Zur Vorhersage ranggeordneter Bewertungen auf „unvollständiger“ Datengrundlage im Marketing

Zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für Wirtschaftswissenschaften der Universität Karlsruhe (TH) vorgelegte Dissertation

von

Dipl.-Phys. Volker Schlecht

Tag der mündlichen Prüfung: 7. November 2007
Referent: Prof. Dr. W. Gaul
Korreferent: Prof. Dr. G. Bol
2007 Karlsruhe

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Die weitgehende Verbreitung interaktiver Medien ermöglicht die kostengünstige Erhebung gigantischer Mengen an Daten über das Verhalten von Konsumenten.

Bedingt durch die fortschreitende Entwicklung der Informationstheorie stehen bereits zum jetzigen Zeitpunkt Mittel zur Verfügung, das Internet zur Datenerhebung zu nutzen, diese Daten zu speichern und mittels quantitativer Verfahren zu analysieren. Die auf diese Weise erhaltenen Ergebnisse ermöglichen den gezielten Einsatz von Marketing-Instrumenten.

Beispielsweise werden bereits via Internet durch diverse kommerzielle Empfehlungssysteme oder Recommender-Systeme Präferenzdaten im Hinblick auf Unterhaltungsprodukte (Filme, Musik, Bücher, Zeitungsartikel) erhoben. Die Nutzer dieser Systeme bewerten einige von sehr vielen möglichen Produkten u.a. in der Hoffnung, daß ihnen das Recommender-System auf Basis dieser Informationen neue Unterhaltungsprodukte vorschlagen kann, die für sie von besonderem Interesse sind. Ein Online-Store, der ein solches Recommender-System benutzt, kann zum einen das Ziel verfolgen, den Abverkauf zu steigern, und zum anderen durch besonders treffende Empfehlungen die Kundenbindung erhöhen. Die so gewonnenen Daten könnten aber auch für den gezielten Einsatz von Marketing-Aktionen verwendet werden. Zudem wäre es wünschenswert, diese Daten wie alle Präferenzdaten zur Neuprodukteinführung nutzen zu können.

Allerdings scheint diesem Unterfangen die besondere Struktur der via Internet erhobenen Daten hinderlich zu sein. Denn meistens kennen die Nutzer dieser Systeme nur einen kleinen Anteil der Items und sind überdies häufig nicht bereit, alle ihnen bekannten Items zu bewerten. Dadurch kommt es dazu, daß für jeden Nutzer nur Bewertungen hinsichtlich einer im Verhältnis zur Gesamtmenge klei-

nen Menge an Items vorliegen. Die auf diese Weise entstandene Fehlendstruktur der Daten erschwert ihre Auswertung erheblich.

Da es sich bei den ranggeordneten Bewertungsdaten um Daten ordinalen Skalenniveaus handelt, ist außerdem fraglich, ob Modelle, die das ordinale Skalenniveau der Bewertungsdaten berücksichtigen, zur Bestimmung von Schätzern für Bewertungsdaten nicht besser geeignet sind.

Die Auswertung dieser Daten kann vielfältigen Zwecken dienen. In der Literatur steht vor allem die Generierung möglichst zutreffender bzw. nützlicher Empfehlungen im Vordergrund. Hierbei bestimmt man i.d.R. für jeden Nutzer auf der Basis der von ihm abgegebenen Bewertungen Schätzer für die Bewertungen der Items, die dieser Nutzer noch nicht bewertet hat. Die Items, deren geschätzte Bewertungen hinsichtlich des betrachteten Nutzers am größten ausfallen oder am wahrscheinlichsten eine hohen Bewertung von dem betrachteten Nutzer erhalten, werden dem Nutzer empfohlen.

Darüberhinaus bieten solche Daten eine kostenlose und breite Grundlage zur Analyse von Präferenzen. Sofern eine Auswertung dieser Daten trotz ihres extrem hohen Fehlendanteils mit hinreichender Genauigkeit gelingt, steht dem Marktforscher eine Vielzahl individueller Information zur Verfügung, die sich zu gezielten Marketing-Aktionen (Direkt-Marketing) nutzen läßt.

Verfahren, die die Schätzung individueller Bewertungen von Items ermöglichen, hinsichtlich derer bisher keine Bewertungen aus dem Kreis der Nutzer abgegeben wurden, kommt besondere praktische Bedeutung zu. Vor allem kann davon ausgegangen werden, daß Empfehlungen, die sich auf neue Items beziehen, hilfreicher als Empfehlungen für allgemein bekannte Items sind.

