Robuste Bayes-Regression mit Mengen von Prioris – Ein Beitrag zur Statistik unter komplexer Unsicherheit

Diplomarbeit

von
Gero Walter

Betreuung:
Prof. Dr. Th. Augustin

27. Oktober 2006

Dank an...

Th. Augustin für die bestmöglichste Betreuung

A. Peters, GSF und KORA für die Hilfe bezüglich des Datensatzes

die Mitarbeiter und Ehemaligen des Instituts, insbesondere C. Strobl, S. Kaiser, Th. Kneib, S. Breitner, S. Pilz, I. Kreuzmair und B. Maxa, für vielfältige Hilfe und Unterstützung,

die Bevölkerung des Projektraums, insbesondere M. Obermeier und M. Schomaker, für moralischen Beistand und Nachsicht

meine Familie

Sicher ist, dass nichts sicher ist.

Karl Valentin

Inhaltsverzeichnis

1 Komplexe Unsicherheit **1**
- 1.1 Einführung . . . 1
- 1.2 Bertrands Paradoxon und nichtinformative Priori-Verteilungen . . . 5
- 1.3 Ellsbergs Paradoxon und die Begrenztheit der Kolmogorovschen Axiomatik 7
- 1.4 Weitere Argumente für Statistik unter komplexer Unsicherheit . . . 9
  - 1.4.1 Argumente aus epistemischer Sicht . . . 9
  - 1.4.2 Argumente aus ontologischer Sicht . . . 11
  - 1.4.3 Prior-data conflict . . . 11
- 1.5 Modellierung komplexer Unsicherheit . . . 13
  - 1.5.1 Prinzipielles Vorgehen . . . 13
  - 1.5.2 Mengen klassischer Wahrscheinlichkeitszuordnungen . . . 13
  - 1.5.3 Intervallwahrscheinlichkeit . . . 14
  - 1.5.4 Die Theoriegebäude von Walley und Weichselberger . . . 14
  - 1.5.5 Abgeleitete Größen . . . 16
- 1.6 Verwandte Theorien . . . 17
- 1.7 Das zur Anwendung kommende Modell . . . 18

2 Beispiel: Das Imprecise Dirichlet Model **22**
- 2.1 Multinomiale Daten . . . 22
- 2.2 Bayes-Lernen . . . 23
- 2.3 Konjugierte Verteilungen . . . 24
- 2.4 Inferenzprinzipien und alternative Modellierungen . . . 27
- 2.5 Das Imprecise Dirichlet Model für $\theta_1, \ldots, \theta_k$ . . . 29
- 2.6 Inferenz mit dem IDM . . . 31
  - 2.6.1 Prädiktive Inferenz . . . 33
  - 2.6.2 Parametrische Inferenz . . . 36

3 Das Stichproben-Modell von Quaeghebeur und de Cooman **40**
- 3.1 Exponentialfamilien . . . 40
- 3.2 Bayes-Lernen in Exponentialfamilien . . . 42
- 3.3 Die ‚Impräzisierung' . . . 45

4 Bayes-Regression unter komplexer Unsicherheit **50**
- 4.1 Einführung . . . 50
  - 4.1.1 Zur Wahl des Intervallwahrscheinlichkeitsmodells . . . 50
  - 4.1.2 Regression . . . 51

4.1.3 Vorgehen . . . 54
4.2 Das Normal-Modell für die lineare Regression . . . 54
4.2.1 Das Modell . . . 54
4.2.2 Die Normalverteilung als Exponentialfamilie . . . 57
4.2.3 Anwendung des Aufdatierungsmodells von Quaeghebeur und de Cooman auf das Normal-Modell . . . 60
4.2.4 Die ‚Impräzisierung' des Normal-Modells . . . 66
4.3 Formulierung des Normal-Modells im Fall $p = 2$ . . . 69
4.3.1 Direkte Prüfung . . . 71
4.3.2 Prüfung mittels eines Hilfssatzes . . . 73
4.3.3 Das Modell für $\rho = 0$ . . . 77
4.4 Anwendung des Modells für $p = 2$, $\rho = 0$ . . . 85
4.4.1 ‚Große' Koeffizienten, kleine Varianz . . . 86
4.4.2 ‚kleine' Koeffizienten, große Varianz . . . 101
4.4.3 Multikollinearität . . . 105
4.5 Datenbeispiel . . . 110
4.5.1 Die AIRGENE-Studie . . . 110
4.5.2 Analyse bei ‚datengeleiteter' Wahl von A . . . 112
4.5.3 Analyse bei der Wahl von $n^{(0)}$ gemäß einer Strategie der Modellierung von sehr schwachem Vorwissen . . . 117

5 Zusammenfassung und Ausblick **122**
5.1 Zusammenfassung . . . 122
5.2 Ausblick . . . 127

Anhang **130**
A.1 Das Normal-InversGamma-Modell für die lineare Regression . . . 130
A.1.1 Das Modell . . . 130
A.1.2 Die Normal-InversGamma-Verteilung als Exponentialfamilie . . . 134
A.1.3 Versuch der Anwendung des Aufdatierungsmodells von Quaeghebeur und de Cooman auf das NIG-Modell . . . 137
A.2 Syntax für die Berechnung der Ergebnisse in Kapitel 4.4 und 4.5 . . . 140

Literatur **141**

Abbildungsverzeichnis **144**

Erklärung zur Urheberschaft **147**

1 Komplexe Unsicherheit

Diese Arbeit beschäftigt sich mit der Berücksichtigung von komplexer Unsicherheit bei der bayesianischen Schätzung von Parametern einer linearen Regression. Der Begriff ‚Modelle unter komplexer Unsicherheit' steht allgemein für statistische Modelle, die den Anspruch haben, die übliche Wahrscheinlichkeitstheorie zu erweitern, indem sie etwa durch die Betrachtung von Mengen von Verteilungen nicht-stochastische Unsicherheit (Ambiguität) adäquat einbeziehen. Aufgrund dieser Motivation wird in dieser Arbeit die potentielle Ungenauigkeit von Priori-Wissen in expliziter Weise in die Modellierung einbezogen, indem statt einer einzigen eine Menge von Priori-Verteilungen bayesianisch aufdatiert wird. Kapitel 1 beschreibt den Begriff der komplexen Unsicherheit näher und stellt die Überlegungen und Argumente dar, die für die Einbeziehung komplexer Unsicherheit in der statistischen Modellierung sprechen. In Kapitel 2 wird als ausführliches Beispiel ein Modell unter komplexer Unsicherheit für multinomial-verteilte Daten beschrieben, dass von [Walley 1996] entwickelt wurde und vielfache Anwendung gefunden hat (siehe z.B. [Bernard 2005] oder [Bernard 2007]). In Kapitel 3 folgt die Beschreibung eines allgemeineren Ansatzes von [Quaeghebeur und de Cooman 2005] für die Berücksichtigung komplexer Unsicherheit bei der bayesianischen Analyse von Stichproben, der in Kapitel 4 auf die in dieser Arbeit untersuchte Fragestellung verallgemeinert wird. In Kapitel 5 werden schließlich die gewonnenen Erkenntnisse zusammengefasst und ein Ausblick auf mögliche Erweiterungen und weitergehende Ansätze gegeben.

1.1 Einführung

> „Statistik unter komplexer Unsicherheit: Der gemeine Statistiker mag sich fragen: Wozu brauche ich so etwas? Ist das nicht unnötig kompliziert? Die ‚normale' Statistik funktioniert doch; sie kann alle unsicheren Phänomene beschreiben. Antwort: Ja, aber nicht immer, manchmal funktioniert sie eben nicht, oder sie funktioniert nur scheinbar. Dann macht man es sich zu einfach, und die Realität ist eben so komplex, dass eine entsprechend komplizierte Modellierung notwendig ist."

Stellen Sie sich vor, Sie kehren von einer Reise in die Tropen zurück und leiden unter seltsamen Beschwerden. Sie suchen den nächsten verfügbaren Arzt auf, der jedoch leider kein Spezialist für Tropenkrankheiten ist. Er hat aber Zugriff auf ein medizinisches Expertensystem, das bei der Eingabe von Symptomen die plausibelsten Krankheitsursachen ausgibt und so Hausärzten hilft, auch seltene Krankheiten für die Diagnose in Betracht ziehen zu können.

Solche medizinischen Expertensysteme werden erstellt, indem mehrere anerkannte Spezialisten Krankheitsursachen und Symptome in Verbindung bringen. Diese Informationen werden in Wahrscheinlichkeitsaussagen übersetzt, die dann in eine Datenbank aufgenommen werden können. Dabei kann es im Rahmen der Erhebung der Wahrscheinlichkeitsaussagen zu folgender Situation kommen:

- „Wir haben hier also Symptome, die mit hoher Sicherheit auf Krankheit A oder Krankheit B schließen lassen; bei einer Skala von 0 bis 1 sagten sie 0.9. Wie sicher sind Sie sich denn, dass es sich nur um Krankheit A handelt?“

• „Nicht sehr sicher, ich würde sagen, so etwa 0.1.“

- „Und wie sicher sind Sie sich, dass tatsächlich Krankheit B vorliegt?“

• „Das ist auch ziemlich unsicher, aber vielleicht ein bisschen wahrscheinlicher als Krankheit A, sagen wir 0.2.“

Wenn das System die Einschätzungen der Experten getreulich wiedergeben soll, kann es also nicht auf ‚echten‘ Wahrscheinlichkeitsaussagen basieren, da hier gilt, dass

$$P(A) + P(B) \neq P(A \cup B)\,,$$

was dem Additivitätsaxiom von Kolmogorov widerspricht.

Nun könnte man einfach davon ausgehen, dass die Wahrscheinlichkeitszuordnungen dieses Spezialisten für Tropenkrankheiten nicht rational sind. Damit macht man es sich aber vermutlich zu einfach, denn der Spezialist wird sich von den oben genannten Aussagen sicher nicht abbringen lassen, da sie auf seinen jahrelangen Erfahrungen in der Diagnose beruhen. Könnte es vielleicht nicht umgekehrt so sein, dass die Regeln, die ein normales Wahrscheinlichkeitsmaß einhalten muss, zu einengend und deshalb nicht in der Lage sind, reale Phänomene in all ihrer Verschiedenheit und Komplexität angemessen zu modellieren?

Die Statistik unter komplexer Unsicherheit hat sich entwickelt, um ebendiese Grenzen der klassischen Statistik (im Sinne der auf den Axiomen von Kolmogorov basierenden Wahrscheinlichkeitsmaße) zu überschreiten und damit auch Situationen modellieren zu können, in denen die klassische Statistik nur unbefriedigende Lösungen ermöglicht.

Die klassische Statistik stellt eine weithin anerkannte Methodik dar, mit der Unsicherheit bezüglich verschiedenster Phänomene in der Realität modelliert werden kann, und ist aus wissenschaftstheoretischer Sicht (bisher) die einzige kodifizierte und nachprüfbare (intersubjektive) Methode, die induktives Schließen möglich macht.

Seit langem schon gibt es aber Hinweise, dass die Instrumente der klassischen Statistik nicht ausreichen, um in allen Situationen eine rationale und begründete Antwort geben zu können. In den letzten drei Jahrzehnten sind aus diesem Grund verschiedene

Ansätze entstanden, die den Methodenkreis der Statistik um Verfahren erweitern sollen, die in Unsicherheitssituationen, in denen die klassische Statistik keine oder nur schlecht begründbare Antworten liefern kann, eine sinnvolle Modellierung möglich machen. Expertensysteme wie das anfangs beschriebene gibt es wirklich (siehe z.B. [Schill 1990]), und sie basieren in den meisten Fällen nicht auf dem kolmogorovschen Wahrscheinlichkeitsbegriff; das Modell, das in Kapitel 2 beschrieben werden wird, hat Auswirkungen auf die Art, wie randomisierte klinische Studien auf ethische Weise durchgeführt werden sollten [Walley 1996, Kap. 5].

[Weichselberger 2001] hat eine Theorie der Wahrscheinlichkeitsmaße entwickelt, die eine konsequente Verallgemeinerung der Axiome von Kolmogorov darstellt. Die zentrale Idee ist dabei, einem Ereignis A nicht einen festen Wahrscheinlichkeitswert zuzuordnen, sondern ein Intervall. Mit dieser Theorie der Intervallwahrscheinlichkeit lässt sich die oben genannte Einschätzung des Spezialisten für Tropenkrankheiten adäquat modellieren; das ebengenannte Modell von Walley kann auch als Intervallwahrscheinlichkeitsmodell angesehen werden.

Im Grunde ist es eine naheliegende Idee, die Wahrscheinlichkeitsbewertung $P(A)$ für ein Ereignis A nicht als eine Zahl, sondern als Intervall anzugeben. Spätestens seit Mitte des 19. Jahrhunderts gibt es schon entsprechende Überlegungen; auch die Kritik an der klassischen Statistik, insbesondere an der Laplace´schen Regel vom unzureichenden Grund (siehe Abschnitt 1.2) hat eine lange Tradition. Einen ideengeschichtlichen Überblick zur Entwicklung des Wahrscheinlichkeitsbegriffs und seiner mathematischen Formalisierung sowie der Ansätze, die über die klassische Modellierung von Unsicherheitssituationen hinausgehen, bietet [Weichselberger 2001].

Die klassische Statistik ignoriert bei der Modellierung von Unsicherheitssituationen für gewöhnlich, dass es verschiedene Arten von Unsicherheit gibt. Es können (mindestens) drei Typen von Unsicherheit unterschieden werden:

- **ideale Stochastizität (Zufälligkeit)**: Situationen unter diesem Typ von Unsicherheit entsprechen ‚idealen Lotterien', bei denen die Gesetzmäßigkeiten, welche die Unsicherheit bezüglich beobachtbarer Phänomene verursachen, genau bekannt sind und perfekten Zufallsmechanismen entsprechen. Ist ein sechsseitiger Spielwürfel fair, so kann die Wahrscheinlichkeit für ein Ergebnis genau angegeben werden: Die Wahrscheinlichkeit, in einem Wurf eine ‚Eins' zu erzielen, liegt genau bei einem Sechstel.

- **Ambiguität (Unbestimmtheit)**: Dieser Begriff meint die Unsicherheit, die sich nicht durch gesetzmäßige Begriffe modellieren lässt und sich über nicht-stochastische Prozesse ergibt, für die sich keine festen Regeln ermitteln lassen. Es handelt sich dabei um Unschärfe in den Wahrscheinlichkeitszuordnungen, und daher oft um Unsicherheit bezüglich der Grundannahmen, die für die Modellierung mittels idealer Stochastizität nötig sind. So kann im obigen Beispiel unklar sein, ob der Würfel wirklich fair ist.

- **Vagheit**: Mit diesem Begriff sei die Unsicherheit beschrieben, die sich ergibt, wenn die Zugehörigkeit zu Mengen nicht eindeutig ermittelbar ist; es handelt sich also um Unschärfe in den Mengen oder Ereignissen. So könnte beispielsweise die Beschriftung der Seiten des Würfels so abgenutzt sein, dass die Unterscheidung der ‚Drei' und der ‚Fünf' bei schlechter Beleuchtung schwierig wird.

Die Berücksichtigung dieser verschiedenen Quellen von Unsicherheit wird in der klassischen Statistik in der Regel vernachlässigt, sie ist auf die Modellierung von idealer Stochastizität beschränkt. Ist unsicher, ob der Würfel fair ist, können Ergebnisse von Modellen, die nur auf idealer Stochastizität beruhen, in die Irre führen. Statistik unter komplexer Unsicherheit macht es hingegen möglich, die anderen beiden Quellen von Unsicherheit in das Kalkül mit einzubeziehen. Speziell für die Modellierung von Vagheit wurde die Theorie der ‚fuzzy sets' entwickelt, die viele Berührungspunkte mit den Theorien zur Modellierung von komplexer Unsicherheit hat, die in diesem Kapitel kurz vorgestellt werden. Einer dieser Ansätze soll in dieser Arbeit seine Anwendung auf die Schätzung der Regressionsparameter in einem linearen Modell finden.

Die Notwendigkeit, über die klassische Statistik hinauszugehen, hat sich an dem Auftreten von verschiedenen Paradoxien bei ihrer Anwendung gezeigt, die in den nächsten beiden Kapiteln kurz beschrieben werden sollen. Weitere Vorteile der Modellierung von komplexer Unsicherheit werden dann im Kapitel 1.4 thematisiert. In Kapitel 1.5 wird skizziert, wie man zu einer allgemeineren statistischen Theorie gelangen kann.

Für Anhänger einer objektivistischen Denkschule führt das Bertrandsche Paradoxon, das die Problematik der Modellierung von Nichtwissen behandelt, zur Entwicklung von verschiedenen Lösungen, Bayes-Inferenz trotz fehlendem Vorwissen zu betreiben, nämlich den sogenannten nichtinformativen Priori-Verteilungen. Die Gedankengänge der verschiedenen Ansätze zur Auffindung einer solchen nichtinformativen Priori-Verteilung sind jeweils für sich rational und nachvollziehbar, können aber zu völlig unterschiedlichen Lösungen führen. Dass alle diese Lösungen, die integraler Bestandteil der Methoden des objektiven Bayesianismus sind, sich aber letztlich als unbefriedigend erweisen, macht die ausführliche Abhandlung von Rüger in [Rüger 1999, Kap. 2.5] deutlich.

Für Vertreter subjektivistischer Denkschulen in der Statistik veranschaulicht das Paradoxon von Ellsberg die Beschränkungen der klassischen Modellierung von Wahrscheinlichkeit. Ellsbergs Untersuchungen zeigen, dass es Präferenzordnungen (erstellt von Wissenschaftlern, deren Rationalität nicht ernsthaft in Frage gestellt werden kann) gibt, die durch keine klassische Wahrscheinlichkeitsbewertung modelliert werden können.

1.2 Bertrands Paradoxon und nichtinformative Priori-Verteilungen

Nichtinformative Priori-Verteilungen sind ein Versuch, die Möglichkeit zu schaffen, bayesianische Methoden auch dann anzuwenden, wenn kein oder nur sehr schwaches Vorwissen herrscht. Die Theorie der nichtinformativen Priori-Verteilungen bietet dabei verschiedene Motivationsansätze, die dann zu Regeln zur Erstellung einer nichtinformativen Priori-Verteilung führen. Das Bertrandsche Paradoxon zeigt beispielhaft, dass man im Zuge solcher Überlegungen zu völlig unterschiedlichen Ergebnissen kommen kann.

Der Ausgangspunkt des Bertrandschen Paradoxons (siehe z.B. [Rüger 1999, Kap. 2.5, Beispiel 2.13]) ist die Fragestellung: „Wie groß ist die Wahrscheinlichkeit dafür, dass eine ‚willkürlich gezogene' Sehne im Einheitskreis länger als die Seite eines gleichseitigen Dreiecks ist, das in den Einheitskreis einbeschrieben wird?" Eine bestimmte Sehne muss also in der Art gewählt werden, dass bezüglich aller anderen möglichen Sehnen eine Gleichverteilung herrscht. Dafür gibt es verschiedene Möglichkeiten:

- Wähle den Mittelpunkt der Sehne so, dass die Mittelpunkte aller Sehnen gleichverteilt sind.
- Wähle die beiden Endpunkte auf dem Kreisrand so, dass die Endpunkte aller Sehnen gleichverteilt sind.
- Wähle die Polarkoordinaten einer Sehne (Länge $\in [0, 1]$ und Winkel $\in [0, 2\pi]$) so, dass die Polarkoordinaten aller Sehnen gleichverteilt sind.

Basis der verschiedenen Ansätze im Bertrandschen Paradoxon ist jeweils eine Gleichverteilungsannahme, die sich aus dem Gebrauch des ‚Prinzips vom unzureichenden Grund' ergibt. Dieses besagt, dass im Falle von Nichtwissen eine Symmetrie vorliegt in dem Sinne, dass sich jede Wahl einer konkreten Annahme (im Beispiel: die Wahl einer bestimmten Sehne) prinzipiell genauso wenig rechtfertigen lässt wie die Wahl irgendeiner beliebigen anderen Annahme (einer beliebigen anderen Sehne).

Die drei Gleichwahrscheinlichkeitsannahmen führen aber zu je einer anderen Antwort auf die gestellte Frage und demonstrieren damit, dass das Prinzip vom unzureichenden Grund nicht dazu geeignet ist, eine nichtinformative Priori-Verteilung zu ermitteln.

Bei der Anwendung des Prinzips vom unzureichenden Grund wird nämlich letztendlich das *Wissen um Symmetrie* mit dem *Nichtwissen um Asymmetrie* gleichgesetzt. Dass Symmetrie herrscht, ist genau gesehen ein ganz spezielles Wissen und damit mit dem Konzept von a priori Nichtwissen unverträglich. Hat man keine Kenntnis über die Beschaffenheit der Verteilung eines Parameters auf einem Parameterraum, sind ja auch nicht-symmetrische Konstellationen denkbar, und die prinzipielle Beschränkung auf symmetrische Konstellationen stellt eine weitreichende Einschränkung dar.

Im Fall von Parameterräumen mit unendlich vielen möglichen Parameterwerten tritt noch ein weiteres Problem auf: Eine ‚Gleichverteilung' auf dem gesamten Parameterraum ist dann gar keine echte Verteilung mehr, da sie nicht normierbar ist; solche unechten Priori-Verteilungen heißen dann *impropere Priori-Verteilungen*.

Die gebräuchlichsten Regeln, die zur Erstellung einer nichtinformativen Priori-Verteilung führen, seien hier kurz aufgezählt:

- Das Prinzip vom unzureichenden Grund führt zur Bayes-Laplace-Regel.
- Festlegung eines Informationsmaßes für Verteilungen. Die nichtinformative Priori-Verteilung ist dann als diejenige Verteilung definiert, für die das Informationsmaß minimiert wird. So führt beispielsweise die Entropie zur Jaynes-Regel.
- Nichtinformativität bedeutet auch Invarianz gegenüber Parametertransformationen. Diese Überlegung führt zur Jeffreys-Regel.
- Die Forderung nach maximalem Informationsgewinn aus der vorliegenden Stichprobe führt zur Lindley-Bernardo-Regel, deren Ergebnisse auch als „reference prior" bezeichnet werden.
- Die Theorie der „data translated likelihoods" führt zur Box-Tiao-Regel.

In vielen Problemstellungen widersprechen sich die Ergebnisse der verschiedenen Regeln; die Anwendung der gleichen Regel kann aber auch, auf unterschiedliche Problemstellungen angewandt, inkonsistente Ergebnisse nach sich ziehen (für ein Beispiel siehe [Rüger 1999, Kap. 2.5, S. 235]). Meist handelt es sich bei den resultierenden nichtinformativen Prioris um impropere Verteilungen, da sie nicht normierbar sind (wie im Beispiel der Gleichverteilung auf $\mathbb{R}$). Die grundsätzlichen Bedenken, Nichtwissen durch eine bestimmte Priori-Verteilung auszudrücken, sind in den Ausführungen zum Bertrandschen Paradoxon schon angeschnitten worden; Walley unterzieht diese Vorgehensweise in [Walley 1991, Kap. 5.5., S. 226–235] einer vernichtenden Kritik. Rüger subsummiert in [Rüger 1999, Kap. 2.5, S. 270–272] die Probleme und Einwände unter den folgenden drei Punkten:

- **Nichteindeutigkeit:** Verschiedene Regeln führen zu verschiedenen Priori-Verteilungen, dabei sollte doch eine nichtinformative Priori-Verteilung gemäß ihrer Intention intersubjektiv und eindeutig festlegbar sein.
- **Charakterisierung von Nichtwissen durch Indifferenz:** Die verwendeten Regeln haben nicht die nicht vorhandene Information, sondern eher Indifferenz- oder Symmetrieüberlegungen als Basis. (Das gilt im Prinzip auch für die Invarianz gegenüber Parametertransformationen.) Das Vorliegen von Symmetrie oder Indifferenz stellt jedoch eine ganz spezielle Art von Vorwissen dar und entspricht eben nicht dem Nichtwissen um Asymmetrie.

- **Konventionscharakter:** Jede nichtinformative Priori-Verteilung enthält letztlich relativ viel Information, so dass die Verwendung von solchen Priori-Verteilungen als eine einfache Konvention angesehen werden kann. Es handelt sich also mehr um eine ‚Standard-Priori' als eine nichtinformative Priori.

Somit ist in Frage zu stellen, ob Nichtwissen grundsätzlich durch eine einzige Priori-Verteilung darstellbar ist. Die mit dieser Formulierung naheliegende Alternative ist klar: Statt einer einzigen könnten Mengen von Priori-Verteilungen Nichtwissen darstellen. Dies stellt einen Weg dar, über den man zu unscharfen Wahrscheinlichkeitsbewertungen gelangen kann.

1.3 Ellsbergs Paradoxon und die Begrenztheit der Kolmogorovschen Axiomatik

Subjektive Bayesianer bestreiten, dass es so etwas wie Nichtwissen bei der Anwendung von Statistik wirklich gibt. Sie gehen davon aus, dass der Anwender immer gewisse latente Präferenzen hat, die sich über ausgeklügelte Assessment-Strategien aus ihm ‚herauskitzeln' und zu einer eindeutigen Priori-Verteilung destillieren lassen.

Ellsberg berichtet 1961 in [Ellsberg 1961] (hier beschrieben gemäß [Weichselberger 2001]) von folgendem Gedankenexperiment, das er mit Ökonomen und Statistikern der Universität in Harvard durchgeführt hatte:

Eine Urne enthält Kugeln der Farben Rot, Gelb und Schwarz. Der Anteil der roten Kugeln beträgt $\frac{1}{3}$, der Anteil der gelben und der Anteil der schwarzen Kugeln ist unbekannt. Es wird eine Kugel zufällig aus der Urne gezogen. Davor hat jeder Proband die Wahl zwischen zwei Aktionen a_1 und a_2, die zu folgenden Konsequenzen führen:

a_1: 1 $ Gewinn, falls eine rote Kugel gezogen wird

a_2: 1 $ Gewinn, falls eine schwarze Kugel gezogen wird

Die Mehrheit der befragten Ökonomen und Statistiker zog die Aktion a_1 der Aktion a_2 vor. Danach wurden die Probanden vor eine weitere Wahl gestellt. Wieder wird eine Kugel zufällig gezogen, doch diesmal sind die Aktionen folgende:

a_3: 1 $ Gewinn, falls eine rote oder eine gelbe Kugel gezogen wird

a_4: 1 $ Gewinn, falls eine schwarze oder eine gelbe Kugel gezogen wird

Die Personen, die sich bei der ersten Fragestellung für a_1 entscheiden hatten, zogen nun meist a_4 vor. Eine klassische Wahrscheinlichkeitsbewertung, die die Präferenzen dieser Gruppe von Probanden modelliert, gibt es nicht. Bei einer Ermittlung der gemäß des subjektiven Bayesianismus vorliegenden Priori-Verteilung ergibt sich nämlich ein Widerspruch: Aus der Entscheidung für a_1 in der ersten Fragestellung folgt

$$\pi(\{\text{rot}\}) > \pi(\{\text{schwarz}\})\,,$$

aus der Entscheidung für a_4 in der zweiten Fragestellung folgt

$$\pi(\{\text{rot, gelb}\}) < \pi(\{\text{schwarz, gelb}\}),$$

was, nach dem Additivitätsaxiom von Kolmogorov, äquivalent sein sollte zu

$$\pi(\{\text{rot}\}) < \pi(\{\text{schwarz}\}),$$

und somit jedoch im Widerspruch zur ersten Ungleichung steht.

Die Probanden blieben auch nach der ‚Aufklärung' darüber, dass ihre Präferenzordnung den damals als allgemein gültig angesehenen Rationalitätsprinzipien (z.B. dem ‚sure thing principle') widersprach, bei ihren Entscheidungen und empfanden sie weiterhin als rational. Auch die Tatsache, dass es sich um ein Gedankenexperiment handelte, spricht gegen diese Rationalitätsprinzipien und die zentrale Annahme subjektiver Bayesianer. Die Überlegungen, die zu den obigen Präferenzordnungen führten, bezogen sich also nicht auf eine konkrete Auszahlung mit daher konkretem Nutzen, sondern sind mehr als prinzipielle Fragestellung anzusehen im Sinne von „was ist besser?".

Der oben gezeigte Widerspruch bei der Ermittlung einer subjektiven Priori π lässt sich auch in folgender Ungleichung ausdrücken:

$$\pi(\{\text{gelb, schwarz}\}) > \pi(\{\text{gelb}\}) + \pi(\{\text{schwarz}\})$$

Das bedeutet: Wenn π die Präferenzordnung eines zur Mehrheit gehörenden Probanden modellieren soll, dann kann diese subjektive Priori-Verteilung π nicht dem Additivitätsaxiom von Kolmogorov gehorchen.

Eine Theorie der Maßfunktionen, die nicht das Additivitätsaxiom von Kolmogorov erfüllen müssen (aber die anderen beiden), ist die Theorie der Fuzzy-Maße. Diese werden, neben anderen Zweigen der Fuzzy-Mathematik, in [Ott 2001] vorgestellt. Sie bieten, in Form des Spezialfalls der Choquet-Kapazitäten, einen zweiten Weg, zu unscharfen Wahrscheinlichkeitsbewertungen zu gelangen, indem sie eine mathematische Beschreibung der Intervallgrenzen liefern.

Ein mit der Unzulänglichkeit der Axiome von Kolmogorov zusammenhängendes starkes Motiv für die Entwicklung von Modellen, die komplexe Unsicherheit miteinbeziehen, ist nach Weichselberger auch das Phänomen der „Unvergleichbarkeit von Wahrscheinlichkeiten". Es gibt Situationen, in denen die Wahrscheinlichkeiten für zwei Ereignisse A und B gemäß einer subjektiven Bewertung nicht vergleichbar sind in dem Sinne, dass weder $P(A) > P(B)$ noch $P(A) < P(B)$ noch $P(A) = P(B)$ eingeschätzt wird.

1.4 Weitere Argumente für Statistik unter komplexer Unsicherheit

Neben den in den letzten beiden Kapiteln erläuterten Hinweisen darauf, dass die Methoden der klassische Statistik letztlich unzureichend sein können, gibt es noch weitere Argumente gegen eine Vernachlässigung von Ambiguität und Vagheit, die in den nächsten Abschnitten beschrieben werden sollen.

Wenn man zulässt, das Wahrscheinlichkeitsbewertungen das Additivitätsaxiom von Kolmogorov nicht notwendigerweise erfüllen müssen, sind nämlich deutlich ‚realistischere' Modellierungen von Unsicherheitssituationen möglich, wobei ‚realistischer' sowohl epistemisch als auch ontologisch verstanden werden kann. Außerdem ermöglicht eine solche Modellierung einen sinnvollen Umgang mit sogenannten ‚prior-data'-Konflikten.

1.4.1 Argumente aus epistemischer Sicht

Aus epistemischer Sichtweise liegt dem zu modellierenden Phänomen ‚in Wahrheit' eine präzise Wahrscheinlichkeitszuordnung zugrunde, aber man ist nicht in der Lage, diese vollständig zu ermitteln. So kann zum Beispiel im Kontext einer subjektivistischen Analyse

- die Ermittlung von allen Präferenz-Bewertungen, die für eine exakte Wahrscheinlichkeitszuordnung nötig sind, viel zu aufwändig sein.
- überhaupt nur mit stark vereinfachten Modellen eine solche Präferenzordnung ermittelbar sein.
- die Gesamtheit der verfügbaren Bewertungen nur einen gewissen Grad an Präzision zulassen.

Oft ist eben nicht wirklich klar, ob die Annahmen in klassischen Modellierungen tatsächlich erfüllt sind. Sind die Beobachtungen wirklich gemäß der Verteilungsannahme verteilt? Sind die einzelnen Beobachtungen wirklich voneinander unabhängig? Unsicheres Wissen über die Gültigkeit solcher grundlegenden Annahmen klassischer Methoden sollte bei der Modellierung nicht einfach ignoriert werden. So ist häufig nur begrenztes Wissen über die Angemessenheit dieser grundlegenden Annahmen klassischer Modellierungen vorhanden; diese Annahmen dann einfach als gegeben vorauszusetzen, stellt eine vielleicht allzu grobe Vereinfachung dar und kann zu fatalen Fehlschlüssen führen.

Insbesondere dann, wenn eventuelle Asymptotik-Argumente noch nicht greifen oder ihre Voraussetzungen fraglich sind, sind mittels klassischer Modellierungen oft keine nutzbaren Wahrscheinlichkeitsaussagen möglich. In vielen solchen Situationen lassen sich aber Modelle unter Einbeziehung von komplexer Unsicherheit erstellen, die nutzbare Aussagen liefern. Oft ist es einfacher, aus den wenigen verfügbaren Informationen ein solches

Modell zu erstellen, als zu versuchen, die starken Annahmen klassischer Modelle in irgend einer Art und Weise zu rechtfertigen.

„Unvollständig determinierte Wahrscheinlichkeit“ beschreibt in der Begrifflichkeit von Weichselberger Situationen, in denen die Wahrscheinlichkeit nicht für alle Elementarereignisse bekannt ist. Solche Situationen führen praktisch automatisch zu unscharfen Wahrscheinlichkeiten für die Elementarereignisse, deren klassische Wahrscheinlichkeit unbekannt ist.

Eine Antwort auf die in den Kapiteln 1.2 und 1.3 beschriebenen Probleme der klassischen bayesianischen Methodik sind Verfahren, die unter dem Namen ‚robuste Bayes-Verfahren‘ zusammengefasst werden, siehe z.B. [Berger (1985)]. Sie versuchen, mögliche Abweichungen von den strengen Voraussetzungen der Bayes-Theorie zu berücksichtigen und so deren Ergebnisse zu ‚robustifizieren‘. Diese Ansätze werden of im Sinne einer Sensitivitätsanalyse interpretiert und können als Spezialfall der Statistik unter komplexer Unsicherheit angesehen werden, da sie mit Mengen von Verteilungen arbeiten.

Aber auch in einem frequentistischen Kontext werden Mengen von Verteilungen zur Analyse von Stichproben verwendet. Meist handelt es sich dabei um sogenannte Umgebungsmodelle, bei denen die Umgebung einer festen Stichproben-Verteilung p_0, die man als Ausgangspunkt wählt, betrachtet wird. Die Umgebung wird dabei über eine geeignete Metrik definiert. Im Folgenden zwei Beispiele für Umgebungsmodelle:

- Bei ε-Kontaminationsklassen (entwickelt von [Huber 1965]) geht man davon aus, dass $(1-\varepsilon)\cdot 100\%$ der Beobachtungen gemäß p_0 verteilt sind, jedoch der Rest der Beobachtungen, also $\varepsilon\cdot 100\%$, ‚gestört‘ ist, also von irgend einer anderen Verteilung stammt.

- Eine allgemeine Form von Umgebungsmodellen kann über eine Verzerrungsfunktion f erhalten werden. Erfüllt f gewisse Voraussetzungen, erzeugt $f\big(p_0(A)\big)$ ein Intervallwahrscheinlichkeitsmodell, das eine Umgebung um p_0 modelliert und somit für jedes Ereignis A ein Wahrscheinlichkeitsintervall liefert (siehe z.B. [Buja 1986] oder [Wallner 2003]).

Mit dem Huber-Strassen-Theorem (siehe [Huber und Strassen 1973] oder [Rüger 1999, Kap. 2.6]) sind Testverfahren für den Vergleich zweier solcher Umgebungsmodelle möglich. Optimale Tests basieren dann auf sogenannten ‚ungünstigsten Paaren‘ von Verteilungen, die am schwierigsten zu trennen sind. Das Huber-Strassen-Theorem lässt sich auch für andere Testsituationen unter komplexer Unsicherheit verallgemeinern (siehe [Augustin 1998]).

Wie in den Ausführungen vielleicht deutlich geworden ist, können auch diese Methoden als Spezialfälle von Intervallwahrscheinlichkeitsmodellen angesehen werden. So bietet die Theorie der Statistik unter komplexer Unsicherheit einen Rahmen für viele Ansätze, in denen Mengen von Verteilungen betrachtet werden.

1.4.2 Argumente aus ontologischer Sicht

In ontologischer Sichtweise sind die Unsicherheiten des zu modellierenden Phänomens ‚real' und nicht durch die Begrenztheit der verfügbaren Informationen, sondern gewissermaßen ‚physikalisch' bedingt. So sind beispielsweise

- Zufallsfolgen denkbar, die nicht in einen Punkt konvergieren, sondern nur in ein Intervall.
- die meisten Größen nur begrenzt genau messbar; ist die Messungenauigkeit groß, kann das zu starken Verzerrungen in den Schlüssen führen.
- die Bewertungen von verschiedenen Experten, die für die Erstellung eines Expertensystems erhoben werden, häufig widersprüchlich. Statistik unter komplexer Unsicherheit macht es dann möglich, diese widersprechenden Informationen trotzdem in *ein* Modell einzupassen.
- nicht komplett übertragbare Informationen von einem Phänomen auf ein zweites zur Modellierung des zweiten Phänomens in einem klassischen Rahmen nur begrenzt nutzbar.

Statistik unter komplexer Unsicherheit macht solche Unsicherheiten, gleich welcher Quelle, quantifizierbar und ermöglicht es somit, sie in die Modellierung einzubeziehen. Tatsächlich sind die Ergebnisse bayesianischer oder klassischer Modelle in Situationen mit wenig Informationen meist viel zu genau. Diese Schein-Genauigkeit kann dann oft zu absurden Aussagen und Interpretationen führen und ist deshalb, zusammen mit anderen Faktoren, vielleicht für das seltsame Bild der Statistik in der Öffentlichkeit mitverantwortlich.

1.4.3 Prior-data conflict

Ein weiteres wesentliches Problem im Rahmen bayesianischer Analysen ist das des sogenannten ‚prior-data conflict', des Falls, in dem das Vorwissen (aus dem die Priori-Verteilung resultiert) nicht zu den Beobachtungen (den Daten) passt.

Eine so erhaltene Posteriori-Verteilung unterscheidet sich dann aber nicht prinzipiell von einer, die in einem Fall das Resultat wäre, wenn die Informationen aus Priori und Likelihood übereinstimmen würden.

Das sei kurz an einem Beispiel illustriert (analog zu [Walley 1991, S. 6]). Angenommen, eine Beobachtung x ist die Realisation einer Normalverteilung mit Varianz 1 und unbekanntem Erwartungswert θ. Das Vorwissen besagt, dass θ in der Nähe eines gewissen Wertes μ liegt, und man bestimmt daher als Priori-Verteilung für θ eine Normalverteilung mit Varianz 1 und Erwartungswert μ. Nach dem Satz von Bayes ist dann die Posteriori-Verteilung wieder eine Normalverteilung, jedoch mit Erwartungswert $\frac{1}{2}(\mu+x)$ und Varianz $\frac{1}{2}$. Betrachtet man nun die beiden folgenden Fälle,

(i) $\mu = 0$, $x = 1$ und

(ii) $\mu = -1000$, $x = 1001$,

so hat die Posteriori-Verteilung in beiden Fällen die gleiche Form, da bei beiden der Posteriori-Erwartungswert $\frac{1}{2}$ beträgt. Die Posteriori-Verteilung ist, gemäß dem 3. Bayes-Postulat (Bezeichnung z.B. in [Rüger 1999]), die Basis aller weiterer a posteriori Schlüsse. Aber die beiden Fälle (i) und (ii) unterscheiden sich fundamental: Bei (i) sind Vorwissen und Beobachtung vergleichsweise konform, und Schlüsse aus der erhaltenen Posteriori-Verteilung scheinen sinnvoll. Bei (ii) widersprechen sich Vorwissen und Beobachtung eklatant, und Schlüsse einzig und allein aus der Posteriori-Verteilung $\mathrm{N}(\frac{1}{2},\frac{1}{2})$ scheinen völlig ungerechtfertigt.

Das wesentliche Problem ist, dass anhand der Posteriori-Verteilung nicht erkennbar ist, in welcher dieser beiden Situationen man sich befindet. Die Information, dass sich Vorwissen und Beobachtung widersprechen, geht bei der bayesianischen Aufdatierung einer einzigen Priori verloren.

Wieder zeigt sich das Problem der ‚Übergenauigkeit' der klassischen bayesianischen Methodik: Aus der Beschränkung auf eine einzige Priori-Verteilung folgt die Beschränkung auf eine einzige Posteriori-Verteilung. Unsicherheiten bezüglich eines Parameters der Priori-Verteilung, z.B. bezüglich dem Erwartungswert μ einer Normalverteilung, sind nur durch die feste Wahl eines oder mehrerer anderer Parameter möglich: Entweder, im Falle eines Lageparameters, durch die Wahl des zugehörigen Skalenparameters in derselben Priori (im Beispiel wäre das ν^2, die Varianz der Normalverteilung um μ), oder, falls das nicht möglich ist (etwa, wenn ν selbst der ‚unsichere' Parameter ist), durch eine Hyper-Priori über den Parameter, bezüglich dessen Unsicherheit herrscht. Die Parameter dieser Hyper-Priori müssen dann aber wiederum fest gewählt werden, so dass man letztendlich vor dem gleichen Problem wie zuvor steht, wobei das Modell aber sehr viel komplexer geworden ist. Außerdem hilft ein solches Vorgehen in der Situation eines ‚prior-data conflict' nicht weiter: ist man sich a priori ziemlich sicher, dass θ in der Gegend um -1000 liegt und wählt man daher $\mu = -1000$ und die zugehörige Varianz ν^2 entsprechend klein, so erhält man zwar einen Posteriori-Erwartungswert, der in gehörigem Abstand zu μ liegt, aber die Posteriori-Varianz verkleinert sich, was bedeutet, dass man sich mindestens so sicher wie zuvor ist, dass θ nun in der Nähe des ‚neuen' μ liegt. Das dürfte jedoch, bei solch eklatanten Widersprüchen, ein ziemlich gewagter Schluss sein.

An dieser Situation und an den Problemen der nichtinformativen Priori-Verteilungen zeigt sich das grundlegende Defizit der klassischen bayesianischen Analyse: In ihrem Rahmen können keine Angaben darüber gemacht werden, wie sicher die priori-Informationen über *alle* Parameter der Priori-Verteilung sind. Daher ist dann auch in der a posteriori-Situation nicht klar, wie unsicher das Wissen nun über die Posteriori-Parameter ist.

Aus diesem Grund ergibt sich oft die seltsame Situation, dass Schlüsse auf der Basis vieler Beobachtungen und Schlüsse ohne jede Datenbasis völlig unterschiedslos sind. Der grundlegende Unterschied zwischen solchen Schlüssen (viele Daten als Basis – keine Daten als Basis) ist in der bayesianischen Analyse nicht erkennbar. Wird im Beispiel des Münzwurfs aufgrund ungenügenden Vorwissens eine nichtinformative Priori verwendet, die für $P(\text{Kopf}) = \frac{1}{2}$ liefert, und ist die relative Häufigkeit in den Versuchen $\frac{1}{2}$, so ergibt sich genau eine solche Situation.

Intervallwahrscheinlichkeitsmodelle lassen es hingegen zu, den Grad an Vertrauen in das priori-Wissen zu modellieren, in dem die Menge der Priori-Verteilungen entsprechend groß gewählt wird oder indem die Priori-Wahrscheinlichkeit für Ereignisse nicht auf einen Punktwert festgelegt, sondern durch Intervalle entsprechender Breite begrenzt wird. Dann kann auch an den Ergebnissen der a posteriori-Inferenz erkannt werden, wie sicher das durch die Aufdatierung gewonnene Wissen ist: je breiter die a posteriori erhaltenen Intervalle einer Größe, desto unsicherer ist das Wissen über diese Größe; eine Modellierung unter Berücksichtigung von komplexer Unsicherheit kann so konstruiert werden, dass diese Intervalle breiter werden, wenn ein ‚prior-data conflict' vorliegt.

1.5 Modellierung komplexer Unsicherheit

1.5.1 Prinzipielles Vorgehen

Wie in den vorigen Kapiteln schon angedeutet wurde, gibt es mehrere Möglichkeiten zur Modellierung komplexer Unsicherheit. Dabei lassen sich prinzipiell zwei Wege unterscheiden:

- Mengen $\mathcal{M}$ klassischer Wahrscheinlichkeitszuordnungen
- Intervallwahrscheinlichkeit

Diese beiden Wege führen unter gewissen Regularitätsvoraussetzungen zu äquivalenten Ergebnissen und stellen dann zwei verschiedene Arten dar, dasselbe Modell zu beschreiben.

Eine kurze Beschreibung der beiden Wege folgt in den nächsten zwei Kapiteln. Der Zusammenhang zwischen ihnen und eine kurze Vorstellung der zwei wichtigsten Monographien, die eine Theorie der Modellierung komplexer Unsicherheit zum Thema haben, folgt dann in Abschnitt 1.5.4.

1.5.2 Mengen klassischer Wahrscheinlichkeitszuordnungen

Beim ersten Weg wird die Ambiguität über eine Menge klassischer Verteilungen modelliert, die Stochastizität über die jeweilige Verteilungsfunktion. Je stärker die

Ambiguität, desto größer kann die Menge der Verteilungsfunktionen gewählt werden; herrscht ideale Stochastizität, so kann diese Menge auf eine einzige Verteilungsfunktion zusammengezogen werden, so dass man sich in der klassischen Situation wiederfindet; bei völliger Ambiguität besteht $\mathcal{M}$ aus allen möglichen Verteilungsfunktionen. Zum Zwecke größerer Praktikabilität kann $\mathcal{M}$ auch nur bestimmte Klassen von Verteilungsfunktionen enthalten; im Kapitel 2 und in dem Modell, das Gegenstand dieser Arbeit ist, wird dieser Weg gewählt. Es handelt sich dann um Mengen, die parametrisch definiert sind: Bestandteil der Menge sind grundsätzlich nur Verteilungen einer bestimmten Verteilungsfamilie (in Kapitel 2 ist das die Familie der Dirichlet-Verteilungen), wobei $\mathcal{M}$ über die Menge der Parameter dieser Verteilungsfamilie erzeugt wird. Über die Plausibilität der einzelnen Elemente von $\mathcal{M}$ werden keine Annahmen gemacht; in diesem Sinne sind alle Verteilungsfunktionen in $\mathcal{M}$ ‚gleich wahrscheinlich'.

1.5.3 Intervallwahrscheinlichkeit

Der zweite Weg geht von der Wahrscheinlichkeit von Ereignissen aus. Die Wahrscheinlichkeit jedes Ereignisses A aus der zum Beobachtungsraum Ω gehörenden σ-Algebra $\mathcal{A}$ wird dann über ein Intervall $[L(A), U(A)]$ modelliert, so dass die Wahrscheinlichkeitsfunktion P definiert wird durch

$$\begin{aligned} P: \mathcal{A} &\longrightarrow \mathcal{Z}_0 \\ A &\longrightarrow P(A) = [L(A), U(A)]\,, \end{aligned}$$

wobei $\mathcal{Z}_0$ die Menge der abgeschlossenen Intervalle in $[0, 1]$ ist. Die Breite des Intervalls stellt dann das Ausmaß der Ambiguität dar; herrscht ideale Stochastizität, so zieht sich dieses Intervall auf einen Wert zusammen und wir befinden uns im klassischen Fall; bei völliger Ambiguität gilt für alle $A \in \mathcal{A}$ dann $P(A) = [0, 1]$, die Wahrscheinlichkeit jedes Ereignisses kann also die Werte ‚sicher' oder ‚unmöglich' sowie alle Abstufungen dazwischen annehmen. Innerhalb der Intervalle herrscht vollkommene Unsicherheit, jeder Wert innerhalb von $[L(A), U(A)]$ ist daher ‚gleich wahrscheinlich'. Die Mengenfunktionen $L(\cdot)$ und $U(\cdot)$ gehören dann zu den in Kapitel 1.3 erwähnten Fuzzymaßen.

1.5.4 Die Theoriegebäude von Walley und Weichselberger

Bisher sind zwei grundlegende Monographien erschienen, die eine Theorie zur Modellierung von komplexer Unsicherheit beschreiben: [Walley 1991] und [Weichselberger 2001].

Walleys Ausgangspunkt sind untere und obere ‚Previsions', die auf Zufallsvariablen definiert sind. Sie entsprechen einem Erwartungswertbegriff; die untere und obere Wahrscheinlichkeit eines Ereignisses ergibt sich über die Indikatorfunktion für das Ereignis. Jede Art von Wahrscheinlichkeitsbewertung ist bei Walley untrennbar mit einer subjektivistischen Interpretation verbunden, die in der Tradition von [de Finetti 1990]

steht. $\underline{P}(A)$, die untere Wahrscheinlichkeit eines Ereignisses A, ist dabei der maximale Einsatz einer Wette, den man als ‚Wettnehmer' bereit zu zahlen wäre, um im Falle des Eintretens von A eine Geldeinheit zu bekommen. Die obere Wahrscheinlichkeit $\overline{P}(A)$ ist hingegen der minimale Einsatz, den man als Anbieter einer Wette akzeptiert, um im Falle des Eintretens von A eine Geldeinheit an einen ‚Wettnehmer' auszuzahlen. Der Bereich zwischen $\underline{P}(A)$ und $\overline{P}(A)$ ist eine Indifferenzzone und umfasst die Geldbeträge, für die man nicht dazu bereit ist, eine Wette auf A anzunehmen, aber ebensowenig dazu bereit ist, eine solche Wette anzubieten. Das bekannteste und einflussreichste Modell von Walley, das IDM (siehe Kapitel 2), wird jedoch über Mengen von Verteilungen, die mit subjektiven Wahrscheinlichkeitsbewertungen vereinbar sind, definiert und entspricht somit einer Vorgehensweise nach dem ersten Weg.

Weichselberger setzt den Schwerpunkt in der Modellierung auf den zweiten Weg; sein zentraler Begriff sind prinzipiell interpretationsfreie obere und untere Wahrscheinlichkeiten für Ereignisse A, die zusammen eine Wahrscheinlichkeitsbewertung $P(A) = [L(A), U(A)]$ ergeben. Diese können dann mit einer Menge von Verteilungen in Beziehung gesetzt werden. Gegebenenfalls müssen die Grenzen $L(A)$ und $U(A)$ des Wahrscheinlichkeitsintervalls modifiziert werden, um eine eineindeutige Beziehung zwischen ihnen und der Menge der Verteilungen zu gewährleisten. Mit Walleys Ansatz, der auf Erwartungswerten als zentralen Begriffen beruht, ist es möglich, allgemeinere Mengen von Verteilungen zu beschreiben als mit der Theorie von Weichselberger, welche direkt auf dem Begriff der Wahrscheinlichkeit eines Ereignisses beruht. Die Theorie von Weichselberger ist jedoch unabhängig von einer Interpretation als subjektivem Wettpreis, der ja implizit auch einen Nutzenbegriff voraussetzt.

Weichselberger kritisiert an Walley, dass dieser, durch die Bindung des Wahrscheinlichkeitsbegriffs an die Interpretation als Wettpreise, ohne Not die Wahrscheinlichkeitsbewertung mit einem Nutzenbegriff verquickt und dabei letztendlich Geldwert mit Geldnutzen gleichsetzt. Auch wenn dieses Problem reduziert werden kann, indem die gedachten Geldbeträge als relativ klein angesetzt werden, bleibt es prinzipell bestehen: Ein Student wird den Gewinn von einem Euro möglicherweise anders bewerten als der Vorstandsvorsitzende eines großen Unternehmens, auch wenn beide ihrem Kalkül die gleiche Wahrscheinlichkeitsbewertung zugrunde legen.

Die Dualität zwischen der Modellierung von komplexer Unsicherheit als Mengen von Verteilungen und als Intervallwahrscheinlichkeit stellt sich genau dann ein, wenn für alle Ereignisse $A \in \mathcal{A}$

- die untere Wahrscheinlichkeit $L(A)$ äquivalent ist zur kleinsten Wahrscheinlichkeit für A unter der Verwendung aller klassischen Verteilungsfunktionen aus $\mathcal{M}$ und
- die obere Wahrscheinlichkeit $U(A)$ äquivalent ist zur größten Wahrscheinlichkeit für A unter der Verwendung aller klassischen Verteilungsfunktionen aus $\mathcal{M}$,

oder in mathematischer Notation

$$L(A) = \inf_{p(\cdot)\in\mathcal{M}} p(A) \quad \text{und} \quad U(A) = \sup_{p(\cdot)\in\mathcal{M}} p(A)\,.$$

In der Axiomatik von Weichselberger [Weichselberger 2001, S. 141ff] heißen Modelle, die diese Dualitätseigenschaft erfüllen, *F-Wahrscheinlichkeit*. Die Menge $\mathcal{M}$ heißt dann *Struktur*, und im Falle dieser eineindeutigen Beziehung zwischen Struktur und Intervallgrenzen sind $L(\cdot)$ und $U(\cdot)$ ‚konjugiert': Es gilt

$$L(A) = 1 - U(A^c)\,,$$

die untere Wahrscheinlichkeit eines Ereignisses A kann als die obere Wahrscheinlichkeit des Gegenereignisses A^c berechnet werden. In diesem Fall reicht daher die Angabe einer der beiden Mengenfunktionen. Die Struktur $\mathcal{M}$ ist eine konvexe Menge; im Falle endlicher Beobachtungsräume Ω ist $\mathcal{M}$ ein konvexer Polyeder, so dass sich viele (in der Theorie häufig anfallende) Minimierungs- und Maximierungsprobleme durch lineare Optimierung lösen lassen.

Wie in der Beschreibung der beiden Wege angemerkt wurde, enthält dieser Ansatz die klassische Modellierung von Unsicherheit (gemäß der Axiome von Kolmogorov) als Spezialfall. Es handelt sich daher nicht um eine Theorie, die die klassische Wahrscheinlichkeitstheorie ersetzen soll, sondern um eine Verallgemeinerung.

1.5.5 Abgeleitete Größen

Geht man von einer Modellierung gemäß Weichselberger aus, ist eine eigene Definition des Erwartungswerts nötig. Dazu gibt es wieder zwei Möglichkeiten:

- Ein unterer und oberer Erwartungswert kann direkt aus $L(\cdot)$ und $U(\cdot)$ über das Choquet-Integral (siehe z.B. [Ott 2001]) berechnet werden. Es gilt dann

  $$\underline{\mathbb{E}}[X] = \int L(\{X > t\})dt \quad \text{und} \quad \overline{\mathbb{E}}[X] = \int U(\{X > t\})dt\,, \quad X \leq 0$$

  Dieser Ansatz ist naheliegend, wenn sich die Modellierung in erster Linie auf $L(\cdot)$ und $U(\cdot)$ stützt.

- Der untere und obere Erwartungswert kann als untere und obere Grenze des ‚normalen' Erwartungswerts gebildet werden, wenn alle klassischen Wahrscheinlichkeiten $p \in \mathcal{M}$ in Betracht gezogen werden:

  $$\underline{\mathbb{E}}[X] = \inf_{p\in\mathcal{M}} \mathbb{E}_p X \quad \text{und} \quad \overline{\mathbb{E}}[X] = \sup_{p\in\mathcal{M}} \mathbb{E}_p X$$

  Geht die Modellierung von Mengen von Verteilungen aus, ist dies der naheliegende Weg.

Beide Wege sind i.A. nicht äquivalent; nur wenn $L(\cdot)$ und $U(\cdot)$ spezielle Bedingungen erfüllen, kommen sie zum gleichen Ergebnis.

Die Art der Definition von bedingten Wahrscheinlichkeiten im Rahmen von komplexer Unsicherheit ist noch umstritten; es gibt dazu verschiedene Ansätze, deren Beschreibung den Rahmen dieser Einleitung sprengen würde (siehe z.B. [Weichselberger 2001]); der naheliegende Ansatz analog zum zweiten Weg bei der Erwartungswert-Definition ist aber natürlich folgender:

$$\underline{P}(A \,|\, B) = \inf_{p \in \mathcal{M}} p(A \,|\, B) \quad \text{und} \quad \overline{P}(A \,|\, B) = \sup_{p \in \mathcal{M}} p(A \,|\, B)$$

1.6 Verwandte Theorien

Es gibt eine Reihe von Theorie-Ansätzen, die in engem Zusammenhang mit der Theorie der Intervallwahrscheinlichkeit stehen; davon sollen die Theorie der mehrwertigen Abbildungen [Dempster 1967] und die der Belief-Funktionen [Shafer 1976] kurz dargestellt werden.

Gewöhnliche Zufallsvariablen können als Abbildung von Elementen der σ-Algebra $\mathcal{A}_1$ eines Raums Ω_1 auf Elemente der σ-Algebra $\mathcal{A}_2$ eines anderen Raums Ω_2 gesehen werden:

$$X : (\Omega_1,\, \mathcal{A}_1) \longrightarrow (\Omega_2,\, \mathcal{A}_2)$$

In diesem Fall ist das Bild eines Elements $\omega_1 \in \Omega_1$ ein Element $\omega_2 \in \Omega_2$. Soll jedoch ein ω_1 auf eine beliebige Teilmenge von Ω_2 abgebildet werden, so muss X als eine mehrwertige Abbildung definiert werden:

$$X : (\Omega_1,\, \mathcal{A}_1) \longrightarrow (\mathcal{A}_2,\, \sigma(\mathcal{A}_2))$$

Grundmenge des Bildraums ist jetzt die σ-Algebra $\mathcal{A}_2$, die im Falle von $|\Omega_2| < \infty$ die Potenzmenge von Ω_2 darstellt.

Das Wahrscheinlichkeitsmaß P auf $(\Omega_1,\, \mathcal{A}_1)$ führt im Fall einer normalen Zufallsvariable zu einer klassischen Wahrscheinlichkeitszuordnung auf dem Bildraum, es gilt

$$P_X(A_2) = P(\{\omega_1 \in \Omega_1 \,|\, X(\omega_1) \in A_2\}) \quad \forall A_2 \in \mathcal{A}_2$$

Das Bildmaß für eine mehrwertige Abbildung X ist dagegen ein Intervall, mit $L_X(A_2) \leq P_X(A_2) \leq U_X(A_2)$, $\forall A_2 \in \mathcal{A}_2$, wobei

$$\begin{aligned} L_X(A_2) &= P(\{\omega_1 \in \Omega_1 \,|\, X(\omega_1) \subset A_2\}) \quad \text{und} \\ U_X(A_2) &= P(\{\omega_1 \in \Omega_1 \,|\, X(\omega_1) \cap A_2 \neq \emptyset\})\,. \end{aligned}$$

$L_X(A_2)$ ist also das Wahrscheinlichkeitsmaß all der Elemente von Ω_1, deren Bild sicher in A_2 liegt, und $U_X(A_2)$ das Wahrscheinlichkeitsmaß all der Elemente von Ω_1, deren

Bild Überschneidungen mit A_2 hat. L_X und U_X erweisen sich dabei als eine spezielle Art der in Kapitel 1.3 erwähnten Choquet-Kapazitäten.

Belief-Funktionen hängen eng mit diesen Bildmaßen zusammen; hier wird das Wahrscheinlichkeitsmaß jedoch direkt auf der Potenzmenge $\mathcal{P}(\Omega)$ des Grundraums definiert, wodurch untere und obere Grenzen generiert werden. Für ein endliches Ω sei $m(\cdot)$ eine Wahrscheinlichkeitszuordnung für $\mathcal{P}(\Omega)$:

$$m : \mathcal{P}(\Omega) \longrightarrow [0,1] \quad \text{mit} \quad m(\emptyset) = 0, \sum_{A \subseteq \Omega} m(A) = 1\,.$$

$m(\cdot)$ heißt dann *basic probability assignment*. Mit dem Übergang von Ω zu $\mathcal{P}(\Omega)$ können unpräzise Beobachtungen und sogar fehlende Werte gut modelliert werden; eine unscharfe Beobachtung, die sowohl dem Elementarereignis A als auch dem Elementarereignis B zugeordnet werden könnte, kann dann hier als ein entsprechendes Gewicht auf $A \cup B$ eingehen. Komplett fehlende Werte entsprechen dann einfach einer Beobachtung des Ereignisses Ω, sie könnten potentiell jedem beliebigen Elementarereignis zugeordnet werden. Natürlich können beliebige Teilmengen von Ω das Wahrscheinlichkeitsgewicht 0 erhalten; gilt nur für die Elemente ω von Ω, dass $m(\omega) \neq 0$, so befinden wir uns im Spezialfall eines gewöhnlichen Wahrscheinlichkeitsmaßes.

Für beliebige Mengen $A \subseteq \Omega$ ist dann

$$L(A) = \sum_{B \subseteq A} m(B) \quad \text{und} \quad U(A) = \sum_{B \cap A \neq \emptyset} m(B)\,.$$

Für $L(A)$ werden somit alle Wahrscheinlichkeitskomponenten aufsummiert, die zwingend für A sprechen, für $U(A)$ hingegen alle, die nicht gegen A sprechen. $L(\cdot)$ heißt Belief-Funktion, $U(\cdot)$ Plausibility, wobei letzterer Begriff auch für ähnliche andere Konzepte verwendet wird.

Basic probability assignments sind eine gute Methode, um zu intervallwertigen Wahrscheinlichkeitsbewertungen zu gelangen. Sie sind daher zu vielfacher Anwendung gekommen, so hat z.B. [Maier 2004] die Anwendung dieses Ansatzes auf vergröberte Daten am Beispiel von Wahlumfragen untersucht; eine weitere Verallgemeinerung des Begriffs wird in [Augustin 2005] vorgestellt.

1.7 Das zur Anwendung kommende Modell

Die in dieser Arbeit verwendete Vorgehensweise zur Modellierung von komplexer Unsicherheit ist die einer Menge von Verteilungen derselben Verteilungsfamilie, die parametrisch definiert wird. $\mathcal{M}$ wird also über die Wahl einer konkreten Verteilungsfamilie und über eine Menge von Parametern erzeugt. Die Verteilungsfamilie repräsentiert dabei mittels einer Verteilungsfunktion für gegebene Parameter die Komponente der idealen

Stochastizität; die Menge der Parameter sorgt für die Einbeziehung von Ambiguität in die Modellierung. Vagheit, die sich durch unscharfe oder fehlende Beobachtungen äußern könnte, soll in dieser Modellierung noch außen vor gelassen werden.

Es handelt sich um eine Erweiterung eines allgemeinen Modells für Verteilungen, die einer Exponentialfamilie entsprechen, das in [Quaeghebeur und de Cooman 2005] vorgestellt wird. Dort wird als Verteilungsfamilie die zu einer Stichprobenverteilung konjugierte Verteilung der Parameter festgelegt und deren Parameter bayesianisch aufdatiert. Das Modell von Quaeghebeur und de Cooman stellt eine Verallgemeinerung eines Vorläufermodells dar, das von Walley [Walley 1996] für multinomiale Daten entwickelt wurde.

Um diese Verfahrensweise bei der Modellierung komplexer Unsicherheit zu veranschaulichen und um das Verständnis der Methodik von Quaeghebeur und de Cooman zu erleichtern, soll in Kapitel 2 das Vorläufermodell von Walley, das Imprecise Dirichlet Model, detailliert vorgestellt werden; das Modell, das dann in dieser Arbeit für eine Verwendung im Rahmen der Regressionsanalyse in Kapitel 4 erweitert wird, soll in Kapitel 3 erläutert werden.

