, und R sei eine einelementige Menge, nämlich das Fischgericht. Der Kern des Axioms ist die Annahme, dass die Wahrscheinlichkeit für die Auswahl des Fischgerichts für den Fall, dass das Restaurant nur die Hauptgerichte S anbietet, dieselbe ist wie die bedingte Wahrscheinlichkeit für die Auswahl des Fischgerichts, wenn die gesamte Speisekarte verfügbar ist.
Desweiteren setzt Teil (1) des Wahlaxioms voraus, dass

$$P(x, y) = \frac{P_S(x)}{P_S(x) + P_S(y)}.$$

Diese Gleichung kann umgeschrieben werden in

$$\frac{P(x,y)}{P(y,x)} = \frac{P_S(x)}{P_S(y)} \tag{2.7}$$

und ist unter der Bezeichnung „Gesetz des konstanten Verhältnisses“ (constant ratio rule) bekannt (Luce, 1959, Coombs, 1964, Luce, 1977, Roskam, 1983). Dieses Gesetz ist eine probabilistische Form der Annahme der „Independence from Irrelevant Alternatives“ (IIA) (Luce, 1959, Arrow, 1963) und impliziert, dass unter dem Wahlaxiom von Luce für eine Menge von Alternativen T und ihre Teilmengen das Verhältnis $P_S(x)/P_S(y)$ unabhängig von S ist. Anders ausgedrückt sollte zwischen der Wahrscheinlichkeit, dass ein bestimmtes Objekt gewählt wird, und der, dass ein anderes Objekt gewählt wird, ein konstantes Verhältnis bestehen, ungeachtet der zugrunde liegenden Alternativenmenge.
Das LCA beschränkt zudem die möglichen paarweisen Wahrscheinlichkeiten. Diese Einschränkung ist auch als Produktregel für Tripel von Paarvergleichen für Alternativen x, y, z bekannt:

$$P(x,y) \cdot P(y,z) \cdot P(z,x) = P(x,z) \cdot P(z,y) \cdot P(y,x). \tag{2.8}$$

Bei dieser Produktregel ist zu beachten, dass die linke und die rechte Seite der Gleichung die jeweils komplementären Paarvergleich-Wahscheinlichkeiten abbilden. Bei Division der linken Seite durch die rechte Seite der Gleichung ergibt sich somit

$$\frac{P(x,y)}{P(y,x)} \cdot \frac{P(y,z)}{P(z,y)} \cdot \frac{P(z,x)}{P(x,z)} = 1. \tag{2.9}$$

Die Produktregel erlaubt somit die Vorhersage für Wahrscheinlichkeiten von Paarvergleichen $P(x,z)$, falls die Wahrscheinlichkeiten $P(x,y)$ und $P(y,z)$ beide bekannt sind (Pleskac, 2012).

Sowohl beim Gesetz des konstanten Verhältnisses (Gleichung (2.7)) als auch bei der Produktregel (Gleichungen (2.8) und (2.9)) lässt sich eine potentielle Schwäche des LCA erkennen, nämlich die psychologische Auswirkung des Kontextes einer Situation (z.B. die Ähnlichkeit zwischen zwei Auswahlalternativen) auf die Auswahlwahrscheinlichkeit eines Objekts. Eines der ersten Beispiele, welche die Gültigkeit des LCA in Frage stellen, gibt Debreu (1960) in seiner veröffentlichten Buchbe-

sprechung von Luces´s „Individual Choice Behavior“ (Luce, 1959). Eine bekannte Abwandlung von Debreus Beispiel ist das „roter Bus/blauer Bus“-Paradoxon (Train, 2003). Betrachtet wird eine Person, die vor der Entscheidung steht, entweder mit dem Auto zur Arbeit zu fahren oder einen blauen Bus dorthin zu nehmen. Für die Auswahlwahrscheinlichkeiten wird in diesem Fall angenommen, dass sie gleich sind: $P_A = P_{bB} = 1/2$, wobei A für das Auto und bB für den blauen Bus steht. Das Verhältnis der Wahrscheinlichkeiten beträgt somit $P_A/P_{bB} = 1$. Nun sei eine dritte Möglichkeit, ein roter Bus (rB), zu den Möglichkeiten in die Arbeit zu gelangen hinzugefügt. Angenommen, der Person sei die Farbe des Busses gleichgültig und sie entscheide sich immer mit der gleichen Wahrscheinlichkeit zwischen der Möglichkeit mit dem Auto zu fahren oder den Bus zu nehmen. Die Auswahlwahrscheinlichkeit für das Auto beträgt somit weiterhin $1/2$, während die Wahrscheinlichkeiten für die jeweiligen Bustypen $1/4$ betragen und das Verhältnis ihrer Wahrscheinlichkeiten somit $P_{rB}/P_{bB} = 1$ ergibt. Das LCA setzt jedoch voraus, dass sich bei gleichzeitiger Vorlage einer Menge S bestehend aus den Möglichkeiten A, bB, rB für die Entscheidung der Person die Wahrscheinlichkeiten $P_A = P_{bB} = P_{rB} = 1/3$ ergeben. Die starke Ähnlichkeit der Alternativen roter Bus/blauer Bus wird nicht berücksichtigt und die Wahrscheinlichkeit, mit dem Auto zu fahren, wenn alle drei Alternativen angeboten werden, ist unrealistisch niedrig. In diesem Fall wird also die Wahrscheinlichkeit dafür, mit einem der beiden Busse zu fahren, über- und die Wahrscheinlichkeit dafür, das Auto zu wählen, unterschätzt (vgl. auch Tutz (2000)).

2.4. Das Wahlaxiom und das BTL-Modell

Weiter auf Luce´s Wahlaxiom aufbauend kann eine Funktion v über die Menge T definiert werden, so dass gilt

$$v(x) = c\, P_T(x),$$

wobei c eine beliebige positive Konstante ist. Für jedes $x, y \in S \subset T$ kann die Wahrscheinlichkeit für die Entscheidung von Alternative x über das folgende mathematische Modell berechnet werden:

$$P_S(x) = \frac{v(x)}{\sum_{y \in S} v(y)}. \tag{2.10}$$

Mit nur zwei Alternativen kann diese fundamentale Gleichung (2.10) umgeschrieben werden zu

$$P(x, y) = \frac{v(x)}{v(x) + v(y)}. \tag{2.11}$$

Diese Skalendefinition in Gleichung (2.11) ist äquivalent zu dem mathematischen Modell für Paarvergleiche von Bradley and Terry (1952), welches bereits unter Abschnitt 2.2 in Gleichung (2.1) vorgestellt wurde (Luce, 1959).

Das dem Wahlaxiom genügende Modell wird üblicherweise *Bradley-Terry-Luce Modell* (BTL-Modell) genannt. Es genügt den Bedingungen der Unabhängigkeit von irrelevanten Alternativen (IIA) (siehe Gleichung (2.7)) und der einfachen Skalierbarkeit und beachtet ebenso die Produktregel für Objekttripel (siehe Gleichungen (2.2), (2.8) und (2.9)) (Roskam, 1983).

2.5. Die Likelihood-Funktion

Um die Likelihood-Funktion für das Bradley-Terry-Luce Modell aufzustellen, wird zunächst Unabhängigkeit zwischen den einzelnen Vergleichen der Objektpaare angenommen. Die fundamentale Gleichung im BTL-Modell $p_{rs} = \pi_r/(\pi_r + \pi_s)$ ist die Wahrscheinlichkeit, dass bei einem Paarvergleich Item r vor Item s bevorzugt wird (siehe auch Gleichung (2.1)). Diese lässt sich ebenfalls ausdrücken als

$$\left(\frac{\pi_r}{\pi_r + \pi_s}\right)^{2-d_{rsk}} \left(\frac{\pi_s}{\pi_r + \pi_s}\right)^{2-d_{srk}} = \frac{\pi_r^{2-d_{rsk}} \pi_s^{2-d_{srk}}}{\pi_r + \pi_s},$$

wobei sich für $d_{rsk} = 1$ und $d_{srk} = 2$ die Wahrscheinlichkeit $\pi_r/(\pi_r + \pi_s)$ für die Präferenz von Objekt r vor Objekt s in der k-ten Wiederholung des Paarvergleichs ergibt. Entsprechend erhält man für $d_{rsk} = 2$ und $d_{srk} = 1$ die Wahrscheinlichkeit $\pi_s/(\pi_r + \pi_s)$, dass Item s in der k-ten Wiederholung Item r vorgezogen wird.
Durch Multiplikation der entsprechenden Ausdrücke für alle Vergleiche innerhalb eines Durchgangs und für alle n Wiederholungen erhält man unter Annahme der Unabhängigkeit für alle Vergleiche die Likelihood-Funktion

$$L(\boldsymbol{\pi}) = \prod_{r<s} \frac{\pi_r^{y_{rs}} \pi_s^{n_{rs}-y_{rs}}}{(\pi_r + \pi_s)^{n_{rs}}}. \tag{2.12}$$

Der Term n_{rs} steht in Gleichung (2.12) für die Anzahl an Vergleichen zwischen den Objekten r und s, y_{rs} bezeichnet die Häufigkeit der Vergleiche, bei denen Objekt r gegenüber Objekt s bevorzugt wurde.
Die Log-Likelihood-Funktion erhält man anschließend durch Logarithmieren der Likelihood-Funktion:

$$\begin{aligned} l(\boldsymbol{\pi}) &= \sum_{r<s} \Big(y_{rs} \, \log(\pi_r) + (n_{rs} - y_{rs}) \, \log(\pi_s) - n_{rs} \, \log(\pi_r + \pi_s) \Big) \\ &= \sum_{r<s} \left(y_{rs} \, \log\left(\frac{\pi_r}{\pi_s}\right) + n_{rs} \, \log\left(\frac{\pi_s}{\pi_r + \pi_s}\right) \right). \end{aligned}$$

Die Log-Likelihood-Funktion ist skaleninvariant, da $l(\boldsymbol{\pi}) = l(\alpha\boldsymbol{\pi})$ für jedes $\alpha > 0$. Aus Gründen der Identifizierung ist es daher sinnvoll anzunehmen, dass $\sum_{r=1}^{m} \pi_r = 1$. Um den Maximum-Likelihood Schätzer (ML-Schätzer) zu erhalten, wird üblicherweise die Log-Likelihood- anstelle der Likelihood-Funktion maximiert, da der Logarithmus lediglich eine monotone Transformation darstellt und rechnerisch leichter handzuhaben ist. Im Fall von Bradley-Terry-Luce Modellen handelt es sich somit um ein Maximierungsproblem unter Nebenbedingungen (Huang et al., 2006):

$$\begin{aligned} &\max_{\boldsymbol{\pi}} \, l(\boldsymbol{\pi}) \\ \text{s.t.} \quad &\pi_r \geq 0, \; r = 1, \ldots, m; \quad \sum_{r=1}^{m} \pi_r = 1. \end{aligned}$$

Dieser Ansatz geht auf Zermelo (1929) zurück und wurde im Laufe der Zeit für andere Situationen angepasst und erweitert. Während Bradley and Terry (1952) das Maximierungsproblem zur Berechnung des ML-Schätzers $\hat{\pi}_r$ mittels Lagrange-Multiplikatorenregel lösen, gehen Hunter (2004) und Huang et al. (2006) das Problem mit einem iterativen Algorithmus an. Eine Alternative zu diesen beiden Möglichkeiten bietet die Schätzung der Itemparameter mittels Einbindung des BTL-Models in ein generalisiertes lineares Modell (GLM) unter Verwendung der logistischen Regression. Detailliertere Informationen dazu finden sich im folgenden Kapitel.

3. Einbindung des BTL-Modells in das GLM

Lineare Regressionsmodelle spielen zweifellos eine Hauptrolle in der Statistik. Zweck einer Regressionsanalyse ist die Beschreibung der Eigenschaften einer Zielvariable y in Abhängigkeit von Kovariablen $x_1, \ldots, x_k$. Die Zielvariable wird dabei auch abhängige Variable oder Response, die Kovariablen werden auch erklärende Variablen oder Regressoren genannt. Im Wesentlichen unterscheiden sich Regressionsmodelle durch verschiedene Arten von Zielvariablen (stetig, binär, kategorial oder Zählvariablen) und unterschiedliche Typen von erklärenden Variablen, welche ebenso stetig, binär oder kategorial sein können. Neben der direkten Anwendung in Regressionsfragestellungen dienen lineare Modelle auch als Grundlage für komplexere Regressionsverfahren, wie beispielsweise die generalisierten linearen Modelle (GLMs). Diese umfassen in einem methodisch einheitlichen Rahmen viele Regressionsansätze für nicht notwendigerweise normalverteilte Zielvariablen. Sie enthalten beispielsweise das Logit-Modell für binäre Zielvariablen, aber auch das klassische lineare Modell mit Normalverteilungsannahme als Spezialfall (Fahrmeir et al., 2009). Ebenso stellt das GLM einen natürlichen Rahmen zur Modellierung von Paarvergeichsexperimenten zur Verfügung, so dass sich Bradley-Terry-Luce Modelle in das GLM einbinden lassen und somit Schätzer für die Itemparameter berechnet werden können. Ein weiterer Vorteil von GLMs liegt in seiner einfachen Ausführung über Standardsoftware.

3.1. Generalisierte lineare Modelle

Generalisierte lineare Modelle wurden eingeführt von Nelder and Wedderburn (1972) und repräsentieren eine Klasse von Regressionsmodellen für feste Effekte für verschiedene Typen abhängiger Variablen, d.h. stetig, dichotom oder zählend. Das traditionelle Buch über GLMs stammt von McCullagh and Nelder (1989), eine kürzere Einführung findet sich bei Dobson (2002). Collett (2003) betrachtet die Modellierung

binärer Daten, eine Bayesianische Perspektive liefert Dey et al. (2000). Zusätzlich sei noch auf Fahrmeir and Tutz (2001) und Agresti (2002) verwiesen.

Sei $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ die Realisation der Stichprobe $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ und $E(\mathbf{Y}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ der Vektor der Erwartungswerte.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}$$

ist die Designmatrix und $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^\top$ sind die Parameter, durch welche das Modell beschrieben werden soll, wobei $\operatorname{rg}(\boldsymbol{X}) = k + 1 = p < n$ gilt.
Das klassische lineare Regressionsmodell hat die Form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}; \sigma^2 \boldsymbol{I}),$$

mit dem Fehlervektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ und der Einheitsmatrix $\boldsymbol{I}$. Für den Erwartungswert $E(\boldsymbol{y})$ gilt der lineare Zusammenhang

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}.$$

Die Klasse der generalisierten linearen Modelle bietet flexible Verallgemeinerungen des einfachen linearen Modells. Die Grundstruktur verallgemeinerter linearer Modelle lässt sich in mehrere Komponenten aufteilen. Die Zufallskomponente spezifiziert die Verteilung des bedingten Response y_i gegeben $\boldsymbol{x}_i$, die strukturelle Annahme spezifiziert die Verbindung zwischen der erwarteten abhängigen Variable und den Kovariablen.

(1) *Zufallskomponente und Verteilungsannahme:*
Für gegebene Kovariablen $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})$ sind die Zielvariablen y_i (bedingt) unabhängige Beobachtungen aus einer einfachen Exponentialfamilie. Für die abhängige Familie wird als Dichte angenommen

$$f(y_i \,|\, \theta_i, \phi_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right).$$

Dabei ist θ_i der *natürliche* oder *kanonische* Parameter, ϕ_i der *Dispersions-* bzw. *Skalenparameter* und $b(.)$ und $c(.)$ sind spezifische Funktionen, dem jeweiligen

Typ der Exponentialfamilie entsprechend. Tabelle 3.1 liefert eine Übersicht über ausgewählte Verteilungen, die der einfachen Exponentialfamilie angehören.
In Exponentialfamilien steht der Erwartungswert in direkter Beziehung zur Funktion $b(\theta_i)$ in folgender Form:

$$\mu_i = b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta}.$$

Für die Varianz ergibt sich somit

$$\sigma_i^2 = \mathrm{Var}(y_i) = \phi_i b''(\theta_i) = \frac{\phi_i \partial^2 b(\theta_i)}{\partial \theta^2}.$$

Die Verknüpfung zwischen Erwartungswert und Varianz beinhaltet den Dispersionsparameter ϕ_i. Für einige Verteilungen ist dieser Parameter fix, für andere Verteilungen wird er in Abhängigkeit der Daten gewählt. In allen Fällen nimmt die Dispersion die allgemeine Form

$$\phi_i = \phi a_i$$

an, wobei a_i grundsätzlich bekannt ist und mit Ausnahme des Vorliegens einer Binomialverteilung ($a_i = 1/m_i$) den Wert 1 annimmt.

(2) *Strukturannahme:*
Die Kovariablen $\boldsymbol{x}_i$ gehen in das Modell in linearer Form durch den sogenannten linearen Prädiktor

$$\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \qquad i = 1, \ldots, n$$

ein, wobei $\boldsymbol{\beta}$ ein unbekannter Parametervektor der Dimension p ist. Diese lineare Komponente gibt dem GLM seinen Namen (Tutz, 2012). Die Beziehung zwischen dem linearen Prädiktor und dem bedingten Erwartungswert $\mu_i = E(y_i \,|\, \boldsymbol{x}_i)$ ist bestimmt durch die Transformation

$$\mu_i = h(\eta_i) = h(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \quad \text{bzw.} \quad g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

wobei h eine (eineindeutige und zweimal differenzierbare) *Responsefunktion* und g die sogenannte *Linkfunktion*, d.h. die Umkehrfunktion oder Inverse $g = h^{-1}$ von h ist.

(a) Dichte

$$f(y_i \mid \theta_i, \phi_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right), \quad \phi_i = \phi a_i$$

(b) Exponentialfamilien

Verteilung	Notation	μ_i	$\theta(\mu_i)$	$b(\theta_i)$	ϕ	a_i
Normal	$\mathrm{N}(\mu_i, \sigma^2)$	μ_i	μ_i	$\theta_i^2/2$	σ^2	1
Bernoulli	$\mathrm{B}(1, \pi_i)$	π_i	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\log(1+\exp(\theta_i))$	1	1
Binomial	$\mathrm{B}(m_i, \pi_i)/m_i$	π_i	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\log(1+\exp(\theta_i))$	1	$\frac{1}{m_i}$
Poisson	$\mathrm{Po}(\lambda_i)$	λ_i	$\log(\mu_i)$	$\exp(\theta_i)$	1	1
Gamma	$\Gamma(\nu, \frac{\nu}{\mu_i})$	μ_i	$-1/\mu_i$	$-\log(-\theta_i)$	$\frac{1}{\nu}$	1
Inverse Gauss	$\mathrm{IG}(\mu_i, \lambda)$	μ_i	$-1/(2\mu_i^2)$	$-(-2\theta_i)^{1/2}$	$1/\lambda$	1

(c) Erwartungswert und Varianz

Verteilung	$\mu_i = b'(\theta_i)$	$b''(\theta_i)$	$\mathrm{Var}(y_i) = \phi_i\, b''(\theta_i)$
Normal	$\mu_i = \theta_i$	1	σ^2
Bernoulli	$\mu_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$	$\pi_i(1-\pi_i)$	$\pi_i(1-\pi_i)$
Binomial	$\mu_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$	$\pi_i(1-\pi_i)$	$\frac{1}{m_i}\pi_i(1-\pi_i)$
Poisson	$\lambda_i = \exp(\theta_i)$	λ_i	λ_i
Gamma	$\mu_i = -\frac{1}{\theta_i}$	μ_i^2	$\frac{\mu_i^2}{\nu}$
Inverse Gauss	$\mu_i = (-2\theta_i)^{-1/2}$	μ_i^3	μ_i^3/λ

Tabelle 3.1.: Ausgewählte Verteilungen und Eigenschaften von Exponentialfamilien. Quelle: In Anlehnung an Fahrmeir et al. (2009, S. 219) und Tutz (2012, S. 61).

Zusammenfassend ist ein generalisiertes lineares Modell vollständig bestimmt durch den Typ der Exponentialfamilie, welcher die Verteilung von $y_i \,|\, \boldsymbol{x_i}$ spezifiziert, durch die Form des linearen Prädiktors, d.h. die Auswahl und Kodierung der erklärenden Variablen, und durch die Response- oder Linkfunktion (Fahrmeir and Tutz, 2001, Tutz, 2012).

3.2. Binäre Regression

Binäre Regressionsmodelle werden für die Analyse der Beziehung zwischen einem binären Response und erklärenden Variablen herangezogen. Wie bereits in Abschnitt 2.1 beschrieben, liegt bei Paarvergleichen für die Zufallsvariable $Y_{(r,s)}$, welche die Präferenz zwischen zwei Objekten beschreibt, ein binärer Response vor, welcher die Werte 0 und 1 annehmen kann:

$$Y_{(r,s)} = \begin{cases} 1, & \text{falls Objekt } r \text{ bei Vergleich der Objekte } r \text{ und } s \text{ bevorzugt wird,} \\ 0, & \text{falls Objekt } s \text{ bei Vergleich der Objekte } r \text{ und } s \text{ bevorzugt wird.} \end{cases}$$

Im Folgenden wird daher kurz auf die Theorie der binären Regressionsmodelle, insbesondere dabei auf die logistische Regression eingegangen.

Als Ausgangspunkt werden von zu n Objekten oder Individuen vorliegende Daten $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$, betrachtet, welche den Beobachtungen einer binären, durch 0 und 1 kodierten Zielvariable y und den Kovariablen $x_1, \ldots, x_k$ entsprechen. Ziel einer binären Regressionsanalyse ist die Modellierung und Schätzung des Effekts der Kovariablen auf die (bedingte) Wahrscheinlichkeit für das Auftreten von $y_i = 1$ bei gegeben Kovariablenwerten $x_{i1}, \ldots, x_{ik}$:

$$\pi_i = P(y_i = 1 \,|\, x_{i1}, \ldots, x_{ik}) = E(y_i \,|\, x_{i1}, \ldots, x_{ik}).$$

Für die Zielvariablen wird dabei (bedingte) Unabhängigkeit angenommen.

In allen üblichen binären Regressionsmodellen wird die Wahrscheinlichkeit π_i durch eine Beziehung der Form

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i_1} + \ldots + \beta_k x_{ik}) = h(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \tag{3.1}$$

mit $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$, $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})^\top$ und dem linearen Prädiktor η_i verknüpft. Dabei ist h eine auf der ganzen reellen Achse streng monoton wachsende Verteilungsfunktion, so dass grundsätzlich $h(\eta) \in [0, 1]$ gilt und die Beziehung (3.1) mit Hilfe der Inversen $g = h^{-1}$ in der Form

$$\eta_i = g(\pi_i)$$

geschrieben werden kann (Fahrmeir et al., 2009). Die bekanntesten unter den binären Regressionsmodellen sind die folgenden:

- **Logit-Modell**:
  Das Logit-Modell ergibt sich durch die Wahl der logistischen Responsefunktion

  $$\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

  bzw. der Logit-Linkfunktion

  $$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i_1} + \ldots + \beta_k x_{ik}.$$

  Damit erhält man ein lineares Modell für die *logarithmierten Chancen (log-odds)*, kurz *logits*

  $$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right).$$

  Die Responsefunktion des Logit-Modells ist in Abbildung 3.1 dargestellt.

- **Probit-Modell**:
  Beim Probit-Modell wird für die Responsefunktion h die Verteilungsfunktion Φ der Standardnormalverteilung verwendet, d.h.

  $$\pi_i = \Phi(\eta_i) = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$

  Das Modell legt dem linearen Prädiktor η_i keinerlei Restriktionen auf, es ist jedoch bei Berechnung der Likelihood ziemlich rechenintensiv aufgrund der notwendigen Auswertung von Φ (Fahrmeir and Tutz, 2001).

- **Komplementäres log-log-Modell**:
  Dieses Modell besitzt für die Responsefunktion die Extremwert-Verteilungsfunktion

$$h(\eta_i) = 1 - \exp(-\exp(\eta_i))$$

  und für die Linkfunktion die Inverse

$$g(\pi_i) = \log(-\log(1 - \pi_i)).$$

  Im Vergleich zum Logit- und Probit-Modell wird dieses Modell in spezielleren Fällen angewendet, beispielsweise für die Modellierung und Analyse von zeitdiskreten Verweildauern (Fahrmeir et al., 2009).

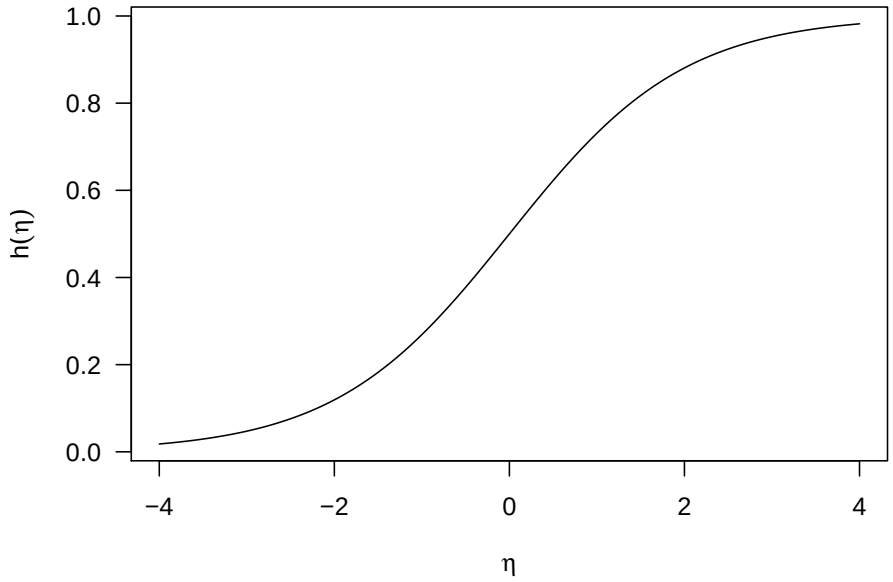

Abbildung 3.1.: Die Responsefunktion für das Logit-Modell.