Die Entwicklung von Verfahren, die sich zur Prognose von Bewertungen eignen, die sich auf bisher von den Nutzern nicht bewertete Items beziehen, ist aber auch vor einem anderen Hintergrund von ökonomischem Interesse. Durch solche Prognosen könnte der Betreiber eines Online-Geschäfts bereits vor der Einführung eines Produkts abschätzen, für wieviele und welche seiner Kunden, die das Empfehlungssystem nutzen, dieses Produkt in Frage kommen könnte. Dies könnte eine vernünftige Grundlage für die Entscheidung über die Aufnahme neuer Produkte ins Sortiment bilden. Sofern das neue Produkt ins Sortiment aufgenommen wird, ist die zugehörige Zielgruppe wenigstens zum Teil identifiziert und kann durch gezielte Marketing-Aktionen angesprochen werden. Zudem erlauben diese Verfahren sogar Prognosen hinsichtlich noch nicht existenter Produkte

	bekannte Items	neue Items
bekannte Nutzer	in der Literatur überwiegend vorausgesetzte (idealisierte) Situation (kollaborative Verfahren sind anwendbar)	Neues-Item Problem (bisher in der Literatur kaum betrachtet)
neue Nutzer	Neuer-Nutzer Problem (Kaltstartproblem)	Neues-Items Problem/ Neuer-Nutzer Problem

Abbildung 1.1: Skizze der unterschiedlichen Vorhersage-Situationen in Anlehnung an Ansari et. al. (2000)

und könnten somit theoretisch selbst zur Entwicklung neuer Produkte genutzt werden.

Im Rahmen dieser Arbeit werden alle Items, hinsichtlich derer keiner der Nutzer eine Bewertung abgegeben hat, die zur Bestimmung von Schätzern eingesetzt werden kann, als neue Items bezeichnet. Insgesamt darf davon ausgegangen werden, daß der Vorhersage individueller Bewertungen neuer Items eine extrem hohe ökonomische Bedeutung zukommt. Entsprechend wird ein Nutzer, der bisher noch keine Bewertung abgegeben hat, als neuer Nutzer bezeichnet. Abbildung 1.1 stellt unterschiedliche Vorhersage-Situationen dar. Die bisherige Literatur befaßt sich beinahe ausschließlich mit dem Szenario, in dem sowohl die betrachteten Nutzer als auch die behandelten Items bekannt sind. Die für die Praxis erheblich wichtigere Situation, in der es um Vorhersagen bezüglich neuer Items geht, bleibt dagegen von der Wissenschaft weitgehend unbeachtet.

Ein möglicher Grund für diese beklagenswerte Praxisferne ist, daß in der Situation, in es nur um die Vorhersage von Bewertungen bekannter Nutzer hinsichtlich bekannter Items geht, die sogenannten kollaborativen Verfahren anwendbar sind. Kollaborative Verfahren benutzen nur gegebene Bewertungen um die nicht

vorhandenen Bewertungen vorherzusagen. Dagegen verwenden andere Verfahren darüberhinaus Zusatzinformation, die oft nur schwer zu beschaffen ist und sind nicht selten mit erheblichem Datenerhebungsaufwand verbunden.

Sofern die Vorhersage von Bewertungen hinsichtlich neuer Items beabsichtigt wird, liegt das Neues-Item Problem vor. Geht es um die Prognose von Bewertungen neuer Nutzer, spricht man von dem Neuer-Nutzer Problem.

Ein wichtiges Ziel diese Arbeit ist, aufbauend auf einer detaillierten Analyse bisher existierender Verfahren neue Verfahren herzuleiten, die zur Lösung des Neues-Item Problems dienen können.

Genau wie die Vorhersage von Bewertungen in Bezug auf neue Items Hintergrundinformation über die Eigenschaften der Items erfordert, ist für die Vorhersage von Bewertungen von neuen Nutzern Information über die Eigenschaften der Nutzer hilfreich. Es kann davon ausgegangen werden, daß viele Nutzer nicht bereit sein werden, einem Empfehlungssystem persönliche Daten preiszugeben. Zudem können Fragen nach persönlichen Eigenschaften (Alter, Geschlecht, Schulbildung, usw.) bei Interessenten leicht zu Widerwillen oder sogar zur Nicht-Benutzung des Empfehlungssystems oder gar zum Verlassen des Online-Geschäfts führen, das solche Information erheben will. Deshalb kann angenommen werden, daß sich die Betreiber kommerzieller Empfehlungssysteme i.d.R. gegen das explizite Erfragen des Hintergrunds der Nutzer entscheiden dürften. Daher kann davon ausgegangen werden, daß in der Praxis nur selten Hintergrund-Information in Bezug auf die Nutzer erhoben werden kann. Somit kommt der Vorhersage von Bewertungen neuer Nutzer (bzw. dem Neuer-Nutzer Problem) untergeordnete praktische Bedeutung zu. I.d.R. werden neuen Nutzern deshalb entweder keine Items oder solche, die sich allgemeiner Beliebtheit erfreuen, empfohlen.

Nichtsdestotrotz ist die Qualität der ersten Empfehlungen meist ausschlaggebend für die weitere Nutzung eines Empfehlungssystems. Daher ist es besonders wichtig, daß ein Verfahren bereits auf Basis weniger Bewertungen zur Generierung möglichst hilfreicher Empfehlungen geeignet ist. Deshalb wird im Rahmen der empirischen Untersuchungen (Kapitel 5 bis 9) insbesondere auch die Eignung der Verfahren zur Schätzung auf Basis weniger Daten eingehend untersucht.