Mengen von Verteilungen lassen sich aber auch nicht-parametrisch definieren. [Coolen 1993] untersucht beispielsweise ein Verfahren, bei dem $\mathcal{M}$ für einparametrische Verteilungen über obere und untere Dichtegrenzen definiert wird. Dabei werden zwei zueinander proportionale Funktionen $u(\theta) = c \cdot l(\theta)$ definiert mit $0 \leq l(\theta) \leq u(\theta)$ für $\theta \in \mathbb{R}$ als obere und untere nicht-normierte ‚Dichten'. Alle Funktionen zwischen diesen Grenzen bilden dann in normierter Form $\mathcal{M}$:

$$\mathcal{M} = \left\{ p(\cdot) \;\middle|\; p(\cdot) = \frac{f(\cdot)}{\int f(\cdot)},\, l(\cdot) \leq f(\cdot) \leq u(\cdot) \right\}$$

Die ‚Abstandskonstante' c darf dabei nicht von θ, dem Argument der Grenzfunktionen abhängen. Diese Voraussetzung stellt sich jedoch als eine schwerwiegende Beschränkung heraus, da es ohne diese Abhängigkeit unmöglich ist, ‚prior-data conflict'-Situationen adäquat zu modellieren. Gerade die Einbeziehung solcher Situationen stellt jedoch ein starkes Argument für die Modellierung komplexer Unsicherheit dar.

Einen Vergleich verschiedener Möglichkeiten zur Generierung von $\mathcal{M}$ bietet [Pericchi und Walley 1991] am Beispiel von Kredibilitätsintervallen für einen unbekannten Lageparameter θ. Dort werden zwei prinzipelle Anwendungsmöglichkeiten unterschieden: Soll a priori-Nichtwissen modelliert werden, schlagen Pericchi und Walley sogenannte translations-invariante Klassen von Verteilungen vor; gibt es substanzielles Vorwissen, so werden verschiedene Umgebungsmodelle vorgeschlagen. Die Qualität der Modellierung wird dann anhand des Verlaufs der Grenzen der resultierenden Kredibilitätsintervalle und der Kredibilität eines klassischen symmetrischen Konfidenzintervalls für θ in Abhängigkeit der Anzahl der verfügbaren Beobachtungen bewertet; bei den Umgebungsmodellen wird zusätzlich das Verhalten des Modells betrachtet, wenn ein

‚prior-data conflict' vorliegt.

Pericchi und Walley kommen dabei zu dem Schluss, dass eine Menge von sinnvollen Verteilungen nicht unbedingt auch eine sinnvolle Menge von Verteilungen ergibt. Ihre Empfehlung ist im Falle von a priori-Nichtwissen eine Menge von Doppelexponentialverteilungen für den unbekannten Mittelwert θ mit Erwartungswert μ und Varianz ν:

$$f_{\mu,\nu}(\theta) = \frac{1}{\sqrt{2}\nu^2} \exp\left\{-\frac{|\theta - \mu|}{\sqrt{2}\nu^2}\right\}$$

Die translations-invariante Menge $\mathcal{M}$ wird dann, für ein festes ν, durch die Variation von μ über ganz $\mathbb{R}$ erzeugt.

Pericchi und Walley vergleichen dieses Modell auch mit einem Modell, dass auf konjugierten Priori-Verteilungen basiert und damit dem in dieser Arbeit zur Anwendung kommenden Modell ähnelt. $\mathcal{M}$ wird dabei als Menge von Normalverteilungen über θ in Abhängigkeit von zwei Parametern μ_0 und ν definiert:

$$\mathcal{M}(\mu_0, \nu) = \left\{ \pi_{\mu,\nu}(\theta) \;\middle|\; |\mu - \mu_0| \leq \frac{c\nu^2}{2\sigma} \right\}$$

$\mathcal{M}$ besteht also aus allen Normalverteilungsdichten $\pi_{\mu,\nu}$ mit Erwartungswert μ und Varianz ν^2, deren Erwartungswert μ keinen ‚zu großen Abstand' zu einem festgelegten μ_0 hat. σ^2 ist die Varianz der Stichprobenverteilung, und c eine nicht näher bestimmte Konstante.

Die betrachtete translations-invariante Menge von konjugierten Verteilungen, die mit der oben beschriebenen Menge von Doppelexponentialverteilungen verglichen wird, wird dann über den Grenzübergang $\nu \to \infty$ erzeugt. Dadurch kann eine explizite Wahl von μ_0 überflüssig gemacht werden, da die Posteriori-Parameter für $\nu \to \infty$ nicht mehr von μ_0 abhängen. In dem Modell von Quaeghebeur und de Cooman, das in dieser Arbeit zur Anwendung kommt, muss die Varianz ν^2 jedoch begrenzt werden. Die Ergebnisse von Pericchi und Walley sind daher nicht direkt darauf übertragbar.

Ein interessanter Ansatz stellt die Umgebungsklasse dar, die Pericchi und Walley für Situationen empfehlen, wenn so viel Priori-Information vorhanden ist, dass eine bestimmte Priori-Verteilung als Ausgangspunkt gewählt werden kann. Die Umgebung wird dabei ähnlich wie bei [Coolen 1993] als „interval of measures", als Menge von (nicht normierten) Maßfunktionen definiert, die zwischen einer oberen Funktion $u(\theta)$ und einer unteren Funktion $l(\theta)$ liegen. Im Gegensatz zu [Coolen 1993] müssen diese Funktionen jedoch nicht proportional zueinander sein. Stattdessen wählen Pericchi und Walley mit $u(\theta) = 1$ eine impropere Dichte, und mit $l(\theta)$ als konjugierte Verteilung eine Normalverteilung $\mathrm{N}(\mu, \nu^2)$, die so normiert wurde, dass am Modus (bei $\theta = \mu$) ebenfalls $l(\cdot) = 1$ gilt. $\mathcal{M}$ enthält damit alle Verteilungen, die einen Modus bei $\theta = \mu$ erreichen und deren Verhalten ‚an den Rändern', also weit entfernt von μ, zwischen dem der

gewählten Normalverteilung und einer Gleichverteilung auf $\mathbb{R}$ entspricht. Im Vergleich mit ε-Kontaminationsklassen stellt sich diese Menge als die für diese Aufgabenstellung besser geeignete Vorgehensweise heraus.

2 Beispiel: Das Imprecise Dirichlet Model

Das Imprecise Dirichlet Model (IDM) ist ein mächtiges Modell für Bayes-Lernen bei multinomiale Daten, das die Vorteile eines Kalküls unter Einbeziehung komplexer Unsicherheit anschaulich machen kann. Es wurde 1996 von Walley vorgeschlagen [Walley 1996] und stellt eine Verallgemeinerung des Bayesianischen Modells für multinomiale Daten dar. Es erfüllt eine Reihe von wünschenswerten Inferenzprinzipien, die von alternativen objektiven Ansätzen verletzt werden.

Das IDM erlaubt es, a priori Nichtwissen über die Wahrscheinlichkeiten der Kategorien zu modellieren, indem statt einer einzigen Priori-Verteilung für diese Wahrscheinlichkeiten eine Menge von Priori-Verteilungen verwendet wird. Das Wissen nach der Beobachtung wird durch eine Menge von Posteriori-Verteilungen über die Kategoriewahrscheinlichkeiten dargestellt. Diese Menge von Verteilungen scheint auf den ersten Blick ein sehr komplexes Ergebnis zu sein, über den Begriff der oberen und unteren Wahrscheinlichkeiten wird diese Komplexität jedoch handhabbar.

Ein weiterer besonderer Vorteil des IDM ist, dass die Inferenz nicht von der Art der Aufteilung der Kategorien abhängt, insbesondere nicht von der Anzahl der zu unterscheidenden Kategorien.

Die Vorstellung des IDMs in diesem Kapitel stützt sich auf den ‚Originalartikel' von Walley [Walley 1996] sowie auf einen Seminarvortrag [Strobl 2005] und einen weiteren Artikel [Bernard 2005].

2.1 Multinomiale Daten

Multinomiale Daten sind Beobachtungen, die in endlich viele nicht geordnete Kategorien fallen, die eindeutig voneinander getrennt sind. Jede Beobachtung kann dabei genau einer Kategorie zugeordnet werden, die einzelnen Beobachtungen sind voneinander unabhängig und besitzen eine jeweils identische Verteilung. Die Multinomialverteilung ist also ein allgemeines Modell für unabhängige Beobachtungen beliebiger diskreter Merkmale, die keine ordinale Information liefern.

Gegeben seien n Beobachtungen, die $j = 1, \ldots, k$ Kategorien zugeordnet werden können, wobei einer Kategorie j die Anzahl n_j Beobachtungen zufallen und $\sum_{j=1}^{k} n_j = n$ gilt. Der

Vektor der beobachteten Häufigkeiten $n_1, \ldots, n_k$ sei mit $\mathbf{n}$, die zugehörigen Zufallsgrößen seien mit $N_1, \ldots, N_k$ bezeichnet. Dann kann die multinomiale Dichte folgenermaßen notiert werden:

$$\begin{aligned} f\big((n_1, \ldots, n_k)\big|(\theta_1, \ldots, \theta_k)\big) &= P_{\theta_1,\ldots,\theta_k}(\{N_1 = n_1, \ldots, N_k = n_k\}) \\ &\propto \prod_{j=1}^{k} \theta_j^{n_j} \end{aligned} \tag{2.1}$$

Dabei gilt für die gegebenen Parameter $\boldsymbol{\theta} = \theta_1, \ldots, \theta_k$ der Multinomialverteilung: $\theta_j \in (0, 1)$, $j = 1, \ldots, k$ und $\sum_{j=1}^{k} \theta_j = 1$. $\boldsymbol{\theta}$ kann also als Vektor der Koeffizienten einer Konvexkombination angesehen werden. Die Menge der zulässigen Parameterkombinationen $\boldsymbol{\theta}$ wird mit Θ bezeichnet. Der Parameter θ_j entspricht der Wahrscheinlichkeit, dass eine einzelne Beobachtung in die Kategorie j fällt.

Es ist möglich, die Voraussetzung der eindeutigen Zuteilbarkeit einer Beobachtung zu einer Kategorie aufzuweichen, wenn statt einer Verteilung über die einzelnen Kategorien eine Verteilung über die Potenzmenge der Kategorien definiert wird. Ein solcher Ansatz führt zu einer generalisierten Version des IDMs, die dann unscharfe Beobachtungen zulässt. Eine kurze Vorstellung und Anwendung auf Entscheidungsprobleme kann beispielsweise in [Utkin und Augustin 2005] gefunden werden. In der in dieser Arbeit vorgestellten ‚klassischen' Version des IDM besteht dagegen die Möglichkeit, die Kategorieaufteilung zu verändern, ohne dass sich die Ergebnisse der Inferenz verändern. Durch diese Eigenschaft erfüllt das IDM das Representation Invariance Principle (siehe Kapitel 2.4).

2.2 Bayes-Lernen

Eine bayesianische Analyse der Stichprobe verlangt nach einer Priori-Verteilung (oder kurz ‚Priori') über die Parameter θ_j. Diese soll, entsprechend dem ersten Bayes-Postulat (Notation z.B. bei [Rüger 1999, Kap 2.4, S.186]), alle Informationen über die Parameter vor der Beobachtung der Stichprobe verkörpern. Diese Priori kann dann, gemäß des Satz von Bayes für Dichten, mittels der Likelihood der Beobachtungen aufdatiert werden:

$$f_{\Theta \,|\, X}(\theta \,|\, x) = \frac{f_{X \,|\, \Theta}(x \,|\, \theta) \cdot f_{\Theta}(\theta)}{f_X(x)} \tag{2.2}$$

Priori-Informationen über eine (auch mehrdimensionale) Zufallsgröße Θ, die durch eine Dichte $f_\Theta(\theta)$ über dem Raum der Realisationen θ von Θ repräsentiert werden, werden mittels weiterer Informationen (Beobachtungen) $X = x$ aufdatiert, die durch die Dichte $f_{X \,|\, \Theta}(x \,|\, \theta)$ dargestellt sind. Diese Dichte, statt als Funktion von x (gegeben θ) als Funktion von θ (gegeben x) interpretiert, wird auch als Likelihood bezeichnet, mit der Notation $L_x(\theta)$. Inferenz, der nur die Likelihood als „Informationsquelle" über

die Beobachtungen dient, entspricht dem zweiten Bayes-Postulat und erfüllt somit das Likelihood-Prinzip (siehe Kapitel 2.4).

Das Resultat des Aufdatierungsschritts ist die Dichte $f_{\Theta \mid X}(\theta \mid x)$, die die Posteriori-Informationen über Θ nach der Beobachtung X repräsentiert. Der Faktor $f_X(x)$ im Nenner stellt nur eine Proportionalitätskonstante dar, die die Posteriori in der Weise normiert, dass sie tatsächlich wieder eine Dichte ist.

Diese Posteriori-Verteilung (oder kurz ‚Posteriori') enthält dann nach dem dritten Bayes-Postulat das gesamte Wissen über Θ nach der Beobachtung X. Der Erwartungswert oder der Modus der Posteriori-Dichte (kurz: Posteriori-Erwartungswert bzw. Posteriori-Modus) kann dann beispielsweise als Vorhersage für eine neue Beobachtung x^* dienen, HPD-Intervalle (highest posterior density) liefern das bayesianische Analogon zu Konfidenzintervallen. Der Prozess einer solchen Aufdatierung von Priori-Wissen heißt Bayes-Lernen. Eine so gewonnene Posteriori kann wieder als Priori zur Verarbeitung neuer Beobachtungen verwendet werden.

Im Falle von multinomialen Beobachtungen entsprechen dem Θ die Parameter $(\theta_1, \ldots, \theta_k)$, über die mehr herausgefunden werden soll. Die Beobachtung X, mittels der die Parameter $(\theta_1, \ldots, \theta_k)$ aufdatiert werden, sind die Häufigkeiten n_j der Kategorien $j = \{1, \ldots, k\}$. $f_{X \mid \Theta}(x \mid \theta)$ entspricht also der Dichte (2.1), $f\big((n_1, \ldots, n_k) \big| (\theta_1, \ldots, \theta_k)\big)$. Gleichung (2.2) wird hier daher zu

$$p\big((\theta_1, \ldots, \theta_k) \big| (n_1, \ldots, n_k)\big) \quad \propto \quad f\big((n_1, \ldots, n_k) \big| (\theta_1, \ldots, \theta_k)\big) \cdot p(\theta_1, \ldots, \theta_k)\,.$$

2.3 Konjugierte Verteilungen

Die Priori-Informationen über $\theta_1, \ldots, \theta_k$ müssen also durch eine Dichte $f(\theta_1, \ldots, \theta_k)$ entsprechend $f_X(x)$ aus (2.2) dargestellt werden. Eine geeignete Priori für diese Parameter ist die Dirichlet-Verteilung:

$$p(\theta_1, \ldots, \theta_k) \quad \propto \quad \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \quad = \quad \prod_{j=1}^{k} \theta_j^{s \cdot t_j - 1}\,, \tag{2.3}$$

mit $\alpha_j > 0$ bzw. $s > 0$ und $t_j \in (0, 1)$, $j = 1, \ldots, k$, $\sum_{j=1}^{k} t_j = 1$. Notation:

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k) \text{ bzw. } \boldsymbol{\theta} \sim \mathrm{Dir}(s, t_1, \ldots, t_k)$$

Die zwei Parametrisierungen mit $(\alpha_1, \ldots, \alpha_k)$ oder $(s, \boldsymbol{t}) = (s, t_1, \ldots, t_k)$ sind äquivalent; die Parametrisierung mit $(s, t_1, \ldots, t_k)$ hat jedoch den Vorteil, dass die Parameter $t_1, \ldots, t_k$ ebenso wie $\theta_1, \ldots, \theta_k$ die Koeffizienten einer Konvexkombination bilden und t_j jeweils dem Erwartungswert von θ_j entspricht:

$$\mathbb{E}[\theta_j] \quad = \quad \frac{\alpha_j}{\sum_{i=1}^{k} \alpha_i} \quad = \quad \frac{s \cdot t_j}{s \cdot \sum_{i=1}^{k} t_i} \quad = \quad t_j \tag{2.4}$$

Der Parameter s stellt einen Hyperparameter da, der die Stärke des Einflusses der Priori-Informationen auf die Posteriori steuert.

Die Dirichlet-Verteilung ist besonders geeignet, da sie die konjugierte Priori zur Multinomialverteilung darstellt. Das bedeutet, dass die Posteriori nach Aufdatierung mit multinomialverteilten Daten wieder vom selben Verteilungstyp wie die Priori ist und sich nur die Parameter der Priori ändern. Dadurch wird die Posteriori analytisch bequem fassbar, da Erwartungswert, Modus und andere Kennwerte der Verteilung i.A. bekannte und einfache Funktionen der Parameter sind. Außerdem ergibt sich die Normierungskonstante der Posteriori-Dichte aus der Identifizierung mit einem bekannten Verteilungstyp. Dies ist deshalb wünschenswert, da $f_X(x)$, die unbedingte Dichte der Beobachtung im Nenner von (2.2), oft nicht analytisch ermittelt werden kann.

Die Verwendung einer nicht-konjugierten Priori ist natürlich ebenso möglich, die Posteriori entspricht dann jedoch nicht notwendigerweise der Dichte einer bekannten Verteilung, sondern einer Mischverteilung. Eine solche Mischverteilung ist jedoch i.A. schwieriger zu handhaben, da die Normierungskonstante nur über Integrale erhältlich ist, die sehr schwierig oder gar nicht analytisch lösbar sind und z.B. obengenannte Kennzahlen nicht unbedingt als Funktionen der Parameter der beiden Ausgangsverteilungen berechnet werden können.

Die Konjugiertheit einer Dirichlet-Priori mit der Multinomialverteilung kann folgendermaßen gezeigt werden:

$$\begin{aligned} p\big((\theta_1,\ldots,\theta_k)\big|(n_1,\ldots,n_k)\big) &\propto f\big((n_1,\ldots,n_k)\big|(\theta_1,\ldots,\theta_k)\big)\cdot p(\theta_1,\ldots,\theta_k) \\ &= \prod_{j=1}^{k}\theta_j^{n_j}\cdot\prod_{j=1}^{k}\theta_j^{s\cdot t_j-1} \\ &= \prod_{j=1}^{k}\theta_j^{n_j+s\cdot t_j-1} \\ &= \prod_{j=1}^{k}\theta_j^{\overbrace{(n+s)}^{=:s^*}\overbrace{\frac{n_j+s\cdot t_j}{n+s}}^{=:t_j^*}-1} \end{aligned}$$

Die Posteriori-Dichte hat also die gleiche Form wie die Priori-Dichte, jedoch mit geänderten, aufdatierten Parametern

$$s^* = n+s \qquad \text{und} \tag{2.5}$$

$$\begin{aligned} t_j^* &= \frac{n_j+s\cdot t_j}{n+s} \\ &= \frac{n_j}{n+s}+\frac{s}{n+s}\cdot t_j\,. \end{aligned} \tag{2.6}$$

Ein aufdatiertes t_j wird also mit $\frac{s}{n+s}$ skaliert und um $\frac{n_j}{n+s}$ verschoben.

Dieses Modell von Bayes-Lernen mit multinomialen Daten und einer dirichletverteilten Priori wird Dirichlet-Multinomial-Modell genannt.

Eine Besonderheit der Dirichlet-Verteilung ist, dass sich nach der Zusammenlegung von zwei oder mehreren Kategorien wieder eine Dirichlet-Verteilung ergibt; ebenso kann jede Kategorie in beliebig viele Unterkategorien aufgeteilt werden, deren Verteilung dann einer von der ursprünglichen Kategorieaufteilung unabhängigen Dirichlet-Verteilung entspricht. Sowohl Aufteilung als auch Zusammenlegung lassen dabei den Parameter s unverändert.

Diese vorteilhafte Eigenschaft lässt sich noch genereller ausdrücken: Die Vergröberung der Kategorieaufteilung kann in mehreren Stufen in beliebiger Weise geschehen, bis nur noch eine einzige Kategorie übrig bleibt. Die verschiedenen Stadien der Vergröberungen können dann, wie in Abbildung 2.1 illustriert, als Baumstruktur aufgefasst werden: Jede ursprüngliche Kategorie der feinsten Aufteilung entspricht einem Blatt des Baumes. Durch die Zusammenfassung mehrerer Kategorien zu einer einzigen entsteht eine Astgabelung darunter, die, bei weiterer Vergröberung der Aufteilung, mit einer noch weiter darunterliegenden Astgabelung verbunden ist. Bei der letzten Zusammenfassung in nur noch eine Kategorie entsteht dann die unterste Astgabelung, die Wurzel des Baumes.

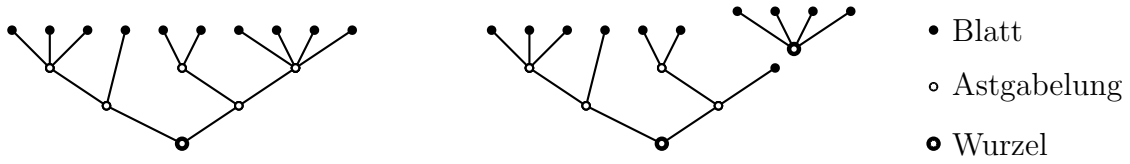

Abbildung 2.1: Baumstruktur von Kategorien: links ein kompletter Baum, rechts nach Unterteilung an einer Astgabelung

Ein Vergröberungsschritt könnte, beispielsweise bei einer Untersuchung des Pasta-Kaufverhaltens italienischer Haushalte, die Zusammenfassung der Pasta-Sorten ‚Linguine', ‚Buccatini', ‚Spaghettini' und ‚Reginette' zur Kategorie ‚Pasta lunga' darstellen; die Pasta-Sorten ‚Penne', ‚Gemelli', ‚Orecchiette'und ‚Anelli Siciliani' könnten zur Kategorie ‚Pasta corta' zusammengefasst werden.

Ein solcher Baum kann an einer beliebigen Astgabelung aufgeteilt werden; der ‚Aststumpf' wird zu einem Blatt des großen Baumes, der abgetrennte Teil ist dann ein eigener kleiner Baum, der an der Schnittstelle seine Wurzel hat. Die so entstandenen Bäume gehorchen dann zwei stochastisch unabhängigen Dirichletverteilungen, deren Kategoriewahrscheinlichkeiten $\boldsymbol{t}$ sich aus denen des ursprünglichen Baums ableiten lassen, wobei jedoch s unverändert bleibt. In Kapitel 3.3 von [Bernard 2005] wird diese Eigenschaft durch zwei Sätze etabliert.

2.4 Inferenzprinzipien und alternative Modellierungen

Klassisches Bayesianisches Schließen stützt sich im Normalfall auf eine konjugierte Priori. Das vorhandene Vorwissen muss also in Parameterwerte der konjugierten Priori übersetzt werden. Idealerweise ist das Vorwissen so präzise, dass (etwa aufgrund der Ergebnisse vorheriger Untersuchungen) eindeutige Werte für die Parameter der Priori naheliegen. Falls das Vorwissen jedoch unpräzise ist oder gar praktisch Nichtwissen herrscht, ist es schwierig, sich auf solche eindeutigen Werte (im Falle multinomialer Daten für $t_1, \ldots, t_k$ und s) festzulegen.

Die bayesianische Lösung in diesem Fall ist die Verwendung einer nichtinformativen Priori. Die allgemeinen Probleme dieser Vorgehensweise wurden schon in Abschnitt 1.2 dargelegt. Für das Dirichlet-Multinomial-Modell wurden verschiedene Vorschläge gemacht, die alle symmetrischen Dirichlet-Verteilungen entsprechen, also mit $t_j =: \frac{1}{k}$, $\forall j = 1, \ldots, k$, jedoch mit unterschiedlichen Werten von s.

$s = k$ ergibt die Laplace'sche Priori entsprechend dem Prinzip vom unzureichenden Grund, die diskrete Gleichverteilung über alle k Kategorien. Die Verwendung dieser Priori entspricht einem frequentistischen Vorgehen, bei dem die Basis der Inferenz die relativen Häufigkeiten der Stichprobe, $\frac{n_j}{n}$ für alle $j = 1, \ldots, k$, also der Anteil der in Kategorie j fallenden Beobachtungen, darstellen.

Die nichtinformative Priori nach der Regel von Haldane, die Rüger als die am wenigsten problematische identifiziert, entspricht hier der Wahl von $t_j = \frac{1}{k}$, $\forall j = 1, \ldots, k$ und $s \to 0$. Sie erfüllt die wichtigsten der nach Walley wünschenswerten Inferenz-Prinzipien, welche im folgenden erläutert werden. Unter bestimmten Voraussetzungen führt sie jedoch zu unsinnigen Ergebnissen, die sich im IDM vermeiden lassen. Darauf wird in Kapitel 2.6 hingewiesen werden.

- *Symmetry principle (SP)*: Unterschiedsloses Nichtwissen über alle Kategoriewahrscheinlichkeiten θ_j sollte invariant gegenüber Permutationen in den Kategorien sein.
- *Embedding principle (EP)*: Aussagen zur Priori-Unsicherheit über ein beliebiges Ereignis B sollten nicht von der Konfiguration oder Art der Aufteilung der Kategorien abhängen, da solche Aufteilungen in den meisten Fällen (auch in der Praxis) rein arbiträr (willkürlich) sind.
- *Representation invariance principle (RIP)*: Aussagen zur Posteriori-Unsicherheit über ein beliebiges Ereignis B sollten nicht von der Konfiguration oder Art der Aufteilung der Kategorien abhängen. Der Unterschied zum EP liegt nur in der Posteriori-Betrachtungsweise, man könnte EP und RIP also auch als ein Prinzip auffassen.
- *Likelihood principle*: Die a posteriori Inferenz sollte bezüglich der Daten nur von deren Likelihood abhängen (und nicht von einer Stopregel beispielsweise).

- *Coherence principle (CP)*: Die Inferenz sollte kohärent sein im Sinne des Kohärenzbegriffs von Walley (siehe [Walley 1991, Kap. 2.2]). Diese Bedingung soll gewährleisten, dass sich verschiedene Inferenz-Aussagen nicht widersprechen.

Diese Inferenzprinzipien sollen nun an einem kurzen Beispiel, dem ‚klassischen', in [Walley 1996] verwendeten „Bag of marbles"- Beispiel erläutert werden:

Aus einem geschlossener Beutel, der mit Murmeln unbekannter, möglicherweise verschiedener Farben gefüllt ist, wird eine Murmel gezogen. Wie groß ist die Wahrscheinlichkeit, dass die gezogene Murmel rot ist?

Wenn zuvor noch keine Murmeln aus dem Beutel gezogen wurden, sind wir in der Situation, Priori-Inferenz zu betreiben. Damit das EP erfüllt ist, sollte also gelten: Die Wahrscheinlichkeit für die Aussage „die gezogene Murmel ist rot" hängt nicht davon ab, wie die Farbkategorien beschaffen sind. Es soll also keinen Unterschied machen, ob die Menge der Kategorien als {rot, andere Farben}, {rot, blau, weiß, andere Farben} oder gar {rot, blau, weiß, weiß mit blauen Streifen und einem roten Wölkchen in der Mitte, andere Farben} definiert wurde.

Wurde zuvor schon eine Sequenz von Murmeln aus dem Beutel gezogen, sind wir in der Situation des RIP, und auch nach dem Zug einer dunkelorangenen Murmel, die es vielleicht nötig macht, die Kategorieaufteilung zu verändern, soll sich die Wahrscheinlichkeitsbewertung für „die nächste gezogene Murmel ist rot" nicht ändern.

Klarerweise verletzen bayesianischen Modelle mit nichtinformativen Priori-Verteilungen mit $s > 0$ das EP und das RIP, da sie von der Anzahl der Kategorien k abhängen. Gemäß ihrem Konstuktionsprinzip erfüllen sie jedoch das SP und, wie alle bayesianischen Methoden, gemäß dem zweiten Bayes-Postulat das LP. Walley zeigt in [Walley 1991, Kap. 5.5] sogar, dass sich das SP und EP / RIP in bayesianischen Modellen mit properen Priori-Verteilungen grundsätzlich gegenseitig ausschließen.

Frequentistische Methoden verletzen in der Regel das LP, da deren Ergebnisse nicht nur von der Likelihood allein, sondern auch von der Stopregel abhängen. Im Murmelbeispiel macht es bei der frequentistischen Inferenz einen Unterschied, ob die Zahl der Züge vorher feststand, oder ob gezogen wurde, bis z.B. drei rote Murmeln gezogen wurden. Ebenso verletzen sie meist das CP.

Auf Unterschiede in den Ergebnissen der Inferenz selbst wird in Kapitel 2.6 eingegangen.

2.5 Das Imprecise Dirichlet Model für $\theta_1, \ldots, \theta_k$

Das Grundmodell des IDM besteht aus allen Dirichlet-Verteilungen mit einem festen $s^{(0)} > 0$ und $\boldsymbol{t}^{(0)}$ mit $t_j^{(0)} \in (0,1)$ für alle $j = 1, \ldots, k$ und $\sum_{j=1}^{k} t_j^{(0)} = 1$. Der obere Index $^{(0)}$ soll die Parameter als Priori-Werte identifizieren.

$$\mathcal{M}^{(0)} = \left\{ \mathrm{Dir}\left(s^{(0)}, t_1^{(0)}, \ldots, t_k^{(0)}\right) \;\middle|\; t_j^{(0)} \in (0,1),\, j = 1, \ldots, k,\, \sum_{j=1}^{k} t_j^{(0)} = 1 \right\}. \tag{2.7}$$

Sei $T^{(0)}$ außerdem die Menge der Parameter $\boldsymbol{t}^{(0)}$ und $s^{(0)}$ aller Dirichlet-Verteilungen in $\mathcal{M}^{(0)}$. Die Priori-Gewichte für jede Kategorie $j \in \{1, \ldots, k\}$ variieren im Intervall $(0,1)$, der ‚Hyperparameter' $s^{(0)}$, der den Einfluss der Priori-Informationen auf die Posteriori-Wahrscheinlichkeiten steuert, muss fest gewählt werden. Auf die Wahl von $s^{(0)}$ wird in den nächsten Absätzen näher eingegangen. Die Menge $\mathcal{M}^{(0)}$ dieser Verteilungen dient als Modell für a priori Nichtwissen über die Parameter $\theta_1, \ldots, \theta_k$ der Multinomialverteilung, da alle vorstellbaren nichttrivialen Verteilungen in $\mathrm{conv}\left(\mathcal{M}^{(0)}\right)$, der konvexen Hülle von $\mathcal{M}^{(0)}$, enthalten sind. Es reicht jedoch, sich für jede weitere Berechnung nur auf alle ‚reinen' Dirichlet-Verteilungen zu stützen, da diese die ‚Ecken' darstellen, welche die konvexe Hülle definieren.

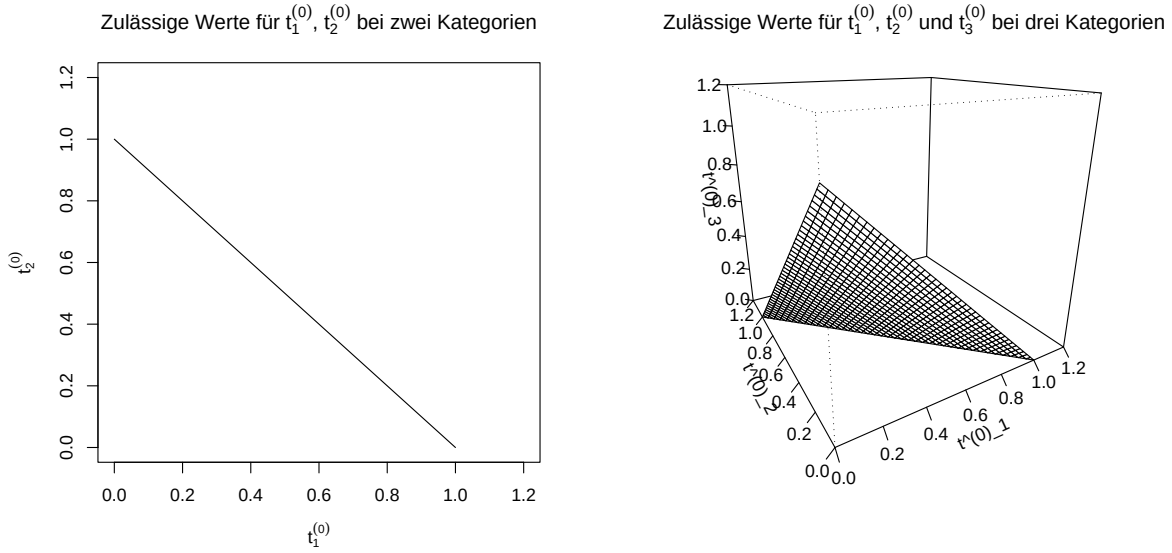

Abbildung 2.2: Darstellung der zulässigen Parameterkombinationen für $k = 2$ und $k = 3$.

Jeder der Priori-Verteilungen aus $\mathcal{M}^{(0)}$ entspricht einer bestimmten Konfiguration der Werte von $t_1^{(0)}, \ldots, t_k^{(0)} \in T^{(0)}$, die als innerer Punkt eines Ausschnitts der $k-1$-dimensionalen Hyperebene im k-dimensionalen Raum dargestellt werden kann, die durch die k Punkte $(1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, ……, $(0, 0, \ldots, 0, 1)$ begrenzt wird.

Bei zwei Kategorien sind alle möglichen Kombinationen von $t_1^{(0)}$ und $t_2^{(0)}$ auf der Strecke zu finden, die auf der Geraden mit der Gleichung $t_2^{(0)} = 1 - t_1^{(0)}$ liegt und durch die Punkte $(0, 1)$ und $(1, 0)$ begrenzt wird. Bei drei Kategorien liegen alle Kombinationen von $t_1^{(0)}, t_2^{(0)}$ und $t_3^{(0)}$ auf einer Fläche, die in der Ebene durch die Punkte $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ liegt und durch ebendiese Punkte begrenzt wird.

Falls a priori mehr als nichts über $\theta_1, \ldots, \theta_k$ bekannt ist, kann $\mathcal{M}^{(0)}$ entsprechend eingeschränkt werden. Ist im oben illustrierten Beispiel mit drei Kategorien bekannt, dass $t_1 \geq 0.2$ und $t_3 \leq 0.6$ gilt, so wähle man $\mathcal{M}^{(0)}$ folgendermaßen:

$$\mathcal{M}^{(0)} = \left\{ \mathrm{Dir}\left(s^{(0)}, \mathbf{t}^{(0)}\right) \;\middle|\; t_1^{(0)} \in [0.2, 1),\, t_2^{(0)} \in (0, 0.8],\, t_3^{(0)} \in (0, 0.6],\, \sum_{j=1}^{3} t_j^{(0)} = 1 \right\} \quad (2.8)$$

Die Bedingung für $t_2^{(0)}$ (also $t_2^{(0)} \leq 0.8$) ergibt sich durch die Bedingung der ‚Mindeststärke' 0.2 für $t_1^{(0)}$.

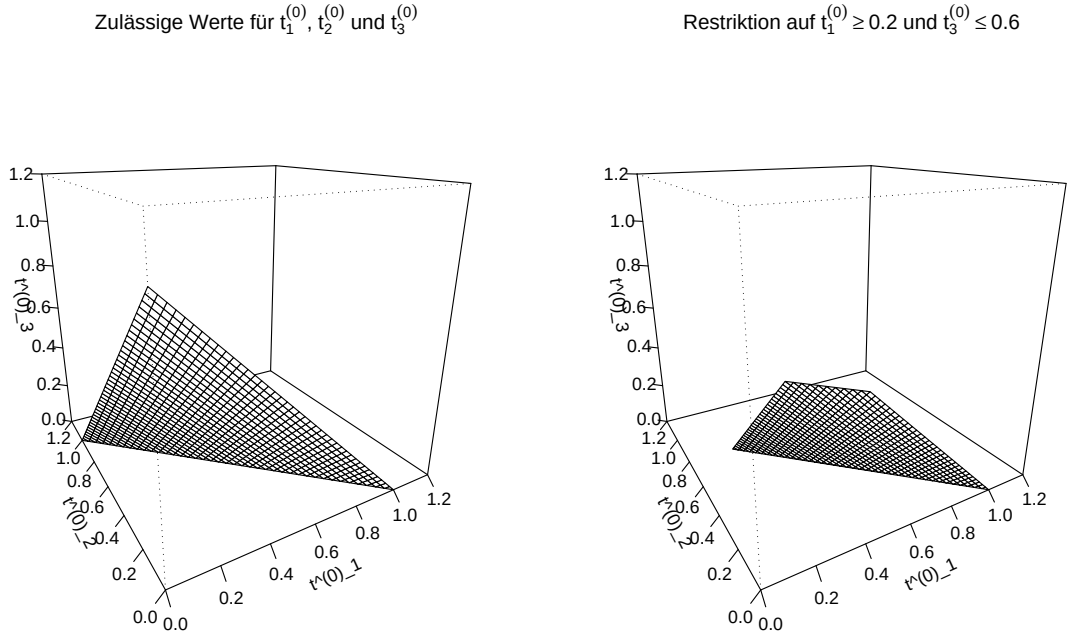

Abbildung 2.3: Darstellung der noch zulässigen Parameterkombinationen für $\mathcal{M}^{(0)}$ aus Gleichung (2.8).

Eine einzige Priori-Verteilung anzugeben, die die Priori-Informationen $t_1^{(0)} \geq 0.2$, $t_3^{(0)} \leq 0.6$ angemessen verkörpert, ist wohl unmöglich.

Der Hyperparameter $s^{(0)}$, der den Grad der Unsicherheit der Priori-Information repräsentiert, muss fest gewählt werden. Größere Werte von $s^{(0)}$ entsprechen größerer Unsicherheit; die Differenz von oberer und unterer Wahrscheinlichkeit für ein Ereignis hängt direkt von $s^{(0)}$ ab (siehe Kapitel 2.6). Im Kontext von a priori Nichtwissen ist die Wahl von $s^{(0)}$ nicht eindeutig gelöst. Von der Intention des IDM her sollten obere und untere Wahrscheinlichkeiten für Ereignisse frequentistische und objektive Bayes-Wahrscheinlichkeiten einschließen; dieses Kriterium legt Werte von $1 \leq s^{(0)} \leq 2$ nahe. Walley schlug in [Walley 1996] $s^{(0)} = 1$ vor, da ihm $s = 2$ als zu vorsichtig schien; laut [Bernard 2005] unterstützen andere Untersuchungen [Bernard 2001, Bernard 2003] jedoch auch $s^{(0)} = 2$, besonders wenn die Anzahl der Kategorien k groß wird. (Eine Wahl von $s^{(0)}$ abhängig von k verletzt jedoch das RIP.) Genauere Abhandlungen bezüglich der Wahl von $s^{(0)}$ im IDM können in den ebengenannten Artikeln sowie in [Walley 1996] gefunden werden. Für eine realistische Modellierung kann es aber auch nötig sein kann, dass auch $s^{(0)}$ in einer Menge variiert. Darauf wird in Abschnitt 2.6.1 kurz eingegangen.

$s^{(0)}$ kann auch als Anzahl der ‚verdeckten' Beobachtungen interpretiert werden. Diese Interpretation wird anhand von Gleichung (2.11) verdeutlicht werden.

Nach n Beobachtungen, die sich auf jeweils n_j Beobachtungen in den Kategorien $j = 1, \ldots, k$ aufteilen, verändert sich $\mathcal{M}^{(0)}$ aus (2.7) mittels der Aufdatierungsregeln (2.5) und (2.6) zu

$$\mathcal{M}^{(n)} = \left\{ \mathrm{Dir}\left(\left(s^{(0)} + n\right), \frac{n_1 + s^{(0)} \cdot t_1^{(0)}}{n+s}, \cdots, \frac{n_k + s^{(0)} \cdot t_k^{(0)}}{n+s} \right) \,\middle|\, s^{(0)}, \boldsymbol{t}^{(0)} \in T^{(0)} \right\}. \quad (2.9)$$

Durch die Aufdatierung werden also die Parameter $t_1^{(0)}, \ldots, t_k^{(0)}$ der Menge $T^{(0)}$ mit $\frac{s^{(0)}}{n+s^{(0)}}$ skaliert und um $\frac{n_j}{n+s^{(0)}}$ verschoben, der Hyperparameter $s^{(0)}$ erhöht sich um n. s stellt also auch eine Art ‚Priori-Stärke' dar, die steuert, mit welchem Gewicht der Wert des Priori-Parameters $t_j^{(0)}$ in den Wert des Posteriori-Parameters eingeht.[1] Diese so aufdatierten Parameter bekommen entsprechend der Anzahl n der Beobachtungen, die zur Aufdatierung verwendet wurden, den oberen Index $^{(n)}$.

2.6 Inferenz mit dem IDM

Da im IDM die Priori-Information über eine Menge von Priori-Verteilungen dargestellt wird, ergibt sich als Posteriori-Information ebenfalls eine Menge von Verteilungen. Aus eindimensionalen Zielgrößen wie beispielsweise dem Posteriori-Erwartungswert oder der Posteriori-Wahrscheinlichkeit für ein Ereignis werden im IDM daher Mengen dieser Zielgrößen, deren Elemente aus der Anwendung je einer der Posteriori-Verteilungen $\pi^{(n)}$ aus $\mathcal{M}^{(n)}$ hervorgehen. Tatsächlich ist aber meistens nicht jedes einzelne dieser Elemente

[1] Auf diese Interpretation wird noch in Kapitel 3.3 näher eingegangen.

von Interesse, sondern vielmehr die Spanne, in der sich die Zielgrößen bewegen. Zur Abschätzung der aus dem IDM resultierenden Posteriori-Wahrscheinlichkeiten eines Ereignisses A_j, dass die nächste Beobachtung zur Kategorie j gehört, kann man sich auf die minimale und maximale der Wahrscheinlichkeiten aus dieser Spanne beschränken, zumal man weiß, dass jeder Wert dazwischen auch angenommen wird. Die minimale wird als untere Wahrscheinlichkeit $\underline{P}(A_j \,|\, \mathbf{n})$, die maximale als obere Wahrscheinlichkeit $\overline{P}(A_j \,|\, \mathbf{n})$ bezeichnet und sind also folgendermaßen definiert:

$$\begin{aligned} \underline{P}(A_j \,|\, \mathbf{n}) &:= \min_{\pi^{(n)} \in \mathcal{M}^{(n)}} P_{\pi^{(n)}}(A_j) \\ \overline{P}(A_j \,|\, \mathbf{n}) &:= \max_{\pi^{(n)} \in \mathcal{M}^{(n)}} P_{\pi^{(n)}}(A_j) \end{aligned} \tag{2.10}$$

Um $\overline{P}(A_j \,|\, \mathbf{n})$ zu erhalten, muss man aber nicht alle (unendlich viele!) Posteriori-Verteilungen durchgehen, die jeweilige Wahrscheinlichkeit für A_j berechnen und dann das Maximum bestimmen. Stattdessen kann diese Suche auf $\mathcal{M}^{(0)}$ zurückgeführt werden. Es gilt für alle $\pi^{(n)}$ aus $\mathcal{M}^{(n)}$, die zu $t_j^{(n)}$ ‚kompatibel' sind, d.h. die so beschaffen sind, dass die Posteriori-Wahrscheinlichkeit für die Kategorie j gemäß (2.4) $t_j^{(n)}$ beträgt:

$$\begin{aligned} P_{\pi^{(n)}}(A_j) &= t_j^{(n)} \quad \text{, also mit (2.6)} \\ &= \frac{n_j + s^{(0)} \cdot t_j^{(0)}}{n + s^{(0)}} \end{aligned}$$

Aufgrund der Linearität des Aufdatierungsschritts muss man daher nur das kleinste bzw. das größte $t_j^{(0)}$ aus $\mathcal{M}^{(0)}$ ermitteln. Im Fall von Priori-Nichtwissen, bei dem also $\mathcal{M}^{(0)}$ aus allen möglichen Dirichlet-Verteilungen besteht (siehe Gleichung (2.7)), lässt sich $\underline{P}(A_j \,|\, \mathbf{n})$ also durch den Grenzübergang $t_j^{(0)} \to 0$ und $\overline{P}(A_j \,|\, \mathbf{n})$ durch $t_j^{(0)} \to 1$ bestimmen. (Die ‚Ränder' 0 und 1 wurden in der Definition ausgeschlossen.) Es gilt also

$$\underline{P}(A_j \,|\, \mathbf{n}) = \frac{n_j}{n + s^{(0)}} \quad \text{und} \quad \overline{P}(A_j \,|\, \mathbf{n}) = \frac{n_j + s^{(0)}}{n + s^{(0)}} \,. \tag{2.11}$$

Hieran wird die Interpretation von $s^{(0)}$ als Anzahl der ‚verdeckten' Beobachtungen deutlich: Im Nenner steht die Summe der n ‚sichtbaren' und $s^{(0)}$ ‚verdeckten' Beobachtungen; die Untergrenze für $P(A_j \,|\, n)$ ergibt sich aus der Annahme, das keine der ‚verdeckten' Beobachtungen zu der Kategorie j gehört, die Obergrenze durch die Annahme, das alle $s^{(0)}$ Beobachtungen dazu gehören.

An (2.11) ist auch erkennbar, dass die obere und untere Wahrscheinlichkeit für A_j nicht von der Anzahl der Kategorien abhängt. In bayesianischen Modellen mit nichtinformativer Priori ist dies im Allgemeinen nicht der Fall. Dort gilt

$$P(A_j \,|\, \mathbf{n}) = \frac{n_j + s^{(0)} \frac{1}{k}}{n + s^{(0)}} \,,$$

und jede Wahl einer nichtinformativen Priori führt zu einer anderen posteriori-Wahrscheinlichkeit, die, außer bei der Verwendung der Priori nach der Regel von Haldane, von der Anzahl der Kategorien k abhängt.

Darüber hinaus sind diese posteriori-Wahrscheinlichkeiten auch nicht in der Lage, die Menge an Beobachtungen, die ihnen zugrunde liegt, angemessen widerzuspiegeln. Dazu ein kurzes Beispiel: Wird nur zwischen zwei Kategorien unterschieden ($k = 2$) und liegen die Hälfte der Beobachtungen in Kategorie j ($n_j = \frac{n}{2}$), ist die bayesianische Posteriori-Wahrscheinlichkeit unter allen genannten nichtinformativen priori-Verteilungen $P(A_j \,|\, \mathbf{n}) = \frac{1}{2}$, und zwar für jede Anzahl von Beobachtungen n. Die oberen und unteren Wahrscheinlichkeiten in (2.11) hingegen hängen von n ab und rücken immer näher zusammen, je größer n wird: Je mehr Daten eine Vorhersage als Grundlage hat, desto genauer wird sie. Sind jedoch wenig Daten vorhanden, liegen $\underline{P}(A_j \,|\, \mathbf{n})$ und $\overline{P}(A_j \,|\, \mathbf{n})$ weit auseinander und verdeutlichen so den unsicheren Charakter einer Prognose, im Gegensatz zu den bayesianischen Wahrscheinlichkeiten, die unverhältnismäßig genaue Vorhersagen liefern.

Analog zum Vorgehen bei der Ermittlung von $\underline{P}(A_j \,|\, \mathbf{n})$ und $\overline{P}(A_j \,|\, \mathbf{n})$ gilt im IDM allgemein: Ist die Zielgröße eine lineare Funktion der Priori-Parameter $\boldsymbol{t}^{(0)}$ (wie hier im Beispiel die Wahrscheinlichkeit für ein Ereignis A_j), lässt sich der obere und untere Rand des Zielgrößenbereichs auf den oberen und unteren Rand der Priori-Verteilung zurückführen und rechnerisch einfach bestimmen.

2.6.1 Prädiktive Inferenz

Unter prädiktiver Inferenz wird die Ermittlung von Wahrscheinlichkeiten für Ereignisse verstanden, die zukünftige Beobachtungen betreffen. In Verallgemeinerung des Einführungsbeispiels gilt für ein Ereignis A, dass die nächste Beobachtung in eine der Kategorien $j_1, \ldots, j_l$ mit $\{j_1, \ldots, j_l\} \subset \{1, \ldots, k\}$ fällt

$$\begin{aligned} P_{\pi^{(n)}}(A) &= \sum_{i \in j_1, \ldots, j_l} t_i^{(n)} \\ &= \frac{\sum_{i \in \{j_1, \ldots, j_l\}} n_i + s^{(0)} \cdot \sum_{i \in \{j_1, \ldots, j_l\}} t_i^{(0)}}{n + s^{(0)}} \end{aligned} \tag{2.12}$$

und damit unter Priori-Nichtwissen, mit der Definition $\sum_{i \in j_1, \ldots, j_l} n_i =: n(A)$

$$\underline{P}(A \,|\, \mathbf{n}) = \frac{n(A)}{n + s^{(0)}} \quad \text{und} \quad \overline{P}(A \,|\, \mathbf{n}) = \frac{n(A) + s^{(0)}}{n + s^{(0)}} \,.$$

Dieses Ergebnis erhält man, wenn man $\sum_{i \in \{j_1, \ldots, j_l\}} t_i^{(0)} \to 0$ bzw. $\sum_{i \in \{j_1, \ldots, j_l\}} t_i^{(0)} \to 1$ anwendet.

Ist $\mathcal{M}^{(0)}$ aufgrund von Vorwissen restringiert wie beispielsweise in (2.8), so kann je nach Art der Restriktion die minimal bzw. maximal mögliche Summe der zu A gehörenden priori-Kategorieparameter,

$$t(A)_{\min} := \min_{t_i^{(0)} \in T^{(0)}} \sum_{i \in \{j_1,\ldots,j_l\}} t_i^{(0)} \quad \text{und} \quad t(A)_{\max} := \max_{t_i^{(0)} \in T^{(0)}} \sum_{i \in \{j_1,\ldots,j_l\}} t_i^{(0)} ,$$

ungleich 0 bzw. 1 sein. Daher gilt allgemein

$$\underline{P}(A \,|\, \mathbf{n}) = \frac{n(A) + s^{(0)} \cdot t(A)_{\min}}{n + s^{(0)}} \quad \text{und} \quad \overline{P}(A \,|\, \mathbf{n}) = \frac{n(A) + s^{(0)} \cdot t(A)_{\max}}{n + s^{(0)}} . \tag{2.13}$$

Sind also A und das Vorwissen so beschaffen, dass $t(A)_{\min} > 0$ gilt, so ist $\underline{P}(A \,|\, \mathbf{n})$ um $\frac{s^{(0)} \cdot t(A)_{\min}}{n + s^{(0)}}$ größer als unter Priori-Nichtwissen. Falls $t(A)_{\max} < 1$ gilt, so ist $\overline{P}(A \,|\, \mathbf{n})$ um $\frac{s^{(0)} \cdot (1 - t(A)_{\max})}{n + s^{(0)}}$ kleiner. Das Vorwissen kann also die Spanne der Wahrscheinlichkeiten für ein Ereignis A verkleinern, wenn es das Ereignis A tangiert.

Ebenso kann auch die obere und untere Wahrscheinlichkeit für ein Ergebnis der nächsten zwei (oder mehr) Züge durch Minimieren bzw. Maximieren der Wahrscheinlichkeit $P_{\pi^{(n)}}(A, B, \ldots)$ bezüglich von $\boldsymbol{t}$ berechnet werden, wobei mit A das Ereignis im nächsten Zug, B das Ereignis im darauf folgenden Zug usw. bezeichnet wird. Die Formelausdrücke werden dann entsprechend komplizierter, und es müssen je nach Beschaffenheit der Ereignisse $A, B, \ldots$ Fallunterscheidungen vorgenommen werden.

Als Maß für die Unsicherheit bezüglich der Vorhersage eines Ereignisses A kann man die Spanne dessen posteriori-Wahrscheinlichkeiten, also die Länge des Wahrscheinlichkeitsintervalls betrachten:

$$\begin{aligned} \overline{P}(A \,|\, \mathbf{n}) - \underline{P}(A \,|\, n) &= \frac{n(A) + s^{(0)} \cdot t(A)_{\max}}{n + s^{(0)}} - \frac{n(A) + s^{(0)} \cdot t(A)_{\min}}{n + s^{(0)}} \\ &= \frac{s^{(0)} \cdot [t(A)_{\max} - t(A)_{\min}]}{n + s^{(0)}} \end{aligned} \tag{2.14}$$

Mit wachsendem n wird diese Differenz geringer; mit $n \to \infty$ wird sie 0, d.h., je mehr Beobachtungen zur Verfügung stehen, desto genauer werden die Ergebnisse, so dass bei genügend vielen Beobachtungen $\underline{P}(A \,|\, \mathbf{n})$ und $\overline{P}(A \,|\, \mathbf{n})$ praktisch zusammenfallen. Die Differenz $t(A)_{\max} - t(A)_{\min}$ beträgt höchstens 1 und ist je nach Grad der Restriktion von $\mathcal{M}^{(0)}$ auch kleiner. Je genauer also das Vorwissen ist, desto genauer werden dann auch Vorhersagen im IDM.

Die Konvergenz von $\overline{P}(A \,|\, \mathbf{n})$ und $\underline{P}(A \,|\, \mathbf{n})$ bezüglich $n \to \infty$ gilt allerdings auch dann in gleichem Maße, wenn das Vorwissen nicht zu den Daten passt. Das Intervall $\left[\overline{P}(A \,|\, \mathbf{n}) \,,\, \underline{P}(A \,|\, \mathbf{n})\right]$ ‚wandert' in diesem Fall zwar in die Richtung, die die Daten nahe legen, aber es kann nicht größer werden, da es nicht mehr von $\frac{n(A)}{n}$ abhängt.

Damit das IDM auch in solchen ‚prior-data conflict'-Situationen (siehe Kapitel 1.4.3) das gewünschte Verhalten zeigt, nämlich eine Verbreiterung des Wahrscheinlichkeitsintervalls, wenn sich Priori-Information und die Beobachtungen widersprechen, muss das IDM modifiziert werden. Die nötige Modifikation im Falle von zwei Kategorien diskutiert Walley in [Walley 1991, Kap. 5.4], sie entspräche im IDM der Möglichkeit der Variation von $s^{(0)}$ zwischen einer unteren Grenze $\underline{s}^{(0)}$ und einer oberen Grenze $\overline{s}^{(0)}$. Die Minimierungen und Maximierungen der jeweiligen Inferenzkalküle müssten dann nicht nur über $\mathbf{t}^{(0)} \in T^{(0)}$, sondern auch über $s^{(0)} \in S^{(0)} = [\underline{s}^{(0)}, \overline{s}^{(0)}]$ gebildet werden.

Im Folgenden soll aber noch die ‚klassische Version' des IDMs Basis der Beschreibung sein. In [Walley 1996] und [Bernard 2005] tritt dieses Problem auch nicht auf, da es dort nur in Situationen, bei denen a priori Nichtwissen herrscht, zur Anwendung kommt.

Vergrößert man den Wert von $s^{(0)}$ für festes $\mathbf{t}$, ergibt sich ein größeres Wahrscheinlichkeitsintervall. Unter priori-Nichtwissen ($t(A)_{\max} - t(A)_{\min} = 1$) gilt allgemein, dass die Intervalle, die sich mit steigendem $s^{(0)}$ vergrößern, umeinander geschachtelt sind, da $\underline{P}(A \,|\, \mathbf{n})$ immer kleiner und $\overline{P}(A \,|\, \mathbf{n})$ immer größer wird. Die Ergebnisse der Inferenz im IDM unter priori-Nichtwissen sind daher konsistent in dem Sinne, dass eine weniger vorsichtige Wahl von $s^{(0)}$ zu einem Intervall führt, das sich innerhalb von allen Intervallen mit größerem $s^{(0)}$ befindet. Ändert sich also die Bewertung des Vorwissens in der Art, dass $s^{(0)}$ verkleinert werden kann, so bleiben danach im IDM nur Werte von $P(A \,|\, \mathbf{n})$ plausibel, die es auch vorher schon waren. Dies steht im Gegensatz zu den Ergebnissen objektiver bayesianischer Verfahren: Bei der Wahl von unterschiedlichen Werten von $s^{(0)}$ ergibt sich jeweils ein anderer Wert für $P(A \,|\, \mathbf{n})$.

Ist $\mathcal{M}^{(0)}$ jedoch a priori restringiert, ist die Verschachtelung bei steigendem $s^{(0)}$ nicht in allen Fällen gewährleistet: Falls $\frac{n(A)}{n} < t(A)_{\min}$, wird $\underline{P}(A \,|\, \mathbf{n})$ mit steigendem $s^{(0)}$ größer, oder falls $\frac{n(A)}{n} > t(A)_{\max}$, wird $\overline{P}(A \,|\, \mathbf{n})$ mit steigendem $s^{(0)}$ kleiner. Die Konsistenzeigenschaft geht verloren, wenn $\mathcal{M}^{(0)}$ ausschließlich Informationen enthält, die den beobachteten Häufigkeiten $\frac{n(A)}{n}$ widersprechen. Im Falle von Priori-Nichtwissen kann dieser Widerspruch jedoch nicht eintreten, da $\mathcal{M}^{(0)}$ aussagelos (nicht-selektiv) ist.

Im IDM ergeben sich auch dann sinnvolle Wahrscheinlichkeiten, falls ein beliebiges Ereignis A nach n Beobachtungen (noch) nicht beobachtet wurde. Dann gilt $n(A) = 0$ und, unter priori-Nichtwissen,

$$\underline{P}(A \,|\, \mathbf{n}) = 0 \quad \text{und} \quad \overline{P}(A \,|\, \mathbf{n}) = \frac{s^{(0)}}{n + s^{(0)}} \,.$$

Die untere Wahrscheinlichkeit für dieses Ereignis ist stets 0, die obere geht gegen 0, wenn $n \to \infty$ und A nie beobachtet wird. Mit jeder weiteren Beobachtung, die nicht zu A gehört, wird eine Beobachtung von A im nächsten Zug unwahrscheinlicher, ein plausibles Verhalten des IDM.

Bayesianische Verfahren bieten keine befriedigende Lösung in solchen Situationen: Mit der (von Rüger als am wenigsten untauglich eingeschätzten) nichtinformativen Priori nach der Regel von Haldane ($\pi^{(0)}$ mit $t_j^{(0)} = \frac{1}{k}\,\forall j = 1,\ldots,k,\ s^{(0)} \to 0$), gilt jedoch, analog zu (2.12),

$$P(A \,|\, \mathbf{n}) \;=\; P_{\pi^{(n)}}(A) \;=\; \frac{\sum_{i\in j_1,\ldots,j_l} n_i + s^{(0)} \cdot \sum_{i\in j_1,\ldots,j_l} t_i^{(0)}}{n + s^{(0)}} \;=\; 0\,,$$

unbeachtet der Anzahl der Beobachtungen n, also auch, wenn n klein ist und es durchaus plausibel scheint, dass eine Beobachtung A im nächsten Zug noch erfolgen kann.

2.6.2 Parametrische Inferenz

Mit parametrischer Inferenz wird die Inferenz über die Parameter $\theta_1,\ldots,\theta_k$ bezeichnet. Dabei können Fragestellungen über $\boldsymbol{\theta}$ selbst, wie z.B. die Wahrscheinlichkeit $P(B)$, dass für ein gewisses $\Theta^* \subset \Theta$ das Ereignis $B = \{\boldsymbol{\theta} \in \Theta^*\}$ eintritt, und Wahrscheinlichkeitsaussagen über einen von $\theta_1,\ldots,\theta_k$ abgeleiteten Parameter $\delta(\boldsymbol{\theta}) \in \mathbb{R}$ unterschieden werden. Im letzteren Fall können untere und obere Erwartungswerte $\underline{\mathbb{E}}[\delta(\boldsymbol{\theta})]$ und $\overline{\mathbb{E}}[\delta(\boldsymbol{\theta})]$ berechnet werden. Umfassender ist die Berechnung der unteren und oberen Verteilungsfunktion $\underline{\mathrm{F}}(\delta(\boldsymbol{\theta}))$ und $\overline{\mathrm{F}}(\delta(\boldsymbol{\theta}))$, mittels welchen das Wahrscheinlichkeitsgewicht beliebiger Intervalle von $\delta(\boldsymbol{\theta})$ auf $\mathbb{R}$ berechnet werden kann. Mit Theorem 3 aus [Bernard 2005] sind die Werte von $\boldsymbol{t}$, die $\mathbb{E}(\delta(\boldsymbol{\theta}))$ und $\mathrm{F}(\delta(\boldsymbol{\theta}))$ minimieren bzw. maximieren, für einen in $\boldsymbol{\theta}$ linearen Parameter δ einfach erhältlich. Nichtlineare Funktionen δ können, mit vertretbarer Fehlerrate, linear approximiert werden.

Mit (2.4) und (2.6) ist der posteriori-Erwartungswert für θ_j für ein bestimmtes $\pi^{(n)} \in \mathcal{M}^{(n)}$

$$\mathbb{E}[\theta_j \,|\, \mathbf{n}] = t_j^{(n)} = \frac{n_j + s \cdot t_j^{(0)}}{n + s}$$

und daher gilt, bei Grenzübergang $t_j^{(n)} \to \min$ bzw. $t_j^{(n)} \to \max$

$$\underline{\mathbb{E}}[\theta_j \,|\, \mathbf{n}] = \frac{n_j + s \cdot \min\limits_{t_j^{(0)} \in T^{(0)}} t_j^{(0)}}{n + s} \quad \text{und} \quad \overline{\mathbb{E}}[\theta_j \,|\, \mathbf{n}] = \frac{n_j + s \cdot \max\limits_{t_j^{(0)} \in T^{(0)}} t_j^{(0)}}{n + s}\,.$$

Hier gelten natürlich die gleichen ceteris paribus-Aussagen wie in Kapitel 2.6.1.

Ebenso lassen sich – nach deutlich komplizierteren Berechnungen – die obere und untere Varianz $\underline{\mathbb{V}}(\theta_j \,|\, \mathbf{n})$ und $\overline{\mathbb{V}}(\theta_j \,|\, \mathbf{n})$ explizit angeben. Sie können in [Walley 1996, S. 17] gefunden werden.

Für ein $B = \{\boldsymbol{\theta} \in \Theta^*\}$ mit $\Theta^* \subset \Theta$ lässt sich die Wahrscheinlichkeit unter Verwendung eines bestimmten $\pi^{(n)}_{(s^{(n)}),\boldsymbol{t}^{(n)})} \in \mathcal{M}^{(n)}$ (mit der für dieses $\pi^{(n)}$ charakteristischen

Konfiguration $\boldsymbol{t}^{(n)}$ im unteren Index) folgendermaßen berechnen:

$$P(B \,|\, \mathbf{n}) = \int_{\boldsymbol{\theta} \in \Theta^*} \pi^{(n)}_{(s^{(n)}), \boldsymbol{t}^{(n)})}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{2.15}$$

$\underline{P}(B \,|\, \mathbf{n})$ und $\overline{P}(B \,|\, \mathbf{n})$ ergeben sich dann als Grenzen bei der Variation von $\boldsymbol{t}^{(n)}$.