3.3. Verknüpfung von BTL und GLM

Um das Bradley-Terry-Luce Modell mittels GLM zu schätzen, muss es zunächst in die Grundform eines GLM

$$g(\mu) = \eta = \boldsymbol{x}^\top \boldsymbol{\beta}$$

gebracht werden. Wird dabei für $g()$ die Logit-Linkfunktion verwendet und werden zudem die Gleichungen (2.3) und (2.4) betrachtet, welche bereits in Abschnitt 2.2 eingeführt wurden, ergibt sich das folgende Regressionsmodell:

$$\begin{aligned}\log\left(\frac{P(r \succ s \,|\, (r,s))}{P(s \succ r \,|\, (r,s))}\right) &= \gamma_r - \gamma_s \\ &= x_1^{(r,s)}\gamma_1 + \ldots + x_{m-1}^{(r,s)}\gamma_{m-1} \\ &= (\boldsymbol{x}^{(r,s)})^\top \boldsymbol{\gamma}\end{aligned}$$

Dabei steht $P(r \succ s \,|\, (r,s))$ für die Wahrscheinlichkeit, dass Objekt r gegenüber Objekt s in einem Paarvergleich der Kombination (r,s) präferiert wird, analog gilt dies für $P(s \succ r \,|\, (r,s))$. Es wird weiterhin davon ausgegangen, dass das Paarvergleichsexperiment m Objekte $\{a_1, \ldots, a_m\}$ beinhaltet, wobei von $i = 1, \ldots, n$ Versuchspersonen n unabhängige Vergleiche des Objektpaares (r,s) gemacht werden. Zudem ist $\gamma_r = \log(\pi_r)$ und π_r bezeichnet wieder die wahre Präferenz für ein Objekt auf einer subjektiven Skala. Für den Itemparametervektor gilt $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{m-1})^\top$. Zu beachten ist, dass die logarithmierte Präferenz für Objekt m aus Gründen der Identifizierbarkeit Null gesetzt wird: $\gamma_m = 0$.

Unter Einbeziehung einer Zufallsvariablen

$$Y_{(r,s)} = \begin{cases} 1, & r \succ s \\ 0, & s \succ r \end{cases}$$

lässt sich der Vergleich von Objekt r mit Objekt s folgendermaßen schreiben (Tutz, 2013):

$$\log\left(\frac{P(r \succ s \,|\, (r,s))}{P(s \succ r \,|\, (r,s))}\right) = \log\left(\frac{P(Y_{(r,s)} = 1)}{P(Y_{(r,s)} = 0)}\right).$$

Die Einträge der Designmatrix ergeben sich durch folgende Kodierung:

$$x_j^{(r,s)} = \begin{cases} 1, & j = r \\ -1, & j = s \\ 0, & \text{sonst.} \end{cases} \tag{3.2}$$

Gleichung (3.2) lässt sich ebenfalls darstellen durch

$$\boldsymbol{x}^{(r,s)} = \mathbf{1}_r - \mathbf{1}_s,$$

wobei $\mathbf{1}_r = (0, \ldots, 0, 1, 0, \ldots, 0)$ die Länge $m-1$ hat und eine 1 an der r-ten Stelle steht. Die Designmatrix besitzt folglich zeilenweise Einträge mit den Werte 1, -1 und 0, welche einen Paarvergleich für eine beurteilende Person i repräsentieren. Da für $\gamma_m = 0$ gilt, kommt der Wert -1 in den entsprechenden Zeilen der Designmatrix bei Paarvergleichen mit dem Objekt m nicht vor.

Bei Paarvergleichssystemen spielt die Anordnung der einzelnen Paarvergleiche für die Aufstellung der Designmatrix eine wichtige Rolle. Bereits im Zigarettenbeispiel in Tabelle 1.1 aus Abschnitt 1.1 ist eine bestimmte Struktur der einzelnen Paarvergleiche erkennbar, welche sich auf folgendes Muster zurückführen lässt:

$$1{:}2,\ 1{:}3,\ 2{:}3,\ 1{:}4,\ 2{:}4,\ 3{:}4,\ 1{:}5,\ 2{:}5,\ 3{:}5,\ 4{:}5,\ \ldots,\ 1{:}m,\ 2{:}m,\ \ldots,\ (m-1){:}m.$$

In Tabelle 3.2 wird diese Paarvergleichsstruktur noch einmal übersichtlicher dargestellt.

Vergleich mit j-tem Objekt						
Vergleich mit dem 2. Objekt	1:2					
Vergleich mit dem 3. Objekt	1:3	2:3				
Vergleich mit dem 4. Objekt	1:4	2:4	3:4			
Vergleich mit dem 5. Objekt	1:5	2:5	3:5	4:5		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		
Vergleich mit dem m. Objekt	1:m	2:m	3:m	4:m	$\ldots$	$(m-1)$:m

Tabelle 3.2.: Anordnungsstruktur eines Paarvergleichssystems mit $j = 1, \ldots, m$ zu vergleichenden Objekten.

Unter Berücksichtigung dieser Paarvergleichsstruktur und bei Annahme von $m = 6$ Objekten sieht eine beispielhafte Designmatrix für ein Paarvergleichssystem bei einer einzelnen beurteilenden Person wie in Abbildung 3.2 aus. Durch paarweises Vergleichen der $m = 6$ Items miteinander ergeben sich $I = \binom{m}{2} = 15$ Paarvergleichsurteile und dementsprechend besitzt die Designmatrix 15 Zeilen. Die Spaltenanzahl der Designmatrix beläuft sich auf $m - 1$. Zeilenweise gelesen befinden sich genau an

denjenigen Positionen Werte ungleich Null, deren korrespondierende Items miteinander verglichen werden. Ist in einer Zeile nur ein Wert ungleich Null vorhanden, handelt es sich um ein Paarvergleichsurteil mit dem m-ten Item. Somit steht jede Zeile für ein Paarvergleichsurteil einer einzelnen Person und jede Spalte für das j-te Item, mit Ausnahme von Item m.

$$\left.\begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}\right\} \text{Vergleich mit Item } m$$

Abbildung 3.2.: Darstellung einer Designmatrix für ein Paarvergleichssystem mit $m = 6$ Objekten bei nur einer einzelnen beurteilenden Person.

Allgemein sind in Paarvergleichssystemen jedoch nicht nur die Beurteilungen von nur einer einzelnen Person von Interesse, sondern die Beurteilungen von allen $i = 1, \ldots, n$ Versuchspersonen, so dass über die Struktur der Paarvergleiche eine eindeutige Rangordnung latenter Variablen erstellt werden kann. Es müssen somit die n-fachen Wiederholungen der einzelnen Paarvergleichsurteile berücksichtigt werden, so dass sich insgesamt $n \cdot \binom{m}{2}$ Paarvergleichsurteile und ebenso viele Zeilen einer Designmatrix ergeben. Die Struktur einer Designmatrix für ein Paarvergleichssystem von $m = 6$ Objekten, die von n Versuchspersonen beurteilt werden, ist in Abbildung 3.3 dargestellt.

3.4. Umsetzung und Erweiterung

Um die Möglichkeit der Schätzung eines Modells für Paarvergleichsdaten mittels generalisierter linearer Modelle zu überprüfen, lässt sich eine Simulation mit der

$$
\left.\begin{pmatrix}
1 & -1 & 0 & 0 & 0 \\
\vdots & & & & \vdots \\
1 & -1 & 0 & 0 & 0
\end{pmatrix}\right\} n\text{-mal Vergleich von Items 1 und 2 (1:2)}
$$

$$
\left.\begin{pmatrix}
1 & 0 & -1 & 0 & 0 \\
\vdots & & & & \vdots \\
1 & 0 & -1 & 0 & 0
\end{pmatrix}\right\} n\text{-mal Vergleich von Items 1 und 3 (1:3)}
$$

$$
\left.\begin{pmatrix}
\vdots & & \ddots & & \vdots \\
0 & 0 & 0 & 0 & 1 \\
\vdots & & & & \vdots \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}\right\} n\text{-mal Vergleich von Items } (m-1) \text{ und } m \text{ } ((m-1){:}m)
$$

Abbildung 3.3.: Strukturelle Darstellung einer Designmatrix für ein Paarvergleichssystem mit $m = 6$ Objekten bei n beurteilenden Personen.

Statistiksoftware `R`, Version 3.0.1 (R Development Core Team, 2013), durchführen. Das hierfür benötigte `R`-Paket `stats` (R Core Team and contributors worldwide, 2013) für grundlegende Statistik-Funktionalitäten wird üblicherweise bei Öffnen des Programms automatisch geladen und muss nicht extra aufgerufen werden. Für die Schätzung eines Bradley-Terry-Luce Modells mittels binärer Regression wird die Funktion `glm.fit` verwendet:

```
glm.fit(x, y, family = binomial(link = „logit"), intercept = FALSE)
```

Für das Argument `x` wird die Designmatrix herangezogen, welche die $n \cdot \binom{m}{2}$ Paarvergleiche enthält. Für das Argument `y` wird die binäre abhängige Variable Y verwendet, welche bei den einzelnen Paarvergleichsurteilen der Wahl für eines der beiden Objekte entspricht. Durch den Ausdruck `family = binomial(link = „logit")` wird für die Funktion bestimmt, dass es sich bei diesem GLM um ein logistisches Regressionsmodell handelt. Da bei Paarvergleichen die Designmatrix bereits den Intercept enthält, wird dieser in der Funktion `glm.fit` auf `FALSE` gesetzt. Mit dem Ausführen dieser Funktion werden Schätzer für die Koeffizienten des Itemparametervektors $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{m-1})^\top$ berechnet und auf Wunsch ausgegeben. Die „wahren“ Werte der Itemparameter wurden zuvor in der Simulationsmaske unter Zuhilfenahme eines Zufallsgenerators aus einer Gleichverteilung gezogen ($\gamma_m = 0$). Der zugehörige `R`-Code zur Simulation der Schätzung von BTL-Modellen mittels generalisierter linearer Regression ist unter Anhang A.1 aufgeführt.
Verschiedene Simulationsdurchläufe zeigen, dass generalisierte lineare Modelle eine gute Möglichkeit bieten, um Bradley-Terry-Luce Modelle zu schätzen. Die Anpas-

sung eines GLMs liefert sinnvolle Schätzer und interpretierbare Ergebnisse. Das GLM liefert zudem einen natürlichen Rahmen für die Modellierung einer Vielzahl von Paarvergleichsexperimenten (Critchlow and Fligner, 1991). Es lässt bei der Schätzung von BTL-Modellen ebenso die Berücksichtigung einer Kategorie „Unentschieden“ zu, wie auch die Einbindung von Kovariablen. Eine Erweiterung des BTL-Modells durch Einbindung subjektspezifischer Kovariablen dient der Berücksichtigung von Heterogenität der unterschiedlichen Versuchspersonen. Das Geschlecht oder auch das Alter der Versuchspersonen können einen Einfluss auf die Wahrnehmung von Objekten haben und somit bei den Paarvergleichsurteilen eine wichtige Rolle spielen. Sind mehrere Personen an einem Paarvergleichssystem beteiligt und beurteilen paarweise die Objekte ist es daher sinnvoll, den Einfluss der Probanden in geeigneter Weise zu berücksichtigen. Das BTL-Modell wird im folgenden Kapitel um einen multiplikativen Parameter für die Heterogenität der Versuchspersonen erweitert. Aufgrund der guten Anpassungsfähigkeit des generalisierten linearen Modells wird es wiederum zur Schätzung des erweiterten personenspezifischen BTL-Modells verwendet.

4. Personenspezifische Paarvergleichsmodelle

Bisher betrachtet wurde das einfache BTL-Modell (Bradley and Terry, 1952), bei dem beliebige Objekte paarweise miteinander verglichen werden. Das Ergebnis des Paarvergleichs ist eine Relation, d.h. das eine Objekt wird dem anderen Objekt vorgezogen, ein Unentschieden wird dabei ausgeschlossen. Für Bradley-Terry-Luce Modelle existieren jedoch viele Erweiterungen. Paarvergleichsmodelle mit einer Kategorie der Entscheidungsenthaltung bzw. eines Unentschiedens wurden von Glenn and David (1960) vorgeschlagen und von Rao and Kupper (1967) und Davidson (1970) speziell für BTL-Modelle diskutiert. Darüber hinaus betrachtete Tutz (1986) die Verallgemeinerung auf mehrstufige kategoriale Entscheidungen, das ordinale BTL-Modell mit k Kategorien (kurz BTL(k)-Modell). Ein Modell für die Berücksichtigung der Reihenfolge bei Paarvergleichen wurde außerdem von Davidson and Beaver (1977) formuliert und von Koehler and Ridpath (1982) zur Analyse der Spielergebnisse professioneller Basketballmannschaften bzw. von Fienberg (1980, Kapitel 8) und Agresti (2002) zur Analyse von Baseballergebnissen verwendet, wobei der Effekt des Heimvorteils einer Mannschaft berücksichtigt wurde. Es kann zudem sinnvoll sein, Kovariablen in das BTL-Modell aufzunehmen, wenn beispielsweise angenommen wird, dass die Präferenzentscheidungen nicht allein von den Eigenschaften der zu beurteilenden Items abhängt, sondern auch von den Charakteristika bezüglich der Versuchspersonen selbst (Strobl et al., 2011). Ansätze für die Einbindung von Kovariablen geben Dittrich et al. (1998, 2001) und Böckenholt (2001a,b).

Im Folgenden wird eine Erweiterung des Bradley-Terry-Luce Modells betrachtet, bei der die Variabilität der Versuchspersonen im Vordergrund steht. Für ein Paarvergleichssystem wurde bisher angenommen, dass Versuchswiederholungen mit identischer Verteilung realisierbar sind. Liegt jedoch eine Objektmenge in der Art und Weise vor, dass einem Probanden dieselben Paare wiederholt zum Vergleich vor-

gelegt werden können, lässt sich eine Aussage über die Reizqualität der Items für eine Person machen. Häufig lassen sich jedoch solche wiederholten Paarvergleiche an einem Probanden nicht sinnvoll durchführen, da sich die Versuchsperson an seine vorhergehende Entscheidung erinnert oder durch die häufigen Wiederholungen des Experiments ermüdet. Werden jedoch mehrere Personen für das Experiment herangezogen, stellt die Annahme von Unabhängigkeit für die Versuchswiederholungen ein Problem dar. Zudem steht häufig die Erlangung von Informationen über die Reizqualitäten von Objekten, die mehreren Personen vorgelegt werden, im Vordergrund. Bei Paarvergleichen, die mehreren Personen vorgelegt werden, ist es daher sinnvoll, den Einfluss des Probanden in geeigneter Weise zu berücksichtigen. Wir sprechen daher im Folgenden von einem *personenspezifischen Paarvergleichssystem* (Tutz, 1989).

4.1. Berücksichtigung von Heterogenität

Das folgende Modell erweitert das bereits bekannte BTL-Modell aus Abschnitt 2 um einen personenspezifischen Parameter α_i:

$$p_{irs} = P(Y_{(r,s)} = 1 \mid (r,s), i) = F(\alpha_i(\gamma_r - \gamma_s)). \tag{4.1}$$

Wird für F wieder die logistische Verteilung gefordert, ergibt sich das Modell

$$p_{irs} = P(Y_{(r,s)} = 1 \mid (r,s), i) = \frac{\exp(\alpha_i(\gamma_r - \gamma_s))}{1 + \exp(\alpha_i(\gamma_r - \gamma_s))}. \tag{4.2}$$

Der Parameter α_i ist der Faktor für die Heterogenität, d.h. er steht für die individuelle Wahrnehmung der Reizstärke der Objekte einer Versuchsperson und wird auch *Personenparameter* genannt. Bereits bei kleiner Differenz $\gamma_r - \gamma_s > 0$ wird für großes α_i die Wahrscheinlichkeit, Objekt r zu bevorzugen, sehr groß. Ist α_i hingegen sehr klein ($\alpha_i \to 0$), konvergiert die Wahrscheinlichkeit für die Präferenz von Item r gegenüber Item s auch für große Differenzen $\gamma_r - \gamma_s$ gegen die Ratewahrscheinlichkeit. Durch α_i wird somit die *Diskriminationsfähigkeit* der i-ten Person ausgedrückt, die für großes α_i ebenfalls groß ist und zur Ratewahrscheinlichkeit wird, wenn $\alpha_i \to 0$. Modell (4.1) stellt für jede einzelne Versuchsperson $i = 1, \ldots, n$ ein BTL-Modell dar, da immer wenn nur eine Versuchsperson betrachtet wird, der Personenparameter α_i den Wert 1 annimmt (Tutz, 1989).

Die Verdeutlichung der Notwendigkeit der Einbeziehung eines Parameters für die Heterogenität der Versuchspersonen in das BTL-Modell erfolgt zunächst anhand einer Datensimulation, der ein *datengenerierender Prozess* (*Data Generating Process, DGP*) zugrunde liegt. In diesem werden neben der Definition der relevanten Größen für die Simulation eines Paarvergleichsmodells die wahren Werte für die Itemparameter aus einer Gleichverteilung gezogen. Außerdem wird für den Personenparameter α_i die Normalverteilung angenommen. Dem datengenerierenden Prozess wird der Name „DGP(α_i)“ gegeben.

Datengenerierender Prozess DGP($\boldsymbol{\alpha_i}$)

$$m = 8 \text{ Items}, \; n = 40 \text{ Personen}, \; I = 28 \text{ Paarvergleiche} \tag{4.3}$$

$$\alpha_i \sim N(1, \sigma^2), \quad i = 1, \ldots, n \tag{4.4}$$

$$\sigma \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\} \tag{4.5}$$

$$\gamma_r \sim U[-3, 3], \quad r = 1, \ldots, m-1, \quad \gamma_m = 0 \tag{4.6}$$

Die Verteilungsannahme (4.4) impliziert, dass α_i auch negative Werte annehmen kann. Ein negatives α_i bedeutet hier, dass die i-te Person Objekt s gegenüber Objekt r präferiert, wenn eine Person mit positivem α_i Objekt r gegenüber Objekt s präferiert. Die Skalenwerte vertauschen dadurch ihre Reihenfolge (Tutz, 1989).
Die steigenden Werte für die Varianz bzw. für die Standardabweichung σ des Personenparameters α_i aus Prozessannahme (4.5) stehen für die zunehmende Unsicherheit in den Daten und simulieren größere Unterschiede der individuellen Wahrnehmung der Reizstärken der Objekte in einem Paarvergleich.
Die Schätzung der Itemparameter erfolgt in der Simulation mit zugrunde liegendem DGP(α_i) gemäß dem einfachen BTL-Modell mittels generalisierter linearer Regression. Der zugehörige `R`-Code zur Simulation mittels DGP(α_i) ist in Anhang A.2 aufgeführt.

Als ein generelles Maß zur Beurteilung der Güte der Schätzung wird der *mittlere quadratische Fehler (Mean Squared Error, MSE)* betrachtet. Dieser ist gegeben durch:

$$\begin{aligned}\mathrm{MSE}(\hat{\Theta}_n) &:= \mathrm{E}\Big[(\hat{\Theta}_n - \Theta)^2\Big] \\ &= \underbrace{\mathrm{E}\Big[\{\hat{\Theta}_n - \mathrm{E}(\hat{\Theta}_n)\}^2\Big]}_{} + \underbrace{\Big\{\mathrm{E}(\hat{\Theta}_n) - \Theta)\Big\}^2}_{} \\ &= \qquad \mathrm{Var}(\hat{\Theta}_n) \qquad + \qquad \mathrm{Bias}^2(\hat{\Theta}_n),\end{aligned}$$

wobei $\hat{\Theta}_n = T(X_1, \ldots, X_n)$ die Schätzfunktion für den Parameter Θ ist (Rinne, 2008). Der MSE gibt gemäß Definition wieder, welche Abweichung zwischen Schätzfunktion $\hat{\Theta}_n$ und wahrem Wert Θ für die Schätzfunktion $\hat{\Theta}_n$ zu erwarten ist. Er lässt sich zudem umformen und darstellen als Summe aus der Varianz von $\hat{\Theta}_n$ und dem quadrierten Bias (Fahrmeir et al., 2011).

Abbildung 4.1 veranschaulicht den mittleren quadratischen Fehler für die Schätzer der Itemparameter bei steigender Standardabweichung des Personenparameters α_i. Bei der zugehörigen Simulation mittels DGP(α_i) für diese Abbildung wurden 100 Simulationsdurchläufe angenommen. In Abbildung 4.1 sind nur die Graphen der MSE von sieben Itemparameterschätzern abgebildet, da das Referenzitem Null gesetzt wird ($\gamma_m = 0$).

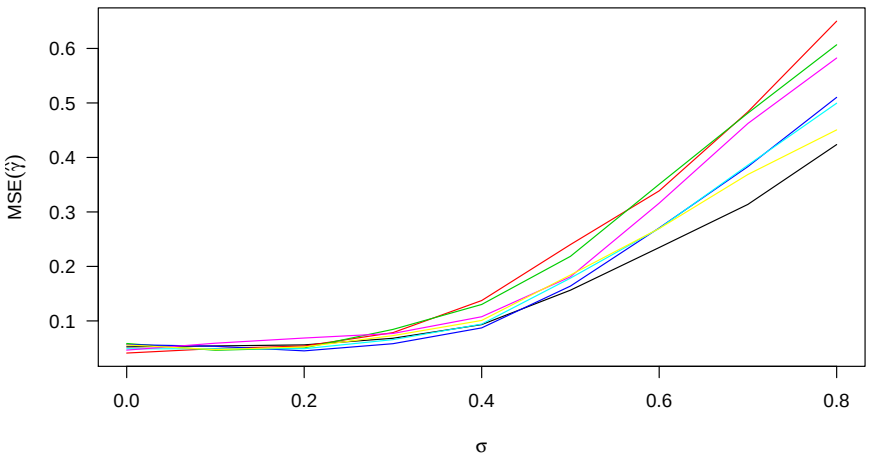

Abbildung 4.1.: Darstellung des MSE der Itemparameterschätzer für die Schätzung eines BTL-Modells in der Simulation mit zugrunde liegendem DGP(α_i). Es wurden 8 zu vergleichende Items, 40 Personen und 100 Simulationsdurchläufe angenommen.

Für $\sigma = 0$ ist $\alpha_i = 1$ und es ergeben sich dieselben Schätzer und somit derselbe MSE wie für die Simulation ohne α_i, also für das einfache BTL-Modell (vergleiche dazu Abschnitt 3.4). Die Unterschiede zwischen zwei zu vergleichenden Objekten werden von allen Versuchspersonen gleich wahrgenommen. Mit steigender Standardabweichung des Parameters, also mit zunehmender Veränderung der Wahrnehmung der Versuchspersonen für die Reizstärke eines Objekts, steigt auch der MSE an und es kommt zu einer schlechteren Präzision der Schätzung. Die Itemparameter $\hat{\boldsymbol{\gamma}}$ werden verzerrt geschätzt und somit entstehen größere Abweichungen von den wahren Werten der Itemparameter $\boldsymbol{\gamma}$.

Anhand der Simulation und der Entwicklung des MSE bei steigender Standardabweichung wird deutlich, dass die Heterogenität der Versuchspersonen in geeigneter Weise bei der Schätzung des BTL-Modells berücksichtigt werden sollte. Durch ein Modell, basierend auf Gleichung (4.1), in dem für jede Versuchperson $i = 1, \ldots, n$ die Schätzung des Personenparameters α_i zusätzlich zur Schätzung der Itemparameter durchgeführt werden soll, wird eine präzisere Schätzung des BTL-Modells und somit eine Verringerung des MSE der Itemparameterschätzer angestrebt. Dieses Modell wird im Folgenden als *Heterogenitätsmodell* bezeichnet. Im nächsten Abschnitt wird die Berechnung dieses Modells genauer dargelegt.

4.2. Schätzverfahren für das Heterogenitätsmodell

Um den Einfluss einer Versuchsperson in geeigneter Weise zu berücksichtigen wird nun das durch einen Personenparameter $\tilde{\alpha}_i$ erweiterte BTL-Modell

$$p_{irs} = P(Y_{(r,s)} = 1 \,|\, (r,s), i) = F(\tilde{\alpha}_i(\gamma_r - \gamma_s)) \tag{4.7}$$

betrachtet (vgl. Gleichung (4.1)). Die zugrunde liegende Verteilungsannahme $\tilde{\alpha}_i \sim N(1, \sigma^2)$ ist analog zu Gleichung (4.4). Unter der Annahme von

$$\tilde{\alpha}_i = 1 + \alpha_i$$

mit entsprechender Verteilungsfunktion

$$\alpha_i \sim N(0, \sigma^2)$$

lässt sich Gleichung (4.7) umformulieren in

$$p_{irs} = P(Y_{(r,s)} = 1 \,|\, (r,s), i) = F((1 + \alpha_i)(\gamma_r - \gamma_s)). \tag{4.8}$$

Sind die Personenparameter α_i alle identisch, liegt Homogenität der Versuchspersonen vor und das Heterogenitätsmodell ist äquivalent zum einfachen BTL-Modell.

