Die approbierte Originalversion dieser Dissertation ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (http://www.ub.tuwien.ac.at).

The approved original version of this thesis is available at the main library of the Vienna University of Technology (http://www.ub.tuwien.ac.at/englweb/).

DISSERTATION

Bayesianischer Vergleich der Qualität endlicher Mischungsmodelle mit bzw. ohne Zufallseffekte anhand künstlicher Daten

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

o. Univ. Prof. Dipl.-Ing. Dr. techn. Manfred Deistler

Inst.Nr.: E105
Institut für Wirtschaftsmathematik

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Dipl.-Ing. Ulrike Schuster

Matr.Nr.: E9725152
Wiesenerstraße 3/1/2
A - 2104 Spillern

WIEN, 15.12.2004
Ort, Datum

Ulrike Schuster
Unterschrift

Kurzfassung

In der Ökonometrie wird man sehr oft mit Problemen konfrontiert, die man nur schwer bzw. manchmal auch gar nicht mit herkömmlichen Schätzmethoden lösen kann. Der bayesianische Ansatz stellt in solchen Fällen eine überaus effiziente Alternative dar. So eignet sich die bayesianische Methode des Markov Chain Monte Carlo Samplings auch zur Schätzung der Parameter des endlichen Mischungsmodells.

In der Arbeit wird erstmals explizit die Performance dreier Modellansätze, die Heterogenität in Konsumentendaten berücksichtigen, unter Verwendung künstlich generierter Daten verglichen. Eine weitere Neuheit stellt der Datengenerierungsprozess dar. Die Daten werden nämlich über ein Simulationsmodell, in dem keine restriktiven Annahmen beispielsweise bezüglich der Verteilung der Parameter getroffen werden müssen, erzeugt. Lediglich die Marktstruktur hinsichtlich der Segmentierung der Konsumenten wird festgelegt.

Konkret handelt es sich bei den Ansätzen einerseits um ein latentes Klassenmodell, mit dem man segmentspezifische Parameter definieren kann. Andererseits wird ein Zufallseffektmodell betrachtet, in dem die Parameter einer kontinuierlichen Verteilung folgend über die Bevölkerung variieren. Das dritte Modell ist schließlich eine Kombination der beiden vorangegangenen Ansätze, das bedeutet die segmentspezifischen Koeffizienten werden als Zufallseffekte spezifiziert und dürfen auch innerhalb der Konsumentengruppen variieren.

Die Modelle werden mithilfe der bayesianischen Methode des Markov Chain Monte Carlo Samplings geschätzt. Insbesondere wird beschränktes Permutation Sampling eingesetzt.
Aus diesem Grund erfolgt im ersten Teil der Arbeit eine Einführung in die Grundlagen der bayesianischen Schätzung und die Präsentation diverser Markov Chain Monte Carlo Algorithmen. Danach werden Werkzeuge zur Konvergenzuntersuchung der Folge von Zufallszügen, die man über Permutation Sampling beispielsweise erhält, präsentiert. Ebenso werden verschiedene Methoden zur Modellselektion behandelt. Außerdem wird das Prinzip von Mischungsmodellen anhand mehrerer literarischer Beiträge vorgestellt, bevor letztendlich der Performancevergleich der Modelle anhand von Simulationsdaten erfolgt.

Abstract

In econometry one is often confronted with problems which can hardly be solved by ordinary estimation methods. In such cases the bayesian approach is an efficient alternative. For example the bayesian method of Markov Chain Monte Carlo Sampling represents a possibility of estimating the parameters of the finite mixture model.

The dissertation comprises three approaches considering heterogeneity in consumer data. For the first time they are explicitly compared with respect to their performance by the use of artificial data. Yet another innovation is the data generation process, because the data are produced by a simulation model where it is not necessary to make restrictive assumptions, for example regarding the distribution of the parameters. Only the market structure concerning the population segmentation is specified.

The first approach is a latent class model that enables the estimation of segmentspecific parameters. On the other hand a random effects model can be considered where the parameters follow a continuous distribution and can vary across the population. The third model is a combination of the first two approaches, that means the segmentspecific coefficients are defined as random effects and therefore are further allowed to vary within the consumer groups.

For the estimation of the models the bayesian method of Markov Chain Monte Carlo Sampling is used. Especially restricted Permutation Sampling is applied.
For that reason the first part of the work concentrates on the basics of the bayesian estimation method and introduces different Markov Chain Monte Carlo algorithms. Then different instruments for the examination of the convergence of the random draws, supplied by a Permutation Sampler for example, are presented. Several methods for model selection are discussed as well. Thereafter the principle of mixture models is described with the help of various literary contributions. Only then the performance comparison follows.

Danksagung

Diese Arbeit wurde am Institut für Wirtschaftsmathematik der Technischen Universität Wien ausgeführt und ist in enger Zusammenarbeit mit dem Institut für Tourismus und Freizeitwirtschaft der Wirtschaftsuniversität Wien entstanden. Ohne der fachlichen und persönlichen Unterstützung von beiden Seiten wäre es nicht möglich gewesen, diese Arbeit in der hier vorliegenden Form auszuführen. Mein besonderer Dank gilt hierbei Herrn Prof. Dr. Manfred Deistler (TU Wien) sowie Herrn Prof. Dr. Josef Mazanec (WU Wien).

Ebenso spreche ich Herrn Dr. Jürgen Wöckl meinen Dank aus. Er hat mir vor allem im Rahmen der Generierung der Simulationsdaten wichtige Anregungen und Hilfestellungen gegeben. Außerdem bedanke ich mich bei Frau Dr. Regina Tüchler für die Bereitstellung diverser Matlab-Routinen zur MCMC Schätzung.

Nicht zuletzt möchte ich noch ein großes Dankeschön an meine Familie richten, die mich über viele Jahre hinweg auf meinem Weg unterstützt und mir diesen Bildungsweg ermöglicht hat.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

AIC	...	Akaike Information Criterion
ASP	...	Aspiration
ATT	...	Attitude
BIC	...	Bayes Information Criterion
BMA	...	Bayesian Model Average
CAIC	...	Consistent Akaike Information Criterion
CMXL	...	Continuous Mixed Logit
CUSUM	...	Cumulated Sum
DAG	...	Directed Acyclic Graph
FM	...	Finite Mixture
HB	...	Hierarchical Bayes
HPD	...	Highest Probability Density
ICOMP	...	Informational Complexity Criterion
LCM	...	Latent Class Model
LOGLV	...	Loglikelihood Validation
LR	...	Likelihood Ratio
MCMC	...	Markov Chain Monte Carlo
ML	...	Maximum Likelihood
MNL	...	Multinomial Logit
MNMXL	...	Mixture-Normal Mixed Logit
MOM	...	Method of Moments
MPMXL	...	Mass-Point Mixed Logit
MSM	...	Method of Simulated Moments
MSS	...	Method of Simulated Scores
MXL	...	Mixed Logit
NEC	...	Normed Entropy Criterion
PCEP	...	Perception
RMSE	...	Root Mean Square Error
UTI	...	Utility

1 Einleitung

In der Ökonometrie stößt man sehr oft auf Probleme, die man nur schwer bzw. manchmal auch gar nicht mit herkömmlichen Schätzmethoden, wie z.B. dem gewöhnlichen Kleinstquadrateschätzer oder dem Maximum Likelihood Schätzer, lösen kann. Der bayesianische Ansatz stellt in derartigen Fällen eine recht praktische und überaus effiziente Alternative dar. So eignet sich die bayesianische Methode des Markov Chain Monte Carlo (MCMC) Samplings auch zur Schätzung der Parameter eines endlichen Mischungsmodells, mit dem es möglich ist gruppenspezifische Koeffizienten zu definieren.

In Kapitel 2 werden die Grundlagen des bayesianischen Ansatzes, insbesondere Bayes' Theorem genauer erläutert. Es werden weiters die Begriffe „a priori" und „a posteriori" in diesem Zusammenhang erklärt.
Im Anschluss wird im Speziellen auf Markov Chain Monte Carlo Sampling eingegangen, und es folgt außerdem die Präsentation einiger wichtiger Algorithmen.
Kapitel 4 befasst sich mit der Analyse und Interpretation der Schätzungen, die MCMC Sampling liefert, wobei das Hauptaugenmerk auf das Konvergenzverhalten des Outputs gerichtet wird.
Abschnitt 5 behandelt das Problem der Modellselektion, wobei unterschiedliche Methoden hierfür vorgestellt werden.
In Kapitel 6 geht es dann um die grundsätzliche Frage der Modellformulierung, insbesondere wird der interessante Aspekt der Relation zu graphischen Modellen angeschnitten.
Im darauffolgenden Abschnitt wird das Prinzip der Mischungsmodelle anhand einer Sammlung mehrerer Veröffentlichungen unterschiedlicher Autoren genauer behandelt.
Der letzte Teil schließlich umfasst eine umfangreiche Studie von Mischungsmodellen unter Verwendung mittels einer Simulation eines Konsumentenmarktes künstlich erzeugter Daten. Es wird versucht, die wahre Struktur speziell hinsichtlich der Segmentierung der Marktteilnehmer in die tatsächlichen Gruppen mittels MCMC Sampling wieder herauszuschätzen.

Begonnen wird nun mit einer allgemeinen Einführung in die Theorie der bayesianischen Schätzmethode.

2 Grundlagen der bayesianischen Analyse

Statistische Inferenz beschäftigt sich mit Schlussfolgerungen über unbeobachtete Größen aus beobachteten Daten. Zu den unbeobachteten Größen können

- Parameter,
- zukünftige Beobachtungen oder fehlende Daten und
- künstlich eingeführte latente Variablen

gehören.

Beim bayesianischen Ansatz werden anfängliche Vorstellungen[1] mittels Bayes' Theorem mit der Information aus den Daten[2] kombiniert, um a posteriori Wahrscheinlichkeiten bezüglich Parameter oder Hypothesen zu erhalten. Die Parameter werden als Zufallsvariablen betrachtet und Schlussfolgerungen aufgrund ihrer Verteilungen bedingt auf die fixierten Daten getroffen. Bayesianische Prozeduren vermeiden speziell zwei der numerischen Schwierigkeiten, welche mit den klassischen Schätzern assoziiert werden:

1. Die Methoden erfordern keine Maximierung einer Funktion.
2. Die erwünschten Schätzeigenschaften wie Konsistenz und Effizienz bleiben unter schwächeren Bedingungen erhalten.

Die bayesianische Inferenz arbeitet auf einem weiten Bereich an Problemen und Modellen sowohl für kleine als auch große Stichproben gut. Ein großer Vorteil des Ansatzes ist, dass er verwendet werden kann, um die Konsequenzen jedes Typs von Wahrscheinlichkeitsmodell zu erforschen ohne sich auf jene mit einer speziellen mathematischen Form beschränken zu müssen.

2.1 Bayes' Theorem

Ein essentielles Element des bayesianischen Ansatzes ist Bayes' Theorem, in der Literatur auch Prinzip der inversen Wahrscheinlichkeit[3] genannt.

[1] durch a priori Wahrscheinlichkeiten repräsentiert

[2] in der Likelihoodfunktion beinhaltet

[3] Bei Problemen, die „inverse Wahrscheinlichkeiten" enthalten, hat man Daten gegeben und versucht aus der Information aus den Daten zu folgern, durch welchen Zufallsprozess diese generiert wurden.

Bayes' Theorem kann für jede Art von Modell angewendet werden, um a posteriori Verteilungen für Parameter zu erhalten.

$y' = (y_1, \ldots, y_n)$ sei ein Vektor von n Beobachtungen, dessen Wahrscheinlichkeitsverteilung $p(y|\theta)$ von den k Parametern $\theta' = (\theta_1, \ldots, \theta_k)$ abhängt. θ selbst besitzt eine Wahrscheinlichkeitsverteilung $p(\theta)$. Dann gilt:

$$p(y|\theta)p(\theta) = p(y, \theta) = p(\theta|y)p(y) \ .$$

Gegeben die beobachteten Daten y ist also die bedingte Verteilung von θ

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \ .$$

Man kann auch schreiben

$$p(y) = \mathbb{E}p(y|\theta) = c^{-1} = \begin{cases} \int p(y|\theta)p(\theta)d\theta & \theta \text{ kontinuierlich} \\ \sum p(y|\theta)p(\theta) & \theta \text{ diskret,} \end{cases}$$

wobei über den zulässigen Bereich von θ integriert bzw. summiert wird.[4]

Man kann Bayes' Theorem somit alternativ schreiben als

$$p(\theta|y) = cp(y|\theta)p(\theta) \ ,$$

wobei $p(\theta)$ beschreibt, was ohne Kenntnis der Daten über θ bekannt ist und wird als a priori Verteilung von θ[5] bezeichnet. $p(\theta|y)$ beschreibt, was gegeben die Daten über θ bekannt ist. Sie wird a posteriori Verteilung von θ gegeben y[6] genannt. c ist eine normalisierende Konstante[7], die sicherstellt, dass sich die a posteriori Verteilung $p(\theta|y)$ auf 1 integriert bzw. summiert.
Gegeben die Daten y kann $p(y|\theta)$ auch als Funktion von θ statt von y betrachtet werden. Sie wird dann Likelihoodfunktion von θ gegeben y genannt und mit $l(\theta|y)$ bezeichnet. Für die Bayes' Formel ergibt sich daraus

$$p(\theta|y) \propto l(\theta|y)p(\theta) \ .$$

---

[4] $p(y)$ ist unabhängig von θ und kann deshalb als konstant angenommen werden.
[5] Verteilung von θ a priori („Prior")
[6] Verteilung von θ a posteriori („Posterior")
[7] Sie wird oft als marginale Verteilung der Daten bezeichnet.

In Worten ausgedrückt ist die a posteriori Wahrscheinlichkeitsverteilung[8] für θ proportional zum Produkt der a priori Verteilung[9] und der Likelihood für θ gegeben y:

$$\text{a posteriori Verteilung} \propto \text{Likelihood} \times \text{a priori Verteilung} .$$

Die a posteriori Verteilung $p(\theta|y)$ dient dazu, um Schlussfolgerungen über die Parameter zu ziehen und beinhaltet das gesamte Vorwissen sowie die Stichprobeninformation. Die a priori Information wird über die a priori Verteilung in die a posteriori Verteilung aufgenommen, die gesamte Stichprobeninformation fließt über die Likelihoodfunktion ein. Die Likelihoodfunktion $l(\theta|y)$ ist die Funktion, mittels der das a priori Wissen bezüglich θ über die Daten modifiziert wird. Sie kann deshalb als die aus den Daten stammende Information über θ betrachtet werden. Sie ist bis auf eine multiplikative Konstante definiert, d.h. Multiplikation mit einer Konstante lässt die Likelihood unverändert.[10] Nur der relative Wert der Likelihood ist also von Bedeutung.

Das Ziel liegt in der Herleitung der a posteriori Verteilung $p(\theta|y)$, die das Wissen über θ bedingt auf die Beobachtung der Daten y repräsentiert. In den meisten Modellen besitzt sie keine analytisch geschlossene Form.
Wie kann man aber dann Stichproben daraus ziehen? Eine der populärsten Methoden hierfür stellt der Gibbs Sampler dar, mit dem es möglich ist, aus $p(\theta|y)$ zu ziehen, ohne jedoch c kennen zu müssen (vgl. Kapitel 3.2.1). Bei der Monte Carlo Integration werden Stichproben aus der gewünschten Verteilung gezogen und daraus Mittel als Approximation der Erwartungswerte gebildet. Beim Markov Chain Monte Carlo (MCMC) Sampling insbesondere werden diese Stichproben mittels Durchlaufen einer Markovkette über eine lange Zeit gezogen. MCMC Sampling ist also eine numerische Methode, um Integrale zu berechnen. Die Simulation besteht aus dem Ziehen aus einer Dichte, dem Berechnen einer Statistik für jeden Zug und dem Mitteln der Ergebnisse. Die Simulation ermöglicht vor allem, dass die Schätzer sogar dann imple-

[8] die Verteilung *nachdem* die Daten bekannt sind

[9] die Verteilung *bevor* die Daten bekannt sind

[10] Das Multiplizieren der Likelihoodfunktion mit einer willkürlichen Konstante wird keinen Effekt auf die a posteriori Verteilung von θ haben.
Die standardisierte Likelihood

$$\frac{l(\theta|y)}{\int l(\theta|y)d\theta}$$

ist so skaliert, dass die Fläche unter der Kurve eins beträgt.

mentiert werden können, wenn das Integral, das den Schätzer definiert, keine geschlossene Form annimmt.

Unterschiedliche Datenmengen

Bayes' Theorem erlaubt es, die Information über eine Menge von Parametern θ fortwährend zu aktualisieren, falls zusätzliche Beobachtungen gemacht werden.
Angenommen man besitzt eine anfängliche Stichprobe von Beobachtungen y_1, dann ergibt Bayes' Formel

$$p(\theta|y_1) \propto p(\theta)l(\theta|y_1) \ .$$

Steht eine zweite Stichprobe von Beobachtungen y_2 unabhängig verteilt von der ersten zur Verfügung, gilt

$$\begin{aligned} p(\theta|y_2, y_1) &\propto p(\theta)l(\theta|y_1)l(\theta|y_2) \\ &\propto p(\theta|y_1)l(\theta|y_2). \end{aligned}$$

Die a posteriori Verteilung $p(\theta|y_1)$ für θ gegeben y_1 spielt hier die Rolle der a priori Verteilung für die zweite Stichprobe.
Dieser Prozess kann beliebig oft wiederholt werden.

2.2 A priori Verteilungen

Während a priori Information hinsichtlich Parameter oder Modelle oft in Schätzanalysen oder bei Vorhersageproblemen in Form von exakten Restriktionen eingebunden werden, können Bayesianer in solchen Situationen weniger restriktiv mittels Einführung einer passenden a priori Verteilung vorgehen.
Die a priori Information bezüglich der Modellparameter wird durch eine geeignet gewählte Verteilung repräsentiert. Falls möglich sollte die Prior $p(\theta)$ aus der natürlich konjugierten Familie gewählt werden, d.h. die Kombination der Prior mit der Likelihood sollte abermals eine a posteriori Verteilung aus der Verteilungsfamilie der Prior ergeben.

2.2.1 Proper und improper Priors

Eine grundlegende Eigenschaft einer Wahrscheinlichkeitsdichte $f(x)$ ist, dass sie sich über ihren zulässigen Bereich auf 1 integriert bzw. summiert:

$$\left.\begin{array}{l}\int f(x)dx \\ \sum f(x)\end{array}\right\} = 1 \quad \left\{\begin{array}{l} x \text{ kontinuierlich} \\ x \text{ diskret} .\end{array}\right.$$

Priors, die diese Anforderung erfüllen, werden „proper" genannt.
Eine a priori Dichte $p(\theta)$ ist somit genau dann proper, wenn sie sich über den Parameterraum auf 1 integriert.

Proper Posteriors aus improper Priors?

Eine improper Prior von θ definiert zwar kein gemeinsames Wahrscheinlichkeitsmodell $p(y, \theta)$, das proper ist, trotzdem führen viele zu einer proper a posteriori Verteilung. Bei Mischungs- und Switchingmodellen, sowie bei Zufallseffektmodellen ist das allerdings nicht der Fall.

2.2.2 Schwach informative Priors

In der bayesianischen Analyse spielt die a priori Verteilung eine wichtige Rolle. Sie dient der Repräsentation des Wissens über die unbekannten Parameter, bevor Daten zur Verfügung stehen.
Es kommt oft vor, dass a priori Wissen gar nicht oder nur sehr vage existiert. Die Herausforderung liegt dann in der Wahl einer Prior, die möglichst wenig Information einfließen lässt. Will man also die Daten „für sich selbst sprechen lassen", wird man nicht-informative oder nur sehr gering informative (schwach informative) a priori Verteilungen einsetzen.
Die Ungewissheit eines Forschers wird in der Varianz der Prior ausgedrückt. Ist diese groß, hat er nur eine vage Vorstellung vom Wert des Parameters. Flache/Diffuse Priors repräsentieren geringe Information, und alle möglichen Werte werden als gleich wahrscheinlich eingestuft. Solche diffusen Priors stellen z.B. Normalverteilungen mit großer Varianz oder Wishart[11]- bzw. invertierte Gammaverteilungen[12] mit einer geringen Anzahl an Freiheitsgraden dar.

---

[11] Wishartverteilung $W(\rho, R)$:

$$\Phi_{ij}^{-1} \sim N(0, R),\ i = 1, \ldots, \rho, \quad \Phi_i^{-1} = \begin{pmatrix} \Phi_{i1}^{-1} \\ \vdots \\ \Phi_{i\rho}^{-1} \end{pmatrix} \quad \Rightarrow\ \Phi_i^{-1'} \Phi_i^{-1} \sim W(\rho, R)$$

[12] invertierte Gammaverteilung $IG(\nu_\varepsilon, G_\varepsilon)$:

$$f(\theta|\nu_\varepsilon, G_\varepsilon) = \left(\tfrac{1}{\theta}\right)^{\nu_\varepsilon + 1} e^{-\frac{G_\varepsilon}{\theta}} c \quad \text{mit} \quad c = \frac{G_\varepsilon^{\nu_\varepsilon}}{\Gamma(\nu_\varepsilon)}$$

Für den Fall dass der Wert eines Parameters völlig unbekannt ist, schlägt Jeffreys zwei Regeln vor. Kann der Parameter jeden Wert in einem endlichen Bereich (oder von $-\infty$ bis ∞) annehmen, sollte seine a priori Wahrscheinlichkeit als gleichverteilt angenommen werden. Kann er jeden Wert von 0 bis ∞ annehmen, sollte die a priori Wahrscheinlichkeit seines Logarithmus als gleichverteilt angenommen werden.[13]
Zögert man in der Verwendung der von Jeffreys vorgeschlagenen improper Verteilungen, kann man „lokal gleichverteilte" oder „sanfte" a priori Verteilungen für unbekannte Parameter einführen. Gemeint sind damit a priori Verteilungen, die über einem Bereich, in dem die Likelihoodfunktion bedeutende Werte annimmt, „einigermaßen flach" sind. Außerhalb dieses Bereiches ist es irrelevant, welche Form die a priori Verteilung besitzt, da ja bei der Herleitung der a posteriori Verteilung die a priori Verteilung mit kleinen Likelihoodwerten multipliziert wird. Eine Prior, die von der Likelihood dominiert wird, ändert sich also in der Region, in der die Likelihood nennenswert ist, nicht sehr stark.
Für eine lokal gleichverteilte Prior (siehe Abb. 1) ist die a posteriori Verteilung annähernd numerisch gleich der standardisierten Likelihood.

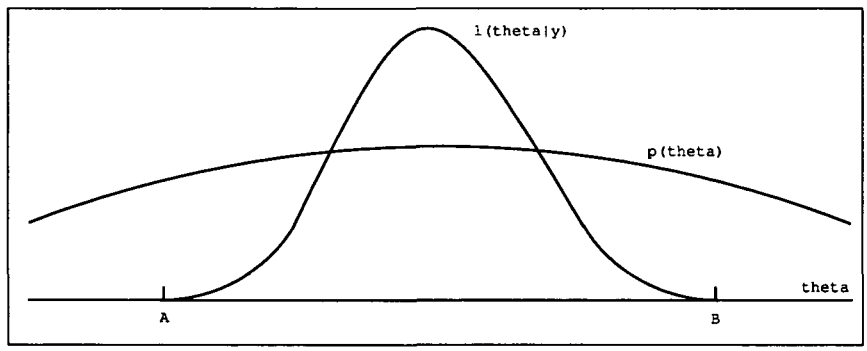

Abbildung 1: Beispiel einer „lokal gleichverteilten" Prior $p(\theta)$
(Quelle: [44], S. 46)

Selbst wenn eine starke Überzeugung bezüglich der Werte eines Parameters vorhanden ist, ist es vorteilhafter und stichhaltiger, wenn man die Daten gegen eine Referenzprior[14], die von der Likelihood dominiert wird, analysiert.

[13] siehe [44], S. 41-53
[14] Als Referenzprior bezeichnet man eine Prior, die sich als Standard eignet.

2.2.3 A priori Unabhängigkeit von Parametern

Bestimmte Parameter oder Mengen von Parametern können a priori als unabhängig verteilt von gewissen anderen Parametern oder Mengen von Parametern beurteilt werden. Manchmal vereinfacht sich dadurch die Wahl der a priori Verteilung, da man die unabhängigen Parametermengen getrennt betrachten kann.
Im Speziellen eignet es sich beispielsweise Lokationsparameter als unabhängig verteilt von Skalenparametern anzunehmen.

2.3 A posteriori Verteilung

Bei der bayesianischen Inferenz werden die Parameter als Zufallsvariablen definiert und Schlussfolgerungen durch Betrachtung ihrer Verteilungen bedingt auf die fixierten Daten getroffen. Die Anzahl und Art der Größen, die notwendig sind, um die a posteriori Verteilung annähernd zu beschreiben, hängen vom Typ und der Komplexität dieser Verteilung ab. Manchmal reichen zusammenfassende Statistiken, welche wichtige Eigenschaften der a posteriori Verteilung umreissen.

2.3.1 Zusammenfassung des a posteriori Wissens

Der Kern der bayesianischen Inferenz liegt in der Entwicklung des Modells $p(y, \theta)$ und der Zusammenfassung der a posteriori Dichte $p(\theta|y)$ in einer geeigneten Art und Weise.
Nachdem man eine Folge von Zufallszahlen $\{\theta^{(i)}\}$ generiert hat, werden die ersten m der insgesamt n Iterationen vernachlässigt. Sie werden als Übergangsperiode oder Burn-in Phase bezeichnet.

1. Punktschätzungen:
   Die a posteriori Parameter werden durch entsprechende zusammenfassende Statistiken aus $\{\theta^{(i)}\}$ approximiert:

   - a posteriori Mittelwert $\approx$ Stichprobenmittelwert:

$$\mathbb{E}(\theta|y) \approx \frac{1}{n-m} \sum_{i=m+1}^{n} \theta^{(i)} ,$$

   - a posteriori Median $\approx$ Stichprobenmedian,

- a posteriori Standardabweichung $\approx$ Stichprobenstandardabweichung,
- a posteriori Kovarianz $\approx$ Stichprobenkovarianz.

2. Randverteilungen:
   Man kann Histogramme basierend auf $\{\theta^{(i)}\}$ erzeugen.
   Besser geeignet aber auch arbeitsaufwendiger ist das Mitteln über die Verteilungen.

2.3.2 Marginale und bedingte a posteriori Verteilungen

Bedingte a posteriori Verteilungen ermöglichen die Untersuchung der Sensitivität der Schlussfolgerungen über bestimmte Teilmengen von Parametern hinsichtlich der für andere Parameter getroffenen Annahmen.
Weiters können sogenannte „Störparameter“[15] aus einer a posteriori Verteilung ausintegriert werden, um die marginale a posteriori Verteilung für die relevanten Parameter zu erhalten. In der Sampling Theorie hängen „optimale“ Schätzer oder Teststatistiken oft von „Störparametern“ ab, deren Werte nicht bekannt sind. In vielen Fällen werden diese „Störparameter“ durch Stichprobenschätzungen ersetzt, was allerdings lediglich eine Approximation des optimalen Schätzers darstellt und üblicherweise nur für große Stichproben gerechtfertigt werden kann. Ein Bayesianer hingegen kann sich ohne Bedenken auch auf eine endliche Stichprobenanalyse beschränken, da folgende Integration eine nützliche Art bietet, um sich derartiger „Störparameter“ zu entledigen.
θ sei beispielsweise partitioniert in $\theta' = (\theta_1', \theta_2')$, und man sucht die marginale a posteriori Verteilung für θ_1. Diese ergibt sich als

$$\begin{aligned} p(\theta_1|y) &= \int_{R_{\theta_2}} p(\theta_1, \theta_2|y) d\theta_2 \\ &= \int_{R_{\theta_2}} p(\theta_1|\theta_2, y) p(\theta_2|y) d\theta_2 \end{aligned}$$

mit R_{θ_2} als Parameterraum für θ_2 und $p(\theta_1|\theta_2, y)$ als bedingte a posteriori Verteilung für θ_1 gegeben θ_2 und die Stichprobeninformation y.

[15] Parameter, die für die Untersuchung nicht von Interesse sind (auch „nuisance parameter“ genannt)

Die marginale a posteriori Verteilung für θ_1 kann als Durchschnitt von bedingten a posteriori Verteilungen $p(\theta_1|\theta_2, y)$ mit der marginalen a posteriori Verteilung für θ_2 als Gewichtsfunktion betrachtet werden.

2.4 Eigenschaften des bayesianischen Schätzers

Bayesianische Prozeduren liefern einen Schätzer, der asymptotisch äquivalent zum Maximum Likelihood (ML) Schätzer ist. Man kann also das Ergebnis interpretieren, als ob es sich um einen ML Schätzer handelt. Anstatt die Likelihoodfunktion zu maximieren kann man den Mittelwert der Posterior berechnen, und der resultierende Schätzer ist im klassischen Sinne gleich gut.
Der Mittelwert der Posterior $\bar{\theta} = \int \theta p(\theta|y)d\theta$ konvergiert gegen das Maximum der Likelihoodfunktion $\hat{\theta}$ (d.h. $\bar{\theta} - \hat{\theta}$ verschwindet asymptotisch), da die Posterior normal wird (d.h. der Mittelwert entspricht dem Maximum) und der Effekt der Prior mit steigender Stichprobengröße verschwindet. Da $\bar{\theta}$ und $\hat{\theta}$ konvergieren, sind ihre asymptotischen Samplingverteilungen die gleichen.[16]

Der Mittelwert der R Züge von θ aus der a posteriori Verteilung simuliert den Mittelwert der Posterior und stellt somit die Parameterschätzungen dar:

$$\check{\theta} = \frac{1}{R}\sum_{r=1}^{R}\theta^r \ .$$

Der simulierte Mittelwert der Posterior ist konsistent und asymptotisch normal für festes R und effizient und äquivalent zum ML Schätzer, falls R mit beliebiger Rate mit der Stichprobengröße N steigt.[17]

3 Markov Chain Monte Carlo Sampling

Es kann vorkommen, dass die gemeinsame a posteriori Verteilung keine einfache Struktur mehr aufweist. Man steht also vor dem Problem, aus einer Verteilung p ziehen zu wollen, es jedoch nicht möglich ist direkt eine Stichprobe daraus zu generieren.
Eine bessere Approximation des Integrals als eine Folge von reinen Zufallszügen

[16] Den exakten Beweis findet man in [41], S. 291 - 294.
[17] Die Beweise findet man in [41], S. 294 - 296.

wie beim Importance Sampling liefern Markov Chain Monte Carlo Methoden, zu denen beispielsweise Metropolis-Hastings Sampling zählt. Stichproben können über eine Markovkette, die $p(\cdot)$ als ihre stationäre Verteilung besitzt, generiert werden. Der Name Markov Chain Monte Carlo stammt daher, dass erstens Zufallszüge vorgenommen werden (Monte Carlo) und zweitens jeder Wert lediglich vom unmittelbar vorangegangenen abhängt (Markovkette).
Man generiert eine Folge von Zufallsvariablen $\{\theta_0, \theta_1, \theta_2, \ldots\}$, sodass zu jedem Zeitpunkt $t \geq 0$ der nächste Zustand θ_{t+1} aus einer nur vom momentanen Zustand abhängigen Verteilung $p(\theta_{t+1}|\theta_t)$ gezogen wird. Der Übergangskern $p^{(t)}(\cdot|\theta_0)$ konvergiert gegen eine eindeutige stationäre Verteilung $p(\cdot)$, die nicht von t oder θ_0 abhängt. Nach einem hinreichend langen Burn-in von m Iterationen werden die Punkte $\{\theta_t; t = m+1, \ldots, n\}$ abhängige Stichproben annähernd aus $p(\cdot)$ sein.
Die Länge m des Burn-ins hängt vom Startwert θ_0, der Konvergenzrate und der verlangten Ähnlichkeit zwischen der stationären und der tatsächlichen Verteilung ab. Hat man ein Kriterium „ähnlich genug“ definiert, kann m analytisch bestimmt werden. Meistens bedient man sich allerdings hierfür der visuellen Inspektion diverser Plots des Monte Carlo Outputs $\{\theta_t\}$. Die Kette sollte lange genug laufen, um die Startposition zu „vergessen“. Im Wesentlichen kann man also bei ausreichend langer Laufzeit von jedem Punkt ausgehen, dennoch ist es empfehlenswert extreme Startwerte zu meiden.

3.1 Importance Sampling

Will man beispielsweise das Integral $\int t(\varepsilon) f(\varepsilon) d\varepsilon$ berechnen, kann jedoch nicht leicht aus f gezogen werden, formt man dieses Integral einfach um zu

$$\int t(\varepsilon) \frac{f(\varepsilon)}{g(\varepsilon)} g(\varepsilon) d\varepsilon \ ,$$

wobei f Zieldichte und g Vorschlagsdichte genannt werden. Bei g handelt es sich um eine Dichte, aus der man leicht ziehen kann. Es müssen lediglich folgende beiden Anforderungen erfüllt sein:

1. Der Support[18] von $g(\varepsilon)$ muss den Support von f überdecken. D.h. jedes beliebige ε, das in f auftreten kann, muss auch in g auftreten können.

2. Es muss gelten
$$\frac{f(\varepsilon)}{g(\varepsilon)} < \infty \qquad \forall \varepsilon \ .$$

Statt aus der Dichte f zieht man nun aus der Dichte g und gewichtet die Züge mit $\frac{f(\varepsilon)}{g(\varepsilon)}$. Die Menge der gewichteten Züge entspricht der Menge der Züge aus f. Für jeden Zufallszug berechnet man dann $t(\varepsilon)\frac{f(\varepsilon)}{g(\varepsilon)}$ und mittelt anschließend die Ergebnisse.

Eine „Akzeptieren-Verwerfen" (accept-reject) Prozedur kann als eine Variante des Importance Samplings betrachtet werden. Die Züge werden in diesem Fall entweder mit 1 (accept) oder mit 0 (reject) gewichtet.

Genaueres über Importance Sampling findet man in [13], S. 125-127.

3.2 Metropolis-Hastings Algorithmus

Metropolis-Hastings Sampling kann bei jeder beliebigen Dichte angewendet werden. Die Methode ist speziell nützlich, wenn die normalisierende Konstante nicht bekannt ist oder aufgrund der dafür notwendigen Integration nur schwer berechnet werden kann. Die Züge werden nämlich mit einer Wahrscheinlichkeit akzeptiert, bei der sich die normalisierende Konstante im Bruch kürzen lässt.
Zuerst wird ein Kandidatenpunkt θ^* aus einer Vorschlagsdichte[19] $q(\vartheta|\theta)$, für die gilt $\int q(\vartheta|\theta)d\vartheta = 1$, gezogen. Dieser wird mit einer Wahrscheinlichkeit $\alpha(\theta, \vartheta)$ akzeptiert. Wird der Kandidatenpunkt akzeptiert, erhält man als nächsten Zustand $\theta_{t+1} = \theta^*$. Wird der Kandidatenpunkt verworfen, bewegt sich die Kette nicht, d.h. $\theta_{t+1} = \theta_t$.

$U(0,1)$ bezeichne die Gleichverteilung über $(0,1)$. Dann ergibt sich eine allgemeine Version des Metropolis-Hastings Algorithmus zum Ziehen aus der a posteriori Verteilung $p(\theta|y)$.

[18] Bildraum der Zufallsvariable, d.h. Y ist der Support, wenn die Zufallsvariable von $X \to Y$ abbildet

[19] Übergangsdichte

Metropolis-Hastings Algorithmus:

Schritt 0: Wähle einen willkürlichen Startpunkt θ_0 und setze $i = 0$.

Schritt 1: Generiere einen Kandidatenpunkt θ^* aus $q(\cdot|\theta_i)$ und ein u aus $U(0,1)$.

Schritt 2: Setze $\theta_{i+1} = \theta^*$, falls $u \leq \alpha(\theta_i, \theta^*)$ und $\theta_{i+1} = \theta_i$ sonst, wobei die Akzeptanzwahrscheinlichkeit durch

$$\alpha(\theta, \vartheta) = \min\left\{\frac{p(\vartheta|y)q(\theta|\vartheta)}{p(\theta|y)q(\vartheta|\theta)}, 1\right\}$$

definiert ist.

Schritt 3: Setze $i = i + 1$ und gehe zu Schritt 1.

Der Metropolis Algorithmus berücksichtigt nur symmetrische Vorschlagsdichten $q(\vartheta|\theta) = q(\theta|\vartheta)\ \forall\theta, \vartheta$. Die Akzeptanzwahrscheinlichkeit reduziert sich in diesem Fall auf

$$\alpha(\theta, \vartheta) = \min\left\{\frac{p(\vartheta|y)}{p(\theta|y)}, 1\right\} .$$

Die Vorschlagsdichte q kann jede Form besitzen, die stationäre Verteilung der Kette wird immer p sein. Allerdings hängt die Performance des Metropolis-Hastings Algorithmus von ihrer Wahl ab. Die Spannweite der Vorschlagsdichte q beeinflusst die „Akzeptanzrate“[20] und die Region des Samplingraumes, der durch die Kette abgedeckt wird. Ist die Spannweite sehr groß, werden einige Kandidaten nur mit geringer Wahrscheinlichkeit akzeptiert. Andererseits wird es bei einer zu kleinen Spannweite lange dauern, bis der Support der Dichte durchlaufen wurde. Beide Situationen spiegeln sich in hohen Autokorrelationen der Stichprobenwerte wider. Grundsätzlich sollte man q derart wählen, dass sie leicht abgetastet und ausgewertet werden kann.

- Eine übliche Wahl für $q(\vartheta|\theta)$ ist eine Zufallsbewegung (Random Walk[21]) $q(\vartheta|\theta) = q^*(\vartheta - \theta)$ für eine Verteilung q^*. Das Metropolis-Hastings

[20] Prozentsatz, wie oft eine Bewegung zu einem neuen Punkt erfolgt

[21] Der Kandidat wird entsprechend $\theta^* = \theta + \omega$ (d.h. momentaner Wert + Störterm) gezogen.

Verhältnis entspricht in diesem Fall

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

Falls $q(\vartheta) = p(\vartheta)$, ist das Metropolis-Hastings Verhältnis identisch 1, und man zieht aus der Zieldichte.

- Beim Independence Sampler hängt die Vorschlagsdichte nicht von θ ab, d.h. $q(\vartheta|\theta) = q(\vartheta)$. Es ergibt sich daher

$$\alpha(\theta, \vartheta) = \min\left\{\frac{p(\vartheta|y)q(\theta)}{p(\theta|y)q(\vartheta)}, 1\right\} .$$

- Beim Einzelkomponenten Metropolis-Hastings Sampler wird θ in Komponenten $\{\theta_{.1}, \theta_{.2}, \ldots, \theta_{.h}\}$ gespalten, welche einzeln aktualisiert werden. Typischerweise arbeitet man mit niedrigdimensionalen oder skalaren Komponenten.
  Es sei $\theta_{.-i} = \{\theta_{.1}, \ldots, \theta_{.i-1}, \theta_{.i+1}, \ldots, \theta_{.h}\}$, sodass $\theta_{.-i}$ das gesamte θ mit Ausnahme der i-ten Komponente umfasst. $\theta_{t.i}$ bezeichne den Zustand von $\theta_{.i}$ am Ende der Iteration t. Für Schritt i der Iteration $t+1$ wird $\theta_{.i}$ mittels Metropolis-Hastings aktualisiert. Der Kandidat $\theta^*_{.i}$ wird aus einer Vorschlagsdichte $q_i(\theta^*_{.i}|\theta_{t.i}, \theta_{t.-i})$ generiert mit $\theta_{t.-i} = \{\theta_{t+1.i}, \ldots, \theta_{t+1.i-1}, \theta_{t.i+1}, \ldots, \theta_{t.h}\}$ als Wert von $\theta_{.-i}$ nach Vollendung von Schritt $i-1$ der Iteration $t+1$. Der Kandidat wird mit Wahrscheinlichkeit

$$\alpha(\theta_{.-i}, \theta_{.i}, \theta^*_{.i}) = \min\left\{\frac{p(\theta^*_{.i}|\theta_{.-i}, y)q_i(\theta_{.i}|\theta^*_{.i}, \theta_{.-i})}{p(\theta_{.i}|\theta_{.-i}, y)q_i(\theta^*_{.i}|\theta_{.i}, \theta_{.-i})}, 1\right\}$$

akzeptiert. Falls $\theta^*_{.i}$ akzeptiert wird, setzt man $\theta_{t+1.i} = \theta^*_{.i}$, ansonsten gilt $\theta_{t+1.i} = \theta_{t.i}$.
Eine fixe Reihenfolge, in der die Komponenten gezogen werden, ist nicht unbedingt erforderlich. Weiters müssen nicht alle Komponenten in jeder Iteration aktualisiert werden. Man kann sich stattdessen beispielsweise auf nur eine Komponente pro Iteration beschränken, wobei Komponente i mit fester Wahrscheinlichkeit $s(i)$ gewählt wird.[22]

[22] Eine natürliche Wahl hierfür stellt $s(i) = \frac{1}{h}$ dar.