Da dieses Integral über ein mehrdimensionales Intervall gebildet werden muss, kann es schwierig zu berechnen sein. Sind in B jedoch nur Bedingungen für ein θ_j enthalten, kann die Berechnung vereinfacht werden: Da die Inferenz im IDM invariant gegenüber der Kategorieaufteilung ist, kann die Anzahl k der Kategorien auf zwei reduziert werden, nämlich ‚j' und ‚nicht-j'. Eine Dirichlet-Verteilung mit zwei Kategorien entspricht aber einer Beta-Verteilung, so dass gilt

$$\begin{aligned} \theta_j \;\sim\; \text{Beta}\left(s^{(0)} \cdot t_j^{(0)} + n_j \,,\; s^{(0)} - s^{(0)} \cdot t_j^{(0)} + n - n_j\right), \quad \text{also} \\ p(\theta_j \,|\, \mathbf{n}) \;\propto\; \theta_j^{s^{(0)} \cdot t_j^{(0)} + n_j - 1} (1-\theta_j)^{s^{(0)} - s^{(0)} \cdot t_j^{(0)} + n - n_j - 1}. \end{aligned}$$

Es muss also nur über $t_j^{(0)}$ minimiert bzw. maximiert werden, um $\underline{P}(B \,|\, \mathbf{n})$ und $\overline{P}(B \,|\, \mathbf{n})$ zu erhalten.

Hypothesentests über $\boldsymbol{\theta}$ können durchgeführt werden, indem die obere und untere Wahrscheinlichkeit für $B = \{\boldsymbol{\theta} \in H_0\}$ berechnet wird. Sind nur Bedingungen für ein θ_j in H_0, vereinfacht sich die Berechnung in obiger Weise. Soll z.B. $H_0 : \theta_j \geq \theta_0 = \frac{1}{2}$ gegen $H_1 : \theta_j < \theta_0 = \frac{1}{2}$ getestet werden, so ist die posteriori-Wahrscheinlichkeit von θ_j wieder betaverteilt, und es gilt

$$\begin{aligned} \underline{P}(H_0 \,|\, \mathbf{n}) &= \int_{\frac{1}{2}}^{1} c_l \cdot \theta_j^{s^{(0)} \cdot t_j^{(0)} + n_j - 1} (1-\theta_j)^{s^{(0)} - s^{(0)} \cdot t_j^{(0)} + n - n_j - 1}\, d\theta_j \quad \text{mit } t_j^{(0)} \to \min \\ \overline{P}(H_0 \,|\, \mathbf{n}) &= \int_{\frac{1}{2}}^{1} c_u \cdot \theta_j^{s^{(0)} \cdot t_j^{(0)} + n_j - 1} (1-\theta_j)^{s^{(0)} - s^{(0)} \cdot t_j^{(0)} + n - n_j - 1}\, d\theta_j \quad \text{mit } t_j^{(0)} \to \max \end{aligned}$$

mit den Normierungskonstanten c_l und c_u, die sich aus den folgenden Gleichungen ergeben:

$$\begin{aligned} \int_0^1 c_l \cdot \theta_j^{s^{(0)} \cdot t_j^{(0)} + n_j - 1} (1-\theta_j)^{s^{(0)} - s^{(0)} \cdot t_j^{(0)} + n - n_j - 1}\, d\theta_j &\stackrel{!}{=} 1 \quad \text{für } t_j^{(0)} \to \min \\ \int_0^1 c_u \cdot \theta_j^{s^{(0)} \cdot t_j^{(0)} + n_j - 1} (1-\theta_j)^{s^{(0)} - s^{(0)} \cdot t_j^{(0)} + n - n_j - 1}\, d\theta_j &\stackrel{!}{=} 1 \quad \text{für } t_j^{(0)} \to \max \end{aligned}$$

H_0 kann kann abgelehnt werden, wenn $\overline{P}(H_0 \,|\, \mathbf{n})$ kleiner als eine zuvor festgelegte Schranke ist. Das Intervall $\left[\underline{P}(H_0 \,|\, \mathbf{n}) \,,\, \overline{P}(H_0 \,|\, \mathbf{n})\right]$ vergrößert sich, wenn $s^{(0)}$ vergrößert wird, und enthält für $s^{(0)} \geq 1$ die üblichen frequentistischen p-Werte unter binomialer oder negativ-binomialer Verteilungsannahme.

Kredibilitätsregionen für $\boldsymbol{\theta}$ zu einem Niveau γ können prinzipiell ebenfalls über die Gleichung (2.15) berechnet werden. Sie sind größer als ihre Bayes-Analoga, da sie sowohl die Unsicherheit aufgrund der endlichen Zahl von Beobachtungen als auch die Unsicherheit über die Präzision der Priori-Informationen wiederspiegeln.

Für die Form einer solchen Region gibt es im Allgemeinen zwei Möglichkeiten:

- Die Kredibilitätsregion kann als ein mehrdimensionales Intervall mit Wahrscheinlichkeitsgewicht von mindestens γ gebildet werden, bei dem die Grenzen für jede Dimension Hyperebenen sind, die zur allen anderen Koordinatenachsen parallel verlaufen. Bei zweiseitigen Kredibilitätsregionen muss die Lage des Intervalls so gewählt werden, dass die mehrdimensionale ‚Länge' minimal wird. Bei drei Kategorien ergäbe sich also ein ‚Kredibilitätswürfel', dessen Kanten parallel zu den Koordinatenachsen sind und dessen Lage so bestimmt ist, dass das Volumen des Würfels minimal ist.

- Eine Kredibilitätsregion kann auch als highest posterior density - Region (HPD-Region) ermittelt werden. Dafür wird die Schranke ξ gesucht, für welche die untere posteriori-Wahrscheinlichkeit aller $\boldsymbol{\theta} \in \Theta$, deren Dichte größer als ξ ist, gerade γ beträgt.
$$\underline{P}\big(\{\boldsymbol{\theta} : \pi^{(n)}_{(s^{(n)},\boldsymbol{t}^{(n)})}(\boldsymbol{\theta}) > \xi\}\big) = \gamma \tag{2.16}$$
Diese Region ist dann entsprechend der ‚Höhenlinien' der Dichte im mehrdimensionalen Raum amorph und kann auch in mehrere unverbundene Teilregionen zerfallen, wenn die posteriori-Dichte nicht unimodal ist.

Bei Kredibilitätsregionen für ein (eindimensionales) θ_j fallen diese beiden Methoden zusammen. Wieder lässt sich die Dirichlet-Verteilung auf eine Beta-Verteilung zurückführen, indem die Kategorieaufteilung entsprechend vereinfacht wird.

Ein einseitiges Intervall $I = [0, \theta_j^*]$ mit $\underline{P}(I \mid \mathbf{n}) = \gamma$ für ein θ_j kann mit Hilfe der unteren Verteilungsfunktion $\underline{\mathrm{F}}$ einer Beta-Verteilung mit den Paramtern $s^{(0)} \cdot t_j^{(0)} + n_j$ und $s^{(0)} - s^{(0)} \cdot t_j^{(0)} + n - n_j)$ berechnet werden. Diese ergibt sich für $t_j^{(0)} \to \max$, und es gilt $\theta_j^* = \underline{\mathrm{F}}^{-1}(\gamma)$. In Abbildung 2.4 ist die Konstruktion eines solchen Kredibilitätsintervalls dargestellt. Ebenso kann ein θ_{j*} für das Kredibilitätsintervall $I = [\theta_{j*}, 1]$ mit Hilfe der oberen Verteilungsfunktion $\overline{\mathrm{F}}$ der gleichen Beta-Verteilung berechnet werden, welche sich für $t_j^{(0)} \to \min$ ergibt. θ_{j*} ist dann die Lösung von $1 - \gamma = \overline{\mathrm{F}}(\theta_{j*})$, weil damit die obere Wahrscheinlichkeit des Komplementärintervalls $[0,\ \theta_{j*}]$ $1 - \gamma$ beträgt.

Zweiseitige Intervalle sind ebenso möglich. Ein Intervall mit der Glaubwürdigkeit von mindestens γ kann berechnet werden, indem die einseitigen Intervallgrenzen θ_j^* und θ_{j*} jeweils mit $\frac{1+\gamma}{2}$ statt γ ermittelt werden. Soll das Intervall den Sicherheitsgrad genau erfüllen, müssen die Grenzen mittels (2.16) berechnet werden.

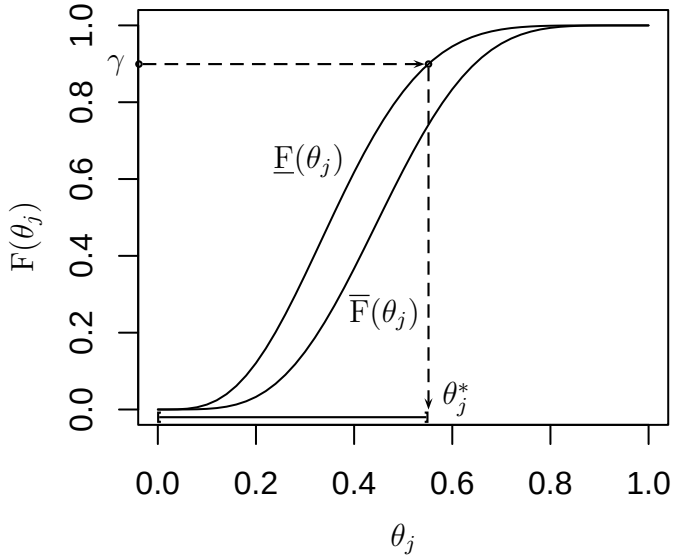

Abbildung 2.4: Konstruktion eines einseitigen Kredibilitätsintervalls für θ_j, mit $s = 1$, $n = 10$ und $n_j = 4$

Walley empfiehlt in [Walley 1996] generell, die unteren und oberen Verteilungsfunktionen für jede Komponente θ_j getrennt zu plotten. Aus diesen lassen sich dann, wie in Abbildung 2.4 beispielhaft illustriert, einseitige Intervalle direkt ablesen.

3 Das Stichproben-Modell von Quaeghebeur und de Cooman

Das Stichproben-Modell von Quaeghebeur und de Cooman [Quaeghebeur und de Cooman 2005] stellt eine direkte Verallgemeinerung des IDM dar. Mit ihm können nicht nur multinomialverteilte, sondern auch anders verteilte Stichproben in gleicher Weise analysiert werden, deren Verteilung zu einer Exponentialfamilie gehört. Ebenso wie beim IDM liegt das in Kapitel 2.2 beschriebene bayesianische Aufdatierungsschema mit konjugierten Priori-Verteilungen zugrunde: Eine Menge von Priori-Verteilungen wird mittels der präzisen Likelihood der Beobachtungen aufdatiert, die so entstandene Menge der Posteriori-Verteilungen bildet die Basis aller Schlüsse aus der Stichprobe x.

Die Menge der Verteilungen wird dabei wieder über eine Menge von Parametern $\mathcal{Y}$ definiert, die linear aufdatiert werden können. Die Parameter $y \in \mathcal{Y}$ sind jedoch nicht die ‚klassischen' Parameter der konjugierten Verteilung, sondern entsprechen der „durchschnittlichen" suffizienten Statistik der Stichprobenverteilung [Quaeghebeur und de Cooman 2005, Kap. 2.2]. In vielen Fällen hat dieses y jedoch eine einfache Interpretation, und oft stellt er sich als Erwartungswert des gesuchten Parameters der Stichprobenverteilung heraus.

Priori- und Posteriori-Verteilung hängen zusätzlich noch von einem Parameter n ab, der die Rolle von s aus dem IDM übernimmt und somit steuert, wie stark die Priori-Annahmen in die Posteriori-Verteilung eingehen. Bei vielen Anwendungen gilt, dass n gleich der Zahl der Beobachtungen ist, die nötig sind, um die Differenz zwischen dem unteren und dem oberen Priori-Erwartungswert a posteriori auf die Hälfte zu reduzieren.

Posteriori-Inferenz über den Stichprobenparameter kann, analog zu (2.10), durch untere und obere Wahrscheinlichkeiten dargestellt werden, die sich als Infimum bzw. Supremum über $y \in \mathcal{Y}$ ergeben.

Die Beschreibung dieses Modells in diesem Kapitel lehnt sich eng an [Quaeghebeur und de Cooman 2005] an.

3.1 Exponentialfamilien

Ausgangspunkt der Stichprobenanalyse ist die Verteilung der Stichprobe, die zum Typ einer Exponentialfamilie gehören muss. Solche Verteilungsfamilien besitzen bestimmte

Regularitätseigenschaften; ob eine Verteilungsfamilie einer Exponentialfamilie entspricht, wird über eine bestimmte Faktorisierung ihrer Dichte gezeigt.

In der Literatur gibt es verschiedene solche Faktorisierungen. Quaeghebeur und de Cooman beziehen sich auf die Notation von [Bernardo und Smith 1993], die verschiedene abgestufte Kriterien verwenden. Die in [Quaeghebeur und de Cooman 2005] verwendete Faktorisierung ist das Kriterium für das Vorliegen einer „regular, linear canonical exponential family“ gemäß der Systematik von Bernardo und Smith [Bernardo und Smith 1993, S. 202 und S. 272f]:

$$f(x \mid \psi) = \mathbf{a}(x) \exp\left\{\langle \psi, \tau(x) \rangle - \mathbf{b}(\psi)\right\} \tag{3.1}$$

Dabei ist $x \in \mathcal{X}$ die Beobachtung einer Zufallsvariable X mit der Dichte $f(x \mid \psi)$. Diese Stichprobe sei zunächst als einelementig angenommen, eine Erweiterung auf Stichproben vom Umfang m folgt dann am Ende des nächsten Kapitels. $\psi \in \Psi$ ist der (eventuell mehrdimensionale) natürliche Parameter der Verteilung, der sich als eine Funktion der ‚klassischen‘ Parameter ergibt. Hat der natürliche Parameter die gleiche Dimension q wie der klassische, so heißt die Verteilung auch „strikt q-parametrisch“ (siehe [Rüger 1999, S. 19]). Test- und Schätzprobleme, bei denen eine strikt q-parametrische Verteilungsannahme nicht vorliegt, sind deutlich schwerer zu lösen. $\tau(x) \in \mathcal{T}$ ist eine Funktion der Beobachtung x, die sich durch die geforderte Faktorisierung im Exponenten ergibt, und ist eine suffiziente Statistik von X. Diese wird im Folgenden als ‚die‘ suffiziente Statistik bezeichnet. Sie zeigt, in welcher Weise der beobachtete Wert der Zufallsvariable X in die Berechnungen für das Inferenzproblem eingeht. Auch wenn X mehrdimensional wird (z.B. durch mehrere unabhängige und identisch verteilte Beobachtungen), behält $\tau(x)$ dieselbe Dimension. Dann zeigt $\tau(x)$, wie die Stichprobe zusammengefasst werden kann, ohne dass die für das Test- oder Schätzproblem notwendigen Informationen verloren gehen. Der Ausdruck $\langle \psi, \tau(x) \rangle$ bezeichnet im Falle von mehrdimensionalen Größen das Skalarprodukt von ψ und $\tau(x)$, daher besitzen ψ und $\tau(x)$ die gleiche Dimension. Sind ψ und $\tau(x)$ eindimensional, handelt es sich um ein normales Produkt. Die Funktionen $\mathbf{a}(x) : \mathcal{X} \to \mathbb{R}^+$ und $\mathbf{b}(\psi) : \Psi \to \mathbb{R}$ sind für die verschiedenen Verteilungsannahmen, die einer Exponentialfamilie entsprechen, charakteristisch.

Diese Faktorisierung sei am Beispiel der Stichprobenverteilung einer Beobachtung $x \in \mathcal{X} = \mathbb{R}$ gemäß $\mathrm{N}(0, \sigma^2)$, mit $\sigma \in \mathbb{R}^+$, veranschaulicht:

$$\begin{aligned} f(x \mid \psi) \;=\; f(x \mid \sigma^2) \;&=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\} \\ &=\; \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}x^2 - \ln(\sigma)\right\} \\ &=\; \frac{1}{\sqrt{2\pi}} \exp\left\{\langle \psi, \tau(x) \rangle + \frac{1}{2}\ln(-2\psi)\right\} \end{aligned}$$

Hier gilt also

$$\psi = -\frac{1}{2\sigma^2}\,, \qquad \tau(x) = x^2\,, \qquad \mathbf{a}(x) = \frac{1}{\sqrt{2\pi}}\,, \qquad \mathbf{b}(\psi) = -\frac{1}{2}\ln(-2\psi)\,,$$

so dass $\Psi = \mathbb{R}^-$ und $\mathcal{T} = \mathbb{R}_0^+$.

Eine vorteilhafte Eigenschaft der in dieser Art definierten Exponentialfamilien ist, dass der Erwartungswert der suffizienten Statistik gegeben ψ der Ableitung (bzw. im Falle eines mehrdimensionalen ψ dem Gradienten) von $\mathbf{b}(\psi)$ nach ψ entspricht:

$$\mathbb{E}\left[\tau(X) \,|\, \Psi = \psi\right] = \nabla \mathbf{b}(\psi)$$

Diese Eigenschaft wird sich später bei der Interpretation der Paramter der konjugierten Verteilung als hilfreich erweisen.

Im Beispiel gilt

$$\mathbb{E}[X^2 \,|\, \Psi = \psi] = \nabla \mathbf{b}(\psi) = -\frac{1}{2\psi} = \sigma^2\,, \tag{3.2}$$

der Erwartungswert von $\tau(x)$ gegeben ψ erweist sich also hier gerade als der gesuchte Parameter der Stichprobenverteilung.

3.2 Bayes-Lernen in Exponentialfamilien

Wird eine Exponentialfamilien-Dichte gemäß (3.1) als Likelihood interpretiert, in der Notation

$$\begin{aligned} L_x : \Psi &\longrightarrow \mathbb{R}^+ \\ \psi &\longrightarrow f(x \,|\, \psi)\,, \end{aligned}$$

dann kann die Dichte der zugehörigen konjugierten Verteilung, die eine besonders einfache bayesianische Analyse möglich macht (siehe Kapitel 2.3), folgendermaßen angegeben werden:

$$f_C(\psi \,|\, n, y) = \mathbf{c}(n, y) \exp\left\{ n \left[\langle \psi, y \rangle - \mathbf{b}(\psi) \right] \right\} \tag{3.3}$$

Es handelt sich um eine Dichte über den (potentiell mehrdimensionalen) Raum Ψ der natürlichen Parameter der Dichte (3.1). In dieser konjugierten Verteilung gibt es zwei Parameter: $n \in \mathbb{R}^+$ kann als eine Art ‚Zählwert' angesehen werden. Er kann a posteriori die Anzahl der vorhandenen Datenpunkte darstellen, die in die Likelihood eingegangen sind, aber auch sogenannte ‚pseudocounts' enthalten; diese können die Stärke der Priori-Informationen symbolisieren. Eine Veranschaulichung dieser Interpretation kann in Kapitel 3.3 gefunden werden. Letztendlich entspricht n dem Parameter s aus dem IDM. Der zweite Parameter, y, nimmt die Rolle von $\tau(x)$ in (3.1) ein; folgerichtig interpretieren ihn Quaeghebeur und de Cooman als eine Art durchschnittliche suffiziente Statistik. y muss also in einem ähnlichen Raum wie $\tau(x)$ variieren; $\mathcal{Y}$ wird daher als die

konvexe Hülle von $\mathcal{T}$, $\mathrm{co}(\mathcal{T})$ definiert, jedoch ohne deren Rand. $\mathbf{c}(n, y)$ ist ein Faktor zur Normierung der Dichte, ähnlich wie $\mathbf{a}(x)$. Er kann über Vergleiche des Exponenten mit Kernen bekannter Verteilungen ermittelt werden, wobei es dazu nötig sein kann, ψ zu transformieren. $\mathbf{b}(\psi)$ wird direkt aus der Dichte der Stichprobenverteilung (3.1) übernommen.

Eine konjugierte Priori-Verteilung (3.3) kann durch die (eindimensionale) Beobachtung x gemäß des Satzes von Bayes für Dichten (in Verallgemeinerung des Vorgehens in Kapitel 2.2) in einfacher Weise aufdatiert werden. Die resultierende Posteriori-Verteilung gehört wieder zur gleichen Verteilungsfamilie wie die Priori und ergibt sich dabei einfach durch die *lineare* Aufdatierung der Parameter n und y:

$$\begin{aligned} f_C(\psi \,|\, n, y, x) &= f_C(\psi \,|\, n, y) \cdot L_x(\psi) \\ &= \mathbf{c}(n, y) \exp\left\{ n \left[\langle \psi, y \rangle - \mathbf{b}(\psi) \right] \right\} \cdot \mathbf{a}(x) \exp\left\{ \langle \psi, \tau(x) \rangle - \mathbf{b}(\psi) \right\} \\ &\propto \exp\left\{ (n+1) \left[\left\langle \psi, \frac{ny + \tau(x)}{n+1} \right\rangle - \mathbf{b}(\psi) \right] \right\} \end{aligned} \tag{3.4}$$

Um von einer Priori-Verteilung in der Form (3.3) zur zugehörigen Posteriori-Verteilung zu gelangen , müssen also nur die Parameter n und y folgendermaßen aufdatiert werden:

$$n \to n + 1 \qquad\qquad y \to \frac{ny + \tau(x)}{n+1}$$

Im Beispiel kann die konjugierte Verteilung ermittelt werden, indem in einem ersten Schritt ψ zu λ, der sogenannten Präzision, transformiert wird:

$$\lambda = \frac{1}{\sigma^2} = -2\psi$$

Dann gilt für die konjugierte Verteilung, statt als Dichte über ψ als Dichte über λ notiert:

$$\begin{aligned} f_C^{\psi}(n, y) &= \frac{1}{2} f_C^{\lambda}(n, y) \\ &= \frac{1}{2} \mathbf{c}(n, y) \exp\left\{ n \left[-\frac{\lambda}{2} y + \frac{1}{2} \ln(\lambda) \right] \right\} \\ &= \frac{1}{2} \mathbf{c}(n, y) \lambda^{\frac{n}{2}} \exp\left\{ -\frac{ny}{2} \lambda \right\} \end{aligned}$$

Dieser Ausdruck ist proportional zu der Dichte einer Gamma-Verteilung über λ mit den Parametern $\alpha = \frac{n+2}{2}$ und $\beta = \frac{ny}{2}$. Damit kann $\mathbf{c}(n, y)$ über den Normierungsfaktor der Gamma-Verteilung berechnet werden:

$$\mathbf{c}(n, y) = 2 \frac{\beta^{\alpha}}{\Gamma(\alpha)} = 2 \frac{\left(\frac{ny}{2}\right)^{\frac{n+2}{2}}}{\Gamma\left(\frac{n+2}{2}\right)}$$

$\mathcal{Y}$, der Raum des Parameters y, entspricht in diesem Beispiel $\mathbb{R}^+$, der konvexen Hülle von $\mathcal{T}$, $\text{co}(\mathcal{T}) = \mathbb{R}_0^+$, jedoch ohne den Rand.

Eine weitere vorteilhafte Eigenschaft eines so definierten Modells ist die folgende: Bildet man den Erwartungswert bezüglich f_C von $\nabla\mathbf{b}(\psi)$ gegeben die Parameter n und y, so ist das Ergebnis y:

$$\mathbb{E}_C\big[\nabla\mathbf{b}(\psi) \,|\, n, y\big] = y \tag{3.5}$$

Zusammen mit der Gleichung (3.2) macht dies eine Interpretation von y möglich. Im Beispiel ergab sich $\sigma^2 = \nabla\mathbf{b}(\psi)$, daher ist y der Erwartungswert von σ^2 gegeben n und y, jedoch für beliebige Werte von n. Also kann y hier als Erwartungswert des gesuchten Parameters σ^2 interpretiert werden.

Quaeghebeur und de Cooman geben auch eine „zugehörige prädiktive Verteilung“ an, die zur prädiktiven Inferenz, also z.B. zur Vorhersage zukünftiger Beobachtungen, herangezogen werden kann. Sie erhält man, indem ψ aus der Posteriori-Verteilung ‚herausintegriert‘ wird, und kann hier durch geschickte Ausnutzung der Normierungsfaktoren der bekannten Dichten berechnet werden:

$$\begin{aligned} f_P(x \,|\, n, y) &= \int f_C(\psi \,|\, n, y, x)\, d\psi \\ &= \int f_C(\psi \,|\, n, y) L_x(\psi)\, d\psi \\ &= \frac{\mathbf{c}(n, y) \cdot \mathbf{a}(x)}{\mathbf{c}\left(n+1, \frac{ny+\tau(x)}{n+1}\right)} \end{aligned}$$

Mit $\mathbb{E}_P$ wird dann der bezüglich dieser Dichte gebildete Erwartungswert bezeichnet.

Bei der Beobachtung einer i.i.d-Stichprobe der Größe m, also von $x_1, \ldots, x_m$ statt nur einer eindimensionalen Stichprobe x, müssen nur die folgenden Änderungen gemacht werden:

$$\begin{aligned} \tau(x) &\to \tau(x_1, \ldots, x_m) = \sum_{j=1}^{m} \tau(x_j) \\ \mathbf{a}(x) &\to \mathbf{a}(x_1, \ldots, x_m) = \prod_{j=1}^{m} \mathbf{a}(x_j) \\ \mathbf{b}(\psi) &\to m \cdot \mathbf{b}(\psi) \end{aligned}$$

Eventuell muss noch $\mathbf{a}(x_1, \ldots, x_m)$ mit einem Faktor multipliziert werden, um beschränktes Wissen über die Reihenfolge der Beobachtungen mit einzubeziehen. Der Aufdatierungsschritt für die Parameter n und y ändert sich dann zu

$$n \to n + m \qquad\qquad y \to \frac{ny + \sum_{j=1}^{m} \tau(x_j)}{n + m}\,. \tag{3.6}$$

Mit diesen Angaben ist die Vorbereitung für die Implementierung eines Intervallwahrscheinlichkeitsmodells für Exponentialfamilien abgeschlossen; Quaeghebeur und de Cooman geben zusätzlich noch eine Tabelle mit den dazu nötigen Angaben für die gebräuchlichsten Stichprobenfamilien an. Sie enthält die Angabe der zugehörigen konjugierten Verteilung, den Stichprobenraum $\mathcal{X}$, den natürlichen Parameter ψ, $\tau(x)$ und die Funktionen $\mathbf{a}(x)$, $\mathbf{b}(\psi)$ und $\mathbf{c}(n, y)$, darüber hinaus aber auch den für die Interpretation wichtigen Gradienten von $\mathbf{b}$; außerdem noch eine in manchen Fällen nötige Beschränkung von $\mathcal{Y}$, auf die im nächsten Abschnitt eingegangen werden wird. Die Einträge für $\nabla\mathbf{b}(\psi)$ zeigen, dass nicht nur in dem hier angeführten Beispiel, sondern auch für viele andere Stichprobenverteilungen gilt, dass y dem Erwartungswert des gesuchten Parameters entspricht.

Anhand dieser Tabelle wird auch ersichtlich, dass das Modell von Quaeghebeur und de Cooman tatächlich eine Verallgemeinerung des IDM darstellt: Zur Multinomialverteilung wird als konjugierte Verteilung die Dirchletverteilung angegeben; im Text wird dieser Zusammenhang gesondert gewürdigt.

3.3 Die ‚Impräzisierung‘

Die bisher beschriebenen Definitionen und Aussagen entsprechen einer klassischen bayesianischen Modellierung mit konjugierter Verteilung mit der Besonderheit, dass die Modellklassen so gewählt sind, dass die Parameter n und y jeweils linear aufdatiert werden. Diese Linearität ist jedoch der Schlüssel zur Generierung eines einfach zu handhabenden Intervallwahrscheinlichkeitsmodells: Werden a priori Unter- und Obergrenzen für die Parameter festgelegt, können aufgrund der linearen Aufdatierung die a posteriori Unter- und Obergrenzen direkt ermittelt werden, so dass sich das Minimierungs- und Maximierungsproblem genau wie in Kapitel 2 auf die Menge der Priori-Parameter zurückführen lässt, falls eine lineare Funktion von y oder y selbst die Zielgröße ist.

Es handelt sich also um ein Intervallwahrscheinlichkeitsmodell, das über eine Menge von Verteilungen erzeugt wird. Die Menge $\mathcal{M}$ der Verteilungen wird dabei genau wie im IDM über eine Menge von Parametern definiert, die durch das Variieren der Parameter der konjugierten Verteilung, also n und y, entsteht. Dabei gibt es mehrere Möglichkeiten:

- Bei einem Ansatz im Kontext von robusten Bayes-Methoden wird $n \cdot y$ um einen bestimmten Wert von $n \cdot y$, der als Ausgangspunkt gewählt wird, variiert (siehe [Boratyńska 1997]).
- Quaeghebeur und de Cooman lassen stattdessen y in einer Menge $\mathcal{Y}$ variieren, die durch die (im Falle eines mehrdimensionalen Parameters y elementweisen) Unter- und Obergrenzen $\underline{y}$ und $\overline{y}$ definiert wird. n muss fest gewählt werden.
- Zusätzlich zu y könnte n variiert werden, um auch ‚prior-data conflict‘-Situationen optimal modellieren zu können; wie im IDM ist dies nämlich nur mit einem variie-

renden n möglich[1]. Quaeghebeur und de Cooman nennen diese Möglichkeit, jedoch ohne Bezug zu ‚prior-data conflict'-Situationen, und betrachten diese Möglichkeit eines variierenden n in ihrem Artikel nicht weiter. Das Modell würde dadurch deutlich komplexer werden, hauptsächlich deshalb, weil der Aufdatierungsschritt von y dabei nichtlinear wird und sich daher Minimierungs- und Maximierungsprobleme nicht direkt auf die Menge der Priori-Parameter zurückführen lassen (wie in Kapitel 2.6 beschrieben).

Die resultierende Menge von Verteilungen heißt ‚credal set' und ist bei Quaeghebeur und de Cooman die Vereinigung aller Konvexkombinationen von Verteilungen, die durch die Variation von $y \in \mathcal{Y}$ entstehen; die generierende Menge der Parameter $\mathcal{Y}$ muss im mehrdimensionalen Fall jedoch selbst nicht konvex sein.

Da die Extrempunkte der ‚credal sets' durch die Menge der ‚reinen', parametrischen Verteilungen definiert werden, reicht es, sich auf die Menge dieser ‚reinen' Verteilungen zu beschränken, wenn bei der Analyse nur Extremwerte eine Rolle spielen. Extremwerte sind dabei all die Werte, die durch eine Minimierung oder Maximierung über die Menge der Verteilungen im ‚credal set' erhalten werden, also z.B. der untere oder obere Posteriori-Erwartungswert.

Wie beim IDM ist es nötig, durch einen oberen Index $^{(0)}$ die Menge der Priori-Parameter $\mathcal{Y}^{(0)}$ und deren Elemente $y^{(0)}$ zu kennzeichnen, ebenso die (elementweisen) Unter- und Obergrenzen $\underline{y}^{(0)}$ und $\overline{y}^{(0)}$ von $\mathcal{Y}^{(0)}$. Der feste Anfangswert von n wird analog mit $n^{(0)}$ bezeichnet. Mit (3.6) kann dann die Menge der Parameter, die sich nach der Aufdatierung auf der Basis von k Beobachtungen a posteriori ergibt, folgendermaßen notiert werden:

$$\mathcal{Y}^{(k)} = \left\{ \frac{n^{(0)} y^{(0)} + \sum_{j=1}^{m} \tau(x_j)}{n^{(0)} + k} \;\middle|\; y^{(0)} \in \mathcal{Y}^{(0)} \right\} \subset \mathcal{Y} \tag{3.7}$$

Wenn die Aufdatierungsregel für y etwas umformuliert wird, ist ersichtlich, dass $\mathcal{Y}^{(k)}$ einfach eine verschobene und skalierte Version der Menge $\mathcal{Y}^{(0)}$ ist:

$$\mathcal{Y}^{(0)} \;\rightarrow\; \frac{n^{(0)}}{n^{(0)} + k} \cdot \mathcal{Y}^{(0)} + \frac{k}{n^{(0)} + k} \cdot \frac{1}{k} \sum_{j=1}^{m} \tau(x_j)$$

$\mathcal{Y}^{(k)}$ kann also als konvexe Mischung von $\mathcal{Y}^{(0)} \subset \mathcal{Y}$ und $\frac{1}{k} \sum_{j=1}^{m} \tau(x_j) \in \mathrm{co}(\mathcal{T})$ angesehen werden, jeweils mit einem Koeffizienten, der das entsprechende Gewicht wiederspiegelt: Während $\frac{1}{k} \sum_{j=1}^{m} \tau(x_j)$ mit einem Gewicht entsprechend der Größe k der Stichprobe eingeht, hat $\mathcal{Y}^{(0)}$ einen Einfluss gemäß der ‚Priori-Stärke' $n^{(0)}$.

Hieran wird deutlich, was es mit dem von Quaeghebeur und de Cooman verwendeten Begriff der ‚pseudocounts' auf sich hat: $n^{(0)}$ spielt für die Priori-Verteilung letztlich die

[1] Siehe dazu die Erläuterungen bei Gleichung (3.8).

gleiche Rolle wie k für die Stichprobe. Man kann daher $n^{(0)}$ als ‚Stichprobenumfangs-Äquivalent' (im Folgenden auch kürzer ‚Stichprobe-Äquivalent' genannt) des Vorwissens ansehen und sich somit bei der konkreten Anwendung des Modells $n^{(0)}$ als die Größe einer imaginären Stichprobe vorstellen, die zum Ergebnis $\mathcal{Y}^{(0)}$ geführt hat. $n^{(0)}$ entspricht dann der Größe einer Stichprobe, zu deren Ergebnissen man ein ähnliches Vertrauen hätte wie in das vorliegende Vorwissen, dass durch $\mathcal{Y}^{(0)}$ verkörpert wird. Bei der Anwendung dieses Modells in Kapitel 4.4 wird auf diese Interpretation zurückgegriffen werden; natürlich ist die bei Gleichung (2.11) erläuterte Interpretation als Anzahl ‚verdeckter' Beobachtungen ebenso möglich.

Posteriori-Inferenz basiert dann auf der mit (3.7) erhaltenen Menge von Parametern. Diese bestimmt dann die Menge der aufdatierten konjugierten Verteilungen $f_C(n^{(k)}, y^{(k)})$, die die Basis für parametrische Inferenz ist, und ebenso die Menge der aufdatierten zugehörigen prädiktiven Verteilung $f_P(n^{(k)}, y^{(k)})$, mit deren Hilfe prädiktive Inferenzaussagen gemacht werden können.

Quaeghebeur und de Cooman geben für diese Zwecke gemäß der Methodik von Walley die unteren und oberen a posteriori-Erwartungswerte an. Diese ergeben sich als Infimum und Supremum der Erwartungswerte, die bezüglich der genannten Mengen von Verteilungen gebildet werden:

$$\underline{\mathbb{E}}_C[\,\cdot\,] = \inf_{y^{(k)} \in \mathcal{Y}^{(k)}} \mathbb{E}_{f_C(n^{(k)}, y^{(k)})}[\,\cdot\,] \qquad \overline{\mathbb{E}}_C[\,\cdot\,] = \sup_{y^{(k)} \in \mathcal{Y}^{(k)}} \mathbb{E}_{f_C(n^{(k)}, y^{(k)})}[\,\cdot\,]$$

$$\underline{\mathbb{E}}_P[\,\cdot\,] = \inf_{y^{(k)} \in \mathcal{Y}^{(k)}} \mathbb{E}_{f_P(n^{(k)}, y^{(k)})}[\,\cdot\,] \qquad \overline{\mathbb{E}}_P[\,\cdot\,] = \sup_{y^{(k)} \in \mathcal{Y}^{(k)}} \mathbb{E}_{f_P(n^{(k)}, y^{(k)})}[\,\cdot\,]$$

Mit den Gleichungen (3.2) und (3.5) können unterer und oberer Posteriori-Erwartungswert von $\nabla\mathbf{b}(\psi)$ also direkt als elementweises Infimum oder Supremum von $\mathcal{Y}^{(k)}$ ermittelt werden; in unserem Beispiel gilt also

$$\underline{\mathbb{E}}_C[\sigma^2] = \underline{y}^{(k)} = \frac{n^{(0)}\underline{y}^{(0)} + \sum_{j=1}^{m} \tau(x_j)}{n^{(0)} + k}$$

und

$$\overline{\mathbb{E}}_C[\sigma^2] = \overline{y}^{(k)} = \frac{n^{(0)}\overline{y}^{(0)} + \sum_{j=1}^{m} \tau(x_j)}{n^{(0)} + k}\,.$$

Der Posteriori-Erwartungswert für σ^2 hängt hier also unmittelbar von der Wahl der Unter- und Obergrenze des Priori-Erwartungswertes für σ^2 ab. Aufgrund der gleichen analytischen Form wie im Falle des IDM (siehe Kapitel 2.6.1, Gleichung (2.13)) gelten die gleichen Überlegungen für das Verhalten des Modells:

- Je ungenauer das Vorwissen und somit je weiter auseinander $\underline{y}^{(0)}$ und $\overline{y}^{(0)}$ liegen, desto breiter wird das Posteriori-Erwartungswertintervall für $\overline{\nabla}\mathbf{b}(\psi)$, desto ungenauer wird also das Ergebnis des Inferenzprozesses.

- Je weniger Vertrauen in das Vorwissen gesetzt wird und somit je kleiner die Wahl von $n^{(0)}$, desto breiter wird das Posteriori-Erwartungswertintervall für $\nabla \mathbf{b}(\psi)$.

- Da $n^{(0)}$ (so wie im IDM $s^{(0)}$) fest gewählt werden muss, ist das Verhalten des Modells im Falle eines ‚prior-data conflict' nicht optimal, da sich dann das Erwartungswertintervall zwar in die richtige Richtung ‚bewegt', aber eben nicht breiter werden kann, weil die Breite des Intervalls unabhängig von $\tau(x)$ ist:

$$\overline{\mathbb{E}}_C[\nabla \mathbf{b}(\psi)] - \underline{\mathbb{E}}_C[\nabla \mathbf{b}(\psi)] = \overline{y}^{(k)} - \underline{y}^{(k)} = \frac{n^{(0)} \left(\overline{y}^{(0)} - \underline{y}^{(0)}\right)}{n^{(0)} + k} \tag{3.8}$$

  Hieran wird jedoch die anfangs erwähnte Interpretation von $n^{(0)}$ deutlich: für $n^{(0)} = k$ halbiert sich durch den Aufdatierungsprozess die Spanne des Erwartungswerts für $\nabla \mathbf{b}(\psi)$.

Wie genau sich die Möglichkeit der Variation von $n^{(0)}$ auf den Inferenzprozess und die Berechenbarkeit von Inferenzaussagen auswirken würde, muss der Gegenstand zukünftiger Untersuchungen bleiben; einen Anhaltspunkt bietet aber die schon in Kapitel 2.6.1 erwähnte Erweiterung von Walley im Falle des IDM mit zwei Kategorien (siehe [Walley 1991, Kap. 5.4]).

Die Wahl von $\mathcal{Y}^{(0)}$ soll das Vorwissen über die Parameter wiederspiegeln. Wenn sehr wenig oder gar kein substantielles Wissen über die Parameter vorhanden ist, muss $\mathcal{Y}^{(0)}$ möglichst groß gewählt werden, idealerweise als Menge aller möglichen Parameter, also $\mathcal{Y}^{(0)} = \mathcal{Y}$. In den meisten Fällen würde aber eine solche Wahl zur Folge haben, dass $\mathcal{Y}^{(0)}$ im Zuge der Aufdatierung unverändert bleibt, egal wie groß die Anzahl der eingegangenen Beobachtungen ist. Dieses unerwünschte Verhalten ist ein Teilaspekt eines Phänomens, das in der Literatur ‚dilation' genannt wird. (Siehe beispielsweise [Seidenfeld und Wasserman 1993]; ‚dilation' umfasst auch den Fall, dass durch eine Beobachtung nicht nur nichts dazugelernt werden kann, sondern die Aufdatierung mit jeder beliebigen Beobachtung zu aussagelosen Wahrscheinlichkeitsintervallen führt.)

Um auszuschließen, dass $\mathcal{Y}^{(0)}$ auch nach der Aufdatierung aussagelos bleibt, muss daher $\mathcal{Y}^{(0)}$ durch die Wahl von endlichen Unter- und Obergrenzen beschränkt werden. Im Spezialfall des IDM ist dies jedoch nicht nötig, da $\mathcal{Y}$ selbst schon beschränkt ist und somit mit $\mathcal{Y}^{(0)}$ gleichgesetzt werden kann. (Siehe dazu Gleichung (2.7) und Abbildung 2.2; $\mathcal{Y}^{(0)}$ entspricht $T^{(0)}$.)

Im Beispiel der zentrierten Normalverteilung ist unmittelbar ersichtlich, dass $\overline{y}^{(0)}$ endlich sein muss, damit die Obergrenze für den Posteriori-Erwartungswert endlich wird. (Die Untergrenze ist schon prinzipiell durch die Form des Raums $\mathcal{Y}^{(0)} = \mathbb{R}^+$ festgelegt.)

Abbildung 1 in [Quaeghebeur und de Cooman 2005] zeigt für die Normalverteilung mit unbekannten Parametern μ und σ^2 beispielhaft, welche Auswirkungen der

Aufdatierungsschritt bei Verteilungen mit mehr als einem Parameter hat. Aus der Darstellung wird deutlich, dass $\mathcal{Y}^{(0)}$ auch nicht konvex sein kann; Quaeghebeur und de Cooman schlagen vor, die Beschränkungen für die einzelnen Dimensionen von $\mathcal{Y}^{(0)}$ miteinander zu verknüpfen. Bei der Anwendung des Modells auf die lineare Regression wird in Kapitel 4.2.4 der Vorschlag für die multivariate Normalverteilung aufgegriffen.

Die Bedingung, dass der Parameterraum beschränkt sein muss, scheint es zu erschweren, das Modell von Quaeghebeur und de Cooman zur Beschreibung von a priori Nichtwissen heranzuziehen. Die Autoren führen jedoch an, dass eine solche Einschränkung möglich sein müsse, wenn es schon möglich war, die Verteilungsfamilie auszuwählen. Vermutlich sind generell Situationen, in denen überhaupt keine Informationen über die ungefähre Größenordnung eines Parameters vorhanden sind, relativ selten.

In der Anwendung auf die lineare Regression, die das Thema dieser Arbeit ist, gibt es für den Regressionsparameter β jedoch gewissermaßen einen ‚natürlichen Nullpunkt', und im Falle von standardisierten Regressionsvariablen dürfte es sicher nicht schwierig sein, die Grenzen der Beträge von β, die a priori vernünftigerweise erwartbar sind, festzulegen. Die Grenzen für die Varianz-Kovarianzmatrix für β sind hingegen deutlich schwieriger zu wählen. Im Rahmen der konkreten Implementierung des Modells werden in Kapitel 4.4 jedoch verschiedene Auswahlstrategien vorgestellt und ihre Ergebnisse kurz diskutiert.

4 Bayes-Regression unter komplexer Unsicherheit

4.1 Einführung

In diesem Kapitel soll nun das Modell von Quaegehebeur und de Cooman auf die Problemstellung der Schätzung der Parameter einer linearen Regression angewendet werden. Statt wie in der klassisch bayesanischen Methodik eine einzige Priori-Verteilung aufzudatieren, soll hier also nun eine Menge von Priori-Verteilungen bayesianisch aufdatiert werden, um nicht die unrealistischen und überpräzisen Annahmen über das Vorwissen treffen zu müssen, die für eine Modellierung gemäß der klassisch bayesanischen Methodik nötig sind. Das resultierende Modell ermöglicht es, die Methode der Bayes-Regression auf Situationen mit komplexer Unsicherheit anwendbar zu machen: Es bietet differenzierte Möglichkeiten, a priori vorhandenes Wissen über die Regressionsparameter in den Inferenzprozess einzubinden; es kann aber auch, bei entsprechender Wahl der Parametermenge $\mathcal{Y}^{(0)}$, für eine deutlich realistischere Modellierung von a priori Nichtwissen sorgen, als es die standardmäßig verwendeten nichtinformativen Priori-Verteilungen erlauben.

4.1.1 Zur Wahl des Intervallwahrscheinlichkeitsmodells

Die Erzeugung der Menge von Priori-Verteilungen erfolgt hier parametrisch mittels der konjugierten Verteilung. Dass es auch andere Möglichkeiten zur Erzeugung von Mengen von Verteilungen gibt, wurde schon in Kapitel 1.7 erwähnt; die Gründe für die Wahl einer Modellierung nach [Quaeghebeur und de Cooman 2005] anstatt eines der dort diskutierten Modelle sollen im Folgenden kurz erläutert werden.

Das hier verwendete parametrische Modell hat gegenüber denjenigen mit oberen und unteren Dichtegrenzen zwar den Nachteil, dass die Verteilungsfamilie der Priori-Verteilung fest gewählt werden muss; dass dies allein aus technischen Gründen geschieht, ist von einem allgemeinen Standpunkt aus schwer rechtfertigbar, ein solches Vorgehen ist aber in der bayesianischen Methodik durchaus völlig üblich und akzeptiert; es ist schwerlich vorstellbar, dass nichtinformative Priori-Verteilungen eine solche Verbreitung gefunden hätten, wenn sie die Berechenbarkeit der Posteriori-Verteilung in hohem Maße erschwert hätten. Andererseits spricht auch gerade die Berechenbarkeit für das hier verwendete Modell; wie in Kapitel 3.2 gezeigt wurde, können die natürlichen Parameter

der Priori-Verteilungen linear aufdatiert werden. Diese ‚exakte' Aufdatierung nach dem Satz von Bayes sorgt auch für recht exakte a posteriori Aussagen, die bei einer Verwendung der in der Literatur zur Modellierung komplexer Unsicherheit diskutierten alternativen Schlussregeln wie z.B. die ‚generalized Bayes´ rule' [Walley 1991] nicht möglich wären.

Bei den Modellen mit oberen und unteren Dichtegrenzen, die in [Coolen 1993] und [Pericchi und Walley 1991] vorgestellt werden, werden hingegen *alle* Verteilungen, deren Dichten zwischen den festgelegten Grenzen liegen, berücksichtigt. Sie bieten daher eine sehr allgemeine und umfassende Modellierung. Problematisch ist dabei jedoch, dass daher auch ‚exotische' und sich ex post als unsinnig erweisende Verteilungen in der Menge enthalten sein können, und daher nur schwache Posteriori-Aussagen möglich sein können.

Speziell für das Modell von Coolen gilt, wie schon in Kapitel 1.7 angemerkt, dass es trotzdem keine Möglichkeit bietet, ‚prior-data conflict'-Situationen angemessen zu modellieren. In dem im vorigen Kapitel vorgestellten und in dieser Arbeit angewendeten Modell ist das ebenfalls noch nicht möglich; in Kapitel 3.3 wurde jedoch schon darauf hingewiesen, auf welche Art das Modell erweitert werden kann, um ein sinnvolles Verhalten in einem solchen Fall zu gewährleisten. In der Untersuchung von Pericchi und Walley wird hingegen nach dem Anwendungszweck unterschieden und für die beiden Situationen (a priori Nichtwissen bzw. Vorwissen, dass zu einem ‚prior data conflict' führen könnte) verschiedene ‚maßgeschneiderte' Modelle verglichen; die Modelle für a priori Nichtwissen sind dann für Situationen mit Vorwissen untauglich und umgekehrt.

Der grundsätzlichste Unterschied zwischen dem hier angewendeten Modell von Quaeghebeur und de Cooman und den in [Coolen 1993] und [Pericchi und Walley 1991] vorgestellten Modellen ist jedoch, dass ersteres auch Mengen von Verteilungen über mehr als einen Parameter beschreibbar macht; die anderen aufgeführten sind nur Modelle für den Fall von Dichten über einen eindimensionalen Parameter, die von Perrichi und Walley sogar nur für den Spezialfall eines Lageparameters.

4.1.2 Regression

Regression ist das zentrale Konzept für asymmetrische Fragestellungen in der statistischen Methodik. Es handelt sich um ein umfassendes Konzept, mit dem der Einfluss einer oder mehrerer erklärender Variablen auf eine oder mehrere Zielvariablen modelliert werden soll. Die erklärenden Variablen heißen Regressoren und können spaltenweise in einer sogenannten Designmatrix $\mathbf{X}$ zusammengefasst werden. Bei dem einfachsten Fall, der linearen Regression, wird der Vektor z der unabhängigen Realisationen einer metrischen Zielvariable Z mit ebenfalls metrischen Regressoren über folgende Regressionsgleichung in Verbindung gesetzt:

$$z = \mathbf{X}\beta + \varepsilon$$

Dabei soll die Anzahl der Beobachtungen mit k, die Anzahl der erklärenden Variablen mit p bezeichnet werden. z ist also ein Vektor der Länge k, $\mathbf{X}$ eine Matrix mit k Zeilen und p Spalten. Der Einfluss der p Regressoren auf die Zielvariable wird über die Schätzungen von p Parametern quantifiziert, die in einem Vektor β zusammengefasst werden. Wenn das Ziel der Regression ist, den Einfluss der Regressoren auf die Zielvariable einzuschätzen, ist die Größenordnung dieser Parameter β von Interesse; die geschätzten Parameter können aber auch dazu genutzt werden, für bestimmte Werte der Regressoren eine Vorhersage für die Zielvariable Z zu erstellen. Mittels des vektoriellen Fehlerterms ε wird berücksichtigt, dass sich die beobachteten Werte z normalerweise nicht in einen perfekt linearen Zusammenhang mit $\mathbf{X}$ bringen lassen. Über die Verteilung von ε müssen im einfachsten Fall folgende Annahmen getroffen werden:

$\mathbb{E}[\varepsilon] = \mathbf{0}\,,$	der Erwartungswert von ε ist Null, und
$\mathbb{V}(\varepsilon_i) = \sigma^2\,,\ i = 1\ldots, k,$	die Varianz jedes Elements des Fehlerterms beträgt σ^2, wobei die einzelnen Elemente unkorreliert sind, so dass
$\mathbb{V}(\varepsilon) = \sigma^2\mathbf{I}$	gilt.

Wird zusätzlich angenommen, dass die Verteilung von ε einer Normalverteilung entspricht, sind Aussagen über die Verteilung der geschätzten Werte von β möglich, die für die Erstellung von Intervallschätzungen und Tests verwendbar sind. In diesem Fall gilt, da die Designmatrix $\mathbf{X}$ als nichtstochastisch angesehen wird, dass auch Z normalverteilt ist.

Die lineare Regression ist der Ausgangspunkt für vielfältige Erweiterungen, z.B. zur Berücksichtigung kategorialer Daten; die erhaltenen Parameter können dann etwa zur Klassifikation von Untersuchungseinheiten genutzt werden. Ebenso gibt es Erweiterungen für Beobachtungen von Z mit komplexer Abhängigkeitsstruktur, etwa im Falle von Zeitreihen oder wiederholter Messungen. Die generalisierten linearen Modelle sind eine Modellklasse, bei der zwischen Prädiktor und Zielvariable eine Funktion geschaltet wird, um eine Vielzahl an Verteilungen für Z zu ermöglichen; sie erlauben daher auch die Behandlung diskreter Zielvariablen.

Für die Schätzung der gesuchten Parameter β gibt es verschiedene Methoden. Nach der klassische Methode der kleinsten Quadrate werden die Parameter so gewählt, dass die Summe der quadrierten Differenzen zwischen den beobachteten Werten von Z und den durch den linearen Prädiktor $\mathbf{X}\beta$ geschätzten Werten minimal wird. (Kurzbezeichnung: KQ-Schätzung.) Im Falle der linearen Regression mit normalverteiltem Fehlerterm fällt das Ergebnis dieser Methode mit dem einer Likelihood-Schätzung zusammen, bei der der maximale Wert der Likelihood von β bei der Beobachtung von z den Schätzwert ergibt. Alternative Schätzmethoden wie Ridge oder Elastic Net wurden für Situationen entwickelt, in denen die Vorhersage das wesentliche Ziel ist oder eine Schätzung nach der Methode der kleinsten Quadrate nicht handhabbar ist.

Die Schätzmethode, die in dieser Arbeit verallgemeinert werden soll, ist die bayesianische Schätzung. Analog zu dem schon in den Kapiteln 2 und 3 beschriebenen Vorgehen wird dabei eventuell vorhandenes Vorwissen in eine Priori-Verteilung über den zu schätzenden Parameter β übersetzt. Die zur Aufdatierung dieser Verteilung herangezogene Likelihood basiert hier auf den gemeinsamen Beobachtungen $\mathbf{X}$ und z; wieder stellt die Posteriori-Verteilung die Basis für alle Schlüsse aus den Daten dar. Als Punktschätzung für den Regressionsparameter β kann beispielsweise der Modus oder der Erwartungswert der Posteriori-Verteilung über β herangezogen werden. Wird die Priori-Verteilung über β als konstant auf $\mathbb{R}$ angenommen und der Modus der Posteriori-Verteilung für die Ermittlung der Punktschätzung verwendet, gilt, da die Posteriori-Verteilung proportional zur Likelihood ist, dass die bayesianische Schätzung der Likelihood-Schätzung und damit auch der KQ-Schätzung entspricht. Im Folgenden soll daher, wenn Ergebnisse des hier untersuchten Modells mit denen der KQ-Schätzung verglichen werden, implizit auch der Vergleich mit der Likelihood-Schätzung und einer klassischen bayesianischen Schätzung unter den erwähnten Bedingungen gemeint sein.

Auch für die Fragestellung der linearen Regression gibt es eine konjugierte Verteilungsfamilie, die die bayesianische Analyse wesentlich vereinfacht. Da normalerweise die Annahme gemacht wird, dass ε normalverteilt ist, ist die Likelihood von z gegeben $\mathbf{X}$, β und σ^2 ebenfalls normalverteilt. Die konjugierte Verteilung zu dieser Likelihood ist dann wiederum eine Normalverteilung. Dieses Modell, das hier als Normal-Modell (für die lineare Regression) bezeichnet werden soll, wird in Kapitel 4.2 zuerst in Abschnitt 4.2.1 beschrieben und in den darauf folgenden Abschnitten verallgemeinert, indem nicht nur eine Normalverteilung über β, sondern eine Menge von Normalverteilungen über β aufdatiert wird.

Um die Situation möglichst umfassend zu modellieren, wird idealerweise jedoch nicht nur eine Priori über β, sondern auch eine Priori über σ^2 definiert. σ^2, die Varianz der Fehlerterme, für als nicht-stochastisch angesehenes $\mathbf{X}$ auch die Varianz von Z, ist im Allgemeinen unbekannt, hat aber weitreichenden Einfluss auf die Genauigkeit der Schätzung von β. Die beiden Priori-Verteilungen lassen sich dann als eine gemeinsame Priori-Verteilung über β und σ^2 auffassen; die konjugierte Verteilung in diesem Fall ist dann die Normal-InversGamma-Verteilung (abgekürzt: NIG-Verteilung). Die Beschreibung dieses Modells ist im Anhang in Kapitel A.1 zu finden; leider ist die Verallgemeinerung analog zur Vorgehensweise beim Normal-Modell nicht möglich. Die Erklärung dafür kann in Abschnitt A.1.3 gefunden werden, Hinweise auf die sich ergebenden Probleme werden aber schon bei der Verallgemeinerung des Normal-Modells in Abschnitt 4.2.3 gegeben. Das Normal-Modell und das NIG-Modell werden in Kapitel 4.2.1 bzw A.1.1 im Detail vorgestellt und deren Konjugiertheits-Eigenschaft gezeigt, da in manchen Lehrbüchern (z.B. [Box und Tiao 1973]) ausschließlich Ansätze mit nicht-informativen Priori-Verteilungen behandelt werden.

4.1.3 Vorgehen

Das Konzept der Verallgemeinerung des Normal-Modells für die lineare Regression, die das Ziel dieser Arbeit darstellt, ist folgendes: Die konjugierte Verteilung des Normal-Modells soll direkt mit der konjugierten Verteilung des Modells von Quaeghebeur und de Cooman in Verbindung gebracht werden, so dass jedes Element einer Menge von multivariaten Normalverteilungen in Anwendung von [Quaeghebeur und de Cooman 2005] bayesianisch aufdatiert werden kann und sich so eine Menge von a posteriori Normalverteilungen ergibt. Die damit erzeugten ‚credal sets', die Mengen der mit diesem Modell vereinbaren klassischen Wahrscheinlichkeitszuordnungen, bestehen dann aus allen Konvexkombinationen aller a posteriori Normalverteilungen.

In Kapitel 4.2.1 wird das Normal-Modell vorgestellt, in Kapitel 4.2.2 wird gezeigt, dass die Normalverteilung einer Exponentialfamilie gemäß der von Quaeghebeur und de Cooman verwendeten Systematik entspricht; die resultierende Form der Bestandteile der Dichte gemäß (3.1) wird angegeben. Danach wird in Abschnitt 4.2.3 gezeigt, wie diese Form angepasst werden muss, um einer konjugierten Verteilung aus dem Modell von Quaeghebeur und de Cooman (siehe Gleichung (3.3)) zu entsprechen. Dann wird in Kapitel 4.2.4 erläutert, wie sich die Strategie, den Parameter y der konjugierten Verteilung zu variieren, auf das hier untersuchte Modell auswirkt und auf die sich ergebenden Probleme bei einer konkreten Implementierung eingegangen. Kapitel 4.3 zeigt dann, wie ein Modell im Falle von zwei Regressoren erstellt werden kann, während in Kapitel 4.4 die Anwendung eines solchen Modells an drei simulierten Datensätzen gezeigt wird und so dessen Aussagekraft einschätzbar gemacht werden soll. In Abschnitt 4.5 erfolgt schließlich die Anwendung dieses Modells auf einen realen Datensatz.

4.2 Das Normal-Modell für die lineare Regression

Die Darstellung des Normal-Modells in diesem Kapitel ist so gewählt, dass die Erweiterung auf das NIG-Modell in Kapitel A.1 nur ein vergleichsweise kleiner Schritt ist. Aus diesem Grund wird σ^2, das eigentlich als bekannt und nichtstochastisch vorausgesetzt wird, in der Notation wie eine Zufallsgröße behandelt und erscheint deswegen z.B. bei der Beschreibung der Priori-Dichte von β in der Bedingung.

4.2.1 Das Modell

Annahmen des Regressionsmodells

Für die multiple Regression sei die Regressionsgleichung folgendermaßen notiert:

$$z = \mathbf{X}\beta + \varepsilon\,, \quad \mathbf{X} \in \mathbb{R}^{k\times p}\,,\ \beta \in \mathbb{R}^{p}\,,\ z \in \mathbb{R}^{k}\,,\ \varepsilon \in \mathbb{R}^{k}$$

Von Z und von jeder Variablen in $\mathbf{X}$ gibt es k Beobachtungen, die wie in klassisch bayesianischen Modellierungen alle als präzise und nicht fehlerbehaftet angesehen

werden. Die Anzahl der erklärenden Variablen und somit die Dimension von β ist p.