Die Schätzung der Itemparameter $\boldsymbol{\gamma}$ wird wie zuvor mittels GLM (siehe Kapitel 3), die Schätzung der Personenparameter α_i über generalisierte lineare gemischte Modelle (*Generalized Linear Mixed Models, GLMM*) durchgeführt.
Wie bereits am Anfang dieses Kapitels erwähnt, stellt die Annahme von Unabhängigkeit bei Versuchswiederholungen an mehreren Personen in einem Paarvergleichsexperiment ein Problem dar. Die Lösung des Problems liegt hier in der Schätzung der Personenparameter mittels gemischter Modelle. Diese beziehen in den Prädikor $\eta = \boldsymbol{x}^\top \boldsymbol{\beta}$ linearer und generalisierter linearer Modelle neben den bislang betrachteten festen Effekten auch zufällige Effekte oder Koeffizienten ein. Es wird deshalb auch von Modellen mit zufälligen Effekten (*random effects models*) gesprochen (Fahrmeir et al., 2009). Eine Definition von generalisierten linearen gemischten Modellen ist in Abschnitt 4.3 zu finden.

Ein mögliches Verfahren zur Schätzung des Heterogenitätsmodells mit zugrunde liegender Gleichung (4.8) ist somit der folgende Algorithmus:

Algorithmus zur Schätzung des Heterogenitätsmodells

---

(1) Starte mit der Schätzung eines einfachen BTL-Modells, d.h. $\alpha_i = 0$ bzw. $\tilde{\alpha}_i = 1$, und ermittle $\hat{\boldsymbol{\gamma}}$ mittels GLM:

$$\begin{aligned} \log\left(\frac{P(Y_{(r,s)} = 1 \,|\, (r,s), i)}{P(Y_{(r,s)} = 0 \,|\, (r,s), i)}\right) = \eta_{irs} \; &= \tilde{\alpha}_i \, (\gamma_r - \gamma_s) \\ &= \tilde{\alpha}_i \, (\boldsymbol{x}^{(r,s)})^\top \boldsymbol{\gamma} \\ &= (\boldsymbol{x}^{(r,s)})^\top \boldsymbol{\gamma}. \end{aligned} \tag{4.9}$$

(2) Fitte ein random effects model unter Einbeziehung der bereits geschätzten Itemparameter $\hat{\boldsymbol{\gamma}}$ und erhalte Schätzer für die zufälligen Effekte $\hat{\alpha}_i$ und die zugehörige

geschätzte Standardabweichung $\hat{\sigma}$:

$$\log\left(\frac{P(Y_{(r,s)}=1\,|\,\hat{\gamma}_r,\hat{\gamma}_s,i)}{P(Y_{(r,s)}=0\,|\,\hat{\gamma}_r,\hat{\gamma}_s,i)}\right)=\eta_{irs} \;=\tilde{\alpha}_i\,(\hat{\gamma}_r-\hat{\gamma}_s)$$

$$=(1+\alpha_i)\,(\hat{\gamma}_r-\hat{\gamma}_s)$$

$$=\hat{\gamma}_r-\hat{\gamma}_s+\alpha_i\,(\hat{\gamma}_r-\hat{\gamma}_s) \tag{4.10}$$

$$=(\boldsymbol{x}^{(r,s)})^\top\hat{\boldsymbol{\gamma}}+\alpha_i\,(\boldsymbol{x}^{(r,s)})^\top\hat{\boldsymbol{\gamma}} \tag{4.11}$$

(3) Fitte ein personenspezifisches BTL-Modell unter Einbeziehung der in Schritt (2) geschätzen Personenparameter $\hat{\alpha}_i$ und ermittle neue Schätzer für die Itemparameter $\hat{\boldsymbol{\gamma}}$ über GLM:

$$\log\left(\frac{P(Y_{(r,s)}=1\,|\,(r,s),\hat{\alpha}_i)}{P(Y_{(r,s)}=0\,|\,(r,s),\hat{\alpha}_i)}\right)=\eta_{irs} \;=(1+\hat{\alpha}_i)\,(\gamma_r-\gamma_s)$$

$$=(1+\hat{\alpha}_i)\,(\boldsymbol{x}^{(r,s)})^\top\boldsymbol{\gamma} \tag{4.12}$$

(4) Iteriere Schritte (2) und (3) bis die Veränderung zum vorangegangen Rechenschritt kleiner ist als ein vorher spezifizierter Grenzwert ε. Analog zur Newton-Raphson Methode in Tutz (2012) wird die Iteration nach dem k-ten Schritt gestoppt, falls

$$\|\hat{\boldsymbol{\gamma}}^{(k)}-\hat{\boldsymbol{\gamma}}^{(k-1)}\|\;/\;\|\hat{\boldsymbol{\gamma}}^{(k-1)}\|\;<\;\varepsilon.$$

Die mit der Konvergenz einhergehenden Schätzer $\hat{\boldsymbol{\gamma}}^{(k)}$ und $\hat{\alpha}_i^{(k)}$ sind die geschätzten Item- und Personenparameter des Heterogenitätsmodells.

Anhand einer Datensimulation in `R` wird dieser Algorithmus in Abschnitt 4.4 näher erläutert.

4.3. Definition und Eigenschaften von GLMMs

In Abschnitt 3.1 wurde ein GLM folgendermaßen definiert: Die bedingte Dichte von y_i, gegeben der lineare Prädiktor $\eta_i=\boldsymbol{x}_i^\top\boldsymbol{\beta}$, gehört zur Klasse der Exponentialfamilien und der lineare Prädiktor η_i ist mit dem bedingten Erwartungswert $\mu_i=E(y_i\,|\,\boldsymbol{x}_i)$

durch die Beziehung

$$\mu_i = h(\eta_i) \quad \text{bzw.} \quad \eta_i = g(\mu_i)$$

verknüpft. Für die Definition von GLMMs wird der lineare Prädiktor $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ eines GLMs durch die Hinzunahme zufälliger Effekte erweitert.
Die Zielvariablen $y_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i$, mit n_i Messwiederholungen pro Individuum i, können nun zum Beispiel binäre Variablen oder Zählvariablen sein. Mit individuenspezifischen zufälligen Effekten $\boldsymbol{b_i}$ ist dann $y_{ij} \,|\, \boldsymbol{b_i}$ zum Beispiel binär, Binomial- oder Poisson-verteilt. Der bedingte Erwartungswert $\mu_{ij} = E(y_{ij} \,|\, \boldsymbol{b}_i)$ ist mit einem linearen (gemischten) Prädiktor

$$\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n_i,$$

durch $\mu_{ij} = h(\eta_{ij})$ mit einer geeigneten Responsefunktion h verbunden.
Für binäre Zielvariablen ergeben sich somit Logit-Modelle mit zufälligen Effekten durch

$$\log \frac{P(y_{ij} = 1 \,|\, \boldsymbol{b}_i)}{P(y_{ij} = 0 \,|\, \boldsymbol{b}_i)} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i$$

und für Zählvariablen $y_{ij} \sim \text{Po}(\lambda_{ij} \,|\, \boldsymbol{b}_i)$ log-lineare Modelle durch

$$\log(\lambda_{ij}) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i.$$

Durch geeignete Wahl der Designmatrizen $\boldsymbol{X}$ und $\boldsymbol{Z}$ kann man den Prädiktorvektor $\boldsymbol{\eta}$ in der Form

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$$

schreiben. Es wird angenommen, dass $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ unabhängig und identisch

$$\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$$

verteilt sind und somit $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G})$ mit $\boldsymbol{G} = \text{diag}(\boldsymbol{D}, \ldots, \boldsymbol{D})$ gilt.

Zusammenfassend gelten also für GLMMs gemäß Fahrmeir et al. (2009) folgende Punkte:

1. *Verteilungsannahme:* Gegeben die zufälligen Effekte $\boldsymbol{b}$ und die Kovariablen $\boldsymbol{x}_i$, sind die Zielvariablen y_i bedingt unabhängig und die bedingte Dichte $f(y_i \,|\, \boldsymbol{b})$ gehört zu einer Exponentialfamilie wie im Abschnitt 3.1 für GLMs.

2. *Strukturannahme:* Der bedingte Erwartungswert $\mu_i = E(y_i \mid \boldsymbol{b})$ ist mit dem linearen Prädiktor

$$\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{b}$$

durch

$$\mu_i = h(\eta_i) \quad \text{bzw.} \quad \eta_i = g(\mu_i)$$

verknüpft, wobei h die Responsefunktion und $g = h^{-1}$ die Linkfunktion ist.

3. Verteilungsannahme für die zufälligen Effekte $\boldsymbol{b} = (b_1, \ldots, b_n)$:

$$\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}),$$

mit positiv definiter Kovarianzmatrix $\boldsymbol{G}$.

Für eine detailliertere Einführung in (lineare) gemischte Modelle sei an dieser Stelle auf Laird and Ware (1982), Diggle et al. (2002) und Verbeke and Molenberghs (2009) verwiesen.

Das im Algorithmus zur Schätzung des Heterogenitätsmodells unter Schritt (2) zu fittende random effects model (siehe Abschnitt 4.2) ist ein reines random slope model. Aufgrund der bereits in Schritt (1) bzw. Schritt (3) ermittelten Itemparameter $\hat{\boldsymbol{\gamma}}$, ist der Term $\hat{\gamma}_r - \hat{\gamma}_s$ bekannt und kann daher als Offset in das Modell aufgenommen werden. Somit beinhaltet Gleichung (4.10) weder feste Effekte noch einen random intercept und dient lediglich zur Schätzung der random effects, also der Personenparameter α_i.

4.4. Umsetzung des Algorithmus zur Schätzung des Heterogenitätsmodells

Aufbauend auf der Datensimulation aus Abschnitt 4.1 werden in diesem Abschnitt dieselben Daten simuliert und der Algorithmus zur Schätzung des Heterogenitätsmodells darauf angewendet. Im anschließenden Abschnitt 4.5 wird die Güte der Schätzung des BTL-Modells mit zugrunde liegendem DGP(α_i) mit der des Heterogenitätsmodells verglichen.

Ebenso wie die Schätzung des BTL-Modells in der Simulation mit DGP(α_i) werden zunächst die wahren Werte für die Itemparameter aus einer Gleichverteilung gezogen, während für die Personenparameter eine Normalverteilung angenommen wird, d.h.

$$\tilde{\alpha}_i \sim N(1, \sigma^2)$$

bzw. ergibt sich mit $\tilde{\alpha}_i = 1 + \alpha_i$

$$\alpha_i \sim N(0, \sigma^2).$$

Diese Verteilungsannahmen implizieren, dass $\tilde{\alpha}_i$ bzw. α_i auch negative Werte annehmen können. Mithilfe des Setzens derselben Seeds wie im DGP(α_i) wird das Ziehen gleicher Zufallszahlen für die Simulationen der Schätzungen des einfachen BTL- und des Heterogenitätsmodells sowie ihr späterer Vergleich miteinander sichergestellt. Ein Seed ist eine beliebige ganze Zahl, welche dazu genutzt wird, einen Pseudozufallszahlengenerator zu initialisieren. Auf diese Weise ist es möglich, dieselbe Folge von Pseudozufallszahlen zu erzeugen und somit reproduzierbare Ergebnisse zu erhalten. In `R` erfolgt dies mit der Funktion `set.seed` (Ligges, 2008).

Ist die Anzahl der zu durchlaufenden Operationen wie in Schritt (4) des Algorithmus zur Schätzung des Heterogenitätsmodells vorher nicht bekannt, z.B. soll ein iterativer Algorithmus zur Maximierung einer Likelihood solange durchgeführt werden, bis eine gewisse Konvergenzgenauigkeit erreicht ist, kann in `R` die `while`-Schleife verwendet werden. Im hier mit `R` zu schätzenden Heterogenitätsmodell werden die Algorithmusschritte (1) bis (3) in eine solche Iterationsschleife eingebaut und die Schätzung des Modells durchgeführt. Der zugehörige `R`-Code zur Schätzung des Heterogenitätsmodells ist in Anhang A.3 aufgeführt.
Für die Schritte (1) und (3) im Algorithmus zur Schätzung des Heterogenitätsmodells erfolgt die Schätzung der Itemparameter $\boldsymbol{\gamma}$ wie im einfachen BTL-Modell mittels GLM (vgl. auch Abschnitt 3.4). Zur Berücksichtigung der Heterogenität in Gleichung (4.12) ist es nötig, die Personenparameter $\hat{\tilde{\alpha}}_i$ mit der Designmatrix $\boldsymbol{X}$ zu multiplizieren und anschließend die Itemparameter zu bestimmen. Daher werden die sich aus der Schätzung eines GLMMs in Schritt (2) ergebenden zufälligen Effekte $\hat{\alpha}_i$ für die Ausführung von Schritt (3) mit 1 addiert und mit der Designmatrix multipliziert. Für Schritt (1) beträgt $\hat{\alpha}_i = 0$ bzw. $\hat{\tilde{\alpha}}_i = 1$.
Für das Schätzen der random effects models, genauer gesagt der random slope mo-

dels, (Schritt (2) des Algorithmus) ist das R-Paket `lme4` (Bates et al., 2013) notwendig.[1] Zur Schätzung des GLMMs wird die Funktion `glmer` verwendet:

```
glmer(formula, family = binomial, offset)
```

mit der folgenden Zuweisung für `formula`:

```
formula <- y ~ 0 + X.noise + (0 + offset | ID).
```

Für das Argument `y` wird analog zum GLM die binäre abhängige Variable Y verwendet, welche bei den einzelnen Paarvergleichsurteilen der Wahl für eines der beiden Objekte entspricht. Die Formelbestandteile nach der Tilde stehen sowohl für die fixen als auch für die zufälligen Effekte eines GLMMs. Da hier wie bereits erwähnt ein random slope model geschätzt wird, beinhaltet das Modell keinerlei fixe Effekte. Auffällig ist daher zunächst das Argument `X.noise`. Dieses Argument ist lediglich eine Zufallsvariable, welche völlig unabhängig vom Response gebildet wird und nur der Sicherstellung der Funktionsfähigkeit der `glmer`-Funktion in `R` dient. Dass das vorliegende Modell auch weder einen fixen noch einen zufälligen Intercept besitzt, wird dargestellt durch die beiden Nullen in `formula`.
Mit dem GLMM soll folgendes Modell gefittet werden (vgl. auch Gleichung (4.10)):

$$\eta_{irs} = \hat{\gamma}_r - \hat{\gamma}_s + \alpha_i \, (\hat{\gamma}_r - \hat{\gamma}_s).$$

Der Term $\hat{\gamma}_r - \hat{\gamma}_s$ ist durch die im Algorithmus in Schritt (1) bzw. (3) geschätzten Itemparameter $\hat{\boldsymbol{\gamma}}$ bekannt und kann daher einfach als bekannter konstanter additiver Term im linearen Prädiktor angesehen werden. `R` handhabt einen solchen Term im linearen Prädiktor, der keine zu schätzenden unbekannten Parameter enthält, als „Offset“. Das Argument `offset` in der `glmer`-Funktion ist somit definiert durch die Matrixmultiplikation der Designmatrix $\boldsymbol{X}$, welche die $n \cdot \binom{m}{2}$ Paarvergleiche enthält, mit dem Vektor der geschätzten Itemparameter $\hat{\boldsymbol{\gamma}}$.
Allgemein muss für die Anwendung der Funktion `glmer` eine Variable für die Personenidentität erstellt werden, welche als Gruppierungsfaktor dient. In diesem Fall wird eine Variable `ID` definiert, welche die n Personen jeweils für die Anzahl der Paarvergleiche wiederholt und somit eine Länge von $n \cdot \binom{m}{2}$ besitzt. Die Terme der

[1] An dieser Stelle sei explizit darauf hingewiesen, dass für die vorliegende Arbeit mit der Paketversion 0.999999-2 gearbeitet wurde, bei Abgabe der Arbeit jedoch bereits eine neuere Paketversion zur Verfügung stand und somit der für diese Arbeit relevante R-Code bei Nutzung einer neueren Paketversion gegebenenfalls angepasst werden muss.

zufälligen Effekte werden im `formula`-Argument der `glmer`-Funktion mittels vertikaler Striche (|) von den Termen der fixen Effekte unterschieden. Sie trennen die Ausdrücke für die Designmatrizen von den Gruppierungsfaktoren.

Im vorliegenden Fall werden die Schätzungen des GLMs und des GLMMs solange durchgeführt bis im k-ten Iterationsschritt die Ungleichung

$$\|\hat{\boldsymbol{\gamma}}^{(k)} - \hat{\boldsymbol{\gamma}}^{(k-1)}\| \;/\; \|\hat{\boldsymbol{\gamma}}^{(k-1)}\| \;<\; \varepsilon$$

gilt (Tutz, 2012). Der vor Beginn der Iterationsschleife festzulegende Grenzwert ε wird hier auf den Wert 0.0001 spezifiziert.

4.5. Simulationsergebnisse

Steigende Werte für die Varianz bzw. für die Standardabweichung des Personenparameters stehen für die zunehmende Unsicherheit in den Daten und simulieren größere Unterschiede der individuellen Wahrnehmung der Reizstärken der Objekte in einem Paarvergleich. Die Heterogenität im Heterogenitätsmodell wird also durch unterschiedliche Werte der Standardabweichung der Personenparameter simuliert. Mit dem für das Heterogenitätsmodell geschriebenem `R`-Code lässt sich jeweils ein Heterogenitätsmodell für einen vorgegebenen Wert der Standardabweichung des Personenparameters schätzen. Die Schätzung des Heterogenitätsmodells wird daher für verschiedene Werte der Standardabweichung des Personenparameters durchgeführt. Ebenso wie in Annahme (4.5) im datengenerierenden Prozess DGP(α_i) in Abschnitt 4.1 werden für die Darstellung der Heterogenität folgende Werte für die Standardabweichung des Personenparameters verwendet:

$$\sigma \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}.$$

Mit dem Wert $\sigma = 0$ bestehen keine Unterschiede der individuellen Wahrnehmung der Objekt-Reizstärken in einem Paarvergleich, für $\sigma = 0.8$ liegen große Unterschiede in der individuellen Wahrnehmung vor, d.h. während eine Person Objekt r gegenüber Objekt s bevorzugt, kann eine andere Person bezüglich dieser zwei Objekte umgekehrte Präferenzen besitzen. Für Modelle mit hoher Standardabweichung wird es also umso schwieriger, die Personen- und Itemparameter des Heterogeni-

tätsmodells zu schätzen. Dieses Problem wird sichtbar in der fallweise fehlenden Konvergenz innerhalb der Iterationsschleife während der Modellschätzung. Da während der Simulation des Heterogenitätsmodells mit zunehmender Standardabweichung des Personenparameters die Anzahl an Iterationen bis zur Konvergenz der Schleife zunimmt und zudem Fälle auftreten, in denen einzelne Iterationsschleifen nicht konvergieren, wird für jede Iterationsschleife ein Maximum von 100 Iterationen festgelegt und die `while`-Schleife so zur Konvergenz gezwungen. Im Anschluss an die P Simulationsdurchläufe werden für die Berechnungen aller Statistiken und die grafische Darstellung der Ergebnisse diejenigen Simulationsdurchläufe entfernt, bei denen die 100 Iterationen erreicht wurden und die Schleife somit nicht konvergiert, um mögliche Verzerrungen der Ergebnisse zu umgehen. Tabelle 4.1 gibt für beispielhafte Seeds die Häufigkeiten an, für welche Standardabweichungswerte die Iterationsschleife der Heterogenitätsmodelle nicht konvergiert.

		$\boldsymbol{\sigma}$								
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Seed	1	1	0	0	1	0	1	2	9	15
	160	0	0	0	0	0	3	1	6	19
	2911	0	0	0	0	1	0	2	7	19

Tabelle 4.1.: Angabe der Häufigkeiten der Nicht-Konvergenz von Simulationen des Heterogenitätsmodells bei je 100 Simulationsdurchläufen.

Die Iterationen der `while`-Schleife bis hin zur Konvergenz und die somit wiederholten Schätzungen der Personen- und Itemparameter dienen der Festigung des Algorithmus. Nach Konvergenz der Schleife wird von präzisen Schätzern für die Personen- und Itemparameter ausgegangen. Um diese Annahme zu überprüfen, wird sowohl für die geschätzten Personenparameter $\hat{\alpha}_i$ als auch für die geschätzten Itemparameter $\hat{\boldsymbol{\gamma}}$ jeder Iteration in jeder Wiederholung des Simulationsszenarios der mittlere quadratische Fehler (MSE) berechnet. Eine abnehmende Folge der MSE-Werte innerhalb eines Simulationsdurchlaufes würde für präzise Schätzungen der Personen- und Itemparameter sprechen.
Abbildung 4.2 stellt die Entwicklung des MSE von Personen- und Itemparameterschätzern für die Schätzung eines Heterogenitätsmodells bei einer Standardabweichung des Personenparameters von $\sigma = 0.4$, 40 Personen und acht zu vergleichenden Objekten dar. Der Übersicht halber wurden für Abbildung 4.2 nur 15 Simulationsdurchläufe angenommen. Sowohl für $\hat{\alpha}_i$ als auch für $\hat{\boldsymbol{\gamma}}$ ist in den Grafiken der oberen

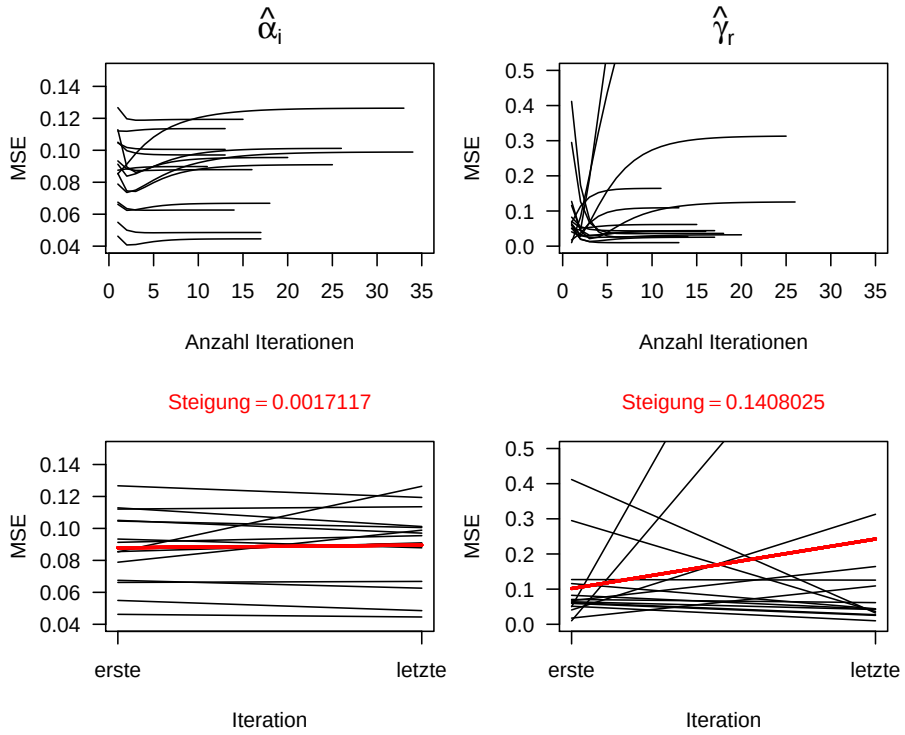

Abbildung 4.2.: Darstellung der MSE-Entwicklung der Parameterschätzer innerhalb der Iterationsschleife für die Simulation eines Heterogenitätsmodells. Eine Kurve entspricht der MSE-Entwicklung des jeweiligen Schätzers $\hat{\alpha}_i$, $i = 1, \ldots, n$ bzw. $\hat{\gamma}_r$, $r = 1, \ldots, m-1$ von der ersten bis zur letzten Iteration innerhalb der Schleife für einen Simulationsdurchlauf. Die rot hervorgehobene Gerade beschreibt den Verlauf der MSE-Mittelwerte der ersten und letzten Iterationen der Simulationsdurchläufe. Für die zugrunde liegende Simulation wurden 8 zu vergleichende Items, 40 Personen, 15 Simulationsdurchläufe und eine wahre Standardabweichung von $\sigma = 0.4$ angenommen.