3.2.1 Gibbs Sampling

Für multinomiale Verteilungen kann es manchmal schwierig sein, direkt aus der gemeinsamen Dichte zu ziehen. Oft ist es einfacher, Stichproben aus den bedingten Dichten für jedes einzelne Element gegeben alle anderen zu generieren. Es wird folglich iterativ aus den bedingten Dichten gezogen, wobei dieser Prozess gegen die gemeinsame Dichte konvergiert.
Gibbs Sampling stellt einen Sonderfall des Einzelkomponenten Metropolis-Hastings Algorithmus mit spezieller Wahl von q dar. Die Vorschlagsdichte für das Update der i-ten Komponente von θ lautet hier speziell

$$q_i(\theta^*_{.i}|\theta_{.i}, \theta_{.-i}) = p(\theta^*_{.i}|\theta_{.-i}, y)$$

mit $p(\theta^*_{.i}|\theta_{.-i}, y)$ als voll bedingte Verteilung. Das führt zu einer Akzeptanzwahrscheinlichkeit identisch 1, was bedeutet dass beim Gibbs Sampler jeder Kandidatenpunkt akzeptiert wird.

Es sei $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$ ein k-dimensionaler Parametervektor und $p(\theta|y)$ dessen a posteriori Verteilung gegeben die Daten y. Im ersten Schritt müssen für die Elemente, auf die bedingt wird, Anfangswerte gewählt werden.

Gibbs Sampling Algorithmus:

Schritt 0: Wähle einen willkürlichen Startpunkt $\theta_0 = (\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{k,0})'$ und setze $i = 0$.

Schritt 1: Generiere $\theta_{i+1} = (\theta_{1,i+1}, \theta_{2,i+1}, \ldots, \theta_{k,i+1})'$ folgendermaßen:

- Generiere $\theta_{1,i+1} \sim p(\theta_1|\theta_{2,i}, \ldots, \theta_{k,i}, y)$;
- generiere $\theta_{2,i+1} \sim p(\theta_2|\theta_{1,i+1}, \theta_{3,i}, \ldots, \theta_{k,i}, y)$;
  $\vdots$
- generiere $\theta_{k,i+1} \sim p(\theta_k|\theta_{1,i+1}, \theta_{2,i+1}, \ldots, \theta_{k-1,i+1}, y)$.

Schritt 2: Setze $i = i + 1$ und gehe zu Schritt 1.

Die Korrelation der Züge aus dem Gibbs Sampling kann reduziert werden, indem man nur einen Teil davon verwendet.
Manchmal wird auch Metropolis-Hastings Sampling in Verbindung mit dem Gibbs Sampler benötigt, nämlich wenn die Posterior für einen Parameter bedingt auf alle anderen keine einfache Form annimmt.

3.2.2 Data Augmentation

Eine ähnliche Idee wie die des Gibbs Samplings wird Data Augmentation[23] genannt. Es handelt sich dabei um eine Methode zur Konstruktion iterativer Algorithmen über die Einführung unbeobachteter Daten oder latenter Variablen.
Beispielsweise kann auch unbeobachtete Vergangenheit gemeinsam mit den Parametern unter Verwendung von Data Augmentation geschätzt werden.

4 Analyse der Schätzungen

Auch bei MCMC Analysen besteht das Risiko ernsthafte Fehler zu begehen:

- Ungeeignete Modellierung:
  Das vorausgesetzte Modell ist nicht realistisch oder eignet sich nicht für die verfügbaren Daten.
  Im Fall von Mischungsmodellen könnte man beispielsweise als diagnostisches Tool untersuchen, ob der Mittelwert der gezogenen bedingten Verteilungen der Konsumentenpräferenzen ähnlich der geschätzten Bevölkerungsverteilung ist. Trifft das zu, kann man sagen, dass das Modell korrekt spezifiziert und exakt geschätzt wurde. Sind die Verteilungen nicht ähnlich, kann es
  - sich um einen Spezifikationsfehler handeln,
  - an einer unzureichenden Anzahl an Zügen in der Simulation oder
  - an einer inadäquaten Stichprobengröße liegen, oder
  - die Folge der Zufallszüge konvergiert nicht gegen ein globales sondern nur gegen ein lokales Optimum.
- Fehler in der Berechnung oder Programmierung:
  Die stationäre Verteilung entspricht nicht der gewünschten Zielverteilung, oder die Folge konvergiert gegen eine improper Verteilung.
- Langsame Konvergenz:
  Die Simulation bleibt über viele Iterationen in einer Region, die stark von der Startverteilung beeinflusst wird.

[23] vgl. [25], S. 173-176

Bayesianische Prozeduren verwenden iterative Prozesse, welche mit einer hinreichend großen Anzahl an Iterationen konvergieren. Es ist allerdings schwierig zu bestimmen, ob Konvergenz tatsächlich schon erreicht wurde. Wenn eine durch den MCMC Algorithmus induzierte Markovkette es verabsäumt zu konvergieren, werden die resultierenden a posteriori Schätzungen verzerrt und unzuverlässig. Es kommt zu einer inkorrekten bayesianischen Datenanalyse und falschen Schlussfolgerungen.
Die Kombination verschiedener diagnostischer Werkzeuge kann bei der Untersuchung, wie schnell bzw. langsam eine Markovkette konvergiert und wie gut bzw. schlecht sie mischt, helfen. Für langsam mischende Markovketten sind Konvergenzdiagnostiken eher unzuverlässig, da ihre Schlussfolgerungen ausschließlich auf dem aus einer kleinen Region des Parameterraumes stammenden Output basieren. Daher ist es wichtig zu betonen, dass man sich nicht nur auf einzelne spezielle Prozeduren zur Konvergenzdiagnostik verlassen sollte.
Manche Konvergenzdiagnostiken eignen sich auch zur Bestimmung der Anzahl an Iterationen n. Hierfür kann man mehrere Ketten mit unterschiedlichen Startwerten parallel laufen lassen und die Schätzungen miteinander vergleichen. Entsprechen diese einander nicht hinreichend, muss n erhöht werden. Außerdem ist es sinnvoll, den Algorithmus mit unterschiedlichen Startpunkten zu replizieren und zu untersuchen, ob die jeweiligen Markovketten gegen den gleichen Punkt konvergieren.

4.1 Trace Plot

Ein einfaches aber effektives Werkzeug, um zu untersuchen, wann die Markovketten ihre Startpunkte „vergessen“ haben, ist der Trace Plot.
Es existieren hier zwei Varianten. Einerseits kann man eine einzelne langfristige Folge plotten, andererseits kann man aber auch mehrere kürzere Folgen mit gestreuten Startpunkten miteinander vergleichen.
Die Plots parallel simulierter Zeitreihen können übereinandergelegt werden, um zu untersuchen, ob sich die beiden Folgen unterscheiden. Annähernde Konvergenz wird diagnostiziert, wenn die Varianz zwischen den Folgen nicht größer ist als die Varianz innerhalb jeder individuellen Folge. Diese Vorgangsweise mag eine bessere Aussagekraft als die Betrachtung einer langen Folge haben, hängt aber stark von der Wahl der Startpunkte ab und leidet weiters unter dem hohen Anstieg der Anzahl „verschwendeter“ Burn-in Simulationen.

Eine langfristige einzelne Folge kann wiederum vorteilhaft in der Erforschung potentieller Kodierungsbugs und des Mischverhaltens der Markovkette sein.

4.2 CUSUM Plot

In der Praxis ist die Dimension des Parameterraumes für gewöhnlich groß, weshalb es nicht möglich ist, Trace Plots für jeden einzelnen Parameter zu untersuchen. In diesem Fall kann man Trace Plots für einige ausgewählte Parameter[24], einen sogenannten CUSUM (CUmulative SUM) Plot konstruieren.

Gegeben den Output $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}\}$ beginnt man damit, die ersten m Iterationen, die man zur Burn-in Periode gehörend vermutet, auszusondern.

CUSUM Plot:

Schritt 1: Berechne

$$\bar{\theta} = \frac{1}{(n-m)} \sum_{i=m+1}^{n} \theta^{(i)} .$$

Schritt 2: Berechne die CUSUM

$$S_t = \sum_{i=m+1}^{t} (\theta_i - \bar{\theta}) \quad \text{für } t = m+1, \ldots, n .$$

Schritt 3: Plotte s_t gegen t für $t = m+1, \ldots, n$ und verbinde aufeinanderfolgende Punkte durch Liniensegmente.

Die Glätte des resultierenden CUSUM Plots gibt Aufschluss über die Geschwindigkeit, mit der die Kette mischt. Ein glatter Plot steht für langsames Mischen, ein „haariger" Plot deutet auf eine schnelle Mischrate für θ hin.

4.3 Varianzverhältnismethode

Eine der populärsten quantitativen Konvergenzdiagnostiken stellt die Varianzverhältnismethode dar. Hierbei werden mehrere unabhängige Folgen analy-

[24] Die Auswahl sollte Parameter, die als langsam konvergierend bekannt sind, Funktionen von Parametern des Interesses, sowie Störparameter inkludieren.

siert, um eine verteilte Schätzung dafür zu erhalten, was über einige Zufallsvariablen bekannt ist gegeben die Beobachtungen, die bis dahin simuliert wurden.[25]

4.4 Autokorrelationsplot

Autokorrelationsplots sind das einfachste Werkzeug, um das Mischverhalten einer Markovkette quantitativ einzuschätzen. Man sollte sowohl die Autokorrelationen innerhalb als auch zwischen den Folgen untersuchen. Eine langsame Dämpfung in den Autokorrelationen deutet auf langsames Mischen hin. Eine einzelne langfristige Folge mag hier nützlicher im Vergleich zu mehreren kurzfristigen Folgen sein.
Ohne Aussonderung der zur Burn-in Periode gehörigen Iterationen kann es zu einer Unter- bzw. Überschätzung der Autokorrelationen kommen, was ein falsches Mischverhalten der Markovkette reflektiert.

4.5 Bayesianische Glaubwürdigkeits- und HPD-Intervalle

Glaubwürdigkeitsbereiche enthalten den unbekannten Parameter mit hoher Wahrscheinlichkeit. Man kann $100(1-\alpha)\%$ a posteriori Glaubwürdigkeitsintervalle für interessante Parameter tabellieren - entweder analytisch oder unter Verwendung von MCMC Methoden. Ist die marginale Verteilung nicht symmetrisch, ist ein $100(1-\alpha)\%$ Highest Probability Density (HPD) Intervall erstrebenswerter.

Für einen skalaren Parameter θ ist C_α ein $100(1-\alpha)\%$ Glaubwürdigkeitsintervall, wenn

$$\int_{C_\alpha} p(\theta|y)d\theta = 1-\alpha \ .$$

Das kürzeste solche Intervall wird Highest Probability Density (HPD) Intervall genannt und besitzt folgende Eigenschaften:

1. Die Dichte für jeden Punkt innerhalb des Intervalls ist größer als jene für jeden Punkt außerhalb.

[25] siehe [13], S. 61-62

2. Für eine gegebene Inhaltswahrscheinlichkeit[26] ist das Intervall jenes mit der kürzesten Länge.

Ein HPD-Intervall ist also ein spezielles Glaubwürdigkeitsintervall, nämlich das kürzeste unter allen möglichen Glaubwürdigkeitsintervallen mit gleichem Wahrscheinlichkeitsinhalt $(1-\alpha)$.

$p(\theta|y)$ bzw. $P(\theta|y)$ bezeichnen die marginale a posteriori Dichte bzw. die marginale kumulierte a posteriori Verteilungsfunktion von θ. Eines der gebräuchlichsten $100(1-\alpha)\%$ bayesianischen Glaubwürdigkeitsintervalle für θ lautet

$$(\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)}) ,$$

wobei

$$P(\theta^{(\alpha/2)}|y) = \frac{\alpha}{2} \quad \text{und} \quad P(\theta^{(1-\alpha/2)}|y) = 1 - \frac{\alpha}{2} .$$

Ist $p(\theta|y)$ symmetrisch und unimodal, ist dieses bayesianische Glaubwürdigkeitsintervall auch ein HPD-Intervall.
Ein $100(1-\alpha)\%$ HPD-Intervall für θ ist gegeben durch

$$R(p_\alpha) = \{\theta : p(\theta|y) \geq p_\alpha\}$$

mit p_α als größte Konstante, sodass $\mathbb{P}(\theta \in R(p_\alpha)) \geq 1-\alpha$.

Es ist möglich, einen MCMC Ansatz zu verwenden, um bayesianische Glaubwürdigkeits- und HPD-Intervalle zu approximieren.[27]

Weitere quantitative Methoden zur Konvergenzdiagnostik bilden die Spektraldichtediagnostik, die L^2-Konvergenzdiagnostiken, geometrische Konvergenzgrenzen und der Konvergenzratenschätzer.[28]

[26] üblicherweise $(1-\alpha)$
[27] Genaueres in [13], S. 216-221
[28] siehe [13], S. 62

5 Modellselektion

Modellbewertung und Modellvergleiche können mittels Bayes Faktoren, Informationskriterien und Goodness-of-Fit Maßen erfolgen.
Im Unterschied zu kriterienbasierten Methoden werden beim bayesianischen Ansatz zur Variablenselektion die a priori Unsicherheiten über Wahrscheinlichkeiten für jedes Modell quantifiziert. Dann wird eine a priori Verteilung für jeden Parameter spezifiziert und schließlich Bayes' Theorem verwendet, um a posteriori Modellwahrscheinlichkeiten zu berechnen.
Bayesianische Methoden für Modellvergleiche beruhen üblicherweise auf a posteriori Modellwahrscheinlichkeiten oder Bayes Faktoren. Hierfür werden proper Priors benötigt, wenn sich die Anzahl an Parametern in den beiden Modellen unterscheiden. Kriterienbasierte Methoden benötigen keine proper a priori Verteilungen, sind aber im Allgemeinen nicht leicht zu kalibrieren und/oder interpretieren.

Beim bayesianischen Ansatz zur Modellselektion werden unterschiedliche Modelle $\mathcal{M}_1, \ldots, \mathcal{M}_K$ über ihre a posteriori Wahrscheinlichkeiten

$$P(\mathcal{M}_k|y) \propto l(y|\mathcal{M}_k)P(\mathcal{M}_k)$$

mit $l(y|\mathcal{M}_k)$ als Modelllikelihood miteinander verglichen.
Zur Approximation der marginalen Likelihood (Randdichte der Daten) aus dem Output der Markovkette wird häufig die Methode von Gelfand und Dey ([23]) verwendet. Für das Modell mit K Komponenten wird die Menge aller Parameter mit θ_K bezeichnet. Die Randdichte der Daten gegeben K Komponenten lautet:

$$p_K(y) = \int_{\theta_K} p_K(y|\theta_K)p_K(\theta_K)d\theta_K = \left\{ \mathbb{E}\left[\frac{g_K(\theta_K)}{p_K(y|\theta_K)p_K(\theta_K)} \right] \right\}^{-1}$$

g_K ... willkürliche Dichte auf dem Support von θ_K.

Der Erwartungswert wird bezüglich der a posteriori Verteilung von θ_K berechnet. Die MCMC Approximation lautet:

$$\tilde{p}_K(y) = \left[\frac{1}{n-m} \sum_{i=m+1}^{n} \frac{g_K(\theta_K^{(i)})}{p_K(y|\theta_K^{(i)})p_K(\theta_K^{(i)})} \right]^{-1}$$

$\theta_K^{(i)}$ ... Wert von θ_K bei der i-ten Iteration der Markovkette.

Verwendet werden nur die letzten $n - m$ von n Iterationen. Falls g_K die a posteriori Dichte von θ_K ist, ist die Approximation exakt.
Man muss ein g_K spezifizieren, das komplett bekannt ist. Die geschätzten a posteriori Wahrscheinlichkeiten sind

$$\tilde{\mathbb{P}}(K|y) \propto p(K)\tilde{p}_K(y)$$

$p(K)$ ... a priori Wahrscheinlichkeit für K Mischungskomponenten.

5.1 Bayesianische Variablenwahl

Das Problem der Variablenselektion tritt vor allem dann auf, wenn eine große Anzahl k an Kovariaten zur Verfügung steht. In diesem Fall existieren 2^k mögliche Modelle.

5.1.1 Kriterienbasierte Methoden

Variablenwahl erfolgt meistens unter Verwendung kriterienbasierter Methoden. Informationskriterien messen die Eignung eines Modells, indem sie Modellanpassung und -komplexität ausbalancieren und sollten demnach minimiert werden. Dazu zählt mitunter das Akaike Informationskriterium (AIC[29]). Die Regularitätsbedingungen, auf denen das AIC beruht, gelten nicht, wenn das Likelihoodverhältnis (Likelihood Ratio[30]) λ so gestaltet ist, dass sich zwei Hypothesen bezüglich der Anzahl der Komponenten unterscheiden sollen. In diesem Fall verwendet man eine Approximation zur Nullverteilung von $-2\ln\lambda$, was zum modifizierten AIC führt (AIC3[31]).
Das Bayes Informationskriterium (BIC[32]) leitet sich von der a posteriori Modellwahrscheinlichkeit ab. Die a posteriori Wahrscheinlichkeit von Modell $\mathcal{M}$ bedingt auf die beobachteten Daten y lautet

$$p(\mathcal{M}|y) = \frac{p(y|\mathcal{M})p(\mathcal{M})}{p(y)}$$

[29] $AIC = -2\ln L + 2k$ (Akaike Information Criterion)
$\ln L$ ... maximierte Loglikelihood
k ... Anzahl der Parameter
[30] $LR = \frac{\max_{\theta\in\omega} L(\theta)}{\max_{\theta\in\Omega} L(\theta)}, \quad \omega \subseteq \Omega$
[31] $AIC3 = -2\ln L + 3k$
[32] $BIC = -2\ln L + \ln(n)k$ (Bayes Information Criterion)
n ... Anzahl der Beobachtungen

mit $p(\mathcal{M}|y)$ als a posteriori Wahrscheinlichkeit für Modell $\mathcal{M}$ gegeben die Daten y, $p(\mathcal{M})$ als Prior für Modell $\mathcal{M}$ und $p(y|\mathcal{M})$ als Likelihood. Da die Posterior auf y bedingt wird, ist $p(y)$ konstant, und obige Formel vereinfacht sich auf

$$p(\mathcal{M}|y) \propto p(y|\mathcal{M})p(\mathcal{M}) \ .$$

Es gilt

$$p(y|\mathcal{M}) = \int p(y|\mathcal{M},\theta)p(\theta|\mathcal{M})d\theta \ ,$$

was auch integrierte Likelihood genannt wird. Das BIC ist eine Approximation von $-2\log$(*integrierter Likelihood*). Eine Minimierung des BIC ist äquivalent zur Maximierung der integrierten Likelihood, was wiederum äquivalent zur Maximierung der a posteriori Modellwahrscheinlichkeit ist, wenn die Priors alle gleich sind.

Ein Leistungsvergleich diverser Kriterien

Ein Paper von Andrews und Currim ([6]) beschäftigt sich mit der Untersuchung der Leistungsfähigkeit sieben verschiedener Kriterien[33] via Simulation, die im Zusammenhang mit endlichen Mischungs-Regressionsmodellen verwendet wurden. Die Performance der verschiedenen Kriterien wurde anhand ihrer Erfolgsrate (oder des Prozentsatzes der Datensets, in denen die wahre Segmentzahl identifiziert wurde) und der Wurzel aus dem mittleren quadratischen Fehler zwischen den wahren und den geschätzten β-Parametern (RMSE(β)[34]) der Modelle, die aufgrund des Kriteriums ausgewählt wurden, gemessen. Ergaben zwei Kriterien ähnliche Erfolgsraten, wurde Underfitting einem Overfitting bevorzugt. Weiters wurden Kriterien, die Modelle mit einem geringeren RMSE(β)-Werten auswählten, präferiert. Außerdem wurden z-Tests eingesetzt, um auf statistisch signifikante Unterschiede zwischen den Gesamterfolgsraten zu testen.
In der Studie stellte sich das AIC3 (das Akaike Informationskriterium mit einem Strafterm von 3 pro Parameter[35]) über eine weite Spannbreite an Modellspezifikationen und multinomialen Datenkonfigurationen als am besten

[33] Insbesondere wurden das AIC, AIC3, BIC, CAIC, ICOMP (Informational Complexity Criterion), LOGLV (Validierungssample-Loglikelihood) und NEC (Normiertes Entropiekriterium) miteinander verglichen.

[34] RMSE ... Root Mean Square Error

[35] anstelle des traditionell üblichen Faktors 2

geeignet heraus. Es lieferte die höchste Erfolgsrate (72%), nur geringes Overfitting (6%) und den niedrigsten Parameterbias. Momentan wird dieses Kriterium allerdings noch sehr selten in der Marketingliteratur verwendet.

5.1.2 Reversible Jump MCMC Algorithmus

Reversible Jump Sampling kann dazu verwendet werden, um Mischrepräsentationen mit einer unbekannten und somit variierenden Anzahl an Komponenten zu ziehen. D.h. speziell bei Mischungsmodellen (siehe 7) kann die a priori unbekannte Anzahl an Segmenten mitgeschätzt und muss nicht erst im Anschluss über die Minimierung von Informationskriterien beispielsweise bestimmt werden. Der Reversible Jump Algorithmus ist attraktiv für bayesianische Variablenselektion bei einer großen Anzahl an Kovariaten. Er basiert auf der Konstruktion einer Markovkette, die zwischen Modellen mit Parameterräumen unterschiedlicher Dimension springen kann.
Angenommen man besitzt Kandidatenmodelle $\{\mathcal{M}_k, k \in \mathcal{M}\}$ mit $\theta^{(k)}$ als unbekannte Parameter mit Dimension p_k für Modell $\mathcal{M}_k$. Unter Modell $\mathcal{M}_k$ sieht die a posteriori Verteilung von $\theta^{(k)}$ folgendermaßen aus:

$$\begin{aligned} p(\theta^{(k)}|y, \mathcal{M}_k) &\propto p^*(\theta^{(k)}|y, \mathcal{M}_k) \\ &= l(\theta^{(k)}|y, \mathcal{M}_k) p(\theta^{(k)}|\mathcal{M}_k) \end{aligned}$$

mit $l(\theta^{(k)}|y, \mathcal{M}_k)$ als Likelihoodfunktion, y als Daten, $p(\theta^{(k)}|\mathcal{M}_k)$ als a priori Verteilung und $p^*(\theta^{(k)}|y, \mathcal{M}_k)$ als nichtnormalisierte a posteriori Dichte. Die gemeinsame Verteilung von $(k, \theta^{(k)})$ gegeben die Daten y lautet dann

$$p(k, \theta^{(k)}|y) \propto \pi(k) p^*(\theta^{(k)}|y, \mathcal{M}_k) \ .$$

Reversible Jump MCMC Algorithmus:

Schritt 0: Der momentane Zustand sei $(k, \theta^{(k)})$.

Schritt 1: Schlage ein neues Modell $\mathcal{M}_{k^*}$ mit Wahrscheinlichkeit $j(k^*|k)$ vor.

Schritt 2: Generiere u aus einer spezifizierten Vorschlagsdichte $q(u|\theta^{(k)}, k, k^*)$.

Schritt 3: Setze $(\theta^{*(k^*)}, u^*) = g_{k,k^*}(\theta^{(k)}, u)$, wobei g_{k,k^*} eine Bijektion zwischen $(\theta^{(k)}, u)$ und $(\theta^{*(k^*)}, u^*)$ definiert. Weiters müssen u und u' die Bedingung $\pi_k + dim(u) = \pi_{k^*} + dim(u^*)$ erfüllen.

Schritt 4: Akzeptiere die vorgeschlagene Bewegung zu $(k^*, \theta^{*(k^*)})$ mit Wahrscheinlichkeit

$$\alpha = \min\left\{\frac{\pi(k^*)p^*(\theta^{*(k^*)}|y, \mathcal{M}_{k^*})j(k|k^*)q(u^*|\theta^{*(k^*)}, k^*, k)}{\pi(k)p^*(\theta^{(k)}|y, \mathcal{M}_k)j(k^*|k)q(u|\theta^{(k)}, k, k^*)} \times \left|\frac{\partial g_{k,k^*}(\theta^{(k)}, u)}{\partial(\theta^{(k)}, u)}\right|, 1\right\}.$$

Ist $q(\cdot)$ die a posteriori Verteilung von $\theta^{*(k^*)}$, vereinfacht sich die Akzeptanzwahrscheinlichkeit auf

$$\alpha = \min\left\{\frac{\pi(k^*)\pi(y|\mathcal{M}_{k^*})j(k|k^*)}{\pi(k)\pi(y|\mathcal{M}_k)j(k^*|k)}, 1\right\}$$

mit $\pi(y|\mathcal{M}_k)$ als marginale Verteilung der Daten y unter Modell $\mathcal{M}_k$.

Wurde erst einmal eine MCMC Stichprobe $\{k_l, l = 1, 2, \ldots, L\}$ durch den Reversible Jump MCMC Algorithmus generiert, kann die a posteriori Modellwahrscheinlichkeit $\pi(k|y)$ über

$$\hat{\pi}(k|y) = \frac{1}{L}\sum_{l=1}^{L} \mathbf{1}_k(k_l)$$

geschätzt werden.

Eine allgemeinere Version des Reversible Jump MCMC Algorithmus stellt der Metropolized Carlin-Chib Algorithmus[36] dar.

[36] siehe [13], S. 303

5.1.3 Jump Diffusions

Bei sogenannten Jump Diffusions handelt es sich um eine Alternative zum zuvor präsentierten Reversible Jump Algorithmus. Auch hier geht es um den Vergleich von Modellen mit eventuell unterschiedlichen Dimensionen.
Die wesentlichen Eigenschaften dieses Ansatzes bestehen darin, dass diskrete Übergänge oder Sprünge zwischen Modellen unterschiedlicher Dimensionalität vorgenommen werden können und dass innerhalb eines Modells mit fixer Dimension die entsprechende bedingte Posterior simuliert wird.[37]

5.2 Vergleich nichtgeschachtelter Modelle

In der Praxis steht man oft vor dem Problem, mehrere Modelle, die nicht geschachtelt sind, miteinander vergleichen zu wollen.

5.2.1 Marginale Likelihoodansätze

Betrachtet werden $\mathcal{K}$ Modelle $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_\mathcal{K}$. Modell $\mathcal{M}_k$ besitzt die a posteriori Verteilung

$$p(\theta_k|y, \mathcal{M}_k) \propto l(\theta_k|y, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)$$

mit $l(\theta_k|y, \mathcal{M}_k)$ als Likelihoodfunktion, y als Daten und $p(\theta_k|\mathcal{M}_k)$ als a priori Verteilung. Dann ist die marginale Likelihood gegeben durch

$$m(y|\mathcal{M}_k) = \int l(\theta_k|y, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)d\theta_k \ .$$

Um verschiedene Modelle miteinander vergleichen zu können, berechnet man $m(y|\mathcal{M}_k)$ für $k = 1, 2, \ldots, \mathcal{K}$ und entscheidet sich für jenes mit der größten marginalen Likelihood.

Bei der marginalen Likelihood $m(y|\mathcal{M}_k)$ handelt es sich im Wesentlichen um die normalisierende Konstante $p(\theta_k|y, \mathcal{M}_k)$ der a posteriori Verteilung

$$m(y|\mathcal{M}_k) = \frac{l(\theta_k|y, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)}{p(\theta_k|y, \mathcal{M}_k)} \ .$$

[37] Details findet man in [25], S. 221-226.

Es sei θ_k^* der a posteriori Mittelwert oder das a posteriori Maximum der a posteriori Verteilung $p(\theta_k|y, \mathcal{M}_k)$. Dann gilt

$$m(y|\mathcal{M}_k) = \frac{l(\theta_k^*|y, \mathcal{M}_k)p(\theta_k^*|\mathcal{M}_k)}{p(\theta_k^*|y, \mathcal{M}_k)} .$$

Aus rechenbetonter Sicht mag es effizienter sein,

$$ln[m(y|\mathcal{M}_k)] = ln[l(\theta_k^*|y, \mathcal{M}_k)] + ln[p(\theta_k^*|\mathcal{M}_k)] - ln[p(\theta_k^*|y, \mathcal{M}_k)]$$

statt direkt $m(y|\mathcal{M}_k)$ zu berechnen. Die einzig schwierige Berechnung betrifft in diesem Fall $ln[p(\theta_k^*|y, \mathcal{M}_k)]$. Im Allgemeinen kann dieser Ausdruck z.B. mittels der Kernmethode[38] über MCMC Sampling geschätzt werden.

Die oben beschriebene Vorgangsweise der Modellselektion entspricht dem Bayes Faktor Ansatz

$$B_{ij} = e^{ln[m(y|\mathcal{M}_i)]-ln[m(y|\mathcal{M}_j)]}$$

mit B_{ij} als Bayes Faktor.

5.2.2 Bayes Faktoren

Beim Bayes Faktor, der im Wesentlichen zwei konkurrierende Modelle miteinander vergleicht, handelt es sich um das Verhältnis der marginalen Likelihoods unter diesen beiden Modellen. Die marginale Likelihood eines Modells kann man aus dem a posteriori Simulationsoutput schätzen (beispielsweise mittels Importance Sampling, siehe Abschnitt 3.1).
Sowohl die Null- als auch die Alternativ-Hypothese werden als parametrische Wahrscheinlichkeitsmodelle repräsentiert. Der Bayes Faktor B_{10} für ein Modell $\mathcal{M}_1$ gegen ein anderes Modell $\mathcal{M}_0$ gegeben Daten y

$$B_{10} = \frac{p(y|\mathcal{M}_1)}{p(y|\mathcal{M}_0)}$$

entspricht dem Verhältnis der marginalen Likelihoods

$$p(y|\mathcal{M}_k) = \int p(y|\theta_k, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)d\theta_k$$

[38] Details in [13], S. 97

mit θ_k als Parameter und $p(\theta_k|\mathcal{M}_k)$ als a priori Dichte.
Zweimal logarithmiert entspricht der Wert der Skala der Likelihood Ratio.

Bayes Faktor	Anzeichen für Modell $\mathcal{M}_1$
$B_{10} < 1$	negativ (Unterstützung von Modell $\mathcal{M}_0$)
$1 \leq B_{10} \leq 3$	kaum erwähnenswert
$3 \leq B_{10} \leq 12$	positiv (Unterstützung von Modell $\mathcal{M}_1$)
$12 \leq B_{10} \leq 150$	stark
$B_{10} > 150$	sehr stark

Tabelle 1: Bayes Faktoren (Quelle: [25], S. 165)

Die Berechnung der marginalen Likelihood stellt ein rechentechnisch gesehen herausforderndes Problem dar. Eine Methode zur Berechnung des Verhältnisses zweier marginaler Likelihoods ist das sogenannte Bridge Sampling.

Bridge Sampling

Beim Bridge Sampling werden die MCMC Simulationen mit Simulationen aus einer Wichtigkeitsdichte kombiniert. Es handelt sich dabei um eine Technik zur Berechnung von Verhältnissen zwischen normalisierenden Konstanten nichtstandardmäßiger Dichten.
Es seien $\pi_1(\theta) = c_1^{-1} q_1(\theta)$ mit $\theta \in \Omega_1$ und $\pi_2(\theta) = c_2^{-1} q_2(\theta)$ mit $\theta \in \Omega_2$ zwei Dichtefunktionen.[39] Bridge Sampling ist eine Monte Carlo Technik zur Schätzung des Verhältnisses der normalisierenden Konstanten $r = \frac{c_1}{c_2}$ basierend auf der Gleichung

$$r = \frac{\mathbb{E}_2\{q_1(\theta)\alpha(\theta)\}}{\mathbb{E}_1\{q_2(\theta)\alpha(\theta)\}} .$$

$\alpha(\theta)$ wird Bridge Funktion genannt und kann willkürlich auf $\Omega_1 \cap \Omega_2$ gewählt werden. Es muss lediglich

$$0 < \left| \int_{\Omega_1 \cap \Omega_2} \alpha(\theta) q_1(\theta) q_2(\theta) d\theta \right| < \infty$$

[39] Ω_1 bzw. Ω_2 bezeichnen den Support der nicht-normalisierten Dichten $q_1(\theta)$ bzw. $q_2(\theta)$.

erfüllt sein.
Beim Bridge Schätzer, der einen konsistenten Schätzer von r darstellt, werden die erwarteten Werte durch die Stichprobenmittel von $q_1(\theta)\alpha(\theta)$ und $q_2(\theta)\alpha(\theta)$ ersetzt:

$$\hat{r}_{BS}(\alpha) = \frac{\frac{1}{n_2}\sum_{i=1}^{n_2} q_1(\theta_{2,i})\alpha(\theta_{2,i})}{\frac{1}{n_1}\sum_{i=1}^{n_1} q_2(\theta_{1,i})\alpha(\theta_{1,i})}$$

mit $\{\theta_{l,1}, \theta_{l,2}, \ldots, \theta_{l,n_l}\}$ als Zufallsstichprobe aus π_l für $l = 1, 2$.

5.2.3 „Super-Modell"- oder „Sub-Modell"-Ansätze

Der in diesem Abschnitt vorgestellte Ansatz eignet sich ebenfalls für den Vergleich zweier nichtgeschachtelter Modelle $\mathcal{M}_1$ und $\mathcal{M}_2$.
Die a posteriori Verteilung für Modell $\mathcal{M}_1$ ist

$$p(\theta, \psi|y, \mathcal{M}_1) \propto l(\theta, \psi|y, \mathcal{M}_1)p(\theta, \psi|\mathcal{M}_1) \ ,$$

jene für Modell $\mathcal{M}_2$ lautet

$$p(\theta, \varphi|y, \mathcal{M}_2) \propto l(\theta, \varphi|y, \mathcal{M}_2)p(\theta, \varphi|\mathcal{M}_2) \ .$$

$p(\theta, \psi|\mathcal{M}_1)$ und $p(\theta, \varphi|\mathcal{M}_2)$ seien proper Priors, und $m(y|\mathcal{M}_k)$ bezeichne die marginale Likelihood. Dann ist der Bayes Faktor zwecks Vergleich von Modell $\mathcal{M}_1$ mit Modell $\mathcal{M}_2$ gegeben durch

$$B_{12} = \frac{m(y|\mathcal{M}_1)}{m(y|\mathcal{M}_2)} \ .$$

Es werden im Folgenden hinreichende Bedingungen für die Anwendung der „Super-Modell"- und „Sub-Modell"-Ansätze gegeben.
Angenommen es existieren ein reduziertes Modell $\mathcal{M}_r$ und ein gesättigtes Modell $\mathcal{M}_s$, sodass ihre a posteriori Verteilungen

$$p(\theta|y, \mathcal{M}_r) \propto l(\theta|y, \mathcal{M}_r)p(\theta|\mathcal{M}_r)$$

und

$$p(\theta, \psi, \varphi|y, \mathcal{M}_s) \propto l(\theta, \psi, \varphi|y, \mathcal{M}_s)p(\theta, \psi, \varphi|\mathcal{M}_s)$$

lauten. Das reduzierte Modell $\mathcal{M}_r$ bzw. das gesättigte Modell $\mathcal{M}_s$ sollen als

„Brücken“ dienen, um zwei nichtgeschachtelte Modelle miteinander zu verbinden.

Folgende Annahmen werden getroffen:

1. $l(\theta|y,\mathcal{M}_r) = l(\theta,\psi=0|y,\mathcal{M}_1) = l(\theta,\varphi=0|y,\mathcal{M}_2)$
2. $p(\theta|\mathcal{M}_r) \propto p(\theta,\psi=0|y,\mathcal{M}_1)$ und $p(\theta|\mathcal{M}_r) \propto p(\theta,\varphi=0|y,\mathcal{M}_2)$
3. $l(\theta,\psi,\varphi=0|y,\mathcal{M}_s) = l(\theta,\psi|y,\mathcal{M}_1)$ und $l(\theta,\psi=0,\varphi|y,\mathcal{M}_s) = l(\theta,\varphi|y,\mathcal{M}_2)$
4. $p(\theta,\psi|\mathcal{M}_1) \propto p(\theta,\psi,\varphi=0|y,\mathcal{M}_s)$ und $p(\theta,\varphi|\mathcal{M}_2) \propto p(\theta,\psi=0,\varphi|y,\mathcal{M}_s)$.

Diese Forderungen implizieren, dass $\mathcal{M}_r$ in $\mathcal{M}_1$ und $\mathcal{M}_2$ geschachtelt ist, während umgekehrt $\mathcal{M}_1$ und $\mathcal{M}_2$ in $\mathcal{M}_s$ geschachtelt sind.
Es seien $m(y|\mathcal{M}_r)$ bzw. $m(y|\mathcal{M}_s)$ die marginalen Likelihoods von $\mathcal{M}_r$ bzw. $\mathcal{M}_s$.

Theorem: Falls die Bedingungen 1 und 2 erfüllt sind, gilt

$$\begin{aligned} B_{12} &= \frac{m(y|\mathcal{M}_1)/m(y|\mathcal{M}_r)}{m(y|\mathcal{M}_2)/m(y|\mathcal{M}_r)} \\ &= \frac{p(\psi=0|y,\mathcal{M}_1)/p(\psi=0|\mathcal{M}_1)}{p(\varphi=0|y,\mathcal{M}_2)/p(\varphi=0|\mathcal{M}_2)} . \end{aligned}$$

Die Bedingungen 3 und 4 führen zu

$$\begin{aligned} B_{12} &= \frac{m(y|\mathcal{M}_1)/m(y|\mathcal{M}_s)}{m(y|\mathcal{M}_2)/m(y|\mathcal{M}_s)} \\ &= \frac{p(\varphi=0|y,\mathcal{M}_s)/p(\varphi=0|\mathcal{M}_s)}{p(\psi=0|y,\mathcal{M}_s)/p(\psi=0|\mathcal{M}_s)} . \end{aligned}$$

Der „Sub-Modell"-Ansatz:

Schritt 1: Generiere vier MCMC Stichproben aus den a posteriori und a priori Verteilungen $p(\theta, \psi|y, \mathcal{M}_1)$, $p(\theta, \psi|\mathcal{M}_1)$, $p(\theta, \varphi|y, \mathcal{M}_2)$ und $p(\theta, \varphi|\mathcal{M}_2)$.

Schritt 2: Verwende z.B. die Kernmethode, um Schätzungen von $p(\psi = 0|y, \mathcal{M}_1)$, $p(\psi = 0|\mathcal{M}_1)$, $p(\varphi = 0|y, \mathcal{M}_2)$ und $p(\varphi = 0|\mathcal{M}_2)$ zu erhalten.

Schritt 3: Berechne B_{12} mithilfe der ersten Formel aus obigem Theorem.

Der „Super-Modell"-Ansatz:

Schritt 1: Generiere zwei MCMC Stichproben aus den a posteriori und a priori Verteilungen $p(\theta, \psi, \varphi|y, \mathcal{M}_s)$ und $p(\theta, \psi, \varphi|\mathcal{M}_s)$.

Schritt 2: Verwende beispielsweise die Kernmethode, um Schätzungen von $p(\psi = 0|y, \mathcal{M}_s)$, $p(\psi = 0|\mathcal{M}_s)$, $p(\varphi = 0|y, \mathcal{M}_s)$ und $p(\varphi = 0|\mathcal{M}_s)$ zu erhalten.