Der Fehlerterm ε ist ein Zufallsvektor der Dimension k; jedes Element mit dem Index $i = 1, \ldots, k$ soll normalverteilt mit Erwartungswert 0 und mit der bekannten Varianz σ^2 verteilt sein; die einzelnen Fehlerterme sind unabhängig.

$$\varepsilon_i \overset{i.i.d}{\sim} \mathrm{N}(0, \sigma^2) \quad \Longrightarrow \quad \varepsilon \sim \mathrm{N}_k(\mathbf{0}, \sigma^2 \mathbf{I})) \quad (\sigma^2 \text{ bekannt})$$

Aus diesen Annahmen folgt, dass die Likelihood ebenfalls multivariat normalverteilt ist:

$$z \,|\, X, \beta, \sigma^2 \sim \mathrm{N}_k(\mathbf{X}\beta, \sigma^2 \mathbf{I})$$

Priori-Verteilung

Als Priori-Verteilung für β wird eine zur obigen Likelihood konjugierte Verteilung herangezogen. Nach [O'Hagan 1994, S. 244ff] kann es sich dabei um eine multivariate Normalverteilung der Dimension p handeln. Die Werte deren Parameter $\beta^{(0)}$ und $\mathbf{\Sigma}^{(0)}$ können dann aus dem Vorwissen bestimmt werden. Der obere Index $^{(0)}$ soll, analog zur Notation in den vorigen Kapiteln, die Parameter der Priori-Verteilung auszeichnen.

$$\beta \,|\, \sigma^2 \;\sim\; \mathrm{N}_p\left(\beta^{(0)}, \sigma^2 \mathbf{\Sigma}^{(0)}\right) \qquad \text{mit } \beta^{(0)} \in \mathbb{R}^p\,,\ \mathbf{\Sigma}^{(0)} \in \mathbb{R}^{p \times p} \text{ positiv definit}$$

d.h.

$$p(\beta \,|\, \sigma^2) \;=\; \frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}} (\sigma^2)^{\frac{p}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\beta - \beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta - \beta^{(0)}\right) \right\} \quad (4.1)$$

Posteriori-Verteilung

Die Posteriori-Verteilung kann nun gemäß des Satzes von Bayes berechnet werden:

$$\begin{aligned} p(\beta \,|\, \sigma^2, \mathbf{X}, z) &= \frac{p(\beta, \mathbf{X}, z \,|\, \sigma^2)}{p(\mathbf{X}, z \,|\, \sigma^2)} \\ &= \frac{p(\mathbf{X}, z \,|\, \beta, \sigma^2) \cdot p(\beta \,|\, \sigma^2)}{p(\mathbf{X}, z \,|\, \sigma^2)} \\ &\propto p(z \,|\, \mathbf{X}, \beta, \sigma^2)\, p(\beta \,|\, \sigma^2) \end{aligned}$$

Die Konjugiertheitseigenschaft und der Aufdatierungsschritt für die Parameter $\beta^{(0)}$ und $\mathbf{\Sigma}^{(0)}$ sollen im Folgenden gezeigt werden. Mit obiger Likelihood und Priori-Verteilung gilt

$$\begin{aligned} p(\beta \,|\, \sigma^2, \mathbf{X}, z) \;\propto\; & \frac{1}{(\sigma^2)^{\frac{k}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} (z - \mathbf{X}\beta)^{\mathsf{T}} (z - \mathbf{X}\beta) \right\} \\ & \cdot \frac{1}{(\sigma^2)^{\frac{p}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\beta - \beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta - \beta^{(0)}\right) \right\} \\ = \; & \frac{1}{(\sigma^2)^{\frac{p}{2}} (\sigma^2)^{\frac{k}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \Big[(z - \mathbf{X}\beta)^{\mathsf{T}} (z - \mathbf{X}\beta) \right. \\ & \left. + \left(\beta - \beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta - \beta^{(0)}\right) \Big] \right\}. \end{aligned}$$

Betrachtet man nun den Term in den eckigen Klammern im Exponenten isoliert, so ergibt sich:

$$\begin{aligned}
\Big[\cdots\Big] &= (z-\mathbf{X}\beta)^{\mathsf{T}}(z-\mathbf{X}\beta) + \left(\beta-\beta^{(0)}\right)^{\mathsf{T}}\boldsymbol{\Sigma}^{(0)^{-1}}\left(\beta-\beta^{(0)}\right) \\
&= z^{\mathsf{T}}z - z^{\mathsf{T}}(\mathbf{X}\beta) - (\mathbf{X}\beta)^{\mathsf{T}}z + (\mathbf{X}\beta)^{\mathsf{T}}(\mathbf{X}\beta) \\
&\quad + \beta^{\mathsf{T}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta - \beta^{\mathsf{T}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)} - \beta^{(0)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta + \beta^{(0)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)} \\
&= z^{\mathsf{T}}z + \beta^{(0)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)} \\
&\quad - \underbrace{\beta^{\mathsf{T}}}_{A^{\mathsf{T}}}\underbrace{\left(\mathbf{X}^{\mathsf{T}}z + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)}_{CB} \\
&\quad - \underbrace{\left(z^{\mathsf{T}}\mathbf{X} + \beta^{(0)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(0)^{-1}}\right)^{\mathsf{T}}}_{B^{\mathsf{T}}C}\underbrace{\beta}_{A} \\
&\quad + \underbrace{\beta^{\mathsf{T}}}_{A^{\mathsf{T}}}\underbrace{\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)}_{C}\underbrace{\beta}_{A}
\end{aligned}$$

Quadratische Ergänzung mit dem fehlenden Term $B^{\mathsf{T}}CB$, also Ergänzung mit

$$\begin{aligned}
&-\left(z^{\mathsf{T}}\mathbf{X} + \beta^{(0)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(0)^{-1}}\right)\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right) \\
&+\left(z^{\mathsf{T}}\mathbf{X} + \beta^{(0)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(0)^{-1}}\right)\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)
\end{aligned}$$

liefert

$$\begin{aligned}
\Big[\cdots\Big] &= z^{\mathsf{T}}z + \beta^{(0)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)} \\
&\quad - \left(z^{\mathsf{T}}\mathbf{X} + \beta^{(0)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(0)^{-1}}\right)\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right) \\
&\quad + \Bigg[\beta - \underbrace{\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)}_{=:\beta^{(1)}}\Bigg]^{\mathsf{T}} \\
&\quad \cdot \underbrace{\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)}_{=:\boldsymbol{\Sigma}^{1^{-1}}} \\
&\quad \cdot \Bigg[\beta - \underbrace{\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)}_{=:\beta^{(1)}}\Bigg] \\
&= z^{\mathsf{T}}z + \beta^{(0)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)} - \beta^{(1)^{\mathsf{T}}}\boldsymbol{\Sigma}^{(1)^{-1}}\beta^{(1)} + \left(\beta-\beta^{(1)}\right)^{\mathsf{T}}\boldsymbol{\Sigma}^{(1)^{-1}}\left(\beta-\beta^{(1)}\right).
\end{aligned}$$

Also ist

$$\begin{aligned} p(\beta \,|\, \sigma^2, \mathbf{X}, z) \;&\propto\; \frac{1}{(\sigma^2)^{\frac{p}{2}}(\sigma^2)^{\frac{k}{2}}} \\ &\quad \cdot \exp\left\{ -\frac{1}{2\sigma^2}\Big[z^\mathsf{T} z + \beta^{(0)\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \beta^{(0)} - \beta^{(1)\mathsf{T}} \mathbf{\Sigma}^{(1)^{-1}} \beta^{(1)} \right. \\ &\qquad \left. + \left(\beta - \beta^{(1)}\right)^\mathsf{T} \mathbf{\Sigma}^{(1)^{-1}} \left(\beta - \beta^{(1)}\right) \Big] \right\} \\ &\propto\; \exp\left\{ -\frac{1}{2\sigma^2} \left(\beta - \beta^{(1)}\right)^\mathsf{T} \mathbf{\Sigma}^{(1)^{-1}} \left(\beta - \beta^{(1)}\right) \right\} . \end{aligned}$$

Der Exponent entspricht dem Kern einer Normalverteilung über β mit Erwartungswert $\beta^{(1)}$ und Varianz-Kovarianzmatrix $\sigma^2\mathbf{\Sigma}^{(1)}$, so dass gilt:

$$p(\beta \,|\, \sigma^2, \mathbf{X}, z) \sim \mathrm{N}_p\left(\beta^{(1)}, \sigma^2\mathbf{\Sigma}^{(1)}\right)$$

Die aufdatierten Parameter sind folgendermaßen erhältlich:

$$\begin{aligned} \beta^{(1)} &= \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Sigma}^{(0)^{-1}}\right)^{-1} \left(\mathbf{X}^\mathsf{T} z + \mathbf{\Sigma}^{(0)^{-1}} \beta^{(0)}\right) \qquad (4.2) \\ &= \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Sigma}^{(0)^{-1}}\right)^{-1} \left(\mathbf{X}^\mathsf{T}\mathbf{X}\hat{\beta} + \mathbf{\Sigma}^{(0)^{-1}} \beta^{(0)}\right) \\ &= (\mathbf{I} - \mathbf{A})\beta^{(0)} + \mathbf{A}\hat{\beta} \qquad \text{mit } \mathbf{A} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Sigma}^{(0)^{-1}}\right)^{-1} \mathbf{X}^\mathsf{T}\mathbf{X} \\ \mathbf{\Sigma}^{(1)} &= \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Sigma}^{(0)^{-1}}\right)^{-1} \qquad (4.3) \end{aligned}$$

Mit der Wahl einer Normalverteilung als Priori für β lassen sich also über die Aufdatierungsschritte (4.2) und (4.3) die Posteriori-Parameter der Verteilung des Regressionsparameters β einfach berechnen.

Aus der Form von (4.2), die in Bezug auf $\hat{\beta}$ umgeformt wurde, ist erkenntlich, dass der Posteriori-Erwartungswert für β als ein gewichtetes Mittel von $\beta^{(0)}$ und der KQ-Schätzung $\hat{\beta}$ angesehen werden kann. Die Gewichte setzen sich aus der für $\beta^{(0)}$ angenommenen Varianz-Kovarianzstruktur $\mathbf{\Sigma}^{(0)}$ und der zu $\hat{\beta}$ gehörenden ‚empirischen' Varianz-Kovarianzstruktur $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ zusammen. Setzt man $\beta^{(0)} = \hat{\beta}$, wird also die KQ-Schätzung des Regressionsparameters als Erwartungswert der konjugierten Priori-Verteilung eingesetzt, so erhält man als Posteriori-Mittelwert wieder $\hat{\beta}$, jedoch mit einer verminderten Varianz.

4.2.2 Die Normalverteilung als Exponentialfamilie

In dem in dieser Arbeit behandelten Ansatz soll überprüft werden, ob sich die konjugierte Verteilung des Normal-Modells direkt als eine konjugierte Verteilung des Modells von Quaeghebeur und de Cooman verstehen lässt. Aus diesem Grund wird im Folgenden die Dichte der konjugierten Verteilung aus dem Normal-Modell schrittweise an die

Form (3.3) angepasst. Eine Alternative zu diesem Vorgehen wäre, sich eine konjugierte Verteilung aus der Likelihood des Normal-Modells in der Weise zu konstruieren, wie das in [Quaeghebeur und de Cooman 2005] geschieht. Diese Möglichkeit soll in dieser Arbeit jedoch nicht untersucht werden.

Zur Anwendung des Modells von Quaeghebeur und de Cooman ist es in einem ersten Schritt nötig, die konjugierte Verteilung in die Form einer Exponentialfamilie gemäß [Quaeghebeur und de Cooman 2005] zu bringen, um die Form deren Bestandteile ψ und $\mathbf{b}(\psi)$ zu erkennen zu können. Diese erscheinen dann unverändert in der konjugierten Verteilung. Aus dem in diesem Schritt ermitteltem $\tau(\cdot)$ kann dann in einem nächsten Schritt ebenfalls die Form von $y^{(0)}$ gefolgert werden.

Sei also $\beta \,|\, \sigma^2$ normalverteilt mit den Parametern $\beta^{(0)} \in \mathbb{R}^p$, $\mathbf{\Sigma}^{(0)} \in \mathbb{R}^{p \times p}$ symmetrisch positiv definit. Soll $\mathbf{\Sigma}^{(0)}$ voll parametrisiert werden, sind dazu $\frac{p(p+1)}{2}$ eindimensionale Parameter nötig. Die Anzahl der Parameter bei einer vollen Parametrisierung beträgt dann $p + \frac{p(p+1)}{2} = \frac{p(p+3)}{2} =: q$. Gilt also

$$\beta \,|\, \sigma^2 \sim \mathrm{N}_p(\beta^{(0)}, \sigma^2 \mathbf{\Sigma}^{(0)}) \,,$$

dann hat die Dichte folgende gemäß (4.1) die folgende Form:

$$p(\beta \,|\, \sigma^2) \;=\; \frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}} (\sigma^2)^{\frac{p}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\beta - \beta^{(0)}\right)^{\mathsf{T}} {\mathbf{\Sigma}^{(0)}}^{-1} \left(\beta - \beta^{(0)}\right) \right\}$$

Diese Dichte soll nun auf die Form einer Exponentialfamilie gemäß der Notation von Quaeghebeur und de Cooman gebracht werden. Die für diesen Zweck konkretisierte Form von (3.1) lautet:

$$p(\psi) = \mathbf{a}\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right) \cdot \exp\left\{ \left\langle \psi \,,\, \tau\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right) \right\rangle - \mathbf{b}(\psi) \right\}$$

Hier handelt es sich um eine Dichte über den Parameter ψ, welcher sich als Funktion von β ergeben muss. ψ soll hier schon wie in der konjugierten Verteilung als der natürlicher Parameter angesehen werden. Aus diesem Grund spielen $\beta^{(0)}$ und $\mathbf{\Sigma}^{(0)}$ in diesem Schritt die Rolle der ‚Beobachtung'.

Umformung:

$$\begin{aligned}
p(\beta) &= \frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}(\sigma^2)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(\beta-\beta^{(0)}\right)^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\left(\beta-\beta^{(0)}\right)\right\} \\
&= \frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}(\sigma^2)^{\frac{p}{2}}} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\left[\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta - \beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)} - \beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta + \beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}\right]\right\} \\
&= \underbrace{\frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}(\sigma^2)^{\frac{p}{2}}} \cdot \exp\left(-\frac{1}{2\sigma^2}\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)}_{=:\,\mathbf{a}(\beta^{(0)},\mathbf{\Sigma}^{(0)})} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta + \frac{1}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)} + \frac{1}{2\sigma^2}\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta\right\}
\end{aligned}$$

Da $\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}$ und $\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta$ Skalare sind und $\left(\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)^{\mathsf{T}} = \beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta$ gilt, können diese beiden Terme zusammengefasst werden.

Zur Vereinfachung der Notation sei im Folgenden $\mathbf{\Sigma}^{(0)^{-1}} =: \mathbf{\Lambda}^{(0)}$ mit den Elementen $\left(\mathbf{\Lambda}^{(0)}\right)_{ij} =: \lambda_{ij}^{(0)}$.

Außerdem seien die Laufindizes folgendermaßen festgelegt:

h, i, j	Laufindex von $1,\ldots,p = \dim(\beta)$
l, m	Laufindex von $1,\ldots,k = \dim(z)$

Somit folgt

$$\begin{aligned}
p(\beta) &= \mathbf{a}(\beta^{(0)},\mathbf{\Sigma}^{(0)}) \cdot \exp\left\{-\frac{1}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta + \frac{1}{\sigma^2}\beta^{\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta^{(0)}\right\} \\
&= \mathbf{a}(\beta^{(0)},\mathbf{\Sigma}^{(0)}) \cdot \exp\left\{-\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\beta_i\beta_j}{2\sigma^2}\cdot\lambda_{ij}^{(0)} + \sum_{i=1}^{p}\frac{\beta_i}{\sigma^2}\cdot\Big(\sum_{j=1}^{p}\lambda_{ij}^{(0)}\beta_j^{(0)}\Big)\right\} \\
&= \mathbf{a}(\beta^{(0)},\mathbf{\Sigma}^{(0)}) \cdot \exp\left\{-\sum_{i=1}^{p}\frac{\beta_i^2}{2\sigma^2}\cdot\lambda_{ii}^{(0)} - \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\frac{\beta_i\beta_j}{\sigma^2}\cdot\lambda_{ij}^{(0)}\right. \\
&\qquad\qquad \left. + \sum_{i=1}^{p}\frac{\beta_i}{\sigma^2}\cdot\Big(\sum_{j=1}^{p}\lambda_{ij}^{(0)}\beta_j^{(0)}\Big)\right\}
\end{aligned}$$

Also gilt:

$$\begin{aligned}
\mathbf{a}\left(\beta^{(0)},\mathbf{\Sigma}^{(0)}\right) &= \frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}(\sigma^2)^{\frac{p}{2}}} \cdot \exp\left(-\frac{1}{2\sigma^2}\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}\right) \\
\mathbf{b}(\psi) &= 0
\end{aligned}$$

$$\Psi = \begin{bmatrix} -\frac{\beta_1^2}{2\sigma^2} \\ \vdots \\ -\frac{\beta_p^2}{2\sigma^2} \\ -\frac{\beta_1\beta_2}{\sigma^2} \\ \vdots \\ -\frac{\beta_{p-1}\beta_p}{\sigma^2} \\ -\frac{\beta_1}{\sigma^2} \\ \vdots \\ -\frac{\beta_p}{\sigma^2} \end{bmatrix}, \quad \tau\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right) = \begin{bmatrix} \lambda_{11}^{(0)} \\ \vdots \\ \lambda_{pp}^{(0)} \\ \lambda_{12}^{(0)} \\ \vdots \\ \lambda_{p-1,p}^{(0)} \\ \sum_{j=1}^{p} \lambda_{1j}^{(0)} \beta_j^{(0)} \\ \vdots \\ \sum_{j=1}^{p} \lambda_{pj}^{(0)} \beta_j^{(0)} \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ \vdots \\ a \end{matrix}}\right\} p \text{ Summanden} \\ \left.\vphantom{\begin{matrix} a \\ \vdots \\ a \end{matrix}}\right\} \frac{p(p-1)}{2} \\ \left.\vphantom{\begin{matrix} a \\ \vdots \\ a \end{matrix}}\right\} p \end{matrix}$$

Dass im Exponenten kein Term auftaucht, der die Rolle von $\mathbf{b}(\psi)$ übernimmt, erstaunt etwas, jedoch macht es gerade dieses Ergebnis im folgenden nächsten Schritt möglich, die Methodik von Queaghebeur und de Cooman auf das Normal-Modell anzuwenden, wie sich in folgenden Abschnitt zeigen wird.

Die Anzahl der natürlichen Parameter beträgt $p + \frac{p(p-1)}{2} + p = \frac{p(p+3)}{2} = q$, die Normalverteilung für β ist also „strikt q-parametrig“ [Rüger 1999, S. 19] und besitzt somit eine Eigenschaft, die positiv für die Behandlung von Test- und Schätzproblemen ist.

4.2.3 Anwendung des Aufdatierungsmodells von Quaeghebeur und de Cooman auf das Normal-Modell

Die in Kapitel 3 schon gezeigte Formel (3.3) für die konjugierte Verteilung soll hier in einer an die konkrete Anwendung angepasste Form notiert werden. Die konjugierte Verteilung hat nach [Quaeghebeur und de Cooman 2005] als Priori-Verteilung mit den Parametern $n^{(0)}$ und $y^{(0)}$ die Form

$$p(\psi) \;=\; \mathbf{c}\left(n^{(0)}, y^{(0)}\right) \exp\left\{n^{(0)}\left[\left\langle \psi\,,\, y^{(0)}\right\rangle - \mathbf{b}(\psi)\right]\right\} \tag{4.4}$$

Die in Kapitel 4.2.2 ermittelte Form der Normalverteilung kann einfach an diese Form angepasst werden, indem

- jeder Summand im Exponenten von $p(\beta \,|\, \sigma^2)$ mit dem Faktor $\frac{1}{n^{(0)}}$ multipliziert wird. Dieser Faktor wird $\tau\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right)$ zugeordnet und entspricht damit dem $y^{(0)}$ des Modells von Quaeghebeur und de Cooman:

$$y^{(0)} \;:=\; \frac{1}{n^{(0)}}\,\tau\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right) \tag{4.5}$$

- $\mathbf{c}\left(n^{(0)}, y^{(0)}\right)$ aus (4.4) mit dem in Kapitel 4.2.2 ermittelten $\mathbf{a}\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right)$ identifiziert wird. $\mathbf{a}\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right)$ hängt dabei tatsächlich auch von der Anzahl der Pseudocounts $n^{(0)}$ ab, wenn man (wie Quaeghebeur und de Cooman) von $y^{(0)}$ ausgeht und $\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right)$ als Funktion von $n^{(0)}$ und $y^{(0)}$ ansieht, mit (4.5) also $\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right) = \tau^{-1}\left[n^{(0)} \cdot y^{(0)}\right]$ definiert und damit folgendes gilt:

$$\mathbf{a}\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right) \;=\; \mathbf{a}\left(\tau^{-1}\left[n^{(0)} \cdot y^{(0)}\right]\right) .$$

Eine mit einer eindimensionalen Beobachtung x aufdatierte Priori, die Posteriori über Ψ, hat in der Notation von Quaeghebeur und de Cooman gemäß (3.4) die Form

$$p(\psi \,|\, x) \;=\; \mathbf{c}\big((n^{(0)}+1), y^{(1)}\big) \exp\Big\{(n^{(0)}+1)\big[\,\langle \psi\,,\, y^{(1)}\rangle - \mathbf{b}(\psi)\big]\Big\}, \qquad (4.6)$$
$$\text{mit } y^{(1)} = \frac{n^{(0)} y^{(0)} + \tau^x(x)}{n^{(0)}+1} .$$

Zu beachten ist hierbei, dass $\tau^x(x)$ der suffizienten Statistik in der Likelihood entspricht, die zu dieser konjugierten Verteilung gehört, und nichts mit $\tau\left(\beta^{(0)}, \mathbf{\Sigma}^{(0)}\right)$ zu tun hat, welches wir aus der konjugierten Verteilung selbst erhalten haben. Die suffiziente Statistik der Likelihood $\tau^x(x)$ ist hier aber (noch) unbekannt, da es sich um die Likelihood bezüglich der ‚beobachteten' Werte von β und $\mathbf{\Sigma}$ handeln würde. Der ‚Beobachtung x' in (4.6) entspricht aber im Falle des Normal-Modells der Designmatrix $\mathbf{X}$ und der Zielvariable z. Im Folgenden soll daher die Stichprobe der Größe k, die in diese Likelihood eingeht, als eine k-dimensionale Stichprobe der Größe 1 (sozusagen in Bezug auf β) aufgefasst werden.

Die Likelihood aus Kapitel 4.2.1 ist eine Likelihood über z und daher auch nicht die passende Likelihood zur Ermittlung von $\tau^x(x)$. Der gesuchte Term $\tau^x(x) =: \tau(\mathbf{X}, z)$ lässt sich aber über die Analyse des Aufdatierungsschritts im Normal-Modell ermitteln, wie im Folgenden gezeigt wird.

Zur Überprüfung, ob sich das Normal-Modell als ein Stichproben-Modell in der Methodik von Quaeghebeur und de Cooman auffassen lässt, muss also gezeigt werden, dass die Dichte einer Normalverteilung mit aufdatierten Parametern $\beta^{(1)}$, $\mathbf{\Sigma}^{(1)}$ (siehe Kapitel 4.2.1) in die Form (4.6) übersetzt werden kann.

Analog zur Form der Priori muss gelten

$$y^{(1)} \;:=\; \frac{1}{n^{(1)}}\,\tau\left(\beta^{(1)}, \mathbf{\Sigma}^{(1)}\right) \;=\; \frac{1}{n^{(0)}+1}\,\tau\left(\beta^{(1)}, \mathbf{\Sigma}^{(1)}\right) ,$$

ebenso lässt sich wieder $\mathbf{c}\left(n^{(0)}+1, y^{(1)}\right)$ aus (4.6) mit $\mathbf{a}\left(\beta^{(1)}, \mathbf{\Sigma}^{(1)}\right)$ identifizieren:

$$\mathbf{a}\left(\beta^{(1)}, \mathbf{\Sigma}^{(1)}\right) \;=\; \mathbf{a}\left(\tau^{-1}\left[n^{(1)} \cdot y^{(1)}\right]\right)$$

Für jeden Summanden im Exponenten kann getrennt geprüft werden, ob der Aufdatierungsschritt im Normal-Modell einem Aufdatierungsschritt im Modell von Quaeghebeur und de Cooman entspricht, gleichzeitig lässt sich $\tau(\mathbf{X}, z)$ ermitteln.

Mit $\mathbf{\Lambda}^{(0)}$ an Stelle von ${\mathbf{\Sigma}^{(0)}}^{-1}$ lauten die in Kapitel 4.2.1 ermittelten Aufdatierungs-Regeln (4.2) und (4.3):

$$\begin{aligned} \beta^{(1)} &= \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Lambda}^{(0)}\right)^{-1} \left(\mathbf{X}^\mathsf{T} z + \mathbf{\Lambda}^{(0)}\beta^{(0)}\right) \\ \Lambda^{(1)} &= \mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Lambda}^{(0)} \end{aligned}$$

Zur Vereinfachung der Notation seien die Elemente von $\mathbf{X}$ folgendermaßen bezeichnet: $(\mathbf{X})_{l,i} =: x_{l,i}$; außerdem gilt folgender Zusammenhang, der für die dritte Gruppe der Summanden nötig ist:

$$\begin{aligned} \beta_j^{(1)} &= \left(\mathbf{\Lambda}^{(1)^{-1}} \left(\mathbf{X}^\mathsf{T} z + \mathbf{\Lambda}^{(0)}\beta^{(0)}\right)\right)_j \\ &= \sum_{i=1}^{p} \left(\mathbf{\Lambda}^{(1)^{-1}}\right)_{ji} \left(\mathbf{X}^\mathsf{T} z + \mathbf{\Lambda}^{(0)}\beta^{(0)}\right)_i \\ &= \sum_{i=1}^{p} \left(\mathbf{\Lambda}^{(1)^{-1}}\right)_{ji} \left(\sum_{l=1}^{k} x_{li} z_l + \sum_{h=1}^{p} \lambda_{ih}^{(0)} \beta_h^{(0)}\right) \end{aligned}$$

$$
\begin{array}{ccccc}
\underbrace{-\frac{\beta_1^2}{2\sigma^2}}_{\psi_1} \cdot \underbrace{\frac{1}{n^{(0)}}\lambda_{11}^{(0)}}_{y_1^{(0)}} & \xrightarrow{\text{Beobachtung}} & -\frac{\beta_1^2}{2\sigma^2} \cdot \underbrace{\frac{1}{n^{(1)}}\lambda_{11}^{(1)}}_{y_1^{(1)}} & = & \underbrace{-\frac{\beta_1^2}{2\sigma^2}}_{\psi_1} \cdot \frac{1}{n^{(0)}+1}\Big(n^{(0)} \cdot \underbrace{\frac{1}{n^{(0)}}\lambda_{11}^{(0)}}_{y_1^{(0)}} + \underbrace{\sum_{l=1}^{k} x_{l1}^2}_{\tau_1(\mathbf{X},z)}\Big) \\
\vdots \quad \vdots & \vdots & \vdots & & \vdots \quad \vdots \quad \vdots \\
\underbrace{-\frac{\beta_p^2}{2\sigma^2}}_{\psi_p} \cdot \underbrace{\frac{1}{n^{(0)}}\lambda_{pp}^{(0)}}_{y_p^{(0)}} & \longrightarrow & -\frac{\beta_p^2}{2\sigma^2} \cdot \underbrace{\frac{1}{n^{(1)}}\lambda_{pp}^{(1)}}_{y_p^{(1)}} & = & \underbrace{-\frac{\beta_p^2}{2\sigma^2}}_{\psi_p} \cdot \frac{1}{n^{(0)}+1}\Big(n^{(0)} \cdot \underbrace{\frac{1}{n^{(0)}}\lambda_{pp}^{(0)}}_{y_p^{(0)}} + \underbrace{\sum_{l=1}^{k} x_{lp}^2}_{\tau_p(\mathbf{X},z)}\Big) \\
\underbrace{-\frac{\beta_1\beta_2}{\sigma^2}}_{\psi_{p+1}} \cdot \underbrace{\frac{1}{n^{(0)}}\lambda_{12}^{(0)}}_{y_{p+1}^{(0)}} & \longrightarrow & -\frac{\beta_1\beta_2}{\sigma^2} \cdot \underbrace{\frac{1}{n^{(1)}}\lambda_{12}^{(1)}}_{y_{p+1}^{(1)}} & = & \underbrace{-\frac{\beta_1\beta_2}{\sigma^2}}_{\psi_{p+1}} \cdot \frac{1}{n^{(0)}+1}\Big(n^{(0)} \cdot \underbrace{\frac{1}{n^{(0)}}\lambda_{12}^{(0)}}_{y_{p+1}^{(0)}} + \underbrace{\sum_{l=1}^{k} x_{l1}\, x_{l2}}_{\tau_{p+1}(\mathbf{X},z)}\Big) \\
\vdots \quad \vdots & \vdots & & \ddots & \ddots \quad \ddots \\
\underbrace{-\frac{\beta_{p-1}\beta_p}{\sigma^2}}_{\psi_{\frac{p(p+1)}{2}}} \cdot \underbrace{\frac{1}{n^{(0)}}\lambda_{p-1,p}^{(0)}}_{y_{\frac{p(p+1)}{2}}^{(0)}} & \longrightarrow & -\frac{\beta_{p-1}\beta_p}{\sigma^2} \cdot \underbrace{\frac{1}{n^{(1)}}\lambda_{p-1,p}^{(1)}}_{y_{\frac{p(p+1)}{2}}^{(1)}} & = & \underbrace{-\frac{\beta_{p-1}\beta_p}{\sigma^2}}_{\psi_{\frac{p(p+1)}{2}}} \cdot \frac{1}{n^{(0)}+1}\Big(n^{(0)} \cdot \underbrace{\frac{1}{n^{(0)}}\lambda_{p-1,p}^{(0)}}_{y_{\frac{p(p+1)}{2}}^{(0)}} + \underbrace{\sum_{l=1}^{k} x_{l,p-1}\, x_{l,p}}_{\tau_{\frac{p(p+1)}{2}}(\mathbf{X},z)}\Big)
\end{array}
$$

$$
\begin{aligned}
\underbrace{-\frac{\beta_1}{\sigma^2}}_{\psi_{\frac{p(p+1)}{2}+1}} \cdot \underbrace{\frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{1j}^{(0)} \beta_j^{(0)}}_{y^{(0)}_{\frac{p(p+1)}{2}+1}} \xrightarrow{\text{Beobachtung}} & -\frac{\beta_1}{\sigma^2} \cdot \frac{1}{n^{(1)}} \sum_{j=1}^{p} \lambda_{1j}^{(1)} \beta_j^{(1)} \\
&= -\frac{\beta_1}{\sigma^2} \cdot \frac{1}{n^{(0)}+1} \sum_{j=1}^{p} \lambda_{1j}^{(1)} \sum_{i=1}^{p} \left(\mathbf{\Lambda}^{(1)^{-1}}\right)_{ji} \cdot \Big(\sum_{l=1}^{k} x_{li} z_l + \sum_{h=1}^{p} \lambda_{ih}^{(0)} \beta_h^{(0)} \Big) \\
&= -\frac{\beta_1}{\sigma^2} \cdot \frac{1}{n^{(0)}+1} \sum_{i=1}^{p} \underbrace{\sum_{j=1}^{p} \left(\mathbf{\Lambda}^{(1)}\right)_{1j} \left(\mathbf{\Lambda}^{(1)^{-1}}\right)_{ji}}_{\left[\begin{array}{ll} i=1: & \sum_{j=1}^{p} \left(\mathbf{\Lambda}^{(1)}\right)_{1j} \left(\mathbf{\Lambda}^{(1)^{-1}}\right)_{ji} = 1 \\ i \neq 1: & \sum_{j=1}^{p} \left(\mathbf{\Lambda}^{(1)}\right)_{1j} \left(\mathbf{\Lambda}^{(1)^{-1}}\right)_{ji} = 0 \end{array}\right]} \cdot \Big(\sum_{l=1}^{k} x_{li} z_l + \sum_{h=1}^{p} \lambda_{ih}^{(0)} \beta_h^{(0)} \Big) \\
&= -\frac{\beta_1}{\sigma^2} \cdot \frac{1}{n^{(0)}+1} \Big(\sum_{l=1}^{k} x_{l1} z_l + \sum_{h=1}^{p} \lambda_{1h}^{(0)} \beta_h^{(0)} \Big) \\
&= \underbrace{-\frac{\beta_1}{\sigma^2}}_{\psi_{\frac{p(p+1)}{2}+1}} \cdot \frac{1}{n^{(0)}+1} \Big(n^{(0)} \cdot \underbrace{\frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{1j}^{(0)} \beta_j^{(0)}}_{y^{(0)}_{\frac{p(p+1)}{2}+1}} + \underbrace{\sum_{l=1}^{k} x_{l1} z_l}_{\tau_{\frac{p(p+1)}{2}+1}(\mathbf{X}, z)} \Big)
\end{aligned}
$$

$$
\begin{array}{ccccccc}
\vdots & \vdots & \vdots & & \ddots & \ddots & \ddots \\
\underbrace{-\frac{\beta_p}{\sigma^2}}_{\psi_{\frac{p(p+3)}{2}}} \cdot \underbrace{\frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{pj}^{(0)} \beta_j^{(0)}}_{y^{(0)}_{\frac{p(p+3)}{2}}} & \longrightarrow & -\frac{\beta_p}{\sigma^2} \cdot \underbrace{\frac{1}{n^{(1)}} \sum_{j=1}^{p} \lambda_{pj}^{(1)} \beta_j^{(1)}}_{y^{(1)}_{\frac{p(p+3)}{2}}} & = & \underbrace{-\frac{\beta_p}{\sigma^2}}_{\psi_{\frac{p(p+3)}{2}}} \cdot \frac{1}{n^{(0)}+1} \Big(n^{(0)} \cdot \underbrace{\frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{pj}^{(0)} \beta_j^{(0)}}_{y^{(0)}_{\frac{p(p+3)}{2}}} + \underbrace{\sum_{l=1}^{k} x_{lp} z_l}_{\tau_{\frac{p(p+3)}{2}}(\mathbf{X}, z)} \Big)
\end{array}
$$

Damit ist gezeigt, dass sich das Normal-Modell für die lineare Regression als Aufdatierungs-Modell im Sinne von Quaeghebeur und de Cooman verstehen lässt und damit deren Mechanismus zur ‚Impräzisierung' und die einfache Aufdatierungsregel in Gleichung (4.6) nutzbar ist.

Dieser Nachweis war jedoch nur möglich, da für $\mathbf{b}(\psi) = 0$ galt, und somit die ‚künstliche' Einführung von $n^{(0)}$ im Exponenten keine Auswirkungen auf $\mathbf{b}(\psi)$ hatte. Andenfalls hätte sich bei der Überprüfung ein Problem ergeben: der letzte Summand im Exponenten, $n^{(0)}\mathbf{b}(\psi)$, wäre nach der Methodik von Quaeghebeur und de Cooman zu $(n^{(0)} + 1)\mathbf{b}(\psi)$ aufdatiert worden, hätte aber gemäß der Aufdatierungsregeln aus Kapitel 4.2.2 unverändert bleiben sollen, weshalb sich Widerspruch ergeben hätte. Im Falle des NIG-Modells tritt genau dieser Fall auf. Wie im Anhang in Abschnitt A.1 zu sehen ist, gilt dort $\mathbf{b}(\psi) \neq 0$.

Zur Übersicht noch einmal die $\frac{p(p+3)}{2}$-dimensionalen Größen ψ, $y^{(0)}$ und $\tau(\mathbf{X}, z)$ im Überblick:

$$\psi = \begin{bmatrix} -\frac{\beta_1^2}{2\sigma^2} \\ \vdots \\ -\frac{\beta_p^2}{2\sigma^2} \\ -\frac{\beta_1\beta_2}{\sigma^2} \\ \vdots \\ -\frac{\beta_{p-1}\beta_p}{\sigma^2} \\ -\frac{\beta_1}{\sigma^2} \\ \vdots \\ -\frac{\beta_p}{\sigma^2} \end{bmatrix}, \quad y^{(0)} = \begin{bmatrix} \frac{\lambda_{11}^{(0)}}{n^{(0)}} \\ \vdots \\ \frac{\lambda_{pp}^{(0)}}{n^{(0)}} \\ \frac{\lambda_{12}^{(0)}}{n^{(0)}} \\ \vdots \\ \frac{\lambda_{p-1,p}^{(0)}}{n^{(0)}} \\ \frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{1j}^{(0)} \beta_j^{(0)} \\ \vdots \\ \frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{pj}^{(0)} \beta_j^{(0)} \end{bmatrix}, \quad \tau(\mathbf{X}, z) = \begin{bmatrix} \sum_{l=1}^{k} x_{l1}^2 \\ \vdots \\ \sum_{l=1}^{k} x_{lp}^2 \\ \sum_{l=1}^{k} x_{l1}x_{l2} \\ \vdots \\ \sum_{l=1}^{k} x_{l,p-1}x_{l,p} \\ \sum_{l=1}^{k} x_{l1}z_l \\ \vdots \\ \sum_{l=1}^{k} x_{lp}z_l \end{bmatrix}$$

Die Elemente von $y^{(0)}$ und $\tau(\mathbf{X}, z)$ lassen sich zu einer schon bekannten Matrix bzw. einem Vektor umsortieren:

$$y^{(0)} = \begin{pmatrix} y_a^{(0)} \\ y_b^{(0)} \end{pmatrix} = \frac{1}{n^{(0)}} \begin{pmatrix} \mathbf{\Lambda}^{(0)} \\ \mathbf{\Lambda}^{(0)}\beta^{(0)} \end{pmatrix} \tag{4.7}$$

$$\tau(\mathbf{X}, z) = \begin{pmatrix} \tau_a(\mathbf{X}, z) \\ \tau_b(\mathbf{X}, z) \end{pmatrix} = \begin{pmatrix} \mathbf{X}^\mathsf{T}\mathbf{X} \\ \mathbf{X}^\mathsf{T}z \end{pmatrix} \tag{4.8}$$

$\mathcal{T}$ ist also die Vereinigung der Menge positiv semidefiniter $(p \times p)$-Matrizen und beliebiger p-dimensionaler Vektoren. $\mathcal{Y}^{(0)}$ kann auch hier als die konvexe Hülle von $\mathcal{T}$ ohne ‚Rand' angesehen werden, da $y_a^{(0)}$ eine beliebige positiv definite Matrix sein kann.

4.2.4 Die ‚Impräzisierung' des Normal-Modells

Die ‚Impräzisierung' erfolgt über die Variation von $y^{(0)}$ bei fest gewähltem $n^{(0)}$. Die Menge $\mathcal{Y}^{(0)}$ der Priori-Werte $y^{(0)}$ soll so gewählt werden, dass sie alle Priori-Informationen über den Regressionsparameter β widerspiegelt. $\mathcal{Y}^{(0)}$ definiert dann die Struktur $\mathcal{M}^{(0)}$, die Menge der klassischen Wahrscheinlichkeiten, die mit den Priori-Informationen über β vereinbar sind.

Die Menge $\mathcal{Y}^{(0)}$ der Parameter $y^{(0)}$ muss aber, wie in Kapitel 3.3 erläutert, beschränkt werden, damit ausgeschlossen werden kann, dass die Posteriori-Inferenz auf der Basis von $\mathcal{Y}^{(1)}$ aussagelos bleibt. Nach [Quaeghebeur und de Cooman 2005] kann die Menge $\mathcal{Y}^{(0)}$ einer multivariaten Normalverteilung, auf die Notation des hier behandelten Falles übersetzt, folgendermaßen sinnvoll beschränkt werden:

$$\frac{1}{n^{(0)}}\mathbf{\Lambda}^{(0)} \quad \text{positiv definit} \tag{4.9}$$

$$\text{und} \qquad \frac{1}{n^{(0)}}\left(\mathbf{\Lambda}^{(0)} - \frac{1}{n^{(0)}}\mathbf{\Lambda}^{(0)}\beta^{(0)}\beta^{(0)\mathsf{T}}\mathbf{\Lambda}^{(0)}\right) \quad \text{positiv definit} \tag{4.10}$$

Soll das Normal-Modell als Intervallwahrscheinlichkeitsmodell angewendet werden, so muss daher folgendermaßen vorgegangen werden:

1. Das Priori-Wissen über β und $\mathbf{\Lambda}$ muss als eine Menge von $\beta^{(0)}$ und $\mathbf{\Lambda}^{(0)}$ ausgedrückt werden.

2. Diese Menge muss in die Form von $y^{(0)}$ übersetzt werden, wobei beachtet werden muss, dass die so entstandene Menge $\mathcal{Y}^{(0)}$ von $y^{(0)}$-Werten die Bedingungen (4.9) und (4.10) erfüllt.

3. Jedes $y^{(0)}$ aus $\mathcal{Y}^{(0)}$ wird dann mittels (4.6) linear zu $y^{(1)}$ aufdatiert.

4. Die so erhaltene Menge $\mathcal{Y}^{(1)}$ muss darauf wieder in die interpretierbare Form von $\beta^{(1)}$ und $\mathbf{\Lambda}^{(1)}$ ‚rückübersetzt' werden.

Die in der obigen Aufstellung vorkommenden Mengen sollen dabei durch elementweise Unter- und Obergrenzen beschrieben werden, also beispielsweise für $\beta^{(0)}$ durch

$$\begin{array}{rcl} \beta_1^{(0)} & \in & \left[\underline{\beta}_1^{(0)}\,,\,\overline{\beta}_1^{(0)}\right] \\ \vdots & \vdots & \vdots \quad \vdots \\ \beta_p^{(0)} & \in & \left[\underline{\beta}_p^{(0)}\,,\,\overline{\beta}_p^{(0)}\right]. \end{array} \tag{4.11}$$

Bei $\beta^{(0)}$ können diese Grenzen prinzipiell unabhängig voneinander gewählt werden, im Falle von $\mathbf{\Lambda}^{(0)}$ jedoch nicht; auf die notwendigen Einschränkungen wird in den folgenden Absätzen eingegangen.

Wenn als Ausgangspunkt statt der Varianz-Kovarianz-Matrix $\mathbf{\Sigma}^{(0)}$ deren Inverse $\mathbf{\Lambda}^{(0)}$, auch als Präzisionsmatrix bekannt, verwendet wird, vereinfachen sich die Übersetzungsschritte 2. und 4. erheblich. Die Einträge einer Präzisionsmatrix $\mathbf{\Lambda}$ zu einem Zufallsvektor β lassen sich nach [Reithinger 2006] folgendermaßen interpretieren:

- Die Diagonalelemente λ_{ii} entsprechen invertierten partiellen Varianzen,

$$\lambda_{ii} = \frac{1}{\mathrm{var}(\beta_i \mid \beta_{\setminus i})}$$

  wobei $\beta_{\setminus i}$ den Vektor β ohne die i-te Komponente bezeichnet, so dass λ_{ii} also die partielle Varianz von β_i unter der Berücksichtigung des linearen Effekts der anderen Variablen in β ist.

- Die Elemente neben der Diagonalen entsprechen der mit den zugehörigen partiellen Varianzen skalierten negativen partiellen Korrelation,

$$\lambda_{ij} = -\sqrt{\lambda_{ii}\lambda_{jj}} \cdot \rho(\beta_i, \beta_j \mid \beta_{\setminus ij})$$

  wobei $\beta_{\setminus ij}$ den Vektor ohne die Komponenten i und j, sowie $\rho(X, Y \mid Z)$ die Korrelation von X und Y unter Berücksichtigung des linearen Effekts von Z bezeichnet. Diese ist folgendermaßen definiert:

$$\rho(X, Y \mid Z) = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}$$

  Dabei ist ρ_{XY} die (gewöhnliche) marginale Korrelation von X und Y. (Es sei daran erinnert, dass auch der marginale Korrelationskoeffizient prinzipiell nur den linearen Zusammenhang messen kann.)

Die Angabe der Varianz-Kovarianzstruktur in Form einer Präzisionsmatrix muss dabei nicht unbedingt von Nachteil sein; die hier anzugebenden partiellen Varianzen und Korrelationen stehen der Intuition unter Umständen sogar näher als deren ‚unbedingte' (marginale) Versionen, die die Einträge der Varianz-Kovarianz-Matrix darstellen. Tatsächlich entspricht die partielle Korrelation besser dem allgemein üblichen „ceteris paribus"-Denkschema, wohingegen die marginale Korrelation den Einfluss anderer Kovariablen ignoriert.[1]

[1] Ein Beispiel aus [Reithinger 2006] kann dies veranschaulichen:
Die Koeffizienten eines linearen Regressionsmodells stellen den Zusammenhang zwischen der Zielvariablen und den Prädiktoren dar. Diese entsprechen im Wesentlichen den partiellen Korrelationen. Ein Datensatz im Rahmen einer Mietspiegelermittlung liefert für das Modell

$$\text{Nettomiete} = \beta_0 + \ \text{Anzahl Räume} \cdot \beta_A + \varepsilon_{N,A}$$

einen Wert von $\hat{\beta}_A = 88.58$. Gültigkeit des Modells vorausgesetzt, bedeutet das, dass der Mietpreis einer Wohnung im Mittel um ca. 90 Euro steigt, wenn die Wohnung einen weiteren Raum hat. Nimmt man die Wohnfläche in das Regressionsmodell auf und betrachtet

$$\text{Nettomiete} = \beta_0 + \ \text{Anzahl Räume} \cdot \beta_A + \ \text{Wohnfläche} \cdot \beta_W + \varepsilon_{N,AW}\,.$$

Sollen die Mengen der Priori-Werte $\beta^{(0)}$ und $\mathbf{\Lambda}^{(0)}$ im Sinne von (4.11) als Intervalle angegeben werden (wobei es sich bei $\beta^{(0)}$ um ein p-dimensionales Intervall und bei $\mathbf{\Lambda}^{(0)}$ um ein $\frac{p(p+1)}{2}$-dimensionales Intervall handelt), stellt der Übersetzungsvorgang nach $\mathcal{Y}^{(0)}$ ein schwer lösbares Optimierungsproblem dar. Zur Definition von $\mathcal{Y}^{(0)}$ müssen die Unter- und Obergrenzen jeder der $\frac{p(p+3)}{2}$ Dimensionen von $y^{(0)}$ ermittelt werden, die sich einerseits durch die Vorgabe der Mengen der $\beta^{(0)}$ und $\mathbf{\Lambda}^{(0)}$, andererseits über die zu erfüllenden Bedingungen (4.9) und (4.10) ergeben.

Die Ermittlung der Unter- und Obergrenzen $\underline{y}_a^{(0)}$ und $\overline{y}_a^{(0)}$ ist noch relativ leicht, da sie sich elementweise direkt aus den elementweisen Unter- und Obergrenzen für $\mathbf{\Lambda}^{(0)}$ ergeben. (Das ist der Grund für den Übergang zur Präzisionsmatrix $\mathbf{\Lambda}^{(0)}$.) Diese Grenzen $\underline{\lambda}_{ij}^{(0)}$ bzw. $\overline{\lambda}_{ij}^{(0)}$ müssen jedoch so beschaffen sein, dass für jeden festen Wert innerhalb der Grenzen einer Dimension sich Werte in allen anderen Dimensionen innerhalb ihrer jeweiligen Grenzen finden lassen, so dass die resultierende Matrix $\mathbf{\Lambda}^{(0)}$ immer noch die Bedingung (4.9) erfüllt, also positiv definit ist. Das gilt natürlich insbesondere für die Grenzen $\underline{\lambda}_{ij}^{(0)}$ und $\overline{\lambda}_{ij}^{(0)}$ selbst, also für Matrizen, die bezüglich einer Komponente minimiert bzw. maximiert sind.

Grenzen für die zweite Komponente $y_b^{(0)}$ zu ermitteln, wird hingegen wesentlich komplexer. Gesucht sind nämlich die elementweisen Minima und Maxima des Vektors $y_b^{(0)} = \frac{1}{n^{(0)}}\mathbf{\Lambda}^{(0)}\beta^{(0)}$ unter der Nebenbedingung (4.10), so dass für eine Komponente $i \in \{1, \ldots, p\}$ von $y_b^{(0)}$ gilt

$$\underline{y}_{bi}^{(0)} = \min_{\beta^{(0)}, \mathbf{\Lambda}^{(0)}} \frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{ij}^{(0)} \beta_j^{(0)} \tag{4.12}$$

$$\overline{y}_{bi}^{(0)} = \max_{\beta^{(0)}, \mathbf{\Lambda}^{(0)}} \frac{1}{n^{(0)}} \sum_{j=1}^{p} \lambda_{ij}^{(0)} \beta_j^{(0)} \tag{4.13}$$

$$\text{unter NB } \frac{1}{n^{(0)}} \left(\mathbf{\Lambda}^{(0)} - \frac{1}{n^{(0)}} \mathbf{\Lambda}^{(0)} \beta^{(0)} \beta^{(0)\mathsf{T}} \mathbf{\Lambda}^{(0)} \right) \text{ positiv definit}$$

Zu beachten ist hierbei, dass für die Optimierung von *einem* $y_{bi}^{(0)}$ alle Elemente von $\beta^{(0)}$ und die i-te Zeile von $\mathbf{\Lambda}^{(0)}$ eingehen. Besonders dadurch wird die Nebenbedingung brisant: es dürfen nur zulässige Kombinationen von $\beta^{(0)}$ und $\mathbf{\Lambda}^{(0)}$ bei der Minimierung bzw. Maximierung verwendet werden; die Zulässigkeit ergibt sich einerseits durch die Zugehörigkeit zur zuvor ausgesuchten Menge von Werten von $\beta^{(0)}$ und $\mathbf{\Lambda}^{(0)}$,

---

so erhält man $\hat{\beta}_A = -41.68$. Naiv interpretiert bedeutet das, dass der Mietpreis einer Wohnung im Mittel um etwa 40 Euro fällt, wenn die Wohnung einen weiteren Raum hat. Diese Interpretation enspricht (wie beim ersten Regressionsmodell) der einer marginalen Korrelation; sie ist in diesem Modell, bei dem eine weitere Enflussgröße berücksichtigt ist, jedoch unzulässig verkürzt. Die Nettomiete nimmt mit wachsender Anzahl von Räumen nur ab, wenn die Wohnfläche gleich bleibt („ceteris paribus“), genauer: wenn der lineare Effekt der Wohnfläche berücksichtigt wird.

darüberhinaus aber auch durch die Nebenbedingung.

Diese Nebenbedingung kann dabei als eine nichtlineare Nebenbedingung (über das Kriterium der Positivität aller Eigenwerte der Matrix) angesehen werden. Die Minimierung bezüglich einer solchen Nebenbedingung kann im Falle eines hochdimensonalen β daher nur numerisch geschehen.

Um einen Eindruck von der Funktionsweise des Modells zu bekommen und um konkrete Ergebnisse zu erhalten, soll daher im nächsten Abschnitt die Analyse auf eine Situation mit zwei abhängigen Variablen beschränkt werden.

4.3 Formulierung des Normal-Modells im Fall $p = 2$

Im folgenden sei die Dimension p des Regressionsparameters β gleich zwei. Die Designmatrix ist somit von der Dimension $(k \times 2)$, es seien also k Beobachtungen von zwei Regressoren vorhanden, deren Einfluss auf die Responsevariable z linear modelliert werden soll.

Dieser Fall zeichnet sich dadurch aus, dass $\mathbf{\Lambda}$ die Dimension (2×2) hat und somit die Eigenschaft der positiven Definitheit (im Folgenden in Rechnungen mit ‚p.d.' abgekürzt) noch analytisch einfach fassbar ist, da für das Kriterium, dass alle Eigenwerte der Matrix positiv sein müssen, nur eine quadratische Gleichung zu lösen ist. Daher lässt sich das Nichtdiagonalelement in eine einfache Beziehung zu den zwei Diagonalelementen setzen, mit welcher dann gesichert werden kann, dass die Matrix positiv definit ist.

Es gilt für beliebige Skalare a, b, d:

$$\begin{pmatrix} a & b \\ b & d \end{pmatrix} \text{ p.d.} \quad \Longleftrightarrow \quad \text{Eigenwerte } \lambda_{1/2} > 0$$

Die Eigenwerte $\lambda_{1/2}$ sind erhältlich als die Lösungen von $(a-\lambda)(d-\lambda) - b^2 = 0$, so dass aus

$$\lambda^2 - (a+d)\lambda + ad - b^2 \;=\; 0$$

folgt

$$\lambda_{1/2} \;=\; \frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad-b^2)}}{2}\,.$$

Diese Eigenwerte sollen nun beide größer Null sein:

$$\begin{aligned}
\frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad-b^2)}}{2} &> 0 \\
\iff \quad (a+d) &> \mp\sqrt{(a+d)^2 - 4(ad-b^2)} \\
\iff \quad (a+d)^2 &> (a+d)^2 - 4(ad-b^2) \\
\iff \quad 4(ad-b^2) &> 0 \\
\iff \quad ad - b^2 &> 0 \\
\iff \quad ad &> b^2
\end{aligned} \tag{4.14}$$

Damit eine Matrix $\begin{pmatrix} a & b \\ b & d \end{pmatrix}$ positiv definit ist, darf also das Nichtdiagonalelement betragsmäßig im Verhältnis zu den Diagonalelementen nicht zu groß sein; es gilt:

$$-\sqrt{ad} < b < +\sqrt{ad}$$

Im Falle einer Varianz-Kovarianzmatrix für zwei Zufallsvariablen X und Y bedeutet dies:

$$\begin{aligned}
& -\sqrt{\mathbb{V}(X) \cdot \mathbb{V}(Y)} < \mathrm{Cov}(X,Y) < +\sqrt{\mathbb{V}(X) \cdot \mathbb{V}(Y)} \\
\iff \quad & -1 < \rho_{XY} < 1
\end{aligned}$$

Die Bedingung der positiven Definitheit bedeutet hier also nichts anderes, als dass das Nichtdiagonalelement eine Kovarianz darstellen muss, die zu einer zulässigen und nicht-deterministischen Korrelation führt.

Leider ist die Situation auch im Falle von nur zwei Regressoren noch immer sehr komplex. Wenn man $\beta^{(0)}$ und $\mathbf{\Lambda}^{(0)}$ voll parametrisiert (mit den fünf Parametern $\beta_1^{(0)}$, $\beta_2^{(0)}$, $\lambda_{11}^{(0)}$, $\lambda_{12}^{(0)}$, $\lambda_{22}^{(0)}$, also $q = 5$) und die Nebenbedingung (4.10) zu einer Bedingung in diesen fünf Parametern umformt, erhält man eine Ungleichung in all diesen fünf Parametern (und zusätzlich in $n^{(0)}$), die sich als wenig aufschlussreich erweist. Dass dabei $\lambda_{11}^{(0)}$, $\lambda_{12}^{(0)}$ und $\lambda_{22}^{(0)}$ selbst auch eine positiv definite Matrix generieren müssen, vereinfacht diese Ungleichung leider auch nicht weiter.

Daher soll die Situation noch ein klein wenig vereinfacht werden: $\mathbf{\Lambda}^{(0)}$ wird nicht voll parametrisiert, stattdessen sei eine Korrelationsstruktur K angenommen, die durch ρ, die Korrelation zwischen den beiden Regressoren, charakterisiert wird. Auch diese Strukturannahme hat noch eine genügend komplexe Nebenbedingung (4.10) zur Folge. Sei also

$$\beta \sim \mathrm{N}_2\left(\beta^{(0)}, \sigma^2 \frac{1}{a} K\right),$$

wobei $\beta^{(0)} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$, $a > 0$ zunächst fest und nicht näher bestimmt seien[2],

sowie $K = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ die Korrelationsstruktur vorgibt

und σ^2 gemäß der Modellannahmen fest gegeben ist.

Dann gilt

$$\Lambda^{(0)} = a \cdot K^{-1} = \frac{a}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}. \tag{4.15}$$

Die Nebenbedingungen, die in das Minimierungs- bzw. Maximierungsproblem zur Ermittlung von $y^{(0)}$ eingehen, sollen nun analysiert werden. Die erste Bedingung (4.9), also $\frac{1}{n^{(0)}}\Lambda^{(0)}$ positiv definit, ist erfüllt für jede beliebige zulässige nicht-deterministische Korrelation ($-1 < \rho < +1$).

Zur Ermittlung von Bedingungen für ρ, a, β_1, β_2 und $n^{(0)}$, die gewährleisten, dass die zweite Nebenbedingung (4.10) erfüllt ist, kann das Kriterium (4.14) entweder direkt auf die Matrix aus (4.10) oder mit Hilfe eines Satzes auf eine weniger komplexe Matrix angewendet werden:

4.3.1 Direkte Prüfung

Zur Ermittlung der Form der Bedingung (4.10) muss zur direkten Anwendung des Kriteriums (4.14) die Matrix aus (4.10) konkret berechnet werden. Dabei kann der erste Faktor $\frac{1}{n^{(0)}}$ wegfallen, da er bei einer Prüfung der positiven Definitheit keine Rolle spielt.

[2]Der obere Index $^{(0)}$ der Komponenten von $\beta^{(0)}$ sei wie hier in den folgenden Berechnungen zur Vereinfachung der Notation unterdrückt.

$$
\begin{aligned}
&\Lambda^{(0)} - \frac{1}{n^{(0)}}\Lambda^{(0)}\beta^{(0)}\beta^{(0)\mathsf{T}}\Lambda^{(0)} \\
&\quad = \frac{a}{1-\rho^2}\left[\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} \beta_1^2 & \beta_1\beta_2 \\ \beta_1\beta_2 & \beta_2^2 \end{pmatrix}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\right] \\
&\quad = \frac{a}{1-\rho^2}\left[\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} \beta_1^2 - \rho\beta_1\beta_2 & \beta_1\beta_2 - \rho\beta_1^2 \\ \beta_1\beta_2 - \rho\beta_2^2 & \beta_2^2 - \rho\beta_1\beta_2 \end{pmatrix}\right] \\
&\quad = \frac{a}{1-\rho^2}\left[\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\begin{pmatrix} \beta_1^2 - \rho\beta_1\beta_2 - \rho\beta_1\beta_2 + \rho^2\beta_2^2 & \beta_1\beta_2 - \rho\beta_1^2 - \rho\beta_2^2 + \rho^2\beta_1\beta_2 \\ \rho^2\beta_1\beta_2 - \rho\beta_1^2 + \beta_1\beta_2 - \rho\beta_2^2 & \rho^2\beta_1^2 - \rho\beta_1\beta_2 + \beta_2^2 - \rho\beta_1\beta_2 \end{pmatrix}\right] \\
&\quad = \frac{a}{1-\rho^2}\left[\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\begin{pmatrix} (\beta_1 - \rho\beta_2)^2 & (1+\rho^2)\beta_1\beta_2 - \rho(\beta_1^2 + \beta_2^2) \\ (1+\rho^2)\beta_1\beta_2 - \rho(\beta_1^2 + \beta_2^2) & (\rho\beta_1 - \beta_2)^2 \end{pmatrix}\right] \\
&\quad = \frac{a}{1-\rho^2}\begin{pmatrix} 1 - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}(\beta_1 - \rho\beta_2)^2 & -\rho - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\left[(1+\rho^2)\beta_1\beta_2 - \rho(\beta_1^2 + \beta_2^2)\right] \\ -\rho - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\left[(1+\rho^2)\beta_1\beta_2 - \rho(\beta_1^2 + \beta_2^2)\right] & 1 - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}(\rho\beta_1 - \beta_2)^2 \end{pmatrix}
\end{aligned}
$$

Da der Faktor $\frac{a}{1-\rho^2} > 0$ ist, kann auch er bei der Prüfung weggelassen werden. Es muss also gelten:

$$
\begin{aligned}
&\left(-\rho - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\left[(1+\rho^2)\beta_1\beta_2 - \rho(\beta_1^2 + \beta_2^2)\right]\right)^2 \;<\; \left(1 - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}(\beta_1 - \rho\beta_2)^2\right)\left(1 - \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}(\rho\beta_1 - \beta_2)^2\right) \\
&\quad \Longleftrightarrow \\
&\left(\rho(\beta_1^2 + \beta_2^2) - (1+\rho^2)\beta_1\beta_2 - \rho(1-\rho^2)\frac{n^{(0)}}{a}\right)^2 \;<\; \left((1-\rho^2)\frac{n^{(0)}}{a} - (\beta_1 - \rho\beta_2)^2\right)\left((1-\rho^2)\frac{n^{(0)}}{a} - (\rho\beta_1 - \beta_2)^2\right) \quad (4.16)
\end{aligned}
$$

Die so erhaltene Gleichung (4.16), die die Nebenbedingung bei der Minimierung (4.12) bzw. Maximierung (4.13) darstellt, ist wenig anschaulich. Daher sei im Folgenden der Versuch unternommen, über einen anderen Weg zu einer etwas einfacheren Gleichung zu gelangen.

4.3.2 Prüfung mittels eines Hilfssatzes

Satz A 51 aus [Toutenburg (2003)] bietet ein Kriterium für die positiv Semidefinitheit einer Matrix, die als Differenz zweier Matrizen entsteht. In der Art modifiziert, dass er ein Kriterium für positiv Definitheit bietet (und nach Korrektur eines Fehlers, der offensichtlich durch eine Vertauschung in Satz A 50 verursacht wird), lautet der Satz folgendermaßen:

Seien A und B zwei symmetrische Matrizen der gleichen Dimension $(p \times p)$, mit A positiv definit und B positiv semidefinit. Dann gilt

$$A - B \text{ positiv definit} \quad \Longleftrightarrow \quad \lambda_i^A(B) < 1 \quad \forall i = 1, \ldots, p\,.$$

Dabei sind $\lambda_i^A(B)$ die Eigenwerte von B bezüglich der Metrik von A. Sie sind die Lösungen von

$$\det(B - \lambda A) = 0\,.$$

Dieser Satz ist anwendbar auf (4.10), da $\Lambda^{(0)}$ gemäß Definition (4.15) positiv definit ist. $\frac{1}{n^{(0)}}\Lambda^{(0)}\beta^{(0)}\beta^{(0)\mathsf{T}}\Lambda^{(0)}$ ist positiv semidefinit, da $\beta^{(0)}\beta^{(0)\mathsf{T}}$ für jedes $\beta^{(0)}$ positiv semidefinit und $\Lambda^{(0)}$ regulär ist.