Reihe ein anfängliches Sinken des MSE vieler Simulationsdurchläufe während der ersten Iterationen erkennbar. Während der MSE von $\hat{\alpha}_i$ in den meisten Simulationsdurchläufen anschließend wieder steigt und den MSE-Wert der ersten Iteration in den folgenden nicht oder kaum übersteigt, sind auch Simulationsdurchläufe mit von Beginn an extrem steigenden MSE-Werten zu beobachten. Bei Betrachtung des MSE von $\hat{\gamma}$ ist ebenso ein anfängliches Fallen der Werte zu erkennen, bei etwa der Hälfte der Fälle steigt im Anschluss der MSE, verbleibt jedoch unter dem Wert der ersten Iteration. In allen anderen Fällen sind extremere Verläufe erkennbar, wie von Beginn an starkes Fallen oder Steigen des MSE. Die Grafiken der unteren Reihe stel-

len lediglich den Unterschied der MSE-Werte zwischen der ersten und der letzten Iteration der einzelnen Simulationsdurchläufe dar und sollen insgesamt das Steigen bzw. Fallen der MSE-Werte im Laufe der Simulationsdurchläufe vedeutlichen. Die hervorgehobene rote Gerade beschreibt dabei den Verlauf der MSE-Mittelwerte der ersten und letzten Iterationen der Simulationsdurchläufe. Ihre zugehörige Steigung ist in der jeweiligen Grafiküberschrift abzulesen und dient als Indikator für eine präzise bzw. unpräzise Schätzung der Parameter. Sowohl für $\hat{\alpha}_i$ als auch für $\hat{\boldsymbol{\gamma}}$ ist diese Steigung positiv, was eine unpräzise Schätzung der Personen- und Itemparameter nahelegt.
Dieses Ergebnis der unpräzisen Schätzung der Parameter lässt sich auch auf die Simulation des Heterogenitätsmodells für Standardabweichungen von 0 bis 0.8 bei 100 Simulationsdurchläufen übertragen, jedoch nur auf die Schätzung der Itemparameter, nicht auf die Schätzung der Personenparameter. Für geschätzte Personenparameter besitzt der Verlauf der MSE-Mittelwerte der ersten und letzten Iterationen der Simulationsdurchläufe für Standardabweichungswerte von $\sigma = 0.6$ bis $\sigma = 0.8$ negative Steigungen und spricht daher für eine präzisere Schätzung der Personenparameter bei starker Heterogenität. Zur Bestätigung dieser Aussage beinhaltet die Tabelle 4.2 die Werte der Steigung der MSE-Mittelwert-Gerade zwischen der ersten und letzten Iteration aus Abbildung 4.2 für beispielhafte Seeds.

Die weiteren Simulationsergebnisse werden im Folgenden für den Seed von 160 präsentiert und erläutert. Damit die Ergebnisse der Simulationen des Heterogenitätsmodells und des BTL-Modells mit datengenerierendem Prozess DGP(α_i) miteinander verglichen werden können, werden dem datengenerierenden Prozess der Simulation des Heterogenitätsmodells u.a. die Prozessannahmen des DGP(α_i) zugrunde gelegt. In beiden Simulationen werden also grundsätzlich acht zu vergleichende Objekte, 40 Personen und $P = 100$ Simulationsdurchläufe angenommen, ebenso eine Normalverteilung für die Personen- und eine Gleichverteilung für die Itemparameter. Als Instrument zum Vergleich der Parameterschätzungen wird erneut der mittlere quadratische Fehler herangezogen. Für die Grundlage seiner Berechnung werden je nach Nicht-Konvergenz der Iterationsschleife gewisse Simulationsdurchläufe nicht berücksichtigt. Die genaue Anzahl der Simulationsdurchläufe für die Simulation eines Heterogenitätsmodells lässt sich berechnen durch das Abziehen der Werte aus Tabelle 4.1 von den ursprünglichen 100 Simulationsdurchläufen.
Zunächst wird die Schätzung des Personenparameters $\hat{\alpha}_i$ genauer betrachtet. In Ab-

	Seed					
	1		**160**		**2911**	
σ	$\hat{\alpha}_i$	$\hat{\gamma}$	$\hat{\alpha}_i$	$\hat{\gamma}$	$\hat{\alpha}_i$	$\hat{\gamma}$
0	0.0007586	0.0042766	0.0011043	0.0034723	0.0010596	0.0034085
0.1	0.0010084	0.0071820	0.0010977	0.0067183	0.0009669	0.0049438
0.2	0.0008409	0.0115097	0.0011312	0.0110202	0.0009567	0.0148774
0.3	0.0008438	0.0180714	0.0021800	0.0347988	0.0013103	0.0341835
0.4	0.0039499	0.0666485	0.0048605	0.0930034	0.0053061	0.0708914
0.5	0.0059520	0.2157226	0.0023636	0.2526040	0.0045986	0.2098837
0.6	-0.0042044	0.5473949	0.0249749	0.6263461	-0.0160375	0.6451917
0.7	-0.0439931	1.1303600	-0.0314169	1.4029960	-0.0592296	1.4368810
0.8	-0.0666539	2.3365430	-0.0565171	2.1107010	-0.1251898	2.2171670

Tabelle 4.2.: Steigungen des Verlaufs der MSE-Mittelwerte der ersten und letzten Iterationen der Simulationsdurchläufe des Heterogenitätsmodells. Die Steigung dient als Indikator für eine präzise Schätzung (negative Werte) bzw. unpräzise Schätzung (positive Werte) der Parameter.

bildung 4.3 sind für die unterschiedlichen Werte der Standardabweichung der Personenparameter jeweils nebeneinander Boxplots für die Verteilung des MSE der geschätzten Personenparameter sowohl für das Heterogenitätsmodell als auch für das BTL-Modell mit zugrunde liegendem DGP(α_i) abgebildet. Dies mag zunächst verwirren, da im BTL-Modell mit zugrunde liegendem DGP(α_i) der Personenparameter lediglich berücksichtigt, jedoch nicht geschätzt wird. Die Berechnung des MSE für den Personenparameterschätzer erfolgt in diesem Fall über die Annahme von Schätzungen $\hat{\alpha}_i = 0$, sodass lediglich der Erwartungswert der quadrierten wahren Personenparameter berechnet wird:

$$\text{MSE}^{BTL}(\hat{\alpha}_i) = \text{E}\Big[(0 - \alpha_i)^2\Big].$$

Für beide Modelle ist sofort ersichtlich, dass mit zunehmender Veränderung der Wahrnehmung der Reizstärke eines Objekts sowohl der MSE als auch die Größe der Boxen zunimmt. Offensichtlich ist ebenso, dass die Verteilung des MSE der Personenparameter im Heterogenitätsmodell eine geringere Steigung besitzt als im BTL-Modell mit DGP(α_i). Dies spricht für eine präzisere Schätzung der Personenparameter im Heterogenitätsmodell als im BTL-Modell mit DGP(α_i).

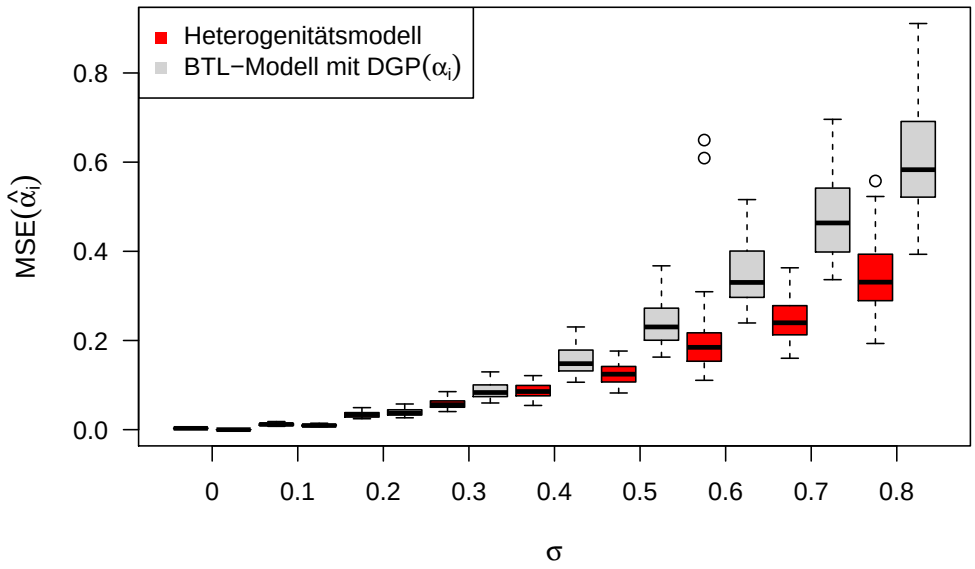

Abbildung 4.3.: Verteilung der MSE der geschätzten Personenparameter für das Heterogenitäts- und das BTL-Modell mit zugrunde liegendem DGP(α_i).

Da in Abbildung 4.3 die Boxplots für die Standardabweichung von 0 bis 0.4 kaum zu unterscheiden sind, werden sie in Abbildung 4.4 noch einmal näher betrachtet. Es wird deutlich, dass ab einer Standardabweichung des Personenparameters von 0.2 das Heterogenitätsmodell präzisere Schätzungen für $\hat{\alpha}_i$ hervorbringt als das BTL-Modell mit DGP(α_i).

Im zweiten Schritt des Algorithmus zur Schätzung des Heterogenitätsmodells werden die Personenparameter mittels random effects models bestimmt. Zusätzlich zu den zufälligen Effekten $\hat{\alpha}_i$ werden auch Schätzer für die zugehörigen Standardabweichungen $\hat{\sigma}$ berechnet. Abbildung 4.5 veranschaulicht die Verteilung der Standardabweichungen der zufälligen Effekte für die jeweiligen Heterogenitätsmodelle und vergleicht sie mit den wahren Werten der Standardabweichung der Personenparameter, gekennzeichnet durch die jeweils rot hervorgehobene horizontale Linie. Während der Median eines jeden Boxplots mit wachsender Standardabweichung erwartungsgemäß steigt, übersteigt sein Wert jedoch nie die wahre Standardabweichung der Personenparameter. Dementsprechend sind ab einer Standardabweichung von 0.3 auch nur Ausreißer nach unten hin erkennbar. Die Werte des Medians liegen für Modelle mit

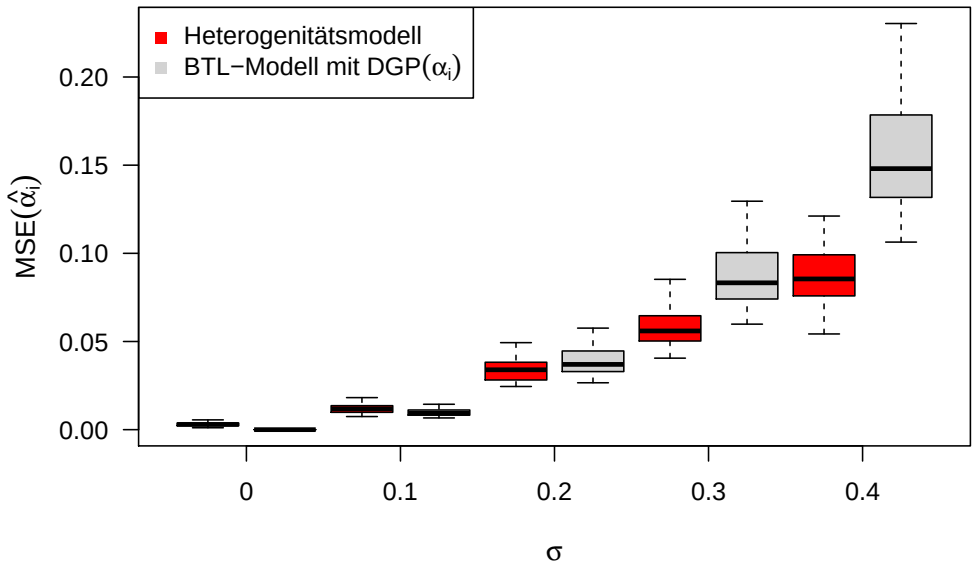

Abbildung 4.4.: Verteilung der MSE der geschätzten Personenparameter für das Heterogenitäts- und das BTL-Modell mit zugrunde liegendem DGP(α_i) bei Standardabweichungen von 0 bis 0.4.

$\sigma = 0.2$ bis $\sigma = 0.5$ sehr nah an den wahren Standardabweichungen, bei $\sigma = 0.3$ und $\sigma = 0.4$ fallen sie sogar fast exakt aufeinander. Für die Standardabweichungen von 0 bis 0.2 weist die Größe der Boxen auf eine größere Streuung der geschätzten Parameter hin. Ab einer wahren Standardabweichung von $\sigma = 0.6$ weisen die Standardabweichungen der zufälligen Effekte eine deutliche Tendenz zur Unterschätzung auf.

Analog zum Vergleich der Schätzung der Personenparameter sowohl im Heterogenitätsmodell als auch im BTL-Modell mit DGP(α_i), werden nun die Schätzungen der Itemparameter in diesen beiden Modellen miteinander verglichen. Abbildungen 4.6 und 4.7 veranschaulichen für die unterschiedlichen Werte der Standardabweichung der Personenparameter jeweils nebeneinander Boxplots für die Verteilung des MSE der geschätzten Itemparameter $\hat{\gamma}$ sowohl für das Heterogenitäts- als auch für das BTL-Modell mit zugrunde liegendem DGP(α_i). Ebenso wie beim Vergleich der Modelle für die Schätzung von α_i ist offensichtlich, dass in beiden Modellen mit zunehmender Standardabweichung sowohl der MSE als auch die Größe der Boxen zunimmt. Eindeutig ist außerdem, dass die Verteilung des MSE der Itemparameter

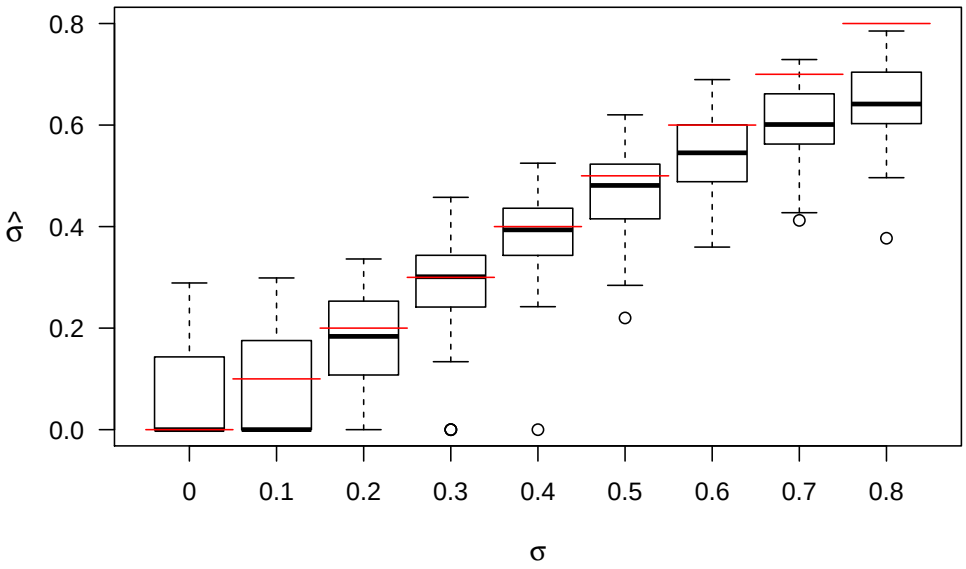

Abbildung 4.5.: Verteilung der Standardabweichung der zufälligen Effekte. Die horizontalen roten Linien kennzeichnen die Werte der jeweiligen wahren Standardabweichungen.

im Heterogenitätsmodell eine stärkere Steigung besitzt als die des BTL-Modells, was für eine unpräzise Schätzung der Itemparameter im Heterogenitätsmodell spricht.

Das in Abschnitt 4.1 spezifizierte Ziel anhand Abbildung 4.1 war die präzisere Schätzung eines BTL-Modells unter Einbeziehung einer Schätzung der Personenparameter α_i neben der Schätzung der Itemparameter $\boldsymbol{\gamma}$ und die damit einhergehende Verringerung des MSE der geschätzten Itemparameter. Zum Erreichen dieses Ziels wurde anschließend das Heterogenitätsmodell und sein zugehöriger Schätzalgorithmus vorgestellt. Im vorliegenden Fall der Abbildungen 4.6 und 4.7 entspricht die Verteilung des MSE von $\hat{\boldsymbol{\gamma}}$ für die Schätzung eines BTL-Modells mittels DGP(α_i) (graue Boxplots) der Darstellung des MSE von $\hat{\boldsymbol{\gamma}}$ aus Abbildung 4.1. Die simulierten Daten für dieses Modell sind dieselben, lediglich die Darstellungsart ist eine andere. Mit der in den Abbildungen 4.6 und 4.7 dargestellten grundsätzlich höheren Verteilung des MSE von $\hat{\boldsymbol{\gamma}}$ im Heterogenitätsmodell (rote Boxplots) als im BTL-Modell mit DGP(α_i) wird somit das Ziel einer präziseren Schätzung der Itemparameter durch das Heterogenitätsmodell nicht erreicht.

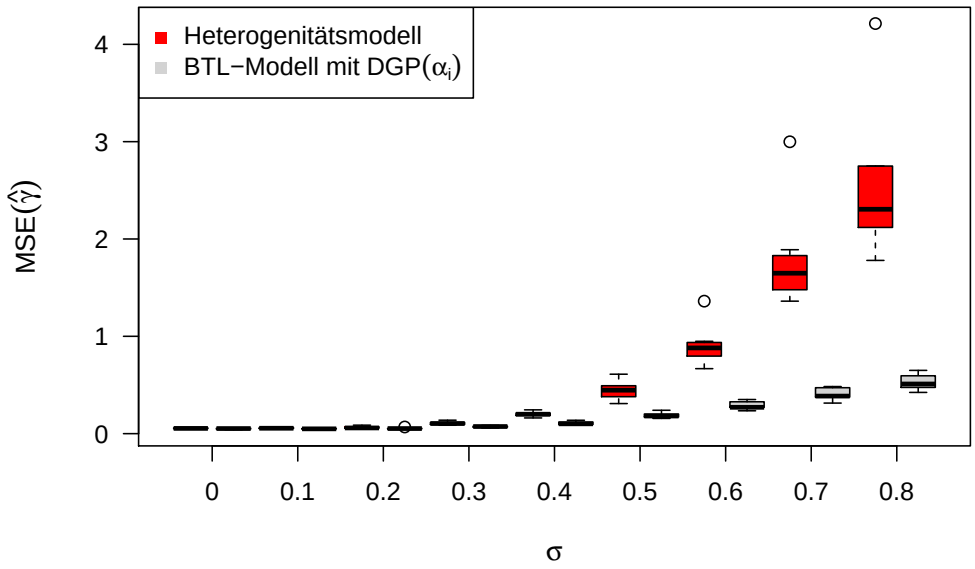

Abbildung 4.6.: Verteilung der MSE der geschätzten Itemparameter für das Heterogenitäts- und das BTL-Modell mit zugrunde liegendem DGP(α_i).

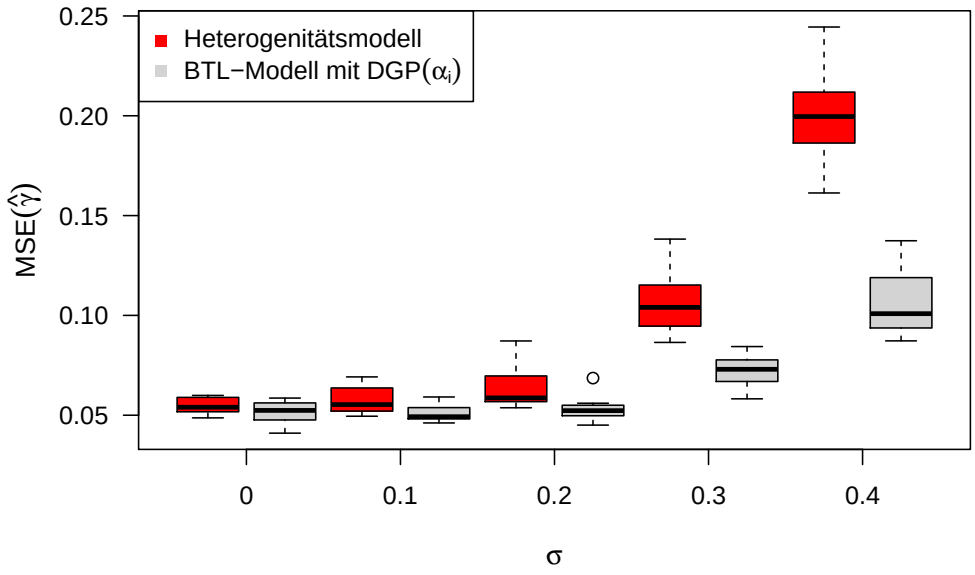

Abbildung 4.7.: Verteilung der MSE der geschätzten Itemparameter für das Heterogenitäts- und das BTL-Modell mit zugrunde liegendem DGP(α_i) bei Standardabweichungen von 0 bis 0.4.

Der Vollständigkeit halber sei an dieser Stelle erwähnt, dass bei der Schätzung des Heterogenitätsmodells in `R` regelmäßig Warnmeldungen der Form

```
In mer_finalize(ans) : singular convergence (7)
```
bzw.
```
In mer_finalize(ans) : false convergence (8)
```

auftreten. Diese Warnmeldungen wurden hier ignoriert, da keinerlei Hinweise auf die Verfälschung der Simulationsergebnisse beobachtet werden konnten.

Zur illustrativen Darstellung der in dieser Arbeit vorgestellten Schätzverfahren werden in den folgenden Anwendungsbeispielen das einfache BTL-Modell ebenso wie das Heterogenitätsmodell geschätzt und miteinander verglichen.

5. Anwendungsbeispiel: Lernmethoden

Für die Anwendung von generalisierten linearen Modellen zur Schätzung des einfachen und des personenspezifischen BTL-Modells wird in folgendem Anwendungsbeispiel der Datensatz `trdel` aus dem `R`-Paket `prefmod` (Hatzinger and Dittrich, 2012) verwendet. Dieser Datensatz beinhaltet u.a. die Daten einer Paarvergleichsstudie, in welcher erforscht werden soll, welche Methode von fünf betrachteten Lernmethoden (training delivery modes) von 198 Schulungsteilnehmern bevorzugt wird. Die Daten wurden in Verbindung mit der Bearbeitung einer Masterarbeit an der Wirtschaftsuniversität Wien erhoben (Schöll and Veith, 2011). Die einzelnen Lernmethoden sind:

- **Computerunterstütztes Lernen (CO):** Es soll nur durch die Unterstützung von Computer, Internet und CD-Rom gelernt werden. Es werden Lernprogramme (Lernsoftware) verwendet, die vom Lernenden zeitlich und räumlich flexibel genutzt werden können und bei denen die Lernenden nicht in direktem Kontakt mit dem Lehrenden und anderen Lernenden stehen.
- **TV-unterstütztes Lernen (TV):** Die Instruktionen erfolgen ausschließlich durch das Fernsehen über Lernsendungen, Videokonferenzen und DVDs.
- **Gedruckte Lernmittel (GL):** Das Lernen erfolgt textbasiert durch Lesematerialien wie Bücher, Skripte oder Arbeitshefte.
- **Audiounterstütztes Lernen (AU):** Es wird mittels Hörbänden und Hörbüchern gelernt. Die Wissensvermittlung erfolgt nur durch das Sprechen.
- **Unterricht/Vortrag (UV):** Bei dieser klassisschen Methode erfolgt die Wissensvermittlung in einer Präsenzveranstaltung durch einen Vortragenden (Lehrer/In).

Die Studienteilnehmer waren arbeitsuchende Personen, die an Arbeitsmarkt-Schulungen teilgenommen haben, welche vom österreichischen Arbeitsmarkservice (AMS) angeboten wurden. Das AMS ist das führende Dienstleistungsunternehmen am Arbeitsmarkt in Österreich. Es trägt im Rahmen der Vollbeschäftigungspolitik der Bundesregierung, im Auftrag des Bundesministers für Arbeit, Soziales und Konsumentenschutz und unter maßgeblicher Beteiligung der Sozialpartner zur Verhütung und Beseitigung von Arbeitslosigkeit in Österreich bei (AMS-Österreich, 2013).

Der Datensatz `trdel` enthält 198 Beobachtungen und 14 Variablen. Mit den fünf zu vergleichenden Items

1. Computerunterstütztes Lernen (CO)
2. TV-unterstütztes Lernen (TV)
3. Gedruckte Lernmittel (GL)
4. Audiounterstütztes Lernen (AU)
5. Unterricht/Vortrag (UV)

ergeben sich $\binom{5}{2} = 10$ mögliche Vergleichspaare, welche als binäre Variablen `V1` bis `V10` mit je zwei Antwortmöglichkeiten in den Datensatz eingehen. Die Ausprägungen der Variablen für die Paarvergleiche `V1` bis `V10` sind numerisch mit der Bedeutung für 1: „erstes Objekt bevorzugt“ und 2: „zweites Objekt bevorzugt“. Die bereits bekannte Anordnungstruktur der Paarvergleiche aus Abschnitt 3.3 lässt sich hier in der Anordnung der Daten wiederfinden. Die beobachteten Häufigkeiten der 198 befragten Versuchsteilnehmer bezüglich ihrer Präferenzen zwischen den Lernmethoden sind in Tabelle 5.1 dargestellt.
Der Datensatz `trdel` beinhaltet neben den Präferenzurteilen auch Variablen über bestimmte Eigenschaften der Studienteilnehmer, von denen angenommen wird, dass sie einen Einfluss auf die Präferenzwahl der Probanden haben könnten. Diese subjektspezifischen Variablen werden in den folgenden Analysen jedoch vernachlässigt, da zunächst in Abschnitt 5.1 das einfache BTL-Modell geschätzt wird, bei welchem nur die Effekte der einzelnen Paarvergleiche interessieren, und anschließend in Abschnitt 5.2 das in Kapitel 4 vorgestellte personenspezifische BTL-Modell mit multiplikativem Personenparameter α_i auf den Datensatz angewendet wird.