Schritt 3: Berechne B_{12} mithilfe der zweiten Formel aus obigem Theorem.

5.2.4 Bayesianische Modellmittelung

Bei der bayesianischen Modellmittelung (BMA) wird Inferenz auf einen Durchschnitt aller möglichen Modelle im Modellraum $\mathcal{M}$ basiert. Die Motivation dahinter ist die Tatsache, dass ein einzelnes „bestes" Modell die Unsicherheit bezüglich dem Modell selbst ignoriert.
Es sei $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$. Das bayesianische Modellmittel besteht aus dem Durchschnitt der a posteriori Verteilungen unter jedem Modell gewichtet mit den entsprechenden a posteriori Modellwahrscheinlichkeiten:

$$p(\theta|y) = \sum_{k=1}^{K} p(\theta|y, \mathcal{M}_k)\pi(\mathcal{M}_k|y)$$

mit y als Daten, $p(\theta|y, \mathcal{M}_k)$ als a posteriori Verteilung von θ unter dem Modell $\mathcal{M}_k$ und $\pi(\mathcal{M}_k|y)$ als a posteriori Modellwahrscheinlichkeit.

6 Relation von MCMC zu graphischen Modellen

Bei der Definition eines Modells, das man mittels MCMC Methoden schätzen will, sollte man folgende Punkte der Reihe nach befolgen:

1. Spezifikation der Modellvariablen und ihrer qualitativen bedingten Unabhängigkeitsstrukturen,
2. Spezifikation der parametrischen Form der direkten Beziehungen zwischen diesen Größen und
3. Spezifikation der a priori Verteilungen für die Parameter.

6.1 Wahrscheinlichkeitsmodellierung

Ein DAG[40] Modell (siehe z.B. [37]) ist äquivalent zur Annahme, dass die gemeinsame Verteilung aller Zufallsgrößen komplett spezifiziert ist in Ausdrücken der bedingten Verteilung jedes Knoten v gegeben seine Eltern:

$$p(V) = \prod_{v \in V} p(v|Eltern[v])$$

mit $p(\cdot)$ als Wahrscheinlichkeitsverteilung und V als Menge aller Knoten im DAG.
Um die Spezifikation des Wahrscheinlichkeitsmodells zu komplettieren, benötigt man a priori Verteilungen für jene Knoten ohne Eltern. Man kann auf diese Art und Weise auch externe Information in Form informativer a priori Verteilungen berücksichtigen.

6.2 Modellanpassung mittels Gibbs Sampling

Im Allgemeinen sind vier Schritte zur Implementierung des Gibbs Samplers erforderlich:

1. Für alle unbeobachteten Knoten müssen Startwerte bereitgestellt werden.

[40] DAG ... Directed Acyclic Graph (gerichteter azyklischer Graph)

2. Für jeden unbeobachteten Knoten müssen voll bedingte Verteilungen konstruiert werden. Des weiteren muss entschieden werden, welche Methoden man verwendet, um aus diesen zu ziehen.

3. Der Output muss überwacht werden, um über die Länge des Burn-in sowie die Gesamtlauflänge zu entscheiden.

4. Für interessante Größen müssen zwecks Inferenz der wahren Werte der unbeobachteten Knoten zusammenfassende Statistiken aus dem Output berechnet werden.

6.3 Ziehen aus den voll bedingten Verteilungen

Gibbs Sampling arbeitet über iteratives Ziehen von Stichproben aus den voll bedingten Verteilungen der unbeobachteten Knoten im Graphen. Diese ist die Verteilung des Knoten gegeben die momentanen oder bekannten Werte für alle anderen Knoten im Graphen.
Für jeden Knoten v werden die verbleibenden Knoten mit V_{-v} bezeichnet, dann gilt

$$\begin{aligned} P(v|V_{-v}) &\propto P(v, V_{-v}) \\ &\propto \textit{Ausdrücke in } P(V), \textit{ die } v \textit{ beinhalten} \\ &= P(v|Eltern[v]) \times \prod_{w \in Kinder[v]} P(w|Eltern[w]) \,. \end{aligned}$$

Die voll bedingte Verteilung für v besteht aus einer a priori Komponente $P(v|Eltern[v])$ und Likelihoodkomponenten, welche aus jedem Kind von v hervorgehen. Die Proportionalitätskonstante, die sicherstellt, dass sich die Verteilung auf 1 integriert, wird im Allgemeinen eine Funktion der verbleibenden Knoten V_{-v} sein.

7 Das Problem unbeobachteter Heterogenität

In diesem Abschnitt wird ein Literaturüberblick über die Verwendung von Mischungs- bzw. hierarchischen Bayesmodellen gegeben. Die präsentierten Beiträge diverser Autoren beschäftigen sich speziell mit Konsumentenmärkten und dem in diesem Zusammenhang häufig auftretenden Phänomen der

unbeobachteten Heterogenität.
Die Frage nach der Erklärung unbeobachteter Heterogenität ist eines der Hauptprobleme, das bei der Analyse von Konsumentendaten auftritt. Man versteht darunter die Heterogenität der Konsumenten, die in empirischen Choicedaten unbeobachtet bleibt. Konsumenten können sich in ihren Wahrnehmungen und Bewertungen der unbeobachteten Eigenschaften der am Markt vorherrschenden Marken unterscheiden und überdies verschiedene Messverlässlichkeiten aufweisen. Sie können weiters heterogen auf Marketingvariablen reagieren. Ignoriert man diese Heterogenität völlig, kann dies zu irreführenden Schlussfolgerungen wie beispielsweise überhöhten Schätzungen der Messverlässlichkeit führen.
Unter Marktstruktur versteht man die Erklärung von Markenpräferenzen ausgedrückt über Präferenzen der Konsumenten für einzelne Markeneigenschaften. Interne Marktstrukturanalyse folgert sowohl die Markenattribute als auch die Konsumentenpräferenzen aus Präferenz- oder Choicedaten. Die Methoden für die interne Analyse der Marktstruktur unterscheiden sich in den erforderlichen Datentypen. Alle diese Modelle erlauben es, dass Markenattribute entweder a priori bekannt sind oder durch das Modell gefolgert werden. Es existieren Modelle für Marktanteilsdaten, welche Änderungen in den Marktanteilen über die Zeit analysieren. Diese Änderungen werden als durch Schwankungen in einer Marketing Mix Variable (wie dem Preis etwa) verursacht angenommen. Weiters gibt es z.B. auch Modelle für binäre Daten. Eine Eins kennzeichnet hierbei die Wahl einer Marke, während eine Null für die Ablehnung dieser steht. Für derartige Modelle existieren zwei Methoden zur Behandlung von Konsumentenheterogenität. Idealpunkte oder -vektoren können für jeden Haushalt einzeln geschätzt werden. Andererseits kann man die Konsumenten in Gruppen clustern und Segment-Idealpunkte betrachten. Neben endlichen Mischungsmodellen, die eine simultane Segmentierung der Daten und Schätzung der Parameter ermöglichen, existiert als Alternative eine sequentielle Analyse, welche zuerst Gruppen mit Hilfe eines Clusteralgorithmus formt und dann multigruppale Modelle definiert. Allerdings sind letztere in vielen Situationen weder effizient noch wirklich zufriedenstellend. Die Mischungsmethode sollte verwendet werden, wenn eine a priori Segmentierung nicht möglich ist und die Theorie vorschlägt, dass die Daten heterogen sind und zu einer endlichen Anzahl an unbeobachteten Gruppen gehören. Man sollte es allerdings vermeiden, ein endliches Mischungsmodell anzupassen, das nicht ausreichend durch die Theorie fundiert ist. Andernfalls kann es nämlich passieren, dass man solange Gruppen hinzufügt, bis eine vernünftige

Anpassung gefunden wurde. Eine gute Anpassung (Fit) für das Mischungsmodell muss jedoch nicht unbedingt bedeuten, dass die vorausgesetzte kausale Struktur korrekt gewählt wurde.

Im Anschluss werden nun verschiedene Ansätze zur Berücksichtigung unbeobachteter Heterogenität in der Bevölkerung vorgestellt.

7.1 Ein Entscheidungsmodell für Marktsegmentierung

[41] Es werden die üblichen Annahmen der Zufallsnutzentheorie getroffen. Stehen die Konsumenten vor einer Kaufentscheidung, wird jeder Marke ein Zufallsnutzen zugeordnet und jene gewählt, die den höchsten Nutzen bringt. Dieser Nutzen setzt sich zusammen aus einer deterministischen Komponente, die von den spezifischen Charakteristiken der Marke und anderen Marketing Mix Variablen abhängt, und einer Zufallskomponente:

$$U_{jkt} = u_{jk} + \beta_k X_{jkt} + \epsilon_{jkt}$$

U_{jkt}	...	Zufallsnutzen der Marke j für Konsument k zur Zeit t
u_{jk}	...	spezifischer Nutzen der Marke j für Konsument k
β_k	...	Parameter der Marketing Mix Variablen für Konsument k
X_{jkt}	...	Marketing Mix Variablen der Marke j für Konsument k zur Zeit t
ϵ_{jkt}	...	Zufallsfehler.

Es wird angenommen, dass die ϵ_{jkt} unabhängig identisch Weibull verteilt[42] sind. Somit ergibt sich die subjektspezifische Wahrscheinlichkeit, dass Marke j zum Zeitpunkt t von Konsument k gewählt wird, als multinomiales Logit Modell:

$$P_{jkt} = \frac{e^{u_{jk}+\beta_k X_{jkt}}}{\sum_{j'} e^{u_{j'k}+\beta_k X_{j'kt}}}$$

mit subjektspezifischen Parametern β_k.
Es wird angenommen, dass M homogene Segmente mit relativen Größen $f_i = \frac{e^{\lambda_i}}{\sum_{i'} e^{\lambda_{i'}}}$ existieren. Die optimale Segmentanzahl M wird in der Praxis mithilfe einer Variante des AIC bestimmt.

[41] [31]: „A Probabilistic Choice Model for Market Segmentation and Elasticity Structure"

[42] Weibullverteilung: $f(\epsilon) = \left(\frac{\eta}{\tau}\right)\left(\frac{\epsilon}{\tau}\right)^{\eta-1} e^{-\left(\frac{\epsilon}{\tau}\right)^{\eta}}$ für $\epsilon > 0$ bzw. 0 sonst
$\tau > 0, \eta > 0$ ($\eta = 1$: Exponentialverteilung)

Angenommen Konsument k gehört zu Segment i. Dann lautet die auf die Segmentzugehörigkeit bedingte (also die segmentspezifische) Wahrscheinlichkeit Marke j zu wählen

$$P_{jit} = \frac{e^{u_{ji}+\beta_i X_{jkt}}}{\sum_{j'} e^{u_{j'i}+\beta_i X_{j'kt}}}$$

mit segmentspezifischen Parametern β_i.
Da f_i die Likelihood einen Konsumenten in Segment i zu finden repräsentiert, kann die unbedingte Wahrscheinlichkeit der Wahl von Marke j durch Konsument k berechnet werden als

$$P_{jkt} = \sum_i f_i P_{jit} \ .$$

Es wird angenommen, dass die unbedingte Entscheidungswahrscheinlichkeit in einen gewichteten Durchschnitt zugrundeliegender („latenter") Entscheidungswahrscheinlichkeiten zerlegt werden kann. D.h. die nicht auf die Klassenzugehörigkeit bedingte Choicewahrscheinlichkeit lässt sich als Summe der mit den jeweiligen relativen Segmentgrößen f_i gewichteten segmentspezifischen Wahrscheinlichkeiten anschreiben.
Man bezeichne die Entscheidungshistorie eines Konsumenten k während eines Zeitintervalls T mit $H_k = c(t)$ mit $c(t)$ als Index der gewählten Marke zur Kaufgelegenheit t. Die nicht auf die Gruppenzugehörigkeit bedingte Likelihood dieser Entscheidungshistorie kann ebenfalls als Durchschnitt der bedingten Likelihoods gewichtet mit den entsprechenden relativen Segmentgrößen berechnet werden:

$$L(H_k) = \sum_i \frac{e^{\lambda_i} L(H_k|i)}{\sum_{i'} e^{\lambda_{i'}}}$$

mit

$$L(H_k|i) = \prod_t P_{c(t)}(\mathbf{u}_i, \beta_i, \mathbf{X}_{kt})$$

als bedingte Likelihood für die Entscheidungshistorie H_k unter der Annahme, dass Konsument k zu Segment i gehört. Obiger Ausdruck für die bedingte Likelihood setzt die Unabhängigkeit zwischen den Entscheidungen eines Konsumenten über alle Gelegenheiten voraus.

Mithilfe der bedingten Likelihood kann auch die a posteriori Wahrscheinlichkeit der Zugehörigkeit eines bestimmten Konsumenten k zu einem speziellen Segment i bedingt auf die beobachtete Entscheidungshistorie über

$$\mathbb{P}(k \in i|H_k) = \frac{L(H_k|i)f_i}{\sum_{i'} L(H_k|i')f_{i'}}$$

ermittelt werden.

7.2 Modellierung von Präferenz- und Strukturheterogenität

[43] In einer Erweiterung des clusterweisen Logit Modells von Kamakura und Russell (siehe vorigen Abschnitt) können sich die Konsumenten neben Differenzen in den Präferenzen und Reaktionen auf Marketing (Präferenzheterogenität) auch in ihren Entscheidungsprozessen unterscheiden (Strukturheterogenität). Segmente werden simultan auf Basis der Präferenzen, der Reaktion der Konsumenten auf Marketing und ihres Entscheidungsprozesses identifiziert. Das Modell besteht aus einer endlichen Mischung von geschachtelten Logit Modellen und beinhaltet die Mischung multinomialer Logits als Spezialfall.
Bei einem zweistufigen Ansatz wird jeder Konsument in der ersten Stufe aufgrund heuristischer Kriterien deterministisch einer der geschachtelten Logit Strukturen zugeordnet. Der zweite Schritt umfasst die Schätzung eines geschachtelten Logit Modells für jede Gruppe von Konsumenten, wobei die Daten aller Segmentmitglieder gepoolt werden. Danach werden die Konsumenten jener Gruppe zugeordnet, die am besten zu ihren beobachteten Entscheidungen passt. Im Anschluss werden die geschachtelten Logit Modelle neu geschätzt. Diese Vorgangsweise wird so lange wiederholt bis sich die Zuordnung der Konsumenten zu jedem strukturellen Segment zwischen den einzelnen Stufen nicht mehr ändert.
Im Gegensatz zu einer zweistufigen Heuristik erfolgen hier Identifizierung der Konsumentensegmente und Parameterschätzung simultan.

Zu jeder Gelegenheit t trifft Haushalt i eine Entscheidung für eine Marke j, welche in k verschiedenen Produktformen verfügbar ist. Es wird angenommen, dass S Segmente existieren. Der indirekte Nutzen bezüglich einer

[43] [30]: „Modeling Preference and Structural Heterogeneity in Consumer Choice“

Wahlalternative zu einem gegebenen Zeitpunkt unter der Annahme, dass Konsument i zu Segment s gehört, lautet

$$U^i_{jkt} = \gamma^s X^i_{jkt} + \epsilon^i_{jkt}$$

X^i_{jkt} ... exogene Variablen assoziiert mit Marke j in Produktform k zum Zeitpunkt t für Konsument i
γ^s ... Reaktionsparameter und alternativenspezifische Interzepte für Segment s
ϵ^i_{jkt} ... extremwertverteilter Zufallsfehler.

Die exogenen Variablen können Marketing Mix Faktoren wie Preis, Display oder Werbung beinhalten. Diskrepanzen zwischen dem vom Modell gelieferten höchsten Nutzen und der Marke, die tatsächlich vom Konsumenten gewählt wird, werden durch eine Zufallskomponente erklärt, die die unbeobachteten Determinanten des Nutzens umfasst, wodurch U_{jkt} ebenfalls zu einer Zufallsvariable wird.
Die Strukturheterogenität fließt über die Entscheidungsprozesse der Konsumenten ein. Es werden mehrere Segmente mit jeweils eigenen Nutzenfunktionen zugelassen.

1. Bei der ersten Variante eines hierarchischen Entscheidungsprozesses (Segment b) wählen die Konsumenten zuerst die Marke j und danach die Produktform k, was zu einem multinomialen geschachtelten Logit Modell mit Choicewahrscheinlichkeit

$$P^b_t(j,k) = P^b_t(k|j) P^b_t(j)$$

für Marke j in Produktform k führt.

$$P^b_t(k|j) = \frac{e^{\gamma^b X^i_{jkt}}}{\sum_{k'} e^{\gamma^b X^i_{jk't}}} \qquad \text{bzw.} \qquad P^b_t(j) = \frac{e^{\lambda^b V^b_j}}{\sum_{j'} e^{\lambda^b V^b_{j'}}}$$

$V^b_j = \ln \sum_k e^{\gamma^b X^i_{jkt}}$ ... eingeschlossener Wert für Marke j
λ^b, γ^b ... zu schätzende Parameter

definieren die multinomiale Choicewahrscheinlichkeit für eine Produktform k unter der Bedingung, dass Marke j gewählt wird bzw. die Choicewahrscheinlichkeit für die Wahl der Marke j.

Das geschätzte Modell erfüllt nur für $0 < \lambda^b \leq 1$ die Annahme stochastischer Zufallsnutzenmaximierung. Für $\lambda^b = 1$ reduziert sich das geschachtelte auf das multinomiale Logit Modell.

2. Manche Konsumenten entscheiden umgekehrt zuerst über die Produktform und danach erst über die Marke (Segment f). Die Choicewahrscheinlichkeiten lauten dann

$$P_t^f(j,k) = \left(\frac{e^{\gamma^f X^i_{jkt}}}{\sum_{j'} e^{\gamma^f X^i_{j'kt}}}\right)\left(\frac{e^{\theta^f V^f_k}}{\sum_{k'} e^{\theta^f V^f_{k'}}}\right) .$$

Für $\theta^f = 1$ erhält man wiederum ein multinomiales Logit Modell.

3. Weiters können Segmente existieren, deren Mitglieder immer die gleiche Alternative wählen und überhaupt nicht auf Marketing Mix Variablen reagieren.

Jeder Konsument besitzt eine Wahrscheinlichkeit π_b, π_f oder φ entweder zu den B Switching-Segmenten des Typs (1), zu den F Switching-Segmenten des Typs (2) oder zum loyalen Segment (3) zu gehören. Konsumenten innerhalb eines Segmentes werden als homogen und ihre Präferenzen und Entscheidungsprozesse als zeitinvariant angenommen.
Für die Schätzung von geschachtelten Logits benötigt man multiple Choicedaten von jedem Konsumenten, wie sie typischerweise von Scannerpanels bereitgestellt werden.
Die Likelihood der beobachteten vergangenen Entscheidungen von Konsument i bedingt auf die Zugehörigkeit zum markentypischen Segment b ist gegeben durch

$$L_{ib} = \prod_t \prod_j \prod_k P_t^b(j,k)^{\Upsilon(i,j,k,t)} .$$

$\Upsilon(i,j,k,t) = 1$ falls Haushalt i Marke j der Produktform k zum Zeitpunkt t kauft und 0 sonst.
Die bedingte Likelihood für das produktformtypische Segment f lautet

$$L_{if} = \prod_t \prod_j \prod_k P_t^f(j,k)^{\Upsilon(i,j,k,t)} .$$

Die bedingte Likelihood für beide Segmente b und f ergibt sich daraus als

$$L_{iM}{}^{44} = \sum_b \pi_b L_{ib} + \sum_f \pi_f L_{if}$$

und die unbedingte Likelihood als

$$L_i = L_{iM}(1 - \varphi) + \varphi D_i$$

mit $D_i = 1$, falls Konsument i im Verlauf der Samplingperiode immer die gleiche Marke kauft und 0, wenn Konsument i zwischen den Marken wechselt.

Die a posteriori Zugehörigkeitswahrscheinlichkeiten zu den Segmenten für jeden Haushalt i erhält man über

$$\tau_{is} = \frac{\pi_s L_{is}}{L_i} .$$

Aufgrund des wohlbekannten Problems lokaler Optima in der Schätzung endlicher Mischungen folgt man der üblichen Praxis, das Modell mit unterschiedlichen Startwerten zu schätzen.
Die Bestimmung der Anzahl der Konsumetensegmente erfolgt auf Basis mehrerer Kriterien wie dem Bayes (BIC) oder dem konsistenten Akaike Informationskriterium (CAIC[45]). Statt einer vollen Permutation kann man eine sequentielle Suche durchführen, indem man immer nur ein weiteres Segment hinzufügt, dann basierend auf den Informationskriterien das beste Modell wählt und mit diesem fortfährt.

7.3 Das diskrete heterogene Logit Modell

[46] In dieser Studie wird ein Logit Modell formuliert, das eine diskrete Verteilung zwecks Repräsentation der Heterogenität in den spezifischen Präferenzen verwendet.

---

[44] M ... Mover

[45] $CAIC = -2 \ln L + (\ln n + 1)k$ (Consistent Akaike Information Criterion)
Das CAIC wurde von Bozdogan (1987, [11]) entwickelt.

[46] [14]: „Heterogeneous Logit Model Implications for Brand Positioning"

Haushalt i besitzt den indirekten Nutzen U_{ijt} für Marke j, wenn er diese bei Kaufgelegenheit t wählt. Entsprechend dem Konzept des Zufallsnutzens gilt

$$U_{ijt} = \nu_{ij} + X_{ijt}\beta_i + \epsilon_{ijt}$$

ν_{ij} ... zeitinvariante Präferenz für Marke j
X_{ijt} ... Marketingvariablen
β_i ... Effekte der Marketingvariablen
ϵ_{ijt} ... Fehlerterm.

Angenommen ϵ_{ijt} sei Typ I extremwertverteilt[47], dann ergibt sich folgende Wahrscheinlichkeit, dass Haushalt i Marke j bei Gelegenheit t wählt bedingt auf $\{\nu_{ij}, \beta_i\}$ und die Werte der Kovariate:

$$P_{ijt} = \frac{e^{\nu_{ij}+X_{ijt}\beta_i}}{\sum_{l=1}^{J} e^{\nu_{il}+X_{ilt}\beta_i}} .$$

Der Vektor der Präferenzen $v_i = [\nu_{i1}, \ldots, \nu_{iJ}]'$ sei eine lineare Funktion der zeitinvarianten Attribute der Marken: $v_i = Aw_i$ mit A als Matrix der Positionen der J Marken im M-dimensionalen Raum und w_i als Wichtigkeitsgewichte für diese Dimensionen. Daraus ergibt sich

$$P_{ijt} = \frac{e^{a_j w_i+X_{ijt}\beta_i}}{\sum_{l=1}^{J} e^{a_l w_i+X_{ilt}\beta_i}} ,$$

wobei a_j für die j-te Zeile der Matrix A steht. Die Positionen der Marken sind invariant, während die Gewichte w_i über die Haushalte variieren können, wodurch die Markenpräferenzen ebenfalls über die Haushalte variieren.
Die Likelihoodfunktion für Haushalt i bedingt auf w_i und β_i, wenn T_i Käufe getätigt werden, ist

$$L_{i|\beta_i,w_i} = \prod_{t=1}^{T_i} \left\{ \prod_{j=1}^{J} P_{ijt}^{\delta_{ijt}} \right\}$$

mit $\delta_{ijt} = 1$, falls Haushalt i Marke j bei Gelegenheit t kauft und $\delta_{ijt} = 0$ sonst. Diese Formulierung setzt voraus, dass $\theta_i = \{w_i, \beta_i\}$ über die Haushalte

[47] Typ I Extremwertverteilung: $f(\epsilon) = \frac{1}{\alpha} e^{\frac{\epsilon-\mu}{\alpha}} e^{-e^{\frac{\epsilon-\mu}{\alpha}}}, \qquad \alpha > 0$

variiert. Man kann also annehmen, dass θ_i die Realisiation der Zufallsvariable θ mit variater Verteilung $G(\theta)$ ist. Somit lautet die unbedingte Likelihood

$$L_i = \int_\theta \left[\prod_{t=1}^{T_i} \left\{ \prod_{j=1}^{J} P_{ijt}^{\delta_{ijt}} \right\} \right] dG(\theta) \ .$$

Das heterogene Logit Modell (insbesondere das Zufallseffektmodell mit diskreter Heterogenitätsverteilung) approximiert $G(\theta)$ durch eine diskrete Verteilung mit einer endlichen Anzahl an Stützen S und ihren assoziierten Wahrscheinlichkeiten $\rho(\theta_s)$, sodass $\sum_{s=1}^{S} \rho(\theta_s) = 1$.
Die Likelihood von Haushalt i lautet somit

$$L_i = \sum_{s=1}^{S} \left[\prod_{t=1}^{T_i} \left\{ \prod_{j=1}^{J} P_{ijt}^{\delta_{ijt}} \right\} \right] \rho(\theta_s) \ .$$

Die Positionen $\theta_s = \{w_s, \beta_s\}$ und Wahrscheinlichkeiten $\rho(\theta_s)$ werden durch Maximieren des Logarithmus der Likelihoodfunktion $L = \prod_{i=1}^{N} L_i$ aus Stichprobendaten geschätzt.

7.4 Ein Zufallseffekt-Logit Modell

[48] Manchmal kann es aufgrund des Mangels an Heterogenität in den Reaktionen auf Marketing Mix Variablen genügen, sich nur auf die Berücksichtigung von Präferenzheterogenität zu beschränken. Wenn man keinerlei a priori Information bezüglich der Natur der Heterogenität über Haushalte hat, ist es allerdings wichtig, sowohl Heterogenität in den Präferenzen als auch in den Reaktionen zu berücksichtigen.
Man kann einerseits endliche Mischungen von Logit Modellen oder ein Zufallseffektmodell formulieren. In letzterem wird angenommen, dass die Zufallsfehlerkomponente des Logit Modells unabhängig von der Verteilung der unbeobachteten Heterogenität ist. Es hat den Vorteil sparsamer im Vergleich zur Schätzung haushaltspezifischer Parameter zu sein, was die Anzahl der zu schätzenden Parameter betrifft.
Bei der Verwendung einer multivariaten Wahrscheinlichkeitsverteilung, um unbeobachtete Heterogenität zu repräsentieren, muss man zwei kritische Punkte beachten:

[48] [28]: „A Random-Coefficients Logit Brand-Choice Model Applied to Panel Data“

1. Der erste Punkt betrifft die Spezifizierung der zugrundeliegenden Verteilung der Heterogenität. In der Praxis kennt man a priori fast nie ihre wahre Form.

2. Das Schätzen der Parameter eines Logit Modells unter derartigen Spezifikationen kann äußerst schwierig sein, weil man multiple Integrale auswerten muss. Da diese nicht in geschlossener Form dargestellt werden können, muss man sich zu diesem Zweck numerischer Methoden bedienen.

Konkret wird hier mit folgender Logit Formulierung gearbeitet:

$$P_{it}(j) = \frac{e^{\beta_{0ij}+\sum_{k=1}^{K}\beta_{ik}x_{ijkt}}}{\sum_{l=1}^{N} e^{\beta_{0il}+\sum_{k=1}^{K}\beta_{ik}x_{ilkt}}}$$

$P_{it}(j)$	...	Wahrscheinlichkeit, dass Haushalt i Marke j zur Zeit t wählt
X_{ijkt}	...	Wert von Kovariat k
β_{0ij}	...	haushaltspezifischer Interzeptterm
β_{ik}	...	haushaltspezifischer Reaktionskoeffizient auf Kovariat k.

Die haushaltspezifischen Koeffizienten folgen einer multivariaten Wahrscheinlichkeitsverteilung $G(\Theta)$. Fehlt jegliche a priori Information hinsichtlich der parametrischen Form von $G(\Theta)$, wird die Verteilung alternativ durch eine endliche Anzahl S von Stützvektoren approximiert und die Positionen dieser Vektoren sowie die Wahrscheinlichkeitsmassen $\pi(\Theta_s)$ geschätzt. Es wird somit jede Verteilung der unbeobachteten Heterogenität zugelassen ohne sie auf eine spezielle parametrische Form einzuschränken.

Eine kritische Größe beim empirischen Schätzen der Wahrscheinlichkeitsverteilung der unbeobachteten Heterogenität stellt die Anzahl an Stützvektoren S dar. Man kann diese Anzahl z.B. mithilfe einer Stoppregel-Prozedur basierend auf Kriterien wie das BIC oder das AIC bestimmen. Es werden so lange Stützvektoren hinzugefügt, solange der BIC- (bzw. AIC-)Wert sinkt. Ein Vorteil des BIC gegenüber dem AIC ist die Tatsache, dass das BIC die Anzahl der in der Analyse verwendeten Beobachtungen mitberücksichtigt und einen Anstieg der Stichprobengröße bestraft.

7.5 Eine dynamische Analyse der Marktstruktur

[49] Die erfolgreiche Entwicklung von Marketingstrategien erfordert die genaue Messung von Haushaltspräferenzen und ihren Reaktionen auf Variablen wie Preis oder Werbung.
Logistische Regression wird häufig verwendet, um die Kaufinformation von Haushalten, wie sie in Scannerpaneldaten zur Verfügung steht, zu modellieren. Das Modell kann von der Zufallsnutzentheorie hergeleitet werden, bei der man annimmt, dass die Konsumenten jene Marke mit dem höchsten Nutzen kaufen. Logistische Regression spezifiziert die Markenwahlwahrscheinlichkeiten als deterministische Funktionen der Marketingvariablen (wie Preis und Werbung) sowie demographischer Variablen (wie Haushaltseinkommen und Familiengröße). Interne Marktstrukturanalyse schließt von Präferenz- und Choicedaten für Konsumenten mit heterogenen Vorlieben bezüglich der Markenattribute auf Positionen der Marken in einem Produkteigenschaftsraum. Der in diesem Artikel verwendete Modellansatz berücksichtigt Heterogenität sowohl in den Präferenzen der Konsumenten als auch in ihren Wahrnehmungen der Markeneigenschaften. Weiters wird angenommen, dass vergangene Käufe die gegenwärtige Kaufentscheidung beeinflussen, d.h. die Konsumentenpräferenzen (Nutzengewichte) werden von den Eigenschaften der Marken, die in der Vergangenheit konsumiert wurden, beeinflusst. Was die Dynamik betrifft, handelt es sich um eine attributbasierte Quelle, d.h. es existiert keine Dynamik in dem Sinn, dass gegenwärtige Präferenzen bestimmter Marken von Marken, die in der Vergangenheit gekauft wurden, abhängen, sondern die Quelle der Dynamik basiert auf den Markeneigenschaften.

Betrachtet wird ein allgemeines Modell, in dem sich jeder der I Konsumenten zu jeder Kaufgelegenheit $t = 1, 2, \ldots, T_i$ für eine einzige der J verfügbaren Marken entscheiden muss. Der Nutzen, den Konsument i dieser Kauf bringt, ist

$$U_{ijt} = \gamma_{ijt} + \alpha_i p_{ijt} + \varepsilon_{ijt}$$

p_{ijt} ... Preis
γ_{ijt} ... Interzept
ε_{ijt} ... zeitvariante stochastische Komponente.

[49] [19]: „A Dynamic Analysis of Market Structure Based on Panel Data"

Der Interzeptterm im Zufallsnutzenmodell kann folgendermaßen spezifiziert werden:

- zeitinvariant:

$$\gamma_{ijt} = \beta_{i1} a_{j1} + \beta_{i2} a_{j2}$$

a_{j1}, a_{j2} ... Eigenschaften der Marke j (unbeobachtbar)
β_{i1}, β_{i2} ... Nutzengewichte bezüglich der ersten bzw. zweiten gemeinsamen Eigenschaften.

Die Nutzengewichte sind unabhängig identisch normalverteilt mit Varianz σ_c^2, d.h.

$$\beta_{ik} = \beta_k + \mu_{cik}$$

mit β_k als mittleres Nutzengewicht für die k-te gemeinsame Eigenschaft und

$$\mu_{cik} \sim N(0, \sigma_c^2) \ .$$

- zeitvariant:

$$\gamma_{ijt} = \beta_{i1} a_{j1} + \beta_{i2} a_{j2} + c d_{i,t-1,j}$$

d_{itj} ... Wahlvariable
c ... Nutzengewicht der verzögerten Wahl $d_{i,t-1,j}$.

$d_{itj} = 1$, falls Konsument i zum Zeitpunkt t Marke j kauft und $d_{itj} = 0$ sonst, d.h. $d_{itj} = 1$ dann und nur dann, wenn $U_{ijt} = \max\{U_{i1t}, \ldots, U_{iJt}\}$.

Obige Nutzengleichung kann man nun folgendermaßen umformulieren:

$$U_{ijt} = V_{ijt} + \varepsilon_{ijt}$$

$$V_{ijt} = \gamma_{ijt} + \alpha_i p_{ijt} \ .$$

Die Parametermittelwerte werden in einem Vektor θ_M zusammengefasst, welcher die mittleren Preiskoeffizienten α, die mittleren gemeinsamen Attribute a_{jk}, die Mittelwerte der Nutzengewichte β_k und den Koeffizienten der verzögerten Wahl c enthält. Auch die Standardabweichungen und Korrelationen der Parameter werden in einem Vektor θ_s zusammengefasst, der insbesondere aus der Standardabweichung des Preiskoeffizienten σ_α, den Standardabweichungen der gemeinsamen Attributwahrnehmungen σ_{a_k}, den Standardabweichungen der Nutzengewichte über die Konsumenten sowie über die

Kaufgelegenheiten σ_c und σ_d und den Korrelationen zwischen den Preiskoeffizienten und den Nutzengewichten ρ_k besteht. Es wird angenommen, dass die Fehlerterme ε_{ijt} unabhängig identisch extremwertverteilt sind. Dann handelt es sich bei der Wahrscheinlichkeit, dass Konsument i Marke j zum Zeitpunkt t kauft, um eine multinomiale Logit Choice Wahrscheinlichkeit.[50] Da die zustandspezifischen Parameter heterogen sein können, ist diese Wahrscheinlichkeit bedingt auf den Raum der Zufallsterme $\mu_{\alpha_i}, \mu_{a_{ik}}, \mu_{1\beta_{it}}$ und $\mu_{2\beta_{it}}$, die wiederum in einem Vektor ν_i zusammengefasst werden, und muss darüber integriert werden.
Die Wahrscheinlichkeit, dass Konsument i Marke j zum Zeitpunkt t wählt, hängt von ν_i und θ_M ab:

$$P_{ijt}(\nu_i, \theta_M) = \frac{e^{V_{ijt}(\nu_i, \theta_M)}}{\sum_{l=1}^{J} e^{V_{ilt}(\nu_i, \theta_M)}} .$$

Die Wahrscheinlichkeit, dass Konsument i die durch d_{ijt} gegebene Folge von Käufen tätigt, hängt von ν_i und θ_M ab:

$$P_i(\nu_i, \theta_M) = \prod_{t=1}^{T_i} \prod_{j=1}^{J} P_{ijt}(\nu_i, \theta_M)^{d_{ijt}} .$$

Somit lautet die von θ abhängige Wahrscheinlichkeit, dass Konsument i die durch d_{ijt} angegebene Kauffolge wählt

$$P_i(\theta) = \int_{\nu_i} P_i(\nu_i, \theta_M) f(\nu_i|\theta_s) d\nu_i$$

$f(\nu_i|\theta_s)$ ... multinomiale normale Wahrscheinlichkeitsverteilungsfunktion für ν_i bedingt auf θ_s.

Die entsprechende Loglikelihoodfunktion ist dann

$$LogL(\theta) = \sum_{i=1}^{I} \ln P_i(\theta) .$$

[50] Nimmt man ε_{ijt} als unabhängig identisch normalverteilt an, ergibt sich stattdessen ein Probit Modell.

Das Modell kann mithilfe von Simulationstechniken geschätzt werden, wobei sich speziell Monte Carlo Methoden zur Simulation der hochdimensionalen Integrale, welche in die Likelihoodfunktion eingehen, besser eignen als eine numerische Auswertung. MCMC Methoden sind zufriedenstellend solange eine große Anzahl an Zufallszügen und entsprechende Glättungstechniken verwendet werden.

7.6 Das endliche Mischungs-Strukturgleichungsmodell

[51] Heterogenität und Marktsegmentierung können simultan über ein allgemeines endliches Mischungs-Strukturgleichungsmodell behandelt werden. Es werden gleichzeitig Cluster (Segmente) gebildet und segmentspezifische Mess- und Strukturparameter einer postulierten Modellstruktur geschätzt. Eine zweistufige Methode, in der zuerst Gruppen ohne Berücksichtigung des strukturellen Modells geformt werden und anschließend eine multigruppale Strukturgleichungsmethode auf die segmentierten Daten angewendet wird, ist in den meisten Fällen statistisch ineffizient. Weiters ist solch eine Vorgangsweise für sehr große Modelle oft nicht durchführbar.
Die Strukturgleichungen bilden ein Grundgerüst aus unbeobachteten Konstrukten und manifesten Indikatoren. Alle beobachteten Variablen werden mit einem Fehler gemessen, wobei sowohl Mess- als auch Strukturfehler zugelassen werden.
Speziell wird folgendes Messmodell verwendet:

$$y|g = \nu_y^g + \Lambda_y^g \eta^g + \varepsilon^g \text{ [52]} \tag{1}$$

$$x|g = \nu_x^g + \Lambda_x^g \xi^g + \delta^g \text{ [53]} \tag{2}$$

g	...	Segmentzugehörigkeit ($g = 1, \ldots, G$)
$y\|g$	...	beobachtbare Indikatorvariablen, die η^g messen
$x\|g$	...	beobachtbare Indikatorvariablen, die ξ^g messen
η^g	...	endogene latente Variablen
ξ^g	...	exogene latente Variablen
ν_y^g, ν_x^g	...	Messinterzeptterme
Λ_y^g, Λ_x^g	...	Koeffizientenmatrizen
ε^g, δ^g	...	Messfehler in $y\|g$ bzw. $x\|g$.

[51] [29]: „Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity"
[52] $y \in (p \times 1), \nu_y \in (p \times 1), \Lambda_y \in (p \times m), \eta \in (m \times 1)$
[53] $x \in (q \times 1), \nu_x \in (q \times 1), \Lambda_x \in (q \times n), \xi \in (n \times 1)$

Außerdem werden für die Parameter diverse Annahmen spezifiziert:

- $\mathbb{E}(\xi^g) = \tau_\xi^g$
- $\mathbb{E}[(\xi^g - \tau_\xi^g)(\xi^g - \tau_\xi^g)'] = \Phi^g$
- $\mathbb{E}(\varepsilon^g \varepsilon^{g'}) = \Theta_\varepsilon^g$
- $\mathbb{E}(\delta^g \delta^{g'}) = \Theta_\delta^g$
- $\mathbb{E}(\varepsilon^g) = \mathbb{E}(\delta^g) = 0$
- Messfehler und latente Variablen sind unkorreliert.

Den zweiten Teil des Strukturgleichungsmodells bildet das Strukturmodell:

$$B_g \eta^g = \alpha^g + \Gamma^g \xi^g + \zeta^g \tag{3}$$

B_g ... Strukturparameter (spezifizieren die Verbindungen zwischen den endogenen latenten Variablen)
α^g ... Interzeptterm
Γ^g ... Koeffizientenmatrix (beschreibt den Effekt von ξ^g auf η^g)
ζ^g ... Störterme.

Auch hier werden Annahmen für die Modellparameter getroffen:

- $\mathbb{E}(\zeta^g \zeta^{g'}) = \psi^g$
- $\mathbb{E}(\zeta^g) = 0$
- ζ^g unkorreliert mit ξ^g
- B_g nichtsingulär.

Spezialfälle des obigen Strukturgleichungsmodells stellen folgende Ansätze dar:

- Finite Mixture Confirmatory Factor Model: Beschränkung entweder nur auf die erste oder nur auf die zweite Gleichung;
- Finite Mixture Simultaneous Equation Model: Beschränkung nur auf die dritte Gleichung (d.h. alle Variablen werden fehlerfrei gemessen);
- Finite Mixture Second-Order Confirmatory Factor Model: Beschränkung auf Gleichungen eins und drei; oder
- allgemeine Typen von Heterogenität: Verwendung aller drei Gleichungen (über die Segmente unterschiedliche Struktur- und Messmodelle sind erlaubt).