Da die Nutzer sinnvollerweise nur Bewertungen in Bezug auf Gegenstände abgeben können, die ihnen bekannt sind, und man meist hauptsächlich nur mit Items vertraut ist, von denen man zumindest irgendwann geglaubt hat, daß sie einen interessieren könnten, muß davon ausgegangen werden, daß die abgegebenen

Bewertungen im allgemeinen deutlich höher oder niedriger sind als sie hinsichtlich der übrigen Items ausgefallen wären, wenn diese bewertet worden wären. Bei Unterhaltungsprodukten werden sich die Konsumenten i.d.R. eher mit Items vertraut machen, von denen sie annehmen, daß sie ihnen gefallen könnten. Da den Konsumenten in Bezug auf Unterhaltungsprodukte meist genügend Quellen zur Verfügung stehen, anhand derer sie sich vor dem Konsum einen Eindruck davon verschaffen können, ob ihnen ein bestimmtes Item gefallen dürfte, ist anzunehmen, daß die konsumierten Items meistens zumindest nicht allzu weit hinter den Erwartungen der Nutzer zurückbleiben. Sofern die Kosten für den Konsum eines solchen Unterhaltungsprodukts als gering empfunden werden, besteht auch nicht die Gefahr, daß in ihren Erwartungen getäuschte Nutzer aus Wut Bewertungen abgeben, die signifikant schlechter sind als die Bewertungen, die sie als neutrale Beurteiler vergeben hätten. Man steht somit meist vor der Aufgabe, auf Basis der Bewertungen für Items, die von den jeweiligen Nutzern i.d.R. als besser als die meisten übrigen Items eingestuft werden, Schlußfolgerungen hinsichtlich aller Items zu ziehen. Diese Verzerrung wurde in der Literatur nicht untersucht. Dennoch ist insbesondere die Eignung der Verfahren zur Generierung nützlicher Empfehlungen auf Basis verzerrter Daten von besonderer praktischer Bedeutung. Deshalb werden die wichtigsten der im Rahmen dieser Arbeit vorgestellten Verfahren auch auf ihre Eignung zur Analyse verzerrter Daten untersucht.

Zu Beginn der Arbeit werden im Rahmen von Kapitel 2 die traditionellen Verfahren zur Analyse ordinaler Daten auf ihre Eignung zur Analyse ranggeordneter Bewertungsdaten untersucht.

Kapitel 3 gibt einen kurzen Überblick über die wichtigsten Verfahren zur Analyse unvollständiger Daten. In Bezug auf die dargestellten Methoden ist problematisch, daß sie für Datensätze entwickelt wurden, in denen der prozentuale Anteil fehlender Werte in den erhobenen Datenmatrix i.d.R. gering ist.

Die Aufgabe der Kapitel 2 und 3 ist, die in Kapitel 4 bis 9 vorzustellenden Verfahren besser in den Kontext bereits vorher existierender Methoden einzugliedern.

Die Kapitel 4 bis 6 widmen sich der Beschreibung der wichtigsten Nicht-Bayes' schen Ansätze zur Schätzung ranggeordneter Bewertungsdaten auf „unvollständiger“ Datengrundlage. Die Methoden werden im Rahmen von Kapitel 7 konzeptionell und empirisch miteinander verglichen. Im Vordergrund steht hierbei die Eignung der resultierenden Schätzer zur Generierung von Empfehlungslisten.

Zu Beginn von Kapitel 8 erfolgt eine kurze Einführung in die Grundlagen der Bayes'shen Statistik. Auf dieser Basis werden dann verschiedene Bayes'sche Verfahren zur Schätzung ranggeordneter Bewertungsdaten auf „unvollständiger" Datengrundlage erklärt. Die Bayes'schen Verfahren werden am Ende des achten Kapitels empirisch eingehend untersucht.

Im Rahmen von Kapitel 9 werden verschiedene Verfahren vorgestellt, die die Schätzung von Bewertungen für neue Items ermöglichen. Diese sind - wie bereits erwähnt - von besonderer ökonomischer Bedeutung. Es werden sowohl Bayes'sche als auch Nicht-Bayes'sche Methoden betrachtet. Am Ende von Kapitel 9 werden alle Verfahren empirisch untersucht. Auf dieser Basis wird eine optimale Strategie zur Empfehlung neuer Items entwickelt. Auch die eventuelle Eignung der vorgestellten Methoden zu Marktforschungszwecken und für Direkt-Marketing Anwendungen wird erörtert.

Kapitel 2

Datengrundlage und Modellierung ordinaler Daten

2.1 Datengrundlage

Ranggeordnete Bewertungsdaten sind im Rahmen der Marktforschung und eines faktenbasierten Marketing-Managements von zentraler Bedeutung. Sie bieten eine Möglichkeit, verschiedene Grade der Zufriedenheit von Kunden in Bezug auf Produkte und Dienstleistungen oder deren Rahmenbedingungen zu quantifizieren. Hierdurch ist es möglich, die diesbezüglichen Auskünfte zahlreicher (potentieller) Kunden zu aggregieren und mittels quantitativer Verfahren zu analysieren.