	>	<
CO : TV	157	41
CO : GL	109	89
TV : GL	82	116
CO : AU	162	36
TV : AU	129	69
GL : AU	137	61
CO : UV	86	112
TV : UV	55	143
GL : UV	72	126
AU : UV	51	147

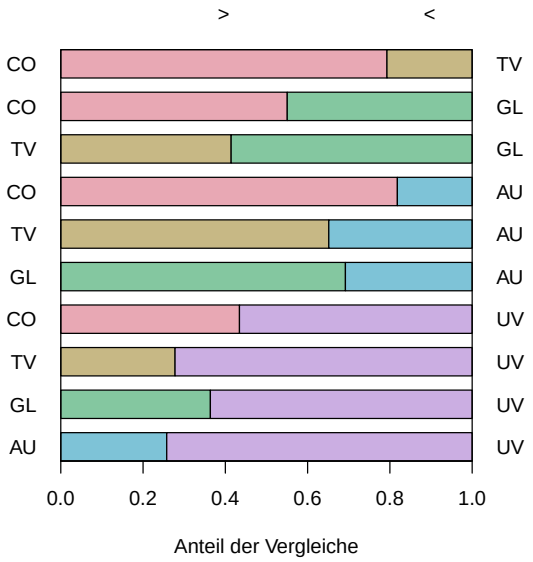

Abbildung 5.1.: Beobachtete Häufigkeiten der Paarvergleiche für den `trdel`-Datensatz.

5.1. Anwendung für das einfache BTL-Modell

Zunächst wird das einfache BTL-Modell mittels generalisierter linearer Regression geschätzt. Für den Datensatz `trdel`, mit $m = 5$ zu vergleichenden Objekten, $I = 10$ Paarvergleichen und $n = 198$ beurteilenden Personen, ergibt sich für die Grundstruktur der Designmatrix $\boldsymbol{X}$ die Dimension (1980×4). Aus Gründen der Identifizierbarkeit enthält die Designmatrix $m - 1 = 4$ Spalten.

Die Schätzung der Paarvergleichsdaten über generalisierte lineare Modelle liefert die Effekte aus den einzelnen Paarvergleichen der fünf Lernmethoden. Ebenso ergibt die Untersuchung der latenten Variable „Bevorzugung einer Lernmethode“ eine eindeutige Rangordnung. Die Schätzer für die Itemparameter sowie die den Lernmethoden zugewiesen Ränge sind in Tabelle 5.1 angegeben.

	CO	TV	GL	AU	UV
Schätzer:	-0.0642	-0.9729	-0.5055	-1.3936	0
Rangzuweisung:	2	4	3	5	1

Tabelle 5.1.: Koeffizientenschätzer und Rangzuordnung des einfachen BTL-Modells im `trdel`-Datensatz.

Der letzte Itemparameter, in diesem Fall der Parameter für die Lernmethode „Unterricht/Vortrag“, wird immer auf Null gesetzt und fungiert somit als Referenzobjekt. Anhand der Schätzwerte lässt sich erkennen, dass die Versuchsteilnehmer das Lernen durch Unterricht (UV) am häufigsten bevorzugen. An zweiter Stelle liegt das computerunterstützte Lernen (CO), gefolgt von gedruckten Lernmitteln (GL) und TV-unterstütztem Lernen (TV). An letzter Stelle steht das Lernen durch audiounterstützten Unterricht (AU).
Eine grafische Darstellung der Koeffizientenschätzer ist in Abbildung 5.2 gegeben. Die horizontale Linie entlang der Null markiert das Referenzobjekt UV.

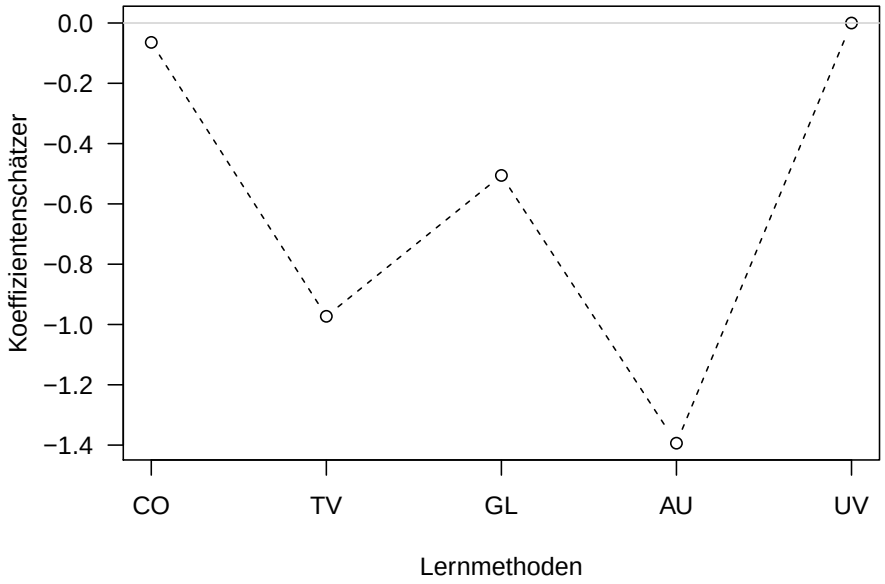

Abbildung 5.2.: Darstellung der Koeffizientenschätzer des einfachen BTL-Modells für den `trdel`-Datensatz. Die horizontale Gerade entlang der Null markiert das Referenzobjekt „Unterricht/Vortrag“ (UV).

Die Ergebnisse der Schätzung von Paarvergleichsdaten mittels generalisierter linearer Regression können mit den Ergebnissen der Schätzung von einfachen Bradley-Terry-Luce Modellen verglichen werden, für die bereits eine Funktion in `R` implementiert wurde, da keine zusätzlichen Kovariablen berücksichtigt werden. Hierfür wird die `R`-Funktion `btReg.fit` aus dem Paket `psychotools` (Zeileis et al., 2012) verwendet. Es resultieren dieselben Schätzer wie bei der Modellierung mittels generalisierter linearer Regression (vgl. Anhang A.4.1).

Über die geschätzten Itemparameter können anschließend die Wahrscheinlichkeiten für den Vergleich der Objekte r und s gemäß Gleichung (2.5) berechnet werden. Werden beispielsweise die Lernmethoden CO und TV miteinander verglichen, ergibt die Wahrscheinlichkeit für die Präferenz von computerunterstütztem Lernen gegenüber TV-unterstütztem Lernen etwa 71.3 %:

$$p_{\text{CO, TV}} = \frac{\exp(-0.0642 - (-0.9729))}{1 + \exp(-0.0642 - (-0.9729))} = 0.7127.$$

Neben den Koeffizientenschätzern lassen sich bei der Modellschätzung durch die Funktion `btReg.fit` auch Schätzer für die wahren Präferenzen $\hat{\pi}_r \geq 0$, $\sum_{r=1}^{m} \hat{\pi}_r = 1$ durch den Aufruf `worth` ermitteln. In Tabelle 5.2 werden die Schätzer der wahren Präferenzen, die auch mit „Worth-Parameter“ bezeichnet werden, für die einzelnen Lernmethoden angegeben.

	CO	TV	GL	AU	UV
$\hat{\pi}_r$:	0.2961	0.1194	0.1904	0.0784	0.3157

Tabelle 5.2.: Worth-Parameter des einfachen BTL-Modells im `trdel`-Datensatz.

Die graphische Darstellung der Worth-Parameter in Abbildung 5.3 zeigt den gleichen Verlauf wie bei der Betrachtung der geschätzten Itemparameter in Abbildung 5.2. Die Gerade für den Referenzwert liegt für den vorliegenden Fall von $m = 5$ Items bei $1/5 = 0.2$, dem Wert der wahren Präferenz $\pi_r, r = 1, \ldots, m$ für den Fall eines indifferenten Beurteilers.

Gemäß Gleichung (2.1) lassen sich ebenso wieder die Wahrscheinlichkeiten p_{rs} für die Bevorzugung einer Lernmethode gegenüber einer anderen Berechnen. Die Wahrscheinlichkeit für die Präferenz von computerunterstütztem Lernen gegenüber TV-unterstütztem Lernen ergibt somit wieder etwa 71.3 %:

$$p_{\text{CO, TV}} = \frac{0.2961}{0.2961 + 0.1194} = 0.7126.$$

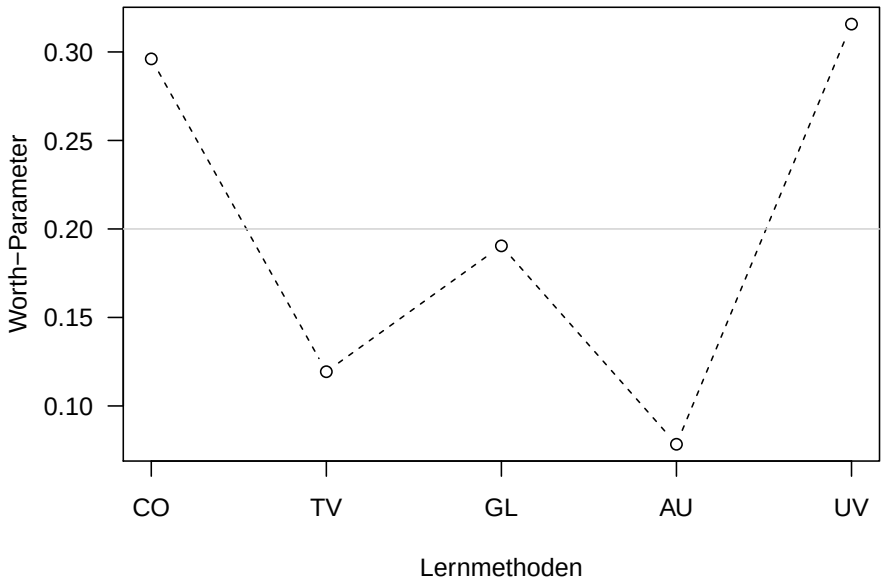

Abbildung 5.3.: Darstellung der Worth-Parameter des einfachen BTL-Modells für den `trdel`-Datensatz. Die horizontale Gerade entlang dem Wert $1/5 = 0.2$ markiert den Referenzwert eines indifferenten Beurteilers.

5.2. Anwendung für das Heterogenitätsmodell

Bei der im letzten Abschnitt ermittelten Rangfolge der Lernmethoden wurde für die Schätzung eines einfachen BTL-Modells angenommen, dass die Versuchspersonen die Reizstärken der Lernmethoden im selben Maße wahrnehmen. Um eine Rangfolge für die Lernmethoden zu erhalten, bei welcher der Einfluss der Versuchspersonen auf geeignete Weise berücksichtigt wird, kann das aus Kapitel 4 bekannte Heterogenitätsmodell geschätzt werden.

Aus Abschnitt 4.4 ist bekannt, dass für die Schätzung der GLMMs (Schritt (2) des Algorithmus zur Schätzung des Heterogenitätsmodells) mit der `R`-Funktion `glmer` das Argument `X.noise` spezifiziert werden muss. Dieses Argument ist eine Zufallsvariable, welche völlig unabhängig vom Response kreiert wird und nur der Sicherstellung der Funktionsfähigkeit der `glmer`-Funktion dient. Für die Anwendung des Heterogenitätsmodells werden für `X.noise` $n \cdot \binom{m}{2} = 198 \cdot 10 = 1980$ standardnormalverteilte Zufallszahlen gezogen, diese jeweils mit der Zahl 10000 multipliziert, anschließend ihr Absolutbetrag genommen und ihr Wert zuletzt auf ganze Zahlen gerundet. Zur Gewährleistung der Reproduzierbarkeit der Ergebnisse der Schätzung

des Heterogenitätsmodells, muss dem Ziehen der Zufallszahlen ein Seed vorhergehen. Die Ergebnisse der Modellschätzung sind mit jedem unterschiedlichen Seed gewissen Schwankungen ausgesetzt, welche sich nicht unerheblich auf die Schätzungen der Itemparameter auswirken. Es ergibt sich für unterschiedliche Seeds zum einen nicht immer dieselbe Rangfolge der zu beurteilenden Items, zum anderen sind die Unterschiede in den aus den Itemparametern berechneten individuellen Wahrscheinlichkeiten p_{irs} für die Bevorzugung von Objekt r gegenüber Objekt s zu groß, um als vernachlässigbar zu gelten. Um die Instabilität des Modells für unterschiedliche Seeds und damit die Bedeutung des gewählten Seeds für das Modell zu verdeutlichen, werden für dieses Anwendungsbeispiel zunächst Heterogenitätsmodelle für drei unterschiedliche Seeds (279, 5117 und 13389) berechnet und ihre Schätzer in Tabelle 5.3 gegenübergestellt. Um letztendlich ein allgemeingültiges Heterogenitätsmodell für die Präferenz der Lernmethoden zu erhalten, werden mit 100 zufälligen Seeds Heterogenitätsmodelle gerechnet und die sich jeweils ergebenden Itemparameterschätzer über die 100 Modelle hinweg gemittelt. Um das Ergebnis reproduzieren zu können, wird der Ziehung der 100 zufälligen Seeds wiederum ein Seed (hier: Seed = 160) vorausgesetzt. Das Ergebnis der gemittelten Schätzer für die Itemparameter ist ebenfalls in Tabelle 5.3 ersichtlich.

	CO	TV	GL	AU	UV
Schätzer bei Seed = 279:	-2.3736	-8.9125	-1.4163	-9.6926	0
Schätzer bei Seed = 5117:	-1.3515	-3.7482	-0.9581	-4.3259	0
Schätzer bei Seed = 13389:	-1.0568	-2.7148	-0.7104	-3.1041	0
gemittelte Schätzer aus 100 Seeds:	-2.0494	-7.0773	-1.2482	-7.7798	0
Rangzuweisung:	3	4	2	5	1

Tabelle 5.3.: Koeffizientenschätzer und Rangzuordnung des Heterogenitätsmodells im `trdel`-Datensatz. Den Schätzern unterliegen Modellschätzungen mit jeweils unterschiedlichen Seeds für die Zufallsvariable `X.noise`, welche für die Funktion `glmer` in `R` benötigt wird.

Die Lernmethode „Unterricht/Vortrag“ ist das Referenzobjekt in allen Modellen und nimmt daher den Wert Null an. Es ergibt sich für jede der drei Modellschätzungen, bei denen sich nur der Seed für die Zufallsvariable `X.noise` ändert, dieselbe Rangordnung wie bei den gemittelten Modellschätzungen: Die Versuchsteilnehmer bevorzugen das Lernen durch Unterricht (UV) am häufigsten. An zweiter Stelle liegt

das Lernen mit gedruckten Lernmitteln (GL), gefolgt von computerunterstütztem Lernen (CO) und TV-unterstütztem Lernen (TV). Den letzten Platz belegt das Lernen durch audiounterstützten Unterricht (AU). In Abbildung 5.4 wird der Verlauf der über die 100 Heterogenitätsmodelle gemittelten Itemparameterschätzer (hervorgehobene schwarze Kurve) dargestellt, an dem sich auch die sich ergebende Rangfolge ablesen lässt. Die Parameterschätzer der 100 Heterogenitätsmodelle, welche den gemittelten Schätzern zugrunde liegen, sind in Abbildung 5.4 um die gemittelten Itemparameterschätzer herum eingezeichnet und sollen noch einmal die sich durch die Wahl des Seeds ergebenden Schwankungen in den Modellschätzungen verdeutlichen. Mit 97 der 100 eingezeichneten Heterogenitätsmodelle mit unterschiedlichen Seeds (graue Kurven) ergibt sich dieselbe Rangfolge, bei drei der den gemittelten Schätzern zugrunde gelegten Modellen ergibt sich als die am häufigsten präferierte Lernmethode diejenige mit gedruckten Lernmitteln (rote Kurven). Diese Modelle, mit vom gemittelten Heterogenitätsmodell abweichender Rangfolge, werden im Folgenden „Rangfolgen-Ausreißer-Modelle“ genannt.

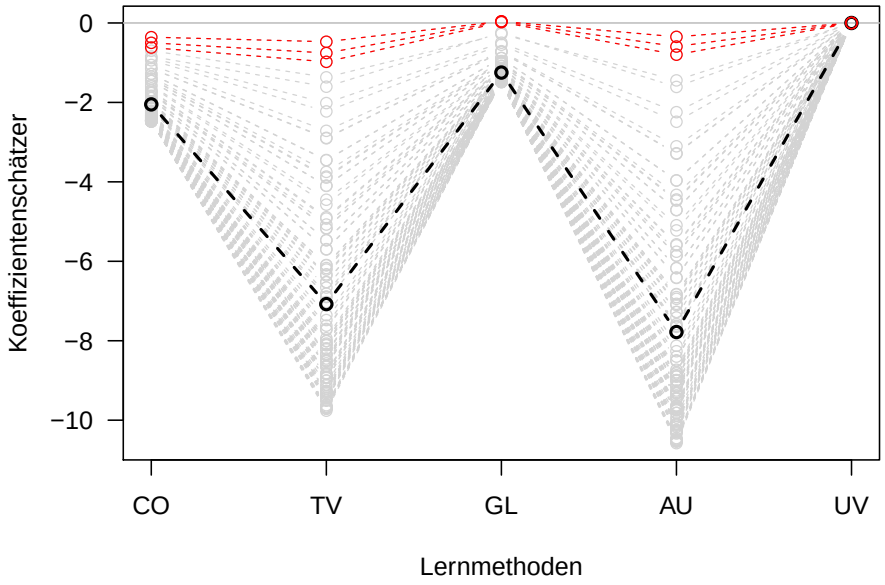

Abbildung 5.4.: Darstellung von Koeffizientenschätzern aus 100 Heterogenitätsmodellen mit unterschiedlichem Seed für den `trdel`-Datensatz. Die schwarze Kurve verbindet die Mittelwerte aller Schätzer. Die drei Modelle, die zu einer anderen Rangfolge führen als die restlichen 97 sind rot hervorgehoben. Die horizontale Gerade entlang der Null markiert das Referenzobjekt „Unterricht/Vortrag“ (UV).

Der Vergleich der Schätzungen zwischen dem Heterogenitätsmodell und dem einfachen BTL-Modell aus Abschnitt 5.1 zeigt eine Änderung in der Rangordnung der Lernmethoden. Während das BTL-Modell ergibt, dass die Versuchspersonen das computerunterstützte Lernen (Rang 2) vor dem Lernen mit gedruckten Lernmitteln (Rang 3) bevorzugen, werden die Präferenzen bezüglich dieser zwei Items im gemittelten Heterogenitätsmodell umgekehrt. Für alle anderen Items ergibt sich keine Verschiebung in der Rangfolge.

Wie zuvor schon erwähnt, bringt die Berechnung der individuellen Wahrscheinlichkeiten p_{irs} für die Bevorzugung von Objekt r gegenüber Objekt s (siehe Gleichung (4.2)) je nach Modellschätzung mit unterschiedlichem Seed auch ein unterschiedliches Ergebnis. Im Vergleich zum BTL-Modell, in dem grundsätzlich für alle $i = 1, \ldots, n$ Versuchpersonen ein Wert von $\alpha_i = 1$ angenommen wird, variieren im Heterogenitätsmodell die Werte der Personenparameter für alle Versuchpersonen und drücken damit ihre individuelle Wahrnehmung der Reizstärke der Objekte aus. Für den Fall von $\alpha_i = 1$ wird angenommen, dass er dem wahren Reizwert der zu vergleichenden Objekte entspricht.
Während bei $\alpha_i = 1$ die Wahrscheinlichkeit für die Präferenz von computerunterstütztem Lernen gegenüber TV-unterstütztem Lernen im BTL-Modell etwa 71.3 % beträgt, ergibt sich für diese Wahrscheinlichkeit im gemittelten Heterogenitätsmodell ein deutlich höherer Wert von etwa 99.4 %:

$$p_{i,\text{CO, TV}} = \frac{\exp(1 \cdot (-2.0494 - (-7.0773)))}{1 + \exp(1 \cdot (-2.0494 - (-7.0773)))} = 0.9935.$$

Mit dem Heterogenitätsmodell mit einem Seed von 6194, ein Rangfolgen-Ausreißer-Modell, dessen Schätzer in Abbildung 5.4 einer der rot hervorgehobenen Kurven entsprechen, lässt sich für $\alpha_i = 1$ eine vom gemittelten Heterogenitätsmodell stark abweichende Wahrscheinlichkeit $p_{i,\text{CO, TV}}$ von 52.9 % berechnen und sich damit nochmals die Instabilität des Heterogenitätsmodells für unterschiedliche Seeds verdeutlichen. In Tabelle 5.4 lassen sich die Wahrscheinlichkeiten $p_{i,\text{CO, TV}}$ für Modelle mit unterschiedlichen Seeds und für das gemittelte Heterogenitätsmodell bei verschiedenen Werten für den Personenparameter ablesen. Für einem Personenparameter von $\alpha_i = 0.5$ wird angenommen, dass eine Versuchsperson die Reizwerte der Objekte nicht so stark wahrnimmt, wie sie tatsächlich vorliegen. Bei einem Heterogenitätsfaktor $\alpha_i = -0.5$ werden von einer Versuchsperson die Reizstärken der Objekte ver-

tauscht wahrgenommen, so dass sie Objekt s gegenüber Objekt r präferiert, wenn eine andere Person mit $\alpha_i = 0.5$ Objekt r gegenüber Objekt s präferiert.

$p_{i,\text{CO, TV}}$	$\alpha_i = -0.5$	$\alpha_i = 0.5$	$\alpha_i = 1$
Heterogenitätsmodell mit Seed = 279:	3.7 %	96.3 %	99.9 %
Heterogenitätsmodell mit Seed = 5117:	23.2 %	76.8 %	91.7 %
Heterogenitätsmodell mit Seed = 6194:	48.6 %	51.4 %	52.9 %
Heterogenitätsmodell mit Seed = 13389:	30.4 %	69.6 %	84.0 %
gemitteltes Heterogenitätsmodell aus 100 Seeds:	7.5 %	92.5 %	99.4 %

Tabelle 5.4.: Berechnete Wahrscheinlichkeiten $p_{i,\text{CO, TV}}$ (`trdel`-Datensatz) für Schätzungen des Heterogenitätsmodells mit unterschiedlichen Seeds bei Annahme unterschiedlicher Personenparameter.

Da das Heterogenitätsmodell nicht auf der `R`-Funktion `btReg.fit` aus dem Paket `psychotools` (Zeileis et al., 2012) basiert, kann die Funktion `worth` zur Bestimmung der Worth-Parameter nicht angewendet werden. Selbstverständlich lassen sich die Worth-Parameter jedoch auch händisch über folgende Gleichung (5.1) bestimmen:

$$\hat{\pi}_r = \frac{\exp(\hat{\gamma}_r)}{\sum_{r=1}^{m} \exp(\hat{\gamma}_r)}. \tag{5.1}$$

Das Heterogenitätsmodell, in welchem der Einfluss der Versuchspersonen durch einen multiplikativen Personenparameter α_i berücksichtigt wird, führt in der vorliegenden Anwendung auf Daten einer Paarvergleichsstudie, in welcher erforscht werden soll, welche Methode von fünf betrachteten Lernmethoden bevorzugt wird, zu einem anderen Ergebnis als das BTL-Modell. Während im einfachen BTL-Modell Homogenität der Versuchpersonen angenommen wird, wird für das Heterogenitätsmodell angenommen, dass die Versuchspersonen die Reizstärken der Objekte in einem Paarvergleich unterschiedlich wahrnehmen. Der im Heterogenitätsmodell berechnete Wert der Standardabweichung des Personenparameters, $\hat{\sigma}$, schätzt diese Unterschiede der Wahrnehmung der Reizstärken der Items. Tabelle 5.5 gibt die Werte der geschätzten Standardabweichungen sowohl für das gemittelte Heterogenitätsmodell als auch für Modelle mit unterschiedlichen Seeds an.

Im Heterogenitätsmodell mit dem Seed von 6194, einem Rangordnung-Ausreißer-Modell, weicht der Wert von $\hat{\sigma} = 3.820$ für die geschätzte Standardabweichung aus

	$\hat{\sigma}$
Heterogenitätsmodell mit Seed = 279:	0.856
Heterogenitätsmodell mit Seed = 5117:	0.868
Heterogenitätsmodell mit Seed = 6194:	3.820
Heterogenitätsmodell mit Seed = 13389:	0.937
gemitteltes Heterogenitätsmodell aus 100 Seeds:	0.918

Tabelle 5.5.: Geschätzte Standardabweichungen des Personenparameters in Heterogenitätsmodellen mit unterschiedlichen Seeds für den `trdel`-Datensatz.