$\Delta|g = \begin{bmatrix} y|g \\ x|g \end{bmatrix}$ bezeichne den gemeinsamen Vektor der beobachtbaren Indikatorvariablen bedingt auf die Zugehörigkeit zu Gruppe g. Dann lauten ihr bedingter Mittelwertsvektor

$$\mu_g = \begin{bmatrix} \nu_y^g + \Lambda_y^g B_g^{-1}(\alpha^g + \Gamma^g \tau_\xi^g) \\ \nu_x^g + \Lambda_x^g \tau_\xi^g \end{bmatrix}$$

und die bedingte Kovarianzmatrix

$$\Sigma_g = \begin{bmatrix} \Lambda_y^g B_g^{-1}(\Gamma^g \Phi^g \Gamma^{g'} + \psi^g) B_g^{-1'} \Lambda_y^{g'} + \Theta_\varepsilon^g & \Lambda_y^g B_g^{-1} \Gamma^g \Phi^g \Lambda_x^{g'} \\ \Lambda_x^g \Phi^g \Gamma^{g'} B_g^{-1'} \Lambda_y^{g'} & \Lambda_x^g \Phi^g \Lambda_x^{g'} + \Theta_\delta^g \end{bmatrix}.$$

Es wird angenommen, dass $\Delta|g$ eine bedingte multivariate Normalverteilung besitzt. Somit stellt die unbedingte Verteilung des beobachteten Vektors $\Delta = \begin{bmatrix} y \\ x \end{bmatrix}$ eine endliche Mischung dieser Verteilungen dar:

$$\Delta \sim \sum_{g=1}^{G} w_g f_g(\Delta|\mu_g, \Sigma_g)$$

$w = (w_1, \ldots, w_G)'$ ... Mischproportionen ($w_g > 0$, $\sum_{g=1}^{G} w_g = 1$)
$f(\cdot)$ ... Dichte der bedingten multivariaten Normalverteilung.

Daraus ergibt sich die Likelihoodfunktion für eine Stichprobe $(\Delta_1, \ldots, \Delta_N)$ zufällig gezogener Beobachtungen als

$$L = \prod_{i=1}^{N} \left[\sum_{g=1}^{G} w_g (2\pi)^{-(p+q)/2} |\Sigma_g|^{-1/2} e^{-1/2(\Delta_i - \mu_g)' \Sigma_g^{-1} (\Delta_i - \mu_g)} \right]$$

p ... Anzahl der Indikatoren der endogenen Konstrukte
q ... Anzahl der Indikatoren der exogenen Konstrukte.

L ist eine Funktion von w_g, B_g, Γ^g, Λ_x^g, Λ_y^g, ν_x^g, ν_y^g, α^g, Φ^g, ψ^g, Θ_δ^g, Θ_ε^g und τ_ξ^g. Das Problem besteht nun in der Maximierung von L gegeben $(\Delta_1, \ldots, \Delta_N)$ und eine spezifizierte Anzahl an Gruppen G unter den Restriktionen für w und $|\Sigma_g| > 0$.[54] Die Maximum Likelihood Schätzer $\hat{\Sigma}_g$ und $\hat{\mu}_g$ sowie die Mischproportionen $\hat{w}_g$ sind Funktionen des postulierten theoretischen Modells.

Die Parameterschätzungen werden im Anschluss dazu verwendet, um jedem der G Segmente mithilfe der Bayes'schen Regel individuelle Beobachtungen i zuzuordnen:

$$\hat{P}_{ig} = \frac{\hat{w}_g f_g(\Delta_i | \hat{\mu}_g, \hat{\Sigma}_g)}{\sum_{k=1}^{G} \hat{w}_k f_k(\Delta_i | \hat{\mu}_k, \hat{\Sigma}_k)}$$

$\hat{P}_{ig}$ ... a posteriori Wahrscheinlichkeit, dass Beobachtung i zu Segment g gehört.

Kennt man die Segmentanzahl G a priori, kann man geschachtelte Modelle einfach über Likelihood Ratio Statistiken miteinander vergleichen. Ist die Anzahl der Segmente allerdings a priori nicht bekannt, ist dies aufgrund der Verletzung der Regularitätsbedingungen nicht mehr möglich. Da in der Praxis jedoch eher letzteres der Fall ist, werden meist statistische Kriterien wie das konsistente Akaike (CAIC) oder das Bayes Informationskriterium (BIC) zur Bestimmung des optimalen G verwendet.
Weiters kann der Grad der Separation der verschiedenen Segmente über ein entropiebasiertes Maß

$$E_G = 1 - \left[\sum_i \sum_g -\hat{P}_{ig} \ln \hat{P}_{ig} \right] / (N \ln G)$$

[54] minimal erforderliche Stichprobengröße: $\frac{(p+q)(p+q+1)}{2}$

geschätzt werden. Es beruht auf den a posteriori Wahrscheinlichkeiten und liegt zwischen 0 und 1, wobei Werte nahe bei 0 auf schlecht separierte a posteriori Wahrscheinlichkeiten hindeuten.

7.7 Über die Heterogenität der Nachfrage

[55] Traditionellerweise wird Heterogenität in der Nachfrage über Segmente bestehend aus homogenen Konsumenten definiert. Allerdings existieren in der realen Welt keine wirklich homogenen Segmente. Mittels Verwendung eines Mischungsmodells normalverteilter Komponenten kann gezeigt werden, dass tatsächlich Heterogenität innerhalb der Segmente auftritt.
In diesem allgemeinen Modell verschachtelt liegt auch das endliche Mischungsmodell.

Das Mischungsmodell mit normalverteilten Komponenten wird im Kontext eines Logit Modells eingeführt, in dem die Wahlwahrscheinlichkeit von Alternative i für Haushalt j als

$$p_{i,j} = \frac{e^{x_i' \beta_j}}{\sum_{l=1}^{L} e^{x_l' \beta_j}}$$

l ... Index der Wahlalternativen
x_i ... erklärende Variablen für Alternative i
β_j ... Markenpräferenzen und Sensibilität bezüglich der Kovariate für Haushalt j

definiert wird. Heterogenität wird mithilfe eines Modells mit normalverteilten Mischungskomponenten der Form

$$\beta_j \sim \sum_k \phi_k N(\bar{\beta}_k, D_k)$$

k ... Anzahl der Komponenten
ϕ_k ... Masse jeder Komponente
N ... Normalverteilung

spezifiziert. Diese Formulierung ermöglicht die Darstellung eines weiten Spektrums an Heterogenitätsverteilungen und erlaubt die Berücksichtigung einer

[55] [1]: „On the Heterogeneity of Demand"

Vielzahl an alternativen Heterogenitätstypen.

Das Normalkomponenten-Mischungsmodell kann mit einer Gibbs Sampling Prozedur geschätzt werden, bei der man eine latente Variable s_{jk} einführt, die die k Komponenten danach indiziert, zu welchem der j Befragten sie gehören. In jeder Iteration werden nur jene Daten verwendet, die von den Befragten einer bestimmten Komponente stammen, um den entsprechenden Komponentenmittelwert ($\bar{\beta}_k$) und die Kovarianzmatrix (D_k) zu ermitteln. Die Zuordnung der Befragten zu den einzelnen Komponenten erfolgt direkt, wenn die Massepunkte der Komponenten $\{\phi_k\}$ bekannt sind.
Statistische Identifizierbarkeit erhält man, indem man für die Komponentenmassen beispielsweise eine ordinale Restriktion der Form $\phi_1 > \phi_2 > \ldots > \phi_k$ über die a priori Verteilung einführt.

7.8 Die hierarchische Bayes Methode in Strukturgleichungsmodellen

[56] Die Annahme invarianter Faktorkovarianzen und Messfehlervarianzen kann in vielen empirischen Anwendungen nicht aufrechterhalten werden, weshalb ein hierarchisches Bayesmodell[57] entwickelt wurde, das die Spezifikation und Schätzung von Heterogenität sowohl in den Mittelwert- als auch den Kovarianzstrukturen erlaubt. Es wird somit Heterogenität in der Faktorkovarianzstruktur, den Messfehlern und den Strukturparametern berücksichtigt.

Zuerst werden für jedes Individuum Struktur- und Messmodelle spezifiziert. Auf der zweiten Stufe wird dann eine Bevölkerungsverteilung festgelegt, über die beobachtete und unbeobachtete Quellen an Heterogenität berücksichtigt werden. Die Bevölkerungsverteilung beschreibt, wie individuelle Parameter in den beiden Gleichungen über die Bevölkerung variieren. Beim latenten Klassenmodell im Speziellen ist diese Verteilung diskret.

[56] [7]: „A Hierarchical Bayesian Methodology for Treating Heterogeneity in Structural Equation Models"

[57] genauer gesagt ein teilweise rekursives Zufallseffekt-Strukturgleichungsmodell

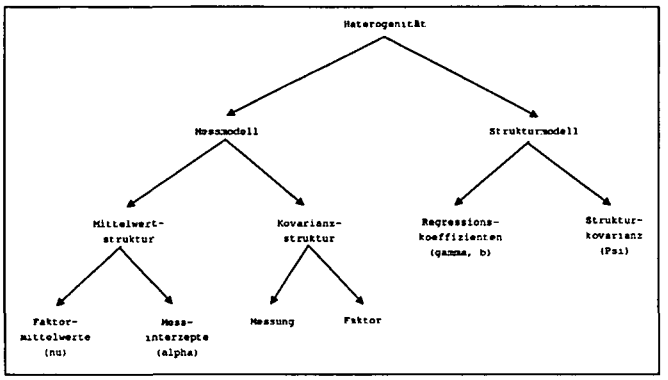

Abbildung 2: Heterogenitätstypen (Quelle: [7])

Die Beziehungen zwischen den manifesten Variablen können über latente Konstrukte beschrieben werden. Das Messmodell für Individuum i, welches die Beziehung zwischen den beobachteten und den latenten Variablen beschreibt, lautet:

$$\begin{aligned} x_{ij} &= \alpha_{i,x} + \Lambda_{i,x}\xi_{ij} + \delta_{ij} \\ y_{ij} &= \alpha_{i,y} + \Lambda_{i,y}\eta_{ij} + \epsilon_{ij} \end{aligned}$$

x_{ij}, y_{ij}	...	exogene bzw. endogene (manifeste) Indikatorvariablen
$i = 1, \ldots, I$	...	Individuen
$j = 1, \ldots, N_i$	...	zu Individuum i gehörende Beobachtungen
$\alpha_{i,x}, \alpha_{i,y}$	...	Messinterzept der exogenen bzw. endogenen Indikatorvariablen
$\Lambda_{i,x}, \Lambda_{i,y}$	...	Faktorladungen
ξ_{ij}, η_{ij}	...	exogene bzw. endogene latente Variablen
$\delta_{ij}, \epsilon_{ij}$	...	Messfehler
p, q	...	Anzahl der exogenen bzw. endogenen (manifesten) Variablen.

Es soll weiters gelten $\delta_{ij} \sim N(0, \Theta_{i,x})$ und $\epsilon_{ij} \sim N(0, \Theta_{i,y})$, wobei $\Theta_{i,x}$ und $\Theta_{i,y}$ diagonal sind und die Messfehlervarianzen enthalten. Es wird angenommen, dass die m latenten Faktoren in ξ_{ij} gemäß $N(\nu_i, \Phi_i)$ normalverteilt sind, wobei ν_i die Faktormittelwerte und Φ_i die Kovarianzmatrix der Faktorauswertungen bezeichnen.
Im Messmodell können sich die Faktormittelwerte ν_i, die Faktorkovarianzen ϕ_i und die Messfehlervarianzen $\Theta_{i,x}$ bzw. $\Theta_{i,y}$ über die Individuen unterscheiden. Was die Messheterogenität betrifft, wird $\alpha_{i,x} = \alpha_x$ und $\alpha_{i,y} = \alpha_y$ gesetzt.

Auch die Faktorladungsmatrizen sollen invariant über die Individuen sein, speziell gilt $\Lambda_{i,x} = \Lambda_x$ und $\Lambda_{i,y} = \Lambda_y$. Da die Skalen der latenten Faktoren willkürlich sind, kann man die entsprechenden Elemente in den Ladungsmatrizen auf 1 setzen. Die individuellen Faktormittelwerte ν_i werden gemäß $N(0, \Delta)$ in der Bevölkerung multivariat normalverteilt angenommen, dürfen also über die Individuen variieren. Ein Mittelwert von 0 wird angenommen, um Identifizierbarkeit zu sichern. Weiters wird vorausgesetzt, dass die Genauigkeitsmatrix Φ_i^{-1} von einer allgemeinen Wishart Bevölkerungsverteilung $W(\rho, R)$ mit ρ Freiheitsgraden und positiv definiter Skalenmatrix R stammt. Spezielle zusätzliche Restriktionen je nach betrachtetem Fall garantieren Identifizierbarkeit. Außerdem soll jede Messfehlervarianz von einer unabhängigen invertierten Gammaverteilung $IG(\Theta; a, b)$ der Bevölkerung stammen.

Das Strukturmodell, welches die latenten Konstrukte ξ_{ij} und η_{ij} jedes Individuums in Beziehung zueinander stellt, wird folgendermaßen definiert:

$$B_i \eta_{ij} = \gamma_{0i} + \Gamma_i \xi_{ij} + \zeta_{ij}$$

- B_i ... Strukturparameter (spezifizieren die Verbindungen zwischen den endogenen latenten Variablen)
- γ_{0i} ... Strukturinterzeptterm
- Γ_i ... Koeffizientenmatrix (bezeichnet die Effekte von ξ_{ij} auf η_{ij})
- ζ_{ij} ... Störungen.

Es wird angenommen, dass die Störungen ζ_{ij} mit ξ_{ij} unkorreliert und gemäß $N(0, \Psi_i)$ verteilt sind. In dem hier betrachteten Fall ist B_i eine Dreiecksmatrix. Beim hierarchischen Ansatz sind die individuellen Parameter $\varphi_i = \{\alpha_{i,x}, \alpha_{i,y}, \Lambda_{i,x}, \Lambda_{i,y}, \nu_i, \Phi_i, \Theta_{i,x}, \Theta_{i,y}, B_i, \gamma_{0i}, \Gamma_i, \Psi_i\}$ als Zufallsvariablen spezifiziert gezogen aus einer allgemeinen Bevölkerungsverteilung $h(\varphi_i)$.
Im Strukturmodell können die Strukturkoeffizienten über die Individuen variieren. Der Vektor π_i, der die Terme in B_i, γ_{i0} und Γ_i enthält, wird als multivariat normalverteilt gemäß $N(Z_i\vartheta, \Upsilon)$ angenommen mit Z_i als Matrix individuell spezifischer Kovariate (Alter, Geschlecht, usw.). Sind diese nicht verfügbar, reduziert sich Z_i auf die Einheitsmatrix. Es ist also möglich, sowohl beobachtete als auch unbeobachtete Heterogenität zu berücksichtigen. Die Parameter in ϑ erklären die individuellen Differenzen in den Strukturparametern in Ausdrücken der individuellen Kovariate. Υ fängt die Kovariation

in den Strukturparametern resultierend aus den unbeobachteten individuellen Variablen ein. Sie wird als invariant über die Individuen angenommen. Aufgrund der Skalenunbestimmtheit der endogenen Faktoren wird $\mathbb{E}(\gamma_{i0}) = 0$ gesetzt, um Identifizierbarkeit zu erhalten.

Das komplette zweistufige Modell unter Berücksichtigung aller oben getroffenen Annahmen reduziert sich demnach auf

1.Stufe:

$$\begin{aligned} x_{ij} &= \alpha_x + \Lambda_x \xi_{ij} + \delta_{ij} \\ y_{ij} &= \alpha_y + \Lambda_y \eta_{ij} + \epsilon_{ij} \\ B_i \eta_{ij} &= \gamma_{0i} + \Gamma_i \xi_{ij} + \zeta_{ij} \end{aligned}$$

$$\begin{aligned} \delta_{ij} &\sim N(0, \Theta_{i,x}) \\ \epsilon_{ij} &\sim N(0, \Theta_{i,y}) \\ \xi_{ij} &\sim N(\nu_i, \Phi_i) \\ \zeta_{ij} &\sim N(0, \Psi) \end{aligned}$$

2.Stufe:

$$\begin{aligned} \nu_i &\sim N(0, \Delta) \\ \Phi_i^{-1} &\sim W(\rho, R) \\ \Theta_{ix} &\sim \prod_{k=1}^{p} IG(a_k, b_k) \\ \Theta_{iy} &\sim \prod_{k=1}^{q} IG(a_k, b_k) \\ \pi_i &\sim N(Z_i \vartheta, \Upsilon). \end{aligned}$$

Zur Fixierung des Ursprungs der Faktoren werden $\mathbb{E}(\nu_i) = 0$ und $\mathbb{E}(\gamma_{i0}) = 0$ gewählt. Außerdem werden die Ladungen einer Indikatorvariable pro Faktor auf eins gesetzt, um die Skala der Faktoren festzulegen.

Genauer betrachtet wird im Folgenden der Fall, in dem B_i eine Dreiecksmatrix ist, aber Ψ_i uneingeschränkt bleibt. $w_{ij} = \{y_{ij}, x_{ij}\}$ bezeichne den gemeinsamen Vektor der manifesten Variablen einer willkürlichen Beobachtung j von Individuum i, welche aus einer multivariaten Normalverteilung $f_i(w_{ij}, \mu_i, \Sigma_i)$ mit bedingtem Mittelwertsvektor

$$\mu_i = \begin{pmatrix} \alpha_y + \Lambda_y B_i^{-1}(\gamma_{i0} + \Gamma_i \nu_i) \\ \alpha_x + \Lambda_x \nu_i \end{pmatrix}$$

und bedingter Kovarianzmatrix

$$\Sigma_i = \begin{pmatrix} \Lambda_y B_i^{-1}(\Gamma_i \Phi_i \Gamma_i' + \Psi) B_i^{-1'} \Lambda_y' + \Theta_{iy} & \Lambda_y B_i^{-1} \Gamma_i \Phi_i \Lambda_x' \\ \Lambda_x \Phi_i \Gamma_i' B_i^{-1'} \Lambda_y' & \Lambda_x \Phi_i \Lambda_x' + \Theta_{ix} \end{pmatrix}$$

stammt.

Die Likelihood lautet

$$L_i = \prod_{j=1}^{N_i} (2\pi)^{-(p+q)/2} |\Sigma_i|^{-1/2} e^{-(1/2)(w_{ij}-\mu_i)'\Sigma_i^{-1}(w_{ij}-\mu_i)},$$

die unbedingte Likelihood für eine zufällige Stichprobe von I Individuen ist gegeben durch eine kontinuierliche Mischung

$$L = \prod_{i=1}^{I} \int \int \dots \int L_i(\mu_i(\varphi), \Sigma_i(\varphi)) h(\varphi) d\varphi$$

mit $h(\varphi)$ als kontinuierliche Bevölkerungsverteilung. Sie ist eine Funktion der Parameter $\varphi = \{\alpha_x, \alpha_y, \Lambda_x, \Lambda_y, \rho, R, \Delta, a_x, b_x, a_y, b_y, \vartheta, \Upsilon, \Psi\}$ und kann nicht in geschlossener Form dargestellt werden, was Maximum Likelihood Schätzung extrem kompliziert macht. MCMC Prozeduren umgehen die Auswertung komplexer mehrdimensionaler Integrale, die bei ML Methoden notwendig wäre.

7.9 Endliche Mischungen verallgemeinerter linearer Modelle mit Zufallseffekten

[58] Komplizierte Heterogenitätsstrukturen können mithilfe einer Erweiterung des traditionellen Zufallseffektmodells beschrieben werden, indem man eine endliche Mischung von Normalverteilungen für die Verteilung der Koeffizienten verwendet.
Endliche Mischungsmodelle unterteilen die Stichproben in Mischkomponenten und schätzen die Mischwahrscheinlichkeiten sowie die unbekannten Parameter jeder Komponentendichte. Sie benötigen oft eine exzessive Anzahl

[58] [32]: „Bayesian Inference for Finite Mixtures of Generalized Linear Models with Random Effects"

an relativ kleinen, latenten Klassen oder Subpopulationen, um die Heterogenität in den Daten entsprechend repräsentieren zu können (Überparametrisierung). Eine Alternative stellt das Zufallseffektmodell dar, welches annimmt, dass die subjektspezifischen Koeffizienten eine zufällige Stichprobe aus einer Normalverteilung repräsentieren. Sie benötigen weniger Parameter, weisen aber Mängel in der Flexibilität auf. Genauer gesagt sind sie nicht imstande nicht-normale Heterogenität (z.B. eine multimodale Verteilung) zu beschreiben. Der hier betrachtete hierarchische Bayes Ansatz zur Formulierung von Parameterheterogenität in verallgemeinerten linearen Modellen kombiniert die Flexibilität endlicher Mischungs- oder latenter Klassenmodelle, welche gemeinsame Parameter für jede Subpopulation voraussetzen, und die Sparsamkeit von Zufallseffektmodellen, welche Normalverteilungen der Regressionsparameter annehmen. Das Mischungs-Zufallseffektmodell erlaubt die Berücksichtigung von Heterogenität innerhalb jeder latenten Klasse oder Subpopulation. Es ist sparsamer als das latente Klassenmodell und flexibler als das traditionelle Zufallseffektmodell.

Im latenten Klassenmodell werden zuerst die Koeffizienten für jedes Segment geschätzt und danach basierend auf den beobachteten Choices die Wahrscheinlichkeit, dass die Beobachtung aus den Segmenten stammt, berechnet. Es stehen Beobachtungen für n Subjekte zur Verfügung:

$$Y_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{bmatrix}, \; X_i = \begin{bmatrix} x'_{i1} \\ \vdots \\ x'_{im_i} \end{bmatrix}$$

Y_{ij} ... j-te abhängige Beobachtung für Subjekt i
x_{ij} ... unabhängige Variablen (mit Einträgen 1 für das Interzept).

Im latenten Klassenmodell wird angenommen, dass die Bevölkerung aus K Subpopulationen besteht, innerhalb derer es separate Regressionsmodelle gibt. Angenommen Subjekt i gehört zu Klasse k, dann spezifiziert das latente Klassenmodell:

$$Y_i = X_i\theta_k + \epsilon_{ik} \; .$$

Die Fehlerterme ϵ_{ik} seien multivariat normalverteilt mit Mittelwert 0 und Kovarianzmatrix $\sigma_k^2 I$ und sollen über die Individuen voneinander unabhängig sein. ψ_k bezeichne das Verhältnis der Population, die zu Klasse k gehört. Die

nicht auf die Klassenzugehörigkeit bedingte Dichte von Y_i bildet dann ein endliches Mischungsmodell mit K Komponentendichten:

$$f_i(Y_i) = \sum_{k=1}^{K} \psi_k q_{m_i}(Y_i | X_i\theta_k, \sigma_k^2 I)$$

mit $q_{m_i}(\cdot | X_i\theta_k, \sigma_k^2 I)$ als m_i-dimensionale multivariate normale Dichte mit Mittelwert $X_i\theta_k$ und Kovarianzmatrix $\sigma_k^2 I$.
Das latente Klassenmodell weist Mängel auf, wenn wesentliche Heterogenität innerhalb der Klassen existiert. Ergäben sich beispielsweise im Mischungs-Zufallseffektmodell zwei Klassen, würde das endliche Mischungsmodell ohne Zufallseffekte[59] mehr als nur diese beiden Klassen benötigen, um die Heterogenität in den Parametern hinreichend gut zu erklären.

Beim normalen Mischungs-Zufallseffektmodell wird angenommen, dass die Regressionskoeffizienten subjektspezifisch sind und innerhalb der Klassen einer Normalverteilung folgen. Für Subjekt i gilt:

$$Y_i = X_i\beta_i + \epsilon_i$$

β_i ... Regressionskoeffizienten
ϵ_i ... Fehlerterme.

Die Fehlerterme seien voneinander unabhängig multivariat normalverteilt mit Mittelwert 0 und Kovarianzmatrix $\sigma_i^2 I$. Weiters wird angenommen, dass die Logfehlervarianzen $\{\phi_i = \log(\sigma_i^2)\}$ eine zufällige Stichprobe aus einer Normalverteilung mit Mittelwert α und Varianz τ^2 bilden.
Falls Subjekt i zur Klasse k gehört, besitzt β_i eine Normalverteilung mit Mittelwert θ_k und Kovarianzmatrix Λ_k, und das Verhältnis der Population, die zu Klasse k gehört, ist ψ_k. Die Regressionskoeffizienten bilden also bei unbekannter Klassenzugehörigkeit eine Zufallsstichprobe aus einer Mischverteilung mit Dichte

$$g(\beta_i) = \sum_{k=1}^{K} \psi_k q_p(\beta_i | \theta_k, \Lambda_k)$$

mit $q_p(\cdot | \theta_k, \Lambda_k)$ als p-dimensionale multivariate normale Dichte mit Mittelwert θ_k und Kovarianzmatrix Λ_k. Unbedingter Mittelwert und Kovarianz von

[59] in dem also keine Variabilität innerhalb der Klassen erlaubt ist

β_i lauten

$$\mathbb{E}(\beta_i) = \theta = \sum_{k=1}^{K} \psi_k \theta_k$$

$$\mathbb{V}(\beta_i) = \Lambda = \sum_{k=1}^{K} \psi_k (\Lambda_k + \theta_k \theta_k') - \theta\theta'.$$

Nach Integration über β_i ergibt sich auch für die Randverteilung von Y_i ein Mischungsmodell:

$$f_i(Y_i) = \sum_{k=1}^{K} \psi_k q_{m_i}(Y_i | X_i\theta_k, \sigma_i^2 I + X_i \Lambda_k X_i').$$

Das latente Klassenmodell erhält man, wenn man die Varianzen innerhalb der Klassen gegen null gehen lässt ($\Lambda_k = 0$), das traditionelle Zufallseffektmodell ergibt sich, wenn man nur eine Klasse oder Komponente ($K = 1$) berücksichtigt. Die Betrachtung von nur einer Klasse ($K = 1$) und zusätzliches Nullsetzen der Varianz ($\Lambda_k = 0$) führt zu einem einfachen aggregierten Modell.[60]

Beim Mischungs-Zufallseffektmodell für verallgemeinerte lineare Modelle stammt die j-te abhängige Variable Y_{ij} für das i-te Subjekt von einem verallgemeinerten linearen Modell mit der Dichte

$$f(Y_{ij}|\beta_i) = e^{\frac{Y_{ij} h(x_{ij}'\beta_i) - b[h(x_{ij}'\beta_i)]}{a(\phi_i)} + c(Y_{ij}, \phi_i)}$$

x_{ij}	...	unabhängige Variablen
β_i	...	Regressionskoeffizienten
$h(x_{ij}'\beta_i) = \xi_{ij}$	...	natürlicher Parameter
ϕ_i	...	Skalenparameter.

Die Funktionen a, b und h sind univariat und reellwertig. Die Beobachtungen sollen voneinander unabhängig sein. Mittelwert und Varianz von Y_{ij} sind $b_1(\xi_{ij})$ und $b_2(\xi_{ij})a(\phi_i)$ mit $b_1 = \frac{d}{d\xi} b(\xi)$ als erste Ableitung von b und

[60] D.h. alle Konsumenten besitzen die gleichen Parameter.

$b_2 = \frac{d^2}{d\xi^2} b(\xi)$ als zweite Ableitung von b.[61]

Das Modell ist nicht identifizierbar, weil Permutationen der Klassenbezeichnungen zum gleichen Wert der Likelihoodfunktion führen. Dieses Problem ist unter dem Begriff „Label Switching" bekannt. Man kann das Modell jedoch beispielsweise durch Ordnen der Mischwahrscheinlichkeiten $\psi_1 < \ldots < \psi_K$ identifizierbar machen.
Die für die MCMC Methode benötigten a priori Verteilungen der Parameter sind voneinander unabhängig und werden üblicherweise wie folgt gewählt:

- ψ besitzt eine Dirichletverteilung[62] beschränkt auf den Bereich $\psi_1 < \ldots < \psi_K$.
- θ_k besitzt eine multivariate Normalverteilung.
- Λ_k besitzt eine invertierte Wishartverteilung[63].
- α besitzt eine Normalverteilung.
- τ^2 besitzt eine invertierte Gammaverteilung.
- K besitzt eine diskrete Wahrscheinlichkeitsfunktion auf den natürlichen Zahlen $1, \ldots, M$[64].

Die spezielle Wahl der a priori Verteilungen erleichtert die a posteriori Analyse. Weiters handelt es sich um ziemlich flexible Familien, und die a priori Parameter können so gewählt werden, dass die a posteriori Analyse relativ unsensibel gegenüber den Priors für Datenmengen mit moderater Anzahl an Subjekten und Beobachtungen pro Subjekt ist. Die Priors werden derart spezifiziert, dass sie nahezu uninformativ sind, d.h. die a priori Standardabweichung wird so gewählt, dass der Bereich der Variabilität in der a priori Verteilung viel größer ist als der erwartete Bereich der Variabilität in den tatsächlichen Parametern. Als Konsequenz davon verläuft die a priori Verteilung im Bereich, in dem die Likelihoodfunktion den Großteil ihrer Masse

[61] Für die Normalverteilung ist $h(x'_{ij}\beta_i)$ der Mittelwert, $a(\phi_i) = e^{\phi_i}$ die Varianz, $b[h(x'_{ij}\beta_i)] = \frac{1}{2}h(x'_{ij}\beta_i)^2$ und $c(y_{ij}, \phi_i) = -\frac{1}{2}(y_{ij}^2 e^{-\phi_i} + \ln(2\pi) + \phi_i)$.

[62] Dirichletverteilung:
$f(\psi) = \frac{1}{Z(u)} \prod_{k=1}^{K} \psi_k^{u_k - 1}$ mit $Z(u) = \frac{\prod_{k=1}^{K} \Gamma(u_k)}{\Gamma(\sum_{k=1}^{K} u_k)}$, $\psi_k > 0$ und $\sum_{k=1}^{K} \psi_k = 1$

[63] invertierte Wishartverteilung: $W \sim Wishart \Rightarrow \Lambda_k = (W-1) \sim invertiert\ Wishart$

[64] M wird vom Modellierer spezifiziert.

besitzt, relativ flach.

Die Anzahl der Mischungskomponenten wird durch Wählen des Modells mit der größten a posteriori Wahrscheinlichkeit selektiert. Falls die Anzahl der Komponenten a priori gleich wahrscheinlich sind, bildet die Wahl des Modells mit dem größten Bayes Faktor eine äquivalente Methode. Beide Kriterien erfordern allerdings die Berechnung der Randdichte der Daten gegeben die Anzahl der Mischungskomponenten. Hierfür wird häufig die Methode von Gelfand und Dey (siehe [23] und vgl. auch Kapitel 5) verwendet, um die Randdichte aus dem Output der Markovkette zu approximieren.

7.10 Vergleich endlicher Mischungs- mit hierarchischen Bayesmodellen

[65] Andrews, Ansari und Currim führten einen Vergleich zwischen endlichen Mischungsmodellen (FM[66]) und hierarchischen Bayesmodellen (HB[67]) hinsichtlich ihrer relativen Effizienz in Ausdrücken der Anpassung, Vorhersage und Parameterwiedergewinnung durch. Die Gegenüberstellung erfolgte für sorgfältig kontrollierte Simulationen, wobei der Vorteil künstlicher Daten darin liegt, dass die wahren Parameterwerte auf individuellem Niveau bekannt sind, sodass die Differenz zwischen den tatsächlichen und den geschätzten Koeffizienten quantifiziert werden kann.

[65] [4]: „Hierarchical Bayes Versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery"

[66] FM ... Finite Mixture

[67] HB ... Hierarchical Bayes

7.10.1 Endliche Mischungsmodelle

Die Dichtefunktion für Y_i kann als Mischung von Verteilungen

$$H(Y_i; \alpha, X, \beta, \Sigma) = \sum_{k=1}^{K} \alpha_k g(Y_i|X, \beta_k, \Sigma_k)$$

i	... Konsumenten
j	... Wahlalternativen
k	... Komponenten
l	... alternativenbeschreibende Variablen
Y_{ij}	... Reaktion von Konsument i auf Wahlalternative j
Y_i	... Reaktion von Konsument i
X_{jl}	... Wert der Variable l für Alternative j
X_j	... Variablen für Alternative j
β_{lk}	... Koeffizient der Variable l für Komponente k
β_k	... Koeffizienten für Komponente k
Σ_k	... Kovarianzmatrix für Komponente k
$\alpha = (\alpha_1, \ldots, \alpha_k)$	... Mischgewichte

$X = [(X_{jl})]$, $\beta = [(\beta_{lk})]$, $\Sigma = (\Sigma_1, \ldots, \Sigma_K)$

modelliert werden. Man kann die Mischgewichte als Segmentgrößen interpretieren. Sie müssen die Bedingungen $0 < \alpha_k < 1$ und $\sum_k \alpha_k = 1$ erfüllen. Sind die Präferenzen für die Alternativen Y_i normalverteilt, gilt:

$$g(Y_i|X, \beta_k, \Sigma_k) = (2\pi)^{-\frac{J}{2}} |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(Y_i - X\beta_k)' \Sigma_k^{-1} (Y_i - X\beta_k)} .$$

Zwecks Sparsamkeit und hinsichtlich dem rechentechnischen Aufwand kann die Varianz für jede Alternative als gleich angenommen werden.[68]
Für eine Stichprobe von I Konsumenten ergibt sich dann die Loglikelihoodfunktion

$$\ln L = \sum_{i=1}^{I} \ln \left[\sum_{k=1}^{K} \alpha_k g(Y_i|X, \beta_k, \Sigma_k) \right] .$$

[68] $\Sigma_k = \sigma_k^2 I$

Im Anschluss an die Parameterschätzung über die Maximierung der Loglikelihood kann die a posteriori Wahrscheinlichkeit, dass Subjekt i zu Komponente k gehört über

$$\mathbb{P}(i \in k) = \frac{\hat{\alpha}_k g(Y_i|X, \beta_k, \Sigma_k)}{\sum_{k=1}^{K} \hat{\alpha}_k g(Y_i|X, \beta_k, \Sigma_k)}$$

ermittelt werden.
Die Schätzungen der Partworths auf individuellem Niveau lauten dann

$$\hat{\beta}_i = \sum_{k=1}^{K} \mathbb{P}(i \in k)\hat{\beta}_k \ .$$

Zur Bestimmung, wie viele Komponenten sich für eine gegebene Datenmenge eignen, können das BIC oder das CAIC verwendet werden. Hierbei wird die Anzahl an Komponenten so lange erhöht, bis das jeweilige Kriterium minimiert wird.

7.10.2 Hierarchische Bayesmodelle

Bei HB Modellen wird eine kontinuierliche Bevölkerungsverteilung zur Modellierung der Variation in den individuellen Partworths angenommen.
Die Stichprobendichte für Individuum i kann geschrieben werden als

$$f(Y_i; \beta_i, X, \sigma^2) = (2\pi\sigma^2)^{-\frac{J}{2}} e^{-\frac{\sigma^{-2}}{2}(Y_i - X\beta_i)'(Y_i - X\beta_i)}$$

β_i ... Partworths für Individuum i
σ^2 ... Fehlervarianz.

Danach wird eine kontinuierliche, unimodale Bevölkerungsverteilung, typischerweise eine Normalverteilung

$$g(\beta_i; \mu, \Lambda) = \sqrt{2\pi}|\Lambda|^{-\frac{1}{2}} e^{-\frac{1}{2}(\beta_i - \mu)'\Lambda^{-1}(\beta_i - \mu)} \ ,$$

zwecks Spezifizierung der Heterogenität über die Individuen definiert. Der Mittelwertsvektor μ repräsentiert die mittleren Partworths in der Bevölkerung, während die Kovarianzmatrix Λ das Ausmaß an Heterogenität und die Korrelation in den individuellen Partworths beinhaltet.

Für die Inferenz werden Priors für die Hyperparameter μ und Λ, sowie für σ^2 benötigt. Meistens werden eine invertierte Gammaverteilung $IG(a, b)$ als Prior für die Residualvarianz σ^2, eine Wishart Prior $W[\rho, (\rho R)^{-1}]$ für die Präzisionsmatrix Λ^{-1} und eine multivariat normalverteilte Prior $N(\eta, C)$ für den Bevölkerungsmittelwert μ verwendet. Die einzelnen Parameter der Priors sollten so gewählt werden, dass diese nicht-informativ aber trotzdem proper sind.[69]

7.10.3 Modellvergleich

Für den Performancevergleich zwischen FM und HB Modellen wurden folgende Größen eingesetzt:

- der Prozentsatz der erklärten Varianz (R^2) als Maß für die Anpassung;
- die Wurzel der quadrierten Fehler (RMSE) zwischen den wahren und den geschätzten Werten der Partworths

$$RMSE(\beta) = \sqrt{\sum_{i=1}^{I}\sum_{l=1}^{L}\frac{(\hat{\beta}_{li} - \beta_{li})^2}{LI}}$$

L ... Vorhersagen
I ... Individuen

  als Maß für die Parameterwiedergewinnung;
- der RMSE zwischen den beobachteten (Y_{ij}) und den vorhergesagten ($\hat{Y}_{ij}$) Präferenzen für die Holdout Stichprobe

$$RMSE(Y) = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(\hat{Y}_{ij} - Y_{ij})^2}{IJ}}$$

J ... Holdout Profile

  als Maß für die Vorhersagegenauigkeit;
- der Prozentsatz der ersten Choice-Treffer in der Holdout Stichprobe[70] als alternatives Maß für die Vorhersagegenauigkeit.

[69] vgl. Kap. 2.2

[70] Prozentsatz der Subjekte, für die die höchste Präferenz unter den Holdout Stimuli korrekt vorhergesagt wird

Es gab keine signifikanten Unterschiede zwischen FM und HB Modellen hinsichtlich Parameterwiedergewinnung oder Vorhersagegenauigkeit. Auch die Gesamtvorhersagegenauigkeit ($RMSE(\beta)$) betreffend konnten keine wesentlichen Differenzen festgestellt werden. Lediglich bezüglich der Anpassung (R^2-Werte) arbeiten HB signifikant besser als FM Modelle, da bei letzteren das Vorhandensein von Heterogenität innerhalb der Komponenten oft zur Anpassung einer größeren als der wahren Anzahl an Komponenten führt. Außerdem wurden FM und HB Modelle individuellen[71] und einem aggregierten[72] Modell gegenübergestellt. Individuelle Conjoint Modelle passen die Daten gut an, produzieren jedoch schlechte Parameterschätzungen und Vorhersagen, was auf Overfitting hindeutet. Aggregierte Modelle schneiden hinsichtlich der betrachteten Performancemaße signifikant schlechter ab als alle anderen Ansätze. HB und FM Modelle arbeiten signifikant besser als individuelle und aggregierte Modelle in Bezug auf Partworth-Wiedergewinnung, Vorhersagegenauigkeit und Anpassung.