Das Internet bietet eine kostengünstige Möglichkeit zur Erhebung von Daten. Vielfach ist es möglich, aus dem Navigationsverhalten der Nutzer einer Website auf deren Zugehörigkeit zu bestimmten Gruppierungen zu schließen. Zu diesem Zweck existieren eine Reihe von Verfahren (z.B. Gaul, Schmidt-Thieme (2000), (2001), (2002)). Hierdurch wird die gezielte Auswahl von Angehörigen spezifischer Personengruppen ermöglicht. Diese bietet ihrerseits die Basis zur Ansprache der Betreffenden via Pop-Ups oder (sofern die E-mail Adressen bekannt sind) per E-mail. Üblicherweise werde die auf diese Weise kontaktierten Personen gebeten, online einen Fragebogen auszufüllen. Um den Aufwand für die Befragten zu minimieren und hierdurch die Rücklaufquote zu erhöhen, wird meist nur darum gebeten, zutreffende Felder zu markieren. Sofern komplexere Zusammenhänge wie z.B. die Zufriedenheit mit einem Produkt oder einer Dienstleistung erfragt werden, wird häufig zur Abgabe einer Bewertung aufgefordert. In vielen Fällen sind nur ganzzahlinge Bewertungen von 1 bis 5 möglich. Auf diese Weise entste-

hen Daten, die mit denselben Methoden wie alle übrigen Fragebögen analysiert werden können.

Via Internet Handel treibende Unternehmen sind mit dem Problem konfrontiert, daß ihren Kunden durch den Wechsel zu einem Konkurrenten oft nur geringfügiger Zeitaufwand und vernachlässigbare Kosten entstehen. Hiedurch gewinnen die Kundenzufriedenheit und die Kundenbindung an Bedeutung. Viele Online-Shops oder Online-Geschäfte betreiben daher automatisierte Empfehlungssysteme (sogenannte Recommender-Systeme), die versuchen, ihren Nutzern auf Basis zuvor abgegebener Bewertungen Items zu empfehlen.

Einige Online-Shops führen ein erheblich größeres Sortiment als Geschäfte, die von ihren Kunden aufgesucht werden können und in denen diese durch Umherschauen oder durch die Hilfe eines Verkäufers finden können, wofür sie sich interessieren. Insbesondere für wenig Internet-affine Personen ist es daher nicht immer leicht, in einem Online-Shop das Gesuchte zu finden - insbesondere, wenn man sich gerade nicht an den Namen des betreffenden Gegenstands erinnert. Daher erfüllen Recommender-Systeme einen wichtigen Zweck. Sie identifizieren hinsichtlich jedes Nutzers bestimmte Gegenstände, die ihn persönlich interessieren könnten und tragen somit dazu bei, daß ihre Nutzer im zugehörigen Online-Shop schneller und bequemer etwas finden, woran sie interessiert sein könnten. Zusätzlich wird hierdurch versucht, die persönliche Beratung durch engagiertes und kompetentes Verkaufspersonal zu ersetzen.

Wie bereits erwähnt, benötigen die Recommender-Systeme Bewertungen, um auf dieser Basis hilfreiche personalisierte Empfehlungen abgeben zu können. Daher werden die Nutzer gebeten, Items, die sie bereits kennen, zu bewerten. Da manche Online-Shops ein großes Sortiment haben, kann ein Nutzer weder alle relevanten Items kennen noch wird er die notwendige Zeit und Arbeit aufwenden wollen, die erforderlich wäre, um alle ihm bekannten relevanten Items zu beurteilen. Daher gibt jeder Nutzer nur hinsichtlich eines sehr kleinen Teils der insgesamt angebotenen Items Bewertungen ab. Hierdurch entstehen ranggeordnete Bewertungsdaten, die sich von den mittels eines Fragebogens erhobenen Bewertungsdaten erheblich unterscheiden. Während es nicht unwahrscheinlich ist, daß wenigstens ein Teil der Fragebögen vollständig beantwortet wird, kann hinsichtlich der Bewertungsdaten, die im Hinblick auf das automatisierte Empfehlungssystem relevant sind, nicht davon ausgegangen werden, daß ein Nutzer mehr als einen geringer Anteil der relevanten Items beurteilt. Darüberhinaus ist

es vergleichsweise unwahrscheinlich, daß auch nur eine kleine Gruppe von Nutzern eines Empfehlungssystems genau dieselben Items beurteilt hat.

Folglich ist es keine leichte Aufgabe, diese online-generierten Bewertungsdaten vorherzusagen oder zu analysieren. Es ist eines der Anliegen dieser Arbeit, genauer zu analysieren, inwiefern existierende statistische Methoden oder deren Modifikationen hierzu einen Beitrag leisten können.

Unglücklicherweise sind selten zu kommerziellen Zwecken erhobene Bewertungsdaten für Forschungszwecke verfügbar. Daher verwenden die meisten auf diesem Gebiet tätigen Wissenschaftler ähnliche Daten, die meist von automatisierten Empfehlungssystemen stammen, die versuchen, ihren Nutzern Empfehlungen im Hinblick auf Filme zu geben, die ihnen gefallen könnten.