Sei nun also (wieder kann der Faktor $\frac{1}{n^{(0)}}$ ignoriert werden)

$$\begin{aligned}
A &= \Lambda^{(0)} \\
&= \frac{a}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \quad \text{und} \\
B &= \frac{1}{n^{(0)}}\Lambda^{(0)}\beta^{(0)}\beta^{(0)\mathsf{T}}\Lambda^{(0)} \\
&= \frac{1}{n^{(0)}}\left(\frac{a}{1-\rho^2}\right)^2 \begin{pmatrix} (\beta_1 - \rho\beta_2)^2 & (1+\rho^2)\beta_1\beta_2 - \rho(\beta_1^2+\beta_2^2) \\ (1+\rho^2)\beta_1\beta_2 - \rho(\beta_1^2+\beta_2^2) & (\rho\beta_1 - \beta_2)^2 \end{pmatrix}.
\end{aligned}$$

$\lambda_1^A(B)$ und $\lambda_2^A(B)$ sind dann die Lösungen von

$$\det\left[\frac{a}{1-\rho^2}\begin{pmatrix} \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}(\beta_1-\rho\beta_2)^2-\lambda & \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}[(1+\rho^2)\beta_1\beta_2-\rho(\beta_1^2+\beta_2^2)]-\lambda\rho \\ \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}[(1+\rho^2)\beta_1\beta_2-\rho(\beta_1^2+\beta_2^2)]-\lambda\rho & \frac{1}{n^{(0)}}\frac{a}{1-\rho^2}(\rho\beta_1-\beta_2)^2-\lambda \end{pmatrix}\right]=0$$

Berechnung der Determinanten:

$$\begin{aligned}\det\Big[\cdots\Big] \;=\; & \left(\frac{a}{1-\rho^2}\right)^2\Bigg[\lambda^2-\lambda\frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\left((\beta_1-\rho\beta_2)^2+(\rho\beta_1-\beta_2)^2\right)\\ & \qquad +\frac{1}{n^{(0)^2}}\left(\frac{a}{1-\rho^2}\right)^2(\beta_1-\rho\beta_2)^2(\rho\beta_1-\beta_2)^2-\left(\frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\left[(1+\rho^2)\beta_1\beta_2-\rho(\beta_1^2+\beta_2^2)\right]+\lambda\rho\right)^2\Bigg]\end{aligned}$$

Der Vorfaktor $\left(\frac{a}{1-\rho^2}\right)^2$ spielt beim Nullsetzen keine Rolle, also sind die Lösungen gesucht von

$$\begin{aligned}&\lambda^2(1-\rho^2)-\lambda\left(\frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\Big((\beta_1-\rho\beta_2)^2+(\rho\beta_1-\beta_2)^2\Big)+2\rho\frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\left[(1+\rho^2)\beta_1\beta_2-\rho(\beta_1^2+\beta_2^2)\right]\right)\\ &\quad+\quad\frac{1}{n^{(0)^2}}\left(\frac{a}{1-\rho^2}\right)^2(\beta_1-\rho\beta_2)^2(\rho\beta_1-\beta_2)^2-\frac{1}{n^{(0)^2}}\left(\frac{a}{1-\rho^2}\right)^2\left[(1+\rho^2)\beta_1\beta_2-\rho(\beta_1^2+\beta_2^2)\right]^2 \qquad =0\end{aligned}$$

$$\Longleftrightarrow$$

$$\begin{aligned}&\lambda^2(1-\rho^2)-\lambda\Bigg(\frac{1}{n^{(0)}}\frac{a}{1-\rho^2}\Big[\beta_1^2-2\rho\beta_1\beta_2+\rho^2\beta_2^2+\rho^2\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2+2\rho\beta_1\beta_2+2\rho^3\beta_1\beta_2-2\rho^2\beta_1^2-2\rho^2\beta_2^2\Big]\Bigg)\\ &\quad+\quad\frac{1}{n^{(0)^2}}\left(\frac{a}{1-\rho^2}\right)^2\Big[(\beta_1^2-2\rho\beta_1\beta_2+\rho^2\beta_2^2)(\rho^2\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2)\\ &\qquad\qquad-(1+\rho^2)^2\beta_1^2\beta_2^2+2\rho(1+\rho^2)\beta_1\beta_2(\beta_1^2+\beta_2^2)-\rho^2(\beta_1^2+\beta_2^2)^2\Big] \qquad =0\end{aligned}$$

Die erste eckige Klammer lässt sich folgendermaßen vereinfachen:

$$\beta_1^2 - 2\rho\beta_1\beta_2 + \rho^2\beta_2^2 + \rho^2\beta_1^2 - 2\rho\beta_1\beta_2 + \beta_2^2 + 2\rho\beta_1\beta_2 + 2\rho^3\beta_1\beta_2 - 2\rho^2\beta_1^2 - 2\rho^2\beta_2^2$$

$$\begin{aligned} &= -\rho^2\beta_1^2 - \rho^2\beta_2^2 + \beta_1^2 + \beta_2^2 - 2\rho\beta_1\beta_2 + 2\rho^3\beta_1\beta_2 \\ &= \beta_1^2(1-\rho^2) + \beta_2^2(1-\rho^2) - 2\rho\beta_1\beta_2(1-\rho^2) \\ &= (1-\rho^2)(\beta_1^2 - 2\rho\beta_1\beta_2 + \beta_2^2) \end{aligned}$$

Die zweite eckige Klammer lässt sich folgendermaßen vereinfachen:

$$(\beta_1^2 - 2\rho\beta_1\beta_2 + \rho^2\beta_2^2)(\rho^2\beta_1^2 - 2\rho\beta_1\beta_2 + \beta_2^2) - (1+\rho^2)^2\beta_1^2\beta_2^2 + 2\rho(1+\rho^2)\beta_1\beta_2(\beta_1^2 + \beta_2^2) - \rho^2(\beta_1^2 + \beta_2^2)^2$$

$$\begin{aligned} &= \rho^2\beta_1^4 - 2\rho\beta_1^3\beta_2 + \beta_1^2\beta_2^2 - 2\rho^3\beta_1^3\beta_2 + 4\rho^2\beta_1^2\beta_2^2 - 2\rho\beta_1\beta_2^3 + \rho^4\beta_1^2\beta_2^2 - 2\rho^3\beta\beta_2^3 + \rho^2\beta_2^4 \\ &\quad -\beta_1^2\beta_2^2 - \rho^2\beta_1^2\beta_2^2 + 2\rho\beta_1^3\beta_2 + 2\rho\beta_1\beta_2^3 + 2\rho^3\beta_1^3\beta_2 + 2\rho^3\beta_1\beta_2^3 - \rho^2\beta_1^4 - 2\rho^2\beta_1^2\beta_2^2 - \rho^2\beta_2^4 \\ &= \rho^2\beta_1^2\beta_2^2 - \rho^4\beta_1^2\beta_2^2 \\ &= \rho^2\beta_1^2\beta_2^2(1+\rho^2) \end{aligned}$$

Also sind $\lambda_1^A(B)$ und $\lambda_2^A(B)$ die Lösungen bezüglich λ von

$$\lambda^2(1-\rho^2) - \lambda\left(\frac{1}{n^{(0)}}\frac{a}{1-\rho^2}(1-\rho^2)(\beta_1^2 - 2\rho\beta_1\beta_2 + \beta_2^2)\right) + \frac{1}{n^{(0)^2}}\left(\frac{a}{1-\rho^2}\right)^2\rho^2\beta_1^2\beta_2^2(1+\rho^2) = 0$$

$$\iff \lambda^2(1-\rho^2) - \lambda\frac{a}{n^{(0)}}(\beta_1^2 - 2\rho\beta_1\beta_2 + \beta_2^2) + \frac{1}{n^{(0)^2}}\left(\frac{a}{1-\rho^2}\right)^2\rho^2\beta_1^2\beta_2^2(1+\rho^2) = 0$$

Diese Lösungen sind

$$\begin{aligned}
\lambda^A_{1/2}(B) &= \frac{1}{2(1-\rho^2)}\left[\frac{a}{n^{(0)}}\left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2\right)\pm\sqrt{\frac{a^2}{{n^{(0)}}^2}\left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2\right)^2-4(1-\rho^2)\frac{1}{{n^{(0)}}^2}\left(\frac{a}{1-\rho^2}\right)^2\rho^2\beta_1^2\beta_2^2(1+\rho^2)}\right] \\
&= \frac{1}{2(1-\rho^2)}\frac{a}{n^{(0)}}\left[\left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2\right)\pm\sqrt{\left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2\right)^2-4\rho^2\frac{1+\rho^2}{1-\rho^2}\beta_1^2\beta_2^2}\right].
\end{aligned}$$

Da gelten soll, dass $\lambda^A_{1/2}(B)<1$, muss also gelten:

$$\begin{aligned}
\left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2\right)\pm\sqrt{\left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2\right)^2-4\rho^2\frac{1+\rho^2}{1-\rho^2}\beta_1^2\beta_2^2} &< 2\frac{n^{(0)}}{a}(1-\rho^2) \\
\iff \left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2\right)^2-4\rho^2\frac{1+\rho^2}{1-\rho^2}\beta_1^2\beta_2^2 &< \left(2\frac{n^{(0)}}{a}(1-\rho^2)-\left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2\right)\right)^2 \\
\iff -4\rho^2\frac{1+\rho^2}{1-\rho^2}\beta_1^2\beta_2^2 &< 4\frac{{n^{(0)}}^2}{a^2}(1-\rho^2)^2-4\frac{n^{(0)}}{a}(1-\rho^2)\left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2\right) \\
\iff \rho^2\beta_1^2\beta_2^2 &> \frac{n^{(0)}}{a}\frac{(1-\rho^2)^2}{1+\rho^2}\left(\beta_1^2-2\rho\beta_1\beta_2+\beta_2^2-\frac{n^{(0)}}{a}(1-\rho^2)\right) \quad (4.17)
\end{aligned}$$

Auch dieser Weg führt zu einer nur unwesentlich anschaulicheren Form der Nebenbedingung (4.10). Daher soll die Situation durch die Betrachtung von speziellen Werten für ρ nochmals vereinfacht werden:

Für den Fall $\rho = 0$ ergibt sich sowohl über (4.16) als auch über (4.17) die Nebenbedingung

$$\beta_1^2 + \beta_2^2 < \frac{n^{(0)}}{a}\,.$$

Das bedeutet, dass der Quotient aus der Stärke des Vorwissens und der Präzision der Priori-Verteilung über β_1 und β_2 größer sein muss als $\beta_1^2 + \beta_2^2$. Die Varianz von β_1 und β_2, $\frac{1}{a}$, darf also nicht zu klein sein, wenn β_1 oder β_2 stark von Null abweichen sollen.

Untersucht man den Fall $\rho \to \pm 1$, so ergibt sich über den zweiten Weg direkt

$$\beta_1^2 \beta_2^2 > 0\,,$$

eine Bedingung, die für beliebige β_1, β_2 erfüllt ist.

Andere Werte für ρ führen zu keiner wirklichen Vereinfachung der Ausdrücke (4.17) oder (4.16).

4.3.3 Das Modell für $\rho = 0$

Nimmt man a priori $\rho = 0$ an, kann man ein einfach berechenbares Modell aufstellen, bei dem a im Nenner der Priori-Varianzen für $\beta_1^{(0)}$ und $\beta_2^{(0)}$ variieren kann und $\beta^{(0)}$ in einem zweidimensionalen Intervall variiert.

Dieses Modell sieht folgendermaßen aus:

$$\begin{aligned}
\beta \;\sim\; \mathrm{N}_2\left(\beta^{(0)},\, \sigma^2 \frac{1}{a} K\right)\,, \\
\text{wobei } \beta^{(0)} = \begin{pmatrix} \beta_1^{(0)} \\ \beta_2^{(0)} \end{pmatrix} \in \begin{pmatrix} B_1 = \left[\underline{b}_1, \overline{b}_1\right] \\ B_2 = \left[\underline{b}_2, \overline{b}_2\right] \end{pmatrix} = B\,, \\
\text{sowie } K = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ und } a \in A = [\underline{a}, \overline{a}] \text{ mit } \underline{a} > 0\,.
\end{aligned}$$

Mit der Festlegung

$$\overline{\underline{b}}_1 = \max\left\{|\underline{b}_1|, |\overline{b}_1|\right\} \quad \text{und} \quad \overline{\underline{b}}_2 = \max\left\{|\underline{b}_2|, |\overline{b}_2|\right\}$$

bleibt für eine korrekte Anwendung der Bayesregression unter komplexer Unsicherheit nur zu beachten, dass

$$\overline{a}\left(\left(\overline{\underline{b}}_1\right)^2 + \left(\overline{\underline{b}}_2\right)^2\right) \;<\; n^{(0)}\,. \tag{4.18}$$

In Abhängigkeit von der Einschätzung des Werts von $n^{(0)}$, der gewissermaßen das Ausmaß des Vertrauens in die Priori-Informationen darstellt, muss also versucht werden, die Grenzen $\underline{b}_1$, $\underline{b}_2$ und $\overline{a}$ so zu wählen, dass sie die Priori-Informationen über β angemessen widerspiegeln. Die Priori-Stärke $n^{(0)}$ muss also als entsprechend groß angesehen werden können, wenn die Präzision in den Annahmen hoch sein soll oder in Betracht gezogen werden soll, dass sich β weit von Null entfernt befindet. Kann $n^{(0)}$ hingegen nur als klein angesehen werden, müssen entweder die Intervalle für $\beta^{(0)}$ schmal und nahe bei Null oder die maximale Priori-Präzision $\overline{a}$ klein gewählt werden.

Wir erhalten also

$$\underline{y}_a^{(0)} = \frac{\underline{a}}{n^{(0)}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \overline{y}_a^{(0)} = \frac{\overline{a}}{n^{(0)}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\underline{y}_{bj}^{(0)} = \frac{\underline{ab}_j}{n^{(0)}}\,, \quad j = 1,2 \qquad \overline{y}_{bj}^{(0)} = \frac{\overline{ab}_j}{n^{(0)}}\,, \quad j = 1,2$$

wobei

$$\underline{ab}_j := \min_{a \in A,\, b_j \in B_j} a \cdot b \;=\; \min\left\{\underline{a} \cdot \underline{b}_j\,,\, \overline{a} \cdot \underline{b}_j\right\}$$

und

$$\overline{ab}_j := \max_{a \in A,\, b_j \in B_j} a \cdot b \;=\; \max\left\{\underline{a} \cdot \overline{b}_j\,,\, \overline{a} \cdot \overline{b}_j\right\}.$$

Der Aufdatierungsschritt ist gemäß (4.6) linear, was bedeutet, dass man von $\underline{y}^{(0)}$ direkt zu $\underline{y}^{(1)}$ und von $\overline{y}^{(0)}$ direkt auf $\overline{y}^{(1)}$ gelangt:

$$
\begin{aligned}
\underline{y}_a^{(1)} &= \frac{n^{(0)}\underline{y}_a^{(0)} + \tau_a(\mathbf{X}, z)}{n^{(0)}+1} \\
&= \frac{1}{n^{(0)}+1}\begin{pmatrix} \sum_{l=1}^k x_{l1}^2 + \underline{a} & \sum_{l=1}^k x_{l1}x_{l2} \\ \sum_{l=1}^k x_{l1}x_{l2} & \sum_{l=1}^k x_{l2}^2 + \underline{a} \end{pmatrix} = \frac{1}{n^{(1)}}\underline{\mathbf{\Lambda}}^{(1)} \\
\overline{y}_a^{(1)} &= \frac{n^{(0)}\overline{y}_a^{(0)} + \tau_a(\mathbf{X}, z)}{n^{(0)}+1} \\
&= \frac{1}{n^{(0)}+1}\begin{pmatrix} \sum_{l=1}^k x_{l1}^2 + \overline{a} & \sum_{l=1}^k x_{l1}x_{l2} \\ \sum_{l=1}^k x_{l1}x_{l2} & \sum_{l=1}^k x_{l2}^2 + \overline{a} \end{pmatrix} = \frac{1}{n^{(1)}}\overline{\mathbf{\Lambda}}^{(1)} \\
\underline{y}_b^{(1)} &= \frac{n^{(0)}\underline{y}_b^{(0)} + \tau_b(\mathbf{X}, z)}{n^{(0)}+1} \\
&= \frac{1}{n^{(0)}+1}\begin{pmatrix} \underline{ab}_1 + \sum_{l=1}^k x_{l1}z_l \\ \underline{ab}_2 + \sum_{l=1}^k x_{l2}z_l \end{pmatrix} \\
\overline{y}_b^{(1)} &= \frac{n^{(0)}\overline{y}_b^{(0)} + \tau_b(\mathbf{X}, z)}{n^{(0)}+1} \\
&= \frac{1}{n^{(0)}+1}\begin{pmatrix} \overline{ab}_1 + \sum_{l=1}^k x_{l1}z_l \\ \overline{ab}_2 + \sum_{l=1}^k x_{l2}z_l \end{pmatrix}
\end{aligned}
$$

$\underline{\mathbf{\Lambda}}^{(1)}$ und $\overline{\mathbf{\Lambda}}^{(1)}$ sind also direkt aus $\underline{y}_a^{(1)}$ und $\overline{y}_a^{(1)}$ durch Multiplikation mit $n^{(1)}$ erhältlich.

Möchte man auch $\underline{\mathbf{\Sigma}}^{(1)}$ bzw. $\overline{\mathbf{\Sigma}}^{(1)}$ aufgrund der gewohnten Interpretation erhalten, so muss wegen der komplexeren Struktur der Wert von a zur Minimierung bzw. Maximierung nicht notwendigerweise für alle Komponenten der gleiche sein; anders als bei $\mathbf{\Lambda}^{(1)}$ muss hier komponentenweise getrennt geprüft werden, welcher Wert von $a \in A$ die jeweilige Komponente minimiert bzw. maximiert.

Es gilt für ein $a \in A$

$$
\mathbf{\Sigma}^{(1)} =: \begin{pmatrix} \sigma_{11}^{(1)} & \sigma_{12}^{(1)} \\ \sigma_{12}^{(1)} & \sigma_{22}^{(1)} \end{pmatrix} = \begin{pmatrix} \sum_{l=1}^k x_{l1}^2 + a & \sum_{l=1}^k x_{l1}x_{l2} \\ \sum_{l=1}^k x_{l1}x_{l2} & \sum_{l=1}^k x_{l2}^2 + a \end{pmatrix}^{-1}
$$

und damit für die einzelnen Einträge von $\mathbf{\Sigma}^{(1)}$

$$\sigma_{11}^{(1)} = \frac{\sum_{l=1}^{k} x_{l2}^2 + a}{\left(\sum_{l=1}^{k} x_{l1}^2 + a\right)\left(\sum_{l=1}^{k} x_{l2}^2 + a\right) - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)^2} \tag{4.19}$$

$$\sigma_{22}^{(1)} = \frac{\sum_{l=1}^{k} x_{l1}^2 + a}{\left(\sum_{l=1}^{k} x_{l1}^2 + a\right)\left(\sum_{l=1}^{k} x_{l2}^2 + a\right) - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)^2} \tag{4.20}$$

$$\sigma_{12}^{(1)} = \frac{-\sum_{l=1}^{k} x_{l1}x_{l2}}{\left(\sum_{l=1}^{k} x_{l1}^2 + a\right)\left(\sum_{l=1}^{k} x_{l2}^2 + a\right) - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)^2}\,. \tag{4.21}$$

Durch Prüfung der Ableitung von $\sigma_{11}^{(1)}$ nach a kann gezeigt werden, dass die Posteriori-Varianz $\sigma_{11}^{(1)}$ des Regressionsparameters β_1 für $\overline{a}$ minimiert und für $\underline{a}$ maximiert wird:

$$\frac{\partial\,\sigma_{11}^{(1)}}{\partial\,a} = \frac{-\left(\sum_{l=1}^{k} x_{l2}^2 + a\right)^2 - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)^2}{\left(\left(\sum_{l=1}^{k} x_{l1}^2 + a\right)\left(\sum_{l=1}^{k} x_{l2}^2 + a\right) - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)^2\right)^2}$$

Dieser Term ist negativ für jeden beliebigen zulässigen Wert von a und für beliebige Werte der Komponenten von $\mathbf{X}^{\mathsf{T}}\mathbf{X}$. Die Posteriori-Varianz $\sigma_{11}^{(1)}$ ist also plausiblerweise eine monotone Funktion der Priori-Präzision a; Je größer die zuvor gewählte Präzision, desto kleiner die resultierende Posteriori-Varianz des Regressionsparameters β_1.

Für $\sigma_{22}^{(1)}$ gilt das gleiche Ergebnis (es muss im Zähler nur $\sum_{l=1}^{k} x_{l2}^2$ durch $\sum_{l=1}^{k} x_{l1}^2$ ersetzt werden), so dass auch $\sigma_{22}^{(1)}$ für $\overline{a}$ minimiert und für $\underline{a}$ maximiert wird.

Für die Posteriori-Kovarianz $\sigma_{12}^{(1)}$ des Regressionsparameters β spielt hingegen das Vorzeichen von $\sum_{l=1}^{k} x_{l1}x_{l2}$ eine entscheidende Rolle. Gilt also beispielsweise $\sum_{l=1}^{k} x_{l1}x_{l2} > 0$, so ist

$$\underline{\sigma}_{12}^{(1)} = \frac{-\sum_{l=1}^{k} x_{l1}x_{l2}}{\left(\sum_{l=1}^{k} x_{l1}^2 + \underline{a}\right)\left(\sum_{l=1}^{k} x_{l2}^2 + \underline{a}\right) - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)^2}$$

und

$$\overline{\sigma}_{12}^{(1)} = \frac{-\sum_{l=1}^{k} x_{l1}x_{l2}}{\left(\sum_{l=1}^{k} x_{l1}^2 + \overline{a}\right)\left(\sum_{l=1}^{k} x_{l2}^2 + \overline{a}\right) - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)^2}\,.$$

Im Falle von $\sum_{l=1}^{k} x_{l1}x_{l2} < 0$ wird $\sigma_{12}^{(1)}$ hingegen für $\overline{a}$ minimiert und für $\underline{a}$ maximiert.

Gilt also $\sum_{l=1}^{k} x_{l1}x_{l2} < 0$, so werden alle Komponenten von $\mathbf{\Sigma}^{(1)}$ für dasselbe $a \in A$ minimiert bzw. maximiert. In diesem Fall wird für $\overline{a}$ in allen Komponenten minimiert und für $\underline{a}$ in allen Komponenten maximiert.

Gilt aber $\sum_{l=1}^{k} x_{l1}x_{l2} > 0$, so besteht ein gegenläufiger Zusammenhang zwischen den Diagonal- und Nichtdiagonalelementen von $\mathbf{\Sigma}^{(1)}$: Sollen $\sigma_{11}^{(1)}$ und $\sigma_{22}^{(1)}$ minimiert werden, so wird damit gleichzeitig $\sigma_{12}^{(1)}$ maximiert und umgekehrt.

Die Ergebnisse aus (4.19) – (4.21) erhalten auch in den Grenzfällen der Wahl von a eine sinnvolle Interpretation: Aus $a \to 0$ folgt $\sigma^2\mathbf{\Sigma}^{(1)} \to \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$; je kleiner die Priori-Präzision gewählt wird, desto mehr entspricht die Posteriori-Varianz im Normal-Modell einer KQ-Schätzung der Varianz der Regressionsparameter. Aufgrund der Monotonität von $\sigma_{jj}^{(1)}$ in a gilt somit, dass die Posteriori-Varianz im Normal-Modell immer kleiner als die Varianz des KQ-Schätzers ist. Für $a \to \infty$ gilt $\mathbf{\Sigma}^{(1)} \to \mathbf{0}$, eine unendlich große Priori-Präzision hat natürlich eine unendlich kleine Posteriori-Varianz zur Folge.

Zum Erhalt von $\beta^{(1)}$ muss $y^{(1)}$ von links mit $n^{(1)}$ und $\mathbf{\Lambda}^{(1)^{-1}} = \mathbf{\Sigma}^{(1)}$ multipliziert werden.

Es gilt also

$$\underline{\beta}^{(1)} = \min_{a\in A, b\in B} n^{(1)}\mathbf{\Sigma}^{(1)}y_b^{(1)} \quad \text{und} \quad \overline{\beta}^{(1)} = \max_{a\in A, b\in B} n^{(1)}\mathbf{\Sigma}^{(1)}y_b^{(1)},$$

wobei

$$\beta_1^{(1)} = \frac{\left(\sum_{l=1}^{k} x_{l2}^2 + a\right)\left[a\cdot b_1 + \sum_{l=1}^{k} x_{l1}z_l\right] - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)\left[a\cdot b_2 + \sum_{l=1}^{k} x_{l2}z_l\right]}{\left(\sum_{l=1}^{k} x_{l1}^2 + a\right)\left(\sum_{l=1}^{k} x_{l2}^2 + a\right) - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)^2} \tag{4.22}$$

$$\beta_2^{(1)} = \frac{\left(\sum_{l=1}^{k} x_{l1}^2 + a\right)\left[a\cdot b_2 + \sum_{l=1}^{k} x_{l2}z_l\right] - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)\left[a\cdot b_1 + \sum_{l=1}^{k} x_{l1}z_l\right]}{\left(\sum_{l=1}^{k} x_{l1}^2 + a\right)\left(\sum_{l=1}^{k} x_{l2}^2 + a\right) - \left(\sum_{l=1}^{k} x_{l1}x_{l2}\right)^2} \tag{4.23}$$

Diese Ausdrücke sind linear in b_1 und in b_2. Da der Nenner immer positiv ist, gilt für die Minimierung und Maximierung von $\beta^{(1)}$:

$$\beta_1^{(1)} \to \max \qquad \text{für } b_1 \to \overline{b}_1 \text{ und } \begin{cases} b_2 \to \overline{b}_2 & \sum_{l=1}^{k} x_{l1}x_{l2} < 0 \\ b_2 \to \underline{b}_2 & \sum_{l=1}^{k} x_{l1}x_{l2} > 0 \end{cases} \tag{4.24}$$

$$\beta_1^{(1)} \to \min \qquad \text{für } b_1 \to \underline{b}_1 \text{ und } \begin{cases} b_2 \to \underline{b}_2 & \sum_{l=1}^{k} x_{l1}x_{l2} < 0 \\ b_2 \to \overline{b}_2 & \sum_{l=1}^{k} x_{l1}x_{l2} > 0 \end{cases} \tag{4.25}$$

und

$$\beta_2^{(1)} \to \max \qquad \text{für } b_2 \to \overline{b}_2 \text{ und } \begin{cases} b_1 \to \overline{b}_1 & \sum_{l=1}^{k} x_{l1}x_{l2} < 0 \\ b_1 \to \underline{b}_1 & \sum_{l=1}^{k} x_{l1}x_{l2} > 0 \end{cases} \tag{4.26}$$

$$\beta_2^{(1)} \to \min \qquad \text{für } b_2 \to \underline{b}_2 \text{ und } \begin{cases} b_1 \to \underline{b}_1 & \sum_{l=1}^{k} x_{l1}x_{l2} < 0 \\ b_1 \to \overline{b}_1 & \sum_{l=1}^{k} x_{l1}x_{l2} > 0 \end{cases} \tag{4.27}$$

Diese Minimierungen und Maximierungen sind unabhängig von a durchführbar, wenn die Parametergrenzen so gewählt werden, dass Gleichung (4.18) erfüllt wird.

Die Terme (4.22) und (4.23) sind Funktionen, die jeweils in Zähler und Nenner quadratisch in a sind. Anders als bei der Rückübersetzung von $y_a^{(1)}$ ist die Ableitung nach a jedoch nicht aufschlussreich, so dass keine analytischen Aussagen bezüglich einer eventuellen Monotonität von $\underline{\beta}^{(1)}$ und $\overline{\beta}^{(1)}$ (abhängig von gewissen Bedingungen) in a möglich sind. Da aber die zur Minimierung und Maximierung von (4.22) und (4.23) jeweils nötigen Werte von b_1 und b_2 bereits durch (4.24) – (4.27) gegeben sind, muss nur noch über $a \in A = [\underline{a}, \overline{a}]$ numerisch minimiert bzw. maximiert werden, um $\underline{\beta}^{(1)}$ und $\overline{\beta}^{(1)}$ zu erhalten.

Asymptotische Aussagen bezüglich a sind jedoch möglich. Aus einer Analyse von (4.22) und (4.23) folgt:

$$a \longrightarrow \infty \qquad \Longrightarrow \qquad \beta_j^{(1)} \longrightarrow b_j$$

und

$$a \longrightarrow 0 \qquad \Longrightarrow \qquad \beta_j^{(1)} \longrightarrow \hat{\beta}_j$$

Diese Ergebnisse sind höchst plausibel. Ein sehr großer Wert der Priori-Präzision a ist gleichbedeutend mit einer sehr niedrigen Priori-Varianz von β. Gleichzeitig impliziert die Wahl von einem sehr großen Wert von a über die Bedingung (4.18), dass das Vertrauen in diese Angaben sehr hoch sein muss, da eine solche Wahl einen großen Wert von $n^{(0)}$ zur Folge hat. Aufgrund dieses hohen Vertrauens in die Priori-Angabe von β werden die Informationen aus der Stichprobe weitgehend ignoriert, so dass das a posteriori-Ergebnis nur b_j, die Priori-Angabe für β_j sein kann. Im Falle einer intervallwertigen Priori-Angabe B_j ergibt sich für $a \to \infty$ also a posteriori wieder B_j.

Ein sehr kleiner Wert von a impliziert umgekehrt, dass die a priori-Varianz von β sehr groß ist und $n^{(0)}$ ebenfalls klein werden kann. Das mangelnde Vertrauen in die Priori-Angaben für β führt dann dazu, den Informationen aus der Stichprobe eine große Bedeutung beizumessen und daher zur klassischen Schätzung von β überzugehen, nämlich zur üblichen Schätzung nach der Methode der kleinsten Quadrate $\hat{\beta}_j$, die hier ja sowohl mit dem Ergebnis einer Likelihood-Schätzung als auch einer bayesianischen HPD-Schätzung bei einer konstanten Priori-Verteilung für β zusammen fällt.

Eine wichtige Fragestellung im Hinblick auf das Verhalten des Modells ist es, ob die Länge der Posteriori-Intervalle für die Parameter von der Beobachtung abhängt. Hängt die Differenz zwischen dem höchsten und dem niedrigsten Posteriori-Parameter nicht von den Daten ab, ist das Verhalten des Modells im Falle eines ‚prior-data conflict' unbefriedigend, da es dann bei der Anwendung auswirkungslos bleibt, wenn das Vorwissen nicht zu den Beobachtungen passt. Wie in Kapitel 1.4.3 schon ausgeführt, wäre es hingegen wünschenswert, dass das Modell im Falle eines ‚prior-data conflict' zu breiteren Intervallen führt.

Erwartungsgemäß und konform zu den Erkenntnissen in Kapitel 3.3 hängt die Länge der Posteriori-Intervalle nicht von der Beobachtung ab, wenn man die Differenz zwischen den oberen und unteren Werten von $y_a^{(1)}$ und $y_b^{(1)}$ berechnet; diese Intervalllängen sind unabhängig von $\tau(\mathbf{X}, z)$. Dieses Ergebnis wurde schon bei der Beschreibung des Modells von Quaeghebeur und de Cooman bei den Erläuterungen zu Gleichung (3.8) erwähnt. Bei den ‚rückübersetzten' Parametern $\mathbf{\Sigma}^{(1)}$ und $\beta^{(1)}$ ist diese Frage deutlich schwieriger zu beantworten:

Die Differenz der Posteriori-Varianzen von β hängt nämlich von den Werten von $\mathbf{X}$ ab. Mit der Vereinbarung von

$$\mathbf{x}_{11} := \sum_{l=1}^{k} x_{l1}^2\,, \quad \mathbf{x}_{12} := \sum_{l=1}^{k} x_{l1}x_{l2}\,, \quad \mathbf{x}_{22} := \sum_{l=1}^{k} x_{l2}^2\,, \quad \text{d.h. } \mathbf{X}^{\mathsf{T}}\mathbf{X} =: \begin{pmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} \\ \mathbf{x}_{12} & \mathbf{x}_{22} \end{pmatrix}$$

gilt beispielsweise für die Posteriori-Varianzen von β_1 und β_2, gemäß (4.19) bzw. (4.20),

$$\begin{aligned} \overline{\sigma}_{11}^{(1)} - \underline{\sigma}_{11}^{(1)} &= \frac{\mathbf{x}_{22} + \underline{a}}{(\mathbf{x}_{11} + \underline{a})\,(\mathbf{x}_{22} + \underline{a}) - \mathbf{x}_{12}^2} - \frac{\mathbf{x}_{22} + \overline{a}}{(\mathbf{x}_{11} + \overline{a})\,(\mathbf{x}_{22} + \overline{a}) - \mathbf{x}_{12}^2} \\ &= (\overline{a} - \underline{a}) \cdot \frac{(\mathbf{x}_{22} + \underline{a})(\mathbf{x}_{22} + \overline{a}) + \mathbf{x}_{12}^2}{\big((\mathbf{x}_{11} + \underline{a})\,(\mathbf{x}_{22} + \underline{a}) - \mathbf{x}_{12}^2\big)\big((\mathbf{x}_{11} + \overline{a})\,(\mathbf{x}_{22} + \overline{a}) - \mathbf{x}_{12}^2\big)} \quad (4.28) \\ \overline{\sigma}_{22}^{(1)} - \underline{\sigma}_{22}^{(1)} &= \frac{\mathbf{x}_{11} + \underline{a}}{(\mathbf{x}_{11} + \underline{a})\,(\mathbf{x}_{22} + \underline{a}) - \mathbf{x}_{12}^2} - \frac{\mathbf{x}_{11} + \overline{a}}{(\mathbf{x}_{11} + \overline{a})\,(\mathbf{x}_{22} + \overline{a}) - \mathbf{x}_{12}^2} \\ &= (\overline{a} - \underline{a}) \cdot \frac{(\mathbf{x}_{11} + \underline{a})(\mathbf{x}_{11} + \overline{a}) + \mathbf{x}_{12}^2}{\big((\mathbf{x}_{11} + \underline{a})\,(\mathbf{x}_{22} + \underline{a}) - \mathbf{x}_{12}^2\big)\big((\mathbf{x}_{11} + \overline{a})\,(\mathbf{x}_{22} + \overline{a}) - \mathbf{x}_{12}^2\big)} \quad (4.29) \end{aligned}$$

Ob sich die Priori-Differenz $(\overline{a} - \underline{a})$ verringert oder vergrößert, hängt also von dem Bruch in (4.28) bzw. (4.29) ab. Dieser enthält in beiden Fällen Elemente von $\mathbf{X}^{\mathsf{T}}\mathbf{X}$, aber nicht von z. Da in der Modellformulierung (siehe Abschnitt 4.2.1) nur z und nicht $\mathbf{X}$ als stochastisch angesehen wird, kann man trotz der Abhängigkeit von der Designmatrix $\mathbf{X}$ konstatieren, dass die Differenz der Posteriori-Varianzen unabhängig von der ‚zufallsunterworfenen' Beobachtung z ist, und dass somit das allgemeine Ergebnis aus Kapitel 3.3 auch für diesen ‚rückübersetzten' Parameter gilt.

Leider ist in der allgemeinen Form in (4.28) bzw. (4.29) nicht ersichtlich, welche Bedingungen erfüllt sein müssen, damit der Faktor, um den sich die Intervalllänge ändert, einen Wert größer als eins annimmt. Im allgemeinen Fall muss daher offen bleiben, ob diese Abhängigkeit von $\mathbf{X}$ bedeutet, dass das Modell im Falle eines ‚priori-data conflict' zu breiteren Intervallen für die Varianz von β führt oder nicht.

Vereinfacht man jedoch die Rahmenbedingungen und nimmt an, dass die Designmatrix $\mathbf{X}$ in der ersten Spalte nur Einser (zur Schätzung des Intercept) enthält und die Regressor-Variable in der zweiten Spalte standardisiert wurde, dann gilt

$$\mathbf{X}^{\mathsf{T}}\mathbf{X} = \begin{pmatrix} k & 0 \\ 0 & k-1 \end{pmatrix}$$

Mit den so vereinfachten Einträgen ergibt sich für die resultierenden Posteriori-Intervalllängen der Varianzen der Schätzverteilung für β_1 und β_2

$$\begin{aligned} \overline{\sigma}_{11}^{(1)} - \underline{\sigma}_{11}^{(1)} &= (\overline{a} - \underline{a}) \cdot \frac{1}{(k+\underline{a})(k+\overline{a})} \quad \text{bzw.} \\ \overline{\sigma}_{22}^{(1)} - \underline{\sigma}_{22}^{(1)} &= (\overline{a} - \underline{a}) \cdot \frac{1}{(k-1+\underline{a})(k-1+\overline{a})}\,; \end{aligned}$$

Der Faktor, um den sich diese Differenz bei der Aufdatierung ändert, ist nur von der Stichprobengröße k abhängig und für $k > 1$ stets kleiner als eins. Daraus folgt, dass sich die Intervalle für die Varianzen auch dann verkleinern, wenn ein ‚prior-data conflict' vorliegt. In diesem Fall gilt die allgemeine Erkenntnis aus Kapitel 3.3 leider auch für die Varianzschätzungen.

Auch die Differenz der Posteriori-Erwartungswerte von β ist abhängig von $\mathbf{X}$, aber nicht von z; somit kann man folgern, dass die Intervalllängen für die Posteriori-Parameter $\beta_1^{(1)}$ und $\beta_2^{(1)}$ ebenfalls unabhängig von der ‚zufallsunterworfenen' Beobachtung z sind. Nimmt man der Einfachheit halber $B_1 = B_2 = [\underline{b}, \overline{b}]$ an, gilt bei festem a (wie in Kapitel 4.4 ersichtlich, sind in den meisten Fällen $\underline{\beta}^{(1)}$ und $\overline{\beta}^{(1)}$ für den selben Wert von a erhältlich):

$$\begin{aligned} \overline{\beta}_1^{(1)} - \underline{\beta}_1^{(1)} &= (\overline{b} - \underline{b}) \cdot \frac{a \cdot (\mathbf{x}_{22} + a + |\mathbf{x}_{12}|)}{(\mathbf{x}_{11} + a)\,(\mathbf{x}_{22} + a) - \mathbf{x}_{12}^2} \\ \overline{\beta}_2^{(1)} - \underline{\beta}_2^{(1)} &= (\overline{b} - \underline{b}) \cdot \frac{a \cdot (\mathbf{x}_{11} + a + |\mathbf{x}_{12}|)}{(\mathbf{x}_{11} + a)\,(\mathbf{x}_{22} + a) - \mathbf{x}_{12}^2} \end{aligned}$$

Hier ist es für den allgemeinen Fall ebenfalls nicht möglich, den Faktor, um den sich die Spanne für den Erwartungswert von β bei der Aufdatierung ändert, in eine einfache Beziehung zum Vorliegen eines ‚priori-data conflict' zu setzen.

Wählt man jedoch die obigen Rahmenbedingungen, die zu einer vereinfachten Designmatrix $\mathbf{X}$ führen, so erhält man

$$\begin{aligned} \overline{\beta}_1^{(1)} - \underline{\beta}_1^{(1)} &= (\overline{b} - \underline{b}) \cdot \frac{a}{k+a} \quad \text{bzw.} \\ \overline{\beta}_2^{(1)} - \underline{\beta}_2^{(1)} &= (\overline{b} - \underline{b}) \cdot \frac{a}{k-1+a}\,; \end{aligned}$$

somit ist auch hier der Faktor, um den sich diese Differenz bei der Aufdatierung ändert, für eine Stichprobe mit mehr als einer einzigen Beobachtung stets kleiner als eins und nur von der Stichprobengröße k abhängig. Genau wie bei den Intervallen für die Varianzen folgt daraus, dass sich die Intervalle für $\beta^{(1)}$ auch dann verkleinern, wenn ein ‚prior-data conflict' vorliegt, so dass die allgemeine Erkenntnis aus Kapitel 3.3 auch in diesem Fall bestätigt werden muss.

Erstaunlicherweise sind auf den ersten Blick sowohl die Terme (4.19) – (4.21) (für die Rückübersetzung nach $\mathbf{\Sigma}^{(1)}$) als auch die Terme (4.22) und (4.23) (für die Rückübersetzung nach $\beta^{(1)}$) unabhängig von $n^{(0)}$ (und damit auch von $n^{(1)}$). Es scheint zunächst unsinnig, den Parameter, der die ‚Stärke' des Vorwissens modelliert, nicht in den genannten Termen zu finden; diese sind die Rückübersetzungen von $y^{(1)}$, was ja von $n^{(1)}$ abhängt, wie in (4.7) analog die Abhängigkeit von $y^{(0)}$ von $n^{(0)}$ zu sehen ist. Der Grund für die scheinbare Unabhängigkeit von $n^{(1)}$ liegt in der Vorgehensweise im Abschnitt 4.2.3: Es wurde gezeigt, dass sich das übliche konjugierte bayesianische Modell als ein Modell in der Notation von [Quaeghebeur und de Cooman 2005] verstehen lässt. Ausgangspunkt war deshalb nicht die Likelihood (wie in Kapitel 3.2), sondern direkt die konjugierte Verteilung. Deshalb musste der Parameter n gewissermaßen ‚künstlich' im Exponenten eingeführt werden und machte deshalb die Definition in Gleichung (4.5) nötig. Aus diesem Grund kürzt sich $n^{(1)}$ in allen Rückübersetzungen sofort wieder heraus und scheint damit zu ‚verschwinden'.

Tatsächlich hängen die Werte der Posteriori-Parameter $\beta^{(1)}$ und $\mathbf{\Sigma}^{(1)}$ aber sehr wohl von $n^{(0)}$ ab, nämlich über die Einschränkung für die Wahl der Mengen der Priori-Parameter $\beta^{(0)}$ und $\mathbf{\Sigma}^{(0)}$, die in Gleichung (4.18) gefordert wird. Diese Einschränkung wirkt auf $\left(\underline{b}_1\right)^2$ und $\left(\underline{b}_2\right)^2$, die absolut größten Werte der Unter- und Obergrenzen von $\beta_1^{(0)}$ und $\beta_2^{(0)}$, sowie auf $\overline{a}$, die maximale angenommene Präzision.

Bei der Anwendung des Modells sollte die Wahl der Priori-Grenzen für $\beta^{(0)}$ keine echte Schwierigkeit darstellen; die Wahl von $A = [\underline{a}, \overline{a}]$ ist im konkreten Fall trotz der Interpretation von $\frac{\sigma^2}{a}$ als Priori-Varianz der Normalverteilung über β_1 und β_2 nicht ganz so einfach.

Anhaltspunkte zur Ermittlung von A und die Inferenz mit diesem Modell sollen zur besseren Anschaulichkeit im nächsten Kapitel anhand von einfachen Simulationsbeispielen vorgeführt werden; in Kapitel 4.5 wird das Modell dann auf einen realen Datensatz angewendet.

4.4 Anwendung des Modells für $p = 2$, $\rho = 0$

Zur Veranschaulichung der Funktionsweise und zur Einschätzung der Aussagekraft sollen im Folgenden für drei einfache simulierte Datensätze Posteriori-Parameter und Kredibilitätsintervalle berechnet werden.

Die Vorgehensweise ist die folgende: Es werden zwei Regressionsvariablen x_1 und x_2 simuliert, standardisiert und in die Designmatrix X geschrieben. Da die Varianz des Fehlerterms im Normal-Modell als bekannt angesehen wird, muss σ^2 zu Anfangs fest gewählt und auf dieser Basis der Fehlerterm $\varepsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2)$ simuliert werden. Mit Hilfe der gewählten Werte von β_1 und β_2 wird dann die Responsevariable z berechnet und zentriert, so dass die Schätzung des Intercepts unnötig wird. Durch dieses Vorgehen sind auch die geschätzten Regressionskoeffizienten in der Stärke ihrer Wirkungsweise vergleichbar. Die Posteriori-Parameter und -Schätzungen werden dann gemäß der in Kapitel 4.3.3 beschriebenen Formeln berechnet.

Unter Abschnitt A.2 ist eine CD beigefügt, die die Programmierungs-Syntax enthält, mit der die Ergebnisse dieses Kapitels berechnet und die Schaubilder gezeichnet wurden.

Es werden drei Datensätze simuliert, die sich hinsichtlich der Stärke des Einflusses der Parameter auf die Zielvariable, die Varianz des Störterms und der Abhängigkeit zwischen den Regressoren unterscheiden. In Abschnitt 4.4.1 wird ein simulierter Datensatz mit relativ großen Regressionsparametern und kleiner Varianz untersucht, wobei die Regressoren unabhängig simuliert werden; im simulierten Datensatz in Abschnitt 4.4.2 sind die Parameter kleiner und dafür die Varianz größer; beim simulierten Datensatz in Abschnitt 4.4.3 sind hingegen die Regressoren stark korreliert.

4.4.1 ‚Große' Koeffizienten, kleine Varianz

In diesem ersten Beispiel besitzen die Regressionskoeffizienten mit $\beta_1 = 1.5$ und $\beta_2 = 1$ einen relativ großen Wert, die Varianz des Fehlerterms ist mit $\sigma^2 = 0.5$ dagegen relativ klein. Der erzeugte Datensatz soll nur 20 Beobachtungen haben, da insbesondere bei kleinen Datensätzen die Verwendung eines Intervallwahrscheinlichkeitsmodells angebracht scheint, weil dann die Asymptotik-Aussagen, die oft Vorraussetzung für klassische Modellierungen sind, nicht oder nur eingeschränkt gelten. Für den mit diesen Angaben simulierten Datensatz ergibt sich ein Wert von 0.8356 für das multiple adjustierte R^2.

Mit den folgenden Abbildungen soll illustriert werden, wie sich die Wahl von $A = [\underline{a}, \overline{a}]$ sowie von $B_1 = [\underline{b}_1, \overline{b}_1]$ und $B_2 = [\underline{b}_2, \overline{b}_2]$ auf die Spanne der Posteriori-Erwartungswerte von β auswirkt.

In diesen Abbildungen gibt es jeweils ein Schaubild für die resultierende Spanne von $\beta_1^{(1)}$ und von $\beta_2^{(1)}$. Da es sich bei diesen Werten um die Spanne der Posteriori-Erwartungswerte handelt, können diese als intervallwertige Punktschätzungen der gesuchten Parameter β_1 und β_2 angesehen werden. Da die Posteriori-Verteilungen Normalverteilungen sind, fallen Erwartungswert und Modus zusammen, so dass diese Intervalle auch einer HPD-Punktschätzung entsprechen.

Die durchgezogenen Linien in horizontalem Verlauf entsprechen dem Wert von $\beta_j^{(1)}$, $j = 1, 2$, in Abhängigkeit von a; die obere entsteht durch das Einsetzen der Werte aus B_1 und B_2, die $\beta_j^{(1)}$ maximieren, die untere durch das Einsetzen der Werte aus B_1 und B_2, die $\beta_j^{(1)}$ minimieren (siehe Gleichungen (4.24) – (4.27)). Für gegebene Unter- und Obergrenzen von A, die durch die senkrechten durchgezogenen Linien markiert werden, ergibt sich somit $\underline{\beta}_j^{(1)}$ als der niedrigste und $\overline{\beta}_j^{(1)}$ als der höchste Schnittpunkt einer Senkrechten mit einer horizontal verlaufenden durchgezogenen Linie. Als Anhaltspunkt ist der ‚echte' Wert von β_j als dick gestrichelte waagrechte Linie eingezeichnet; zum Vergleich ist die Schätzung nach der Methode der kleinsten Quadrate dünn gestrichelt, sowie die Unter- und Obergrenze des zugehörigen klassischen Konfidenzintervalls mit Linien aus dünnen Strichen und Punkten eingetragen.

Für die Wahl von A werden verschiedene Vorgehensweisen untersucht. In einem ersten Versuch soll ein ‚zentraler' Wert für a gewissermaßen datengeleitet ermittelt werden, indem die angenommene Priori-Varianz für β_1 und β_2 mit der Schätzung dieser Varianzen im normalen linearen Modell gleichgesetzt wird; daraus lässt sich ein Wert für a ableiten:

$$\begin{aligned} \mathrm{Var}^{KQ}(\hat{\beta}_j) &= \sigma^2 \left((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)_{jj} \\ \mathrm{Var}^{priori}(\beta_j) &= \sigma_{jj}^{(0)} = \frac{\sigma^2}{a} \\ \Longrightarrow \qquad & a_j = \frac{1}{\left((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)_{jj}} \end{aligned} \tag{4.30}$$

$\underline{a}$ und $\overline{a}$ werden dann folgendermaßen aus a_1 und a_2 erzeugt:

$$\begin{aligned} \underline{a} &= \min(a_1 - 0.5a_1,\, a_2 - 0.5a_2) \\ \overline{a} &= \max(a_1 + 0.5a_1,\, a_2 + 0.5a_2) \end{aligned}$$

Eine entsprechende Gleichsetzung mit dem Posteriori-Wert $\sigma_{jj}^{(1)}$ führt nur zu negativen Werten für a und bietet damit keinen anwendbaren Vorschlag für A.

In Abbildung 4.1 ist die Wahl von A gemäß dieser datengeleiteten Strategie dargestellt; sie führt zu relativ weiten Intervallen von $\beta^{(1)}$, es gilt

$$\begin{aligned} \mathbb{E}[\beta_1 \mid z] = \beta_1^{(1)} &\in [-0.60,\, 2.03] & \mathbb{V}(\beta_1 \mid z) = \sigma^2 \cdot \sigma_{11}^{(1)} &= [0.011,\, 0.019] \\ \mathbb{E}[\beta_2 \mid z] = \beta_2^{(1)} &\in [-0.73,\, 1.90] & \mathbb{V}(\beta_2 \mid z) = \sigma^2 \cdot \sigma_{22}^{(1)} &= [0.011,\, 0.019]\,. \end{aligned}$$

Der Grund für eine solch unscharfe Schätzung für β liegt darin, dass A relativ große Werte umfasst und somit von einer niedrigen Varianz der Werte in B_1 und B_2 ausgegangen wird; wie in der Unterschrift der Schaubilder schon angegeben, kann aus $\overline{a}$ (bei gegebenem B_1 und B_2) über die Beziehung (4.18) das ‚Stichprobenäquivalent' $n^{(0)}$ errechnet werden, also die Mindestgröße einer Stichprobe, in deren Ergebnisse man das gleiche Vertrauen setzen würde wie in das durch A, B_1 und B_2 ausgedrückte Vorwissen.

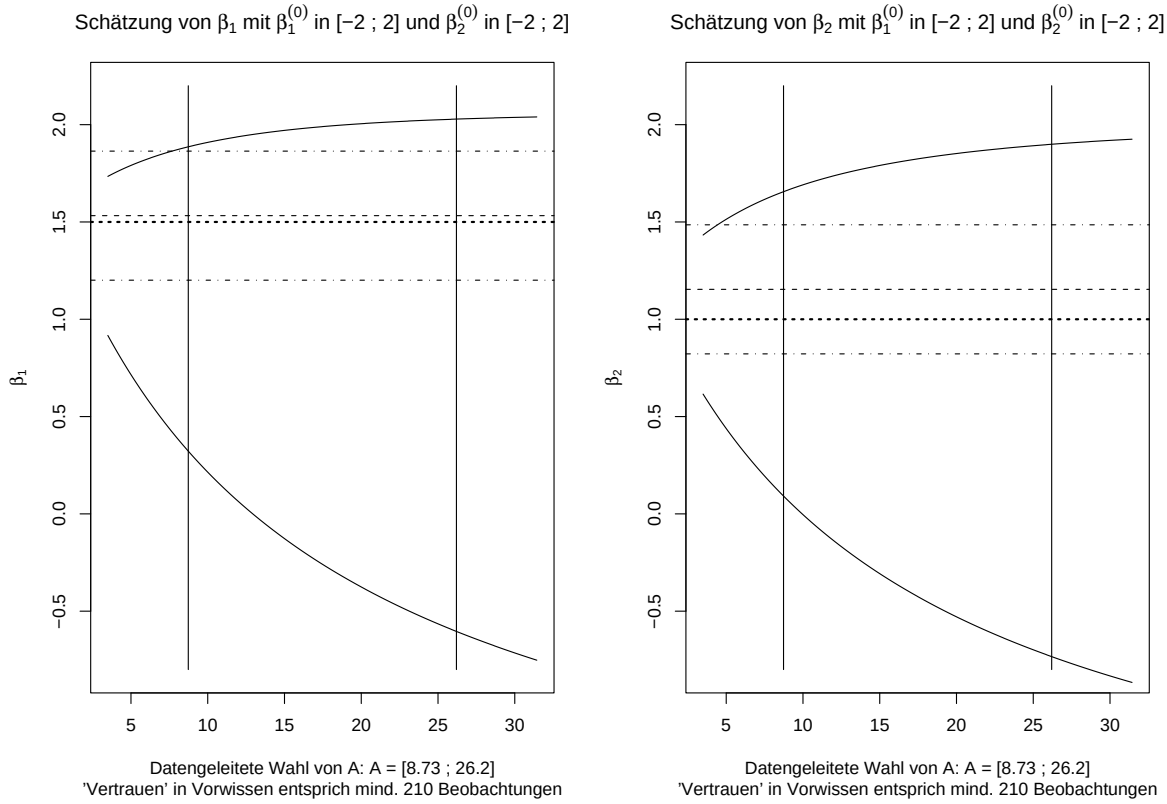

Abbildung 4.1: Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A datengestützt

Die hier vorgenommene Wahl von A, B_1 und B_2 entspricht also einem ‚Vertrauensäquivalent' von mindestens 210 Beobachtungen, aufdatiert wurde hingegen nur auf einer Basis von 20 Beobachtungen. Weil also bei dieser Konstellation das Vertrauen in das unscharfe Vorwissen im Vergleich so viel größer ist als dasjenige, das in eine auszuwertende Stichprobe vom Umfang 20 gesetzt wird, kann man nur wenig ‚dazulernen' und die Intervalle für $\beta^{(1)}$ bleiben noch relativ weit.

Andererseits führt eine so hohe Wahl von $\underline{a}$ und $\overline{a}$ zu einer relativ kleinen Posteriori-Varianz; die Varianz des KQ-Schätzers für β_1 und β_2 ist mit 0.029 deutlich höher. Die Varianzen für β_1 und β_2 sind in beiden Fällen gleich, da beide Regressoren standardisiert wurden.

Wenn hingegen der Stichprobe ein großes Gewicht im Vergleich zur Priori-Verteilung gegeben werden soll, scheinen niedrigere Werte für a angebracht. Über die Bedingung (4.18) kann umgekehrt für ein gegebenes ‚Stichprobenäquivalent' $n^{(0)}$ (bei gegebenem B_1 und B_2) $\overline{a}$ berechnet werden; $\underline{a}$ wird hier (etwas ad hoc) als halb so groß wie $\overline{a}$ de-

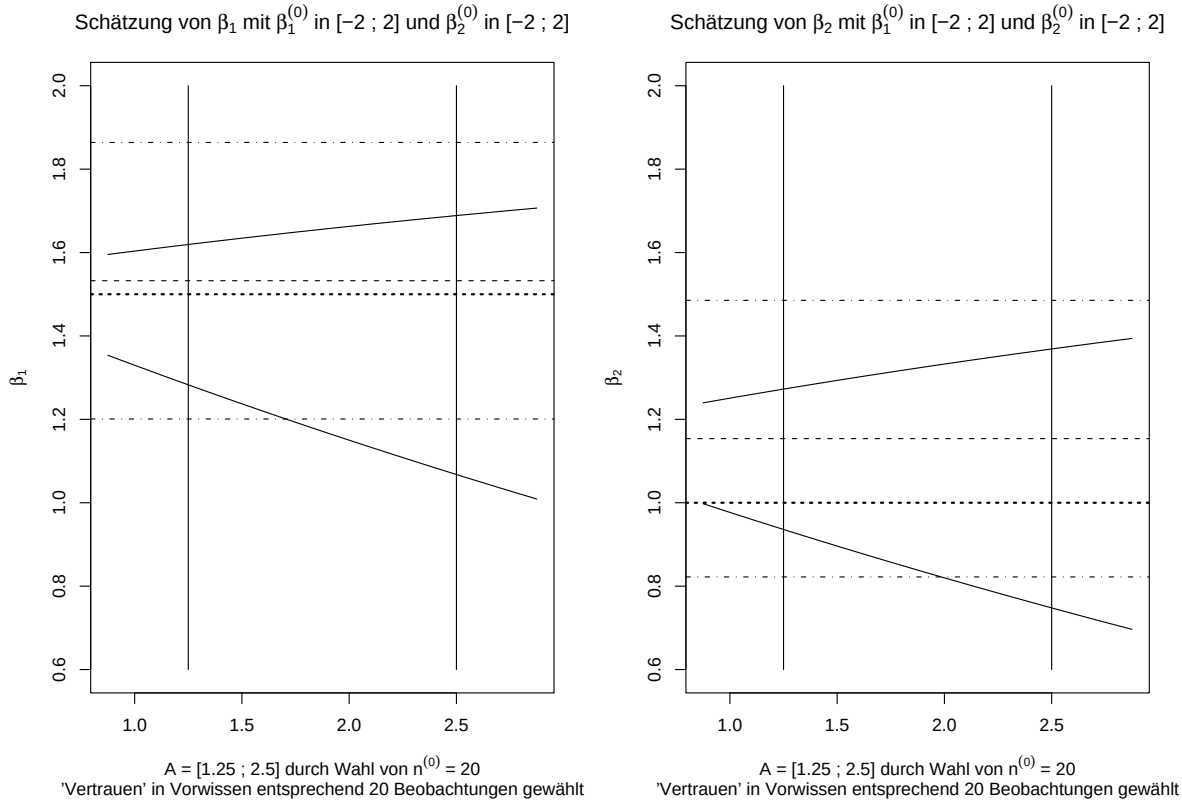

Abbildung 4.2: Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$

finiert.[3] In Abbildung 4.2 ist das Resultat bei einer solchen Ermittlung von A dargestellt.

Wird A also so festgelegt, dass $n^{(0)} = 20$ gilt, wodurch Vorwissen und Beobachtungen mit dem gleichen Gewicht in die Posteriori-Schätzungen eingehen, dann ergibt sich:

$$\mathbb{E}[\beta_1 \mid z] = \beta_1^{(1)} \in [1.07,\, 1.69] \qquad \mathbb{V}(\beta_1 \mid z) = \sigma^2 \cdot \sigma_{11}^{(1)} = [0.025,\, 0.027]$$
$$\mathbb{E}[\beta_2 \mid z] = \beta_2^{(1)} \in [0.75,\, 1.37] \qquad \mathbb{V}(\beta_2 \mid z) = \sigma^2 \cdot \sigma_{22}^{(1)} = [0.025,\, 0.027]$$

Hier sind die resultierenden Intervalle für β deutlich schmäler und von der Größe mit den klassischen Konfidenzintervallen vergleichbar; allerdings handelt es sich bei den

[3] Natürlich könnte auch eine kleinere Wahl von $\underline{a}$ vorgenommen werden, da die Bedingung (4.18) nur für $\overline{a}$ gilt und es daher offen bleibt, wie groß die minimale Präzision bei einer Vorgabe von $n^{(0)}$ gewählt werden kann. Die Wahl von $\underline{a}$ hat jedoch in den meisten Fällen keinen Einfluss auf die Intervalle von $\mathbb{E}[\beta \mid z]$, wie in den folgenden Abbildungen deutlich werden wird; der Einfluss auf die Intervalle von $\mathbb{V}(\beta \mid z)$ ist auf eine Angleichung an die KQ-Schätzung der Varianz beschränkt, die für $\underline{a} \to 0$ erreicht wird. Wählt man also prinzipiell einen Wert von $\underline{a}$ sehr nahe bei Null, so erhält man als Obergrenze von $\mathbb{V}(\beta \mid z)$ den Wert der KQ-Schätzung für die Varianz von β. Um aber zu zeigen, welche Auswirkungen Werte von $\underline{a}$ haben, die größer als Null sind, soll im Folgenden die obengenannte ad hoc - Definition beibehalten werden.

Intervallen für $\beta^{(1)}$ um eine intervallwertige Punktschätzung. Ein Kredibilitätsintervall für diese Punktschätzung entspricht einem nochmals weiteren Intervall; ein Vergleich der klassischen simultanen Konfidenzregion von β_1 und β_2 mit der entsprechenden Kredibilitätsregion erfolgt in Abbildung 4.4.

Die Posteriori-Varianzen von β sind, aufgrund der niedrigeren Werte in A, etwas größer geworden. Eine Veränderung von A hat, bei gleichbleibenden Intervallen B_1 und B_2, Auswirkungen auf $\mathbb{E}[\beta \mid k]$ und $\mathbb{V}(\beta \mid k)$: bei Annahme einer niedrigeren Priori-Präzision werden die Intervalle für $\mathbb{E}[\beta \mid k]$ kleiner, dafür aber die Werte in $\mathbb{V}(\beta \mid k)$ größer. Ein solcher Zusammenhang wird oft mit dem Begriff ‚trade-off' bezeichnet; eine kleinere Varianz einer Schätzung wird hier mit größeren Schätz-Intervallen erkauft. Im Gegensatz zu dem Fall der Schätzung von β über Ridge, Lasso oder Elastic Net, bei dem eine kleine systematische Verzerrung des Schätzwerts zugelassen wird, um dessen Varianz deutlich reduzieren zu können, scheint hier der Vorteil einer kleineren Varianz durch die deutlich größeren Schätzintervalle wieder ‚aufgefressen' zu werden; das wird im Vergleich der beiden Kredibilitätsregionen für β in den Abbildungen 4.3 und 4.4, die durch zwei verschiedene Festlegungen von A erzeugt wurden, deutlich.

Die Erzeugung einer simultanen Kredibilitätsregion für eine zweidimensionale intervallwertige Punktschätzung ist äußerst schwierig; die in (2.16) angegebene Gleichung zur Erzeugung lautet hier folgendermaßen:

$$\underline{P}\left(\left\{\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} : \pi_{\beta^{(1)},\mathbf{\Sigma}^{(1)}}\left(\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}\right) > \xi\right\}\right) = \gamma \tag{4.31}$$

Mit $\pi_{\beta^{(1)},\mathbf{\Sigma}^{(1)})}(\,\cdot\,)$ sei hier die zweidimensionale Normalverteilungsdichte mit Erwartungswert $\beta^{(1)}$ und Kovarianzmatrix $\sigma^2 \cdot \mathbf{\Sigma}^{(1)}$ bezeichnet.

Dies stellt ein äußerst schwer lösbares Problem dar, obwohl die Wahl von $\beta^{(1)}$ für die Minimierungen letztendlich keine Rolle spielt, da sie nur die Lage im Koordinatensystem, aber nicht die Größe oder das Wahrscheinlichkeitsgewicht einer Region beeinflusst. Jedoch ist die untere Wahrscheinlichkeit in Gleichung (4.31) nur über die Minimierung eines Wahrscheinlichkeitsbetrags bezüglich $\mathbf{\Sigma}^{(1)}$ erhältlich, der nur indirekt über eine Optimierung von ξ berechenbar ist, wobei diese Optimierung wiederum von $\mathbf{\Sigma}^{(1)}$ abhängt.

Als Surrogat für eine Kredibilitätsregion soll daher die Vereinigung von Regionen dienen, die jeweils für einen bestimmten Wert von $\beta^{(1)}$ in den oben angegebenen Intervallen für $\beta_1^{(1)}$ und $\beta_2^{(1)}$ gebildet werden und deren Wahrscheinlichkeitsgewicht bezüglich ξ und $\mathbf{\Sigma}^{(1)}$ minimiert wird, so dass es gerade γ beträgt. Diese Regionen werden folgendermaßen generiert:

Gesucht ist der größte Wert von ξ, für den für alle zulässigen $\mathbf{\Sigma}^{(1)}$ und für ein festes $\beta^{(1)}$ die Wahrscheinlichkeit aus (4.31) größer als γ ist. Die zulässigen Werte von $\mathbf{\Sigma}^{(1)}$ sind

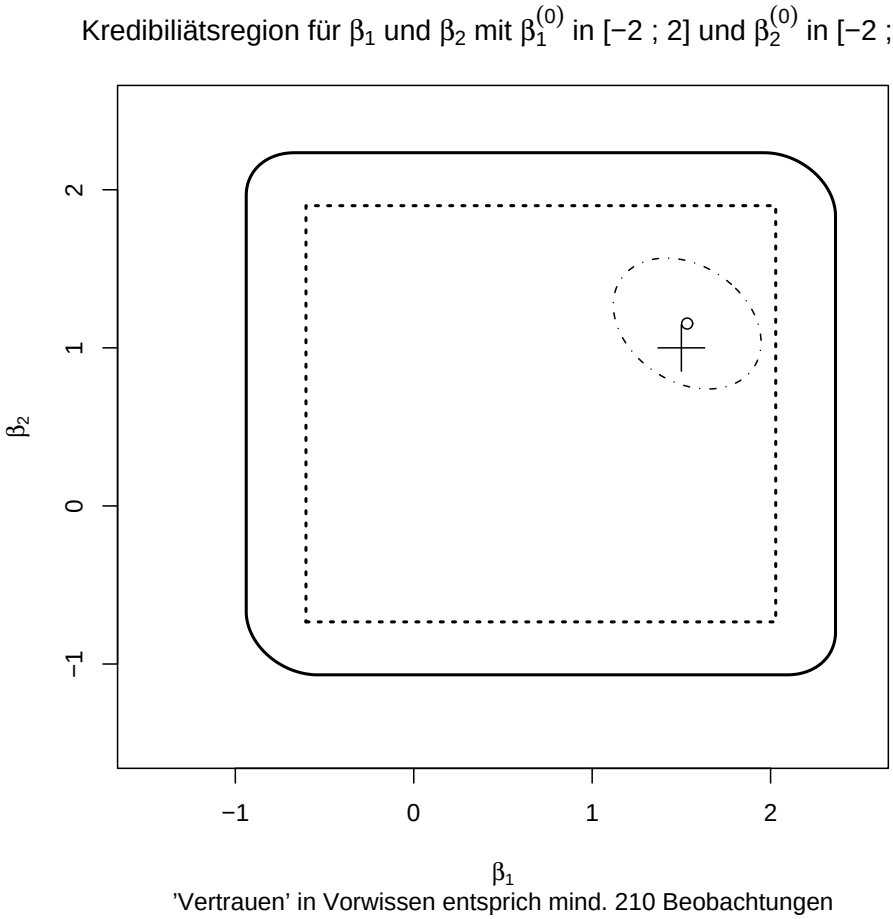

Abbildung 4.3: Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A datengeleitet

jedoch über die Werte von a gemäß (4.19) – (4.21) erhältlich, so dass die Minimierungen nur über $a \in A$ getätigt werden müssen.