7.11 Multivariate latente Klassenmodelle

[73] In diesem Abschnitt wird das multivariate Regressionsmodell

$$y_i = Z_i\alpha + W_i\beta_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, R_i)$$

y_i	...	T_i Messungen für Subjekt i
Z_i, W_i	...	Designmatrizen für fixe Effekte α bzw. Zufallseffekte β_i
α	...	fixe Effekte (für alle Subjekte gleich)
β_i	...	Zufallseffekte (unterschiedlich für jedes Subjekt).

betrachtet.
Die unbekannte Verteilung der Heterogenität $\pi(\beta_i)$ wird über eine diskrete Verteilung mit unbekanntem Stützvektor $\beta_1^G, \ldots, \beta_K^G$ und unbekannten Gruppenwahrscheinlichkeiten $\eta = (\eta_1, \ldots, \eta_K)$ approximiert

$$\beta_i = \begin{cases} \beta_1^G, & \text{falls } S_i = 1 \\ \vdots & \\ \beta_K^G, & \text{falls } S_i = K, \end{cases}$$

[71] unabhängige Regressionen für jedes einzelne Individuum
[72] ein einziges Modell für alle Individuen
[73] [21] „A Fully Bayesian Analysis of Multivariate Latent Class Models with an Application to Metric Conjoint Analysis"

wobei die latenten Gruppenindikatoren S_i Werte in $\{1, \ldots, K\}$ mit unbekannter Wahrscheinlichkeitsverteilung $\mathbb{P}(S_i = k) = \eta_k$[74] annehmen. Man erhält so eine multivariate Mischung von Normalverteilungen.
Die unbekannten Parameter sind $\phi = (\alpha, \beta_1^G, \ldots, \beta_K^G, \eta, \theta)$, wobei θ in der Definition der Beobachtungsvarianz R_i auftritt. Der latente Gruppenindikator $S^N = (S_1, \ldots, S_N)$ wird als fehlende Dateninformation betrachtet und gemeinsam mit dem Modellparameter ϕ geschätzt.[75]

Die Parameter werden über eine MCMC Methode geschätzt, wofür Priors gewählt werden müssen:

- Bedingt auf ϕ ist die Prior von S^N gegeben durch $\mathbb{P}(S_i = k) = \eta_k$. Die Annahme, dass S_i und S_j paarweise unabhängig sind, führt zu $\pi(S^N|\phi) = \prod_{k=1}^K \eta_k^{N_k}$ mit $N_k = \#\{S_i = k\}$.
- Es wird angenommen, dass η unabhängig von den restlichen Parametern von ϕ ist. Eine natürliche a priori Verteilung $\pi(\eta)$ ist eine Dirichlet Prior $D(e_{01}, \ldots, e_{0K})$. Üblicherweise wählt man $e_{0j} = 1$, was zu einer gleichverteilten Prior auf dem Einheitskreis führt.[76]
- Für α wird eine normale Prior $N(c_0, C_0)$ verwendet.
- $\beta_1^G, \ldots, \beta_K^G$ seien a priori unabhängig. Es ist üblich, hierarchische Priors $(\pi(\beta_k^G) \propto N(b_0, B_0))$ zu verwenden, da sie schwach informativ bezüglich gruppenspezifischer Parameter sind. Diese Prior ist invariant gegenüber Relabelling der Gruppennummern.
- Was die Prior von θ betrifft, ist eine invertiert gammaverteilte Prior $IG(\nu_{\varepsilon,0}, G_{\varepsilon,0})$ eine natürliche Wahl für σ_ε^2, falls $R_i = \sigma_\varepsilon^2 I$.

Unter Verwendung von Bayes' Theorem ist die nichtnormalisierte a posteriori Verteilung $\pi(\phi, S^N|y^N)$ des erweiterten Vektors $\psi = (\phi, S^N)$ proportional zu:

$$\pi(\phi, S^N|y^N) \propto f(y^N|S^N, \phi)\pi(S^N|\phi)\pi(\phi)$$

$$f(y^N|S^N, \phi) = \prod_{i=1}^N f(y_i|\alpha, \beta_{S_i}^G, \theta)$$

[74] η_k ... Klassifikationswahrscheinlichkeiten
[75] Data Augmentation: vgl. Abschnitt 3.2.2
[76] Man kann auch e_{0j} größer als 1 wählen, um leere Klassen a priori auszuschließen.

mit $y^N = (y_1, \ldots, y_N)$. Bedingt auf ϕ und S^N faktorisiert sich die komplette Datenlikelihood $f(y_1, \ldots, y_N | S^N, \phi)$ in ein Produkt von Normalverteilungen. Die komplette Datenlikelihood $f(y^N | S^N, \phi)$ und die Prior $\pi(S^N | \phi)$ sind invariant gegenüber Relabelling. Falls also die Prior $\pi(\phi)$ auch invariant gegenüber Relabelling der Gruppen ist, ist es auch die unbeschränkte Posterior.

Die Vorgangsweise zur Schätzung der Parameter sieht zusammengefasst folgendermaßen aus:

1. unbeschränktes Random Permutation Sampling:[77]
   Es handelt sich dabei um einen unbeschränkten Gibbs Sampler gefolgt von einer zufällig gewählten Permutation $\rho(1), \ldots, \rho(K)$ des momentanen Labellings.
   Der gemeinsame unbekannte Parameter wird in Blöcke gespalten, dann wird aus den bedingten a posteriori Dichten jedes Blocks gegeben die fixierten Werte der anderen Blöcke gezogen:

   (a) Ziehe S^N aus $\pi(S^N | \eta, \alpha, \beta_1^G, \ldots, \beta_K^G, \theta, y^N)$.

   (b) Ziehe η aus $\pi(\eta | S^N)$.

   (c) Ziehe fixe und gruppenspezifische Effekte aus $\pi(\alpha, \beta_1^G, \ldots, \beta_K^G | \theta, S^N, y^N)$. Gemeinsames Ziehen aller Effekte ist möglich, da bedingt auf S^N das latente Klassenmodell als klassisches Regressionsmodell geschrieben werden kann.

   (d) Ziehe die Varianzparameter θ aus $\pi(\theta | \alpha, \beta_1^G, \ldots, \beta_K^G, S^N, y^N)$.

   Danach werden die gruppenabhängigen Parameter permutiert:

$$\begin{aligned} (\beta_1^G, \ldots, \beta_K^G) &:= (\beta_{\rho(1)}^G, \ldots, \beta_{\rho(K)}^G) \\ (\eta_1, \ldots, \eta_K) &:= (\eta_{\rho(1)}, \ldots, \eta_{\rho(K)}) \\ (S_1, \ldots, S_N) &:= (\rho(S_1), \ldots, \rho(S_N)). \end{aligned}$$

   Die permutierten Parameter bilden die Startpunkte für den nächsten Gibbs Schritt. Gruppenunabhängige Parameter wie α und θ werden nicht permutiert.
   Der Sampler erforscht auf diese Art den gesamten Raum der nicht restringierten Posterior, wobei alle Labelling Unterräume mit der gleichen Wahrscheinlichkeit besucht werden.

[77] Schätzung der bezüglich Relabelling invariant bleibenden Parameter

2. Modellselektion:
Die a priori unbekannte Anzahl an Klassen wird durch einen formalen bayesianischen Vergleich der Modelllikelihoods bestimmt, welche die Stützung eines Modell gegeben die Daten quantifiziert. Verschiedene Modelle $\mathcal{M}_1, \ldots, \mathcal{M}_K$ werden über ihre a posteriori Wahrscheinlichkeit

$$\mathbb{P}(\mathcal{M}_l|y^N) \propto f(y_1, \ldots, y_N|\mathcal{M}_l)\mathbb{P}(\mathcal{M}_l)$$

miteinander verglichen.
Die Modelllikelihood $L(y^N|\mathcal{M}_l) := f(y_1, \ldots, y_N|\mathcal{M}_l)$ ist gegeben durch das Integral der marginalen Likelihood $L(y_1, \ldots, y_N|\phi)$ bezüglich der Prior $\pi(\phi)$:

$$L(y^N) = \int L(y_1, \ldots, y_N|\phi)\pi(\phi)d\phi$$

bzw.

$$L(y_1, \ldots, y_N|\phi) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} f(y_i|\beta_k^G, \alpha, \theta)\eta_k \right).$$

Eine Methode zur Berechnung der Modelllikelihood ist das sogenannte Bridge Sampling. Hierfür muss das Modell nicht notwendigerweise identifizierbar sein. Die MCMC Stichprobe, die man durch Random Permutation Sampling erhält, wird mit einer unabhängig identisch verteilten Stichprobe aus einer Wichtigkeitsdichte $q(\phi)$ kombiniert, welche aus dem MCMC Output $(\phi^{(1)}, \ldots, \phi^{(M)})$ des Random Permutation Samplers unter Verwendung einer Mischung aus kompletten Datenposteriors konstruiert wird:

$$q(\phi) = 1/M_L \sum_{m=1}^{M_L} \pi(\phi|(S^N)^{(m)}, \phi^{(m)}, y^N)$$

mit M_L als Anzahl der Mischungen. Bridge Sampling ist robuster bezüglich dem Randverhalten der Wichtigkeitsdichte $q(\phi)$, führt zu präziseren Schätzungen als andere Methoden und liefert das beste Resultat mit dem kleinsten Standardfehler. Es übertrifft mitunter auch die Standardmethode des reziproken Importance Samplings von Gelfand und Dey (siehe [23] und vgl. auch Kapitel 5).[78]

[78] bezüglich Bridge Sampling vgl. auch Abschnitt 5.2.2

3. Suche nach Identifizierbarkeits-Restriktionen:[79]
   Zusätzliche Restriktionen sind notwendig, da das unbeschränkte Modell aufgrund von Label Switching[80] nicht identifizierbar ist. Prinzipiell ist das Modell nämlich nur bis auf Permutationen des Labellings der Gruppen identifizierbar. Der unbeschränkte Parameterraum enthält $K!$ Unterräume mit unterschiedlichem Labelling. Das Label Switching zwischen den Labelling Unterräumen verursacht, dass die unbeschränkte Posterior multiple (maximal $K!$) Modi besitzt. Diese Modi sind äquivalent, falls die Prior $\pi(\phi)$ invariant gegenüber Relabelling der Gruppen ist. Um Funktionale $f(\psi)$ von Koeffizienten $\psi = (\beta^N, S^N, \phi)$, welche nicht invariant gegenüber Relabelling sind, zu schätzen, muss das Modell identifizierbar gemacht werden, d.h. es werden nur Simulationen eines eindeutigen Labelling Unterraums zugelassen.
   Nur eine Restriktion, die die Geometrie der Posterior respektiert, wird diese auf einen Unterraum mit eindeutigem Labelling einschränken. Willkürliche Restriktionen[81] müssen nicht notwendigerweise ein eindeutiges Labelling zur Folge haben. Passende datengetriebene Identifizierbarkeits-Restriktionen kann man durch Untersuchung der MCMC Simulationen aus der unbeschränkten a posteriori Verteilung finden. Man verwendet zu diesem Zweck z.B. Plots der Randdichten der Parameter und/oder Scatterplots von MCMC Simulationen eines nicht-identifizierbaren Modells.[82]
   Jede passende Restriktion ist nur eine indirekte Formulierung, um ein eindeutiges Labelling zu identifizieren und somit nicht notwendigerweise eindeutig.

[79] Schritt 3 erfolgt nur für das laut Schritt 2 „beste" Modell

[80] Hin- und Herspringen zwischen verschiedenen Labelling Unterräumen

[81] Eine derartige Alternative, welche nicht von den Daten gestützt werden muss, stellt die häufig verwendete Standardrestriktion $\eta_1 < \ldots < \eta_K$ dar.

[82] Die Plots können auch als empirischer Indikator zur Festlegung der optimalen Anzahl an Gruppen verwendet werden.

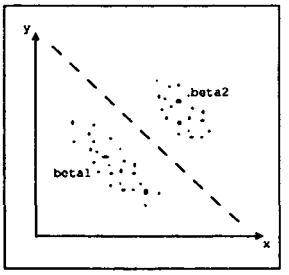

Abbildung 3: Scatterplots der MCMC Simulationen (Quelle: [21])

4. restringiertes Permutation Sampling.[83]

Modellselektion bezieht sich nicht nur auf die Wahl der korrekten Modellstruktur oder der korrekten Anzahl an Gruppen, sondern betrifft auch die Spezifikation einzelner Variablen. Falls die Randdichten einer bestimmten Komponente $\beta^G_{\cdot,r}$ der gruppenspezifischen Parameter $\beta^G_1, \ldots, \beta^G_K$ für alle Gruppen überlappen, kann das auf eher fixe als zufällige Komponenten hindeuten. Man kann auf Heterogenität der Komponenten testen, indem man die Hypothese formuliert, dass eine Komponente $\beta^G_{\cdot,r}$ fest statt zufällig ist. Sie wird gegen die Hypothese einer zufälligen Komponente getestet. Um die Modelllikelihood der neuen Hypothese zu bestimmen, muss die MCMC Schätzung unter der Annahme, dass $\beta^G_{\cdot,r}$ fix ist, noch einmal durchlaufen werden. Die Hypothese wird angenommen, wenn sie eine höhere Modelllikelihood zur Folge hat.
Umfasst die marginale Dichte einer bestimmten Komponente α_s den Wert 0, könnte es sein, dass dieser Effekt nicht signifikant ist. Diese Hypothese wird wiederum gegen die Hypothese, dass α_s im Modell bleiben sollte, getestet. Um die Modelllikelihood der neuen Hypothese zu bestimmen, muss auch in diesem Fall die MCMC Schätzung unter Berücksichtigung von $\alpha_s = 0$ noch einmal durchlaufen werden.

[83] Die Permutation erfolgt entsprechend der in Schritt 3 gefundenen Identifizierbarkeits-Restriktionen.

7.12 Berücksichtigung von Heterogenität in Logit Modellen

[84] Es gibt verschiedene Methoden der Berücksichtigung von Heterogenität unter Verwendung von gemischten Logit (MXL[85]) Modellen.
Die allgemeine Form eines MXL Modells lautet

$$P_n(i) = \int \frac{e^{X_i\beta}}{\sum_{j\in C_n} e^{X_j\beta}} dF(\beta)$$

n ... Individuum
i, j ... Alternativen
C_n ... Menge der für Individuum n verfügbaren Alternativen
X ... erklärende Variablen
β ... zu schätzende Parameter.

Die marginale Wahrscheinlichkeit, dass Individuum n Alternative i wählt berechnet sich als Integral über die Verteilung des Parameters β.

1. Wird für $F(\beta)$ eine kontinuierliche Verteilung mit Dichtefunktion $f(\beta)$ angenommen, erhält man ein kontinuierliches gemischtes Logit (CMXL[86]) Modell:

$$P_n(i) = \int \frac{e^{X_i\beta}}{\sum_{j\in C_n} e^{X_j\beta}} f(\beta) d\beta \ .$$

Zur Berechnung der Wahrscheinlichkeit benötigt man einen Simulationsansatz, der es erlaubt obige Formel durch

$$\tilde{P}_n(i) = \frac{1}{NR} \sum_{r=1}^{NR} \frac{e^{X_i\beta^r}}{\sum_{j\in C_n} e^{X_j\beta^r}}$$

β^r ... r-ter Zufallszug aus der Verteilung $f(\beta)$
NR ... Gesamtanzahl an Zufallszügen

zu ersetzen. Die simulierte Wahrscheinlichkeit wird berechnet, indem man zuerst ein β^r zufällig aus der angenommenen Dichtefunktion $f(\beta)$

[84] [17]: „Comparison of Methods Representing Heterogeneity in Logit Models"
[85] MXL ... Mixed Logit
[86] CMXL ... Continuous Mixed Logit

zieht und dann die Logit Wahrscheinlichkeit bedingt auf die Parameterwerte berechnet. Dieser Vorgang wird NR-mal wiederholt, der Mittelwert davon entspricht dem simulierten Wert der marginalen Wahrscheinlichkeit. Danach können die Parameter der Verteilung $f(\beta)$, sowie die anderen als homogen über die Bevölkerung angenommenen Koeffizienten geschätzt werden.
Der Nachteil bei diesem Ansatz liegt darin, dass eine große Anzahl an Zufallszügen NR notwendig ist, was die Rechenzeit erheblich erhöht. Außerdem ist eine a priori Annahme der Form der Verteilung $f(\beta)$ erforderlich, auf der die Schätzung beruht.
Eine Alternative zum simulierten Maximum Likelihood Schätzer, der die simulierte Likelihood maximiert, stellt die Methode der simulierten Momente dar. Der Schätzer ist konsistent, wenn die Anzahl der Zufallszüge entweder fest ist oder mit jeglicher Rate mit der Stichprobengröße steigt. Er ist zusätzlich effizient und asymptotisch äquivalent zum ML Schätzer, falls die Anzahl der Züge schneller wächst als die Wurzel aus der Stichprobengröße ($\sqrt{N}$).

Momentenmethode (MOM):[87]
Einen ML Schätzer erhält man, indem man die erste Ableitung der Loglikelihoodfunktion gleich null setzt.
Im Allgemeinen sind die Wahlwahrscheinlichkeiten schwierig zu berechnen, was ML Schätzung kompliziert bis unmöglich machen kann. In diesem Fall kann ein MOM Schätzer unter Verwendung der Eigenschaft, dass sich Wahlwahrscheinlichkeiten über einander ausschließende Alternativen auf eins summieren müssen, hergeleitet werden. Ein MOM Schätzer setzt die (gewichtete) Differenz zwischen beobachteter Reaktionen und ihren erwarteten Werten gleich null.
Gegeben optimale Gewichte sind MOM Schätzer asymptotisch äquivalent zu ML Schätzern.

Methode simulierter Momente (MSM):[88]
Hierbei handelt es sich um die Simulationsvariante von MOM, d.h. die Wahlwahrscheinlichkeiten werden simuliert, wenn sie nicht in geschlossener Form dargestellt werden können.

[87] Method of Moments, siehe [18]: Appendix A, [41]: S. 243 - 245
[88] Method of Simulated Moments, siehe [18]: Appendix A, [41]: S. 243 - 246

Ein MSM Schätzer ist asymptotisch äquivalent zum MOM Schätzer, wenn die Anzahl der Zufallszüge mit der Stichprobengröße wächst. Die MSM Schätzungen sind für eine feste Anzahl an Simulationen pro Haushalt konsistent, was bei simulierter ML Schätzung nicht der Fall ist. Allerdings geht die Effizienz verloren, falls nicht-ideale Gewichte verwendet werden.[89] Weiters ist es schwierig, kleine Wahrscheinlichkeiten mittels Simulation exakt zu schätzen.

2. Alternativ dazu kann mit der Hypothese, dass die Verteilung nichtparametrisch in ein oder zwei Dimensionen ist, gearbeitet werden, was zum sogenannten Massepunkt Mischungs-Logit (MPMXL[90]) Modell führt.
Durch Ersetzen der Wahrscheinlichkeitsverteilung $f(\beta)$ durch eine Massepunktverteilung, deren Wahrscheinlichkeitsgewicht des m-ten Massepunkts λ^m ist, und der Integration durch eine Summation über eine feste Anzahl an Massepunkten M erhält man

$$P_n(i) = \sum_{m=1}^{M} \frac{e^{X_i\beta^m}}{\sum_{j\in C_n} e^{X_j\beta^m}} \lambda^m \ .$$

D.h. die Wahrscheinlichkeit, dass Individuum n Alternative i wählt, kann als gewichteter Durchschnitt von M Logit Wahrscheinlichkeiten betrachtet werden. Die Konsumenten werden also über Marktsegmente, innerhalb derer für alle Individuen die gleichen Präferenzen angenommen werden, repräsentiert. Da die Wahrscheinlichkeit in geschlossener Form dargestellt werden kann, ist keine Simulation erforderlich.

Das Modell ähnelt dem latenten Klassenmodell (LCM[91]), da beide Modelle annehmen, dass Heterogenität durch eine endliche Anzahl an Massepunkten repräsentiert werden kann. Allerdings konzentriert sich das MPMXL Modell auf die Verteilung der Zufallsparameter, während die Anzahl der gemeinsamen Massepunkte sowohl über Differenzen in der

[89] Eine Möglichkeit Konsistenz zu erhalten ohne Effizienz zu verlieren bietet der MSS (Method of Simulated Scores) Schätzer (vgl. [41]: S. 246 - 248).
[90] MPMXL ... Mass-Point Mixed Logit
[91] LCM ... Latent Class Model

Lage als auch in den Wahrscheinlichkeitsgewichten für die gemeinsamen Massepunkte definiert wird.
Im Gegensatz dazu richtet das LCM den Fokus auf die Klassen der Bevölkerung und befasst sich nicht mit den unterschiedlichen Werten für jeden einzelnen Parameter. Alle Parameter, die als heterogen angenommen werden, werden zusammengefasst, und es wird versucht, die Anzahl an unterschiedlichen Klassen für die gesamte Menge an Parametern zu finden.
Besitzen das MPMXL Modell und das LCM die gleiche Anzahl an Klassen, sind allerdings üblicherweise beim MPMXL Modell weniger Parameter zu schätzen.

3. Bei der dritten betrachteten Formulierung schließlich werden die Konsumenten innerhalb desselben Segmentes nicht als homogen in ihren Präferenzen angenommen, sondern die Markenpräferenzen der Bevölkerung werden durch die Mischung mehrerer kontinuierlicher Verteilungen beschrieben.
Repräsentiert man die Heterogenität innerhalb der Segmente über Normalverteilungen, erhält man das normale Mischungs-Logit (MNMXL[92]) Modell. Es handelt sich um eine Kombination aus dem CMXL und dem MPMXL Modell, weshalb hier abermals ein Simulationsansatz zur Berechnung der Wahrscheinlichkeit

$$P_n(i) = \sum_{m=1}^{M} \left[\int \frac{e^{X_i \beta^m}}{\sum_{j \in C_n} e^{X_j \beta^m}} f_m(\beta) d\beta \right] \lambda^m$$

notwendig ist. Die Heterogenität innerhalb jedes Marktsegmentes wird durch den Integralterm, die Unterschiede zwischen den Segmenten durch die gewichtete Summe repräsentiert.
Bei diesem Modell erhöht sich zwar der erforderliche Rechenaufwand, jedoch können kompliziertere Heterogenitätsmuster identifiziert werden.

[92] MNMXL ... Mixture-Normal Mixed Logit

8 Experiment

Die überwiegende Anzahl von Autoren verwendet die Mischungs- bzw. die hierarchischen Bayesmodelle im Zusammenhang mit empirischen Daten. Jedoch hat diese Vorgangsweise ihre Mängel. Die erfolgreiche Anwendung auf empirische Daten könnte beispielsweise von der speziell betrachteten Produktklasse abhängen. Funktioniert also ein Modell in einer bestimmten Situation, heißt das noch nicht, dass man diese Ergebnisse einfach verallgemeinern darf. Weiters liefert eine Analyse empirischer Daten keinerlei Information, unter welchen Umständen ein Modell gut arbeitet oder nicht.
Das ist der Grund, warum viele Autoren zusätzlich versuchen, das verwendete Modell mittels Anwendung auf künstlich generierte Daten zu verifizieren. Genauer gesagt generiert man hierbei Daten entsprechend der im Modell getroffenen Grundannahmen, beispielsweise durch Ziehen aus festgelegten Verteilungen. Die Analyse sollte dann wiederfinden, was ursprünglich angenommen wurde.
Es existiert aber auch noch ein dritter Ansatz, der auch mit künstlichen Daten arbeitet, jedoch werden diese mittels einer Simulation generiert. In diesem Fall ist mit Simulation allerdings nicht das Ziehen aus Verteilungen gemeint, sondern die Erzeugung von Daten über ein künstliches Modell, welches reale Marktsituationen nachbilden soll. D.h. statt Parameter zu generieren, produziert man Nutzen- bzw. Choicedaten der Konsumenten, wie sie üblicherweise auch über Konsumentenpanels zur Verfügung stehen. Derartige künstliche Daten entsprechen demzufolge von der Beschaffenheit her empirischen Daten mit dem Unterschied, dass man zusätzlich die Marktstruktur (d.h. die Segmentierung) kennt. Der Vorteil ist jener, dass man z.B. keine restriktiven Annahmen wie bezüglich der Verteilungen bestimmter Variablen treffen muss, sondern einfach gewisse Marktcharakteristika voraussetzt (hier insbesondere betreffend die Segmentierung des Marktes). Mithilfe von Simulationsdaten kann der Grad der Verlässlichkeit eines Modells untersucht werden. Der Vorteil liegt darin, dass die Beschaffenheit des Marktes und alle Annahmen, die vom Experimentator getroffen werden, bekannt sind und kontrolliert werden können. Diese Tatsache erlaubt die Validierung des Modells, indem man die Voraussetzungen und Annahmen, die in der Simulation definiert wurden, mit den Resultaten und der Schätzung des Modells vergleicht. Sowohl das Potential als auch die Grenzen einer Methode oder eines Ansatzes können in diesem Zusammenhang erforscht werden. Der Zweck liegt in der Stärkung des Vertrauens des Forschers in Lösungen, die mit dem Modell oder Algorithmus

gefunden werden, wenn man diese später auf empirische Daten anwendet. Einen weiteren Vorteil bietet die Möglichkeit den Input variieren lassen zu können, was in der Realität nicht immer der Fall sein muss. Beispielsweise bleiben Preise oft über einen langen Zeitraum nahezu unverändert.

In diesem Kapitel soll nun die Performance dreier Ansätze unter Datensätzen unterschiedlicher Beschaffenheit verglichen werden. Betrachtet werden ein endliches Mischungsmodell ohne Zufallseffekte, ein endliches Mischungsmodell mit Zufallseffekten und ein reines Zufallseffektmodell. Die Auswirkungen der Variation verschiedener Experimentalfaktoren (Überlappung der Segmente $\leftrightarrow$ streng separierte Segmente, starke Variation $\leftrightarrow$ schwache Variation der Parameter innerhalb der Segmente) auf die Anwendung dieser Modelle werden näher studiert. Die Vorgangsweise sollte im Folgenden bei der Entscheidung helfen, welches Modell unter welchen speziellen Bedingungen und Voraussetzungen die beste Anpassung liefert. Die Frage besteht hauptsächlich darin, ob es sich lohnt, ein aufwendigeres Modell, das Zufallseffekte innerhalb der Segmente erlaubt und somit zu einer größeren Anzahl an zu schätzenden Parametern führt[93], zu verwenden, oder ob einfachere Modelle (homogene Segmente oder ein reines Zufallseffektmodell) unter gewissen Voraussetzungen ausreichen.

Zur Generierung der Daten wurde ein Simulationsmodell eines künstlichen Konsumentenmarktes verwendet, das im nun folgenden Abschnitt etwas genauer erklärt wird.

8.1 Der künstliche Konsumentenmarkt

Es handelt sich hierbei um die Simulation eines Marktes, der in vier Segmente unterteilt ist. In jedem einzelnen Segment befinden sich Konsumenten mit gleichen Erwartungen und Wünschen hinsichtlich bestimmter Produkteigenschaften. Diese Präferenzen werden mittels eines Segment-Idealpunktes[94] repräsentiert. Es wird angenommen, dass sich die Idealpunkte im Laufe der Zeit nicht ändern und die Konsumenten somit ihr ursprüngliches Segment nicht wechseln. Konkret besteht jede Gruppe aus 250 Konsumenten, es gibt

[93] Es müssen nämlich zusätzlich die Varianzen der Parameter innerhalb jeder Gruppe geschätzt werden.

[94] ASP ... Aspiration

folglich insgesamt 1000 Konsumenten im Markt. Auf der Firmenseite existieren vier Marken, die jeweils eines der Segmente bearbeiten, d.h. sie bewerben nur die entsprechenden Produkteigenschaften, die in diesem Segment von Bedeutung sind und prägen durch Werbung die Markenwahrnehmung[95] der Konsumenten aus. Abb. 4 stellt den Produkteigenschaftsraum graphisch dar.
Die Konsumenten stehen in jedem diskreten Zeitpunkt vor dem Problem, sich für genau eine Marke entscheiden zu müssen. Zu diesem Zweck ziehen sie die preisgewichteten Wahrnehmungen[96] heran. Als Nutzenmaß dient der inverse Abstand zwischen den preisgewichteten Wahrnehmungen und dem entsprechenden Idealpunkt des Konsumenten, wobei speziell die Euklidische Distanz verwendet wird.

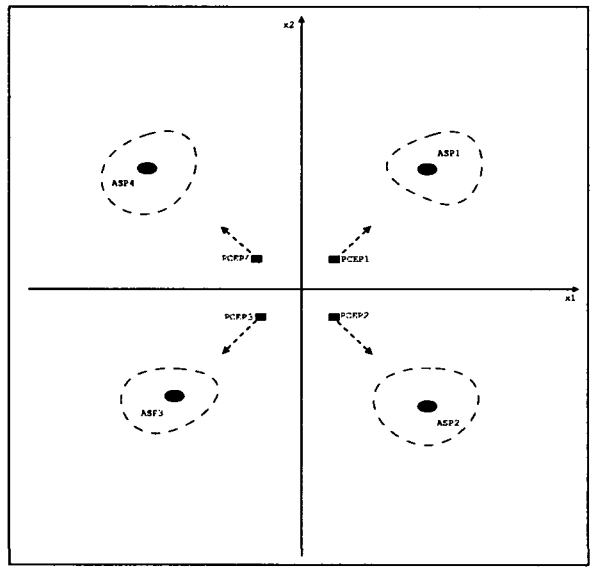

Abbildung 4: Produkteigenschaftsraum (mit xi als Produkteigenschaft, $ASPj$ als Segment-Idealpunkt (Aspiration) und $PCEPk$ als Markenwahrnehmung (Perception))

In einem Flussdiagramm (Abb. 5) wird der Simulationsverlauf noch einmal verdeutlicht.

[95] PCEP ... Perception
[96] ATT ... Attitude

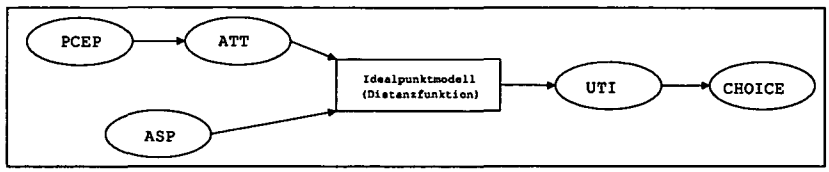

Abbildung 5: Flussdiagramm des Simulationsmodells (Ermittlung der Utilities (UTI) und letztendlich der Choices über die Distanz zwischen Aspirations (ASP) und Attitudes (ATT), welche die preisgewichteten Perceptions (PCEP) bezeichnen)

Die Details und die exakte mathematische Formulierung des künstlichen Konsumentenmarktes können in [40] bzw. [43] nachgelesen werden.

8.2 Das Modell

Geschätzt wurde ein Zufallsnutzenmodell mit den Nutzenwerten der Konsumenten für die vier Marken zu drei verschiedenen Zeitpunkten als Output. Als unabhängige Variablen dienten die Preise und Werbebudgets der Firmen pro beobachtetem Zeitpunkt.
Speziell wird mit folgendem Modell gearbeitet:

$$U_{ijt} = \alpha_i p_{jt} + \beta_i b_{jt} + \varepsilon_{ijt}$$

U_{ijt}	...	Nutzen von Konsument i für Marke j $(j = 1, \ldots, J)$ zur Zeit t $(t = 1, \ldots, T)$
α_i, β_i	...	zu schätzende Parameter
p_{jt}	...	Preis der Marke j zur Zeit t
b_{jt}	...	von Marke j eingesetztes Werbebudget zur Zeit t
ε_{ijt}	...	Störterme.

Die Fehlerterme seien voneinander unabhängig multivariat normalverteilt mit Mittelwert 0 und Kovarianzmatrix $\sigma_i^2 I$.[97] Die logarithmierten Fehlervarianzen werden als zufällige Stichprobe aus einer Normalverteilung mit Mittelwert γ und Varianz τ^2 angenommen.[98]

[97] $I \in (JT \times JT)$
[98] vgl. Abschnitt 7.9

Fasst man die Parameter α_i und β_i in einem Vektor $\theta_i = (\alpha_i, \beta_i)'$ zusammen, besitzt θ_i die Dichte

$$g(\theta_i) = \sum_{k=1}^{K} \psi_k q(\theta_i | \theta_k, \Lambda_k)$$

mit q als multivariate normale Dichte mit Mitelwert θ_k und Kovarianzmatrix Λ_k.

Folgende drei Modelle werden miteinander verglichen:

1. Modell 1:
   Modell 1 bezeichnet ein endliches Mischungsmodell mit fixen Effekten innerhalb der Gruppen, d.h. hier gilt $\Lambda_k = 0$.

2. Modell 2:
   Bei Modell 2 handelt es sich um ein Mischungsmodell mit Zufallseffekten. D.h. der Ansatz unterscheidet sich vom ersten nur dahingehend, dass die Parameter in den Segmenten nicht mehr fix sondern Zufallseffekte sind, d.h. $\Lambda_k \neq 0$.

3. Modell 3:
   Der dritte Ansatz schließlich ist ein reines Zufallseffektmodell ohne diskrete Segmentierung. Hier erfolgt also keine Trennung der Konsumenten in Gruppen. Die Parameter werden als Zufallseffekte modelliert, und der Index k fällt weg[99].

Für den Nutzen sind das Niveau und die Skala irrelevant. Bezüglich dem Niveau kann jede beliebige Konstante addiert werden ohne zu ändern, welche Alternative den höchsten Nutzen bringt. Auch was die Skala betrifft, kann der Nutzen mit jeder beliebigen positiven Konstante multipliziert werden ohne die Ordnung der Alternativen hinsichtlich ihrer Nutzenwerte zu zerstören. Damit das Modell also identifizierbar ist, darf keine Konstante berücksichtigt werden.

Graphische Darstellung des Modells

Mithilfe eines graphischen Modells ist es einfacher, die voll bedingten Verteilungen der einzelnen Parameter zu bestimmen.

[99] $g(\theta_i) = q(\theta_i | \theta, \Lambda)$

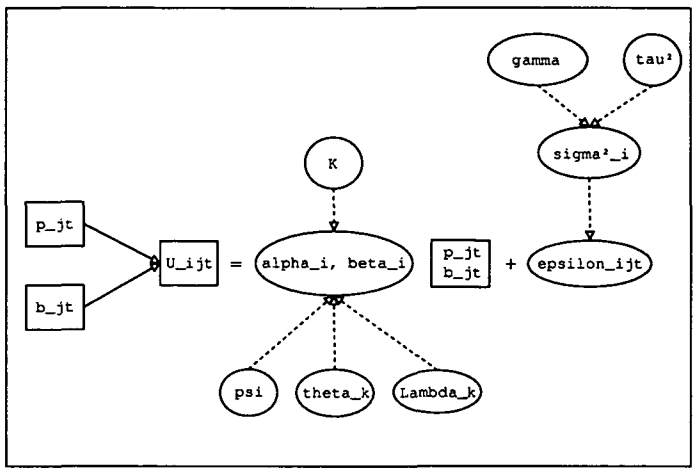

Abbildung 6: Graphische Darstellung des Modells aus 8.2

Das komplette hierarchische Modell kann als gerichteter azyklischer Graph (DAG) dargestellt werden, in dem quadratische Kästchen fixe oder beobachtete Variablen repräsentieren und Kreise für unbekannte Größen stehen.[100] Für letztere müssen beim MCMC Sampling jeweils Startwerte bzw. Priors gewählt werden. Für γ und τ^2 muss man sogenannte Hyperpriors (d.h. Priors für die Parameter der Prior von σ_i^2) festlegen.
Aus obiger Spezifikation ergibt sich die abgebildete graphische Repräsentation des Modells (siehe Abb. 6).

8.3 Ergebnisse

Mithilfe des in Abschnitt 8.1 vorgestellten künstlichen Konsumentenmarktes wurden Simulationsdaten erzeugt und im Anschluss für die Schätzung verwendet. Die Parameter werden mithilfe einer Matlab-Routine zur Schätzung von Mischungsmodellen ermittelt. Speziell handelt es sich um Permutation Sampling (vgl. Abschnitt 7.11), bei dem diverse Restriktionen zur Verhinderung von Label Switching verwendet werden können. Diese Restriktionen sollen vermeiden, dass während des MCMC Samplings die Gruppennummerierung gewechselt wird. Stattdessen sollen sie garantieren, dass man bei einem eindeutigen Labelling bleibt. Hier wurde diesbezüglich nicht die Standard-

[100] vgl. Kap. 6

restriktion der Ordnung der Segmente nach ihrer Größe[101] verwendet, sondern die Permutation erfolgt dahingehend, dass die Parameter entsprechend ihrer ersten Komponente geordnet werden[102]. Diese Restriktion ergab sich aufgrund der Analyse der Scatterplots der von einem unbeschränkten Random Permutation Sampler stammenden Parameterschätzungen.
Weiters wurden immer nur zwei Segmente angenommen. Ein Versuch die Segmentanzahl auf drei zu erhöhen, führte im Wesentlichen nur zu einer Erhöhung der zu schätzenden Parameter ohne allerdings die Performance zu verbessern. Außerdem kam es zu einer Trennung der ersten Gruppe[103] in zwei Untergruppen, während die zweite[104] unverändert blieb. Aus diesem Grund wurde nicht weiters auf die optimale Wahl der Segmentanzahl eingegangen.

Priors werden für die Parameter $\psi = (\psi_1, \ldots, \psi_K)$, θ_k, Λ_k und K, sowie für die beiden Hyperparameter γ und τ^2 benötigt. D.h. diese Variablen stellen im DAG Knoten ohne Eltern dar (vgl. Abb. 6). Die Priors sollten nahezu uninformativ sein, weshalb folgende Wahl getroffen wurde[105]:

- $\psi \sim D(e_{01}, \ldots, e_{0K})$ (Dirichlet) mit $e_{0j} = 1$,
- $\theta_k \sim N(0, 100I)$ (multivariat normal)[106],
- $\Lambda_k \sim IW(f_{0,k}, G_{0,k})$ (invertiert Wishart) mit $f_{0,k} = p + 1$ als Shapeparameter und $G_{0,k} = pI$ als Skalenparameter mit p als Anzahl der Regressoren[107],
- $\gamma \sim N(0, 10)$ (normal),
- $\tau^2 \sim IG(\frac{r_0}{2}, \frac{s_0}{2})$ (invertiert Gamma) mit $\frac{r_0}{2} = \frac{1}{2}$ und $\frac{s_0}{2} = \frac{1}{2}$,
- $K \in \{1, \ldots, M\}$ mit M beliebig (diskret)[108].

[101] Dies erscheint nur sinnvoll, wenn man sich im Voraus sicher ist, dass man es mit unterschiedlich großen Segmenten zu tun hat.
[102] $\alpha_1(1) < \alpha_2(1)$
[103] die Zielgruppe der betrachteten Marke
[104] die Nichtkäufer
[105] Die Priors entsprechen den hierfür üblicherweise verwendeten Verteilungen (vgl. diverse Literaturbeiträge, z.B. [21] oder [32]).
[106] $\theta_k = \theta$ in Modell 3
[107] $\Lambda_k = 0$ in Modell 1 bzw. $\Lambda_k = \Lambda$ in Modell 3
[108] $K = 1$ in Modell 3

Es wurden jeweils die Nutzenwerte für die vier Marken separat geschätzt. Pro Datensatz[109] und Modell[110] waren somit vier MCMC Schätzungen notwendig. Es werden lediglich die Ergebnisse zweier Marken pro Datensatz vorgestellt bzw. die Plots nur von einer Marke repräsentativ für alle anderen dargestellt, da sich die Resultate kaum zwischen den Marken unterscheiden, vor allem was die Performance und die Wahl des am besten geeigneten Modells betrifft.

Der Vergleich kurzer Simulationsfolgen mit unterschiedlichen Startwerten ergab keine wesentlichen Unterschiede, weshalb in den folgenden Berechnungen jeweils nur eine langfristige Simulationsfolge betrachtet wurde. Anhand dieser wurde mithilfe von Autokorrelations- und Trace Plots nun das Konvergenz- und Mischverhalten der Simulationen untersucht.[111] Weiters wurde zusätzlich eine Matlab-Routine („raftery") zwecks Diagnostik der erforderlichen Anzahl an Iterationen bzw. der vorgeschlagenen Länge der Burn-In Phase eingesetzt. Die Funktion „raftery" beruht auf einem Paper von Raftery und Lewis[112] und benötigt lediglich die Gibbs Iterationen eines Durchlaufs des Samplers, weitere externe Spezifikationen der Charakteristiken der Posterior sind nicht notwendig.
Alle MCMC Schätzungen wurden mit 1000 Iterationen und einer Burn-In Phase von 500 Iterationen durchgeführt. Für den Großteil der Schätzungen war diese willkürlich gewählte Laufzeit auch mehr als hinreichend, lediglich in vereinzelten Fällen deuten die Trace Plots sowie die Autokorrelationen darauf hin, dass mehr Iterationen notwendig wären.

Auch die Segmentierung wird graphisch dargestellt. Da man ja aufgrund der Simulationsdaten weiß, welche Konsumenten welchem Segment angehören, kann man untersuchen, wie viele davon mittels der MCMC Schätzung richtig zugeordnet werden konnten. Geplottet wurde hierbei lediglich die Information der letzten 100 Iterationen.