Tabelle 5.5 extrem von denen der geschätzten Standardabweichung aus Modellen mit anderen Seeds ab und stellt damit einen Ausreißer-Wert dar, welcher die geschätzte Standardabweichung des gemittelten Heterogenitätsmodells verzerrt. Abbildung 5.5 veranschaulicht die Verteilung der geschätzten Standardabweichung aller 100 Modelle mit unterschiedlichen Seeds, die dem gemittelten Heterogenitätsmodell zugrunde liegen. Es wird deutlich, dass die extrem hohen Werte der drei Rangordnung-Ausreißer-Modelle (rot gekennzeichnet) eine Erhöhung des Mittelwerts mit sich bringen. Um einen allgemeingültigen Wert für die Schätzung der Unterschiede in der Wahrnehmung der Reizstärken von den fünf Lernmethoden verwenden zu können, sollte daher anstelle der gemittelten Standardabweichung von 0.9175 (in der Abbildung blau gekennzeichnet), welche auf Grundlage aller 100 Modelle mit verschiedenen Seeds berechnet wurde, eher der niedrigere Median (in der Abbildung grün gekennzeichnet) von 0.8504 verwendet werden.
Es lässt sich schlussfolgern, dass mit dem vorangestellten Seed für die Ziehung der Zufallsvariablen `X.noise`, welche die Funktionsfähigkeit der `R`-Funktion `glmer` gewährleistet, Modelle geschätzt werden können, die einen so starken Ausreißer-Charakter besitzen, dass sie die Mittelwerte der Schätzergebnisse von 100 zugrunde liegenden Modellen stark verzerren können.
Aus der Simulation des Heterogenitätsmodells in Kapitel 4 ergab sich mit Abbildung 4.5 die Erkenntnis, dass ab einer wahren Standardabweichung von $\sigma = 0.6$ die geschätzten Standardabweichungen der Personenparameter eine deutliche Tendenz zur Unterschätzung aufweisen. Es kann also davon ausgegangen werden, dass die wahre Standardabweichung für die Versuchspersonen im Paarvergleichsexperiment bezüglich der Beurteilung der Lernmethoden in Wirklichkeit größer ist als die in Ta-

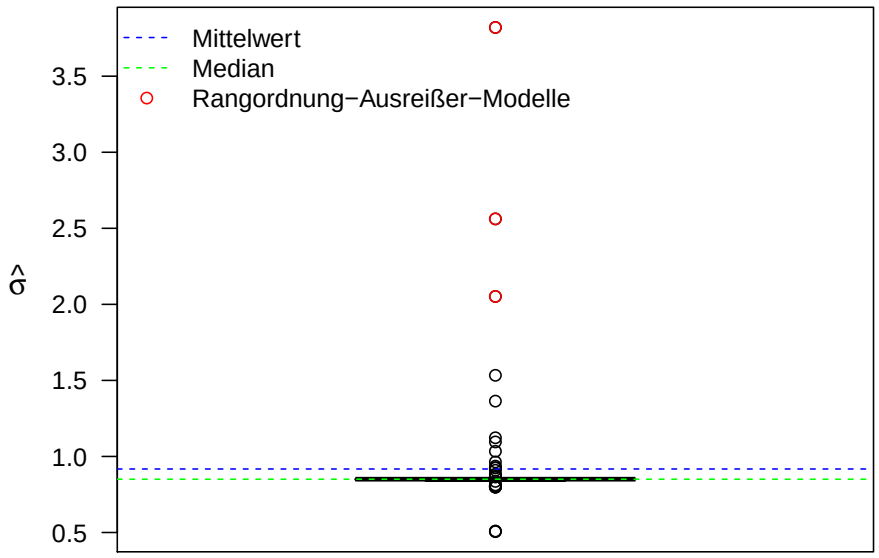

Abbildung 5.5.: Verteilung der geschätzten Standardabweichung von Heterogenitätsmodellen mit unterschiedlichen Seeds für den `trdel`-Datensatz. Die hohen geschätzten Standardabweichungen der Rangordnung-Ausreißer-Modelle führen zu einer Verzerrung der geschätzten Standardabweichung des gemittelten Heterogenitätsmodells.

belle 5.5 angegeben geschätzten Werte. Die Tatsache, dass die Iterationsschleife bei keiner der Schätzungen des Heterogenitätsmodells mit unterschiedlichem Seed konvergiert, spricht ebenso für eine wahre Standardabweichung größer als 0.8 (vgl. Abschnitt 4.5). Aus diesem Grund kann im vorliegenden Anwendungsbeispiel bezüglich der Lernmethoden auch nicht von präzisen Schätzern ausgegangen werden. Obwohl das Heterogenitätsmodell sowohl aufgrund der starken Auswirkungen der Wahl des Seeds für die Zufallsvariable `X.noise` als auch aufgrund der Nicht-Konvergenz der Iterationsschleife nicht stabil ist, deutet das Ergebnis der geänderten Rangfolge der Lernmethoden im Vergleich zum geschätzten BTL-Modell ebenso wie die Tatsache einer von Null verschieden geschätzten Standardabweichung darauf hin, dass die im Modell angenommene Heterogenität der Versuchspersonen tatsächlich vorliegt. Das Heterogenitätsmodell ist jedoch nicht in der Lage, diese genau zu erfassen.

6. Anwendungsbeispiel: Parteipräferenzen

Ein weiteres Anwendungsbeispiel für die Schätzung des einfachen und des personenspezifischen BTL-Modells mittels generalisierter linearer Modelle behandelt die Präferenzen von 192 Personen bezüglich ihrer Paarvergleiche bei der Wahl von fünf deutschen politischen Parteien und der Möglichkeit, sich der Wahl zu enthalten. Der betrachtete Datensatz `GermanParties2009` ist im `R`-Paket `psychotools` (Zeileis et al., 2012) verfügbar und beinhaltet 192 Beobachtungen für sechs Variablen. Im Folgenden wird ausschließlich die Variable „preference“ betrachtet, welche die Urteile der Paarvergleiche der Versuchspersonen beinhaltet.
Die Datenerhebung wurde vom Fachbereich Psychologie der Universität Tübingen im Juni 2009, drei Monate vor der 17. deutschen Bundestagswahl, durchgeführt. Die zu vergleichenden Parteien waren Die Linke, Bündnis 90/Die Grünen, SPD, CDU/CSU und die FDP. Zusätzlich zu diesen 5 Parteien wurde als weiteres Item die Option der Wahlenthaltung (im Datensatz kodiert mit „none“ und im Folgenden mit „keine“ bezeichnet) miteinbezogen. Den Versuchsteilnehmern wurden $I = 15$ Paarvergleiche der sechs Items in zufälliger Reihenfolge präsentiert. Bei jedem Vergleich bestand ihre Aufgabe darin, diejenige Partei auszuwählen, für welche sie auch bei der kommenden Bundestagswahl 2009 stimmen würden. Die beobachteten Präferenzen der Versuchsteilnehmer sind in Abbildung 6.1 aufgeführt.
Die Interviewer waren Master-Studenten des Fachs Psychologie, welche die Daten innerhalb eines bewerteten Kurses erhoben. Da sie hauptsächlich Personen befragten, die sie kannten, sind die Ergebnisse des Paarvergleichsexperiments für die politische Meinung der deutschen Bevölkerung nicht repräsentativ.

	>	<
keine : Linke	119	73
keine : Gruene	25	167
Linke : Gruene	23	169
keine : SPD	38	154
Linke : SPD	34	158
Gruene : SPD	124	68
keine : CDU/CSU	69	123
Linke : CDU/CSU	76	116
Gruene : CDU/CSU	128	64
SPD : CDU/CSU	125	67
keine : FDP	83	109
Linke : FDP	70	122
Gruene : FDP	137	55
SPD : FDP	134	58
CDU/CSU : FDP	106	86

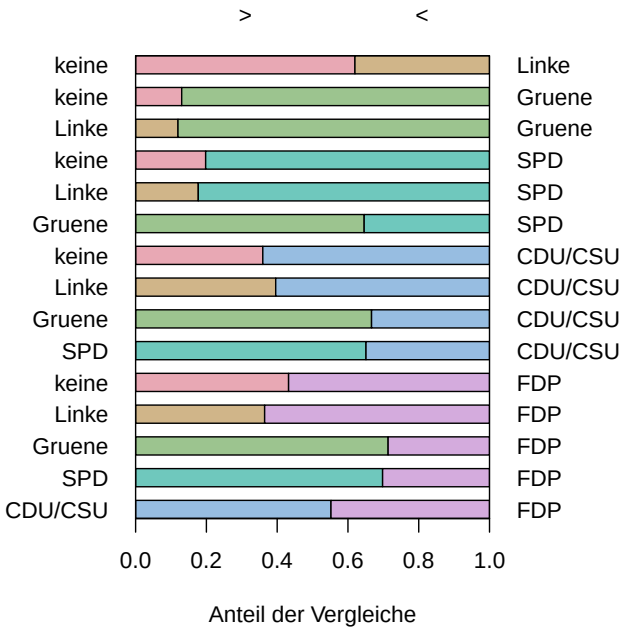

Abbildung 6.1.: Beobachtete Häufigkeiten der Paarvergleiche für den `GermanParties2009`-Datensatz.

6.1. Anwendung für das einfache BTL-Modell

Für die Paarvergleichsdaten des Datensatzes `GermanParties2009` wird mit $m = 6$ zu vergleichenden Objekten, $I = 15$ Paarvergleichen und $n = 192$ beurteilenden Personen zunächst wieder das einfache Bradley-Terry-Luce Modell mittels generalisierter linearer Regression geschätzt. Die Werte der geschätzten Itemparameter sowie die sich damit ergebende Rangordnung der Untersuchung der latenten Variable „Parteipräferenz“ sind in Tabelle 6.1 abzulesen. Als Referenzobjekt dient das Item FDP, dessen Schätzer daher auf Null gesetzt wurde.

	keine	Linke	Gruene	SPD	CDU/CSU	FDP
Schätzer:	-0.3756	-0.6161	1.1858	0.8131	0.1756	0
Rangzuweisung:	5	6	1	2	3	4

Tabelle 6.1.: Koeffizientenschätzer und Rangzuordnung des einfachen BTL-Modells im `GermanParties2009`-Datensatz.

Anhand der Schätzer ergibt sich, dass von den Versuchspersonen bei der zum Zeitpunkt der Befragung anstehenden Bundestagswahl am häufigsten Bündnis 90/Die

Grünen gewählt werden würde, gefolgt von der SPD, CDU/CSU und der FDP. Desweiteren präferieren die Versuchsteilnehmer eine Wahlenthaltung vor einer Stimmabgabe für die Partei Die Linke.
Eine grafische Darstellung der Schätzer gibt Abbildung 6.2. Die horizontale Linie entlang der Null markiert das Referenzobjekt FDP.

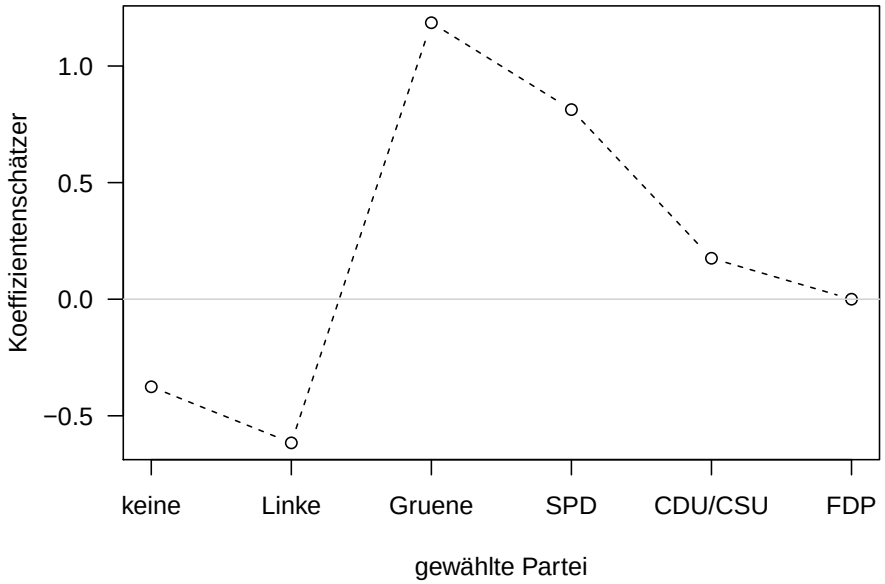

Abbildung 6.2.: Darstellung der Koeffizientenschätzer des einfachen BTL-Modells für den `GermanParties2009`-Datensatz. Die horizontale Gerade entlang der Null markiert das Referenzobjekt FDP.

Bei Betrachtung des „Gewinners“ der BTL-Schätzung (Die Grünen), stimmen im Übrigen die Ergebnisse des Paarvergleichsexperiments mit denen der Bundestagswahl 2009 für die Wähler der Stadt Tübingen überein. Die Ergebnisse der Wahl vom 27. September 2009 (Anzahl der sogenannten Zweitstimmen in Prozent) enthält Tabelle 6.2. Die Wahlbeteiligung betrug 70.8 % in Deutschland und 80.5 % in der Stadt Tübingen (Bundeswahlleiter, 2013, Universitätsstadt Tübingen, 2013).

Bei Vergleich der Schätzung von Paarvergleichsdaten mittels generalisierter linearer Regression und der `R`-Funktion `btReg.fit` aus dem Paket `psychotools` (Zeileis et al., 2012) resultieren wie auch im Anwendungsbeispiel der Lernmethoden (siehe Kapitel 5) erneut dieselben Schätzer (vgl. Anhang A.4.2). Mit der `worth`-Funktion lassen sich ebenso wieder für Ergebnisse der Modellschätzung durch die Funktion

	Deutschland	Tübingen
CDU/CSU	33.8	23.0
SPD	23.0	21.1
FDP	14.6	13.9
Die Linke	11.9	8.5
Grüne	10.7	27.9
Sonstige	6.0	5.7

Tabelle 6.2.: Ergebnisse der Bundestagswahl 2009: Anzahl der Zweitstimmen in Prozent. Quelle: Bundeswahlleiter (2013), Universitätsstadt Tübingen (2013).

`btReg.fit` Schätzer für die wahren Präferenzen $\hat{\pi}_r \geq 0$, $\sum_{r=1}^{m} \hat{\pi}_r = 1$ ermitteln. In Tabelle 6.3 werden diese Worth-Parameter angegeben.

	keine	Linke	Gruene	SPD	CDU/CSU	FDP
$\hat{\pi}_r$:	0.0768	0.0604	0.3658	0.2520	0.1332	0.1118

Tabelle 6.3.: Worth-Parameter des einfachen BTL-Modells im `GermanParties2009`-Datensatz.

Die grafische Darstellung der Worth-Parameter in Abbildung 6.3 stellt einen ähnlichen Verlauf dar wie die grafische Darstellung der Koeffizientenschätzer in Abbildung 6.2. Die Gerade für den Referenzwert liegt für $m = 6$ Objekte bei $1/6 = 0.1\bar{6}$, dem Wert der wahren Präferenz $\pi_r, r = 1, \ldots, m$ für den Fall eines indifferenten Beurteilers.

Die Berechnung der Wahrscheinlichkeiten p_{rs} für die Präferenz von Objekt r gegenüber Objekt s lässt sich erneut sowohl aus den Werten der geschätzten Itemparameter als auch auch aus den Werten der Worth-Parameter ermitteln. Sowohl die Anwendung von Gleichung (2.5) als auch von Gleichung (2.1) ergeben eine Präferenzwahrscheinlichkeit von ca. 65.4 % für die bevorzugte Wahl der SPD vor der CDU/CSU:

$$p^{Koeff}_{\text{SPD, CDU}} = \frac{\exp(0.8131 - 0.1756)}{1 + \exp(0.8131 - 0.1756)} = 0.6542,$$

$$p^{Worth}_{\text{SPD, CDU}} = \frac{0.2520}{0.2520 + 0.1332} = 0.6542.$$

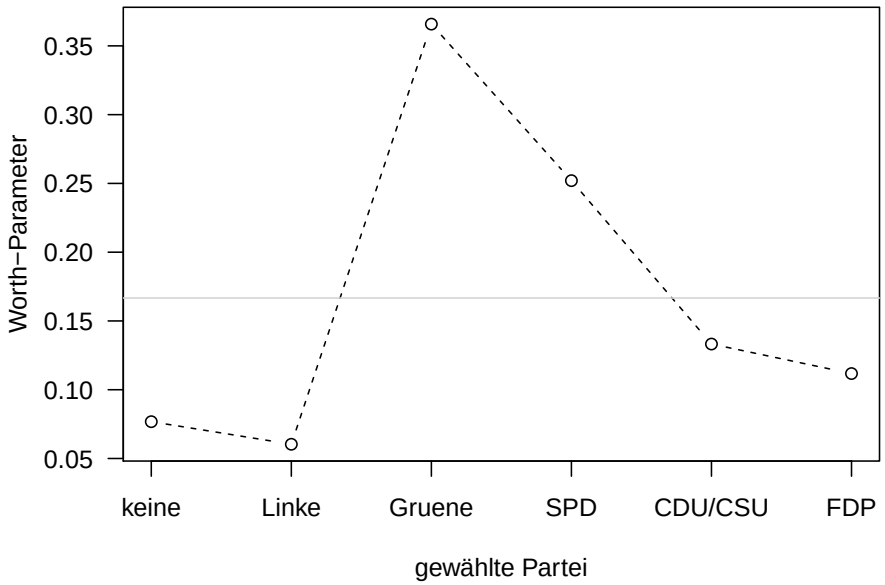

Abbildung 6.3.: Darstellung der Worth-Parameter des einfachen BTL-Modells für den `GermanParties2009`-Datensatz. Die horizontale Gerade entlang dem Wert 1/6 markiert den Referenzwert eines indifferenten Beurteilers.

6.2. Anwendung für das Heterogenitätsmodell

Um eine Rangfolge der Parteien über ein Paarvergleichsexperiment zu erhalten, bei dem der Einfluss der Versuchspersonen berücksichtigt wird, wird das Heterogenitätsmodell berechnet. Für den zweiten Schritt des Algorithmus zur Schätzung des Heterogenitätsmodells, der Schätzung des GLMMs, werden ebenso wie im Anwendungsbeispiel bezüglich der Lernmethoden in Kapitel 5 für das Argument `X.noise` der `R`-Funktion `glmer` $n \cdot \binom{m}{2} = 192 \cdot 15 = 2880$ responseunabhängige Zufallszahlen gezogen. Zur Gewährleistung der Reproduzierbarkeit der Ergebnisse der Schätzung des Heterogenitätsmodells geht dieser Ziehung ein Seed voran. Aufgrund von sich ergebenden Schwankungen in den Schätzergebnissen durch die Wahl unterschiedlicher Seeds, werden mit 100 zufälligen Seeds, welchen für die Reproduzierbarkeit der Ergebnisse wiederum ein Seed von 160 unterliegt, Heterogenitätsmodelle gerechnet und die sich jeweils ergebenden Itemparameterschätzer über die 100 Modelle hinweg gemittelt, um ein allgemeingültiges Heterogenitätsmodell zu erhalten. Die Schätzer dieses gemittelten Modells und die geschätzten Itemparameter für beispielhafte Modelle mit den zugrunde liegenden Seeds von 279, 6185 und 13389 werden in Ta-

belle 6.4 gegenüber gestellt. Referenzobjekt in allen Modellen ist das Item FDP, weshalb der Wert für den zugehörigen Schätzer grundsätzlich Null annimmt. Für 92 der dem gemittelten Heterogenitätsmodell zugrunde liegenden 100 Modelle mit unterschiedlichem Seed, und somit auch für das gemittelte Modell selbst, ergibt sich als Rangfolge aus den Werten der Schätzer, dass die Versuchspersonen zum Zeitpunkt der Befragung anstehenden Bundestagswahl 2009 am häufigsten die Partei Bündnis 90/Die Grünen wählen würden, gefolgt von der SPD, den Linken und der CDU/CSU. Zudem präferieren die Versuchsteilnehmer eine Wahlenthaltung vor der Abgabe ihrer Stimme für die FDP.

	keine	Linke	Gruene	SPD	CDU/CSU	FDP
Schätzer bei Seed = 279:	0.0008	0.2132	4.1682	3.2317	0.0351	0
Schätzer bei Seed = 6185:	0.0196	0.2557	4.2839	3.3293	0.0302	0
Schätzer bei Seed = 13389:	0.0169	0.2556	4.2997	3.3391	0.0271	0
gemittelte Schätzer aus 100 Seeds:	0.0125	0.2435	4.2620	3.3086	0.0289	0
Rangzuweisung:	5	3	1	2	4	6

Tabelle 6.4.: Koeffizientenschätzer und Rangzuordnung des Heterogenitätsmodells im `GermanParties2009`-Datensatz. Den Schätzern unterliegen Modellschätzungen mit jeweils unterschiedlichen Seeds für die Zufallsvariable `X.noise`, welche für die Funktion `glmer` in `R` benötigt wird.

Aufgrund der nah beieinander liegenden Schätzwerte für die Items der Wahlenthaltung („keine“), CDU/CSU und FDP, ergibt sich für acht der 100 Modelle, die dem gemittelten Heterogenitätsmodell zugrunde liegen und bei denen lediglich der Seed verschieden ist, eine etwas abweichende Rangordnung. Die ersten drei Ränge bleiben für alle 100 Modelle mit unterschiedlichen Seeds unverändert, jedoch präferieren die Versuchsteilnehmer in sechs Modellen die Wahl der CDU/CSU vor einer Wahl der FDP und diese wiederum vor einer Wahlenthaltung:

Rangordnung-Ausreißer-Modell I:

$$\text{Gruene} \succ \text{SPD} \succ \text{Linke} \succ \text{CDU/CSU} \succ \text{FDP} \succ \text{keine}.$$

In zwei der 100 Modellen wird sich lieber der Wahl enthalten bevor eine Stimme für die CDU/CSU bzw. FDP abgegeben wird:

Rangordnung-Ausreißer-Modell II:

$$\text{Gruene} \succ \text{SPD} \succ \text{Linke} \succ \text{keine} \succ \text{CDU/CSU} \succ \text{FDP}.$$

Da die Werte der Koeffizientenschätzer für die Items CDU/CSU, FDP und „keine“ so nah beieinander liegen, sind in der grafischen Darstellung der Koeffizientenschätzer für alle 100 Modelle in Abbildung 6.4 die Rangordnung-Ausreißer-Modelle mit bloßem Auge kaum identifizierbar.

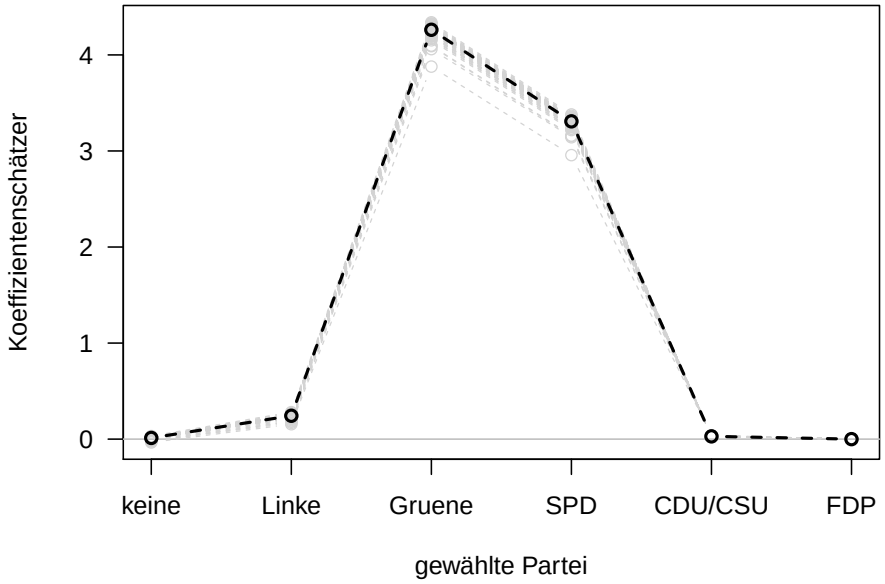

Abbildung 6.4.: Darstellung von Koeffizientenschätzern aus 100 Heterogenitätsmodellen mit unterschiedlichem Seed für den `GermanParties2009`-Datensatz. Die schwarze Kurve verbindet die Mittelwerte aller Schätzer. Die horizontale Gerade entlang der Null markiert das Referenzobjekt FDP.

Der Vergleich der Schätzungen zwischen dem gemittelten Heterogenitätsmodell und dem einfachen BTL-Modell aus Abschnitt 6.1 zeigt eine Änderung in der Rangordnung der Parteipräferenzen. Während die Ränge eins (Bündnis 90/Die Grünen), zwei (SPD) und fünf („keine“) in beiden Modellen gleich vergeben werden, nehmen die anderen Wahlmöglichkeiten unterschiedliche Ränge ein.

Im Gegensatz zur Darstellung der Koeffizientenschätzer im Anwendungsbeispiel bezüglich der Lernmethoden (Abbildung 5.4) in Kapitel 5, bei welcher anhand der starken Streuung der grauen Kurven deutlich zu sehen ist, dass die Werte der Schätzer stark variieren, ist in Abbildung 6.4 zwischen den Werten der Schätzer der 100 Modelle mit unterschiedlichem Seed mit bloßem Auge kaum ein Unterschied zu erkennen. Ähnliche Schätzwerte für die Modelle mit unterschiedlichem Seed ergeben selbstverständlich auch ähnliche individuelle Wahrscheinlichkeiten p_{irs} für die Bevorzugung von Objekt r gegenüber Objekt s (siehe Gleichung (4.2)), daher beschränkt sich hier die beispielhafte Berechnung für individuelle Wahrscheinlichkeiten auf das gemittelte Heterogenitätsmodell.
Während im BTL-Modell grundsätzlich für alle $i = 1, \ldots, n$ Versuchspersonen ein Wert von $\alpha_i = 1$ angenommen wird, für welchen angenommen wird, dass er dem wahren Reizwert der zu vergleichenden Items entspricht, variieren im Heterogenitätsmodell die Werte der Personenparameter für alle Versuchpersonen und drücken damit die individuelle Wahrnehmung der Reizstärke der Objekte aus. In Tabelle 6.5 sind die Wahrscheinlichkeiten $p_{i,\text{SPD, CDU}}$ für die Präferenz der SPD gegenüber der CDU/CSU sowohl für das BTL-Modell als auch für das gemittelte Heterogenitätsmodell bei verschiedenen Werten für den Personenparameter angegeben.