Der am häufigsten verwendete Datensatz ist der nach dem Recommender-System MovieLens benannte Datensatz, der aus im Zusammenhang mit diesem Empfehlungssystem generierten Bewertungsdaten besteht. Dieses automatisierte Empfehlungssystem wurde vom GroupLens-Forschunsprojekt der Universität Minnesota betrieben. Die MovieLens-Daten (D0) können von der Website dieses Forschungsprojekts heruntergeladen werden ($http://www.grouplens.org/data$). Dieser Datensatz enthält ca. 1 Millionen Bewertungen in Bezug auf 3872 verschiedene Filme, die von 6040 verschiedenen Nutzern des Recommender-Systems MovieLens im Jahr 2000 abgegeben wurden. D0 kann daher als eine Matrix mit 6040 Zeilen und 3872 Spalten aufgefaßt werden. In dieser Matrix fehlen ca. 95,7 % der Einträge. Man spricht in diesem Zusammenhang von einem Fehlendanteil von ca. 95,7 %. Die Filme wurden auf einer ganzzahligen Skala von 1 bis 5 bewertet. Hierbei ist 1 die schlechteste und 5 die bestmögliche Bewertung. Alle 6040 Nutzer haben sich im Jahr 2000 bei MovieLens angemeldet und wurden nach ihrem Geschlecht, ihrem Alter, ihrer Zugehörigkeit zu einer von 20 Berufsgruppen und ihrem ZIP-Code befragt. Allen Daten außer dem Geschlecht und dem ZIP-Code wurden nur als Intervalldaten erhoben. Im Hinblick auf das Alter gehört jeder Nutzer einer von 7 verschiedenen Altersgruppen an. Bezüglich der Berufsgruppe gibt es 20 verschiedene Gruppen. Viele dieser Gruppen fassen nur Personen zusammen, die in derselben Branche arbeiten. So ist aufgrund der Angabe „doctor/health care“ beispielsweise nicht zu erkennen, ob es sich bei der Person um einen Arzt oder eine Krankenschwester handelt. In vielen Fällen kann man nicht anhand der Berufsbezeichnung auf den Bildungsstand schließen. Auch eine eindeutige Zuordnung zu einer bestimmten sozialen Schicht ist auf Basis die-

Abbildung 2.1: Histogramme des MovieLens-Datensatzes (D0) und des Ausschnitts D1 aus dem MovieLens-Datensatz

ser Daten nicht möglich. Jeder der 6040 Nutzer hat mindestens 20 Bewertungen abgegeben.

Von den 3872 Filmen sind der Titel und die Zugehörigkeit zu einem oder mehreren Genres bekannt. Es existieren 18 verschiedene Genres. Auf diese Weise ist nur sehr grobe Information bezüglich der Filme gegeben. Da die Zugehörigkeit eines Films zu einem bestimmten Genre im allgemeinen nicht auf die Qualität des jeweiligen Films schließen läßt, fehlt ein wichtiger Teil der bewertungsrelevanten Eigenschaften der betrachteten Items.

Allen empirischen Untersuchungen in dieser Arbeit liegen Teilmengen des MovieLens-Datensazes (D0) zugrunde. Aus den Bewertungsdaten D0 wurden die drei verschiene Teile D1, D2 und D3 ausgewählt. Die Histogramme zu D0 und D1 finden sich in Abbildung 2.1. Der Teildatensatz D1 umfaßt die Bewertungen von 1067 Nutzern in Bezug auf 418 Items. Der Fehlendanteil des Datensatzes D1 liegt bei 78,9 %. Beide Histogramme unterscheiden sich nicht wesentlich voneinander.

Der Teildatensatz D2 besteht aus den von 2000 verschiedenen Nutzern vorgenommenen Bewertungen in Bezug auf 3043 unterschiedliche Items. Der Fehlen-

Abbildung 2.2: Histogramme der Ausschnitte D2 und D3 aus dem MovieLens-Datensatz

danteil dieses Datensatzes beträgt ca. 94,5 % und ist damit fast genauso hoch wie der Fehlendanteil der gesamten verfügbaren Datenmenge D0.

D3 besteht aus 130164 Bewertungen, die von 2020 verschiedenen Nutzern stammen und sich auf dieselben 418 Filme beziehen, wie die Bewertungen des Teildatensatzes D1. Damit beträgt der Fehlendanteil hinsichtlich D3 84,6 %.

Die Histogramme der MovieLens-Daten und ihrer im Rahmen der empirischen Untersuchungen (in den Abschnitten 5.10, 6.2, 8.4 und 9.3) verwendeten Teilmengen D1, D2 und D3 unterscheiden sich praktisch nicht voneinander. Man erkennt, daß verstärkt hohe und mittlere Bewertungen abgegeben worden sind. Eine geeignete Erklärung hierfür ist, daß die Nutzer nur Filme bewerten können, die sie kennen. Da den meisten Nutzern hauptsächlich Items bekannt sind, von denen sie sich zumindest in der Vergangenheit etwas versprochen haben, ist anzunehmen, daß einem Nutzer Items, die er bewertet hat, durchschnittlich besser gefallen, als ihm die Items gefallen würden, mit denen er sich bislang nicht befaßt hat. Dies erklärt die Tatsache, daß relativ wenig schlechte Bewertungen vorhanden sind, obwohl die meisten Menschen darin übereinstimmen würden, daß es viel mehr

schlechte als gute Filme gibt.

Eine hieraus resultierende Schwierigkeit, der sich erstmals im Rahmen dieser Arbeit gestellt wird, ist, daß sich die Menge der hinsichtlich eines Nutzers bekannten Bewertungen von der zu schätzenden bzw. vorherzusagenden Menge der übrigen Bewertungen wesentlich unterscheiden dürfte.