Die Selektion des besten Modells erfolgte mithilfe des über Bridge Samp-

[109] (1) homogene Segmente, (2) heterogene Segmente mit kleiner Varianz, (3) heterogene Segmente mit großer Varianz, (4) überlappende Segmente

[110] (1) fixe Effekte pro Segment, (2) Zufallseffekte pro Segment, (3) Zufallseffekte ohne diskrete Segmentierung

[111] CUSUM-Plots sind nicht notwendig, weil ohnehin nur vier Parameter (zwei pro Segment) geschätzt werden.

[112] siehe [38]: „How Many Iterations in the Gibbs Sampler?"

ling[113] berechneten Logarithmus der Modelllikelihood (Modellloglikelihood). Auch anhand des RMSE[114] wurden die Ansätze miteinander verglichen und beurteilt.

8.3.1 Simulationsdaten 1 (Homogene Segmente)

Betrachtet werden zuerst homogene Segmente, d.h. innerhalb eines Segmentes besitzen alle Konsumenten den gleichen Idealpunkt (Segment-Idealpunkt), die tatsächlichen Parameter sind also für alle Konsumenten derselben Gruppe identisch (siehe Abb. 7).

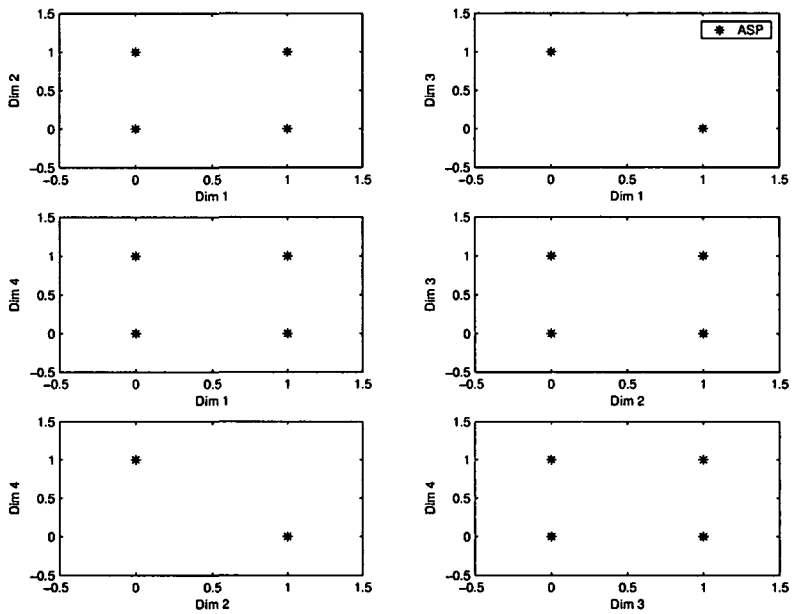

Abbildung 7: Darstellung der Aspirations für alle Attributdimensionen für homogene Segmente (Dim i bezeichnet das jeweilige Markenattribut im Produkteigenschaftsraum)

[113] siehe Abschnitt 5.2.2 bzw. vgl. [21]
[114] bezüglich des Fehlers über die Iterationen

Parameter- und Varianzschätzungen

Wie man aus den Tabellen 2 und 3 ablesen kann, unterscheiden sich die Parameterwerte[115] sowie ihre Standardabweichungen kaum zwischen Modell 1 und 2. In Modell 3 liegen die Schätzungen sowohl der Preis- als auch der Budgetkoeffizienten zwischen den Werten in den beiden Segmenten von Modell 1 bzw. 2, die Standardabweichungen sind größer als jene in den anderen beiden Modellen.
Für Marke 1 sind die Parameter sowie die Standardabweichungen in Segment 2 betragsmäßig kleiner, d.h. die Marketing Mix Variablen üben offensichtlich im Segment der „Nichtkäufer" eine geringere Wirkung auf die Konsumenten aus. Die zu Segment 1 gehörenden Konsumenten reagieren stärker auf Preisänderungen bzw. erhöhten Werbeeinsatz der sie ansprechenden Marke.
Bei der zweiten Marke fällt auf, dass die Parameter betragsmäßig kleinere Werte als bei Marke 1 annehmen und der Preiskoeffizient in Segment 2 sogar positiv wird.
Im Vergleich zwischen Preis und Budget sind die Budgetkoeffizienten ihrem Betrag nach deutlich kleiner als jene der Preise.

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	-2.1641 (0.0187)	-2.1629 (0.0189)	-0.7019 (0.0271)
Marke 1, Segment 2	-0.2123 (0.0112)	-0.2135 (0.0107)	
Marke 2, Segment 1	-0.1079 (0.0082)	-0.1079 (0.0090)	0.1201 (0.0058)
Marke 2, Segment 2	0.1963 (0.0051)	0.1962 (0.0052)	

Tabelle 2: Parameterwerte für den Preis, Standardabweichung in Klammer

[115] α bezeichne einen Parameter des Interesses, $\alpha^{(i)}$ den Zug in der i-ten Iteration und y die Daten. Dann kann der a posteriori Mittelwert einer Funktion g von α mit

$$\mathbb{E}(g(\alpha)|y) = \frac{1}{n-m} \sum_{i=m+1}^{n} g(\alpha^{(i)})$$

geschätzt werden, wobei nur die letzten $n - m$ von n Iterationen verwendet werden. m bezeichnet somit die Anzahl der zur Burn-In Phase gehörenden Iterationen.

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	685 (4.0609)	684 (4.1573)	272 (7.4841)
Marke 1, Segment 2	134 (2.3989)	134 (2.3201)	
Marke 2, Segment 1	212 (1.9120)	212 (2.0874)	104 (2.0671)
Marke 2, Segment 2	67 (1.1896)	67 (1.2223)	

Tabelle 3: Parameterwerte für das Budget, Standardabweichung in Klammer (alle Werte $*10^{-4}$)

Wie man anhand der Trace Plots der MCMC Iterationen der Parameterwerte (Abb. 8) sieht, wurden hinreichend viele Iterationen bzw. eine ausreichend lange Burn-In Phase gewählt, sodass die Schätzungen bereits konvergieren. Eine Betrachtung der Autokorrelationsplots (siehe Abb. 9) zeigt eine rasche Dämpfung, was auf ein schnelles Mischverhalten schließen lässt. Wiederum führt dies zum Schluss, dass bei der MCMC Schätzung genügend lange simuliert wurde.

Die Abbildungen 10 bzw. 11 und 12 bzw. 13 indizieren ebenfalls wie die Grafiken 8 bzw. 9 erreichte Konvergenz bzw. rasches Mischverhalten.

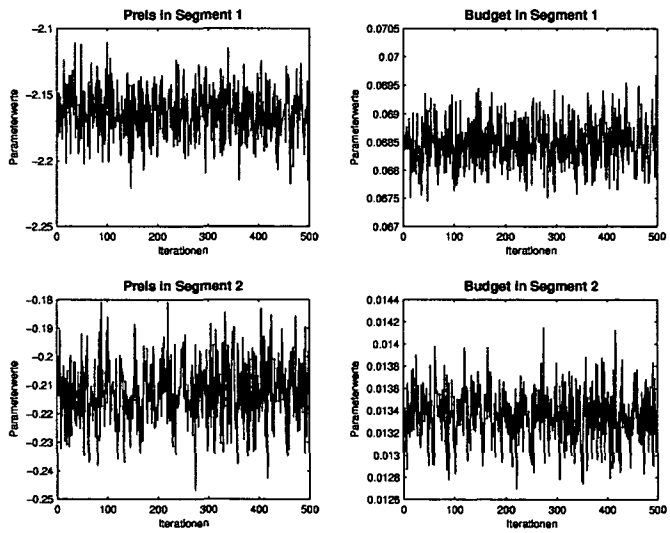

Abbildung 8: Trace Plots der MCMC Iterationen der Parameter (Modell 1)

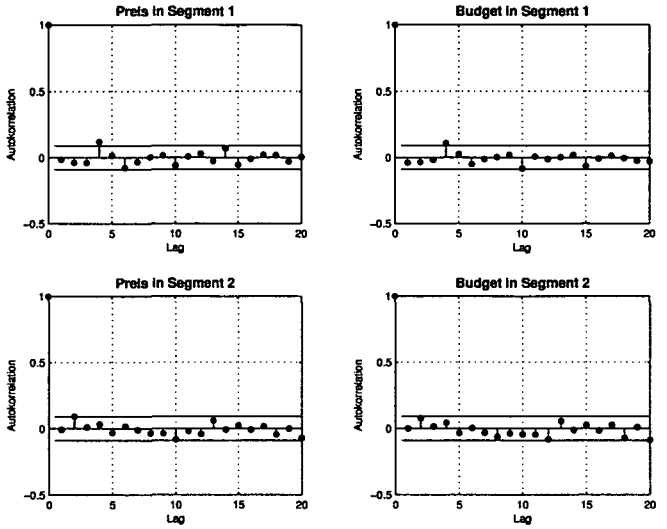

Abbildung 9: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 1)

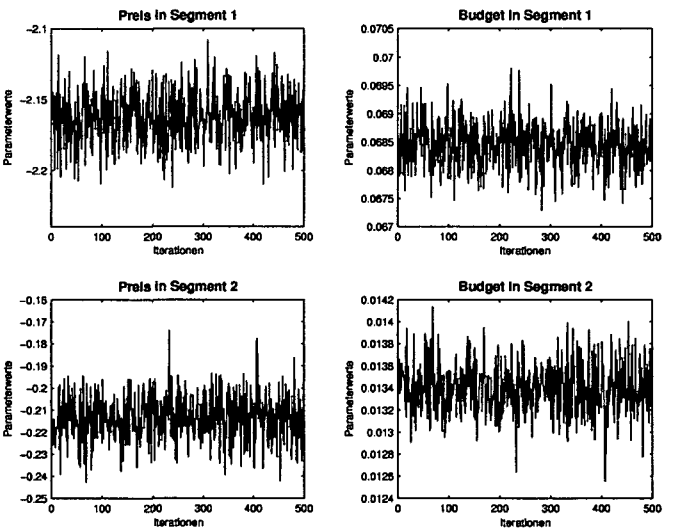

Abbildung 10: Trace Plots der MCMC Iterationen der Parameter (Modell 2)

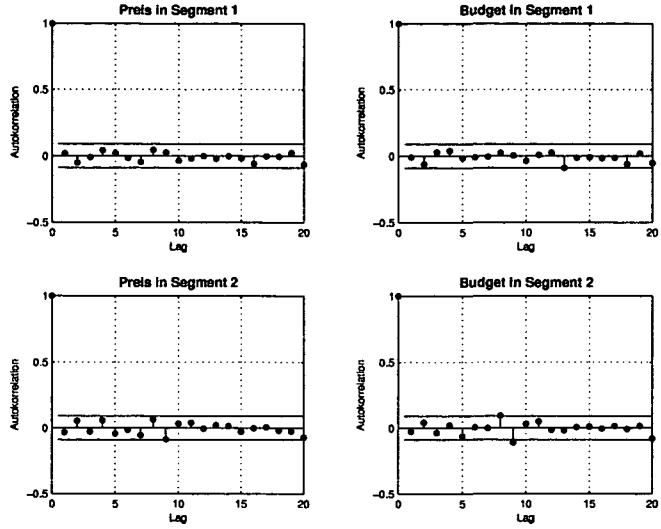

Abbildung 11: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 2)

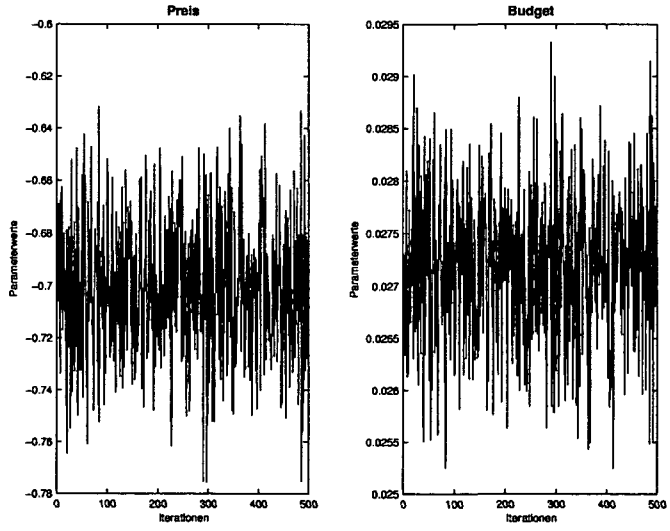

Abbildung 12: Trace Plots der MCMC Iterationen der Parameter (Modell 3)

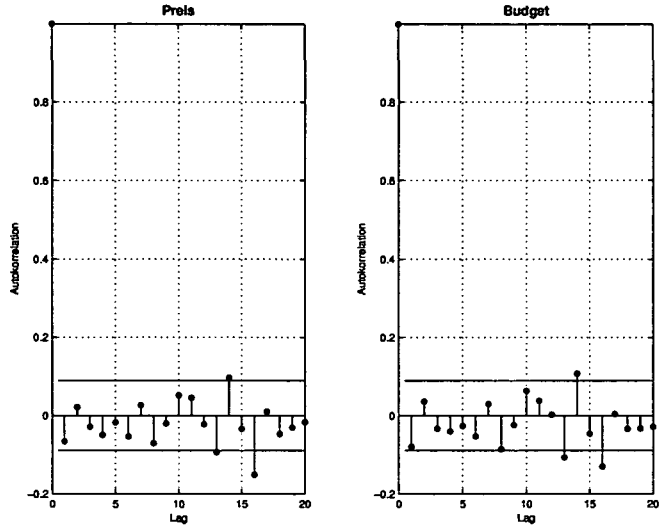

Abbildung 13: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 3)

Wie man anhand der Tabellen 4 und 5 sieht, ist für die erste Marke die Varianz in Segment 2, für die zweite Marke jene in Segment 1 größer. Weiters übertreffen die Varianzen in Modell 3 deutlich jene im zweiten Modell, die dort relativ klein ausfallen.

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	0	3.2863	6294
Marke 1, Segment 2	0	5.8729	
Marke 2, Segment 1	0	3.3989	145
Marke 2, Segment 2	0	0.1525	

Tabelle 4: Parameterwerte der Preisvarianz (alle Werte $*10^{-4}$)

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	0	0.5052	5137.4
Marke 1, Segment 2	0	1.9218	
Marke 2, Segment 1	0	1.4709	347.6
Marke 2, Segment 2	0	0.4059	

Tabelle 5: Parameterwerte der Budgetvarianz (alle Werte $*10^{-7}$)

Modell 1 ist jenes mit diskreten Segmenten, in denen keine Variation der Parameter über die Gruppenmitglieder erlaubt ist, weshalb der Wert der Varianz 0 beträgt. Das kann man auch in den Trace Plots (Abb. 14) und den Autokorrelationen (Abb. 15) sehen.

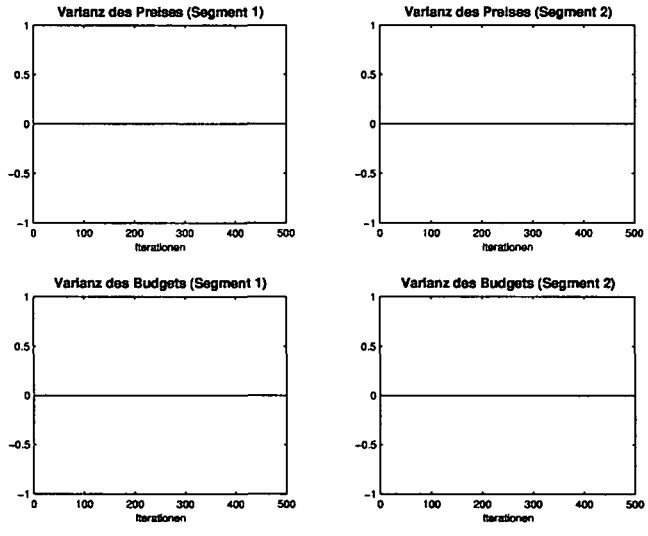

Abbildung 14: Trace Plots der MCMC Iterationen der Parametervarianzen (Modell 1)

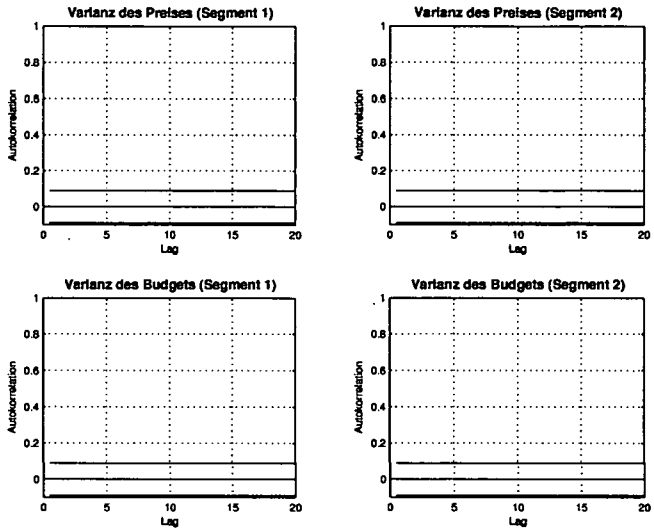

Abbildung 15: Autokorrelationsplots der MCMC Iterationen der Parametervarianzen (Modell 1)

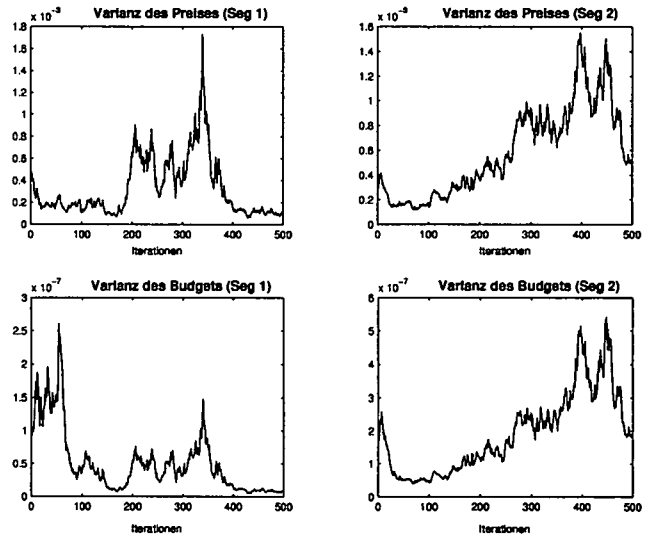

Abbildung 16: Trace Plots der MCMC Iterationen der Parametervarianzen (Modell 2)

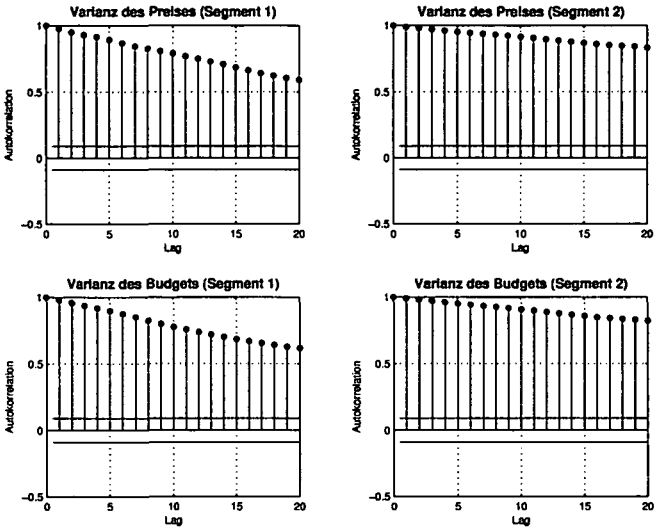

Abbildung 17: Autokorrelationsplots der MCMC Iterationen der Parametervarianzen (Modell 2)

In Abbildung 16 wird deutlich, dass die ad hoc gewählte Anzahl an Iterationen nicht ausreichend war, um Konvergenz zu erreichen bzw. zeigt der Autokorrelationsplot (Abb. 17), dass langsames Mischen erfolgt und somit eine längere Laufzeit der MCMC Simulation notwendig wäre. Die Anwendung der Funktion „raftery“ bestätigt diese Vermutung und empfiehlt eine Mindestanzahl an Iterationen von ca. 2500 für die Preisvarianz in Segment 1 und die Budgetvarianz in beiden Segmenten bzw. sollten für die Preisvarianz in Segment 2 sogar mehr als 7000 statt nur 1000 Iterationen gewählt werden.

Im dritten Modell scheint mit 1000 Iterationen wiederum eine genügend lange Laufzeit gewählt worden zu sein, um Konvergenz zu erreichen (vgl. Abb. 18 und 19), was sich auch bei Anwendung der Matlab-Routine „raftery“ bestätigt.

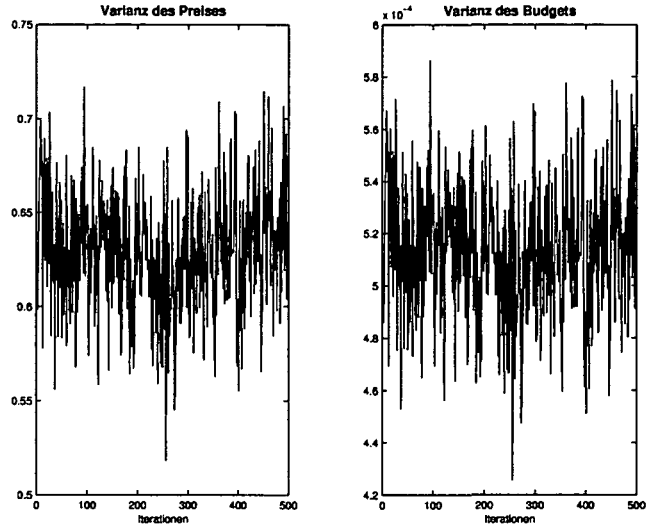

Abbildung 18: Trace Plots der MCMC Iterationen der Parametervarianzen (Modell 3)

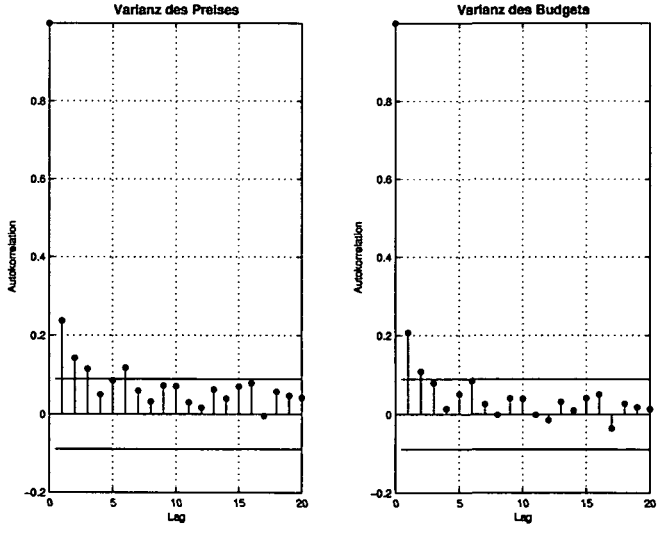

Abbildung 19: Autokorrelationsplots der MCMC Iterationen der Parametervarianzen (Modell 3)

Zwecks Bewertung der Performance der drei Modelle wird neben der Modellloglikelihood (siehe später) auch der Fehler in den MCMC Iterationen betrachtet. Ein Vergleich der RMSEs (siehe Tab. 6) zeigt, dass diesbezüglich Modell 1 am besten arbeitet dicht gefolgt von Modell 2 bzw. schneiden beide Ansätze nahezu gleich gut ab. Modell 3 arbeitet hinsichtlich des RMSE am schlechtesten.

	Modell 1	Modell 2	Modell 3
Marke 1	0.3115	0.3125	0.4453
Marke 2	0.0631	0.0631	0.0777

Tabelle 6: RMSE

Ein direkter Vergleich der Fehler anhand ihrer Trace Plots bzw. Histogramme illustriert obige Beobachtung. Die Fehlerterme für Modell 1 und 2 sind annähernd gleich (vgl. Abb. 20 mit Abb. 21), Modell 3 hingegen weist einen deutlich größeren Fehler auf (siehe Abb. 22).

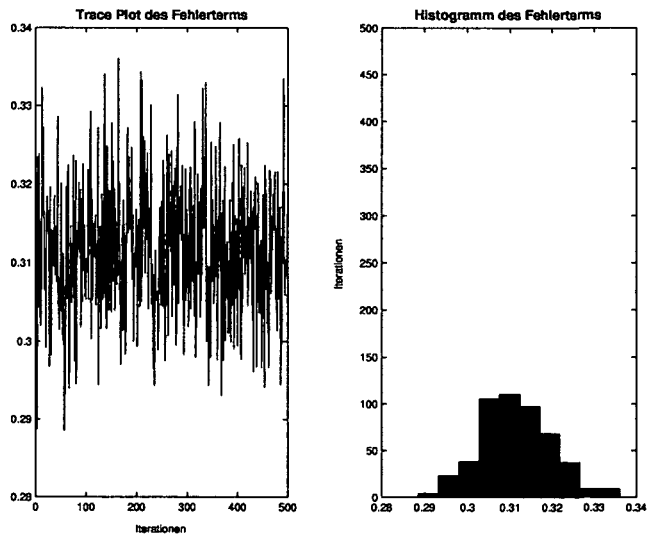

Abbildung 20: Trace Plot und Histogramm des Fehlerterms in den MCMC Iterationen (Modell 1)

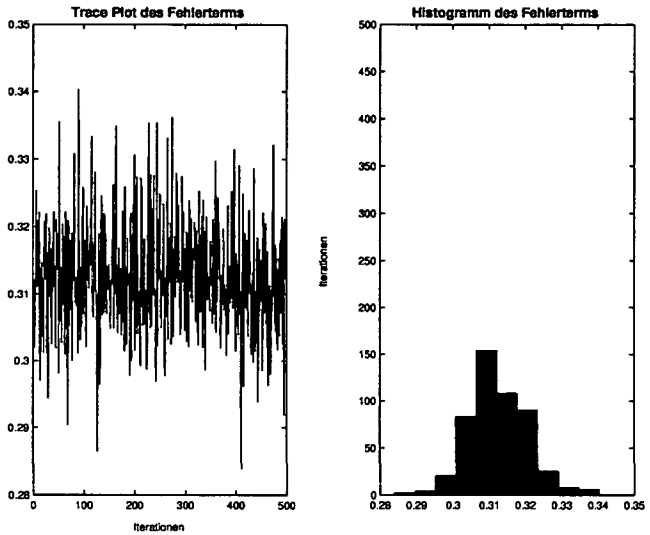

Abbildung 21: Trace Plot und Histogramm des Fehlerterms in den MCMC Iterationen (Modell 2)

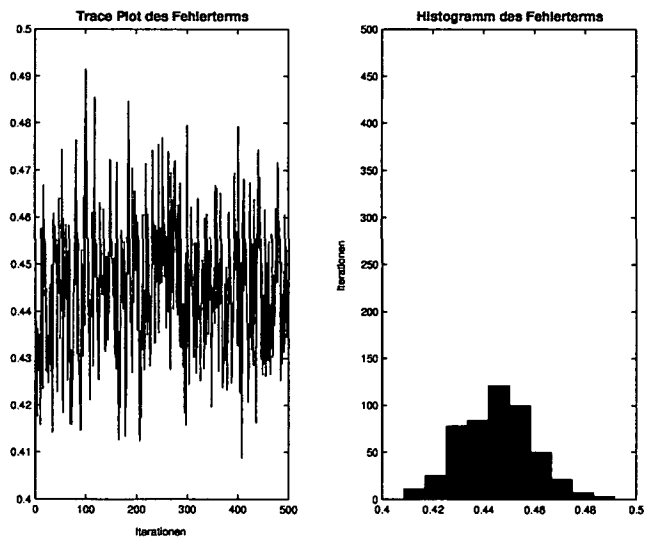

Abbildung 22: Trace Plot und Histogramm des Fehlerterms in den MCMC Iterationen (Modell 3)

Segmentierung

Zwecks Beurteilung der Modelle bezüglich ihrer Performance in Sachen Segmentierung wurde der Prozentsatz der richtigen Zuordnungen der Konsumenten zu ihren Gruppen berechnet.

	Modell 1	Modell 2
Marke 1, richtig	100	100
Marke 1, tw. richtig	0	0
Marke 1, falsch	0	0
Marke 2, richtig	100	100
Marke 2, tw. richtig	0	0
Marke 2, falsch	0	0

Tabelle 7: Trefferwahrscheinlichkeit in Segment 1 (in %)

	Modell 1	Modell 2
Marke 1, richtig	100	100
Marke 1, tw. richtig	0	0
Marke 1, falsch	0	0
Marke 2, richtig	100	100
Marke 2, tw. richtig	0	0
Marke 2, falsch	0	0

Tabelle 8: Trefferwahrscheinlichkeit in Segment 2 (in %)

Genauer gesagt wurde überprüft, wie viele der Konsumenten richtig[116], teilweise richtig[117] oder falsch[118] zugeteilt wurden.

[116] Konsumenten, die in allen MCMC Iterationen dem richtigen Segment zugeordnet wurden

[117] Konsumenten, die nicht ausnahmslos in allen MCMC Iterationen dem richtigen Segment zugeordnet wurden

[118] Konsumenten, die in allen MCMC Iterationen dem falschen Segment zugeordnet wurden

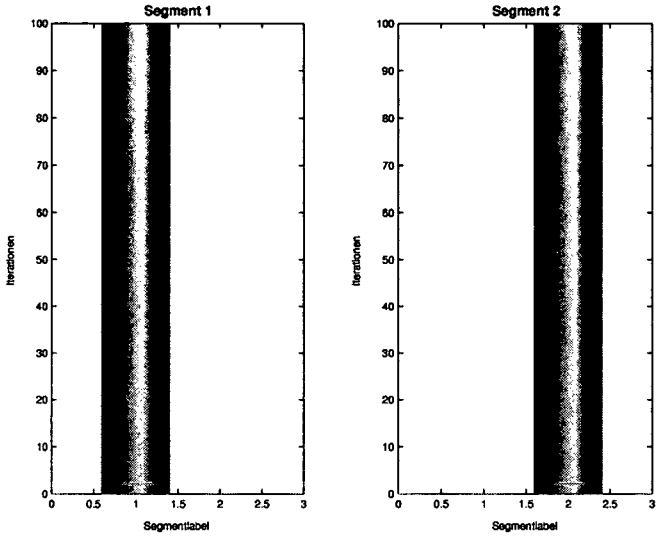

Abbildung 23: Segmentzuordnung der Konsumenten (Modell 1)

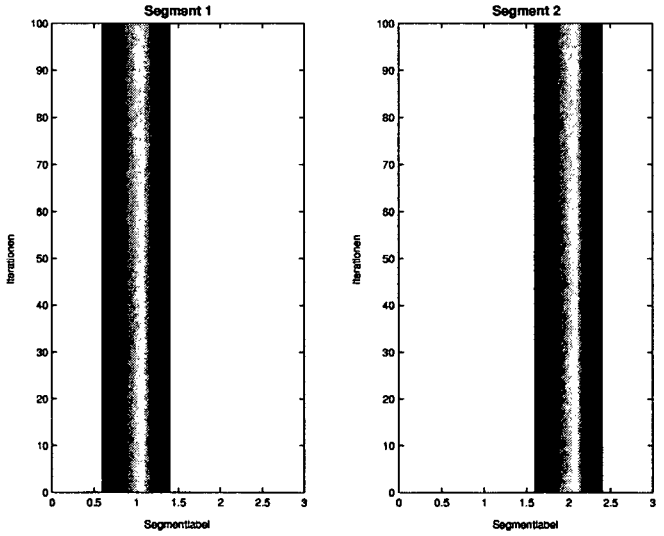

Abbildung 24: Segmentzuordnung der Konsumenten (Modell 2)

Wie man in den Tabellen 7 und 8, sowie den Abbildungen 23 und 24[119] erkennen kann, sind beide Modellansätze in der Lage alle Konsumenten ausnahmslos ihren richtigen Segmenten zuzuordnen. Da dies im Falle von Modell 1 mit weniger Parametern und dadurch mit einem geringeren Rechenaufwand möglich ist[120], scheint rein intuitiv dieses Modell effizienter zu sein.

	Modell 1	Modell 2	Modell 3
Marke 1	-3.1080	-3.3823	-4.3596
Marke 2	-0.7115	-1.1130	-1.5118

Tabelle 9: Modellloglikelihoods (alle Werte $*10^3$)

Eine Modellselektion auf Basis der Modellloglikelihoods bekräftigt obige Intuition (vgl. dazu Tab. 9). Die Reihung der Modelle entspricht jener aufgrund der RMSE-Werte.

Da also die Performance der Modelle 1 und 2 bezüglich Segmentierung gleich gut ist, allerdings Modell 1 eine etwas höhere Loglikelihood hat, einen kleineren RMSE besitzt und zusätzlich weniger Parameter zu schätzen sind, sollte für Datensätze mit homogenen Segmenten optimalerweise Modell 1 eingesetzt werden.

[119] In der linken Grafik (Segment 1) ist die Zuordnung korrekt, wenn die Konsumenten das Label 1 bekommen, in der rechten Grafik (Segment 2) lautet das richtige Label 2. Ein Balken bei Label 2 in der linken Abbildung (bzw. ein Balken bei Label 1 in der rechten Abbildung) repräsentiert Konsumenten, die laut Simulationsmodell eigentlich zu Segment 1 (bzw. zu Segment 2) gehören, vom Sampler allerdings dem falschen Segment zugeordnet wurden.

[120] Man erspart sich nämlich die Schätzung der Varianzen innerhalb der Segmente.

8.3.2 Simulationsdaten 2 (Schwache Heterogenität)

Im zweiten betrachteten Simulationsdatensatz existieren heterogene Segmente, allerdings streuen die individuellen Idealpunkte nur schwach um den Segmentidealpunkt.

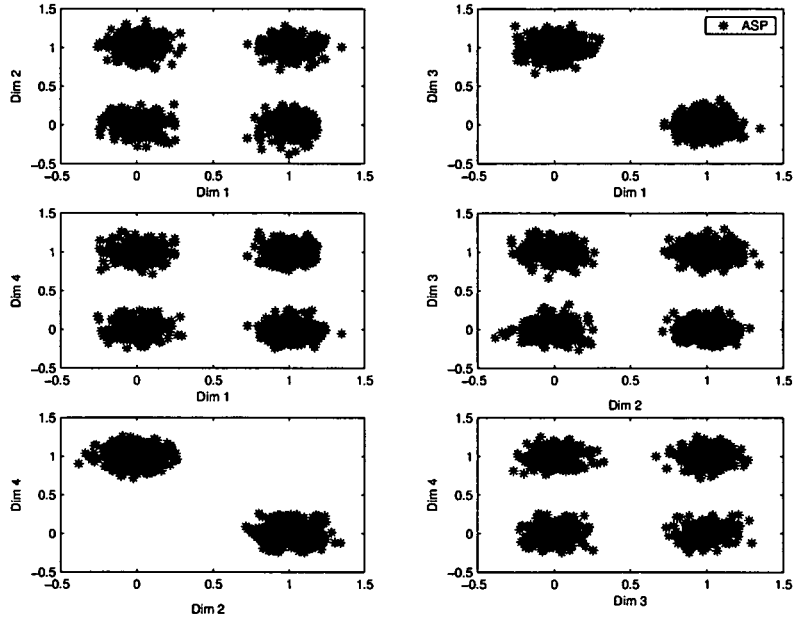

Abbildung 25: Darstellung der Aspirations für alle Attributdimensionen bei schwacher Heterogenität (Dim i bezeichnet das jeweilige Markenattribut im Produkteigenschaftsraum)

Parameter- und Varianzschätzungen

Die untenstehenden Tabellen (Tab. 10 und 11) zeigen, dass Modell 2 betragsmäßig etwas kleinere Preis- und Budgetwerte als Modell 1 liefert. In Segment 1 sind die Standardabweichungen größer, in Segment 2 kleiner als im ersten Modell. Außerdem weist die zweite Marke betragsmäßig größere Werte und höhere Standardabweichungen als Marke 1 auf. Die Parameterwerte in Modell 3 liegen wieder zwischen jenen der beiden Segmente in Modell 1 bzw. 2.