$\boldsymbol{p_{i,\text{SPD, CDU}}}$	$\alpha_i = -0.5$	$\alpha_i = 0.5$	$\alpha_i = 1$
einfaches BTL-Modell:	–	–	65.4 %
gemitteltes Heterogenitätsmodell aus 100 Seeds:	16.2 %	83.8 %	96.4 %

Tabelle 6.5.: Berechnete Wahrscheinlichkeiten $p_{i,\text{SPD, CDU}}$ (`GermanParties2009`-Datensatz) für die Schätzung des BTL- und des gemittelten Heterogenitätsmodell bei Annahme unterschiedlicher Personenparameter.

Im Vergleich zum BTL-Modell ist die Wahrscheinlichkeit für eine bevorzugte Wahl der SPD gegenüber der CDU im gemittelten Heterogenitätsmodell um 31 Prozentpunkte höher und liegt damit bei

$$p_{i,\text{SPD, CDU}} = \frac{\exp(1 \cdot (3.3086 - 0.0289))}{1 + \exp(1 \cdot (3.3086 - 0.0289))} = 0.9637 \approx 96.4\%.$$

Nimmt eine Versuchsperson die Reizwerte der Parteien nicht so stark wahr, wie sie tatsächlich vorliegen ($\alpha_i = 0.5$), verringert sich die Wahrscheinlichkeit für eine be-

vorzugte Wahl der SPD gegenüber der CDU im gemittelten Heterogenitätsmodell auf 83.8 %. Eine Person mit Heterogenitätsfaktor $\alpha_i = -0.5$ nimmt im Gegensatz dazu die Reize der Parteien SPD und CDU/CSU vertauscht wahr, wodurch sich der Gegenwert, also eine Wahrscheinlichkeit von 16.2 % ergibt.

Da das Heterogenitätsmodell nicht auf der `R`-Funktion `btReg.fit` aus dem Paket `psychotools` (Zeileis et al., 2012) basiert, kann die Funktion `worth` zur Bestimmung der Worth-Parameter nicht angewendet werden. Es lassen sich die Worth-Parameter aber natürlich wieder händisch über Gleichung (5.1) bestimmen:

$$\hat{\pi}_r = \frac{\exp(\hat{\gamma}_r)}{\sum_{r=1}^{m} \exp(\hat{\gamma}_r)}.$$

Bei Betrachtung der im Heterogenitätsmodell geschätzten Standardabweichung des Personenparameters, welcher die Unterschiede der Wahrnehmung der Reizstärken der Items schätzt, ergeben sich für alle gefitteten Heterogenitätsmodelle mit unterschiedlichem Seed ähnliche Werte. Trotzdem lässt sich ähnlich wie in Abbildung 5.5 für das Anwendungsbeispiel bezüglich der Lernmethoden in der folgenden Abbildung 6.5 der Sachverhalt einer durch die Rangordnung-Ausreißer-Modelle verzerrten gemittelten geschätzten Standardabweichung veranschaulichen. Es wird deutlich, dass die höheren Werte für die geschätzte Standardabweichung der zwei Rangordnung-Ausreißer-Modelle II (orange eingefärbt) die erheblich niedrigeren Werte der sechs Rangordnung-Ausreißer-Modelle I (rot gekennzeichnet) nicht ausgleichen und somit eine Verzerrung des Mittelwerts der geschätzten Standardabweichung nach unten entsteht. Der Blick auf den Wertebereich der Ordinate in Abbildung 6.5 zeigt jedoch, dass sich der Wert der gemittelten geschätzten Standardabweichung von 0.7527 lediglich ab der dritten Nachkommastelle vom Median der 100 geschätzten Standardabweichungen mit einem Wert von 0.7533 unterscheidet. Um einen allgemeingültigen Wert für die Schätzung der Unterschiede in der Wahrnehmung der Reizstärken der sechs Wahlmöglichkeiten zu erhalten, lassen sich daher beide statistischen Kennzahlen verwenden.

Es kann für das vorliegende Anwendungsbeispiel aufgrund der Erkenntnis aus der Simulation des Heterogenitätsmodells in Kapitel 4, dass ab einer wahren Standardabweichung von $\sigma = 0.6$ die geschätzten Standardabweichungen der Personenparameter eine deutliche Tendenz zur Unterschätzung aufweisen, wieder davon ausge-

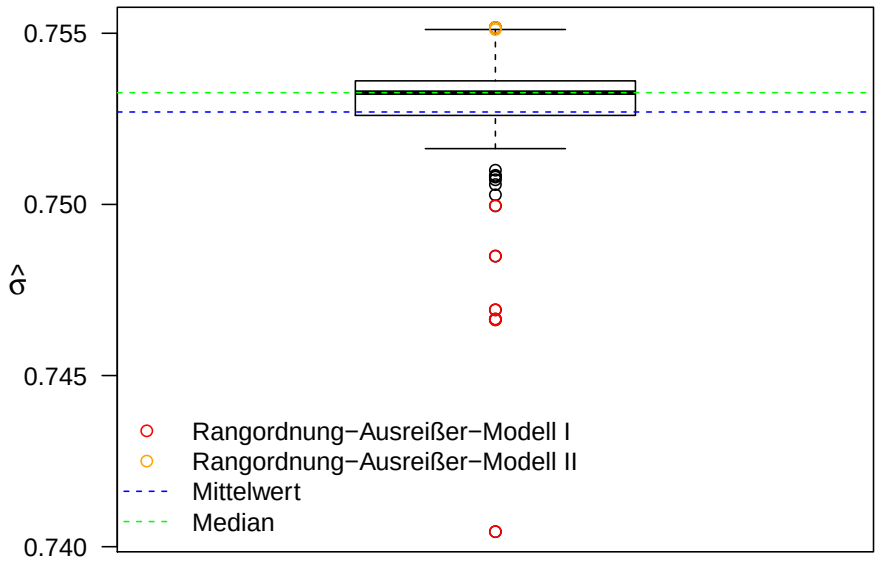

Abbildung 6.5.: Verteilung der geschätzten Standardabweichung von 100 Heterogenitätsmodellen mit unterschiedlichen Seeds für den `GermanParties2009`-Datensatz. Die geschätzten Standardabweichungen der Rangordnung-Ausreißer-Modelle I und II führen zu einer Verzerrung der geschätzten Standardabweichung des gemittelten Heterogenitätsmodells.

gangen werden, dass die wahre Standardabweichung für die Versuchspersonen im Paarvergleichsexperiment bezüglich der Parteipräferenzen in Wirklichkeit größer ist als der allgemeingültige Wert der geschätzten Standardabweichung von ca. 0.75. Für Modelle mit hoher wahrer Standardabweichung wird es umso schwieriger, die Personen- und Itemparameter des Heterogenitätsmodells zu schätzen, was sich in der zunehmenden Anzahl an benötigten Iterationen bis zur Konvergenz der Schleife im Heterogenitätsmodell widerspiegelt. Die durchschnittliche Anzahl an Iterationen in der Iterationsschleife bei den Schätzungen des Heterogenitätsmodells im vorliegenden Anwendungsbeispiel beträgt 93 und spricht daher für eine wahre Standardabweichung größer als 0.8.

Aus den Ergebnissen des Anwendungsbeispiels bezüglich der Parteipräferenzen lässt sich schlussfolgern, dass mit dem vorangestellten Seed für die Ziehung der Zufallsvariablen `X.noise`, welche die Funktionsfähigkeit der `R`-Funktion `glmer` gewährleistet, sehr ähnliche Modelle geschätzt werden können. Die sich ergebenden Rangordnung-Ausreißer-Modelle unterscheiden sich lediglich in der Vergabe der letzten drei Ränge

in der Rangordnung aufgrund der nah beieinander liegenden Schätzwerte für die Parameter der Items CDU/CSU, FDP und „keine“. Es ergeben sich jedoch für alle 100 Modelle sowohl für die individuellen Wahrscheinlichkeiten p_{irs} für die Bevorzugung von Objekt r gegenüber Objekt s als auch für die geschätzte Standardabweichung $\hat{\sigma}$ ähnliche Werte. Für das Paarvergleichsexperiment bezüglich der Parteipräferenzen kann also das gemittelte Heterogenitätsmodell durchaus als allgemeingültiges Modell herangezogen werden. Es sollten jedoch die Simulationsergebnisse bezüglich des Heterogenitätsmodells aus Abschnitt 4.5 im Hinterkopf behalten werden, die besagen, dass das Heterogenitätsmodell numerisch nicht stabil ist und im Vergleich zum BTL-Modell unpräzisere Itemparameterschätzer hervorbringt.
Die Tatsache einer von Null verschieden geschätzten Standardabweichung weist erneut darauf hin, dass die im Modell angenommene Heterogenität der Versuchspersonen tatsächlich vorliegt. Ob das Heterogenitätsmodell jedoch in der Lage dazu ist, diese genau zu erfassen, bleibt fraglich.

7. Zusammenfassung und Schlussfolgerung

In der angewandten Statistik werden Paarvergleiche, die zur Bewertung von Objekten bzw. Items über nicht direkt messbare, subjektive Kriterien (z.B. Geschmack oder Attraktivität) durchgeführt werden, häufig über Bradley-Terry-Luce Modelle angepasst. Mit Hilfe der forced-choice-Befragungstechnik (Möhring and Schlütz, 2010), in welcher kein „Unentschieden“ zulässig ist, resultiert für die zu bewertenden Objekte über das subjektive Kriterium eine eindeutige Rangordnung.
Im Bradley-Terry-Luce Modell wird angenommen, dass jedes Objekt wahre Präferenzen auf einer subjektiven Skala besitzt. Für alle Versuchspersonen wird dabei die gleiche Präferenzskala vorausgesetzt und angenommen, dass sie die Reizstärken der Objekte in einem Paarvergleich gleich wahrnehmen. In einem einfachen Bradley-Terry-Luce Modell wird also Homogenität der Versuchpersonen angenommen. Da jedoch in der Realität jede Person die Reizstärken der zu vergleichenden Objekte unterschiedlich wahrnimmt, ist es sinnvoll, bei Paarvergleichen, die mehreren Personen vorgelegt werden, den Einfluss des Probanden in geeigneter Weise zu berücksichtigen. Durch Erweiterung des Bradley-Terry-Luce Modells um einen personenspezifischen Parameter α_i entsteht somit das Heterogenitätsmodell.

Dass sich das GLM als alternatives Schätzverfahren für BTL-Modelle eignet, wird sowohl im Anwendungsbeispiel bezüglich der Lernmethoden in Abschnitt 5.1 als auch im Anwendungsbeispiel der Parteipräferenzen in Abschnitt 6.1 gezeigt, da bei der Schätzung des einfaches BTL-Modells durch Verwendung einer bereits in `R` implementierten Funktion dieselben Schätzer resultieren. Die Schätzung von Paarvergleichsdaten über generalisierte lineare Modelle in Kapitel 3 ermöglicht zudem eine einfache Einbindung des Personenparameters.

Anhand einer Simulation der Schätzung eines BTL-Modells wird in Kapitel 4 zu-

nächst die Notwendigkeit der Einbeziehung eines Parameters für die Heterogenität der Versuchspersonen in das BTL-Modell verdeutlicht. Der Simulation liegt der datengenerierende Prozess DGP(α_i) zugrunde, in dem neben der Definition der relevanten Größen für die Simulation eines Paarvergleichsmodells die wahren Werte für die Itemparameter aus einer Gleichverteilung gezogen werden und für den Personenparameter α_i die Normalverteilung angenommen wird. Die grafische Darstellung des MSE der Itemparameterschätzer für diese Modellsimulation zeigt deutlich, dass mit zunehmender Veränderung der Wahrnehmung der Versuchspersonen für die Reizstärke eines Objekts auch der MSE ansteigt und es somit zu einer schlechteren Präzision der Schätzung kommt. Mit dem Heterogenitätsmodell, in dem die Schätzung des Personenparameters α_i zusätzlich zur Schätzung der Itemparameter durchgeführt werden soll, wird eine präzisere Schätzung des BTL-Modells und somit eine Verringerung des MSE der Itemparameterschätzer angestrebt.

Die Schätzung des Heterogenitätsmodells erfolgt durch einen Algorithmus, in dem abwechselnd bis zur Konvergenz die Itemparameter mittels GLM und die Personenparameter über generalisierte lineare gemischte Modelle angepasst werden. Aufbauend auf der Datensimulation mit zugrunde liegendem DGP(α_i) wird in einer erneuten Datensimulation der Algorithmus zur Schätzung des Heterogenitätsmodells angewendet und jeweils ein Heterogenitätsmodell für vorgegebene Werte der Standardabweichung des Personenparameters geschätzt, da die Heterogenität im Heterogenitätsmodell durch unterschiedliche Werte der Standardabweichung der Personenparameter simuliert wird. Mit $\sigma = 0$ bestehen keine Unterschiede der individuellen Wahrnehmung der Objekt-Reizstärken in einem Paarvergleich, für $\sigma = 0.8$ liegen große Unterschiede in der individuellen Wahrnehmung vor. Für Modelle mit hoher Standardabweichung wird es somit schwieriger, die Personen- und Itemparameter des Heterogenitätsmodells zu schätzen, was sowohl durch die steigende Anzahl an Iterationen bis zur Konvergenz der Schleife als auch an fallweise fehlender Konvergenz der Iterationsschleife während der Modellschätzung sichtbar wird.
Im Algorithmus zur Schätzung des Heterogenitätsmodells werden die Schritte zur Schätzung der Item- und Personenparameter so oft iteriert bis die Veränderung zum vorangegangen Rechenschritt kleiner ist als ein vorher spezifizierter Grenzwert. Nach Konvergenz der Schleife wird von präzisen Schätzern für die Personen- und Itemparameter ausgegangen. Die Überprüfung dieser Annahme durch Betrachtung des MSE der Parameterschätzer jeder Iteration in jedem Simulationsdurchlauf lässt je-

doch aufgrund der im Mittel steigenden Werte auf unpräzise Schätzungen schließen. Lediglich bei Personenparameterschätzern in Modellen mit Standardabweichungen von $\sigma = 0.6$ bis $\sigma = 0.8$ kann aufgrund der abnehmende Folge der MSE-Werte innerhalb eines Simulationsdurchlaufes von präzisen Schätzungen gesprochen werden.

Der Vergleich der Schätzungen der Item- und Personenparameter im Heterogenitätsmodell und im BTL-Modell mit zugrunde liegendem DGP(α_i) durch die MSE der Parameterschätzer ergibt für die Schätzung von α_i, dass das Heterogenitätsmodell präzisere Schätzungen $\hat{\alpha}_i$ hervorbringt als das BTL-Modell mit DGP(α_i). Es ist also durchaus sinnvoll, einen Personenparameter für die Heterogenität der Versuchspersonen in das BTL-Modell mitaufzunehmen. Für die Schätzung der Itemparameter ergibt der Vergleich zwischen Heterogenitätsmodell und BTL-Modell mit DGP(α_i) eine stärkere Steigung des MSE der Parameterschätzer im Heterogenitätsmodell, was für eine unpräzise Schätzung der Itemparameter im Heterogenitätsmodell spricht. Das Ziel einer Verringerung des MSE der Itemparameterschätzer im modifizierten BTL-Modell wurde also nicht erreicht.

Die Schätzung der Standardabweichung der Personenparameter, also die geschätzte individuelle Wahrnehmung der Unterschiede in den Reizstärken der Objekte, erfolgt in der Simulation des Heterogenitätsmodells für wahre Standardabweichungen von 0.2 bis 0.5 ziemlich genau. Es ist jedoch eine deutliche Tendenz zur Unterschätzung ab einer wahren Standardabweichung von 0.6 erkennbar. In den mit den Daten des `trdel`-Datensatzes und des `GermanParties2009`-Datensatzes angepassten Heterogenitätsmodellen ergeben sich allgemeingültige Werte der geschätzten Standardabweichung von ca. 0.85 und 0.75. Es kann das Simulationsergebnis für das Heterogenitätsmodell, dass dieser Wert unterschätzt wird und der den Daten zugrunde liegende wahre Wert für die Standardabweichung höher ist, durch die fehlende Konvergenz der Iterationsschleife bzw. der sehr hohen Anzahl an benötigten Iterationen bis zur Konvergenz der Schleife aller für diese Datensätze berechneten Heterogenitätsmodelle bestätigt werden.

Aus dem Vergleich des Heterogenitätsmodells mit dem BTL-Modell mit zugrunde liegendem DGP(α_i) lässt sich schlussfolgern, dass das Heterogenitätsmodell numerisch nicht stabil ist. Mit steigender Standardabweichung wird das Modell zunehmend instabiler.

Der für die Reproduzierbarkeit der Ergebnisse verwendete Seed zur Ziehung der für die R-Funktion `glmer` benötigten Zufallsvariable `X.noise` kann in den Anwendungsbeispielen gegebenenfalls zu stark voneinander abweichenden Ergebnissen der Schätzungen des Heterogenitätsmodells und somit ebenso zu unterschiedlichen Rangordnungen der Objekte führen. Insgesamt ergibt sich grundsätzlich in den beiden Anwendungsbeispielen jeweils eine unterschiedliche Rangfolge aus den zu bewertenden Items für das Heterogenitätsmodell im Vergleich zum einfachen BTL-Modell.

Die numerische Instabilität des Modells und die zusätzlich gegebene Unsicherheit einer unpräzisen Modellschätzung, welche aus dem benötigten Seed zur Ziehung der Zufallsvariablen folgt, die für die Modellberechnung in R unabkömmlich ist, sprechen gegen eine Anwendung des in dieser Arbeit vorgestellten Algorithmus zur Schätzung des Heterogenitätsmodells. Letztendlich wird aus den Anwendungsbeispielen die Erkenntnis gewonnen, dass sowohl die geänderte Rangfolge der Objekte für das Heterogenitätsmodell im Vergleich zum geschätzten BTL-Modell als auch die Tatsache einer von Null verschieden geschätzten Standardabweichung darauf hinweisen, dass die im Modell angenommene Heterogenität der Versuchspersonen tatsächlich vorliegt. Das Heterogenitätsmodell, welches auf dem in dieser Arbeit vorgestellten Algorithmus zur Modellschätzung basiert, ist jedoch nicht in der Lage, diese genau zu erfassen.

Eine weitere Möglichkeit zur Schätzung des Heterogenitätsmodells unter zugrunde liegender Verteilungsannahme (4.4) ist die Anwendung einer numerischen Annäherung durch Gauss-Hermite Integration auf die log-Likelihood für alle Personen. Ein Ansatz für dieses Schätzverfahren ist in Tutz (1989) gegeben.

A. Anhang

A.1. Simulationsmaske: einfaches BTL-Modell über GLM

Zunächst muss die Funktion geladen werden, mit der die Designmatrix erstellt wird.

```
> source("Designmatrix.R")
```

Für die Simulation sind folgende Größen anzugeben:

- `m`: Anzahl der Objekte/Items,
- `n`: Anzahl der Personen,
- `P`: Anzahl der Simulationsdurchläufe.

Mit diesen Angaben kann die Designmatrix `X` an die gewünschte Simulation angepasst werden.

```
> X <- designmatrix(m, n, lastcol=FALSE)
```

Die Anzahl der Paarvergleiche wird mit `I` bezeichnet und über die Gleichung $m \cdot (m-1)/2$ berechnet.
Für die Reproduzierbarkeit der Simulationsergebnisse wird ein Startwert mit der Funktion `set.seed` gesetzt.

```
> for(i in 1:P){

   # Ziehen von gleichverteilten Reizstärken
   gamma <- runif(m-1, min=-3, max=3)

   # Linearer Prädiktor
```

```
  eta <- X %*% gamma
  pi <- exp(eta)/(1+exp(eta))

  # Responsevektor
  yi <- rbinom(n*I, 1, pi)
  # GLM
  model <- glm.fit(x = X, y = yi, family = binomial(link="logit"),
                   intercept = FALSE)
  # Ausgabe der Modellkoeffizienten
  print(model$coefficients)
}
```

Im Ergebnis werden für die einzelnen Simulationsdurchläufe die mittels GLM geschätzten Itemparameter ausgegeben.

A.2. Simulationsmaske: BTL-Modell mit DGP(α_i) über GLM

Zunächst muss die Funktion geladen werden, mit der die Designmatrix erstellt wird.

```
> source("Designmatrix.R")
```

Für die Simulation sind folgende Größen anzugeben:

- `m`: Anzahl der Objekte/Items,
- `n`: Anzahl der Personen,
- `P`: Anzahl der Simulationsdurchläufe.

Mit diesen Angaben kann die Designmatrix `X` an die gewünschte Simulation angepasst werden.

```
> X <- designmatrix(m, n, lastcol=FALSE)
```

Die Anzahl der Paarvergleiche wird mit `I` bezeichnet und über die Gleichung $m \cdot (m-1)/2$ berechnet.
Für die Reproduzierbarkeit der Simulationsergebnisse wird ein Startwert mit der Funktion `set.seed` gesetzt.

Für den Vergleich der Simulation für die Schätzung eines BTL-Modells mit DGP(α_i) und der Simulation des Heterogenitätsmodells (vgl. die Simulationsmaske in Anhang A.3) werden für jeden Simulationsdurchlauf für beide Simulationen die gleichen Zufallszahlen gezogen mittels

```
> seeds <- round(abs(rnorm(P)*10000)).
```

Anschließend wird eine Sequenz für verschiedene Werte von Standardabweichungen des Personenparameters α_i aufgestellt

```
> sigma.seq <- seq(0,0.8, by=0.1)
```

und eine leere Matrix zum Abspeichern der Simulationsergebnisse für den MSE der einzelnen Itemparameter erstellt

```
> Z <- matrix(0, nrow=length(sigma.seq), ncol=m-1).
```

Über die folgende Simulationsmaske können daraufhin BTL-Modelle mittels generalisierter linearer Modelle geschätzt werden, wobei durch die Einbeziehung des Parameters α_i die Personenspezifität berücksichtigt werden soll.

```
> for(u in 1:length(sigma.seq)){

   # Hilfsmatrix für die MSE-Berechnung
   mse.help <- matrix(0, nrow=P, ncol=(m-1))

   for(i in 1:P){

     # Ziehen von normalverteilten Personenparametern zur
     # Simulation der Heterogenität der Personen
     set.seed(seeds[i])
     alpha <- rnorm(n, 0, sd=sigma.seq[u])
     alpha.u <- 1+alpha

     # Ziehen von gleichverteilten Reizstärken
     gamma <- runif(m-1, min=-3, max=3)

     # Hilfsmatrix, um alpha_i vor eta einzubauen
     alpha.help <- rep(alpha.u, times=I)
```

```
    # Linearer Prädiktor
    eta <- X %*% gamma
    pi <- exp(alpha.help * eta)/(1+exp(alpha.help * eta))

    # Responsevektor
    yi <- rbinom(n*I, 1, pi)
    # GLM
    model <- glm.fit(x = X, y = yi, family = binomial(link="logit"),
                     intercept = FALSE)
    mse.help[i,] <- (model$coefficients - gamma)^2
  }
  print(u)
  Z[u,] <- colMeans(mse.help)
}
```

Anschließend kann der MSE der Itemparameterschätzer wie in Abbildung 4.1 grafisch dargestellt und interpretiert werden.

A.3. Simulationsmaske: Heterogenitätsmodell

Im Folgenden wird die Simulationsmaske zur Schätzung eines Heterogenitätsmodells für eine wahre Standardabweichung von 0.4 vorgestellt. Analog kann zur Gewinnung der Schätzer für andere wahre Standardabweichungen vorgegangen werden.

Für die Schätzung der Personenparameter mittels random effects models wird das Paket `lme4` (Bates et al., 2013) geladen (hier verwendet: Version 0.999999-2).

```
> library(lme4)
```

Anschließend muss die Funktion geladen werden, mit der die Designmatrix erstellt wird.

```
> source("Designmatrix.R")
```

Für die Simulation sind folgende Größen anzugeben:

- `m`: Anzahl der Objekte/Items,
- `n`: Anzahl der Personen,

- `P`: Anzahl der Simulationsdurchläufe.

Mit diesen Angaben kann die Designmatrix `X` an die gewünschte Simulation angepasst werden.

```
> X <- designmatrix(m, n, lastcol=FALSE)
```

Die Anzahl der Paarvergleiche wird mit `I` bezeichnet und über die Gleichung $m \cdot (m-1)/2$ berechnet.