Ein weiteres Problem im Hinblick auf online-generierte Bewertungsdaten ist, daß aufgrund der Erhebung der Daten Inkonsistenzen in den Daten nicht auszuschließen sind.

Üblicherweise geben die Nutzer zu verschiedenen Zeitpunkten Bewertungen hinsichtlich unterschiedlicher Items ab. Die einmal abgegebenen Bewertungen werden i.d.R. nicht nochmal von dem betreffenden Nutzer miteinander verglichen und auf ihre Konsistenz hin untersucht. Es ist daher möglich, daß ein Nutzer zu einem Zeitpunkt mehrere Bewertungen für Items abgibt, die er überwiegend als mittelmäßig oder schlechter wahrnimmt. Sofern in dieser Menge ein Item enthalten ist, das ihm immerhin besser als die übrigen gefällt, ist denkbar, daß er im Hinblick auf dieses Item eine hohe Bewertung abgibt. Falls derselbe Nutzer später mehrere Items bewertet, die zum größten Teil seinem Geschmack entsprechen, könnte der betreffende Nutzer das Item weniger positiv beurteilen, das ihm in der Vergangenheit besser gefallen hat als verschiedene Items, die er als weniger gut empfunden hat. Da die Nutzer i.d.R. die Gesamtheit der von ihnen abgegebenen Bewertungen nicht auf ihre Konsistenz hin überprüfen, kann es auf diese Weise zu Schwankungen im Bewertungsverhalten kommen. Diese führen zur Ungenauigkeit der Bewertungsdaten und erschweren somit ihre Prognose nicht unwesentlich.

Da die Überprüfung der gesamten Bewertungen auf ihre Konsistenz hin nicht in allen Fällen einfach und i.d.R. zeitaufwendig ist, wird sie von den Nutzern der Recommender-Systeme nicht verlangt. Es ist aber aus Erfahrungen im Zusammenhang mit der Erhebung anderer Bewertungsdaten bekannt, daß es in Bezug auf die Bewertungen einer größeren Anzahl von Alternativen leicht zu Inkonsistenzen kommen kann. Im Rahmen vieler empirischer Studien wird daher versucht, durch die persönliche Befragung des Bewertenden sein Bewertungsverhalten besser zu verstehen und ihm auf dieser Basis dabei zu helfen, seine abgegebenen Bewertungen auf ihre Konsistenz hin zu überprüfen. Dieses Vorgehen erscheint im Zusammenhang mit automatisierten Empfehlungssystemen kaum erfolgversprechend zu sein, da es kostenintensiv ist und außerdem mit

erhöhtem Zeit- und Arbeitsaufwand seitens der Nutzer verbunden ist und diese daher verärgern könnte.

2.2 Modellierung ordinaler Daten

Zu Marktforschungszwecken werden häufig ordinale Daten erhoben. Das sind Daten, deren Werte in einer Rangreihenfolge angeordnet werden können, wobei die Abstände der Daten untereinander nicht gemessen wurden oder generell nicht gemessen werden können (Krantz et. al. (1971), Roberts (1979)). Beispiele für solche Daten im Marketing sind Präferenzdaten für Produkte. So bittet der Online-Retailer Amazon.de häufig seine Kunden, Produkte mit 1, 2, 3, 4 oder 5 Sternen zu bewerten, wobei 5 Sterne die bestmögliche Bewertung sind und ein Stern die extremste Möglichkeit ist, mit welcher der Bewertende seinem Mißfallen hinsichtlich des zu beurteilenden Gegenstands Ausdruck verleihen kann. Die Sterne, die ein Kunde bei Amazon.de hinsichtlich ausgewählter Produkte vergibt, spiegeln zunächst einmal wieder, zu welcher Gruppe von Produkten der Gegenstand der Bewertung aus seiner Sicht gehört: Zu den Produkten, mit denen er „sehr zufrieden“, „zufrieden“, „neutral eingestellt“, „unzufrieden“ oder sogar „sehr unzufrieden“ ist. Indem er Sterne verleiht, ordnet der Kunde das Produkt einer Gruppe von Produkten zu, mit denen er sicherlich nicht gleich aber zumindest ähnlich zufrieden ist. Es macht auch keinen Unterschied, ob der Bewertende Sterne oder diskrete Zahlen vergibt. In beiden Fällen ordnet er den Bewertungsgegenstand lediglich einer Kategorie zu. Die Abstände zwischen den Zahlen entsprechen nicht notwendig den Abständen zwischen den Kategorien aus der Sicht des Befragten. Es ist mehr als fraglich, ob die Differenz der Zufriedenheit eines Kunden mit zwei verschiedenen Produkten, wovon er einem 4 und dem anderen 5 Sterne zuerkannt hat, der Differenz der Zufriedenheit desselben Kunden mit zwei Produkten entspricht, von denen er eins mit 3 Sternen und das andere mit 4 Sternen bewertet hat. Unabhängig davon, ob die Bewertung durch Zahlen oder Sterne erfolgt, sollten daher Bewertungsdaten immer als ordinale Daten aufgefaßt werden.