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	-0.9545 (0.0151)	-0.9413 (0.0192)	-0.2765 (0.0136)
Marke 1, Segment 2	-0.0533 (0.0083)	-0.0509 (0.0060)	
Marke 2, Segment 1	-2.9514 (0.0492)	-2.5868 (0.1008)	-0.5647 (0.0451)
Marke 2, Segment 2	0.0502 (0.0235)	0.1289 (0.0101)	

Tabelle 10: Parameterwerte für den Preis, Standardabweichung in Klammer

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	36.9 (0.2665)	36.6 (0.5143)	17.3 (0.3815)
Marke 1, Segment 2	10.8 (0.1502)	10.7 (0.1154)	
Marke 2, Segment 1	103.7 (1.2000)	92.7 (3.0000)	30.1 (1.4000)
Marke 2, Segment 2	11.1 (0.5545)	8.7 (0.2346)	

Tabelle 11: Parameterwerte für das Budget, Standardabweichung in Klammer (alle Werte $*10^{-3}$)

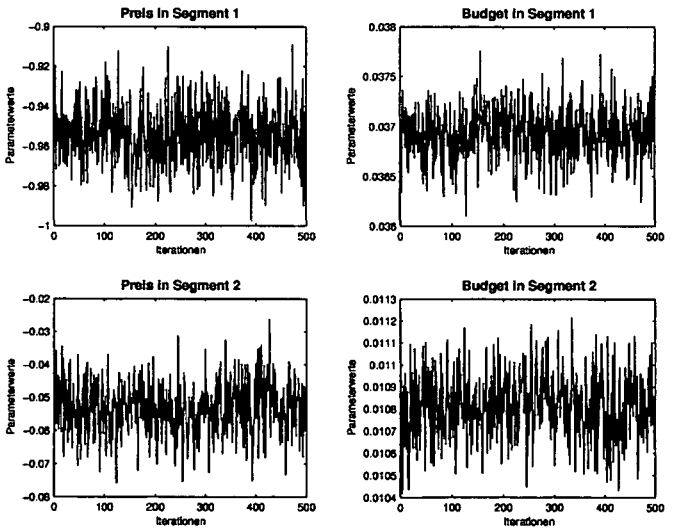

Abbildung 26: Trace Plots der MCMC Iterationen der Parameter (Modell 1)

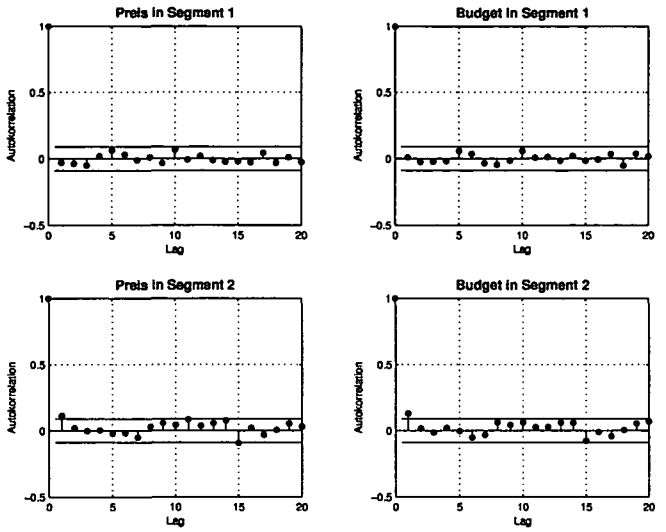

Abbildung 27: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 1)

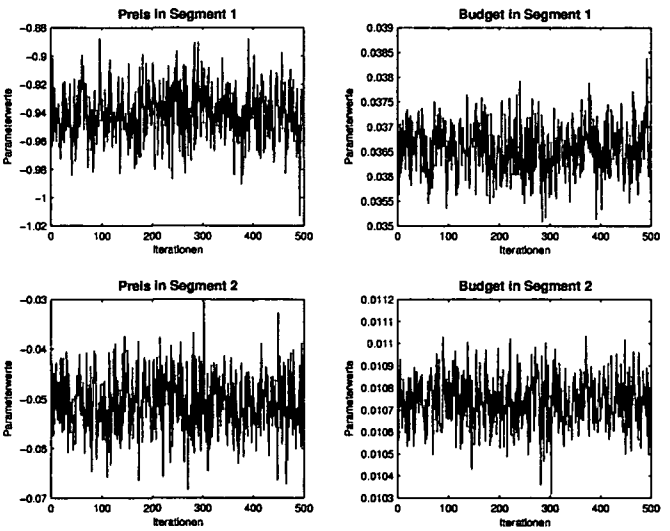

Abbildung 28: Trace Plots der MCMC Iterationen der Parameter (Modell 2)

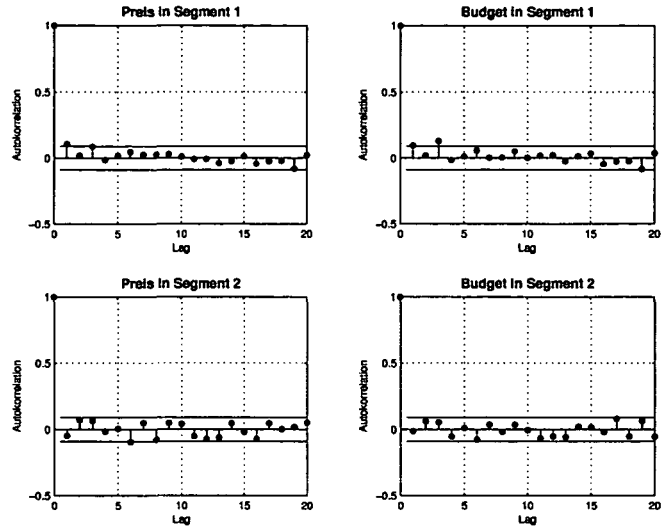

Abbildung 29: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 2)

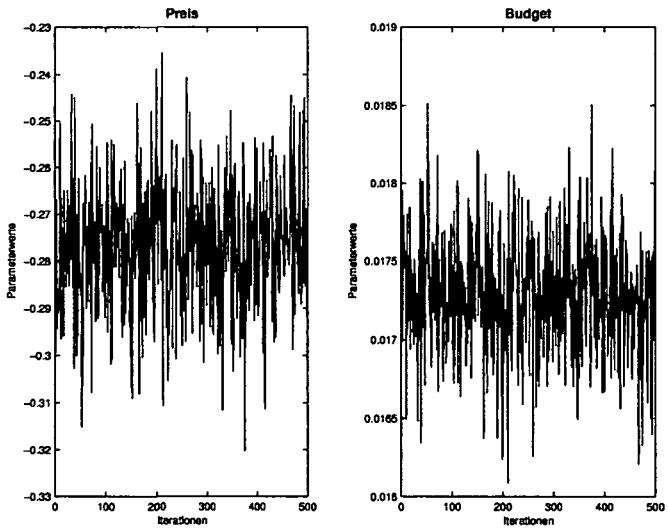

Abbildung 30: Trace Plots der MCMC Iterationen der Parameter (Modell 3)

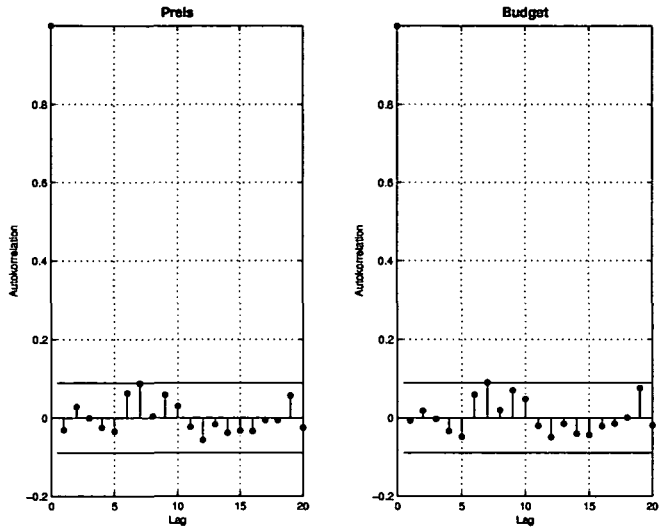

Abbildung 31: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 3)

Die obigen Abbildungen bestätigen, dass alle drei Modelle gutes Konvergenz- und Mischverhalten aufweisen und man daher folgern kann, dass die MCMC Simulation nach 1000 Iterationen bereits lange genug gelaufen ist.

Ein Vergleich der Parameterwerte der Preis- und Budgetvarianzen (siehe Tab. 12 und 13) ergibt, dass für die erste Marke die Varianzwerte des dritten Modells um einiges höher sind als jene von Modell 2. Allerdings ist der Unterschied nicht mehr so enorm wie im Falle homogener Segmente. Die Werte der ersten Marke sind außerdem deutlich kleiner als jene der zweiten Marke. Weiters erkennt man, dass die Varianz im Segment der potentiellen Käufer im Vergleich zum Nichtkäufer-Segment deutlich steigt und wesentlich höhere Werte annimmt.

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	0	63.0000	158.2
Marke 1, Segment 2	0	1.1000	
Marke 2, Segment 1	0	2229.5000	1945.2
Marke 2, Segment 2	0	0.1915	

Tabelle 12: Parameterwerte der Preisvarianz (alle Werte $*10^{-3}$)

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	0	564.0300	1384.9
Marke 1, Segment 2	0	19.7420	
Marke 2, Segment 1	0	20000.0000	18000.0
Marke 2, Segment 2	0	0.6725	

Tabelle 13: Parameterwerte der Budgetvarianz (alle Werte $*10^{-7}$)

Die Trace Plots bzw. Autokorrelationen der Varianzen für Modell 1 werden nicht mehr präsentiert, da bei diesem Ansatz keine Variation der Parameter innerhalb der Klassen erlaubt ist und deshalb die Plots immer den Abbildungen 14 und 15 entsprechen.
Was Modell 2 betrifft, erkennt man anhand der Trace Plots und Autokorrelationen, dass in Segment 1 nach 1000 Iterationen bereits Konvergenz erreicht wurde, in Segment 2 allerdings weitere Iterationen notwendig wären[121].
Modell 3 zeigt hingegen wie Modell 1 sowohl gutes Konvergenz- als auch rasches Mischverhalten.

[121] nämlich laut „raftery“ ca. 1300

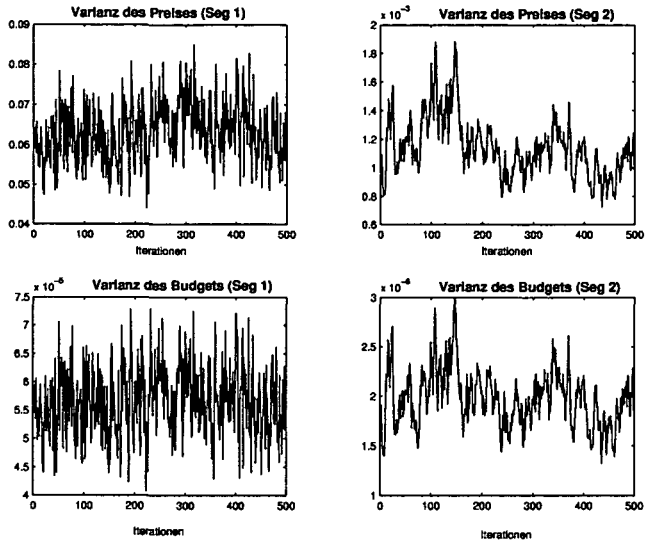

Abbildung 32: Trace Plots der MCMC Iterationen der Parametervarianzen (Modell 2)

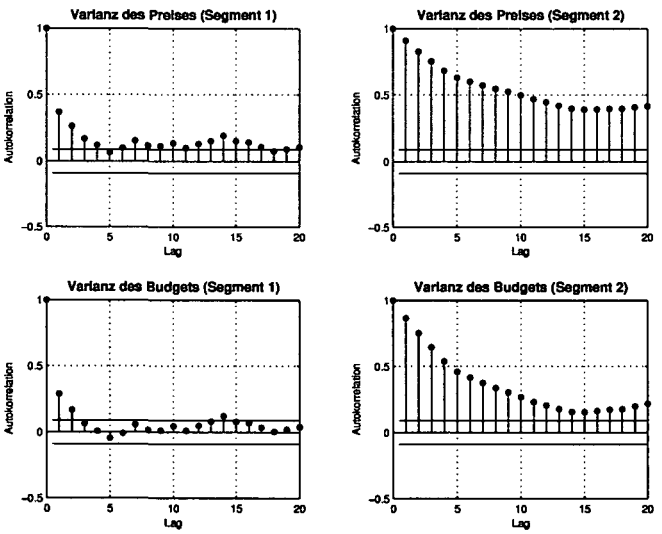

Abbildung 33: Autokorrelationsplots der MCMC Iterationen der Parametervarianzen (Modell 2)

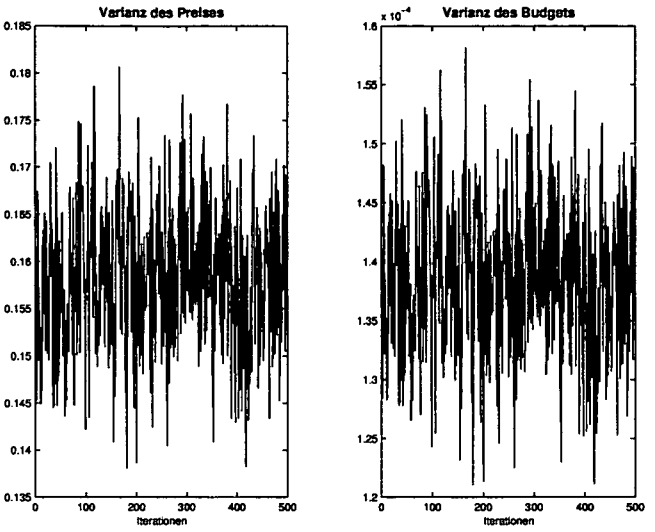

Abbildung 34: Trace Plots der MCMC Iterationen der Parametervarianzen (Modell 3)

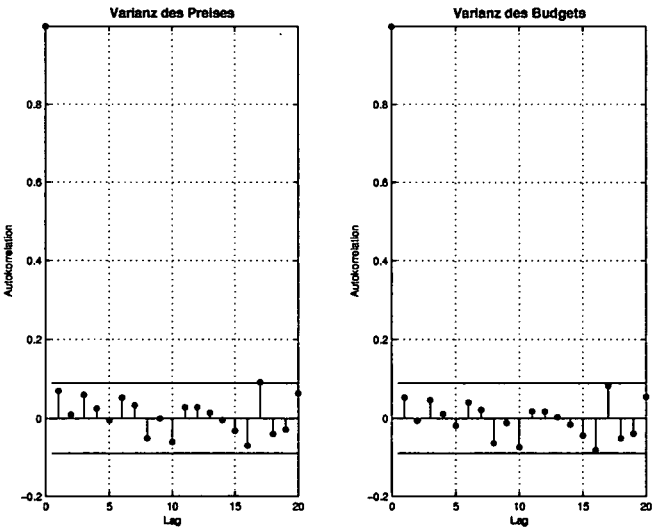

Abbildung 35: Autokorrelationsplots der MCMC Iterationen der Parametervarianzen (Modell 3)

Dem RMSE nach weist Modell 2 den kleinsten Wert auf, gefolgt von Modell 3 und Modell 1 an letzter Stelle.

	Modell 1	Modell 2	Modell 3
Marke 1	0.1656	0.0785	0.0818
Marke 2	1.9556	0.3784	0.5030

Tabelle 14: RMSE

Auch auf die Repräsentation der Trace Plots und Histogramme der Fehlerterme wird in weiterer Folge verzichtet.

Segmentierung

Was die Zuordnung zu Segment 1 betrifft, arbeitet Modell 1 sehr gut, Modell 2 ordnet alle Konsumenten ausnahmslos richtig zu, arbeitet also auch hier wieder perfekt (vgl. Tab. 15).
Was Segment 2 betrifft, arbeiten beide Modelle in etwa gleich gut (vgl. Tab. 16[122]).

	Modell 1	Modell 2
Marke 1, richtig	95.6	100.0
Marke 1, tw. richtig	4.4	0.0
Marke 1, falsch	0.0	0.0
Marke 2, richtig	69.2	99.6
Marke 2, tw. richtig	26.4	0.4
Marke 2, falsch	4.4	0.0

Tabelle 15: Trefferwahrscheinlichkeit in Segment 1 (in %)

[122]Modell 2: Marke 2, teilweise richtig: Lediglich in einigen wenigen Iterationen erfolgte eine falsche Zuordnung.

	Modell 1	Modell 2
Marke 1, richtig	100	84.93
Marke 1, tw. richtig	0	15.07
Marke 1, falsch	0	0.00
Marke 2, richtig	100	52.27
Marke 2, tw. richtig	0	47.73
Marke 2, falsch	0	0.00

Tabelle 16: Trefferwahrscheinlichkeit in Segment 2 (in %)

In Abb. 36 sieht man, dass alle Konsumenten des Segmentes 2 vollkommen korrekt zugeordnet wurden. Bei der Identifizierung der zu Segment 1 gehörenden Konsumenten kam es allerdings zu teilweisen Fehlzuordnungen. Jedoch muss man betonen, dass kein einziger Konsument über alle Iterationen durchgehend der falschen Gruppe zugeteilt wurde.

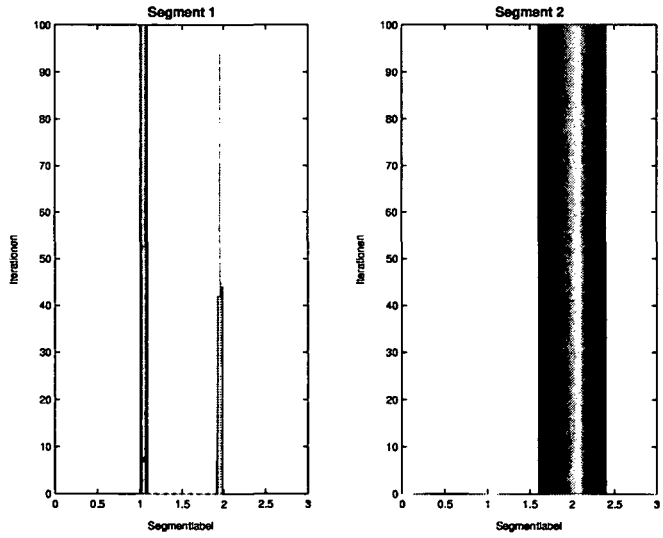

Abbildung 36: Segmentzuordnung der Konsumenten (Modell 1)

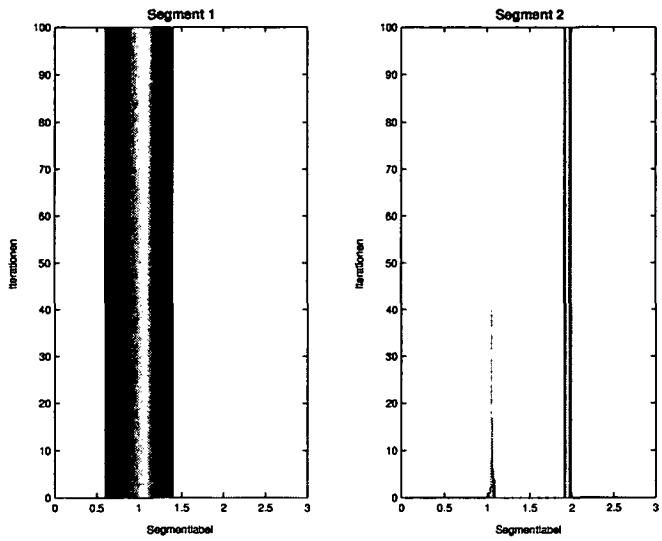

Abbildung 37: Segmentzuordnung der Konsumenten (Modell 2)

Auch Modell 2 schafft es, die Zielgruppe der entsprechenden Marke komplett richtig zu identifizieren. Im Gegensatz zu Modell 1 können jedoch auch fast alle Konsumenten des Segmentes 2 durchgehend korrekt zugeteilt werden. Es treten lediglich vereinzelte Fehlzuordnungen in einigen wenigen Iterationen auf. Weiters wurde kein einziger Konsument in allen Iterationen falsch zugeordnet.

Reiht man die Modelle anhand ihrer Loglikelihoods (vgl. Tab. 17), erweist sich Modell 2 als bestes. Zweitbestes Modell für den hier betrachteten Datentyp war Modell 1, relativ weit abgeschlagen blieb im Fall von Marke 1 das Zufallseffektmodell ohne Segmentierung (Modell 3).

	Modell 1	Modell 2	Modell 3
Marke 1	-2.1541	-1.7101	-2.3863
Marke 2	-5.7833	-4.2602	-5.2184

Tabelle 17: Modellloglikelihoods (alle Werte $*10^3$)

Insgesamt gesehen ergab sich aufgrund des RMSE, der Modellloglikelihood

und der Performance hinsichtlich der Segmentierung Modell 2 als bestes. Für Marke 2 erwies sich zwar angesichts der Loglikelihood[123] Modell 3 als etwas besser als Modell 1, allerdings liefert es keinerlei Segmentierungsinformation. Es benötigt zwar weniger Parameterschätzungen, jedoch weisen die Parameterwerte sehr große Varianzen auf, da gut separierte Segmente über lediglich einen Parameter und dessen Varianz repräsentiert werden müssen. Deshalb lohnt es sich prinzipiell mehr Parameter in Kauf zu nehmen dafür aber nicht auf die Segmentinformation zu verzichten, weshalb insgesamt betrachtet Modell 1 als zweitbester Ansatz interpretiert werden kann.

8.3.3 Simulationsdaten 3 (Starke Heterogenität)

Die Heterogenität innerhalb der Segmente wurde weiter erhöht, allerdings nur so stark, dass die Gruppen noch gut separiert bleiben.

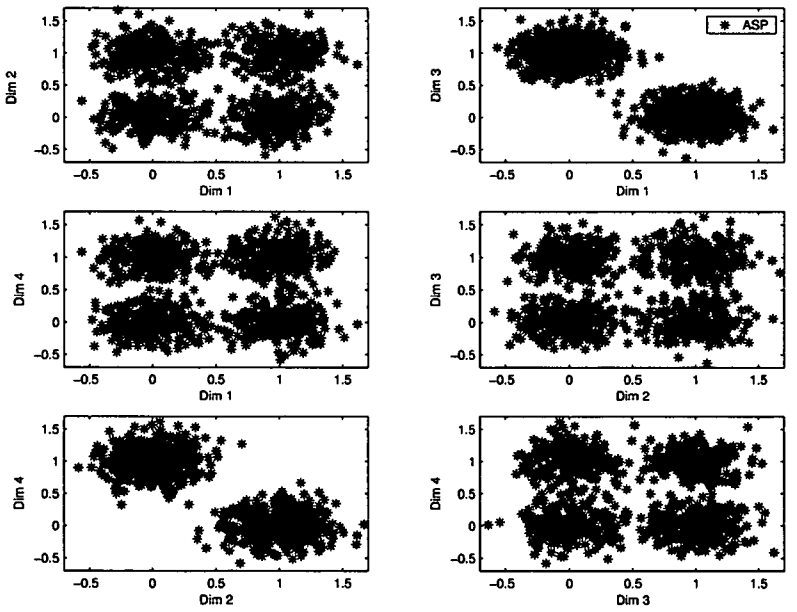

Abbildung 38: Darstellung der Aspirations für alle Attributdimensionen bei starker Heterogenität (Dim i bezeichnet das jeweilige Markenattribut im Produkteigenschaftsraum)

[123]Dem RMSE nach schnitt für beide Marken Modell 3 besser ab.

Parameter- und Varianzschätzungen

Die Unterschiede in den Parameterwerten zwischen den Modellen 1 und 2 werden mit steigender Heterogenität in den Segmenten zunehmend größer. In Modell 2 weist Segment 1 betragsmäßig kleinere und Segment 2 größere Preiskoeffizienten auf als Modell 1 (siehe Tab. 18).

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	-0.7845 (0.0308)	-0.4808 (0.0305)	-0.0681 (0.0109)
Marke 1, Segment 2	0.0365 (0.0114)	0.0959 (0.0030)	
Marke 2, Segment 1	-0.8912 (0.0292)	-0.5672 (0.0338)	-0.0958 (0.0133)
Marke 2, Segment 2	0.0328 (0.0094)	0.0953 (0.0024)	

Tabelle 18: Parameterwerte für den Preis, Standardabweichung in Klammer

Die Budgetwerte in Modell 2 sind kleiner als in Modell 1, die Standardabweichungen sind in Segment 1 größer, dafür in Segment 2 kleiner.
Die Parameterwerte und Standardabweichungen von Preis und Budget liegen bei Modell 3 zwischen jenen in den beiden Segmenten von Modell 1 bzw. Modell 2. Marke 1 und 2 besitzen annähernd die gleichen Preis- und Budgetwerte. Dies gilt für beide Segmente, wobei der Preiskoeffizient im zweiten Segment positiv ist.

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	46.3 (0.8033)	33.7 (1.2000)	18.1 (0.4207)
Marke 1, Segment 2	14.0 (0.2694)	11.9 (0.1288)	
Marke 2, Segment 1	50.7 (0.9087)	37.4 (1.3000)	19.1 (0.5181)
Marke 2, Segment 2	14.0 (0.2664)	11.6 (0.1165)	

Tabelle 19: Parameterwerte für das Budget, Standardabweichung in Klammer (alle Werte $*10^{-3}$)

In allen drei Modellen ist es bereits mit 1000 Iterationen möglich bei den Parameterschätzungen Konvergenz zu erreichen. Weiters deuten die Autokorrelationsplots über die rasche Dämpfung auf ein schnelles Mischverhalten hin.

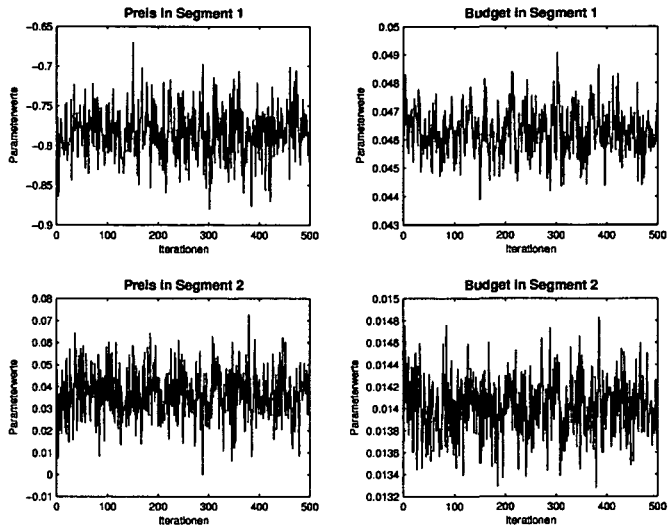

Abbildung 39: Trace Plots der MCMC Iterationen der Parameter (Modell 1)

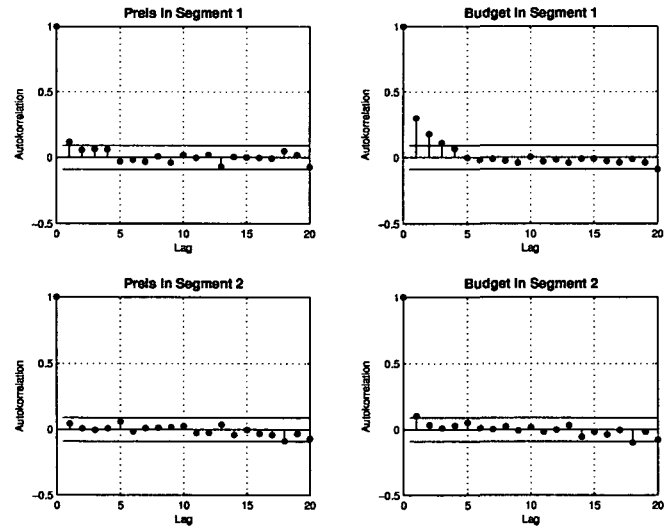

Abbildung 40: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 1)

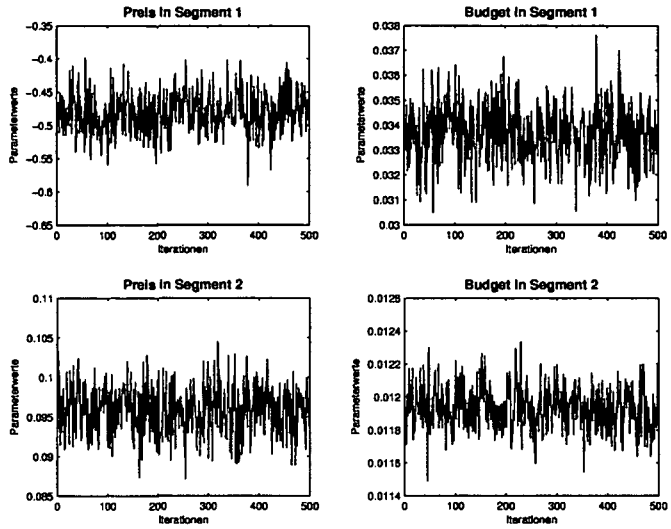

Abbildung 41: Trace Plots der MCMC Iterationen der Parameter (Modell 2)

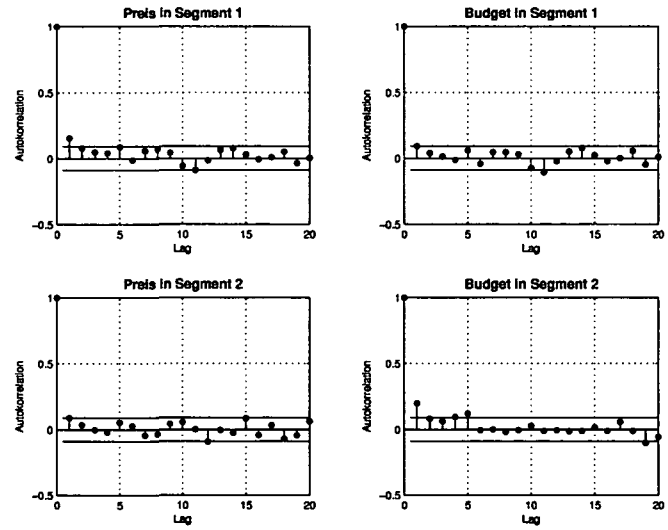

Abbildung 42: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 2)

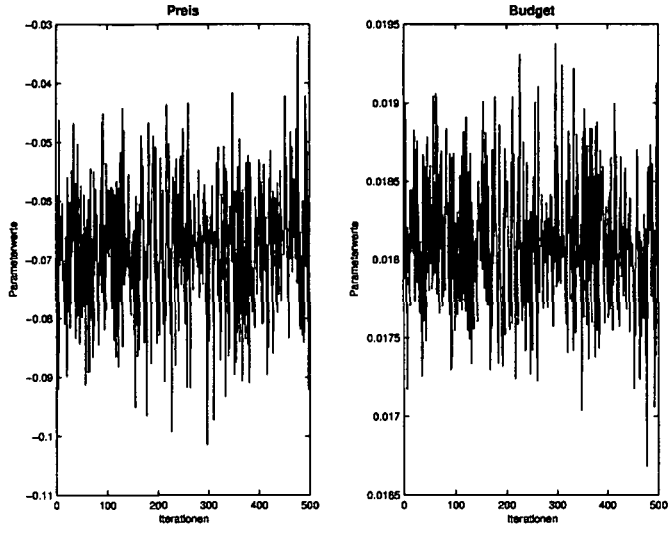

Abbildung 43: Trace Plots der MCMC Iterationen der Parameter (Modell 3)

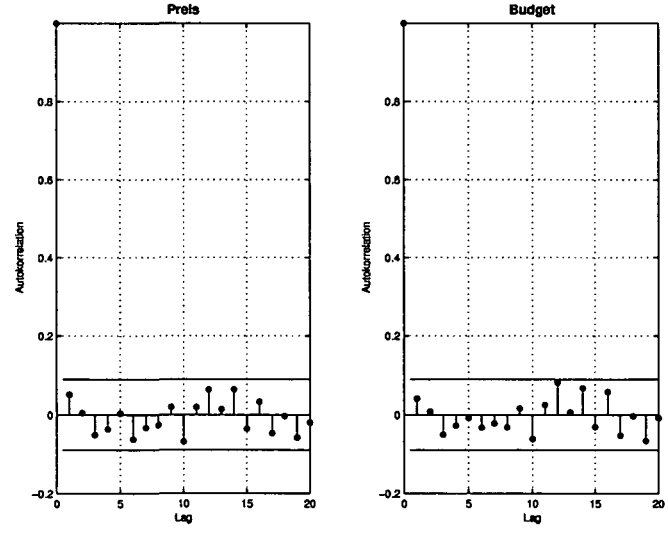

Abbildung 44: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 3)

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	0	1935.0000	1172
Marke 1, Segment 2	0	0.3049	
Marke 2, Segment 1	0	2651.0000	1668
Marke 2, Segment 2	0	0.2019	

Tabelle 20: Parameterwerte der Preisvarianz (alle Werte $*10^{-4}$)

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	0	29.6170	18.100
Marke 1, Segment 2	0	0.6497	
Marke 2, Segment 1	0	40.4560	25.812
Marke 2, Segment 2	0	0.5770	

Tabelle 21: Parameterwerte der Budgetvarianz (alle Werte $*10^{-5}$)

Wie man in den Tabellen 20 und 21 herauslesen kann, besitzen Preis und Budget auch hier in Segment 1 eine wesentlich größere Varianz, wobei der Unterschied für das Budget nicht so extrem ausfällt wie für den Preis. Allgemein betrachtet ist nun die Varianz noch etwas höher als unter schwach heterogenen Konsumentensegmenten. Relativ zu Modell 2 nimmt die Varianz im reinen Zufallseffektmodell mit steigender Heterogenität ab.

Die Varianzen konvergieren nach 1000 Iterationen gut mit Ausnahme der Preisvarianz in Segment 2, für die laut „raftery“ in etwa 3000 Iterationen notwendig wären. Auch für die Budgetvarianz in Segment 2 sollte mit 1300 Iterationen etwas länger simuliert werden (vgl. Abb. 45).
Das schlechte Mischverhalten der Preisvarianz in Segment 2 bzw. das etwas langsamere Mischen für die Budgetvarianz in Segment 2 im Vergleich zu den Varianzen des ersten Segmentes erkennt man auch in den Autokorrelationsplots (siehe Abb. 46).
Für Modell 3 sind sowohl Konvergenz- als auch Mischverhalten der Varianzen für 1000 Iterationen zufriedenstellend (siehe Abb. 47 und 48).

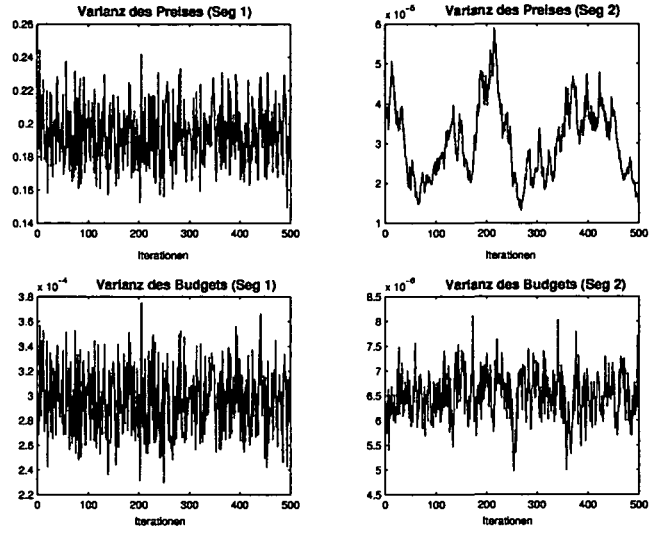

Abbildung 45: Trace Plots der MCMC Iterationen der Parametervarianzen (Modell 2)

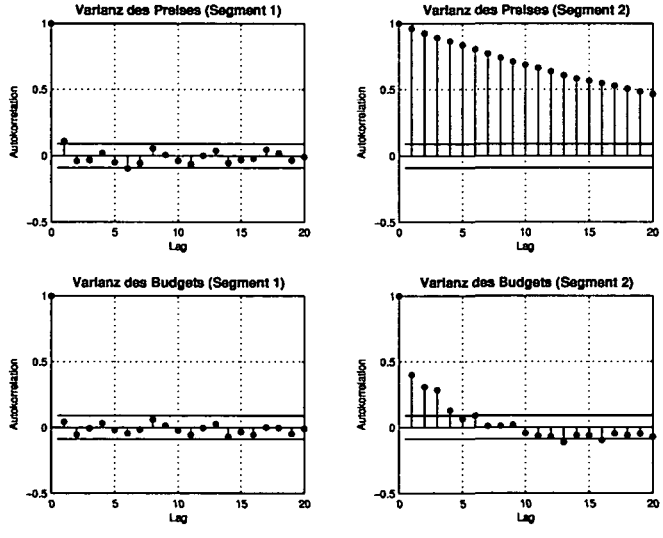

Abbildung 46: Autokorrelationsplots der MCMC Iterationen der Parametervarianzen (Modell 2)

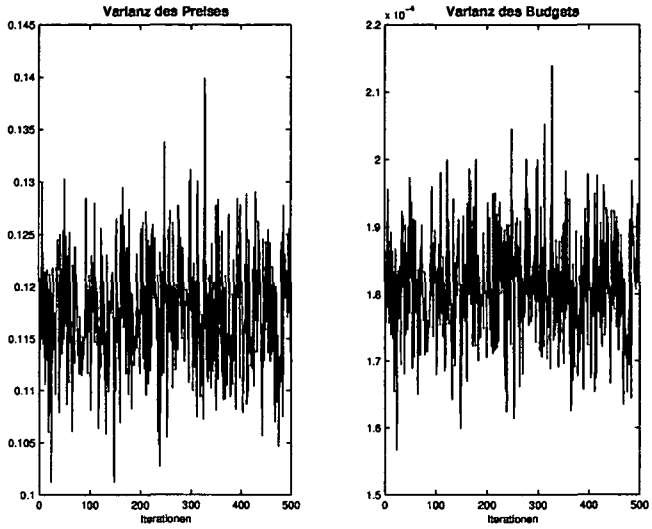

Abbildung 47: Trace Plots der MCMC Iterationen der Parametervarianzen (Modell 3)

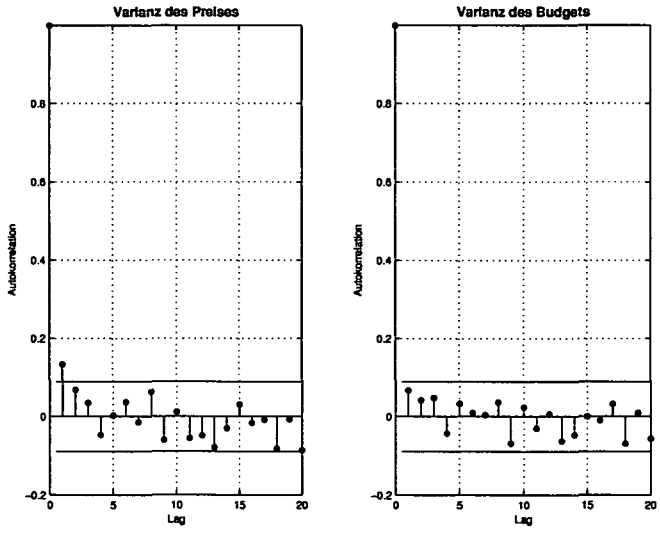

Abbildung 48: Autokorrelationsplots der MCMC Iterationen der Parametervarianzen (Modell 3)

	Modell 1	Modell 2	Modell 3
Marke 1	0.2975	0.0155	0.0210
Marke 2	0.3537	0.0167	0.0181

Tabelle 22: RMSE

Dem RMSE nach ist Modell 2 für diesen Datensatz zu bevorzugen (siehe Tab. 22). An zweiter Stelle steht Modell 3, d.h. ab einer gewissen Streuung wird es vorteilhafter Modell 3 anstelle von Modell 1 zu verwenden.[124]

Segmentierung

Betrachtet man Tabelle 23, sieht man, dass Modell 1 bezüglich der Zuordnung der Konsumenten zu Segment 1 ziemlich konfuse Ergebnisse liefert und nicht einmal die Hälfte der Bevölkerung den richtigen Segmenten zuordnet. Modell 2 hingegen schafft es fast 90% bzw. sogar 100% der Konsumenten der korrekten Gruppe zuzuteilen.
Was Segment 2 betrifft, arbeitet Modell 1 eine Spur besser als Modell 2, allerdings nicht wesentlich (siehe Tab. 24[125]).

	Modell 1	Modell 2
Marke 1, richtig	34.8	89.6
Marke 1, tw. richtig	28.8	10.4
Marke 1, falsch	36.4	0.0
Marke 2, richtig	40.8	100.0
Marke 2, tw. richtig	35.6	0.0
Marke 2, falsch	23.6	0.0

Tabelle 23: Trefferwahrscheinlichkeit in Segment 1 (in %)

[124] Das bestätigt sich auch bei Betrachtung der Modelllloglikelihoods.
[125] Modell 2, teilweise richtig: Lediglich in einigen wenigen Iterationen wurden die Konsumenten falsch zugeordnet.

	Modell 1	Modell 2
Marke 1, richtig	98.67	4.40
Marke 1, tw. richtig	1.33	95.47
Marke 1, falsch	0.00	0.13
Marke 2, richtig	99.73	8.40
Marke 2, tw. richtig	0.27	91.33
Marke 2, falsch	0.00	0.27

Tabelle 24: Trefferwahrscheinlichkeit in Segment 2 (in %)

Die Performance bezüglich Segmentierung ist in Modell 1 äußerst schlecht was Segment 1 betrifft (siehe Abb. 49). Die Gruppe der Nichtkäufer wurde allerdings sehr gut erkannt.
Modell 2 identifiziert beide Gruppen relativ gut, wobei Segment 1 nahezu perfekt entdeckt wird. In Segment 2 kommt es zu kleinen Fehlzuordnungen, allerdings wird kein einziger Konsument über alle Iterationen durchgehend falsch zugeteilt (siehe Abb. 50).

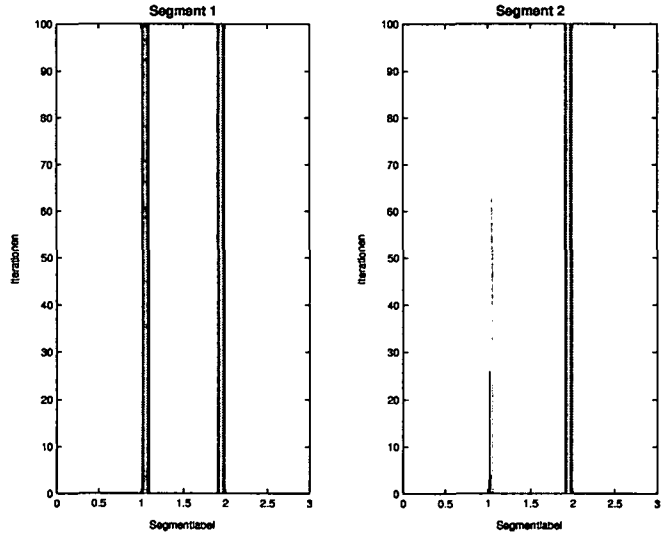

Abbildung 49: Segmentzuordnung der Konsumenten (Modell 1)

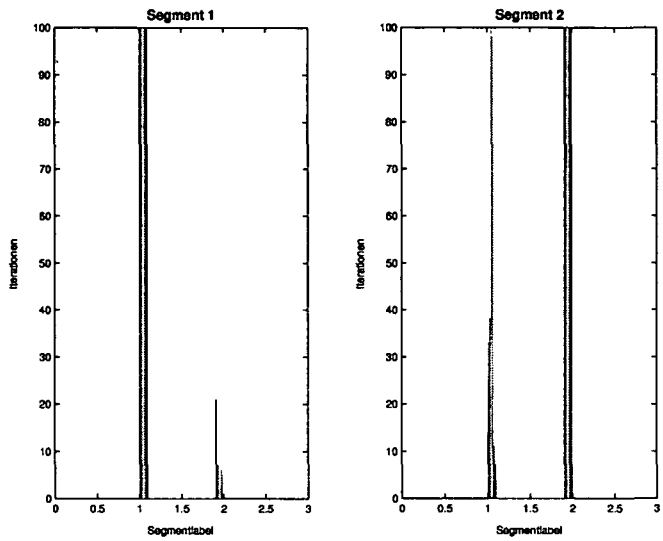

Abbildung 50: Segmentzuordnung der Konsumenten (Modell 2)

Was die Modellloglikelihood betrifft, schlägt Modell 2 die anderen beiden Modelle, wobei Modell 3 besser abschneidet als Modell 1. Auch die Performance hinsichtlich der Segmentierung ist für Modell 2 besser als für Modell 1, da das Segment der potentiellen Käufer (Segment 1) das wichtigere repräsentiert. Modell 1 schafft es allerdings nur ein Drittel der zu diesem Segment gehörenden Konsumenten richtig zu identifizieren.