1. Ermittle für jedes ξ die Werte von $a \in A$, für die das Wahrscheinlichkeitsgewicht der durch ξ und a definierten Region größer als γ ist.

2. Ermittle nun diejenigen Werte von ξ, für die die in Schritt 1 erzeugte Liste von a-Werten alle Werte aus A umfasst.

Der größte Wert von ξ charakterisiert dann die gesuchte Kredibilitätsregion zusammen mit dem Wert von a, der für dieses ξ das Wahrscheinlichkeitsgewicht der Region minimiert (und welches somit den Wert γ hat). Schritt 2 entspricht dann der Minimierung, über die sich normalerweise die untere Wahrscheinlichkeit ergibt.

Da das in Schritt 1 zu ermittelnde Wahrscheinlichkeitsgewicht nicht analytisch berechenbar ist, sollen die möglichen Werte von ξ und a über ein Gitter durchlaufen werden. Für jede Wertekombination von ξ und a wird dann dieses Wahrscheinlichkeitsgewicht über einen gesonderten Algorithmus berechnet, der eine numerische Optimierung beinhaltet und sich auf ein schon implementiertes Verfahren zur Berechnung von simultanen Konfidenzregionen stützt. Der Algorithmus wurde in `R` implementiert, die Syntax kann auf der im Anhang unter A.2 eingefügten CD gefunden werden.

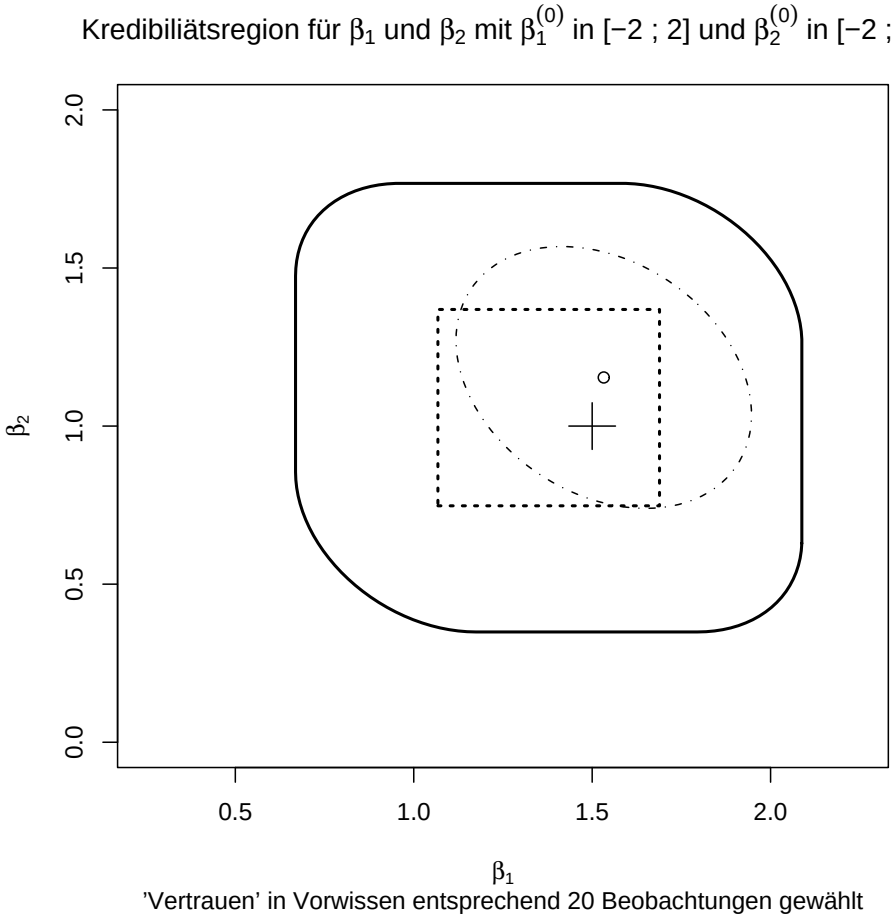

Abbildung 4.4: Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$

In den Abbildungen 4.3 und 4.4 sind jeweils die klassische Konfidenzregion für die KQ-Schätzung und die Kredibilitätsregion für die intervallwertige Punktschätzung von β dargestellt. Die dicke gestrichelte Linie markiert die a posteriori Punktschätzung von β, die als zweidimensionales Intervall vorliegt. Die Kredibilitätsregion für diese intervallwertige Punktschätzung wird durch die dicke durchgezogene Linie umrandet.

Zum Vergleich ist die ellipsenförmige simultane Konfidenzregion mit einer Linie aus dünnen Strichen und Punkten eingetragen. Im Zentrum dieser Konfidenzregion liegt der Wert der KQ-Schätzung, der mit einem kleinen Kreis gekennzeichnet ist. Der ‚wahre' Wert von β ist mit einem etwas größeren Kreuz markiert.

In Abbildung 4.4 ist deutlich erkennbar, dass die Kredibilitätsregion aus mehreren Ellipsen zusammengesetzt ist, die eine ähnliche Form wie die Konfidenzregion aufweisen. Die intervallwertige a posteriori Punktschätzung enthält den ‚wahren' Wert von β und hat in dieser Konstellation in etwa den gleichen Flächeninhalt wie die klassische Konfidenzregion; die Kredibilitätsregion ist etwas weiter, überdeckt aber nicht den Nullpunkt.

In Abbildung 4.3 hingegen ist die Kredibilitätsregion aufgrund des Ausmaßes des zweidimensionalen Intervalls von $\mathbb{E}[\beta \mid z]$ sehr groß; dafür ist aber der ‚Abstand' zwischen den Rändern des zweidimensionalen Intervalls und dessen Kredibilitätsregion (trotz des unterschiedlichen Maßstabs!) deutlich kleiner. Aus diesem Vergleich wird aber ersichtlich, dass die Varianzreduktion durch die höheren Werte in A sehr ‚teuer erkauft' wird; trotz der niedrigeren Varianz ist die resultierende Kredibilitätsregion so groß, dass es die Null überdeckt. Dieser Sachverhalt könnte auch einen Anhaltspunkt liefern, wie die Entscheidung eines Tests unter komplexer Unsicherheit auf den Einfluss der Parameter β_1 und β_2 ausfallen würde: für die Hypothesen H_0: $\beta_1 = 0$, $\beta_2 = 0$ gegen H_1: $\beta_1 \neq 0$, $\beta_2 \neq 0$ gibt es bei dieser Wahl von A, B_1 und B_2, anders als in Abbildung 4.4, deutliche Hinweise darauf, dass H_0 beibehalten werden könnte.

Das Verhalten der Schätzintervalle für β_1 und β_2, wenn sich die Datenbasis vergrößert, ist in den Abbildungen 4.5 und 4.6 dargestellt. Der grau schraffierte Bereich markiert die Fläche zwischen der oberen und unteren Schätzungen für $\mathbb{E}[\beta_j \mid z]$; die KQ-Schätzung ist hier als durchgezogene Linie eingetragen, die zugehörigen klassischen Konfidenzintervalle mit dünnen Linien aus Punkten und Strichen. Als Referenz ist außer dem der ‚wahre' Wert von β_j mit einer waagrechten dick gestrichelten Linie markiert. Deutlich ist zu erkennen, wie die Intervallgrenzen für $k \to \infty$ näher an die KQ-Schätzung rücken, und, zusammen mit dieser, sich dem ‚wahren' Wert von β_j annähern. Der Grund dafür, dass die obere Intervallgrenze sehr bald nur knapp über der KQ-Schätzung zu liegen kommt, die untere Intervallgrenze jedoch länger einen weiteren Abstand zur KQ-Schätzung hält, liegt in der Asymmetrie der Priori-Angaben B_1 und B_2, die einen weiteren Bereich unterhalb des ‚wahren' Wertes abdecken. Am ähnlich gezackten Verlauf der Intervallgrenzen und der KQ-Schätzung wird deutlich, welch enger Zusammenhang zwischen den beiden Schätzungen besteht.

Durch den Vergleich von Abbildung 4.6 mit Abbildung 4.5 wird nochmals deutlich, welchen Einfluss die Wahl des ‚Stichprobenäquivalents' $n^{(0)}$ und, wenn B_1 und B_2 nicht verändert werden, damit die Wahl von A hat: Je mehr Vertrauen in das (intervallwertige) Vorwissen gesetzt wird, desto breiter sind die resultierenden Intervalle, und desto mehr Beobachtungen sind nötig, um die Intervalllänge auf das gleiche Maß zu verkleinern.

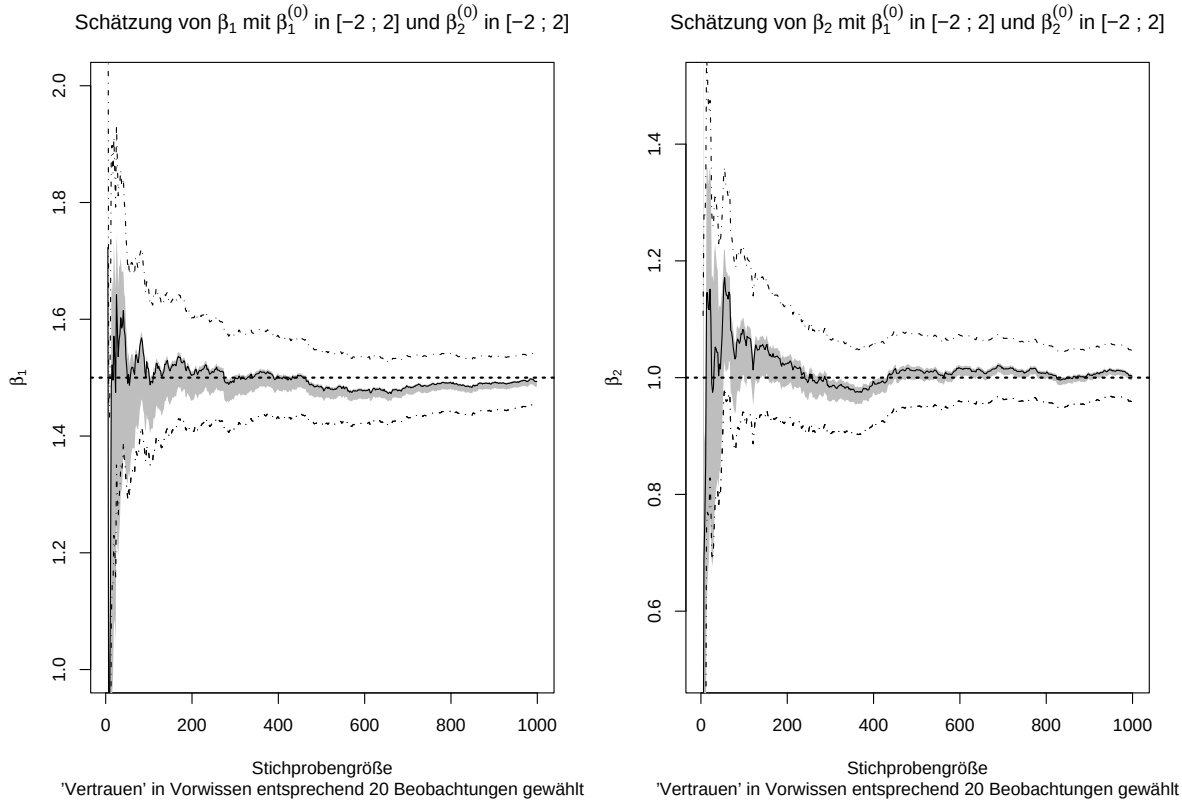

Abbildung 4.5: Schätzintervalle für β bei $k \to \infty$, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$

Wenn der Einfluss der Beobachtungen auf das Posteriori-Ergebnis der Inferenz vergrößert werden soll, könnten auch die Priori-Mengen B_1 und B_2 verkleinert werden, um das ‚Stichprobenäquivalent' $n^{(0)}$ zu verringern. Bei dieser Strategie läuft man jedoch Gefahr, sich in einer ‚prior-data conflict'- Situation wiederzufinden. Auch wenn im vorigen Kapitel das Ergebnis war, dass die Intervalle für von $\mathbf{\Sigma}^{(1)}$ und $\beta^{(1)}$ schon in gewisser Weise von der Beobachtung abhängen, sollte eine solche Situation, wenn möglich, vermieden werden, da nicht klar ist, in welcher Weise das Modell genau darauf reagiert.

Um aber auch den Effekt von kleineren Priori-Intervallen für $\mathbb{E}[\beta \,|\, z]$ beispielhaft zu illustrieren, sollen in den Abbildungen 4.7 und 4.8 die resultierenden Intervalle dargestellt werden, wenn angenommen wird, dass die Intervalle für $\beta_1^{(0)}$ mit $B_1 = [1, 2]$ und für $\beta_2^{(0)}$ mit $B_2 = [0.5, 1.5]$ angegeben werden können. Wieder sollen die beiden oben vorgestellten Methoden zur Ermittlung von A zur Anwendung kommen.

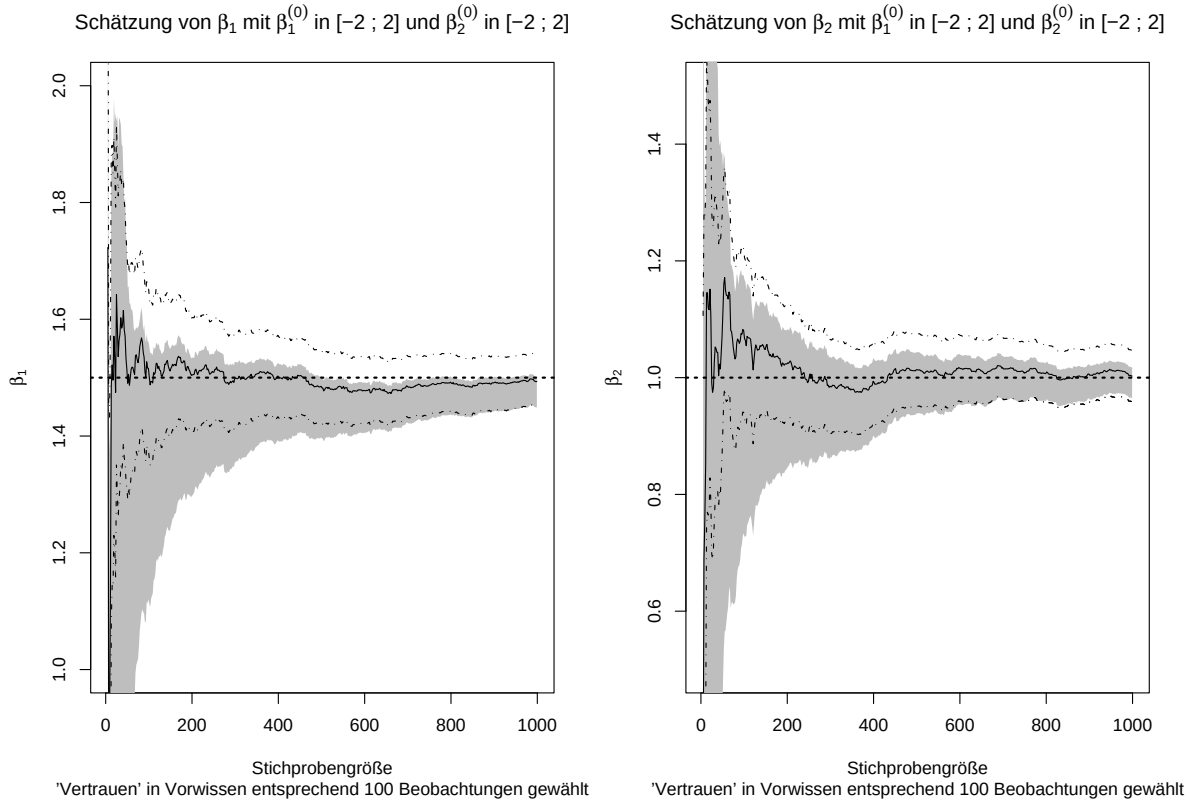

Abbildung 4.6: Schätzintervalle für β bei $k \to \infty$, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 100$

Bei einer ‚datengeleiteten' Wahl von A ergibt sich somit:

$$\mathbb{E}[\beta_1 \mid z] = \beta_1^{(1)} \in [1.20,\, 1.85] \qquad \mathbb{V}(\beta_1 \mid z) = \sigma^2 \cdot \sigma_{11}^{(1)} = [0.011,\, 0.019]$$
$$\mathbb{E}[\beta_2 \mid z] = \beta_2^{(1)} \in [0.74,\, 1.39] \qquad \mathbb{V}(\beta_2 \mid z) = \sigma^2 \cdot \sigma_{22}^{(1)} = [0.011,\, 0.019]$$

Da bei einer ‚datengeleiteten' Wahl von A die Werte von $\underline{b}_1$ und $\underline{b}_2$ nicht eingehen, ergeben sich hier die gleichen Werte für $\underline{a}$ und $\overline{a}$ wie bei der Wahl von umfassenderen Intervallen B_1 und B_2; der Wert des ‚Stichprobenäquivalents' $n^{(0)}$ reduziert sich hingegen. Die in Abbildung 4.7 resultierenden intervallwertigen Punktschätzungen für β_1 und β_2 sind deutlich schmäler als bei Abbildung 4.1 und haben jetzt ungefähr den Umfang der klassischen Konfidenzintervalle. Die Reduzierung der Länge der Priori-Intervalle B_1 und B_2 auf jeweils ein Viertel hat hier also einen ähnlichen Effekt wie die Reduzierung des ‚Stichprobenäquivalents' auf etwa ein Zehntel.

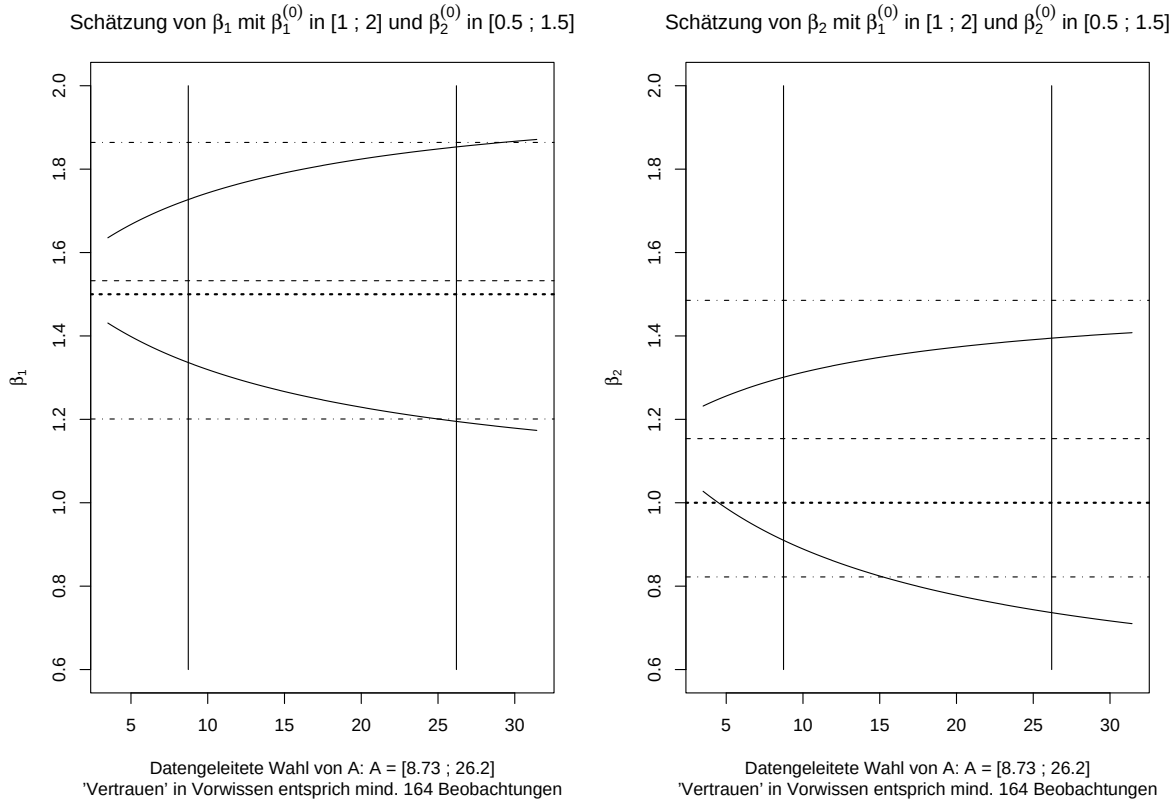

Abbildung 4.7: Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, schmälere Intervalle für $\beta^{(0)}$, Wahl von A datengestützt

Bei einer Wahl von A gemäß $n^{(0)} = 20$ folgt $\underline{a} = 1.6$ und $\overline{a} = 3.2$ und es ergibt sich

$$\mathbb{E}[\beta_1 \mid z] = \beta_1^{(1)} \in [1.44,\, 1.63] \qquad \mathbb{V}(\beta_1 \mid z) = \sigma^2 \cdot \sigma_{11}^{(1)} = [0.024,\, 0.026]$$
$$\mathbb{E}[\beta_2 \mid z] = \beta_2^{(1)} \in [1.04,\, 1.23] \qquad \mathbb{V}(\beta_2 \mid z) = \sigma^2 \cdot \sigma_{22}^{(1)} = [0.024,\, 0.026]$$

Die resultierenden Intervalle sind jetzt, wie in Abbildung 4.8 erkennbar, sehr viel kleiner als die klassischen Konfidenzintervalle geworden, was dadurch ermöglicht wird, dass das Wissen aus den Daten mit dem gleichen Gewicht wie das Vorwissen in die Schätzungen eingeht. Allerdings zeigen sich hier mögliche Probleme bei einer Reduzierung der Priori-Intervalle B_1 und B_2 für β: Die intervallwertige Punktschätzung für β_2 enthält, anders als bei den bisherigen Eingabewerten, nicht mehr den ‚wahren' Wert, der durch die dicke gestrichelte Linie markiert ist, und das, obwohl theoretisch kein ‚prior-data conflict' vorliegt. Die Asymptotik-Aussage bezüglich $\beta^{(1)}$ für $a \to 0$ ist hier besonders deutlich sichtbar: je kleiner der Wert für a, desto enger ziehen sich die Posteriori-Intervalle für β um $\hat{\beta}$ zusammen. Je ähnlicher die intervallwertige Punktschätzung $\hat{\beta}$ wird, desto eher treffen die Eigenschaften der KQ-Schätzung auch auf die intervallwertige Punktschätzung zu, die zwar erwartungstreu ist, aber unter Umständen eine hohe Varianz aufweist und deswegen, wie hier der Fall, auch deutlich

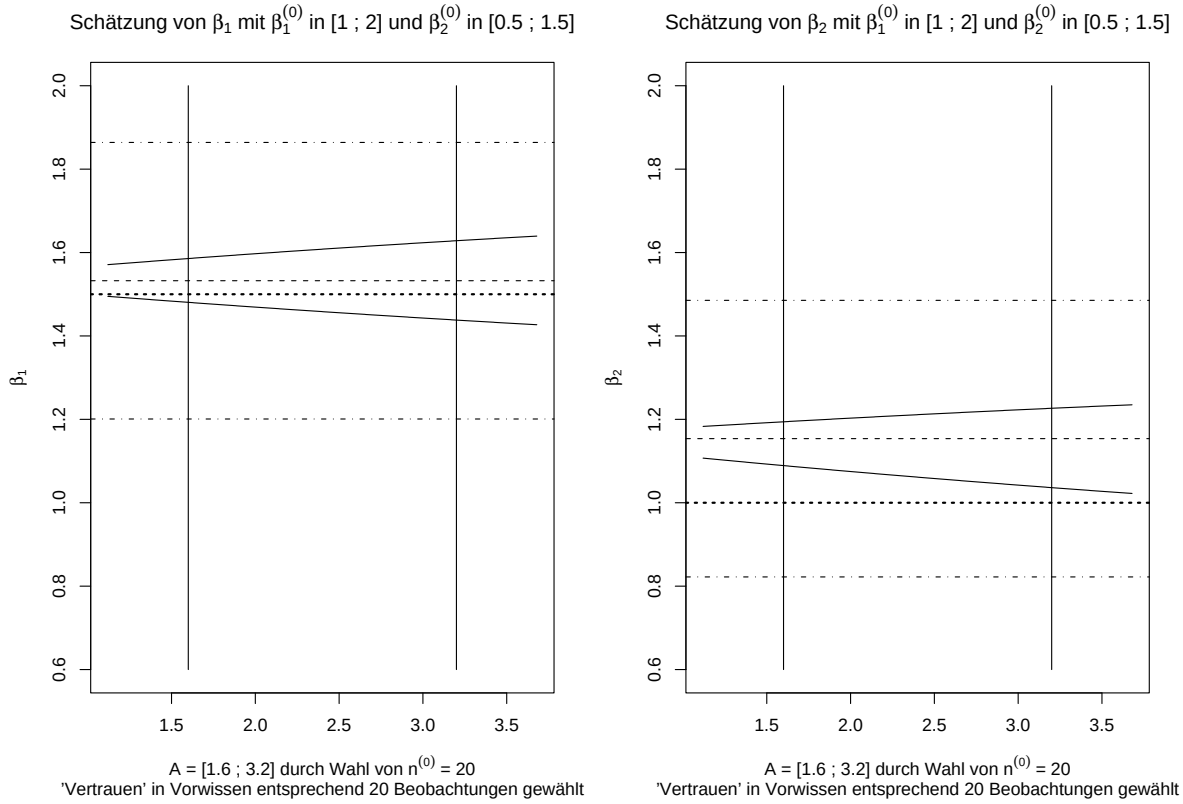

Abbildung 4.8: Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, schmälere Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$

‚daneben liegen' kann.

Die Kredibilitätsregion, die sich für diese Wahl von A, B_1 und B_2 ergibt, ist in Abbildung 4.9 zu sehen. Auch hier ist ersichtlich, dass die Punktschätzung für β als zweidimensionales Intervall nicht mehr den ‚wahren' Wert überdeckt; deutlich ist auch zu sehen, wie sich das zweidimensionale Intervall auf $\hat{\beta}$ zusammenzieht.

Schließlich soll noch gezeigt werden, wie sich eine Wahl von A, B_1 und B_2 auswirkt, die getroffen wird, um sehr schwaches Vorwissen zu modellieren, um einer Modellierung von a priori Nichtwissen nahe zu kommen. A soll gemäß der Vorschläge für $n^{(0)}$ gewählt werden, die für das IDM vorliegen (siehe Abschnitt 2.5, $s^{(0)}$ entspricht $n^{(0)}$), also $1 \leq n^{(0)} \leq 2$. Mit einer sicherheitshalber noch weiteren Wahl von $B_1 = B_2 = [-3\,, 3]$ führt dies zu den Werten $\underline{a} = \frac{1}{18} \approx 0.056$ und $\overline{a} = \frac{1}{9} \approx 0.111$. Die Posteriori-Schätzungen für β lauten dann

$$\mathbb{E}[\beta_1 \,|\, z] = \beta_1^{(1)} \in [1.50,\, 1.55] \qquad \mathbb{V}(\beta_1 \,|\, z) = \sigma^2 \cdot \sigma_{11}^{(1)} = [0.028,\, 0.029]$$
$$\mathbb{E}[\beta_2 \,|\, z] = \beta_2^{(1)} \in [1.12,\, 1.17] \qquad \mathbb{V}(\beta_2 \,|\, z) = \sigma^2 \cdot \sigma_{22}^{(1)} = [0.028,\, 0.029]$$

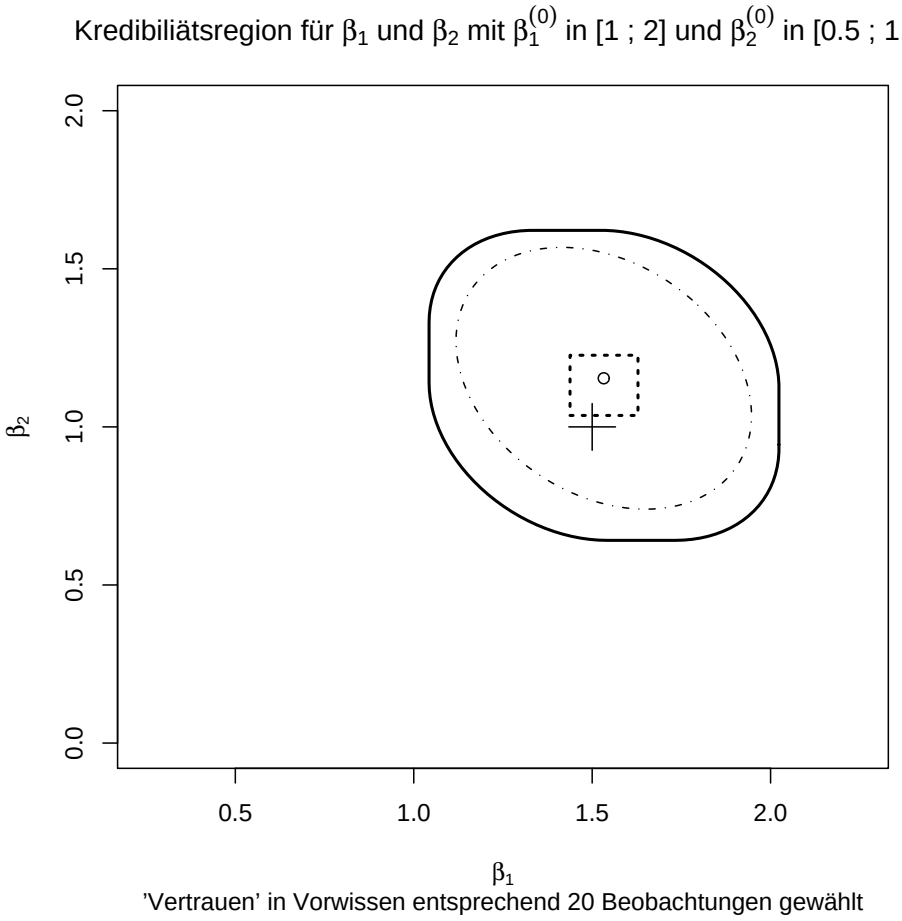

Abbildung 4.9: Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, schmälere Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$

die graphische Darstellung ist in Abbildung 4.10 zu sehen, die zugehörige Kredibilitätsregion ist in Abbildung 4.11 dargestellt.

Da $n^{(0)}$ sehr klein gewählt wurde, liegen die resultierenden Intervallgrenzen für $\beta^{(1)}$ trotz der recht umfangreichen Priori-Intervalle B_1 und B_2 sehr nahe bei $\hat{\beta}$. Das sehr unscharfe Vorwissen bezüglich der Lage von β hat also nicht zur Folge, dass das Wissen nach der Aufdatierung immer noch sehr unscharf ist; dieses Verhalten ist sehr zufriedenstellend. Außerdem ist hier auch ersichtlich, dass durch die Wahl von einem sehr kleinen Wert von $n^{(0)}$ sich die Posteriori-Varianz für β tatsächlich an die Varianz des KQ-Schätzers annähert. In Abbildung 4.11 ist die Punktschätzung für β kaum noch erkennbar, da es sich um ein sehr kleines zweidimensionales Intervall handelt; auch die Kredibilitätsregion ist nun kaum noch weiter als das klassische Konfidenzintervall.

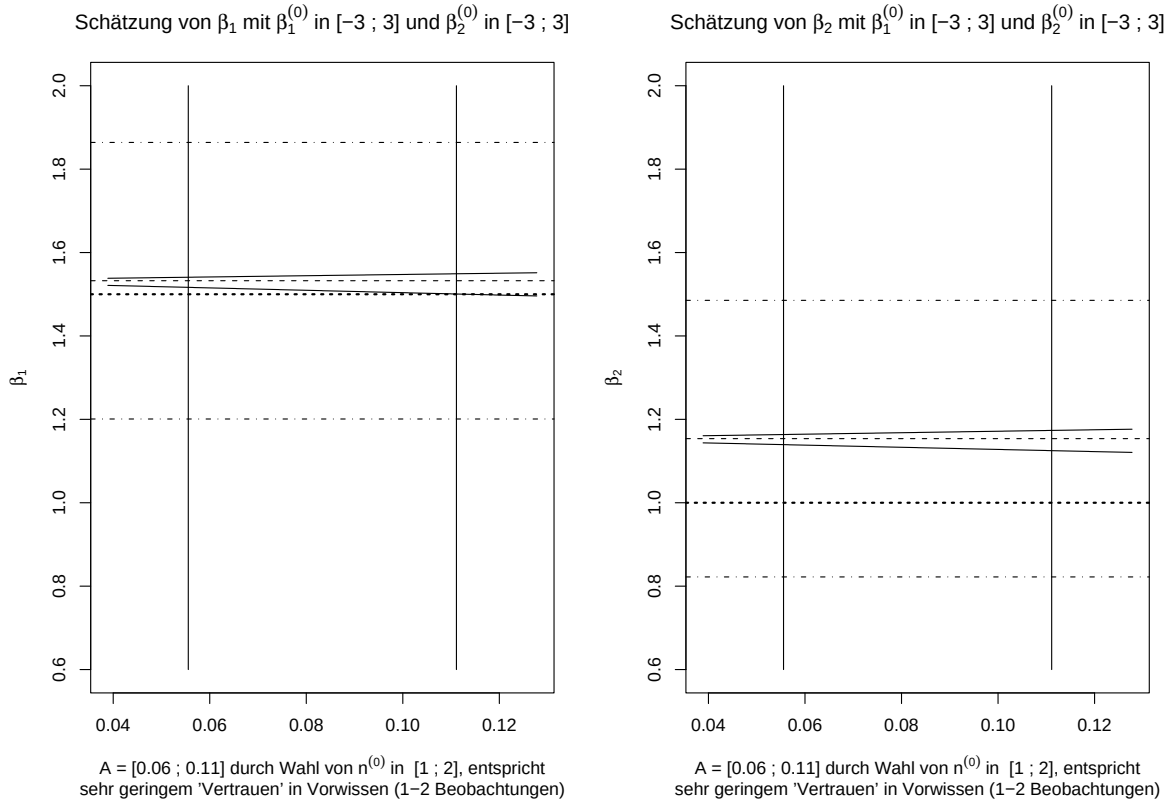

Abbildung 4.10: Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren

Es kann also konstatiert werden, dass das Modell ein vernünftiges und nachvollziehbares Verhalten bezüglich der gezeigten Fälle des Vorwissens zeigt. Auch bei der Modellierung von a priori Nichtwissen, bei der B_1 und B_2 weit und der Bereich der Priori-Präzision A klein gewählt wird, können relativ exakte Ergebnisse gefunden werden. Möchte man hingegen vorsichtigere Inferenzergebnisse erzielen, kann man entweder die Priori-Intervalle für den Regressionsparameter verbreitern, oder $\overline{a}$ höher ansetzen und somit davon ausgehen, dass man sich der unscharfen Angaben von B_1 und B_2 sicherer ist. Es scheint erst einmal kontra-intuitiv, die Priori-Präzision zu erhöhen, um unschärfere Posteriori-Aussagen zu erhalten; man sollte aber A nicht mit der Posteriori-Präzision der Schätzungen für β verwechseln und sich klar machen, dass eine hohe Wahl von $\overline{a}$ einen hohen Wert von $n^{(0)}$ zur Folge hat, weshalb das unscharfe Vorwissen bezüglich β stärker ins Gewicht fällt und somit zur Verbreiterung der Posteriori-Intervalle führt. Eine Vergegenwärtigung der Asymptotik-Aussagen in Kapitel 4.3.3 kann ebenfalls zum Verständnis dieses nicht sofort einleuchtenden Sachverhalts beitragen. In jedem Fall liefert $n^{(0)}$ eine zentrale Interpretation der gesamten Stärke des in A, B_1 und B_2 formulierten Vorwissens und kann über den Vergleich mit der Anzahl der Beobachtungen, die für die Schätzung zur Verfügung stehen, veranschaulicht werden.

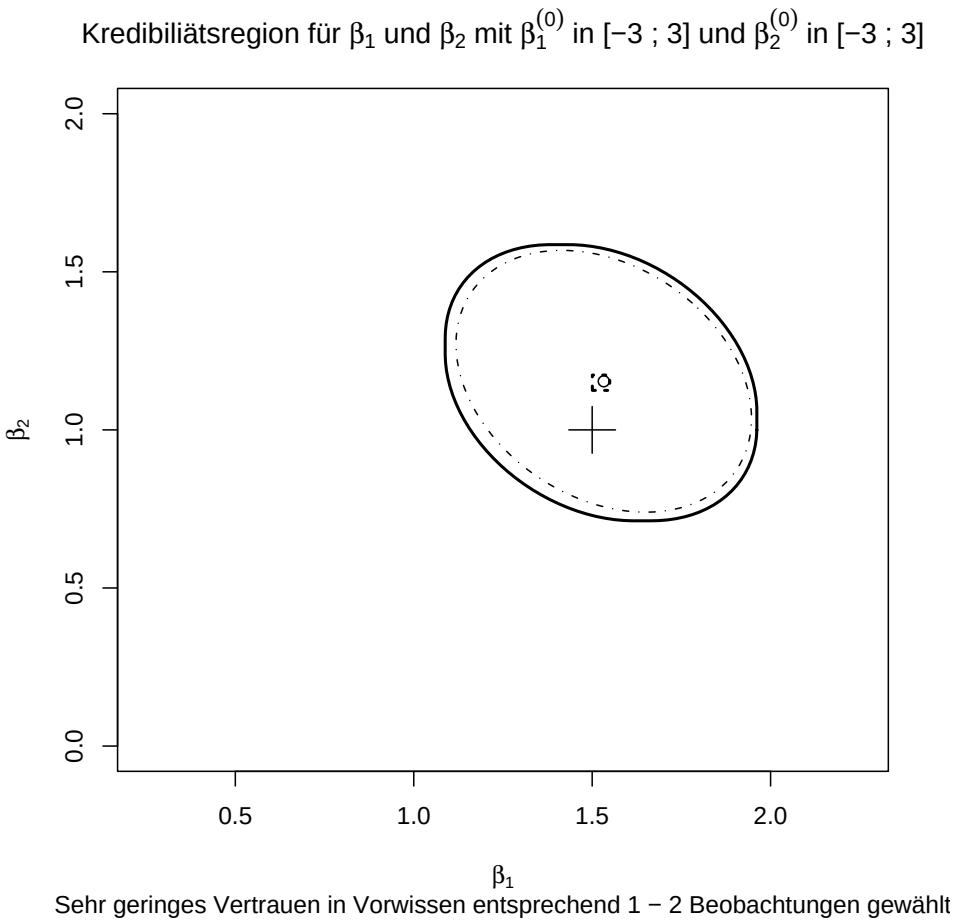

Abbildung 4.11: Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren

In den folgenden beiden Abschnitten soll nun das Verhalten des Modells in zwei anderen Datensituationen beschrieben und gezeigt werden. Eine abschließende Interpretation und eine Zusammenfassung der Ergebnisse findet dann in Kapitel 5.1 statt.

4.4.2 ‚kleine' Koeffizienten, große Varianz

In diesem Abschnitt soll das Verhalten des Modells bei Regressoren mit schwächerem Einfluss und einem größeren Störterm betrachtet werden. Der Datensatz wird in gleicher Weise wie im letzten Abschnitt erzeugt, nur dass hier die ‚wahren' Werte von β_1 0.5 und von β_2 0.1 betragen, und dass $\sigma^2 = 3$ gilt. Damit ergibt sich ein Wert von nur 0.0317 für das multiple adjustierte R^2.

Zuerst soll A aus der Festlegung $n^{(0)} = 20$ abgeleitet werden. Für die Vergleichbarkeit mit den Beispielen aus dem vorigen Abschnitt soll wieder $B_1 = B_2 = [-2, 2]$ gewählt werden, so dass sich die gleichen Grenzen für A, $\underline{a} = 1.25$ und $\overline{a} = 2.5$ ergeben.

In Abbildung 4.12 ist die Ermittlung der Parameter dargestellt; Abbildung 4.13 zeigt die Kredibilitätsregion.

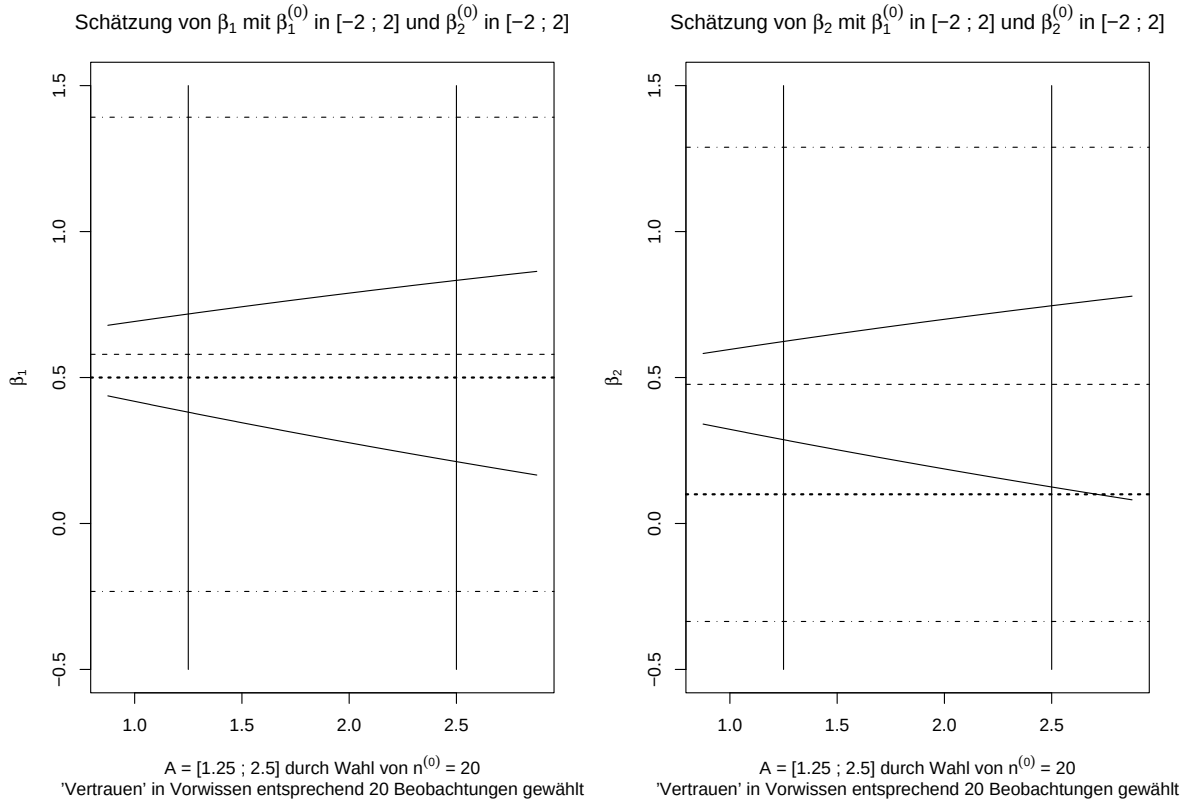

Abbildung 4.12: Simulierter Datensatz mit 20 Beobachtungen und kleinen Werten von β, Wahl von A nach Vorgabe von $n^{(0)} = 20$

Mit diesem neuen Datensatz gilt nun

$$\mathbb{E}[\beta_1 \mid z] = \beta_1^{(1)} \in [0.21,\, 0.83] \qquad \mathbb{V}(\beta_1 \mid z) = \sigma^2 \cdot \sigma_{11}^{(1)} = [0.149,\, 0.159]$$
$$\mathbb{E}[\beta_2 \mid z] = \beta_2^{(1)} \in [0.13,\, 0.75] \qquad \mathbb{V}(\beta_2 \mid z) = \sigma^2 \cdot \sigma_{22}^{(1)} = [0.149,\, 0.159]\,.$$

Trotz der hohen Varianz ändert sich die Länge des Posteriori-Intervalls für β nicht im Vergleich zum ersten Datensatz, da nur $\mathbf{X}$ und nicht z einen Einfluss auf die Länge des Intervalls hat, und für beide Datensätze die Regressoren standardisiert wurden. Die resultierende Länge ist dieses Mal jedoch nicht vergleichbar mit der Länge der Konfidenzintervalle für die KQ-Schätzung, da in deren Berechnung ja σ^2 eingeht. Aufgrund der hohen Varianz, der niedrigen ‚wahren' Parameter und dem geringen Stichprobenumfang ist eine korrekte Schätzung von β natürlich schwierig; daher ist die KQ-Schätzung für β_2 so weit von 0.1 entfernt, und die intervallwertige Punktschätzung für β_2 enthält knapp nicht mehr diesen ‚wahren' Wert.

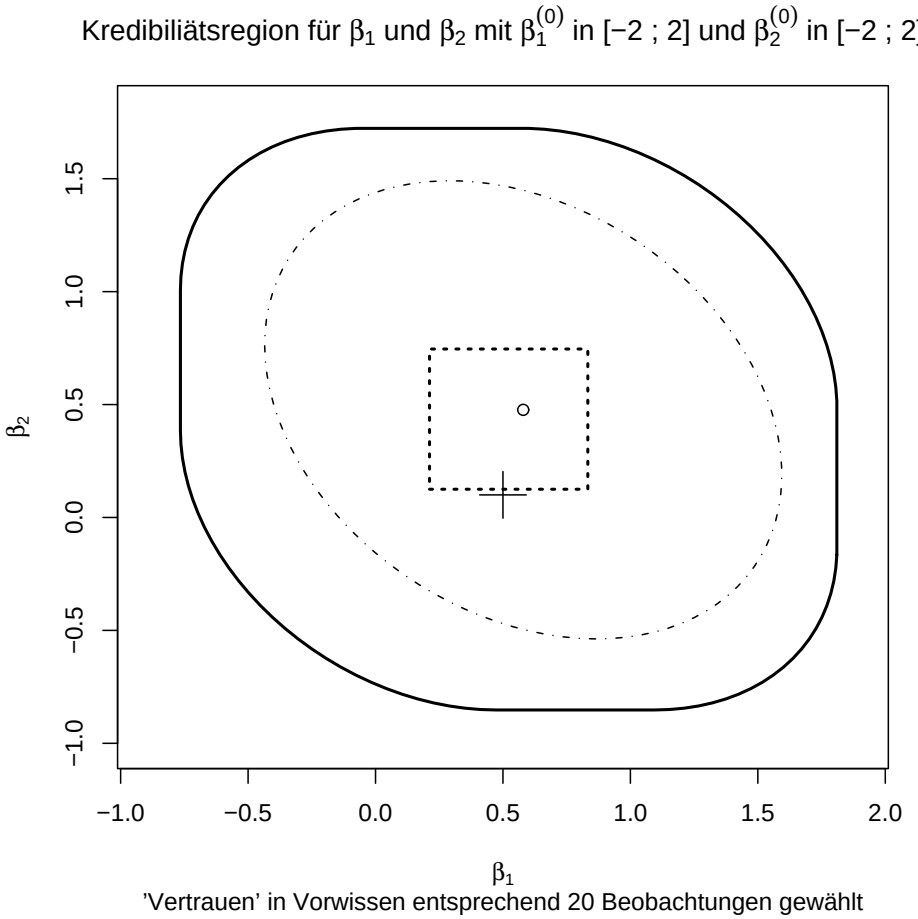

Abbildung 4.13: Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und kleinen Werten von β, Wahl von A nach Vorgabe von $n^{(0)} = 20$

In Abbildung 4.13 ist die klassische simultane Konfidenzregion aufgrund der höheren Varianz σ^2 deutlich größer als in Abbildung 4.4, ebenso die Kredibilitätsregion, obwohl das zugrundeliegende zweidimensionale Intervall die gleiche Fläche besitzt.

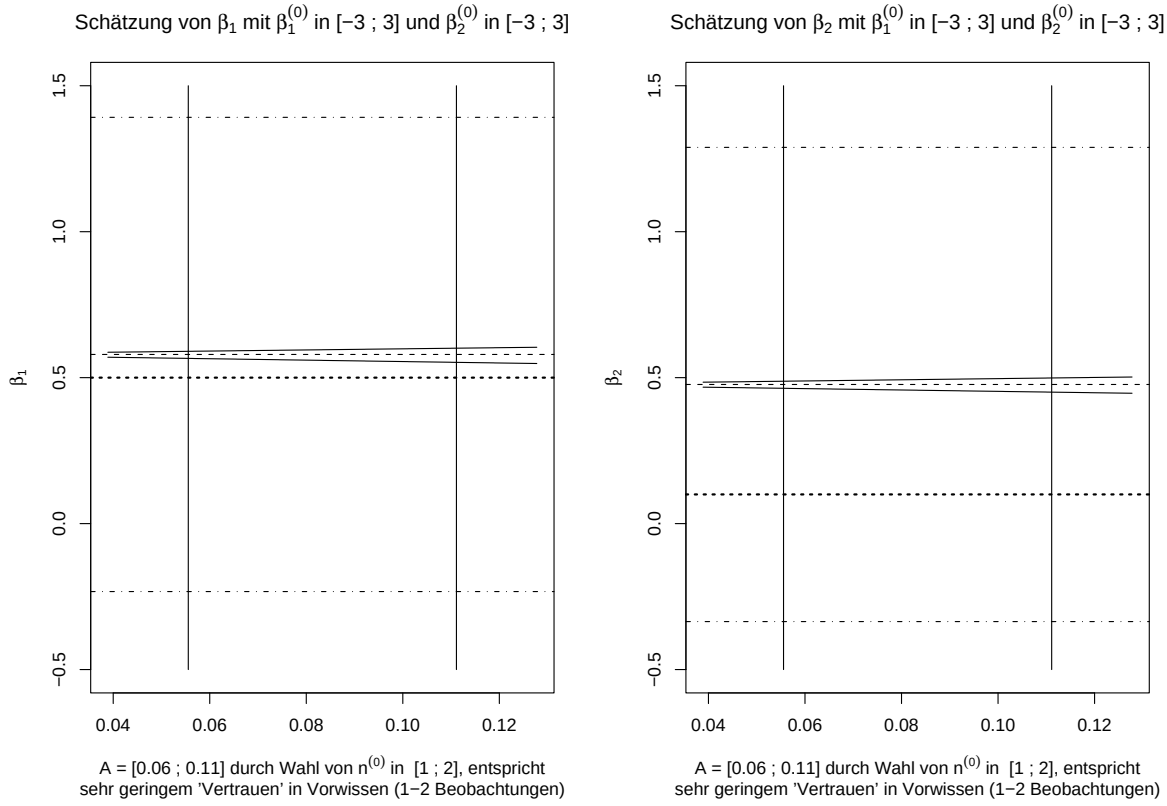

Abbildung 4.14: Simulierter Datensatz mit 20 Beobachtungen und kleinen Werten von β, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren

Nun soll auch für diesen simulierten Datensatz sehr schwaches Vorwissen modelliert werden, indem die gleichen Werte für A, B_1 und B_2 wie im letztem Beispiel des vorigen Abschnitts angenommen werden sollen.

Mit $B_1 = B_2 = [-3,\, 3]$, $n^{(0)} \in [1,\, 2]$ und somit $A = [0.056,\, 0.111]$ gilt

$$\mathbb{E}[\beta_1 \,|\, z] = \beta_1^{(1)} \in [0.55,\, 0.60] \qquad \mathbb{V}(\beta_1 \,|\, z) = \sigma^2 \cdot \sigma_{11}^{(1)} = [0.1706,\, 0.1712]$$
$$\mathbb{E}[\beta_2 \,|\, z] = \beta_2^{(1)} \in [0.45,\, 0.50] \qquad \mathbb{V}(\beta_2 \,|\, z) = \sigma^2 \cdot \sigma_{22}^{(1)} = [0.1706,\, 0.1712]\,.$$

Die Posteriori-Intervalle werden nun wieder sehr schmal; aufgrund der äußerst niedrigen Wahl von $n^{(0)}$ sind die Ergebnisse sehr ähnlich zu denen einer KQ-Schätzung (die Varianz der KQ-Schätzung beträgt 0.1718). In Abbildung 4.14 und 4.15 sind die Ergebnisse graphisch dargestellt; die Intervalle für β_1 und β_2 sind offensichtlich sehr klein; die simultane Konfidenzregion und die Kredibilitätsregion sind praktisch deckungsgleich.

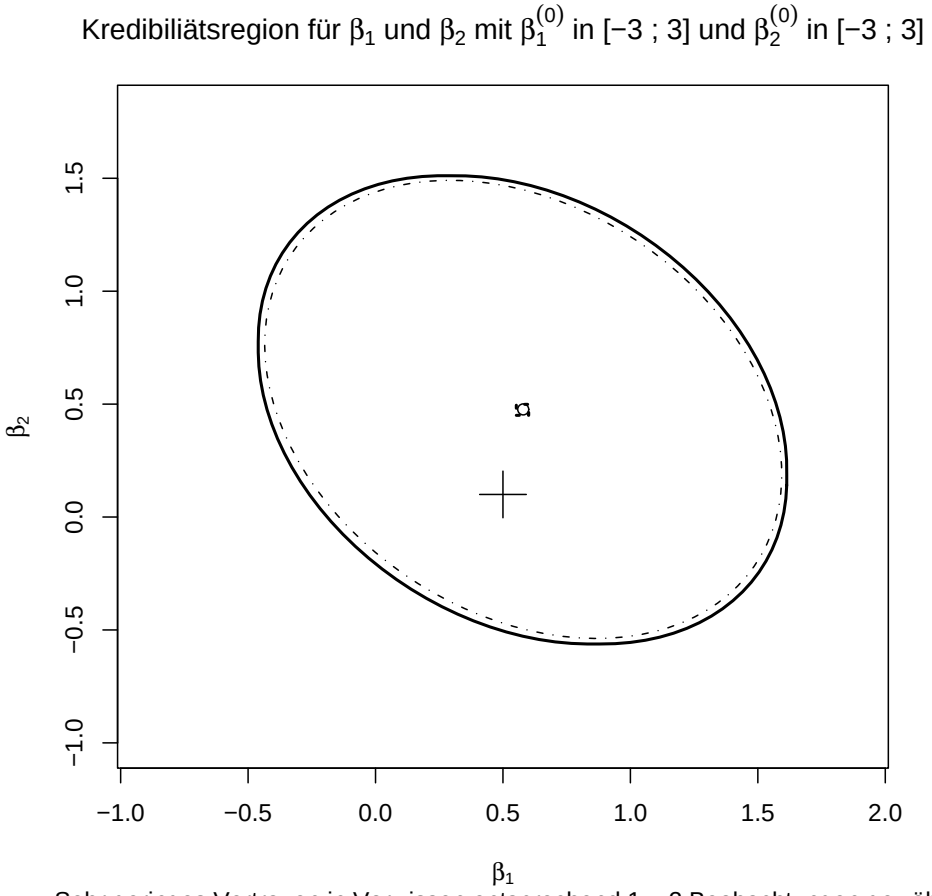

Abbildung 4.15: Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und kleinen Werten von β, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren

4.4.3 Multikollinearität

Ein häufiges Problem bei der Regressionsanalyse ist das der Multikollinearität. Mit diesem Begriff werden Datensituationen bezeichnet, in denen die Regressoren untereinander stark korreliert sind und somit starke lineare Abhängigkeiten zwischen diesen bestehen. In solchen Situationen ist die KQ-Schätzung zwar noch immer erwartungstreu, deren Varianz wird aber so hoch, dass das Ergebnis im konkreten Fall stark daneben liegen kann; oft kann es sogar passieren, dass das Vorzeichen von $\hat{\beta}$ falsch ist. Die Konfidenzintervalle bleiben korrekt, sind dann aber dementsprechend weit.

Der Datensatz wird für diesen Fall folgendermaßen simuliert: Die Werte der Regressionsvariablen sind standardisierte Realisationen einer zweidimensionalen Normalverteilung mit Korrelation $\rho_x = 0.9$; in der Simulation führt dies zu einem Wert der Konditionszahl von 6.12. Die Regressionskoeffizienten werden mit $\beta_1 = \beta_2 = 0.85$ mittelgroß gewählt, die Varianz des Fehlerterms soll mit $\sigma^2 = 1$ einen moderat Wert haben.

Damit ergibt sich im simulierten Datensatz ein multiples adjustiertes R^2 von 0.6552. Die Schätzung für die Varianz von $\hat{\beta}$ ist nun tatsächlich sehr viel größer, als es die moderate Varianz erwarten ließe, nämlich 0.521.

Wieder sollen die Intervalle für $\beta^{(1)}$ zunächst mit denselben Werten für A, B_1 und B_2 wie im ersten Beispiel des vorigen Kapitels berechnet werden. Dort wurde $B_1 = B_2 = [-2, 2]$ gewählt, und durch die Festsetzung von $n^{(0)} = 20$ $A = [1.25, 2.5]$ ermittelt.

$$\begin{aligned} \mathbb{E}[\beta_1 \mid z] = \beta_1^{(1)} &\in [-0.60,\ 2.27] & \mathbb{V}(\beta_1 \mid z) = \sigma^2 \cdot \sigma_{11}^{(1)} &= [0.156,\ 0.237] \\ \mathbb{E}[\beta_2 \mid z] = \beta_2^{(1)} &\in [-0.47,\ 2.40] & \mathbb{V}(\beta_2 \mid z) = \sigma^2 \cdot \sigma_{22}^{(1)} &= [0.156,\ 0.237]\,. \end{aligned}$$

Die Ergebnisse sind auch in den Abbildungen 4.16 und 4.17 dargestellt. Wie in Abbildung 4.16 deutlich wird, sind nun die Posteriori-Intervalle für $\beta^{(1)}$ trotz der gleichen Werte für A, B_1 und B_2 auf einmal sehr viel breiter als in Abbildung 4.12; auch das klassische Konfidenzintervall für β_1 und β_2 ist breiter geworden, jedoch nicht im gleichen Maße wie die intervallwertigen Punktschätzungen des Normal-Modells. Hier zeigt sich ein Vorteil der in Abschnitt 4.3.3 etwas verstörenden Abhängigkeit der Intervalllänge von $\beta_j^{(1)}$ von $\mathbf{X}$. Die Intervalle für $\beta_j^{(1)}$ sind in der Lage, die Unsicherheit der Schätzung aufgrund der Kollinearität direkt zu reflektieren, während die Schätzung $\hat{\beta}$ allein diese Unsicherheit nicht vermitteln kann, sondern nur über die Angabe der Konfidenzintervalle.

Auch die in Abbildung 4.17 dargestellte zugehörige Kredibilitätsregion macht im Vergleich zu Abbildung 4.13 deutlich, dass man sich hier in einer völlig anderen Situation befindet. Die Form sowohl der Kredibilitäts- als auch der Konfidenzregion liefert einen starken Hinweis darauf, dass zwischen den Regressoren eine starke Abhängigkeit besteht. Die tendenziell negative Korrelation zwischen den Schätzwerten von β_1 und β_2, die sich in den anderen simulierten Datensätzen auch schon durch die Form der

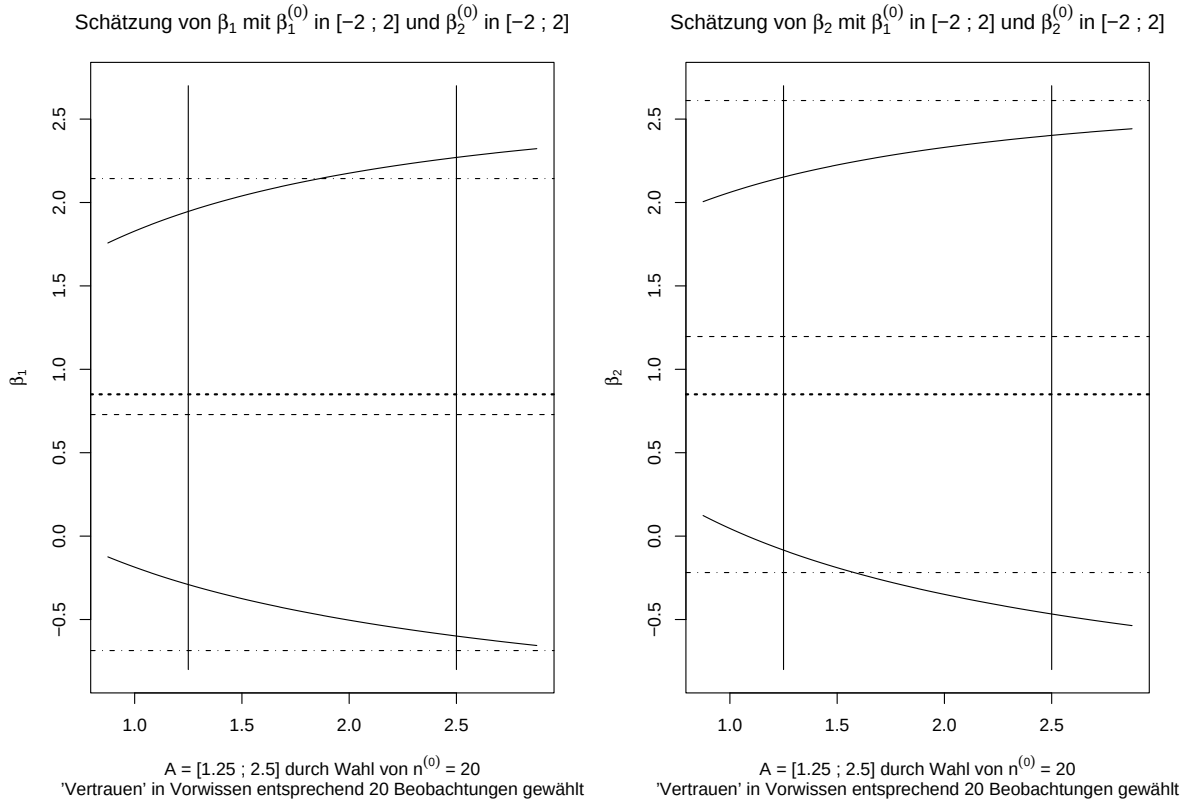

Abbildung 4.16: Simulierter Datensatz mit 20 Beobachtungen und hoher Kollinearität der Regressoren, Wahl von A nach Vorgabe von $n^{(0)} = 20$

Kredibilitäts- und Konfidenzregion angedeutet hatte, ist hier besonders deutlich sichtbar.

Für diesen dritten Datensatz sollen nun auch die Schätzintervalle für sehr schwaches Vorwissen, das in gleicher Weise wie bei den vorigen simulierten Datensätzen modelliert wird, angegeben werden.

$$\mathbb{E}[\beta_1 \mid z] = \beta_1^{(1)} \in [0.45,\, 1.05] \qquad \mathbb{V}(\beta_1 \mid z) = \sigma^2 \cdot \sigma_{11}^{(1)} = [0.470,\, 0.494]$$
$$\mathbb{E}[\beta_2 \mid z] = \beta_2^{(1)} \in [0.87,\, 1.47] \qquad \mathbb{V}(\beta_2 \mid z) = \sigma^2 \cdot \sigma_{22}^{(1)} = [0.470,\, 0.494]$$

Wie bisher werden im Vergleich zur vorigen Wahl von A die Intervalle wesentlich kleiner. Sie sind im Resultat hier jedoch nicht so klein wie etwa in Abbildung 4.14, so dass auch bei dieser Wahl von A die Andersartigkeit der Situation deutlich wird. Auffällig ist ebenso, dass der Unterschied der Posteriori-Varianzen zur entsprechenden KQ-Schätzung hier noch deutlich größer bleibt.