Ordinalen Daten können unterschiedliche Entstehungsmechanismen zugrundeliegen. Dementsprechend unterscheidet Anderson (1984) zwischen Gruppierten Kardinalen und Erhoben-Ordinalen Daten. Während Gruppierte Kardinale Daten prinzipiell auch als kardinale Daten hätten erhoben werden können, wäre dies bei den Erhoben-Ordinalen Daten unmöglich gewesen. Ein Beispiel für Gruppier-

te Kardinale Daten sind Zuordnungen zu Einkommensintervallen. Hier wird eine kardinale Variable nur auf andere Weise abgefragt. Im Gegensatz dazu entstehen Erhoben-Ordinale Daten immer dann, wenn aus Sicht des Befragten nicht eine einzige kardinale Größe existiert, deren Ausprägung entscheidend für seine Beurteilung sein sollte. Daher wären die MovieLens-Daten ein Beispiel für Erhoben-Ordinale Daten.

Prinzipiell kann es sich bei Bewertungsdaten sowohl um Gruppierte Kardinale Daten als auch um Erhoben-Ordinale Daten handeln. Bewertungsdaten sind dann Gruppierte Kardinale Daten, wenn eine einzige kardinale Variable alleinige Grundlage der Bewertung ist. Beispiel hierfür sind die Bewertung des Benzinverbrauchs eines Automobils oder die Noten in einer Klausur.

In diesem Kapitel werden ordinale Zufallsvariablen mit Y^o bezeichnet. Die Zufallsvariable Y^o hat die möglichen Realisationen $\{1, ..., C\}$.

Insbesondere wenn $C \geq 5$ ist, wird in Literatur und Praxis häufig die ordinale Struktur der Daten vernachlässigt. Die Zufallsvariable Y^o wird in diesen Fällen als kardinalskalierte Größe mit unsystematischem Meßfehler aufgefaßt.

Den meistbenutzten Ansatz zur Regression ordinaler Daten Y^o stellt der sogenannte Schwellenwertansatz nach Edwards, Thurstone (1952) dar. Der Grundgedanke des Schwellenwertansatzes ist, daß jeder ordinalen Zufallsvariable Y^o eine latente kontinuierliche Variable Z zugrundeliegt, durch die der Wert von Y^o vollständig determiniert ist. Für Erhoben-Ordinale Daten ist Z eine rein theoretische Größe. Im Falle Gruppierter Kardinaler Daten ist die latente Variable diejenige kardinale Größe, die die Befragten einem bestimmten Intervall zugeordnet haben.
Wie erwähnt kann die ordinale Zufallsvariable Y^o die Werte $1, ..., C$ annehmen. Mit

$$-\infty < \gamma_1 < \gamma_2 < ... < \gamma_C \equiv \infty$$

und $\gamma_0 = -\infty$ als Klasseneinteilung postulieren Edwards, Thurstone (1952) zwischen der Zufallsvariable Y^o und der latenten Variable Z den Zusammenhang

$$Y^o = c \quad \Leftrightarrow \quad \gamma_{c-1} < Z \leq \gamma_c. \tag{2.1}$$

Hier ist $c \in \{1, ..., C\}$. Auf der Formel (2.1), die als Schwellenwertansatz bezeich-

net wird, basieren alle üblicherweise verwendeten Modelle zu Beschreibung und Verfahren zur Prognose ordinaler Daten.

2.2.1 Die Kumulativen Modelle

Die Kumulativen Modelle beruhen neben dem Schwellenwertansatz auf der Annahme, daß es sich bei den Komponenten des Vektors $\underline{x} = (\underline{x}_1, ..., \underline{x}_\Omega)'$ um Einflußgrößen handelt, die die latente Variable Z in linearer Form bestimmen. Daher gilt

$$Z = \underline{x}'\breve{b}^o + \epsilon^o,$$

wobei $\breve{b}^o = (\breve{b}^o_1, ..., \breve{b}^o_\Omega)'$ ein Parametervektor ist und ϵ^o eine Störvariable, für die die Beziehung $E(\epsilon^o) = 0$ gilt und deren Verteilungsfunktion F gegeben sei. Damit folgt auf der Basis des Schwellenwertansatzes (Formel (2.1)) nach Edwards, Thurstone (1952) das sogenannte Kulumative Modell

$$P(Y \leq c|\underline{x}) = F(\gamma_c - \underline{x}'\breve{b}^o).$$

Je nachdem, welche Verteilungsfunktion F verwendet wird, erhält man eine andere Variante dieses Kumulativen Modells. Es gilt

$$F(\gamma_c - \underline{x}'\breve{b}^o) = \begin{cases} \Phi(\gamma_c - \underline{x}'\breve{b}^o), & \text{für das Ordinale Probit-Modell} \\ \dfrac{\exp(\gamma_c - \underline{x}'\breve{b}^o)}{1 + \exp(\gamma_c - \underline{x}'\breve{b}^o)}, & \text{für das Kumulative Logit-Modell} \end{cases}.$$

Sofern F die Verteilungsfunktion der Standardnormalverteilung Φ ist, ergibt sich das Ordinale Probit-Modell. Bei Wahl der logistischen Verteilungsfunktion für F ergibt sich Kumulative Logit-Modell, für das die Beziehung

$$\log\left(\frac{P(Y^o \leq c|\underline{x})}{P(Y^o > c|\underline{x})}\right) = \gamma_c - \underline{x}'\breve{b}^o, \;\; c = 1, ..., C-1,$$

charakteristisch ist. Das Ordinale Probit-Modell und das Kumulative Logit-Modell sind die meistverwendeten Varianten des Kumulativen Modells. Im Rahmen des

Kumulativen Modells müssen sowohl Schätzer für $\gamma_1, ..., \gamma_{C-1}$ als auch Schätzer für $\breve{b}^o$ ermittelt werden.