	Modell 1	Modell 2	Modell 3
Marke 1	-2.8351	-0.5781	-1.3032
Marke 2	-3.1088	-0.6714	-1.5570

Tabelle 25: Modellloglikelihoods (alle Werte $*10^3$)

Im Falle eines Datensatzes mit großer Streuung in gut separierten Segmenten ist also eindeutig ein Ansatz, der die Heterogenität innerhalb der Segmente berücksichtigt, zu bevorzugen. Ein reiner Zufallseffektansatz ist weiters besser als ein Modell, das diskrete Segmente mit fixen Effekten berücksichtigt.

8.3.4 Simulationsdaten 4 (Überlappende Segmente)

Beim vierten und letzten Datensatz wurde die Heterogenität innerhalb der Segmente so weit durch stärkeres Streuen der individuellen Präferenzen der Konsumenten um den Segment-Idealpunkt erhöht, dass die Segmente einander schließlich überlappen.

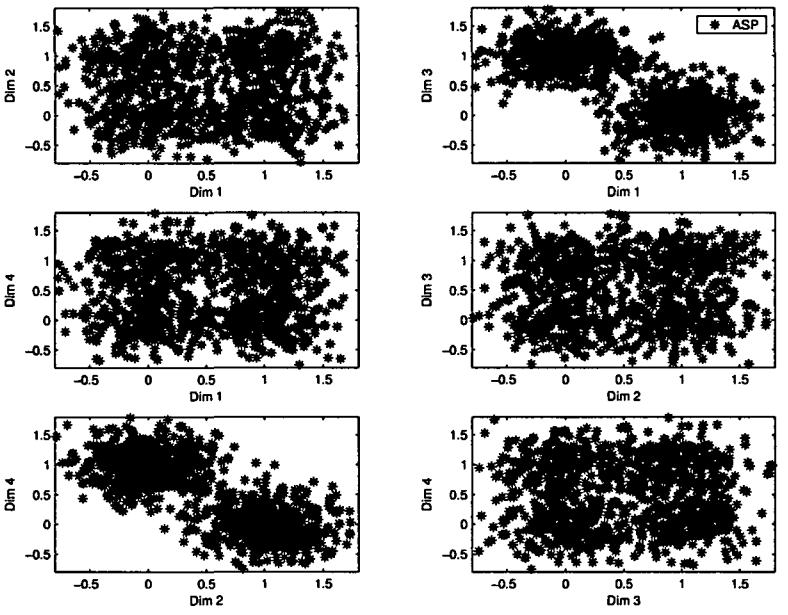

Abbildung 51: Darstellung der Aspirations für alle Attributdimensionen für überlappende Segmente (Dim i bezeichnet das jeweilige Markenattribut im Produkteigenschaftsraum)

Parameter- und Varianzschätzungen

Für die erste Marke sind die Parameterwerte des Preises betragsmäßig kleiner als für die zweite Marke (siehe Tab. 26). In Modell 2 ist weiters die Standardabweichung der Preisparameter kleiner als in Modell 1.

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	-3.1784 (0.1527)	-1.7625 (0.1137)	-0.7474 (0.0254)
Marke 1, Segment 2	-0.6451 (0.0255)	-0.5081 (0.0124)	
Marke 2, Segment 1	-2.2996 (0.1156)	-2.4105 (0.2179)	-0.8871 (0.0218)
Marke 2, Segment 2	-0.7389 (0.0311)	-0.7759 (0.0176)	

Tabelle 26: Parameterwerte für den Preis, Standardabweichung in Klammer

Was die Budgetwerte betrifft, besitzt Modell 1 für die zweite Marke kleinere und für Modell 2 größere Werte als die erste Marke (siehe Tab. 27).
Die Preis- und Budgetwerte in Modell 3 liegen wieder zwischen jenen der beiden Segmente in Modell 1 bzw. Modell 2.

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	109.3 (4.1000)	62.4 (3.3000)	32.9 (0.7503)
Marke 1, Segment 2	29.6 (0.5841)	25.9 (0.3758)	
Marke 2, Segment 1	69.2 (2.5000)	72.9 (5.2000)	30.6 (0.5649)
Marke 2, Segment 2	26.5 (0.6112)	27.5 (0.4512)	

Tabelle 27: Parameterwerte für das Budget, Standardabweichung in Klammer (alle Werte $*10^{-3}$)

Das Konvergenzverhalten nach einer Laufzeit von 1000 Iterationen ist vor allem in Segment 2 relativ gut (vgl. Abb. 52). Für die Parameter in Segment 1 wäre es eventuell besser, die Simulation etwas länger laufen zu lassen. Laut „raftery“ werden speziell für das Budget in Segment 1 ca. 1300 Iterationen empfohlen. Auch die Autokorrelationsplots in Abb. 53 bestätigen diese Beobachtungen.

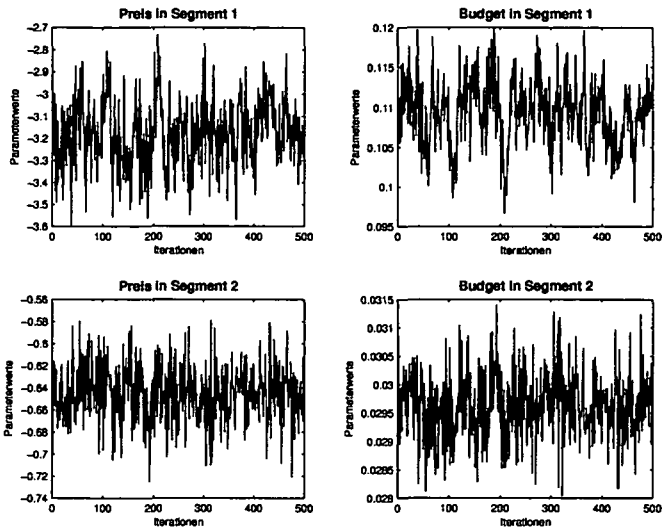

Abbildung 52: Trace Plots der MCMC Iterationen der Parameter (Modell 1)

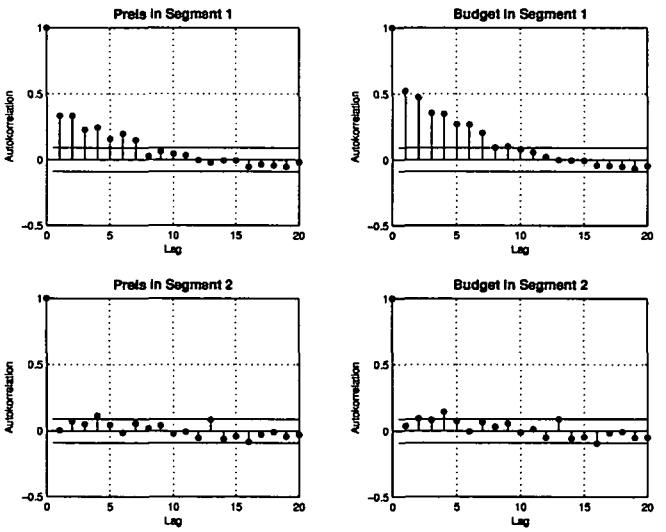

Abbildung 53: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 1)

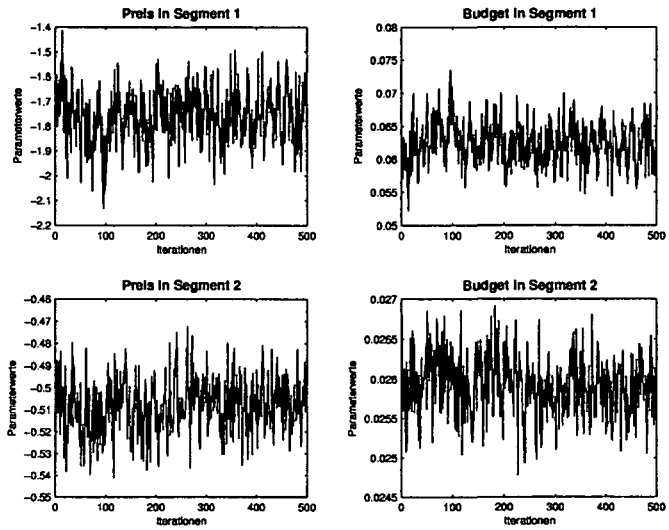

Abbildung 54: Trace Plots der MCMC Iterationen der Parameter (Modell 2)

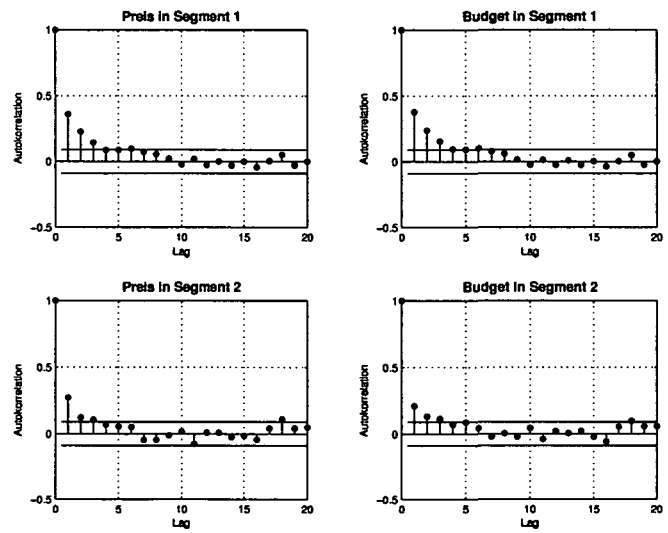

Abbildung 55: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 2)

In Modell 2 und 3 erweist sich die ad hoc gewählte Laufzeit als ausreichend.

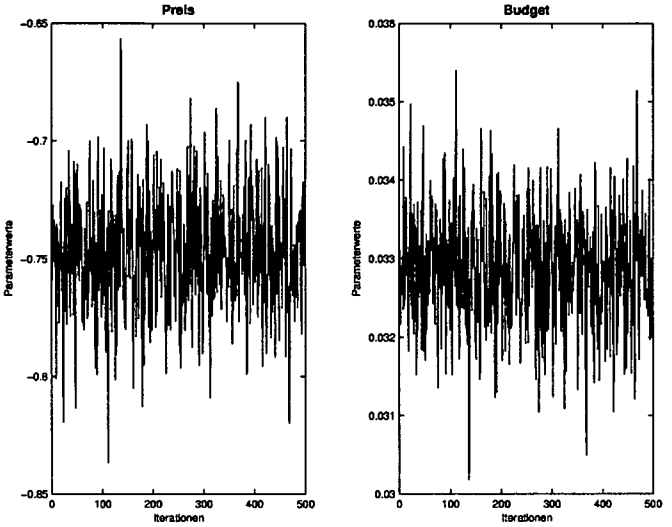

Abbildung 56: Trace Plots der MCMC Iterationen der Parameter (Modell 3)

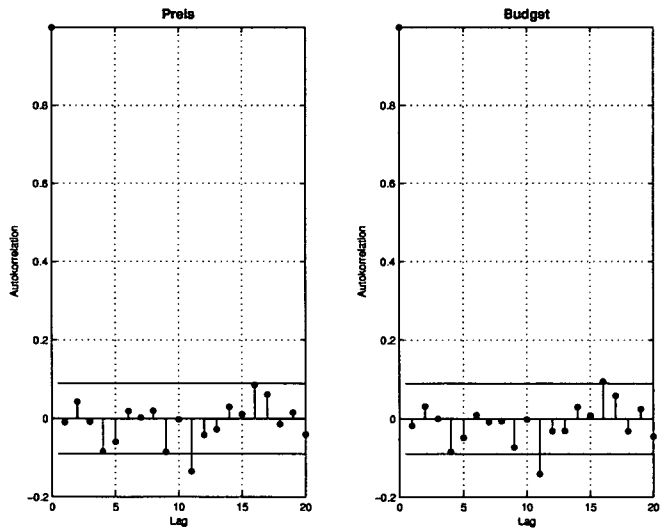

Abbildung 57: Autokorrelationsplots der MCMC Iterationen der Parameter (Modell 3)

In Modell 2 übersteigen die Varianzen des Preises der zweiten Marke jene der

ersten (siehe Tab. 28). Wieder sind die Varianzen im ersten Segment größer als jene in Segment 2, allerdings ist dieser Unterschied nicht mehr signifikant.

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	0	1.7487	0.5411
Marke 1, Segment 2	0	0.0135	
Marke 2, Segment 1	0	2.2520	0.3581
Marke 2, Segment 2	0	0.1153	

Tabelle 28: Parameterwerte der Preisvarianz

Die Varianzen im dritten Modell liegen zwischen den segmentspezifischen Werten von Modell 2 (siehe Tab. 28 und 29).

	Modell 1	Modell 2	Modell 3
Marke 1, Segment 1	0	15.0000	5.024
Marke 1, Segment 2	0	0.5040	
Marke 2, Segment 1	0	12.0000	2.687
Marke 2, Segment 2	0	0.9314	

Tabelle 29: Parameterwerte der Budgetvarianz (alle Werte $*10^{-4}$)

Das Konvergenzverhalten ist für alle Varianzen bis auf die des Preises in Segment 2 zufriedenstellend (siehe Abb. 58).
Ebenso erkennt man anhand der Dämpfung in den Autokorrelationen, dass für die Preisvarianz in Segment 2 eine längere Laufzeit von Vorteil wäre (siehe Abb. 59). Eventuell sollte man in Anbetracht der im Vergleich zu den Varianzen in Segment 1 noch stärkeren Autokorrelationen auch für die Budgetvarianz in Segment 2 eine längere Laufzeit wählen.

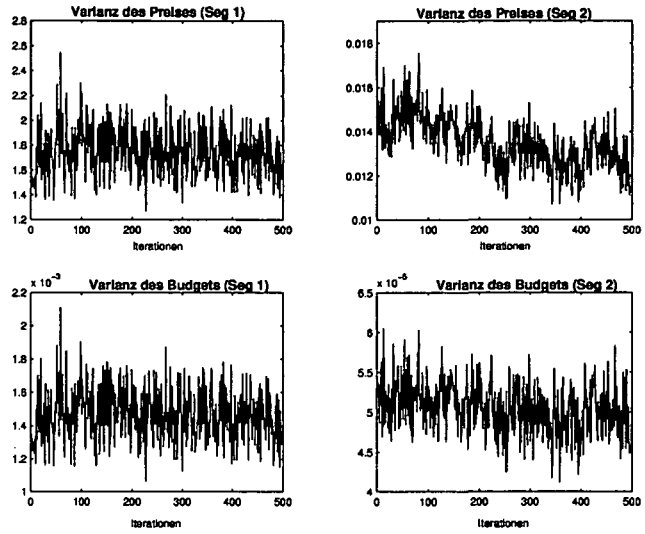

Abbildung 58: Trace Plots der MCMC Iterationen der Parametervarianzen (Modell 2)

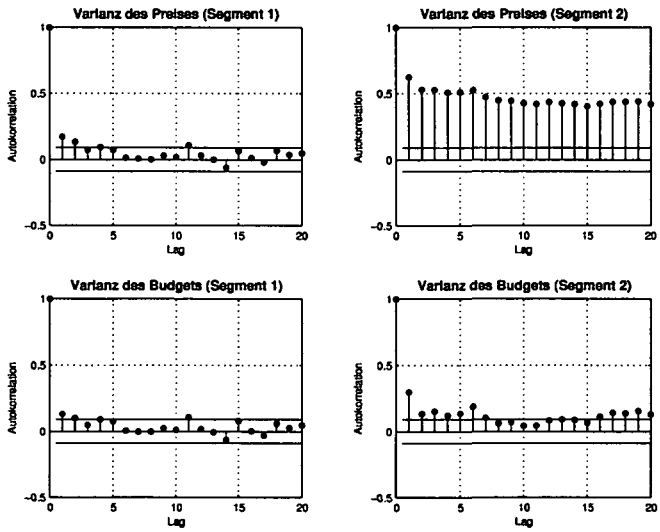

Abbildung 59: Autokorrelationsplots der MCMC Iterationen der Parametervarianzen (Modell 2)

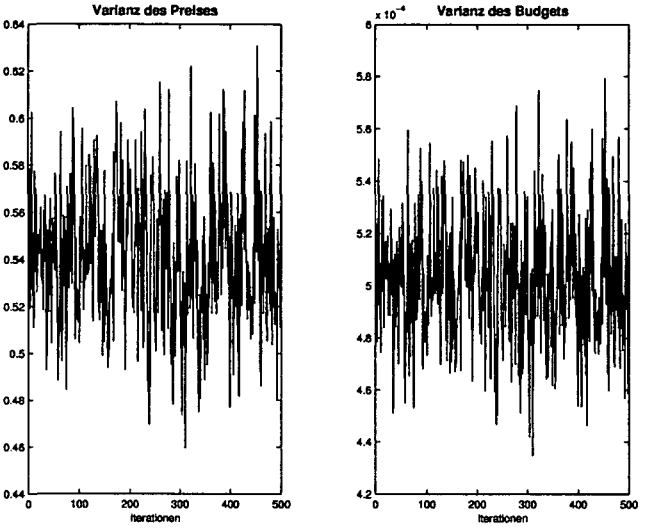

Abbildung 60: Trace Plots der MCMC Iterationen der Parametervarianzen (Modell 3)

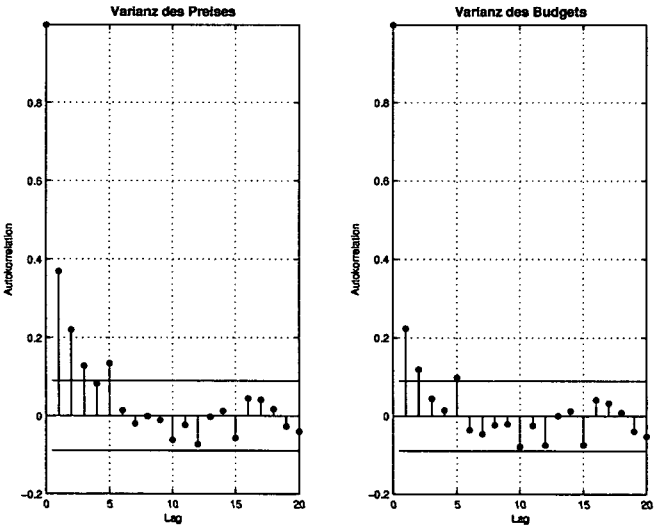

Abbildung 61: Autokorrelationsplots der MCMC Iterationen der Parametervarianzen (Modell 3)

Beim dritten Modell scheinen 1000 Iterationen wieder ausreichend zu sein in Bezug auf Konvergenz (Abb. 60) und Autokorrelation (Abb. 61).

Wie man aus Tabelle 30 ablesen kann, ist hinsichtlich des RMSE Modell 2 zu bevorzugen, gefolgt von Modell 3 und Modell 1 an letzter Stelle. Modell 3 arbeitet bei überlappenden Segmenten fast so gut wie Modell 2. Lediglich Modell 1, das fixe Parameter innerhalb diskreter Segmente annimmt, erweist sich als deutlich schlechter mit größeren Fehlern.

	Modell 1	Modell 2	Modell 3
Marke 1	0.4983	0.0668	0.0831
Marke 2	0.3496	0.0277	0.0398

Tabelle 30: RMSE

Segmentierung

Was die Trefferwahrscheinlichkeit hinsichtlich der Zuordnung der Konsumenten zu den Segmenten (siehe Tabellen 31 und 32) betrifft, arbeitet Modell 2 eine Spur besser als das erste Modell, allerdings nicht wesentlich. Modell 2 teilt speziell für die erste Marke dem ersten Segment kaum Konsumenten komplett falsch zu, allerdings wird gerade einmal ein Drittel über alle Iterationen durchgehend richtig erkannt. Modell 1 erkennt nicht einmal ein Viertel komplett richtig und versagt bei der Zuordnung von mehr als 75% der Konsumenten.

	Modell 1	Modell 2
Marke 1, richtig	12.4	31.6
Marke 1, tw. richtig	10.0	66.8
Marke 1, falsch	77.6	1.6
Marke 2, richtig	19.2	10.8
Marke 2, tw. richtig	24.4	58.0
Marke 2, falsch	56.4	31.2

Tabelle 31: Trefferwahrscheinlichkeit in Segment 1 (in %)

Für Segment 2 arbeiten beide Modelle relativ gut, wobei Modell 1 sogar etwas

zu bevorzugen ist, da es fast alle Konsumenten komplett richtig zuordnet. In Modell 2 kommt es vor, dass fast alle Konsumenten nur in einigen wenigen Iterationen falsch zugeordnet werden.[126]

	Modell 1	Modell 2
Marke 1, richtig	98.93	1.47
Marke 1, tw. richtig	0.93	98.53
Marke 1, falsch	0.13	0.00
Marke 2, richtig	95.07	70.80
Marke 2, tw. richtig	4.13	28.53
Marke 2, falsch	0.80	0.67

Tabelle 32: Trefferwahrscheinlichkeit in Segment 2 (in %)

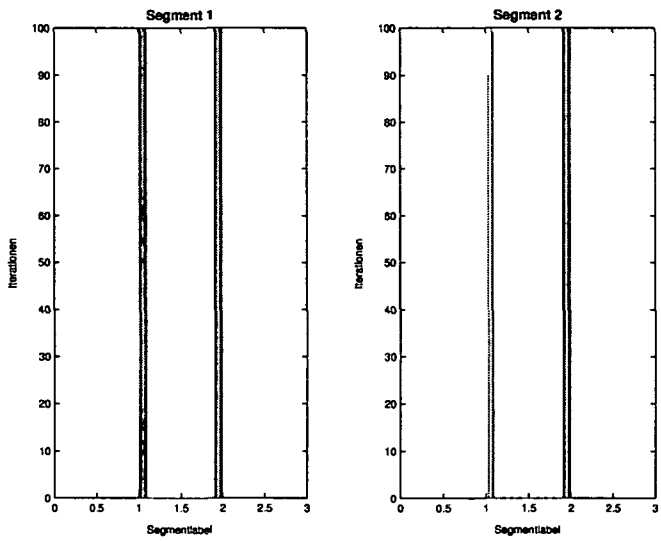

Abbildung 62: Segmentzuordnung der Konsumenten (Modell 1)

[126] vgl. hierzu auch die Abbildungen 62 und 63

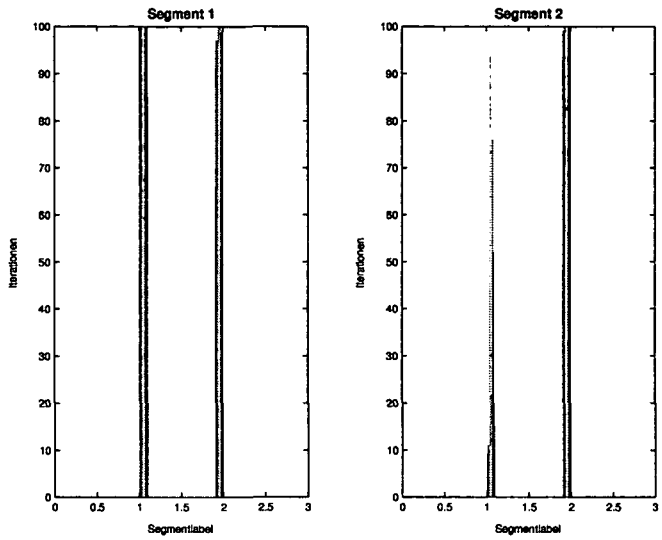

Abbildung 63: Segmentzuordnung der Konsumenten (Modell 2)

Der Modellloglikelihood nach eignet sich von den drei Ansätzen Modell 2 am besten für Datensätze, in denen die Segmente einander überschneiden. Direkt danach mit einer nur etwas geringeren Likelihood folgt Modell 3. Weit abgeschlagen liegt Modell 1 an dritter und letzter Stelle. Im Grunde ist es im Falle überlappender Konsumentensegmente äußerst schwierig, die Bevölkerung den richtigen Gruppen zuzuteilen, was die bessere Performance von Modell 3 im Vergleich zu Modell 1 erklärt.

	Modell 1	Modell 2	Modell 3
Marke 1	-3.4117	-2.1064	-2.5737
Marke 2	-3.0032	-1.1526	-1.5996

Tabelle 33: Modellloglikelihoods (alle Werte $*10^3$)

Ist man an der Segmentinformation interessiert, sollte man Modell 2 verwenden, da die Performance diesbezüglich etwas besser ist als in Modell 1, insbesondere in Anbetracht des interessanteren Segmentes der potentiellen Käufer der entsprechenden Marke. Ansonsten wäre es empfehlenswert, vor allem bei sehr stark überlappenden Segmenten, Modell 3 zu verwenden und

auf die Berücksichtigung diskreter Segmente gänzlich zu verzichten. Modell 3 arbeitet hinsichtlich RMSE und Modellloglikelihood nur eine Spur schlechter als Modell 2, benötigt allerdings weniger Parameter, da quasi nur die Koeffizienten *einer* Gruppe geschätzt werden müssen.

8.4 Zusammenfassung

Betrachtet man die Ergebnisse aller vier Datensätze, erkennt man, dass mit 1000 Iterationen im Großteil der Fälle eine ausreichend lange Laufzeit für die MCMC Schätzungen gewählt wurde. Dies trifft vor allem für die Schätzung der Parameterwerte zu. Was die Varianzen betrifft, sollte allerdings in einigen Situationen länger simuliert werden, da oft noch keine zufriedenstellende Konvergenz der Koeffizienten aufgrund des langsamen Mischverhaltens erreicht werden konnte. Auf diesen Schluss kommt man bei Betrachtung der Trace Plots bzw. der Autokorrelationen. Das gilt vor allem für Modell 2 und hier insbesondere für die Koeffizienten im zweiten Segment, bei Modell 3 sind nur selten längere Laufzeiten notwendig.
Mit zunehmender Heterogenität in den Daten entstehen deutlichere Unterschiede in den Parameterschätzungen der segmentspezifischen Parameter zwischen den Modellen 1 und 2. Weiters wächst die Varianz in Modell 2 mit steigender Heterogenität in den Konsumentengruppen. Die Varianz in Segment 1 ist stets größer als jene in der zweiten Gruppe, was sowohl für den Preis als auch für das Budget zutrifft. Für Modell 3 nimmt die Varianz der Parameter relativ zur Varianz im ersten Segment des zweiten Modells mit zunehmender Heterogenität ab.

Was die Empfehlung eines Modelltyps in Abhängigkeit von der Beschaffenheit der Daten betrifft, ergab sich wie erwartet das folgende Ergebnis. Im Falle sehr homogener Konsumenten innerhalb der Segmente reicht ein diskretes Mischungsmodell (Modell 1). Es ist nicht notwendig, eine Variation der Parameter innerhalb der Gruppen zuzulassen, da dies lediglich die Parameteranzahl erhöht ohne jedoch die Performance zu verbessern. Bezüglich des Aufdeckens der richtigen Segmentstruktur arbeitet Modell 1 ebenso gut wie ein Mischungsmodell mit Zufallseffekten.
Bei steigender Heterogenität innerhalb der Segmente kristallisiert sich der Modelltyp mit innerhalb der Segmente variierenden Parametern (Modell 2) als am besten geeignet heraus. Interessant ist die Tatsache, dass auch bei steigender Streuung dieser Ansatz optimal bleibt, sogar bis zu einem gewis-

sen Überlappungsgrad der Segmente. Im Falle sich überschneidender Konsumentengruppen bleibt zwar die Performance besser als die eines reinen Zufallseffektmodells ohne diskrete Klassen (Modell 3), jedoch können die Konsumenten den richtigen Gruppen kaum mehr verlässlich zugeordnet werden. Berücksichtigt man also neben der besseren Performance von Modell 2 auch die Anzahl der zu schätzenden Parameter, sowie die Segmentierungsergebnisse, wäre eventuell sogar Modell 3 zu bevorzugen, da man bei letzterem nur Parameter für ein Segment zu schätzen hat.
Überschneiden die Gruppen einander schließlich sehr stark, sollte man sich auf den Einsatz von Modell 3 beschränken.

Zusammengefasst wird Modell 1 verhältnismäßig rasch von Modell 2 abgelöst. Schon eine relativ schwache Streuung innerhalb der Segmente reicht, dass ein Mischungsmodell mit Zufallseffekten besser abschneidet. Selbst im Falle sehr homogener Segmente ist die Performance des Mischungsmodells mit Zufallseffekten kaum schlechter als die eines Mischungsmodells, das fixe Effekte in den Gruppen annimmt. Ersteres bleibt auch in einem breiten Bereich an Datensätzen optimal, selbst wenn die Streuung so groß wird, dass die Segmente einander überlappen. Weiters wird Modell 1 ab einem gewissen Heterogenitätsgrad in den Gruppen durch Modell 3 vom zweiten Platz abgelöst. Ein reines Zufallseffektmodell arbeitet erst dann besser als Modell 2, wenn sich die Gruppen so stark überlagern, dass auch eine Segmentierung mithilfe eines Clusteralgorithmus keine sinnvollen Ergebnisse mehr liefert.
Im Großen und Ganzen lassen diese Ergebnisse darauf schließen, dass man bei Vermutung eines segmentierten Marktes am besten Modell 2 verwendet, da es sich selbst bis zu einer sehr starken Streuung in den Gruppen gegen ein reines Zufallseffektmodell durchsetzt und auch bei völlig homogenen Segmenten nicht viel schlechter arbeitet als Modell 1.

8.5 Diskussion

Im Prinzip könnte noch eine Fülle an Experimenten mit weiteren Simulationsdatensätzen[127] durchgeführt werden, allerdings würde dies den Rahmen der Arbeit sprengen. Z.B. könnte genauer untersucht werden, ab welchem Überlappungsgrad sich Modell 3 tatsächlich gegen Modell 2 durchsetzt. Man könnte auch testen, wie groß die Streuung sein muss, damit Modell 2 dem ersten

[127] produziert über den künstlichen Konsumentenmarkt aus Abschnitt 8.1

Modell in Sachen Performance überlegen ist.

Einen weiteren Schritt würde die anschließende Überprüfung der gefundenen Resultate anhand empirischer Daten darstellen. D.h. man könnte versuchen, verschiedenste Datensätze (beispielsweise Paneldaten von Märkten mit homogenen, gut separierten bis hin zu sehr heterogenen, sich überlappenden Konsumentensegmenten) zu bekommen. Man kann damit versuchen, die mithilfe der Simulationsdaten gefundenen Ergebnisse zu reproduzieren. Diese Art von Experimenten war leider nicht durchführbar, da keine entsprechenden Daten für solch eine umfassende Studie zur Verfügung standen.

Literatur

[1] G. M. Allenby, N. Arora, and J. L. Ginter. On the Heterogeneity of Demand. *Journal of Marketing Research*, 35:384–389, August 1998.

[2] G. M. Allenby and P. J. Lenk. Modeling Household Purchase Behavior With Logistic Normal Regression. *Journal of the American Statistical Association*, 89(428):1218–1231, Dezember 1994.

[3] R. L. Andrews, A. Ainslie, and I. S. Currim. An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity. *Journal of Marketing Research*, 39:479–487, November 2002.

[4] R. L. Andrews, A. Ansari, and I. S. Currim. Hierarchical Bayes Versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery. *Journal of Marketing Research*, 39:87–98, Februar 2002.

[5] R. L. Andrews and I. S. Currim. Recovering and Profiling the True Segmentation Structure in Markets: An Empirical Investigation. *International Journal of Research in Marketing*, 20:177–192, 2003.

[6] R. L. Andrews and I. S. Currim. Retention of Latent Segments in Regression-Based Marketing Models. *International Journal of Research in Marketing*, 20:315–321, 2003.

[7] A. Ansari, K. Jedidi, and S. Jagpal. A Hierarchical Bayesian Methodology for Treating Heterogeneity in Structural Equation Models. *Marketing Science*, 19(4):328–347, Herbst 2000.

[8] N. Arora, G. M. Allenby, and J. L. Ginter. A Hierarchical Bayes Model of Primary and Secondary Demand. *Marketing Science*, 17(1):29–44, 1998.

[9] C. Biernacki and G. Govaert. Choosing Models in Model-Based Clustering and Discriminant Analysis. Rapport de recherche 3509, Institut National de Recherche en Informatique et en Automatique, Oktober 1998.

[10] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, 1973.

[11] H. Bozdogan. Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. *Psychometrika*, 52:345–370, September 1987.

[12] R. E. Bucklin, S. Gupta, and S. Siddarth. Determining Segmentation in Sales Response Across Consumer Purchase Behaviors. *Journal of Marketing Research*, 35:189–197, Mai 1998.

[13] M. Chen, Q. Shao, and J. G. Ibrahim. *Monte Carlo Methods in Bayesian Computation.* Springer, NY, 2000.

[14] P. K. Chintagunta. Heterogeneous Logit Model Implications for Brand Positioning. *Journal of Marketing Research*, 31:304–311, Mai 1994.

[15] L.G. Cooper and M. Nakanishi. *Market-Share Analysis.* Kluwer Academic Publishers, Boston, MA, 1988.

[16] W. S. DeSarbo, W. A. Kamakura, and M. Wedel. Applications of Multivariate Latent Variable Models in Marketing. April 2002.

[17] X. Dong and F. S. Koppelman. Comparison of Methods Representing Heterogeneity in Logit Models. In *10th International Conference on Travel Behaviour Research*, August 2003.

[18] T. Elrod and M. P. Keane. A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data. *Journal of Marketing Research*, 32:1–16, Februar 1995.

[19] T. Erdem. A Dynamic Analysis of Market Structure Based on Panel Data. *Marketing Science*, 15(4):359–378, 1996.

[20] S. Frühwirth-Schnatter. MCMC Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association*, 96:194–209, 2001.

[21] S. Frühwirth-Schnatter, T. Otter, and R. Tüchler. A Fully Bayesian Analysis of Multivariate Latent Class Models with an Application to Metric Conjoint Analysis. Working Paper 89, Juni 2002.

[22] S. Frühwirth-Schnatter, R. Tüchler, and T. Otter. Bayesian Analysis of the Heterogeneity Model. *Journal of Business and Economic Statistics*, 22(1):2–15, 2004.

[23] A. E. Gelfand and D. K. Dey. Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society B*, 58:501–514, 1994.

[24] J. Geweke and R. Meese. Estimating Regression Models of Finite but Unknown Order. *International Economic Review*, 22(1):55–70, Februar 1981.

[25] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman & Hall, 1996.

[26] H. Hruschka. Market Share Analysis Using Semi-Parametric Attraction Models. *European Journal of Operational Research*, 138:212–225, 2002.

[27] J. W. Hutchinson, W. A. Kamakura, and J. G. Lynch jr. Unobserved Heterogeneity as an Alternative Explanation for „Reversal" Effects in Behavioral Research. *Journal of Consumer Research*, 27:324–344, Dezember 2000.

[28] D. C. Jain, N. J. Vilcassim, and P. K. Chintagunta. A Random-Coefficients Logit Brand-Choice Model Applied to Panel Data. *Journal of Business and Economic Statistics*, 12(3):317–328, Juli 1994.

[29] K. Jedidi, H. S. Jagpal, and W. S. DeSarbo. Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity. *Marketing Science*, 16(1):39–59, 1997.

[30] W. A. Kamakura, B. Kim, and J. Lee. Modeling Preference and Structural Heterogeneity in Consumer Choice. *Marketing Science*, 15(2):152–172, 1996.

[31] W. A. Kamakura and G. J. Russell. A Probabilistic Choice Model for Market Segmentation and Elasticity Structure. *Journal of Marketing Research*, 26:379–390, November 1989.

[32] P. J. Lenk and W. S. DeSarbo. Bayesian Inference for Finite Mixtures of Generalized Linear Models with Random Effects. *Psychometrika*, 65(1):93–119, März 2000.

[33] P. J. Lenk, W. S. DeSarbo, P. E. Green, and M. R. Young. Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. *Marketing Science*, 15(2):173–191, 1996.

[34] J. P. LeSage. Applied Econometrics using MATLAB. Technical report, Oktober 1999.

[35] D. McFadden and K. Train. Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, 15:447–470, 2000.

[36] T. Otter, R. Tüchler, and S. Frühwirth-Schnatter. Bayesian Latent Class Metric Conjoint Analysis - A Case Study from the Austrian Mineral Water Market. Report 71, Report series, Juni 2002.

[37] Judea Pearl. *Causality.* Cambridge University Press, 2000.

[38] A. E. Raftery and S. Lewis. How Many Iterations in the Gibbs Sampler? *Bayesian Statistics*, 4:763–773, 1992.

[39] S. Richardson and P. J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society B*, 59(4):731–792, 1997.

[40] U. Schuster and J. Wöckl. Optimal Defense Strategies Under Varying Consumer Distributional Patterns and Market Maturity. In *AMS Annual Conference Proceedings*, volume 27, pages 140–144, 2004.

[41] K. E. Train. *Discrete Choice Methods with Simulation.* Cambridge University Press, 2003.

[42] M. Wedel and W. S. DeSarbo. Market Segment Derivation and Profiling Via a Finite Mixture Model Framework. *Marketing Letters*, 13(1):17–25, 2002.

[43] J. Wöckl and U. Schuster. Derivation of Stationary Optimal Defensive Strategies Using a Continuous Market Model. In *AMS Annual Conference Proceedings*, volume 27, pages 305–311, 2004.

[44] A Zellner. *An Introduction to Bayesian Inference in Econometrics.* Wiley, 1971.

Curriculum Vitae

Persönliche Daten: DI Ulrike Schuster
geboren am 4. März 1979 in Stockerau
Österreichische Staatsbürgerschaft
ledig

Schulausbildung:

- 2002-2005: Doktoratsstudium der Technischen Wissenschaften
- 1997-2002: Studium der Wirtschaftsmathematik, TU Wien
  2. Diplomprüfung: 31. 05. 2002 (Abschluss mit Auszeichnung)
  1. Diplomprüfung: 27. 12. 1999
- 1989-1997: Bundesrealgymnasium Stockerau

Beruflicher Werdegang:

- seit Dezember 2004: BAWAG-Bankangestellte
- 2002-2004: Forschungsassistentin an der WU bzw. TU Wien
- Juli 2001: Haas Waffelmaschinen (Ferialpraxis – Buchhaltung)
- Juli/August 2000 sowie Juli 1995: ITT Industries (Ferialpraxis - Registratur)

Publikationen:

- Conference Paper: U. Schuster and J. Wöckl: *„Optimal defense strategies under varying consumer distributional patterns and market maturity"*, AMS Annual Conference, vol. XXVII, pp. 140-144, 2004.
- Conference Paper: J. Wöckl and U. Schuster: *„Derivation of stationary optimal defensive strategies using a continuous market model"*, AMS Annual Conference, vol. XXVII, pp. 305-311, 2004.
- Buchbeitrag: C. Buchta, J. Mazanec, U. Schuster and J.Wöckl: *„Modeling Artificial Consumer Markets"*, Abschlussbericht des SFB010, 2003.
- Paper: U. Schuster and J. Wöckl: *„Optimal defense strategies under varying consumer distributional patterns and market maturity"*, Journal of Economics and Management.
- Diplomarbeit: *„Ökonometrische Analyse des Einflusses der Werbung auf den Umsatz: Der Deutsche Weichspülermarkt"*, TU Wien, 2002.

Fremdsprachen: Englisch - fließend in Wort und Schrift
Französisch – in Wort und Schrift
Italienisch - Grundkenntnisse