Deshalb ist, wie aus Abbildung 4.19 ersichtlich, der Unterschied zwischen der Konfidenz- und der Kredibilitätsregion hier ebenfalls viel größer als noch in Abbildung 4.15, die dort fast deckungsgleich waren; die Schätzung unter Berücksichtigung von komplexer

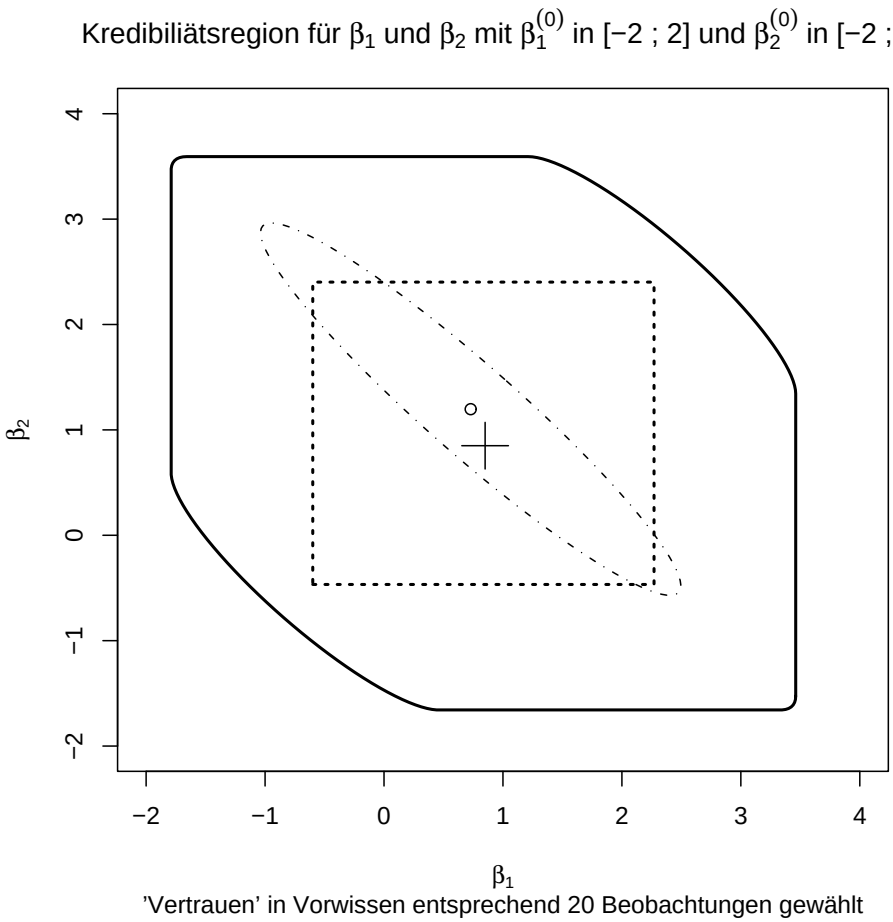

Abbildung 4.17: Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und hoher Kollinearität der Regressoren, Wahl von A nach Vorgabe von $n^{(0)} = 20$

Unsicherheit verdeutlicht also in viel stärkerem Maße, wie unsicher und ungenau die Schätzung des Regressionsparameters im Falle von Multikollinearität sein kann.

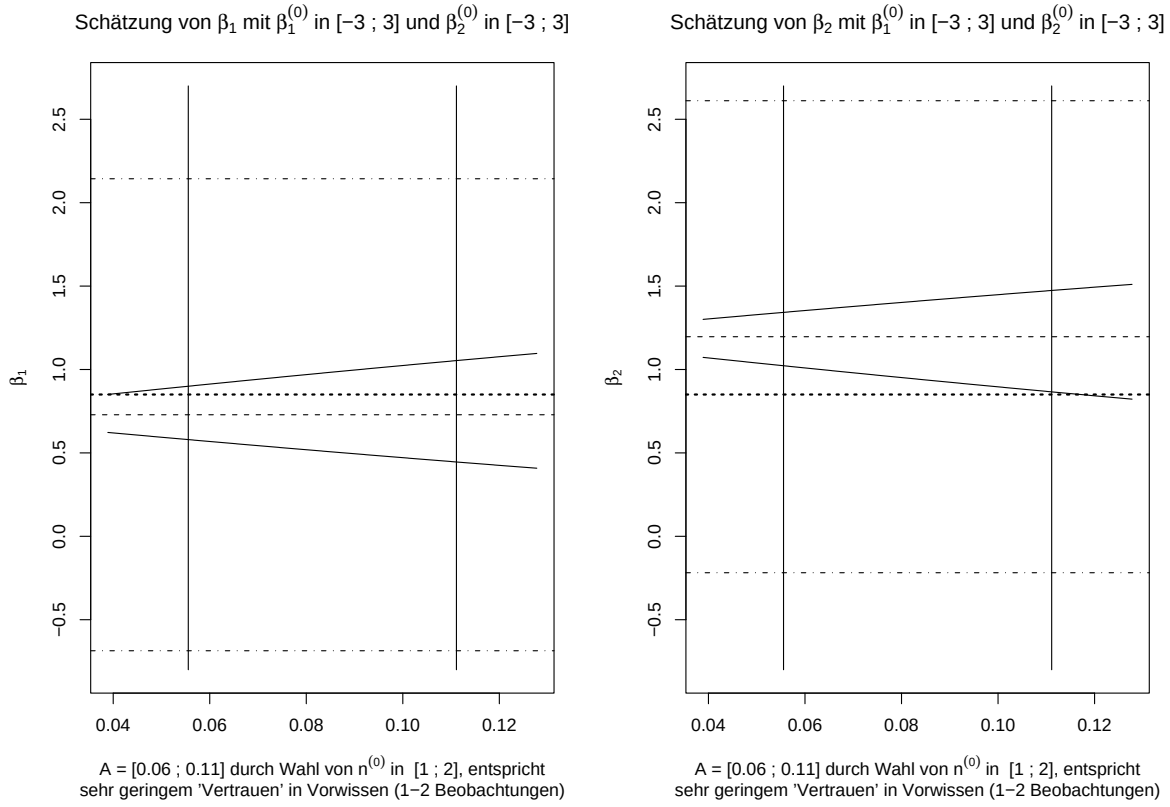

Abbildung 4.18: Simulierter Datensatz mit 20 Beobachtungen und hoher Kollinearität der Regressoren, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren

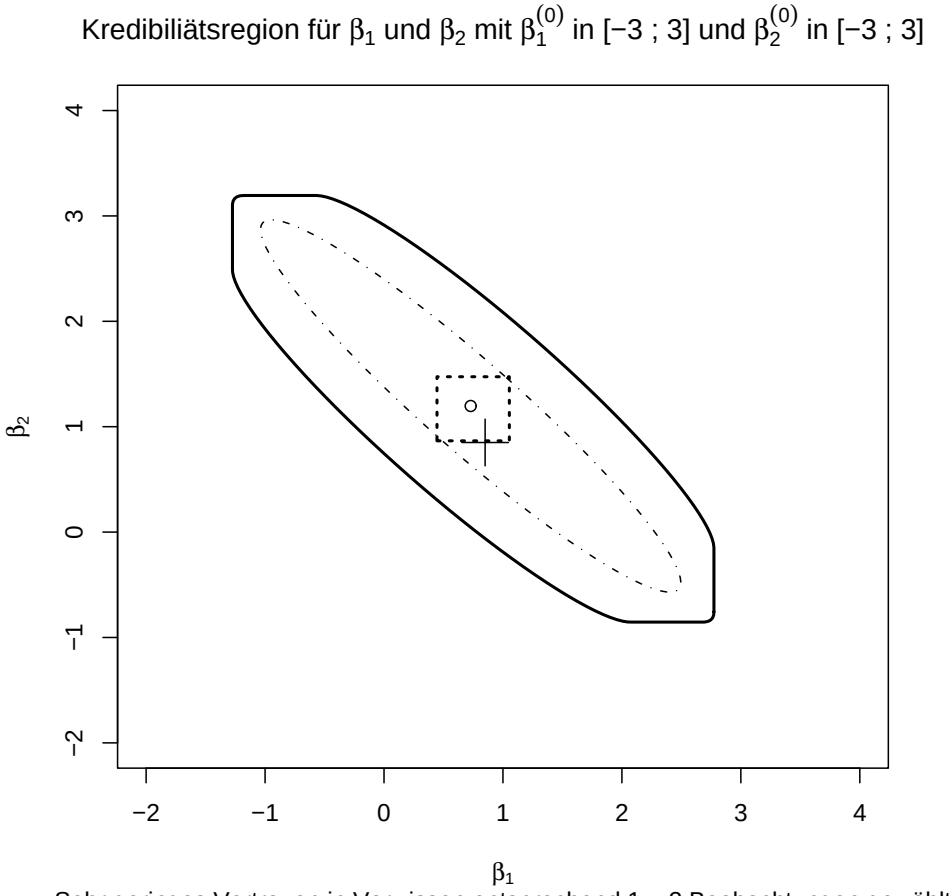

Abbildung 4.19: Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und hoher Kollinearität der Regressoren, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren

4.5 Datenbeispiel

4.5.1 Die AIRGENE-Studie

Die Daten, anhand der die Funktionsweise des Normal-Modells unter komplexer Unsicherheit unter realen Bedingungen demonstriert werden soll, stammen von der AIRGENE-Studie. Die AIRGENE-Studie untersucht den Zusammenhang zwischen Luftschadstoffen und Entzündungsmarkern im Blut [Peters et al., eingereicht] und basiert unter anderem auf dem KORA Herzinfarkt-Register Augsburg [Löwel et al. 2005]. Hier wurde ein Auszug aus dem Datensatz, nämlich die mittleren Fibrinogen-Konzentrationen, das Alter und der Body-Mass-Index (Gewicht/(Körpergröße)^2) von Probanden der Studie als Beispiel verwendet.

AIRGENE steht für „Air Pollution and Inflammatory Response in Myocardial Infarction Survivors: Gene-Environment Interaction in a High Risk Group“ [Peters et al., eingereicht]. Es handelt sich um eine von der EU finanzierte Panel-Studie, die sich auf Erhebungen in sechs europäischen Städten stützt. In Athen, Augsburg, Barcelona, Helsinki, Rom und Stockholm wurden jeweils ca. 200 Probanden rekrutiert, die schon einen Herzinfarkt überlebt hatten. Die Probanden wurden in einer ersten Basis-Untersuchung genau nach ihren Lebensumständen befragt, und es wurde ihnen eine erste Blutprobe abgenommen; weitere fünf bis sieben Untersuchungen pro Proband erfolgten dann im Abstand von etwa vier Wochen. Bei diesen Folgeuntersuchungen wurde jeweils ein Kurz-Fragebogen ausgefüllt und eine erneute Blutabnahme vorgenommen. Während der Studiendauer wurden verschiedene Wetter- und Luftschadstoffvariablen, unter anderem auch die Feinstaubkonzentration, auf stündlicher Basis gemessen.

Entzündungsmarker wie Fibrinogen sind Indikatoren für Entzündungsprozesse im menschlichen Körper. Um eine entzündliche Erkrankung handelt es sich auch bei der Arteriosklerose, bei der sich Ablagerungen in den Blutgefäßen bilden. Brechen solche Ablagerungen in den Herzkranzgefäßen auf, können Blutgerinnsel einzelne Gefäße verstopfen und eine Herzregion von der Blutzufuhr abschneiden, so dass es zum Herzinfarkt kommt.

Epidemiologische Studien haben beobachtet, dass auch Übergewicht mit einem Anstieg der Entzündungsparameter im Blut assoziiert ist [Thorand et al. 2006]. Ebenso steigen Entzündungsmarker im Blut häufig mit dem Alter an.

Das in dieser Arbeit entwickelte Normal-Modell unter komplexer Unsicherheit in der in Kapitel 4.3.3 vorgestellten rechnerisch direkt anwendbaren Version soll nun hier verwendet werden, um den Einfluss des Alters und des BMI, einer Maßzahl für Übergewicht, auf den Entzündungsmarker-Spiegel zu modellieren. Zielvariable ist die über die wiederholten Blutuntersuchungen gemittelte Fibrinogen-Konzentration `fib` der Probanden, die über das KORA-Herzinfarktregister am Zentralklinikum Augsburg rekrutiert wurden; als Prädiktoren werden das Alter in Jahren (`age`) und der Body-

Mass-Index (bmi), gemessen als ‚Gewicht in kg/(Körpergröße in m)^2' in das Modell aufgenommen.

Um einer Voraussetzung des linearen Regressionsmodells, nämlich der Normalverteilung der Residuen, zu entsprechen, muss die Zielvariable fib logarithmiert werden. Damit eine Schätzung des Intercepts im Normal-Modell unnötig gemacht werden kann, werden analog zum Vorgehen bei den simulierten Datensätzen die Regressoren standardisiert und die Zielvariable zentriert. Die mit einer solchen Datenbasis ermittelten sogenannten standardisierten Regressionskoeffizienten lassen sich zu den eigentlichen, interpretierbaren Koeffizienten folgendermaßen transformieren:

$$\beta_{\texttt{age}} = \frac{1}{\sqrt{\mathbb{V}(\texttt{age})}} \cdot \tilde{\beta}_{\texttt{age}} \tag{4.32}$$

$$\beta_{\texttt{bmi}} = \frac{1}{\sqrt{\mathbb{V}(\texttt{bmi})}} \cdot \tilde{\beta}_{\texttt{bmi}} \tag{4.33}$$

$$\beta_0 = \mathbb{E}[\log(\texttt{fib})] - \mathbb{E}[\texttt{age}] \cdot \beta_{\texttt{age}} - \mathbb{E}[\texttt{bmi}] \cdot \beta_{\texttt{bmi}} \tag{4.34}$$

$\beta_{\texttt{age}}$ bezeichnet dabei hier die gewöhnliche Schätzung des Regressionsparameters für die (nicht standardisierte) Variable age aus dem linearen Modell (mit Intercept β_0), $\tilde{\beta}_{\texttt{age}}$ bezeichnet hingegen die zugehörige Schätzung aus dem Modell mit den standardisierten Regressoren und zentrierter Zielvariable. Die Erwartungswerte und Varianzen können in der Anwendung durch ihre jeweiligen erwartungstreuen Schätzungen ersetzt werden.

Nach der Schätzung von σ^2, der Varianz es Fehlerterms ε, die über die bekannte erwartungstreue Schätzung $\hat{\sigma}^2$ mittels der Residuen eines gewöhnlichen linearen Modells erfolgt, kann das Normal-Modell unter komplexer Unsicherheit wie im vorigen Kapitel angewendet werden.

Zur inhaltlichen Ermittlung des benötigten Priori-Intervalls für $\tilde{\beta}_{\texttt{age}}$ wird als Ausgangspunkt der minimal und maximal mögliche Wert für die Variablen age gemäß der Einschlusskriterien der Studie mit 35 bzw. 80 angenommen; diese Grenzwerte werden unter Verwendung der im Datensatz ermittelten Größen standardisiert und in Beziehung zu der zentrierten Zielvariable log(fib) gesetzt, deren minimal und maximal möglicher Wert sehr vorsichtig mit -1 bzw. 1 festgesetzt werden kann. Somit ergibt sich a priori, bei vergrößernder Rundung des Intervalls, $\tilde{\beta}_{\texttt{age}} \in [-0.5, 0.5]$. Zur Ermittlung des Priori-Intervalls für $\tilde{\beta}_{\texttt{bmi}}$ wird analog vorgegangen; in Ermangelung eines offiziellen Ein- oder Ausschlusskriteriums (jedoch unter der Annahme, dass massiv über- oder untergewichtige Personen nicht die Studie aufgenommen wurden) werden die Unter- und Obergrenze für die Variable bmi mit 15 bzw. 50 festgesetzt.[4] Daraus folgt a priori $\tilde{\beta}_{\texttt{bmi}} \in [-0.3, 0.3]$.

[4] Eine $1,6$ m große Person wöge mit einem Body-Mass-Index von 15 nur 38.4 kg; eine $1,7$ m große Person wöge mit einem Body-Mass-Index von 50 hingegen 144.5 kg.

4.5.2 Analyse bei ‚datengeleiteter' Wahl von A

Zunächst soll noch einmal für die Wahl von A die Strategie verfolgt werden, einen ‚zentralen' Wert für a über einen Vergleich mit der Varianzschätzung im KQ-Modell zu erhalten; im darauf folgenden Abschnitt soll dann die ebenfalls in 4.4 erläuterte Strategie zur Modellierung von sehr schwachem Vorwissen angewendet werden.

Da im vorigen Abschnitt schon der vernünftigerweise zu erwartende Rahmen für die Parameter $\tilde{\beta}_{\texttt{age}}$ und $\tilde{\beta}_{\texttt{bmi}}$ in konservativer Weise ermittelt wurde und damit die Mengen B_1 und B_2 feststehen, kann ein ‚zentraler' Wert von a bei einer ‚datengeleiteten' Wahl direkt mittels der Gleichung (4.30) abgeleitet werden. Bei gleichem Vorgehen wie im Falle der simulierten Datensätze ergibt sich $A = [96.5,\, 289.5]$, was zu einem Wert von $n^{(0)} = 99$ führt. Das in A, $B_{\texttt{age}}$ und $B_{\texttt{bmi}}$ ausgedrückte Vorwissen wird also mit einem etwa halb so großen Gewicht wie die Stichprobe, die den Umfang 199 hat, in die Schätzungen eingehen. Die Schätzung der sogenannten standardisierten Regressionskoeffizienten, also der Parameter bezüglich der linearen Regression mit standardisierten Regressoren, liefert im Normal-Modell folgendes Ergebnis:

$$\mathbb{E}[\tilde{\beta}_{\texttt{age}} \mid \log(\texttt{fib})] = [-0.28,\, 0.34] \qquad \mathbb{V}(\tilde{\beta}_{\texttt{age}} \mid \log(\texttt{fib})) = [0.000050,\, 0.000083]$$
$$\mathbb{E}[\tilde{\beta}_{\texttt{bmi}} \mid \log(\texttt{fib})] = [-0.18,\, 0.22] \qquad \mathbb{V}(\tilde{\beta}_{\texttt{bmi}} \mid \log(\texttt{fib})) = [0.000050,\, 0.000083]$$

Die resultierende Intervalle für die Schätzungen der Regressionskoeffizienten sind relativ breit, die Posteriori-Grenzen haben sich noch nicht sehr weit von ihren Priori-Werten entfernt. Die resultierenden Posteriori-Varianzen sind jedoch geradezu winzig, da aufgrund der Wahl von A die a priori angenommenen Varianzen für den standardisierten Regressionsparameter ebenfalls sehr klein waren. (Die Unter- und Obergrenzen der Priori-Varianz bei dieser Wahl von A betragen 0.000084 bzw. 0.00025.) In Abbildung 4.20 kann eine Darstellung der Ermittlung der Intervallgrenzen analog zu den Darstellungen im vorigen Kapitel gefunden werden; die waagrechten Referenz-Linien stellen wieder die KQ-Schätzung und das zugehörige Konfidenzintervall dar.

Werden die erhaltenen Parameter gemäß der Gleichungen (4.32) – (4.34) transformiert, ergeben sich folgende Werte:

$$\mathbb{E}[\beta_{\texttt{age}} \mid \log(\texttt{fib})] = [-0.0314,\, 0.0371] \qquad \mathbb{V}(\beta_{\texttt{age}} \mid \log(\texttt{fib})) = [6.13 \cdot 10^{-7},\, 1.02 \cdot 10^{-6}]$$
$$\mathbb{E}[\beta_{\texttt{bmi}} \mid \log(\texttt{fib})] = [-0.0451,\, 0.0544] \qquad \mathbb{V}(\beta_{\texttt{bmi}} \mid \log(\texttt{fib})) = [3.15 \cdot 10^{-6},\, 5.26 \cdot 10^{-6}]$$
$$\mathbb{E}[\beta_0 \mid \log(\texttt{fib})] = \quad [-2.67,\, 4.43]$$

Diese messen jetzt den Einfluss auf der ‚ursprünglichen' Skala und können mit dem Ergebnis der KQ-Schätzung verglichen werden:

$$\hat{\beta}_{\texttt{age}} = 0.00762 \qquad \mathbb{V}(\hat{\beta}_{\texttt{age}}) = 1.54 \cdot 10^{-6}$$
$$\hat{\beta}_{\texttt{bmi}} = 0.0131 \qquad \mathbb{V}(\hat{\beta}_{\texttt{bmi}}) = 7.93 \cdot 10^{-6}$$
$$\hat{\beta}_0 = 0.342$$

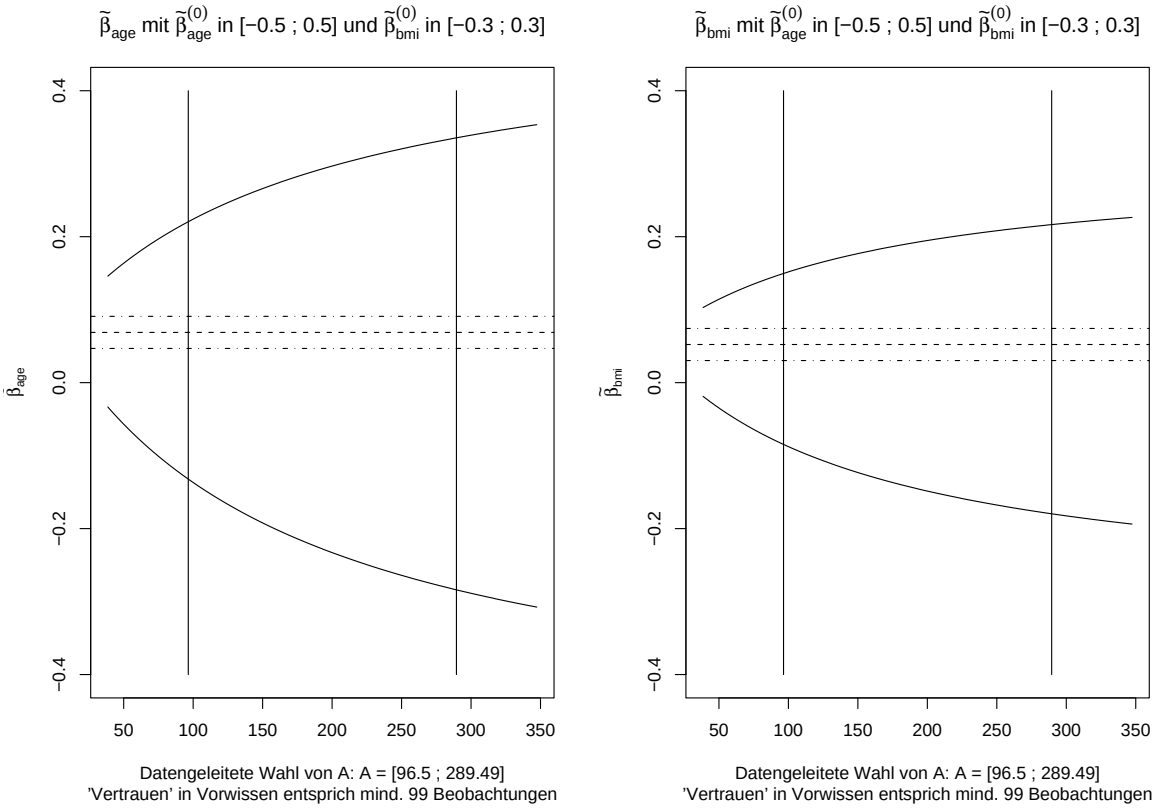

Abbildung 4.20: AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A ‚datengeleitet', Koeffizienten zu standardisierten Regressoren

Wie sich zeigt, können bei der datengleiteten Wahl von A die Parameter noch nicht von Null unterschieden werden, da die intervallwertige Punktschätzung für beide Regressoren die Null überdeckt. Da aufgrund des Wertebereichs von `age` und `bmi` die Parameterschätzungen für die standardisierten Variablen in sehr kleine Werte transformiert werden müssen, bewegen sich auch die zugehörigen Varianzen nach der Transformation auf einer noch kleineren Skala als zuvor. Gemäß der theoretischen Erkenntnisse aus Kapitel 4.3.3 sind sie jeweils kleiner als die Varianzschätzungen für $\hat{\beta}$.

In Abbildung 4.21 ist die Situation für die ‚echten' Parameter des Modells dargestellt; die oben aufgeführten KQ-Schätzungen sind zum Vergleich eingetragen. Auch hier zeigt sich die Weite der Posteriori-Intervalle im Vergleich zu den klassischen Konfidenzintervallen; zur besseren Anschaulichkeit sind die Schätzungen für $\beta_{\texttt{age}}$ und $\beta_{\texttt{bmi}}$ auf verschiedenen Skalen angegeben. Natürlich enthalten auch die ‚zurücktransformierten' intervallwertigen Punktschätzungen die Null.

Wie in Kapitel 4.4 sollen auch hier die simultane Konfidenzregion und die erhaltene Kredibilitätsregion dargestellt werden; für die standardisierten Koeffizienten können

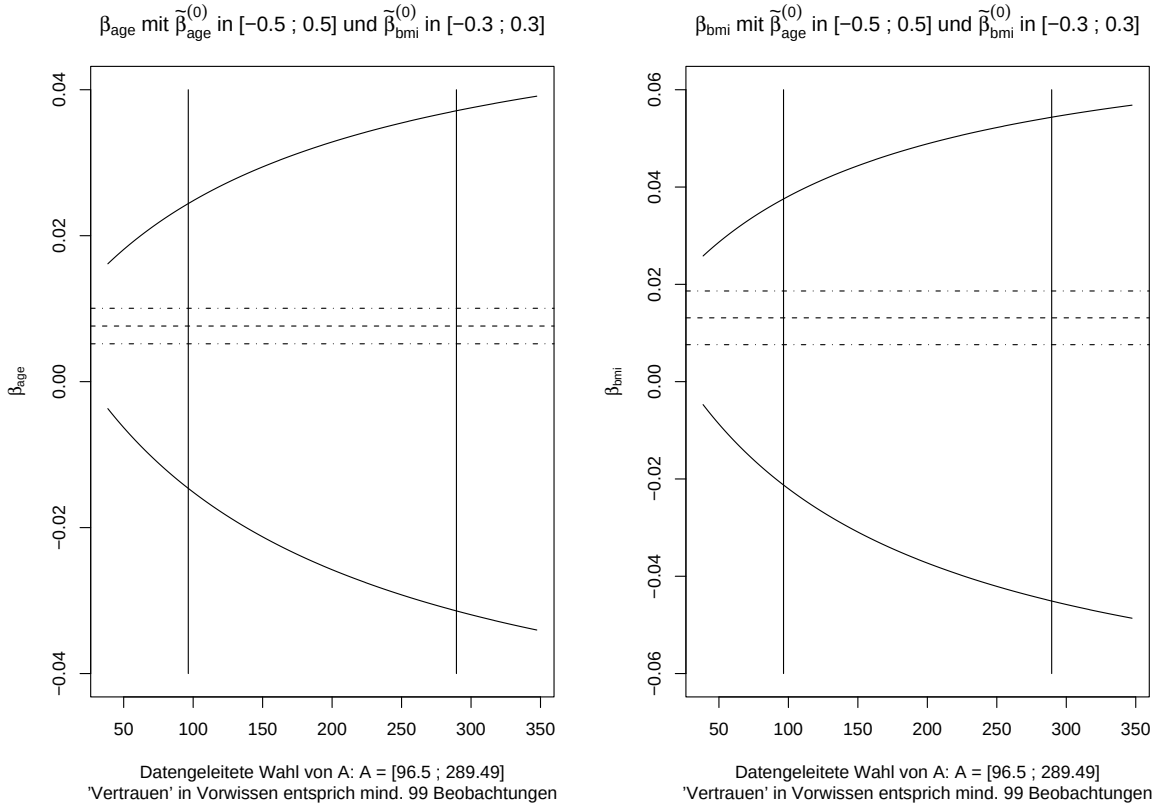

Abbildung 4.21: AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A ‚datengeleitet', Koeffizienten zu ‚normalen' Regressoren des Confounder-Modells

diese in Abbildung 4.22, für die ‚echten', nicht standardisierten Koeffizienten in Abbildung 4.23 gefunden werden.

An Abbildung 4.22 wird noch einmal der eklatante Größenunterschied zwischen der simultanen Konfidenzregion und der Kredibilitätsregion bei der ‚datengeleiteten' Wahl von A deutlich. Die Tatsache, dass die simultane Konfidenzregion fast kreisförmig ist, bestätigt die inhaltliche Vermutung, dass die Korrelation zwischen den Regressoren `age` und `bmi` als sehr niedrig angesehen werden kann.

In Abbildung 4.23 ist die Kredibilitätsregion für die Parameter dargestellt, die zu den ‚echten', nicht standardisierten Regressoren gehören. Aufgrund der unterschiedlichen Transformations-Faktoren für $\tilde{\beta}_{\texttt{age}}$ und $\tilde{\beta}_{\texttt{bmi}}$ ändert sich die Form der intervallwertigen Punktschätzung von einem ‚liegenden' zu einem ‚stehenden' Rechteck; aus diesem Grund ist, anders als in Abbildung 4.22, auch die Interpretation der Form der simultanen Konfidenzregion im Hinblick auf Kollinearität nicht mehr so einfach möglich.

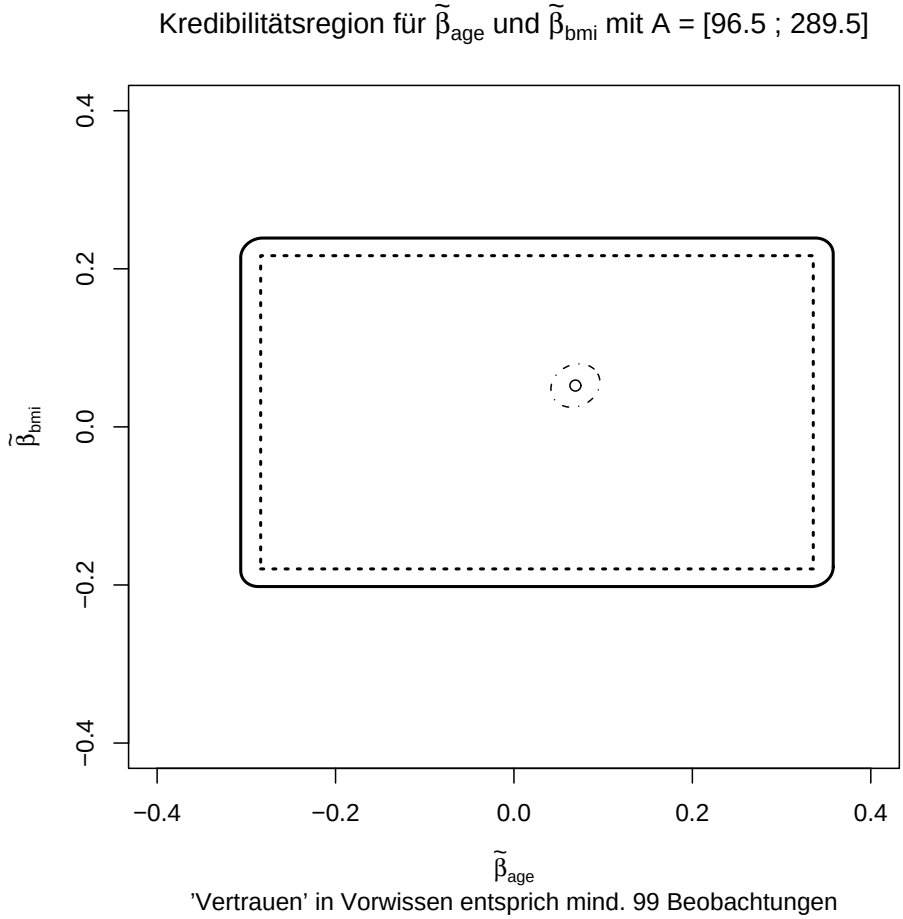

Abbildung 4.22: Kredibilitätsregion für $\tilde{\beta}_{\texttt{age}}$ und $\tilde{\beta}_{\texttt{bmi}}$, AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A ‚datengeleitet', Koeffizienten zu standardisierten Regressoren

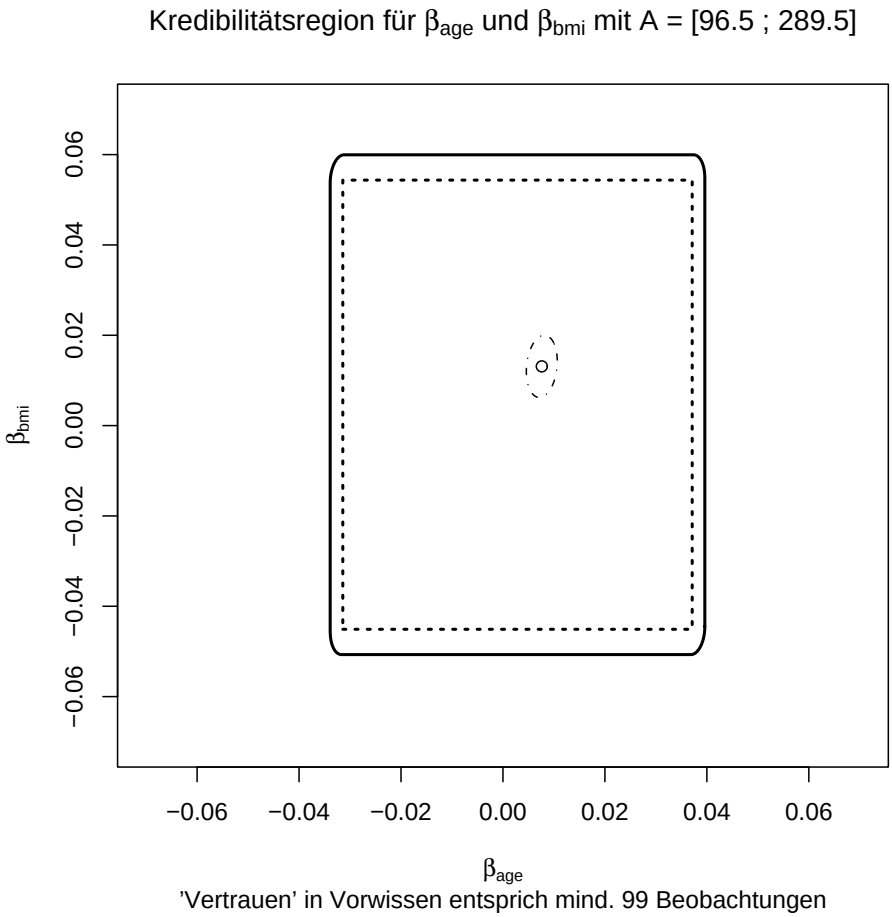

Abbildung 4.23: Kredibilitätsregion für $\beta_{\texttt{age}}$ und $\beta_{\texttt{bmi}}$, AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A ‚datengeleitet', Koeffizienten zu ‚normalen' Regressoren des Confounder-Modells

4.5.3 Analyse bei der Wahl von $n^{(0)}$ gemäß einer Strategie der Modellierung von sehr schwachem Vorwissen

In diesem Abschnitt soll nun eine Auswertung vorgenommen werden, bei der den Informationen in der Stichprobe mehr vertraut wird als bei der Auswertung im vorigen Abschnitt. Analog zur Vorgehensweise in Kapitel 4.4 soll nun das ‚Stichprobenäquivalent' $n^{(0)}$ auf Werte zwischen einer und zwei hypothetischen Beobachtungen festgelegt werden, um der Modellierung von a priori Nichtwissen nahe zu kommen; daraus ergibt sich, bei der Beibehaltung der gut begründeten Wahl von $B_{\texttt{age}} = [-0.5, 0.5]$ und $B_{\texttt{bmi}} = [-0.3, 0.3]$, ein Wertebereich von $A = [2.94\,, 5.88]$. Unter diesen Voraussetzungen können die standardisierten Koeffizienten wie folgt geschätzt werden:

$$\mathbb{E}[\tilde{\beta}_{\texttt{age}} \mid \log(\texttt{fib})] = [0.050,\ 0.083] \qquad \mathbb{V}(\tilde{\beta}_{\texttt{age}} \mid \log(\texttt{fib})) = [0.000122,\ 0.000123]$$
$$\mathbb{E}[\tilde{\beta}_{\texttt{bmi}} \mid \log(\texttt{fib})] = [0.039,\ 0.062] \qquad \mathbb{V}(\tilde{\beta}_{\texttt{bmi}} \mid \log(\texttt{fib})) = [0.000122,\ 0.000123]$$

Die zugehörige Darstellung befindet sich in Abbildung 4.24.

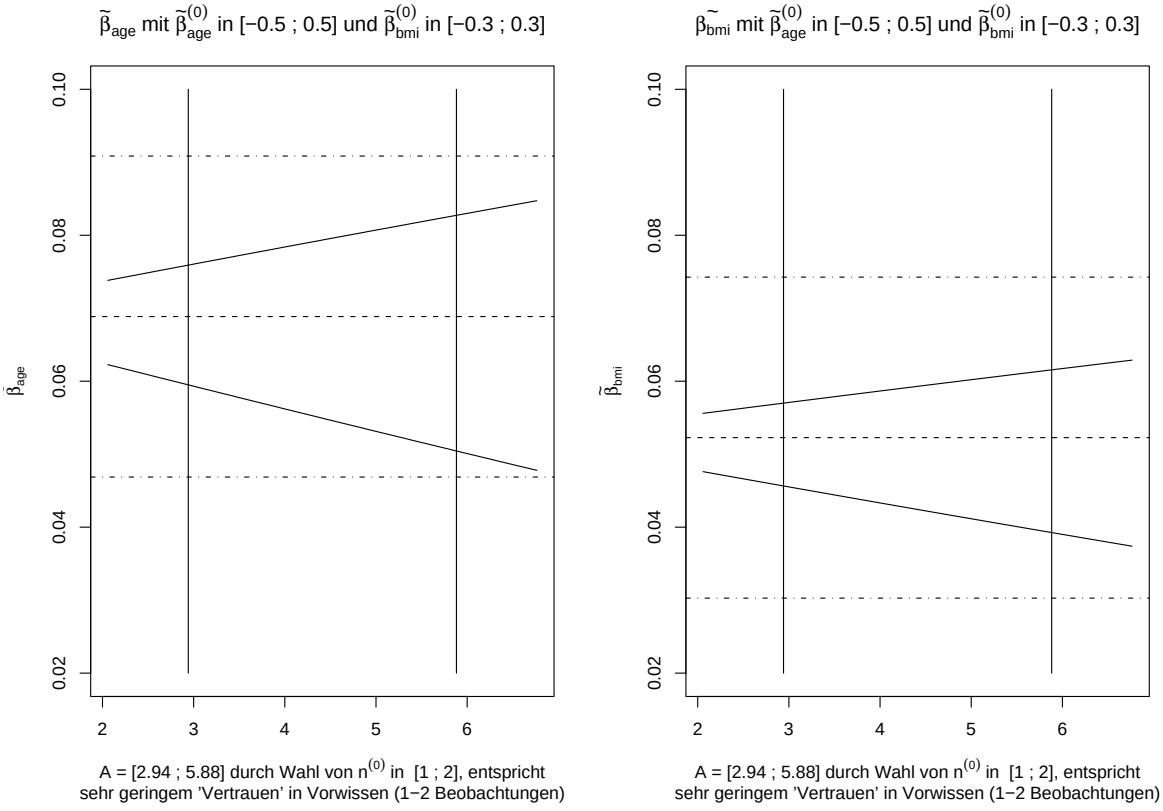

Abbildung 4.24: AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A, um sehr schwaches Vorwissen zu modellieren, Koeffizienten zu standardisierten Regressoren

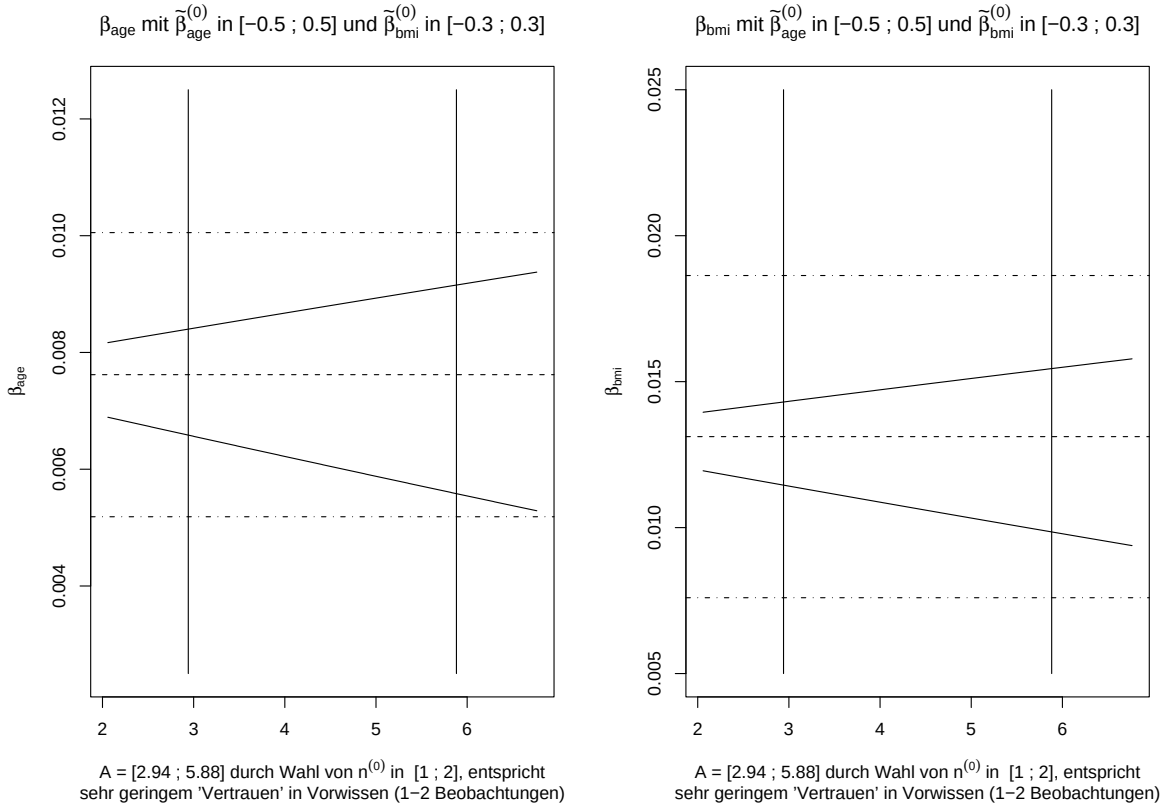

Abbildung 4.25: AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A, um sehr schwaches Vorwissen zu modellieren, Koeffizienten zu ‚normalen' Regressoren des Confounder-Modells

Die Grenzen der intervallwertigen Parameterschätzungen haben nun einen ähnlichen Wertebereich wie die Ränder der Konfidenzintervalle erreicht; bei dieser Wahl von A überdecken die Schätzungen der Koeffizienten beider Regressoren nicht mehr die Null. Für eine bessere Interpretierbarkeit der erhaltenen Koeffizienten sollen die ‚rücktransformierten' Parameterintervalle notiert und in Abbildung 4.25 graphisch dargestellt werden.

$$\begin{aligned}
\mathbb{E}[\beta_{\texttt{age}} \mid \log(\texttt{fib})] &= [0.00558,\, 0.00915] & \mathbb{V}(\beta_{\texttt{age}} \mid \log(\texttt{fib})) &= [1.49 \cdot 10^{-6},\, 1.52 \cdot 10^{-6}] \\
\mathbb{E}[\beta_{\texttt{bmi}} \mid \log(\texttt{fib})] &= [0.00985,\, 0.01545] & \mathbb{V}(\beta_{\texttt{bmi}} \mid \log(\texttt{fib})) &= [7.69 \cdot 10^{-6},\, 7.81 \cdot 10^{-6}] \\
\mathbb{E}[\beta_0 \mid \log(\texttt{fib})] &= \quad [0.180,\, 0.562]
\end{aligned}$$

Das Modell unter Berücksichtigung komplexer Unsicherheit kann also bei dieser Wahl von A folgendermaßen notiert werden:

$$\log(\texttt{fib})_i \;=\; [0.180,\, 0.562] + \texttt{age}_i \cdot [0.00558,\, 0.00915] + \texttt{bmi}_i \cdot [0.00985,\, 0.01545] + \varepsilon_i$$

Der Eindruck aus Abbildung 4.25 bestätigt das zuvor über die standardisierten Parameter gewonnene Bild. Auch hier wird deutlich, dass von einem Einfluss der

Regressoren `age` und `bmi` auf den Entzündungsmarker-Spiegel im Blut ausgegangen werden kann. Allerdings ist die absolute Größe der Parameterschätzungen hier nicht direkt interpretierbar; aus den Schätzungen der standardisierten Koeffizienten kann jedoch geschlossen werden, dass der Einfluss von `age` vermutlich etwas stärker als der von `bmi` ist.

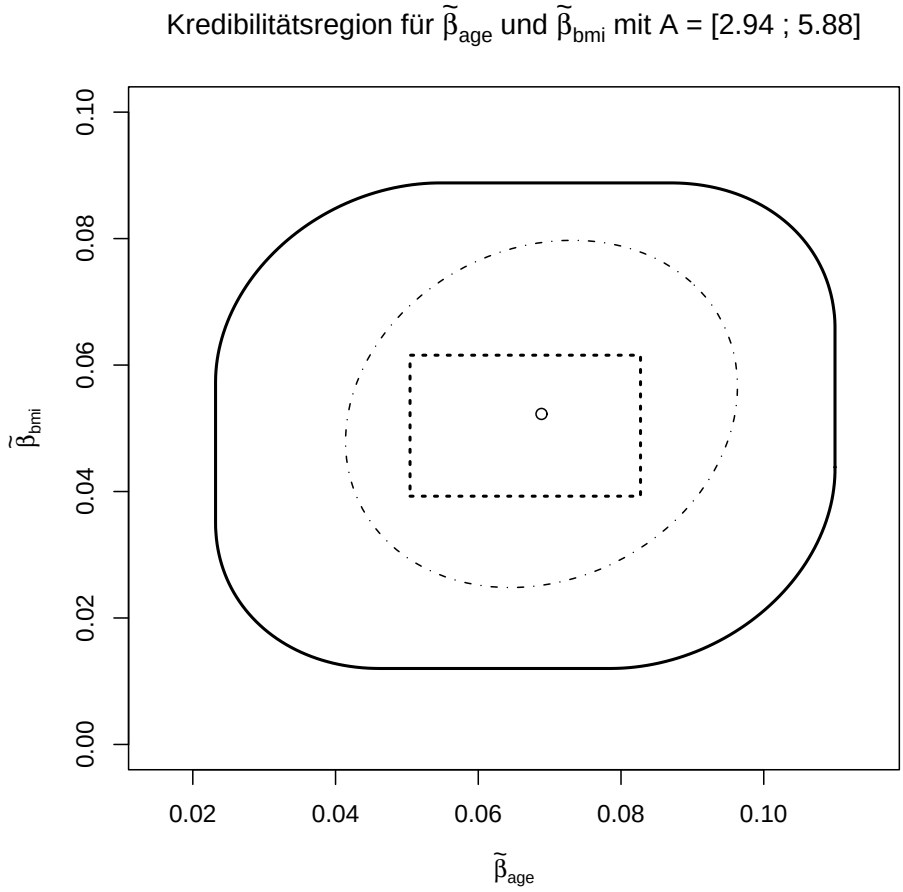

Abbildung 4.26: Kredibilitätsregion für $\tilde{\beta}_{\texttt{age}}$ und $\tilde{\beta}_{\texttt{bmi}}$, AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A, um sehr schwaches Vorwissen zu modellieren, Koeffizienten zu standardisierten Regressoren

Die ‚rücktransformierten' Varianzen sind im Vergleich zur vorsichtigeren Modellierung im vorigen Abschnitt etwas größer geworden; für beide Regressoren ist die Obergrenze schon recht nahe an die KQ-Schätzung der Varianz herangerückt, die Intervalllänge hat sich deutlich reduziert.

Der ‚Abstand' zwischen der Kredibilitätsregion und dem jeweils zugrundeliegenden zweidimensionalen Parameterintervall unterscheidet sich im Vergleich der hier erhalte-

nen Region, die in Abbildung 4.26 dargestellt ist, mit der in Abbildung 4.22 erhaltenen Region kaum, wenn man den unterschiedlichen Maßstab beachtet. Der schon in Kapitel 4.4 erwähnte Effekt des ‚trade-off' für die Varianz im Normal-Modell unter komplexer Unsicherheit ist also beim Vergleich der hier erhaltenen mit der bei einer vorsichtigeren Modellierung erhältlichen Kredibilitätsregion wesentlich weniger deutlich als im Falle der simulierten Datensätze. Der Grund dafür liegt in dem sehr niedrigen Varianzniveau der Koeffizientenschätzungen für diesen Datensatz, so dass die Erhöhung der Varianz bei der Verkleinerung der Werte in A kaum ins Gewicht fällt.

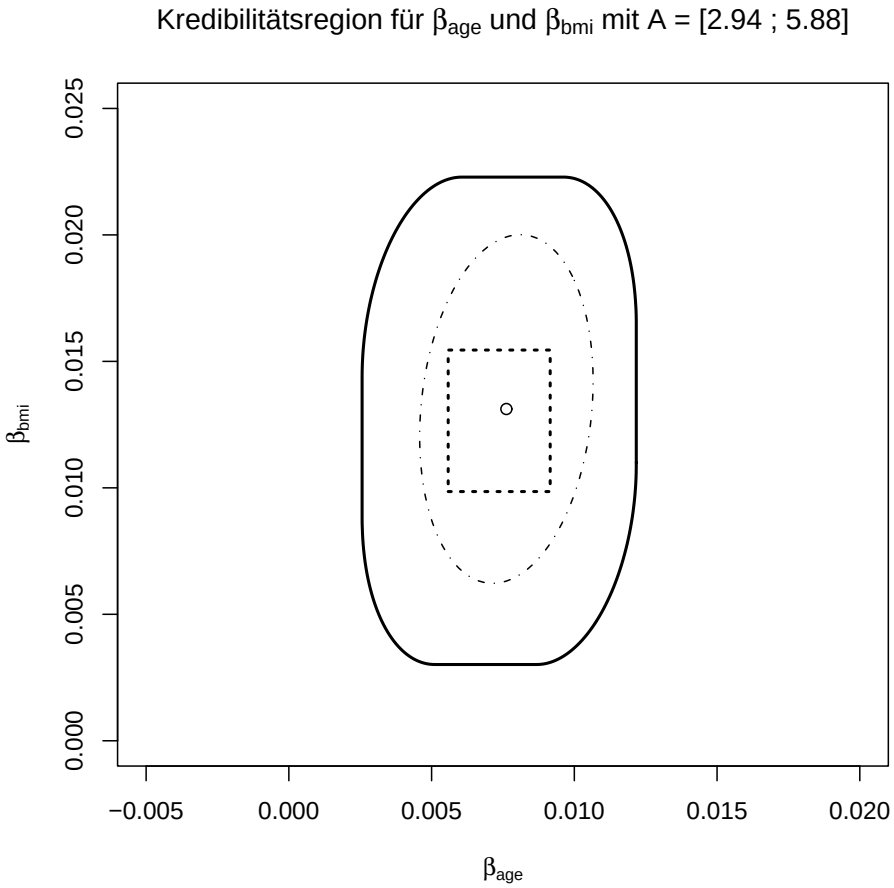

Abbildung 4.27: Kredibilitätsregion für $\beta_{\texttt{age}}$ und $\beta_{\texttt{bmi}}$, AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A, um sehr schwaches Vorwissen zu modellieren, Koeffizienten zu ‚normalen' Regressoren des Confounder-Modells

Abbildung 4.27 zeigt schließlich die Konfidenz- und Kredibilitätsregion für die ‚rücktransformierten' Parameter. Auch hier wird durch die unterschiedlich skalierten Regressoren die Kredibilitätsregion gewissermaßen verzerrt; die Parameter werden durch die Transformierung interpretierbar, an den schon in Abbildung 4.26 ables-

baren grundsätzlichen Aussagen ändert sich jedoch nichts: Hier wie in Abbildung 4.26 überdeckt die Kredibilitätsregion nicht den Nullpunkt, hier wie dort liegt die intervallwertige Punktschätzung vollständig in der klassischen Konfidenzregion, die jeweils von der Kredibilitätsregion umschlossen wird.

Daraus, dass die Kredibilitätsregion nicht den Nullpunkt enthält, könnte geschlossen werden, dass sowohl `age` als auch `bmi` einen nicht zu vernachlässigenden Einfluss auf log(`fib`) haben, obwohl die intervallwertigen Parameterschätzungen für die standardisierten Regressoren nur jeweils Werte kleiner als 0.1 enthalten und der Einfluss der Variablen also als recht klein angesehen werden kann.

Die bestehenden Erkenntnisse bezüglich des Einflusses des Alters und des BMI auf das Niveau des Entzündungsmarkers Fibrinogen erweisen sich also auch bei der Berücksichtigung von komplexer Unsicherheit als stabil. Die hier vorgenommene Untersuchung kann also die bestehenden Erkenntnisse bekräftigen und auf eine bezüglich der Modell-Annahmen unangreifbarere Basis stellen.

5 Zusammenfassung und Ausblick

In diesem Kapitel sollen die Ergebnisse dieser Arbeit zusammengefasst und Ausblicke auf mögliche Erweiterungen und prinzipielle Überlegungen bei der Modellierung komplexer Unsicherheit in der Regressionsanalyse beschrieben werden.

5.1 Zusammenfassung

In Kapitel 1 wurde eine kurze Einführung in die Materie der Modellierung komplexer Unsicherheit gegeben. Einige Probleme und Paradoxien, die sich bei der Anwendung von Methoden der ‚konventionellen' Statistik ergeben können, wurden thematisiert und daraus die Motivation für eine allgemeinere statistische Theorie abgeleitet, sowie eine Reihe von zusätzlichen Argumenten für eine solche Erweiterung der Statistik genannt. Die prinzipiellen Vorgehensweisen bei einer Modellierung komplexer Unsicherheit, insbesondere die Ansätze von [Walley 1991] und [Weichselberger 2001], wurden dann in Abschnitt 1.5 vorgestellt, sowie mit diesen Ansätzen verwandte Theorien kurz erwähnt. In Abschnitt 1.7 wurde dann umrissen, nach welcher Vorgehensweise in dieser Arbeit die Berücksichtigung komplexer Unsicherheit in der Regressionsanalyse erfolgt, und diese mit verschiedenen alternativen Vorgehensweisen verglichen.

Kapitel 2 lieferte dann als detailliertes Beispiel eines Modells, das komplexe Unsicherheit berücksichtigt, die Beschreibung des Imprecise Dirichlet Models (IDM, [Walley 1996]). Anhand dieses Modells für multinomial verteilte Daten wird deutlich, wie solche Modelle implementiert werden können, auf welche Weise damit Inferenz betrieben werden kann und welche Vorteile sich dabei im Vergleich mit konventionellen Modellierungen ergeben. Aufgrund seiner vorteilhaften Inferenzeigenschaften und der einfachen Handhabbarkeit hat das IDM vielfache Anwendung gefunden (siehe z.B. [Bernard 2001] oder [Bernard 2003]). In Abschnitt 2.2 und 2.3 wurde dabei das Prinzip der bayesianischen Aufdatierung im Falle von konjugierten Priori-Verteilungen erläutert, das auch in der in dieser Arbeit vorgestellten Schätzung von Regressionsparametern unter komplexer Unsicherheit zur Anwendung kommt.

In Kapitel 3 erfolgte dann die Vorstellung des Modells von [Quaeghebeur und de Cooman 2005], das eine Verallgemeinerung des IDM darstellt, da es allgemeine Stichproben-Modelle unter Berücksichtigung komplexer Unsicherheit beschreibt. Wie in Abschnitt 3.3 deutlich wurde, kann im Falle der meisten üblichen Stichprobenverteilungen damit die konventionelle bayesianische Parameterschätzung auf

einfache Weise erweitert werden, so dass die Berücksichtigung komplexer Unsicherheit bezüglich der Priori-Verteilung möglich wird. Es wurde aber auch darauf hingewiesen, dass das Modell in der in [Quaeghebeur und de Cooman 2005] vorgestellten Form eine wesentliche Anforderung für eine sinnvolle Modellierung komplexer Unsicherheit noch nicht erfüllt, da es im Falle des in Abschnitt 1.4.3 erläuterten ‚prior-data conflict' unerwünschtes Verhalten zeigt. Auf welche Weise eine dahingehende Erweiterung vorgenommen werden kann, wurde jedoch angedeutet.

Das Ziel dieser Arbeit, ein Modell unter Berücksichtigung komplexer Unsicherheit für die Schätzung der Parameter β einer linearen Regression zu entwickeln, wurde nach Abschluss der umfangreichen Vorarbeiten in den vorigen Kapiteln in Kapitel 4 verfolgt. In Abschnitt 4.2.1 wurde das Normal-Modell, das konjugierte Modell für die Schätzung von β vorgestellt. Danach wurde in Kapitel 4.2.3 die konjugierte Verteilung des Normal-Modells direkt in die Form der konjugierten Verteilung bei [Quaeghebeur und de Cooman 2005] gebracht. Auf dieser Basis wird es möglich, die Menge der Priori-Verteilungen analog zum Vorgehen bei Quaeghebeur und de Cooman zu generieren. (Eine andere Möglichkeit zur Erstellung eines Modells zur Schätzung von β unter Berücksichtigung komplexer Unsicherheit mittels des Modells von Quaeghebeur und de Cooman wäre, statt von der konjugierten Verteilung des Normal-Modells von der Likelihood der Daten auszugehen. Diese Vorgehensweise wurde in dieser Arbeit jedoch nicht untersucht, sie könnte aber ebenso zielführend wie der in dieser Arbeit untersuchte Ansatz sein.)

Leider erweist sich, dass das umfassendere konjugierte Modell für die lineare Regression, das Normal-InversGamma- (NIG-) Modell, bei dem auch eine Priori-Verteilung über σ^2 (die unbekannte Varianz des Störterms ε) gelegt wird, nicht in der gleichen Weise wie in Kapitel 4.2.3 verallgemeinerbar ist. Eine Vorstellung dieses Modells und eine Begründung für die Unmöglichkeit der direkten Erweiterung analog zum Normal-Modell erfolgt im Anhang, Kapitel A.1.

Wie in Abschnitt 4.2.3 gezeigt, kann man das Normal-Modell als ein Aufdatierungsmodell im Sinne von [Quaeghebeur und de Cooman 2005] verstehen; im allgemeinen Fall für eine beliebige Anzahl p von zu schätzenden Regressionsparametern ist die tatsächliche Implementierung jedoch schwierig, wie in Abschnitt 4.2.4 deutlich gemacht wird. Der Grund dafür ist, dass der natürliche Parameter y der konjugierten Verteilung in der Form, in der er variiert wird, um die Menge der Priori-Verteilungen zu generieren, nicht unmittelbar interpretierbar ist: Für eine nachvollziehbare Anwendung muss nämlich das Vorwissen in Intervalle für die klassischen Parameter formuliert und diese Intervalle in eine Menge $\mathcal{Y}^{(0)}$ der natürlichen Parameter ‚übersetzt' werden. Die Menge der natürlichen Parameter kann dann zwar linear zu $\mathcal{Y}^{(1)}$ aufdatiert werden, sie muss aber wieder in interpretierbare Intervalle für die klassischen Parameter ‚rückübersetzt' werden.

Problematisch bei diesen ‚Übersetzungsvorgängen' ist nun, dass nach [Quaeghebeur und de Cooman 2005] die Menge $\mathcal{Y}^{(0)}$ der natürlichen Parameter, in die die Menge der klassischen Parameter im ersten Schritt gewissermaßen übersetzt wird, bestimmten Bedingungen gehorchen muss, damit das Modell sinnvolle Aussagen liefern kann. Das bedeutet, dass die Intervallgrenzen der klassischen Parameter nur so gewählt dürfen, dass bei der Übersetzung die Bedingungen für $\mathcal{Y}^{(0)}$ nicht verletzt werden können. Weil es sich im Falle der multivariaten Normalverteilung um schwierig zu handhabende Bedingungen der positiven Definitheit handelt (siehe Gleichungen (4.9) und (4.10)), können die Bedingungen für eine zulässige Wahl der Intervallgrenzen für die Priori-Parameter in ihrer klassischen, interpretierbaren Form nur über Minimierungen bzw. Maximierungen mit sehr komplexen Nebenbedingungen erhalten werden.

Diese Schwierigkeiten beim rechnerischen Übergang nach $\mathcal{Y}^{(0)}$ waren nicht vorherzusehen; es handelt es sich dabei jedoch nicht um ein theoretisches Problem des Modells, sondern um eine Hürde bei seiner numerischen Umsetzung.

Aufgrund der ‚Übersetzungsschwierigkeiten' ist es daher nicht möglich, im Falle von vielen Regressoren die Bedingungen an die Priori-Intervalle der klassischen Parameter analytisch zu ermitteln. In dieser Arbeit wurde daher im Detail nur der Fall von zwei Regressoren rechnerisch umgesetzt, da sich hier die Minimierungs- und Maximierungsprobleme unter den erwähnten Nebenbedingungen stark vereinfachen und analytisch zugänglich werden. Diese Analyse wurde in Abschnitt 4.3 vorgenommen. Dort wird aber auch deutlich, dass das Modell für eine praktikable Anwendung weiter vereinfacht werden muss, indem die Korrelation der beiden Regressoren untereinander a priori als Null angenommen wird.

Unter diesen Vereinfachungen wird das Modell rechnerisch simpel anwendbar, da so auch die ‚rückübersetzten' a posteriori Grenzen für die klassischen Parameter direkt als einfache Funktionen der Mengen der klassischen Priori-Parameter berechenbar sind. Sie können z.T. auch direkt in geschlossener Form angegeben werden, da aufgrund von Monotonitätseigenschaften Minimierungen und Maximierungen ohne Kenntnis der konkreten Priori-Grenzwerte und Beobachtungen möglich sind.

Die Berechnung von unteren und oberen Grenzen von Wahrscheinlichkeitsgewichten für die Erstellung von Kredibilitätsintervallen oder für die Testentscheidung bei Hypothesentests ist jedoch noch immer schwierig. Es müssen dafür mehrdimensionale Minimierungs- und Maximierungsprobleme bezüglich der Parameter der konjugierten Verteilung gelöst werden, wobei der zu minimierende Term ein mehrdimensionales Integral darstellt, das nicht analytisch berechenbar ist. Aus diesem Grund wird bei der Anwendung des Modells in den Kapiteln 4.4 und 4.5 nur ein Surrogat für ein korrektes Kredibilitätsintervall berechnet, dessen Erzeugung in Abschnitt 4.4.1 näher erläutert wird. Aus diesen Näherungen für ein korrektes Kredibilitätsintervall können aber auch mögliche Testentscheidungen abgelesen werden.