Es werden vorab leere Vektoren, Matrizen und Listen erstellt, um in ihnen die im Laufe der Simulation geschätzten (Zwischen-)Ergebnisse zu speichern.

```
> GLM.Koeff <- matrix(0, nrow=P, ncol=(m-1))
> mse.help <- matrix(0, nrow=P, ncol=(m-1))
> mse.alpha.help <- matrix(0, nrow=P, ncol=n)
> mse.btl.help <- matrix(0, nrow=P, ncol=n)
> A <- matrix(0, nrow=n, ncol=P)
> B <- matrix(0, nrow=n, ncol=P)
> S04 <- mse.alpha <- mse.gamma <- c()
> mse.alpha.last <- mse.gamma.last <- Iter <- c()
> list.mse.alpha <- list.mse.gamma <- list()
```

Für die Schätzung mittels random effects model wird ein Vektor für die Personenidentität erstellt.

```
> ID <- rep(1:n, times=I)
```

Für die Reproduzierbarkeit der Simulationsergebnisse wird ein Startwert mit der Funktion `set.seed` gesetzt. Für den Vergleich der Simulation für die Schätzung eines BTL-Modells mit DGP(α_i) (vgl. die Simulationsmaske in Anhang A.2) und der Simulation des Heterogenitätsmodells werden für jeden Simulationsdurchlauf für beide Simulationen die gleichen Zufallszahlen gezogen mittels

```
> seeds <- round(abs(rnorm(P)*10000)).
```

In der folgenden Schleife werden in `P` Simulationsdurchläufen sowohl die Itemparameter $\boldsymbol{\gamma}$ über generalisierte lineare Modelle als auch die Personenparameter α_i

mittels generalisierter linearer gemischter Modelle (GLMM) geschätzt.

```
> for(j in 1:P){

   print(j)

   # Ziehen von normalverteilten Personenparametern zur
   # Simulation der Heterogenität der Personen
   set.seed(seeds[j])
   alpha <- rnorm(n, 0, sd=0.4)
   alpha1 <- 1+alpha
   alpha1.help <- rep(alpha1, times=I)

   # Ziehen von gleichverteilten Reizstärken
   gamma <- runif(m-1, min=-3, max=3)

   # Linearer Prädiktor
   eta <- alpha1.help * (X %*% gamma)
   pi <- exp(eta)/(1+exp(eta))

   # Responsevektor (beinhaltet die Reizstärken und die
   # Personenheterogenität)
   yi <- rbinom(n*I, 1, pi)

   alpha.new <- alpha.old <- rep(0, n)
   gamma.hat <- rep(0, m-1)

   # Erzeugen einer Zufallsvariable, welche unabhängig vom Response
   # ist, für die Sicherstellung der Funktionsfähigkeit der folgenden
   # glmer-Funktion
   X.noise <- rnorm(length(yi))

   # Setzen einer Grenze für das Konvergieren der folgenden
   # Iterationsschleife
   epsilon <- 1e-4
   thresh <- 1

 ###################################
 # Beginn der Iterationsschleife #

   g <- 1
   g.vec <- c()
   g.max <- 101

   while((g < g.max) & (epsilon < thresh)){
```

```
    gamma.old <- gamma.hat
    alpha.old <- alpha.new

    X.new <- diag(rep(1+alpha.new, I)) %*% X
    # im ersten Durchlauf ist X.new = X, also alpha = 0

    # GLM
    model <- glm.fit(x = X.new, y = yi, family = binomial(link = "logit"),
                     intercept = FALSE)
    # Modellkoeffizienten
    gamma.hat <- (model$coefficients)

    # Aufbereitung für das random effects model
    offset <- X %*% gamma.hat

    formula.RE <- yi ~ 0 + X.noise + (0 + offset|ID)

    # random effects model
    model.RE <- glmer(formula.RE, family = binomial, offset = offset)

    # Extrahieren der random effects
    alpha.new <- ranef(model.RE)$ID[,1]

    mse.alpha[g] <- mean((alpha.new - alpha)^2)
    mse.gamma[g] <- mean((gamma.hat - gamma)^2)

    # Wird die Iterationsschleife weiter ausgeführt?
    thresh <- sqrt(sum((gamma.old - gamma.hat)^2))/sqrt(sum(gamma.old^2))
    # Ausgabe des Wertes der Iterationsbedingung
    cat("threshold =", thresh, "\n")

    g <- g+1

  }

  g.vec[j] <- g
  Iter[j] <- g.vec[j]

# Ende der Iterationsschleife #
###############################

  # Ausgabe der MSE der Schätzer und der Anzahl der Iterationen
  cat("mse.alpha =", mse.alpha, "\n",
      "mse.gamma =", mse.gamma, "\n",
      "Iterationen =", g.vec[j]-1, "\n")
```

```
   # Speichern der Ergebnisse der letzten Iteration für alle P
   # Simulationsdurchläufe in den vorbereiteten Matrizen, Vektoren
   # und Listen
   list.mse.alpha[[j]] <- mse.alpha
   list.mse.gamma[[j]] <- mse.gamma

   A[,j] <- alpha.new
   S04[j] <- as.numeric(summary(model.RE)@REmat[,4])
   GLM.Koeff[j,] <- gamma.hat

   # MSE der Werte der letzten Iteration
   mse.alpha.last[j] <- mean((A[,j]-alpha)^2)
   mse.gamma.last[j] <- mean((GLM.Koeff[j,]-gamma)^2)

   mse.alpha.help[j,] <- (A[,j]-alpha)^2
   mse.btl.help[j,] <- (B[,j]-alpha)^2
   # alle Elemente in Matrix B sind 0 und entsprechen "geschätzten"
   # alphas im BTL-Modell mit DGP(alpha_i)

   mse.help[j,] <- (GLM.Koeff[j,]-gamma)^2

 # Vektor-reset
 mse.alpha <- mse.gamma <- c()
 }
```

Anschließend werden aus den Ergebnismatrizen und -vektoren diejenigen Simulationsdurchläufe herausgefiltert, in denen die Iterationsschleife nicht konvergiert.

```
> if(any(Iter==g.max)){
   nc <- which(Iter==g.max)
   list.mse.alpha <- list.mse.alpha[-nc]
   list.mse.gamma <- list.mse.gamma[-nc]
   A <- A[,-nc]
   B <- B[,-nc]
   S04 <- S04[-nc]
   GLM.Koeff <- GLM.Koeff[-nc,]
   mse.alpha.last <- mse.alpha.last[-nc]
   mse.gamma.last <- mse.gamma.last[-nc]
   mse.alpha.help <- mse.alpha.help[-nc,]
   mse.btl.help <- mse.btl.help[-nc,]
   mse.help <- mse.help[-nc,]
 }
```

A.4. Modelloutputs

A.4.1. Einfaches BTL-Modell: `trdel`-Datensatz

Output zur Modellschätzung über GLM

```
Call:
NULL

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.779  -1.627   1.265   1.643   5.029

Coefficients:
     Estimate Std. Error z value Pr(>|z|)
[1,] -0.06424    0.09582  -0.670    0.503
[2,] -0.97285    0.09782  -9.946  < 2e-16 ***
[3,] -0.50554    0.09536  -5.302 1.15e-07 ***
[4,] -1.39358    0.10256 -13.588  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2744.9  on 1980  degrees of freedom
Residual deviance: 2441.2  on 1976  degrees of freedom
AIC: 2449.2

Number of Fisher Scoring iterations: 4
```

Output zur Modellschätzung über `R`-Funktion `btReg.fit`

```
Bradley-Terry regression model

Parameters:
   Estimate Std. Error z value Pr(>|z|)
CO -0.06424    0.09582  -0.670    0.503
TV -0.97285    0.09782  -9.946  < 2e-16 ***
GL -0.50554    0.09536  -5.302 1.15e-07 ***
AU -1.39358    0.10256 -13.588  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -1221 (df = 4)
```

A.4.2. Einfaches BTL-Modell: `GermanParties2009`-Datensatz

Output zur Modellschätzung über GLM

```
Call:
NULL

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.273  -1.540  -1.165   1.689   7.061

Coefficients:
     Estimate Std. Error z value Pr(>|z|)
[1,] -0.37556    0.08902  -4.219 2.45e-05 ***
[2,] -0.61607    0.09102  -6.768 1.30e-11 ***
[3,]  1.18576    0.09507  12.473  < 2e-16 ***
[4,]  0.81310    0.09067   8.967  < 2e-16 ***
[5,]  0.17562    0.08750   2.007   0.0447 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3992.5  on 2880  degrees of freedom
Residual deviance: 3433.6  on 2875  degrees of freedom
AIC: 3443.6

Number of Fisher Scoring iterations: 4
```

Output zur Modellschätzung über `R`-Funktion `btReg.fit`

```
Bradley-Terry regression model

Parameters:
        Estimate Std. Error z value Pr(>|z|)
keine   -0.37556    0.08902  -4.219 2.45e-05 ***
Linke   -0.61607    0.09102  -6.768 1.30e-11 ***
Gruene   1.18576    0.09507  12.473  < 2e-16 ***
SPD      0.81310    0.09067   8.967  < 2e-16 ***
CDU/CSU  0.17562    0.08750   2.007   0.0447 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -1717 (df = 5)
```

A.5. Elektronischer Anhang

Der elektronische Anhang bezieht sich auf die dieser Arbeit beigefügten CD und beinhaltet die Ordner **Bilder**, **Masterarbeit**, **R** und **Sweave**.

Der Ordner **Bilder** enthält die Dateien derjenigen Abbildungen, die nicht mittels der R-Funktion `Sweave` (Leisch, 2002) erstellt wurden, sondern eigens mit Microsoft Office Word kreiert oder aus einer anderen Quelle entnommen wurden.

Der Ordner **Masterarbeit** beinhaltet gleichnamige Dateien mit dem Inhalt der hier dem Leser vorliegenden Masterarbeit sowohl im TEX-Format als auch im mit dem Programm LATEX konvertierten pdf-Format. Die Datei „Literatur.bib“ beinhaltet die Liste aller für die Arbeit verwendeten Quellen. Die Datei „chicago.bst“ dient der Formatierung des Literaturverzeichnisses und wird ebenso wie die Datei „Sweave.sty“, welche der Einbindung von R-Output dient, zwingend zur Erzeugung der Masterarbeit im pdf-Format benötigt. Die Datei „titlepage.tex“ erzeugt die Titelseite der Masterarbeit. Neben diesen aufgeführten Dateien ist der Unterordner **Kapitel** vorhanden, welcher das Abkürzungsverzeichnis, den Abstract sowie die einzelnen Kapitel der Masterarbeit in TEX-Format enthält. Diese Dateien werden zur besseren Übersicht separat in die Masterarbeit eingefügt.

Die Auswertungen für diese Arbeit mit der Statistik-Software R befinden sich im Ordner **R**, welcher folgende Unterordner enthält:

- **Anwendungsbeispiele:** In diesem Ordner sind Dateien mit R-Code zu den beiden Anwendungsbeispielen aus den Kapiteln 5 und 6 enthalten. Dateien mit dem Namensbestandteil „trdel“ beziehen sich auf das Anwendungsbeispiel der Lernmethoden, Dateien mit dem Namensbestandteil „GP09“ auf das Beispiel mit den Parteipräferenzen.

- **HetMod:** Dieser Ordner beinhaltet die Dateien mit R-Code für die Simulationen des Heterogenitätsmodells zu unterschiedlichen Werten der wahren Standardabweichung des Personenparameters bei je 100 Simulationsdurchläufen („HetMod0.R“ bis „HetMod08.R“). Ebenso befinden sich in diesem Ordner ein R-Workspace mit allen Ergebnissen dieser 9 Simulationen des Heterogenitätsmodells („HetMod.RData“), der Code zur Simulation des Heterogenitäts-

modells bei einer Standardabweichung von $\sigma = 0.4$ und nur 15 Simulationsdurchläufen („HetMod04_P15.R“) und der zu all diesen Simulationen zugehörige R-Code für die Grafiken des Kapitels 4 („Grafiken_HetMod.R“).

Nicht in den Unterordnern enthalten ist neben dem Code der Funktion für die Designmatrix („Designmatrix.R“) auch der Code für die grafische Darstellung des Logit-Modells („Grafiken_allgemein.R“). Mit den weiteren R-Dateien „Simulation_BTL.R“ und „simulation_BTL_alpha.R“ lassen sich sowohl das einfache BTL-Modell als auch das BTL-Modell mit zugrunde liegendem DGP(α_i) über generalisierte lineare Regression simulieren.

Die Dateien zur Einbindung von in R erstellten Outputs und Grafiken in LaTeX mittels der Funktion `Sweave` (Leisch, 2002) befinden sich im Ordner **Sweave**. Zu unterscheiden ist hier zwischen R-Dateien mit dem Präfix „Sweave_...“ und Dateien gleichen Namens ohne diesen Präfix sowohl im R- als auch im TeX-Format. Dateien mit gleichem Namensbestandteil gehören zueinander. Die TeX-Dateien sind Erzeugnis der Ausführung der „Sweave_...“-Dateien. Im Unterordner **Grafiken** sind die pdf-Dateien der erzeugten R-Grafiken enthalten.

Literaturverzeichnis

Agresti, A. (2002). *Categorical Data Analysis* (2. ed.). Wiley Series in Probability and Statistics. John Wiley & Sons.

Altobelli, C. F. (2011). *Marktforschung: Methoden - Anwendungen - Praxisbeispiele* (2. ed.). Konstanz: UVK.

AMS-Österreich (2013). Arbeitsmarktservice Österreich: Dienstleistungsunternehmen des öffentlichen Rechts. Available from: http://www.ams.at, last accessed on 18.12.2013.

Arrow, K. J. (1963). *Social Choice and Individual Values* (2. ed.). New York: John Wiley & Sons.

Bates, D., M. Maechler, and B. Bolker (2013). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2. Available from: http://CRAN.R-project.org/package=lme4, last accessed on 18.12.2013.

Böckenholt, U. (2001a). Hierarchical Modeling of Paired Comparison Data. *Psychological Methods 6*(1), 49–66.

Böckenholt, U. (2001b). Thresholds and Intransitivities in Pairwise Judgments: A Multilevel Analysis. *Journal of Educational and Behavioral Statistics 26*(3), 269–282.

Bortz, J. and N. Döring (2009). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. ed.). Heidelberg: Springer Verlag.

Bradley, R. A. (1976). Science, Statistics, and Paired Comparison. *Biometrics 32*(2), 213–239.

Bradley, R. A. and M. E. Terry (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika 39*(3/4), 324–345.

Bundeswahlleiter (2013). Ergebnisse der Bundestagswahl 2009. Available from: http://www.bundeswahlleiter.de/de/bundestagswahlen/BTW_BUND_09/ergebnisse/, last accessed on 18.12.2013.

Cattelan, M. (2012). Models for Paired Comparison Data: A Review with Emphasis on Dependent Data. *Statistical Science 27*(3), 412–433.

Chan, V. (2011). Prediction Accuracy of Linear Models for Paired Comparisons in Sports. *Journal of Quantitative Analysis in Sports 7*(3). Artikel 18.

Cohen, A. (2004). Models of Choice. Folien zur Vorlesung „Modeling Behavior". Department of Psychology, University of Massachusetts. Available from: http://people.umass.edu/alc/course_pages/fall_2004/modeling_behavior/lectures/choice.ppt, last accessed on 18.12.2013.

Collett, D. (2003). *Modelling Binary Data* (2. ed.). Boca Raton, USA: Chapman & Hall/CRC.

Colonius, H. (1980). Representation and Uniqueness of the Bradley-Terry-Luce Model for Pair Comparisons. *British Journal of Mathematical and Statistical Psychology 33*(1), 99–103.

Coombs, C. H. (1964). *A Theory of Data*. New York: John Wiley & Sons.

Coombs, C. H., R. M. Dawes, and A. Tversky (1970). *Mathematical Psychology: An Elementary Introduction*. Englewood Cliffs, N.J.: Prentice-Hall.

Courcoux, P. and M. Semenou (1997). Preference Data Analysis Using a Paired Comparison Model. *Food Quality and Preference 8*(5/6), 353–358.

Critchlow, D. E. and M. A. Fligner (1991). Paired Comparison, Triple Comparison, and Ranking Experiments as Generalized Linear Models, and their Implementation. *Psychometrika 56*(3), 517–533.

David, H. (1988). *The Method of Paired Comparisons* (2. ed.). Number 41 in Griffin´s statistical monographs. New York: Oxford University Press.

Davidson, R. R. (1970). On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments. *Journal of the American Statistical Association 65*(329), 317–328.

Davidson, R. R. and R. J. Beaver (1977). On Extending the Bradley-Terry Model to Incorporate Within-Pair Order Effects. *Biometrics 33*(4), 693–702.

Davidson, R. R. and P. H. Farquhar (1976). A Bibliography of the Method of Paired Comparisons. *Biometrics 32*(2), 241–252.

Debreu, G. (1960). Review of Individual Choice Behavior: A Theoretical Analysis by R. D. Luce . *The American Economic Review 50*(1), 186–188.

Decker, R. and R. Wagner (2002). *Marketingforschung: Methoden und Modelle zur Bestimmung des Käuferverhaltens*. München: Redline Wirtschaft bei Verlag Moderne Industrie.

Dey, D., S. K. Ghosh, and B. K. Mallick (2000). *Generalized linear models: A Bayesian Perspective*. New York: Dekker.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data* (5. ed.). Oxford University Press, USA.

Dittrich, R., B. Francis, R. Hatzinger, and W. Katzenbeisser (2006). Modelling Dependency in Multivariate Paired Comparisons: A Log-linear Approach. *Mathematical Social Sciences 52*, 197–209.

Dittrich, R., R. Hatzinger, and W. Katzenbeisser (1998). Modelling the Effect of Subject-Specific Covariates in Paired Comparison Studies With an Application to University Rankings. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 47*(4), 511–525.

Dittrich, R., R. Hatzinger, and W. Katzenbeisser (2001). Corrigendum: Modelling the Effect of Subject-Specific Covariates in Paired Comparison Studies With an Application to University Rankings. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 50*(2), 247–249.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models* (2. ed.). Boca Raton, USA: Chapman & Hall/CRC.

Dörsam, P. (2007). *Grundlagen der Entscheidungstheorie: anschaulich dargestellt* (5. ed.). Heidenau: PD-Verlag.

Fahrmeir, L., T. Kneib, and S. Lang (2009). *Regression - Modelle, Methoden und Anwendungen* (2. ed.). Heidelberg: Springer Verlag.

Fahrmeir, L., R. Künster, I. Pigeot, and G. Tutz (2011). *Statistik: Der Weg zur Datenanalyse* (7. ed.). Berlin: Springer Verlag.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2. ed.). New York: Springer Verlag.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data* (2. ed.). Cambridge, Mass.: MIT Press.

Frisenfeldt Tuesen, K. (2007). Analysis of Ranked Preference Data. Masterthesis, Technical University of Denmark, Informatics and Mathematical Modelling, Kongens Lyngby. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.152&rep=rep1&type=pdf, last accessed on 18.12.2013.

Gaul, W. (1979). Zur Methode der paarweisen Vergleiche und ihrer Anwendung im Marketingbereich. In R. Henn (Ed.), *Operations Research Verfahren = Methods of Operations Research*, Volume 35, pp. 123–139. Königstein/Ts.: Verlag Anton Hain Meisenheim GmbH. III. Symposium on Operations Research: Tagungsbericht.

Gediga, G. (1998). *Skalierung: Eine Einführung in die Methodik zur Entwicklung von Test- und Meßinstrumenten in den Verhaltenswissenschaften*. Number 5 in Osnabrücker Schriften zur Psychologie. Münster: Lit Verlag.

Glenn, W. A. and H. A. David (1960). Ties in Paired-Comparison Experiments Using a Modified Thurstone-Mosteller Model. *Biometrics 16*(1), 86–109.

Guttman, L. (1946). An Approach for Quantifying Paired Comparisons and Rank Order. *The Annals of Mathematical Statistics 17*(2), 144–163.

Hatzinger, R. and R. Dittrich (2012). prefmod: An R Package for Modeling Preferences Based on Paired Comparisons, Rankings, or Ratings. *Journal of Statistical Software 48*(10), 1–31. Available from: http://www.jstatsoft.org/v48/i10/, last accessed on 18.12.2013.

Huang, T.-K., R. C. Weng, and C.-J. Lin (2006). Generalized Bradley-Terry Models and Multi-Class Probability Estimates. *Journal of Machine Learning Research 7*(1), 85–115.

Hunter, D. R. (2004). MM Algorithms for Generalized Beradley-Terry Models. *The Annals of Statistics 32*(1), 384–406.

Kendall, M. G. and B. Babington Smith (1940). On the Method of Paired Comparisons. *Biometrika 31*(3/4), 324–345.

Koehler, K. J. and H. Ridpath (1982). An Application of a Biased Version of the Bradley-Terry-Luce Model to Professional Basketball Results. *Journal of Mathematical Psychology 25*(3), 187–205.

Laird, N. M. and J. H. Ware (1982). Random-Effects Models for Longitudinal Data. *Biometrics 38*(4), 963–974.

Leisch, F. (2002). Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis. In W. Härdle and B. Rönz (Eds.), *Compstat 2002 — Proceedings in Computational Statistics*, pp. 575–580. Physica Verlag, Heidelberg. Available from: http://www.stat.uni-muenchen.de/~leisch/Sweave, last accessed on 18.12.2013.

Ligges, U. (2008). *Programmieren mit R* (3. ed.). Berlin, Heidelberg: Springer Verlag.

Luce, R. D. (1959). *Individual Choice Behavior*. New York: John Wiley & Sons.

Luce, R. D. (1977). The Choice Axiom after Twenty Years. *Journal of Mathematical Psychology 15*(3), 215–233.

Lukas, J. (1991). BTL-Skalierung verschiedener Geschmacksqualitäten von Sekt. *Zeitschrift für experimentelle und angewandte Psychologie 38*(4), 605–619.

Malhotra, N. K. (2010). *Marketing Research: An Applied Orientation* (6. ed.). Pearson Education.

Mallows, C. L. (1957). Non-Null Ranking Models. I. *Biometrika 44*(1/2), 114–130.

Matthews, J. N. S. and K. P. Morris (1995). An Application of Bradley-Terry-type Models to the Measurement of Pain. *Applied Statistics 44*(2), 243–255.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models, Second Edition* (2. ed.). Monographs on Statistics and Applied Probability. Taylor & Francis.

Möhring, W. and D. Schlütz (2010). *Die Befragung in der Medien- und Kommunikationswissenschaft* (2. ed.). Wiesbaden: VS Verlag.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Oberfeld, D., H. Hecht, U. Allendorf, and F. Wickelmaier (2009). Ambient Lighting Modifies the Flavor of Wine. *Journal of Sensory Studies 24*(6), 797–832.

Pleskac, T. J. (2012, September). Decision and Choice: Luce´s Choice Axiom. Department of Psychology, Michigan State University. Available from: https://www.msu.edu/~pleskact/research/papers/pleskac_inpress_Luce.pdf, last accessed on 18.12.2013.

R Core Team and contributors worldwide (2013). *stats: The R Stats Package*. R Foundation for Statistical Computing. R package version 3.0.1.

R Development Core Team (2013). R: A Language and Environment for Statistical Computing. Version 3.0.1. Available from: http://www.R-project.org, last accessed on 18.12.2013.

Rao, P. V. and L. L. Kupper (1967). Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model. *Journal of the American Statistical Association 62*(317), 194–204.

Rinne, H. (2008). *Taschenbuch der Statistik* (4. ed.). Frankfurt am Main: Verlag Harri Deutsch.

Roskam, E. E. (1968). *Metric Analysis of Ordinal Data in Psychology*. Holland: Voorschoten.

Roskam, E. E. (1983). Allgemeine Datentheorie. In H. Feger and J. Bredenkamp (Eds.), *Enzyklopädie der Psychologie*, Volume 3 Messen und Testen of *Forschungsmethoden der Psychologie*, Chapter 1. Göttingen: Hogrefe.

Schöll, B. and S. Veith (2011). Lernstilerhebung und bevorzugte Lernmethoden im arbeitsmarktpolitischen Schulungsbereich - eine Präferenzanalyse mittels Paarvergleichsmethode in R. Masterthesis, Wirtschaftsuniversität Wien, Institut für Statistik und Mathematik.

SDI-Research (2013). Forced-Choice Befragungstechnik. Available from: http://www.sdi-research.at/lexikon/forced-choice.html, last accessed on 18.12.2013.

Strobl, C., F. Wickelmaier, and A. Zeileis (2011). Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning. *Journal of Educational and Behavioral Statistics 36*(2), 135–153.

Suppes, P. and J. L. Zinnes (1963). Basic Measurement Theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology*, Volume 1, pp. 1–76. New York: Wiley.

Tack, W. H. (1983). Psychophysische Methoden. In H. Feger and J. Bredenkamp (Eds.), *Enzyklopädie der Psychologie*, Volume 3 Messen und Testen of *Forschungsmethoden der Psychologie*, Chapter 6. Göttingen: Hogrefe.

Thurstone, L. L. (1927). A Law of Comparative Judgement. *Psychological Review 34*, 273–286.

Train, K. E. (2003). *Discrete Choice Methods With Simulation*. Cambridge University Press.

Tutz, G. (1986). Bradley-Terry-Luce Models with an Ordered Response. *Journal of Mathematical Psychology 30*(3), 306–316.

Tutz, G. (1989). *Latent Trait-Modelle für ordinale Beobachtungen*. Number 30 in Lehr- und Forschungstexte Psychologie. Berlin: Springer Verlag.

Tutz, G. (2000). *Die Analyse kategorialer Daten: Anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression*. München, Wien: Oldenbourg Wissenschaftsverlag.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge u.a.: Cambridge University Press.

Tutz, G. (2013, April). BTL. Institut für Statistik, Ludwig-Maximilians-Universität München, unveröffentlichtes Manuskript.

Universitätsstadt Tübingen (2013). Ergebnisse der Bundestagswahl 2009. Available from: http://www.tuebingen.de/wahl/html/bt09zs.html, last accessed on 18.12.2013.

Verbeke, G. and G. Molenberghs (2009). *Linear mixed models for longitudinal data*. New York: Springer Verlag.

Zeileis, A., C. Strobl, and F. Wickelmaier (2012). *psychotools: Infrastructure for Psychometric Modeling*. R package version 0.1-4. Available from: http://CRAN.R-project.org/package=psychotools, last accessed on 18.12.2013.

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift 29*, 436–460.

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit mit dem Titel

Modellierung der Heterogenität in Bradley-Terry-Luce Modellen

selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen (einschließlich elektronischer Quellen und dem Internet) direkt oder indirekt übernommene Gedanken sind ausnahmslos als solche kenntlich gemacht.
Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und auch noch nicht physisch oder elektronisch veröffentlicht.

Ort, Datum (Melanie Poppe)