Ansonsten ist das Modell in der Anwendung, wie schon erwähnt, bezüglich der Berechnung der Posteriori-Grenzen für die klassischen Parameter recht einfach handhabbar. Das Verhalten bei der Angabe der Priori-Intervalle B_1 und B_2 für $\beta^{(0)}$, den Erwartungswert der konjugierten Verteilung, ist sehr intuitiv und unmittelbar einleuchtend, da deren Größe einen direkten Einfluss auf die Größe der Intervalle nach der Aufdatierung hat: Werden die Priori-Intervalle vergrößert, vergrößern sich die Posteriori-Intervalle ebenfalls.

Der Einfluss der Wahl von A, der Menge der Priori-Präzision bezüglich der Werte in B_1 und B_2 (die als inverse Varianz verstanden werden kann), scheint hingegen auf den ersten Blick kontra-intuitiv. Üblicherweise gilt nämlich[1]: je größere Werte sich in A befinden, desto weiter werden die a posteriori erhaltenen Intervalle für $\beta^{(1)}$; je höher man die genannte Priori-Präzision ansetzt, desto größer ist die Unschärfe bezüglich β nach der Aufdatierung. Unter einer solch verkürzten Interpretation, wie sie im letzten Halbsatz wiedergegeben wurde, scheint dieses Verhalten tatsächlich widersinnig; bei genauerer Betrachtung erschließt sich aber die Logik hinter diesem Verhalten des Modells. Mit höheren Werten in A verkleinert man ja die Varianz der Werte in B_1 und B_2 (den a priori angegebenen Intervallen für $\beta^{(0)}$) und vergrößert somit das Vertrauen, dass man in diese Intervalle setzt. Da die Angaben von B_1 und B_2 also als ‚sicherer' wie zuvor gelten, fallen sie bei der Aufdatierung stärker ins Gewicht; die Posteriori-Grenzen müssen also näher an den Priori-Grenzen zu liegen kommen als zuvor, das Intervall wird somit weiter. Weniger technisch formuliert heißt das, dass man mit der Angabe einer Menge von ‚schwammigen' Verteilungen für $\beta^{(0)}$ mehr dazulernen kann, als wenn man schon ein sehr pointiertes Vorwissen hat und die Varianz der im Vorwissen angegebenen Einschätzungen bezüglich $\beta^{(0)}$ für klein hält.

Besonders deutlich wird das Verhalten des Modells bezüglich A, wenn man den resultierenden Wert von $n^{(0)}$ in die Überlegungen miteinbezieht. $n^{(0)}$ stellt die Stärke des in den Priori-Angaben A, B_1 und B_2 formulierten Vorwissens dar und kann im Vergleich mit dem Umfang k der zu analysierenden Stichprobe interpretiert werden: je höher $n^{(0)}$ im Vergleich zu k, desto stärker gehen die Informationen aus dem Vorwissen in die Posteriori-Inferenz ein. Aufgrund der Bedingung (4.18), die im untersuchten Fall die Nebenbedingung bei der obengenannten ‚Übersetzung' der klassischen in die natürlichen Parameter darstellt, impliziert eine höhere Wahl von $\overline{a}$ bei gleichbleibenden Priori-Intervallen B_1 und B_2, dass der Wert des ‚Stichprobenäquivalents' $n^{(0)}$ steigt. Das bedeutet, dass das ‚Vertrauen' in das Vorwissen im Vergleich mit dem ‚Vertrauen' in die Stichprobe der Größe k steigt, dass deshalb das Gewicht der exakten Beobachtung bei der Aufdatierung im Vergleich mit den unscharfen Priori-Angaben sinkt und somit die resultierenden Intervalle breiter werden. Setzt man hohe Werte in A an, vertraut man den Priori-Angaben unter Umständen mehr als den Erkenntnissen aus der Stichprobe und ist somit nicht bereit, bei der Analyse von kleineren und mittleren Stichproben viel

[1] Nur in Ausnahmefällen und bei einer extremen Wahl von A, B_1 und B_2 kann sich das Posteriori-Intervall verkleinern, wenn sich die Werte in A vergrößern.

dazuzulernen; die Priori-Intervalle können sich so nur wenig verschmälern.

Die Wahl der Intervalle A, B_1 und B_2, die für die Inferenz in dem hier entwickelten Modell nötig ist, sollte natürlich auf die jeweilige Anwendung und das vorhandene Vorwissen abgestimmt werden. Hilfreich ist dabei immer die Berechnung des ‚Stichprobenäquivalents' $n^{(0)}$, das eine Interpretation der gemeinsamen Aussagekraft des in A, B_1 und B_2 ausgedrückten Vorwissens ermöglicht.

Erfreulicherweise sind die Resultate dieses Modells annähernd exakt, wenn man A, B_1 und B_2 gemäß einer Strategie der Modellierung von Nichtwissen wählt, bei welcher das Stichprobenäquivalent' $n^{(0)}$ sehr niedrig angesetzt wird. Ist man aber überzeugt davon, dass zu schmale Posteriori-Intervalle für β der Intention der Modellierung von komplexer Unsicherheit widersprechen und nur eine ‚Pseudo-Genauigkeit' darstellen, sollte man den Wert von $n^{(0)}$ erhöhen, um (im Verhältnis zum Vorwissen) weniger Vertrauen in die Daten zu setzen. Gegen eine Strategie, B_1 und B_2 weniger umfassend zu wählen, um $\overline{a}$ zu vergrößern und somit zu weiteren Posteriori-Intervallen gelangen zu können, spricht die Problematik des ‚prior-data conflict'. Da dieses Modell, wie sich aus den allgemeinen Ergebnissen in Kapitel 3.3 ergibt und die bei der Untersuchung des konkret erhaltenen Modells in Abschnitt 4.3.3 bestätigt werden, in solchen Situationen nur ein unbefriedigendes Verhalten zeigen kann, sollten die Priori-Intervalle B_1 und B_2 möglichst umfassend gewählt werden.

Prinzipiell wäre es wünschenswert, das konkrete Verhalten des Modells in den verschiedensten Datensituationen und Priori-Annahmen detaillierter zu untersuchen, als es in dieser Arbeit aufgrund ihres theoretischen Schwerpunkts möglich war. Ein Anfang wurde in Kapitel 4.4 gemacht; einen interessanten Hinweis auf das Verhalten in Situationen mit Multikollinearität liefert das Beispiel in Abschnitt 4.4.3.

Die Erkenntnisse bezüglich der resultierenden Mengen für den Posteriori-Parameter $\beta^{(1)}$ und $\mathbf{\Sigma}^{(1)}$, der Varianz-Kovarianzmatrix der Posteriori-Verteilung über β, die sich für die Grenzwerte bei der Wahl von a ergeben, wurden in Abschnitt 4.3.3 beschrieben und liefern jedoch schon einen relativ genauen und höchst plausiblen Rahmen. Je kleiner die Priori-Präzision der Werte in B_1 und B_2 gewählt wird, desto näher liegen die Grenzen des Intervalls für $\beta^{(1)}$ an $\hat{\beta}$; je größer a gewählt wird, desto höher ist das Vertrauen in das Vorwissen und desto näher liegen in den meisten Fällen die Grenzen des Intervalls für $\beta^{(1)}$ an denen der Priori-Intervalle B_1 und B_2. Ebenso gilt für die Posteriori-Menge der Varianz-Kovarianzmatrizen $\mathbf{\Sigma}^{(1)}$: Je kleiner die Priori-Präzision, desto näher liegt diese Menge an der KQ-Schätzung der Varianz von β; je größer a gewählt wird, desto höher ist das Vertrauen in das Vorwissen, und damit gilt $\mathbf{\Sigma}^{(1)} \to \mathbf{0}$.

Aufgrund dieses Verhaltens kann daher das in dieser Arbeit entwickelte Modell, insbesondere bei der Wahl von A, B_1 und B_2 gemäß einer Modellierung von sehr schwachem Vorwissen, auch als eine Art Robustifizierung der Ergebnisse der konventionellen Schätzmethoden angesehen werden. Da im untersuchten Fall die Schätzungen nach

der Methode der kleinsten Quadrate, die Likelihood-Schätzung und die bayesianische Schätzung bei der Annahme einer konstanten Priori-Verteilung für β zusammenfallen, handelt es sich um eine Robustifizierung aller dieser Methoden gleichermaßen, obwohl es sich eigentlich um einen bayesianischen Ansatz handelt.

Trotzdem ist das Modell natürlich noch nicht optimal anwendbar, da für die konkrete Untersuchung eines Datensatzes die Varianz des Fehlerterms σ^2 zuvor geschätzt werden muss. Auch wenn die Unsicherheit dieser Schätzung nicht explizit in das Normal-Modell eingeht, sollte doch dadurch, dass in den anderen Modellkomponenten durch die Einbeziehung von komplexer Unsicherheit eine sehr viel größere Variabilität als bei den konventionellen Methoden zugelassen wird, dieses Manko zumindest teilweise kompensiert werden können.

Eine andere Schwierigkeit bei der Anwendung ist, dass im Vergleich zu anderen Modellierungen unter komplexer Unsicherheit a priori relativ viele Festlegungen gemacht werden müssen, namentlich die Grenzen für die Mengen A, B_1 und B_2. Die Komplexität der Wahl kann reduziert werden, indem zunächst das ‚Stichprobenäquivalent' $n^{(0)}$ und die potentiell mögliche maximale Größenordnung des Regressionsparameters β durch $B = B_1 = B_2$ festgelegt wird, und in einem zweiten Schritt daraus A ermittelt wird. Auf diese Weise wurde in Kapitel 4.5 bei der Analyse eines Teildatensatzes, der im Rahmen der AIRGENE-Studie [Peters et al., eingereicht] erhoben wurde, vorgegangen. In jedem Falle kann eine konkrete Wahl der Priori-Grenzen für die Mengen A, B_1 und B_2 als diskussionswürdig angesehen werden.

Abschließend kann also festgehalten werden, dass das in dieser Arbeit entwickelte Modell zufriedenstellende Ergebnisse liefert, aber nur ein erster Versuch zur Modellierung von komplexer Unsicherheit bei der Schätzung von Regressionskoeffizienten sein kann. Es sind auch andere Möglichkeiten zur Modellierung von komplexer Unsicherheit für diese Fragestellung denkbar; ebenso könnte das hier vorgestellte Modell als Ausgangspunkt für vielfältige Erweiterungen oder Fortentwicklungen dienen. Sowohl auf ausgewählte Aspekte alternativer Modellierungen als auch auf sich anbietende Fortentwicklungen und wünschenswerte Erweiterungen soll im nachfolgenden Ausblick kurz eingegangen werden; begonnen wird dabei mit sich unmittelbar anbietenden weiterführenden Untersuchungen, um danach den Blick auf perspektivische Weiterentwicklungen zu richten.

5.2 Ausblick

Es wäre vermutlich lohnend, sich in Zukunft näher mit den im Normal-Modell auftretenden Optimierungsproblemen zu beschäftigen. Vielleicht können Algorithmen entwickelt werden, die die Berechnung der Wahrscheinlichkeitsgewichte für die Erstellung von Kredibilitätsintervallen oder für die Testentscheidung bei Hypothesentests beschleunigen

oder vereinfachen; ebenso wäre es denkbar, dass sich Algorithmen entwickeln lassen, die die Optimierungsprobleme im Zusammenhang mit den ‚Übersetzungsvorgängen' beherrschbarer machen, so dass auch Modellformulierungen im Falle von mehr als zwei Regressionsparametern rechnerisch praktikabel würden.

Außerdem könnte die Analyse des Modells bei der Verwendung von kategorialen Regressoren aufschlussreich sein, da sich für solche Regressoren die Schätzung der Intervallgrenzen für die Parameter wahrscheinlich wesentlich vereinfacht. Allgemein stellt sich natürlich die Frage, ob die bekannten Verallgemeinerungen des ‚scharfen' linearen Modells, etwa die Erweiterung auf nicht-normalverteilte Zielvariablen wie etwa bei den generalisierten linearen Modellen, auch auf die hier vorgenommene Modellierung unter Einbeziehung von komplexer Unsicherheit anwendbar sind. Einen möglichen Weg beispielsweise für binäre Zielvariablen bietet der Ansatz von [Holmes und Held 2006], der binäre Größen mittels einer Hilfsvariablen auf die Normalverteilungs-Situation zurückführt.

Ebenfalls könnte der schon im vorigen Abschnitt erwähnte Versuch gemacht werden, in analogem Vorgehen zu Quaeghebeur und de Cooman von der Likelihood der Daten auszugehen und sich die zugehörige konjugierte Verteilung in der gleichen Weise, wie es in Abschnitt 3.2 geschieht, zu konstruieren, anstatt die schon bekannte konjugierte Verteilung des Normal-Modells direkt an die konjugierte Verteilung im Modell von Quaeghebeur und de Cooman anzupassen. Vielleicht wären dann die etwas verstörenden Ergebnisse dieser Arbeit, wie beispielsweise die nur indirekte Abhängigkeit der Posteriori-Schätzungen von $n^{(0)}$, vermeidbar.

Im Zuge der Untersuchung der in Abschnitt 1.7 vorgestellten alternativen Modelle, insbesondere der Vorschläge von [Pericchi und Walley 1991], im Vergleich mit dem Modell von Quaeghebeur und de Cooman stellt sich eine vielleicht zentrale Frage bei der Einbeziehung komplexer Unsicherheit in bayesianische Modelle: Gibt es einen Zielkonflikt bezüglich der Modellierung, je nachdem ob prinzipiell a priori Nichtwissen oder aber ein gewisses, aber unscharfes Vorwissen modelliert werden soll? Walley und Pericchi scheinen diese Frage zu bejahen, da sie in der obengenannten Veröffentlichung je nach Situation des Vorwissens verschiedene Modellierungsansätze präsentieren und somit einerseits Modelle für den Einsatz unter a priori-Nichtwissen (translations-invariante Klassen), andererseits Modelle für den Einsatz unter relativ eingeschränktem unscharfen Vorwissen (Umgebungsmodelle) praktisch maßschneidern.

Oder sind Modelle möglich, die für beide Vorwissenssituationen gleichermaßen geeignet sind? Das Stichproben-Modell von Quaeghebeur und de Cooman ist in der in Kapitel 3 vorgestellten Form noch nicht für den Einsatz in letzterer Situation tauglich, da das Verhalten bei vorliegendem ‚prior-data conflict' ungenügend ist und noch nicht die Stärken, die Intervallwahrscheinlichkeitsmodelle in solchen Situationen haben können, ausspielt. Wie jedoch schon mehrmals erwähnt, scheint eine Erweiterung des Modells von Quaeghebeur und de Cooman auf direkte Weise durch die zusätzliche Variation

eines im Modell bisher als fest angesehenen Parameters möglich. Andererseits kann auch die Modellierung der ersteren Vorwissenssituation im Stichprobenmodell von Quaeghebeur und de Cooman als nicht ganz einwandfrei angesehen werden, da die unendlichen Räume der Parameter der Priori-Verteilung beschränkt werden müssen, so dass völliges Nichtwissen nicht perfekt modelliert werden kann. Im Falle der translations-invarianten Klassen von Pericchi und Walley scheint hingegen eine perfekte Modellierung von a priori Nichtwissen über einen Lageparameter möglich zu sein, da eine Einschränkung des Parameterraums nicht vorgenommen werden muss.

Es könnten also auch die in [Pericchi und Walley 1991] vorgestellten und in Abschnitt 1.7 beschriebenen Modelle als Basis für die Schätzung von Parametern einer linearen Regression unter komplexer Unsicherheit dienen. Bei den Empfehlungen von Pericchi und Walley handelt es dabei jeweils gewissermaßen um ‚Extremfälle' bei der Modellierung komplexer Unsicherheit. Für die Modellierung von völligem a priori Nichtwissen empfehlen sie, eine translations-invariante Klasse von Doppelexponentialverteilungen bayesianisch aufzudatieren. Bei einer näheren Untersuchung dieser Menge von Priori-Verteilungen in der Anwendung auf die lineare Regression könnten sich interessante Zusammenhänge herausstellen, da die Schätzmethode Lasso [Tibshirani 1996] in der bayesianischen Formulierung ebenfalls auf doppelexponentialverteilten Priori-Verteilungen beruht. Bei der Empfehlung für den anderen ‚Extremfall' handelt es sich um eine Umgebungsklasse für eine feste Normalverteilung, die möglicherweise auch als konjugierte Verteilung des Normal-Modells interpretierbar wäre. Vielleicht ist es auch möglich, die im Rahmen von Intervallwahrscheinlichkeitsmodellen relativ starken Anforderungen an dieses Modell (es muss eine konkrete Normalverteilung spezifiziert werden) aufzuweichen, indem eine Menge von Verteilungen als Basis für die Umgebungsklasse dient.

Problematisch könnte bei diesen Ansätzen jedoch sein, dass es sich bisher nur um Mengen von einparametrigen Verteilungen handelt, und eine Erweiterung auf mehrere Regressionsparameter nicht direkt ersichtlich scheint. Andererseits ist gerade dies im Modell von Quaeghebeur prinzipiell nicht problematisch; auch wenn es gewissermaßen, von den Vorwissenssituationen aus betrachtet, momentan noch ‚zwischen den Stühlen sitzt', hat es viele positive Eigenschaften, wie etwa die einfache Berechenbarkeit im Falle der meisten Stichprobenverteilungen oder seine Offenheit und umfassende Modellformulierung.

So scheint das Potential des umfassenden Ansatzes von Quaeghebeur und de Cooman zur Modellierung komplexer Unsicherheit noch lange nicht ausgereizt, da sich viele Erweiterungen anbieten. In erster Linie bietet sich natürlich die mehrfach erwähnte Erweiterung zur besseren Modellierung von ‚prior-data conflict'-Situationen an; potentiell könnte auch die dritte, bisher noch unberücksichtigte Dimension von komplexer Unsicherheit, die Vagheit, einbezogen werden, indem eventuelle Unbestimmtheit bezüglich der beobachtbaren Mengen durch eine Art ‚unscharfe' Likelihood in den Aufdatierungsprozess einginge.

Anhang

A.1 Das Normal-InversGamma-Modell für die lineare Regression

Das Normal-InversGamma-Modell (kurz: NIG-Modell) stellt ein konjugiertes bayesianisches Modell für die lineare Regression dar, wenn σ^2, die Varianz des Fehlerterms ε, unbekannt ist [O'Hagan 1994, S. 244ff]. Wie schon in Kapitel 4.2.3 erwähnt, ist eine direkte Verallgemeinerung analog zum dortigen Vorgehen nicht möglich, so dass das NIG-Modell auf diese Weise nicht als ein Aufdatierungsmodell im Sinne von Quaeghebeur und de Cooman verstanden werden kann. Warum das so ist, soll in diesem Teil des Anhangs erläutert werden. Zunächst muss dazu das NIG-Modell beschrieben werden; die Notation in Kapitel 4.2 wurde so gehalten, dass erkennbar ist, dass das Normal-Modell tatsächlich einen Teil des NIG-Modells darstellt. Die Unterschiede betreffen daher nur die Beifügungen, die durch die zusätzliche Priori-Verteilung über σ^2 gemacht werden müssen. Die Priori-Verteilung aus Abschnitt 4.2.1 wird multiplikativ um wenige Terme ergänzt; die Aufdatierungsschritte (4.2) und (4.3) des Normal-Modells bleiben unverändert, es müssen nur zwei weitere Parameter aufdatiert werden, die durch die Hinzufügung der InversGamma-Verteilung notwendig werden (siehe Gleichungen (A.1) – (A.4)).

A.1.1 Das Modell

Annahmen des Regressionsmodells

Die Annahmen des Regressionsmodells unterscheiden sich nur dadurch von den in Kapitel 4.2.1 getroffenen Annahmen, dass σ^2 nun nicht als bekannt vorausgesetzt wird.

- Regressionsgleichung:

$$z = \mathbf{X}\beta + \varepsilon\,, \quad \mathbf{X} \in \mathbb{R}^{k\times p}\,,\ \beta \in \mathbb{R}^p\,,\ z \in \mathbb{R}^k\,,\ \varepsilon \in \mathbb{R}^k$$

- Charakterisierung des Fehlerterms:

$$\varepsilon_i \overset{i.i.d}{\sim} N(\mathbf{0}, \sigma^2) \quad \Longrightarrow \quad \varepsilon \sim \mathrm{N}_k(0, \sigma^2\mathbf{I}))$$

- Die resultierende Likelihood ist multivariat normalverteilt:

$$z \,|\, X, \beta, \sigma^2 \sim \mathrm{N}_k(\mathbf{X}\beta, \sigma^2\mathbf{I})$$

Priori-Verteilung

Die Priori-Verteilung soll nun eine gemeinsame Verteilung über β und σ^2 sein; diese lässt sich nach dem Satz von Bayes aufspalten in eine Dichte über σ^2 und in eine Dichte über β gegeben σ^2:

$$p(\beta, \sigma^2) = p(\beta \,|\, \sigma^2) \cdot p(\sigma^2)\,,$$

Um zu einem konjugierten Modell zu gelangen, muss die Verteilung von $\beta \,|\, \sigma^2$ gemäß O'Hagan einer Normalverteilung entsprechen, bei der es sich um die gleiche Verteilung wie in Kapitel 4.2.1 handelt. σ^2 muss hingegen gemäß einer inversen Gamma-Verteilung verteilt sein. Wieder soll der obere Index $^{(0)}$ die Parameter der Priori-Verteilung bezeichnen.

$$\begin{aligned} \beta &\sim \mathrm{N}_p(\beta^{(0)}, \sigma^2 \mathbf{\Sigma}^{(0)}) \quad \text{mit } \beta^{(0)} \in \mathbb{R}^p\,,\ \mathbf{\Sigma}^{(0)} \in \mathbb{R}^{p\times p} \text{ positiv definit}\,, \\ \sigma^2 &\sim \mathrm{IG}(a^{(0)}, b^{(0)}) \qquad \text{mit } a^{(0)} > 0\,,\ b^{(0)} > 0\,, \end{aligned}$$

d.h.

$$\begin{aligned} p(\beta) &= \frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}} (\sigma^2)^{\frac{p}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\beta - \beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta - \beta^{(0)}\right) \right\}, \\ p(\sigma^2) &= \frac{(b^{(0)})^{(a^{(0)})}}{\Gamma(a^{(0)})} \frac{1}{(\sigma^2)^{a^{(0)}+1}} \exp\left\{ -\frac{b^{(0)}}{\sigma^2} \right\}. \end{aligned}$$

Die gemeinsame Verteilung von β und σ^2 ist unter dem Namen Normal-InversGamma-Verteilung (oder kurz: NIG-Verteilung)bekannt, mit der Notation

$$(\beta, \sigma^2) \sim \mathrm{NIG}(\beta^{(0)}, \mathbf{\Sigma}^{(0)}, a^{(0)}, b^{(0)})\,.$$

Die Dichte einer solchen Verteilung hat gemäß der obigen Herleitung folgende Form:

$$p(\beta, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{p}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\beta - \beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta - \beta^{(0)}\right) \right\} \frac{1}{(\sigma^2)^{a^{(0)}+1}} \exp\left\{ -\frac{b^{(0)}}{\sigma^2} \right\}$$

Posteriori-Verteilung

Die Posteriori-Verteilung kann nun gemäß des Satzes von Bayes ermittelt werden:

$$\begin{aligned} p(\beta, \sigma^2 \,|\, \mathbf{X}, z) &= \frac{p(\beta, \sigma^2, \mathbf{X}, z)}{p(\mathbf{X}, z)} \\ &= \frac{p(\mathbf{X}, z \,|\, \beta, \sigma^2)\, p(\beta, \sigma^2)}{p(\mathbf{X}, z)} \\ &= \frac{p(z \,|\, \mathbf{X}, \beta, \sigma^2)\, p(\mathbf{X} \,|\, \beta, \sigma^2)\, p(\beta, \sigma^2)}{p(\mathbf{X}, z)} \\ &= \frac{p(\mathbf{X})}{p(\mathbf{X}, z)} \cdot p(z \,|\, \mathbf{X}, \beta, \sigma^2)\, p(\beta, \sigma^2) \\ &\propto p(z \,|\, \mathbf{X}, \beta, \sigma^2)\, p(\beta, \sigma^2) \end{aligned}$$

Im Folgenden soll die Konjugiertheitseigenschaft gezeigt und der Aufdatierungsschritt für die Parameter $\beta^{(0)}$, $\mathbf{\Sigma}^{(0)}$, $a^{(0)}$ und $b^{(0)}$ ermittelt werden. Mit der oben genannten Likelihood und der NIG-Verteilung als Priori-Verteilung gilt also

$$\begin{aligned}
p(\beta, \sigma^2 \mid \mathbf{X}, z) \quad &\propto \quad \frac{1}{(\sigma^2)^{\frac{k}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(z-\mathbf{X}\beta)^{\mathsf{T}}(z-\mathbf{X}\beta)\right\} \\
&\quad \cdot \frac{1}{(\sigma^2)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(\beta-\beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta-\beta^{(0)}\right)\right\} \\
&\quad \cdot \frac{1}{(\sigma^2)^{a^{(0)}+1}} \exp\left\{-\frac{b^{(0)}}{\sigma^2}\right\} \\
&= \quad \frac{1}{(\sigma^2)^{\frac{p}{2}}} \cdot \frac{1}{(\sigma^2)^{a^{(0)}+\frac{k}{2}+1}} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma^2}\Big[(z-\mathbf{X}\beta)^{\mathsf{T}}(z-\mathbf{X}\beta)\right. \\
&\qquad\qquad \left.+\left(\beta-\beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta-\beta^{(0)}\right) + 2b^{(0)}\Big]\right\}.
\end{aligned}$$

Betrachtet man nun den Term in den eckigen Klammern im Exponenten isoliert, so ergibt sich:

$$\begin{aligned}
\Big[\cdots\Big] \quad &= \quad (z-\mathbf{X}\beta)^{\mathsf{T}}(z-\mathbf{X}\beta) + \left(\beta-\beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta-\beta^{(0)}\right) + 2b^{(0)} \\
&= \quad z^{\mathsf{T}}z - z^{\mathsf{T}}(\mathbf{X}\beta) - (\mathbf{X}\beta)^{\mathsf{T}}z + (\mathbf{X}\beta)^{\mathsf{T}}(\mathbf{X}\beta) \\
&\quad + \beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta - \beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)} - \beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta + \beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)} \\
&\quad + 2b^{(0)} \\
&= \quad z^{\mathsf{T}}z + \beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)} + 2b^{(0)} \\
&\quad - \underbrace{\beta^{\mathsf{T}}}_{A^{\mathsf{T}}} \underbrace{\left(\mathbf{X}^{\mathsf{T}}z + \mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)}_{CB} \\
&\quad - \underbrace{\left(z^{\mathsf{T}}\mathbf{X} + \beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\right)^{\mathsf{T}}}_{B^{\mathsf{T}}C} \underbrace{\beta}_{A} \\
&\quad + \underbrace{\beta^{\mathsf{T}}}_{A^{\mathsf{T}}} \underbrace{\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{\Sigma}^{(0)^{-1}}\right)}_{C} \underbrace{\beta}_{A}
\end{aligned}$$

Quadratische Ergänzung mit dem fehlenden Term $B^{\mathsf{T}}CB$, also Ergänzung mit

$$\begin{aligned}
&-\left(z^{\mathsf{T}}\mathbf{X} + \beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\right)\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}\right) \\
&+\left(z^{\mathsf{T}}\mathbf{X} + \beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\right)\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)
\end{aligned}$$

liefert

$$\begin{aligned}
\Big[\cdots\Big] &= z^{\mathsf{T}}z + \beta^{(0)\mathsf{T}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)} + 2b^{(0)} \\
&\quad - \left(z^{\mathsf{T}}\mathbf{X} + \beta^{(0)\mathsf{T}}\boldsymbol{\Sigma}^{(0)^{-1}}\right)\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right) \\
&\quad + \Bigg[\beta - \underbrace{\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)}_{=:\beta^{(1)}}\Bigg]^{\mathsf{T}} \\
&\quad \cdot \underbrace{\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)}_{=:\boldsymbol{\Sigma}^{(1)^{-1}}} \\
&\quad \cdot \Bigg[\beta - \underbrace{\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)}_{=:\beta^{(1)}}\Bigg] \\
&= z^{\mathsf{T}}z + \beta^{(0)\mathsf{T}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)} + 2b^{(0)} \; - \; \beta^{(1)\mathsf{T}}\boldsymbol{\Sigma}^{(1)^{-1}}\beta^{(1)} \\
&\quad + \left(\beta - \beta^{(1)}\right)^{\mathsf{T}}\boldsymbol{\Sigma}^{(1)^{-1}}\left(\beta - \beta^{(1)}\right).
\end{aligned}$$

Also ist

$$\begin{aligned}
p(\beta, \sigma^2 \,|\, \mathbf{X}, z) \;\propto\; & \frac{1}{(\sigma^2)^{\frac{p}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\left(\beta - \beta^{(1)}\right)^{\mathsf{T}}\boldsymbol{\Sigma}^{(1)^{-1}}\left(\beta - \beta^{(1)}\right)\right\} \\
& \cdot \frac{1}{(\sigma^2)^{a^{(0)}+\frac{k}{2}+1}}\exp\Bigg\{-\frac{1}{\sigma^2}\Bigg[b^{(0)} + \frac{1}{2}\Big(z^{\mathsf{T}}z + \beta^{(0)\mathsf{T}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)} \\
& \qquad\qquad - \beta^{(1)\mathsf{T}}\boldsymbol{\Sigma}^{(1)^{-1}}\beta^{(1)}\Big)\Bigg]\Bigg\}.
\end{aligned}$$

Dieser Term entspricht bis auf Normierungskonstanten der Dichte einer NIG-Verteilung, somit gilt

$$p(\beta, \sigma^2 \,|\, \mathbf{X}, z) \sim \mathrm{NIG}(\beta^{(1)}, \boldsymbol{\Sigma}^{(1)}, a^{(1)}, b^{(1)}),$$

mit den aufdatierten Parametern

$$\begin{aligned}
\beta^{(1)} &= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}z + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right) && \text{(A.1)} \\
&= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\hat{\beta} + \boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)}\right) \\
&= (\mathbf{I} - \mathbf{A})\beta^{(0)} + \mathbf{A}\hat{\beta} \qquad \text{mit } \mathbf{A} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X} \\
\boldsymbol{\Sigma}^{(1)} &= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}^{(0)^{-1}}\right)^{-1} && \text{(A.2)} \\
a^{(1)} &= a^{(0)} + \frac{k}{2} && \text{(A.3)} \\
b^{(1)} &= b^{(0)} + \frac{1}{2}\left(z^{\mathsf{T}}z + \beta^{(0)\mathsf{T}}\boldsymbol{\Sigma}^{(0)^{-1}}\beta^{(0)} - \beta^{(1)\mathsf{T}}\boldsymbol{\Sigma}^{(1)^{-1}}\beta^{(1)}\right). && \text{(A.4)}
\end{aligned}$$

Mit der Wahl einer NIG-Verteilung als gemeinsame Priori-Verteilung für β und σ^2 lassen sich also über die Aufdatierungsschritte (A.1) – (A.4) die Parameter der Posteriori-Verteilung von β und σ^2 einfach berechnen. Offensichtlich unterscheiden sich dabei die Aufdatierungsschritte (A.1) und (A.2) nicht von ihren Pendants für das Normal-Modell (siehe Gleichungen (4.2) und (4.3)).

A.1.2 Die Normal-InversGamma-Verteilung als Exponentialfamilie

Wie in Kapitel 4.2.2 soll nun in einem ersten Schritt gezeigt werden, dass eine NIG-Verteilung einer Exponentialfamilie gemäß der Notation von Quaeghebeur und de Cooman entspricht, um die Form von ψ und $\mathbf{b}(\psi)$ herauszufinden.

Sei also (β, σ^2) NIG-verteilt mit den Parametern $\beta^{(0)} \in \mathbb{R}^p$, $\mathbf{\Sigma}^{(0)} \in \mathbb{R}^{p \times p}$ symmetrisch positiv definit, $a^{(0)} > 0$ und $b^{(0)} > 0$. Die Anzahl der Parameter bei voller Parametrisierung von $\mathbf{\Sigma}^{(0)}$ beträgt $p + \frac{p(p+1)}{2} + 2 = \frac{p(p+3)}{2} + 2 =: q$. Im Vergleich zum Normal-Modell kommen also einfach die beiden Parameter $a^{(0)}$ und $b^{(0)}$ für die InversGamma-Verteilung hinzu. Gilt also

$$(\beta, \sigma^2) \sim \mathrm{NIG}(\beta^{(0)}, \mathbf{\Sigma}^{(0)}, a^{(0)}, b^{(0)}) \,,$$

dann hat ihre (gemeinsame) Dichte folgende Form:

$$\begin{aligned} p(\beta, \sigma^2) \;\; = \;\; & \frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}} (\sigma^2)^{\frac{p}{2}}} \cdot \frac{(b^{(0)})^{(a^{(0)})}}{\Gamma(a^{(0)})} \cdot \frac{1}{(\sigma^2)^{a^{(0)}+1}} \\ & \cdot \exp\left\{ -\frac{1}{2\sigma^2} \Big[\left(\beta - \beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta - \beta^{(0)}\right) + 2b^{(0)} \Big] \right\} \end{aligned}$$

Die Form einer Exponentialfamilie für die Priori-Verteilung in der Notation gemäß Quaeghebeur und de Cooman ist dann, schon auf die vorzunehmende Umformung hin angepasst:

$$p(\psi) = \mathbf{a}(\beta^{(0)}, \mathbf{\Sigma}^{(0)}, a^{(0)}, b^{(0)}) \cdot \exp\left\{ \left\langle \psi \,,\, \tau(\beta^{(0)}, \mathbf{\Sigma}^{(0)}, a^{(0)}, b^{(0)}) \right\rangle - \mathbf{b}(\psi) \right\}$$

Es handelt sich um eine Dichte über ψ, welches sich als Funktion von β und σ^2 ergibt. Wieder spielen die Priori-Parameter die Rolle der ‚Beobachtung'.

Umformung:

$$
\begin{aligned}
p(\beta,\sigma^2) &= \frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}}\cdot\frac{(b^{(0)})^{(a^{(0)})}}{\Gamma(a^{(0)})}\cdot \\
&\quad \exp\Bigg\{-(a^{(0)}+1)\log(\sigma^2)-\frac{p}{2}\log(\sigma^2) \\
&\qquad -\frac{1}{2\sigma^2}\Big[\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta-\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}-\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta \\
&\qquad\qquad +\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}+2b^{(0)}\Big]\Bigg\} \\
&= \underbrace{\frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}}\cdot\frac{(b^{(0)})^{(a^{(0)})}}{\Gamma(a^{(0)})}}_{=:\,\mathbf{a}(\beta^{(0)},\mathbf{\Sigma}^{(0)},a^{(0)},b^{(0)})}\cdot \\
&\quad \exp\Bigg\{-\frac{1}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta+\frac{1}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}+\frac{1}{2\sigma^2}\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta \\
&\qquad -\frac{1}{2\sigma^2}\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}-\frac{b^{(0)}}{\sigma^2}-(a^{(0)}+1)\log(\sigma^2)-\underbrace{\frac{p}{2}\log(\sigma^2)}_{=:\,\mathbf{b}(\psi)}\Bigg\}
\end{aligned}
$$

Da $\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}$ und $\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta$ Skalare sind und $\left(\beta^{\mathsf{T}}\mathbf{\Sigma}^{(0)^{-1}}\beta^{(0)}\right)^{\mathsf{T}}=\beta^{(0)^{\mathsf{T}}}\mathbf{\Sigma}^{(0)^{-1}}\beta$ gilt, können diese beiden Terme zusammengefasst werden.

Zur Vereinfachung der Notation sei im Folgenden wieder $\mathbf{\Sigma}^{(0)^{-1}} =: \mathbf{\Lambda}^{(0)}$ mit den Elementen $\left(\mathbf{\Lambda}^{(0)}\right)_{ij} =: \lambda_{ij}$; außerdem seien die Laufindices folgendermaßen festgelegt:

h, i, j	Laufindex von $1,\ldots,p=\dim(\beta)$
l, m	Laufindex von $1,\ldots,k=\dim(z)$

Somit folgt

$$\begin{aligned}
p(\beta,\sigma^2) &= \mathbf{a}(\beta^{(0)},\mathbf{\Sigma}^{(0)},a^{(0)},b^{(0)})\cdot \\
&\qquad \exp\Big\{-\frac{1}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta+\frac{1}{\sigma^2}\beta^{\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta^{(0)}-\frac{1}{2\sigma^2}\beta^{(0)\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta^{(0)} \\
&\qquad\qquad -\frac{b^{(0)}}{\sigma^2}-(a^{(0)}+1)\log(\sigma^2)-\mathbf{b}(\psi)\Big\} \\
&= \mathbf{a}(\beta^{(0)},\mathbf{\Sigma}^{(0)},a^{(0)},b^{(0)})\cdot \\
&\qquad \exp\Big\{-\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\beta_i\beta_j}{2\sigma^2}\cdot\lambda_{ij}^{(0)}+\sum_{i=1}^{p}\frac{\beta_i}{\sigma^2}\cdot\Big(\sum_{j=1}^{p}\lambda_{ij}^{(0)}\beta_j^{(0)}\Big)-\frac{1}{2\sigma^2}\beta^{(0)\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta^{(0)} \\
&\qquad\qquad -\frac{b^{(0)}}{\sigma^2}-(a^{(0)}+1)\log(\sigma^2)-\mathbf{b}(\psi)\Big\} \\
&= \mathbf{a}(\beta^{(0)},\mathbf{\Sigma}^{(0)},a^{(0)},b^{(0)})\cdot \\
&\qquad \exp\Big\{-\sum_{i=1}^{p}\frac{\beta_i^2}{2\sigma^2}\cdot\lambda_{ii}^{(0)}-\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\frac{\beta_i\beta_j}{\sigma^2}\cdot\lambda_{ij}^{(0)}+\sum_{i=1}^{p}\frac{\beta_i}{\sigma^2}\cdot\Big(\sum_{j=1}^{p}\lambda_{ij}^{(0)}\beta_j^{(0)}\Big) \\
&\qquad\qquad -\frac{1}{2\sigma^2}\Big(\beta^{(0)\mathsf{T}}\mathbf{\Lambda}^{0}\beta^{(0)}+2b^{(0)}\Big)-\log(\sigma^2)\cdot(a^{(0)}+1)-\mathbf{b}(\psi)\Big\}
\end{aligned}$$

Also gilt:

$$\begin{aligned}
\mathbf{a}(\beta^{(0)},\mathbf{\Sigma}^{(0)},a^{(0)},b^{(0)}) &= \frac{1}{|\mathbf{\Sigma}^{(0)}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}}\cdot\frac{(b^{(0)})^{(a^{(0)})}}{\Gamma(a^{(0)})} \\
\mathbf{b}(\psi) &= \frac{p}{2}\log(\sigma^2)
\end{aligned}$$

$$
\psi \;=\; \begin{bmatrix} -\frac{\beta_1^2}{2\sigma^2} \\ \vdots \\ -\frac{\beta_p^2}{2\sigma^2} \\ -\frac{\beta_1\beta_2}{\sigma^2} \\ \vdots \\ -\frac{\beta_{p-1}\beta_p}{\sigma^2} \\ -\frac{\beta_1}{\sigma^2} \\ \vdots \\ -\frac{\beta_p}{\sigma^2} \\ -\frac{1}{2\sigma^2} \\ -\log(\sigma^2) \end{bmatrix}, \quad \tau(\,\cdot\,) \;=\; \begin{bmatrix} \lambda_{11}^{(0)} \\ \vdots \\ \lambda_{pp}^{(0)} \\ \lambda_{12}^{(0)} \\ \vdots \\ \lambda_{p-1,p}^{(0)} \\ \sum_{j=1}^{p} \lambda_{1j}^{(0)} \beta_j^{(0)} \\ \vdots \\ \sum_{j=1}^{p} \lambda_{pj}^{(0)} \beta_j^{(0)} \\ \left(\beta^{(0)\mathsf{T}} \mathbf{\Lambda}^{(0)} \beta^{(0)} + 2b^{(0)}\right) \\ (a^{(0)} + 1) \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} p \text{ Summanden} \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \\ a \\ a \end{matrix}}\right\} \frac{p(p-1)}{2} \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}}\right\} p \\ \left.\vphantom{\begin{matrix} a \\ a \end{matrix}}\right\} 2 \end{matrix}
$$

Hier ergibt sich nun, in Abweichung zu dem Ergebnis in Kapitel 4.2.2, ein $\mathbf{b}(\psi) \neq 0$, was die Probleme bei dem Versuch, das NIG-Modell als Aufdatierungsmodell gemäß Quaeghebeur und de Cooman zu verstehen, verursacht. Darauf wird im nächsten Kapitel im Detail eingegangen.

Die Form von ψ und $\tau(\beta^{(0)}, \mathbf{\Sigma}^{(0)}, a^{(0)}, b^{(0)})$ stimmt hingegen in den ersten $\frac{p(p+3)}{2}$ Zeilen vollkommen mit der sich im Normal-Modell ergebenden Form überein; es kommen nur die letzten beiden Zeilen, die ihren Ursprung in der InversGamma-Verteilung haben, hinzu.

Ebenso entspricht die Anzahl der natürlichen Parameter der Anzahl der ‚klassischen' Parameter; es gilt $p + \frac{p(p-1)}{2} + p + 2 = \frac{p(p+3)}{2} + 2 = q$, die NIG-Verteilung erweist sich also genau wie die Normalverteilung in Kapitel 4.2.2 als ‚strikt q-parametrig' (jedoch mit einem anderen Wert von q).

A.1.3 Versuch der Anwendung des Aufdatierungsmodells von Quaeghebeur und de Cooman auf das NIG-Modell

Es ist zwar möglich, die NIG-Verteilung in gleicher Weise wie in Kapitel 4.2.3 in die Form einer Priori-Verteilung gemäß [Quaeghebeur und de Cooman 2005] umzuformen. Die Überprüfung, ob diese Priori-Verteilung aber nach einer Aufdatierung gemäß der Formeln (A.1) – (A.4) einer Posteriori-Verteilung gemäß der Notation von Quaeghebeur und de Cooman entspricht, muss jedoch scheitern, wie im Folgenden erläutert wird.

Das ‚Ziel-Format' ist durch die Gleichung (3.3) gegeben, die der Anschaulichkeit halber hier noch einmal wiedergegeben werden soll:

$$p(\psi) \;=\; \mathbf{c}\left(n^{(0)}, y^{(0)}\right) \exp\left\{ n^{(0)} \left[\left\langle \psi \,,\, y^{(0)} \right\rangle - \mathbf{b}(\psi) \right] \right\}$$

Um die in Kapitel A.1.2 ermittelte Form der Normal-InversGamma-Verteilung nun an diese Form anzupassen, muss jeder Summand im Exponenten, also auch $\mathbf{b}(\psi)$, mit dem Faktor $\frac{1}{n^{(0)}}$ multipliziert werden, damit es möglich ist, $n^{(0)}$ im Exponenten auszuklammern. Bei allen Termen außer $\mathbf{b}(\psi)$ ist es möglich, diesen Faktor $\tau(\beta^0, \mathbf{\Sigma}^0, a^0, b^0)$ zuzuordnen, so dass die schon in Kapitel A.1.2 gefundene Form von ψ beibehalten werden kann. Für den letzten Term im Exponenten, der nur aus $\mathbf{b}(\psi)$ besteht, gibt es diese Möglichkeit nicht. Analog zur Notation des Aufdatierungsschritts in Kapitel 4.2.3 gilt hier

$$\underbrace{\frac{1}{n^{(0)}} \cdot \frac{p}{2} \log(\sigma^2)}_{\substack{\text{letzter Term im} \\ \text{Exponenten } \textit{vor} \\ \text{der Aufdatierung} \\ \text{im NIG-Modell}}} = \frac{1}{n^{(0)}} \cdot \mathbf{b}(\psi) \quad \xrightarrow{\text{Beobachtung}} \quad \frac{1}{n^{(1)}} \cdot \mathbf{b}(\psi) = \frac{1}{n^{(0)}+1} \cdot \frac{p}{2} \log(\sigma^2)$$

$$\neq \underbrace{\frac{1}{n^{(0)}} \cdot \frac{p}{2} \log(\sigma^2)}_{\substack{\text{letzter Term im} \\ \text{Exponenten } \textit{nach} \\ \text{der Aufdatierung} \\ \text{im NIG-Modell}}} .$$

Nach der Ausklammerung von $n^{(0)}$ beträgt der letzte Summand im Exponenten $\frac{1}{n^{(0)}} \cdot \mathbf{b}(\psi)$; nach dem Aufdatierungsschritt gemäß [Quaeghebeur und de Cooman 2005] müsste er $\frac{1}{n^{(0)}+1} \cdot \mathbf{b}(\psi)$ betragen. $\mathbf{b}(\psi)$ hat jedoch, anders als die anderen Summanden im Exponenten, keine Bestandteile, die diese Veränderung ‚kompensieren' könnten, und somit ergibt sich ein Widerspruch.

Erstaunlicherweise stellt aber der Aufdatierungsschritt für alle anderen Summanden, die sich aus den Elementen von ψ und $\tau(\beta^{(0)}, \mathbf{\Sigma}^{(0)}, a^{(0)}, b^{(0)})$ zusammensetzen, kein Problem dar. Bis auf die letzten beiden Summanden, die durch den Übergang zum NIG-Modell hinzugekommenen sind, unterscheiden diese sich ja nicht von den Summanden im Normal-Modell, für die in Kapitel 4.2.3 die Äquivalenz der Aufdatierungsschritte schon gezeigt wurde. Aber auch für die restlichen beiden Summanden kann diese Äquivalenz gezeigt werden. Diese Summanden, die durch die beiden jeweils letzten Elemente von ψ und $\tau(\beta^{(0)}, \mathbf{\Sigma}^{(0)}, a^{(0)}, b^{(0)})$ gebildet werden, können wie alle anderen die Ausklammerung von $n^{(0)}$ ‚kompensieren' und stünden der Interpretation des NIG-Modells als Aufdatierungsmodell gemäß Quaeghebeur und de Cooman nicht im Wege, da mit der oben vereinbarten Notation $\mathbf{\Sigma}^{(0)^{-1}} =: \mathbf{\Lambda}^{(0)}$ die folgenden Beziehungen gelten:

$$\underbrace{-\frac{1}{2\sigma^2}}_{\Psi_{\frac{p(p+3)}{2}+1}} \cdot \underbrace{\frac{1}{n^{(0)}}\left(\beta^{(0)\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta^{(0)} + 2b^{(0)}\right)}_{y^{(0)}_{\frac{p(p+3)}{2}+1}} \xrightarrow{\text{Beobachtung}} -\frac{1}{2\sigma^2} \cdot \frac{1}{n^{(1)}}\left(\beta^{(1)\mathsf{T}}\mathbf{\Lambda}^{(1)}\beta^{(1)} + 2b^{(1)}\right)$$

$$= -\frac{1}{2\sigma^2} \cdot \frac{1}{n^{(0)}+1}\Big(\beta^{(1)\mathsf{T}}\mathbf{\Lambda}^{(1)}\beta^{(1)} + 2b^{(0)} + z^{\mathsf{T}}z + \beta^{(0)\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta^{(0)} - \beta^{(1)\mathsf{T}}\mathbf{\Lambda}^{(1)}\beta^{(1)}\Big)$$

$$= \underbrace{-\frac{1}{2\sigma^2}}_{\Psi_{\frac{p(p+3)}{2}+1}} \cdot \frac{1}{n^{(0)}+1}\left(n^{(0)} \cdot \underbrace{\frac{1}{n^{(0)}}\left(\beta^{(0)\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta^{(0)} + 2b^{(0)}\right)}_{y^{(0)}_{\frac{p(p+3)}{2}+1}} + \underbrace{z^{\mathsf{T}}z}_{\tau_{\frac{p(p+3)}{2}+1}}\right)$$

$$\underbrace{-\log(\sigma^2)}_{\Psi_{\frac{p(p+3)}{2}+2}} \cdot \underbrace{\frac{1}{n^{(0)}}(a^{(0)}+1)}_{y^{(0)}_{\frac{p(p+3)}{2}+2}} \xrightarrow{\text{Beobachtung}} -\log(\sigma^2) \cdot \frac{1}{n^{(1)}}(a^{(1)}+1)$$

$$= \underbrace{-\log(\sigma^2)}_{\Psi_{\frac{p(p+3)}{2}+2}} \cdot \frac{1}{n^{(0)}+1}\left(n^{(0)} \cdot \underbrace{\frac{1}{n^{(0)}}(a^{(0)}+1)}_{y^{(0)}_{\frac{p(p+3)}{2}+2}} + \underbrace{\frac{k}{2}}_{\tau_{\frac{p(p+3)}{2}+2}}\right)$$

A.2 Syntax für die Berechnung der Ergebnisse in Kapitel 4.4 und 4.5

Die in den Kapiteln 4.4 und 4.5 dargestellten Inferenzergebnisse und Schaubilder wurden mit der Statistik-Software `R`, Versionsnummer 2.3.1, berechnet. Insbesondere die Funktionen, die zur Berechnung der Kredibilitätsregionen dienen, wurden im Rahmen der Implementierung jedoch nicht laufzeitoptimiert, da der Schwerpunkt dieser Arbeit auf der theoretischen Entwicklung des Normal-Modells unter komplexer Unsicherheit lag.

Die Programmierungssyntax kann auf der beigefügten CD gefunden werden. Der Ordner `R` enthält die Unterordner `Syntax` und `Libraries`; während in ersterem drei Syntax-Dateien mit den Befehlen zur Berechnung der Ergebnisse in 4.4.1 abgelegt sind, die stellvertretend für die Syntax der anderen Beispiele stehen sollen, enthält letzterer (auch online erhältliche) Packages für `R`, die für die Lauffähigkeit der Syntax nötig sind.

`kredibilfunktionen.r` enthält selbstgeschriebene Funktionen zur Berechnung der in den Kapiteln 4.4 und 4.5 angegebenen zweidimensionalen Kredibilitätsregionen sowie weitere Hilfsfunktionen.

`bsp1-teil1.r` enthält den Hauptteil der Syntax für das Beispiel in Abschnitt 4.4.1, nämlich die Syntax für die Berechnung der verschiedenen Posteriori-Inferenzergebnisse je nach Annahme der Werte für A, B_1 und B_2 sowie für die Erstellung aller dort eingefügten Abbildungen bis auf zwei.

`bsp1-teil2.r` enthält hingegen nur die Syntax für die Erstellung der Abbildungen 4.5 und 4.6, die die Asymptotik der Posteriori-Intervalle für $k \to \infty$ illustrieren.

Literaturverzeichnis

[Augustin 1998] Augustin, T. (1998). *Optimale Tests bei Intervallwahrscheinlichkeit*, Vandenhoeck und Ruprecht, Göttingen.

[Augustin 2005] Augustin, T. (2005). Generalized basic probability assignments, *International Journal of General Systems* **34**: 451–463.

[Berger (1985)] Berger, J.O. (1985). Robust Bayesian Analysis: Sensitivity to the prior, *Journal of Statistical Planning and Inference* **25**: 303–328.

[Bernard 2007] Bernard, J.-M. (2007). Special Issue on the Imprecise Dirichlet Model, *International Journal of Approximate Reasoning*, in Vorbereitung.

[Bernard 2005] Bernard, J.-M. (2005). An introduction to the imprecise Dirichlet Model for multinomial Data, *International Journal of Approximate Reasoning* **39**: 123–150.

[Bernard 2001] Bernard, J.-M. (2001). Non-parametric inference about an unknown mean using the imprecise Dirichlet model, in: de Cooman, G., Fine, T., Seidenfeld, T. (Hrsg.), *Proceedings of the 2nd International Symposium on Imprecise Probabilities and their Applications (ISIPTA'01)*, Shaker, Ithaca, New York.

[Bernard 2003] Bernard, J.-M. (2003). Analysis of local or asymetric dependencies in contigency tables using the imprecise Dirchlet model, in: Bernard, J.-M., Seidenfeld, T., Zaffalon, M. (Hrsg.), *Proceedings of the 3rd International Symposium on Imprecise Probabilities and their Applications (ISIPTA'03)*, Carleton Scientific, Waterloo, Ontario.

[Bernardo und Smith 1993] Bernardo, J. und Smith, A. (1993). *Bayesian Theory*, Wiley & Sons, Chichester, New York.

[Boratyńska 1997] Boratyńska, A. (1997). Stability of Bayesian inference in exponential families, *Statistics & Probability Letters* **36**: 173–178.

[Box und Tiao 1973] Box, G.E.P. und Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA.

[Buja 1986] Buja, A. (1986). On the Huber-Strassen theorem, *Probability Theory and Related Fields*, **73**: 149–152.

[Coolen 1993] Coolen, F.P.A. (1993). Imprecise conjugate prior densities for the one-parameter exponential family of distributions, *Statistics & Probability Letters* **16**: 337–342.

[de Finetti 1990] de Finetti, B. (1990). *Theory of Probability*, Wiley Classics, Chichester, New York. (zweibändig, Übersetzung aus dem Italienischen.)

[Dempster 1967] Dempster, A.P. (1967). Upper and lower probability inferences induced by a multivalued mapping, *Annals of Mathematical Statistics* **38**: 325–339.

[Ellsberg 1961] Ellsberg, D. (1961). Risk, Ambiguity and the Savage axioms, *Quarterly Journal of Economics* **75**: 643–669.

[Holmes und Held 2006] Holmes, C. C. und Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression, *Bayesian Analysis* **1**, Vol. 1: 145–168.

[Huber 1965] Huber, P.J. (1965). A robust version of the probability ratio test, *Annals of Mathematical Statistics* **36**: 1753–1758.

[Huber und Strassen 1973] Huber, P.J. und Strassen, V. (1973). Minimax Tests and the Neyman-Pearson lemma for capacities, *The Annals of Statistics* **1**: 251–263.

[Löwel et al. 2005] Löwel, H., Meisinger, C., Heier, M., und Hormann, A. (2005). The population-based acute myocardial infarction (AMI) registry of the MONICA/KORA study region of Augsburg, *Gesundheitswesen* **67**, Supplement 1: 31–S37.

[Maier 2004] Maier, S. (2004). Entscheidung und Analyse bei unscharfer Information, *Diplomarbeit, Institut für Statistik, Ludwig-Maximilians-Universität München.*

[O'Hagan 1994] O'Hagan, A. (1994) *Bayesian Inference*, Kendall´s Advanced Theory of Statistics Vol. 2B, Arnold, London.

[Ott 2001] Ott, N. (2001). *Unsicherheit, Unschärfe und rationales Entscheiden. Die Anwendung von Fuzzy-Methoden in der Entscheidungstheorie*, Wirtschaftswissenschaftliche Beiträge Vol. 179, Physica-Verlag, Heidelberg.

[Pericchi und Walley 1991] Pericchi, L.P. und Walley, P. (1991). Robust Bayesian Credible Intervals and Prior Ignorance, *International Statistical Review* **58**: 1–23.

[Peters et al., eingereicht] Peters, A., Schneider, A., Greven, S., Bellander, T., Forastiere, F., Ibald-Mulli, A., Illig, T., Jacquemin, B., Katsouyanni, K., Koenig, W., Lanki, T., Pekkanen, J., Pershagen, G., Picciotto, S., Rückerl, R., Schaffrath-Rosario, A., Stefanadis, C. und Sunyer, J. (eingereicht). Air pollution and inflammatory response in myocardial infarction survivors: gene-environment-interactions in a high-risk group: study design of the airgene study, eingereicht bei: *Inhalation Toxicology.*

[Quaeghebeur und de Cooman 2005] Quaeghebeur, E. und de Cooman, G. (2005). Imprecise probability models for inference in exponential families, in: Cozman, F.G., Nau, R., Seidenfeld, T. (Hrsg.) *Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications (ISIPTA'05)*, Pittsburgh (Carnegie Mellon University), SIPTA, Manno (CH).

[Reithinger 2006] Reithinger, F. (2006). Zusammenhangsstrukturen, *Technischer Report, Institut für Statistik, Ludwig-Maximilians-Universität München.*

[Rieder 1994] Rieder, H. (1994). *Robust Asymptotic Statistics*, Springer, New York, Berlin, Heidelberg.

[Rüger 1999] Rüger, B. (1999). *Test- und Schätztheorie. Band I: Grundlagen*, Oldenbourg, München, Wien.

[Schill 1990] Schill, K. (1990). *Medizinische Expertensysteme. Methoden und Techniken*, Oldenbourg, München, Wien.

[Seidenfeld und Wasserman 1993] Seidenfeld, T. und Wasserman, L. (1993). Dilation for sets of probabilities, *Annals of Statistics* **21**: 1139–1154.

[Shafer 1976] Shafer, G. (1976). *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, New Jersey.

[Strobl 2005] Strobl, C. (2005). Das Imprecise Dirichlet Modell, *Seminarvortrag, Institut für Statistik, Ludwig-Maximilians-Universität München.*

[Thorand et al. 2006] Thorand, B., Baumert, J., Döring, A., Herder, C., Kolb, H., Rathmann, W., Giani, G., Koenig, W.; KORA Gruppe; GSF – Forschungszentrum für Umwelt und Gesundheit, Institut für Epidemiologie, Neuherberg (2006). Sex differences in the relation of body composition to markers of inflammation, *Atherosclerosis* **184**, No. 1: 216–224.

[Tibshirani 1996] Tibshirani, R. (2003). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B* **58**, No. 1: 267–288.

[Toutenburg (2003)] Toutenburg, H. (2003). *Lineare Modelle*, 2. Auflage, Physica-Verlag, Heidelberg.

[Utkin und Augustin 2005] Utkin. L.V., Augustin, T. (2005). Decision making under imperfect measurement using the imprecise Dirichlet model, in: Cozman, F.G., Nau, R., Seidenfeld, T. (Hrsg.), *Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications (ISIPTA'05)*, Pittsburgh (Carnegie Mellon University), SIPTA, Manno (CH).

[Walley 1991] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London, New York.

[Walley 1996] Walley, P. (1996). Inferences from Multinomial Data: Learning about a Bag of Marbles, *Journal of the Royal Statistical Society B* **58**, No. 1: 3–57.

[Wallner 2003] Wallner, T. (2003) Bi-elastic Neighbourhood Models, in: Bernard, J.-M., Seidenfeld, T., Zaffalon, M. (Hrsg.), *Proceedings of the 3rd International Symposium on Imprecise Probabilities and their Applications (ISIPTA'03)*, Carleton Scientific, Waterloo, Ontario.

[Weichselberger 2001] Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeinen Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*, Physica-Verlag, Heidelberg.

[Whittaker 1990] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, Wiley & Sons, Chichester, New York.

Abbildungsverzeichnis

2.1 Vergröberung von Kategorien als Baumstruktur 26
2.2 Darstellung der zulässigen Parameterkombinationen für $k = 2$ und $k = 3$. 29
2.3 Darstellung der noch zulässigen Parameterkombinationen für $\mathcal{M}^{(0)}$ aus Gleichung (2.8). 30
2.4 Konstruktion eines einseitigen Kredibilitätsintervalls für θ_j, mit $s = 1$, $n = 10$ und $n_j = 4$ 39

4.1 Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A datengestützt 88
4.2 Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$ 89
4.3 Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A datengeleitet 91
4.4 Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$ 92
4.5 Schätzintervalle für β bei $k \to \infty$, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$ 94
4.6 Schätzintervalle für β bei $k \to \infty$, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, weite Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 100$ 95
4.7 Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, schmälere Intervalle für $\beta^{(0)}$, Wahl von A datengestützt 96
4.8 Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, schmälere Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$ . . 97
4.9 Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, schmälere Intervalle für $\beta^{(0)}$, Wahl von A nach Vorgabe von $n^{(0)} = 20$ 98
4.10 Simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren . . 99
4.11 Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und großen Werten von β, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren 100
4.12 Simulierter Datensatz mit 20 Beobachtungen und kleinen Werten von β, Wahl von A nach Vorgabe von $n^{(0)} = 20$ 101

4.13 Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und kleinen Werten von β, Wahl von A nach Vorgabe von $n^{(0)} = 20$ . . . 102
4.14 Simulierter Datensatz mit 20 Beobachtungen und kleinen Werten von β, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren . . 103
4.15 Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und kleinen Werten von β, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren 104
4.16 Simulierter Datensatz mit 20 Beobachtungen und hoher Kollinearität der Regressoren, Wahl von A nach Vorgabe von $n^{(0)} = 20$ 106
4.17 Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und hoher Kollinearität der Regressoren, Wahl von A nach Vorgabe von $n^{(0)} = 20$ 107
4.18 Simulierter Datensatz mit 20 Beobachtungen und hoher Kollinearität der Regressoren, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren 108
4.19 Kredibilitätsregion für β, simulierter Datensatz mit 20 Beobachtungen und hoher Kollinearität der Regressoren, Wahl von A, B_1 und B_2, um sehr schwaches Vorwissen zu modellieren 109
4.20 AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A ‚datengeleitet', Koeffizienten zu standardisierten Regressoren 113
4.21 AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A ‚datengeleitet', Koeffizienten zu ‚normalen' Regressoren des Confounder-Modells . . 114
4.22 Kredibilitätsregion für $\tilde{\beta}_{\texttt{age}}$ und $\tilde{\beta}_{\texttt{bmi}}$, AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A ‚datengeleitet', Koeffizienten zu standardisierten Regressoren 115
4.23 Kredibilitätsregion für $\beta_{\texttt{age}}$ und $\beta_{\texttt{bmi}}$, AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A ‚datengeleitet', Koeffizienten zu ‚normalen' Regressoren des Confounder-Modells 116
4.24 AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A, um sehr schwaches Vorwissen zu modellieren, Koeffizienten zu standardisierten Regressoren 117
4.25 AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A, um sehr schwaches Vorwissen zu modellieren, Koeffizienten zu ‚normalen' Regressoren des Confounder-Modells 118
4.26 Kredibilitätsregion für $\tilde{\beta}_{\texttt{age}}$ und $\tilde{\beta}_{\texttt{bmi}}$, AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A, um sehr schwaches Vorwissen zu modellieren, Koeffizienten zu standardisierten Regressoren 119
4.27 Kredibilitätsregion für $\beta_{\texttt{age}}$ und $\beta_{\texttt{bmi}}$, AIRGENE-Datensatz mit 199 Beobachtungen, Wahl von A, um sehr schwaches Vorwissen zu modellieren, Koeffizienten zu ‚normalen' Regressoren des Confounder-Modells 120

Erklärung zur Urheberschaft

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

München, den 27.10.2006

(Gero Walter)