2.2.2 Die Intervalldaten-Ansätze

Besteht die Datenbasis aus Gruppierten Kardinalen Daten und sind außerem die Intervallgrenzen $\gamma_1, ..., \gamma_{C-1}$ bekannt, wird der sogenannte Intervalldaten-Ansatz verwendet, der ebenfalls auf dem Schwellenwertansatz nach Edwards, Thurstone (1952) basiert (siehe Formel (2.1)). Die einfachste und bekannteste Variante des Intervalldaten-Ansatzes geht zudem von der Annahme aus, daß die latente Variable Z normalverteilt ist (Hasselblad et. al. (1980), Stewart (1983)). Im Rahmen dieses Ansatzes gilt

$$Z = \breve{\underline{b}}_0 + \underline{x}'\breve{\underline{b}} + \underline{\epsilon}, \quad \text{mit} \quad E(\underline{\epsilon}) = 0 \text{ und } var(\underline{\epsilon}) = \underline{\sigma}^2.$$

Der Parameter $\breve{\underline{b}}_0$ beschreibt den Achsenabschnitt und für die übrigen Parameter gilt $\breve{\underline{b}} = (\breve{\underline{b}}_1, ..., \breve{\underline{b}}_\Omega)'$. Ebenso wie die Kumulativen Modelle gehen auch die Intervalldatenansätze (Hasselblad et. al.(1980), Stewart (1983), Cameron (1987)) von einer der beobachtbaren Variable Y^o zugrundeliegenden kardinalen Variable Z aus. Anders als bei den Kumulativen Modellen ist dies aber nicht eine latente kardinale Variable, deren Wert weder der Versuchsperson noch dem Experimentator bekannt sein muß, sondern es wird beim Intervalldaten-Ansatz vorausgesetzt, daß die zugrundeliegende latente Größe zwar der Versuchsperson aber nicht dem Experimentator bekannt ist. Die Versuchsperson ordnet die ihr bekannte zugrundeliegende Variable einem Intervall zu. Dem Experimentator sind dann neben dem gewählten Intervall, das die Bewertung $c \in \{1, ..., C\}$ enthält, auch die Grenzen γ_{c-1} und γ_c dieses Intervalls bekannt. Somit weiß der Experimentator über zwei voneinander verschiedene Intervallen zwar immer, welchem von beiden Intervallen die größere latente kardinale Größe zugeordnet ist, aber er kann nicht sagen, wie groß der genaue Unterschied zwischen den zugrundeliegenden latenten Variablen ist. Aus der Sicht des Experimentators handelt es sich daher um ein Modell zur Beschreibung ordinaler Variablen. Meistens bestimmt der Experimentator die Intervallgrenzen (zum Beispiel: Long, Caudill (1991)). Es existieren aber auch Modifikationen, die es erlauben, daß jede Versuchsperson eigene Intervallgrenzen vorgibt (Poon et. al. (1993)).

Generell sind die Intervalldaten-Ansätze insbesondere für Marktforschungszwecke eine geeignete Methode. Häufig werden keine absoluten Beträge erhoben, sondern nur, in welchem Intervall sich der betreffende Betrag befindet. Die Motivation für diese Datenerhebungsstrategie ist in der Regel die Vermeidung fehlender oder fehlerhafter Daten. Insbesondere bei Daten, bei deren Preisgabe der Befragte um seine Privatsphäre fürchten könnte (wie beispielsweise das verfügbare Haushaltseinkommen), will man so Antwortverweigerungen oder bewußte Falschinformationen unterbinden. Desweiteren kann auch nicht immer davon ausgegangen werden, daß allen Versuchspersonen die zu erfragenden Sachverhalte genau bekannt sind.

2.2.3 Die lineare Approximation

In Anwendungen wird häufig die ordinale Struktur der Daten einfach ignoriert. Statt dessen wird meist ein linearer Zusammenhang zwischen der (in Wahrheit ordinalen) endogenen Variable und den exogenen Variablen unterstellt (z.B. Mild, Natter (2002), Dillon et. al. (2001), Ansari et. al. (2000)). Diese Näherung muß nicht in allen Fällen zu schlechteren Ergebnissen führen. Fraglich ist aber, ob und inwieweit sich durch Modelle, die dem ordinalen Skalenniveau der Bewertungsdaten Rechnung tragen, im Hinblick auf die Vorhersage ranggeordneter Bewertungsdaten bessere Ergebnisse erzielen lassen. Im Rahmen dieser Arbeit werden mehrere Methoden zur Vorhersage von Bewertungsdaten vorgestellt (vgl. Abschnitt 5.5 und Abschnitt 5.8), die auf dem Schwellenwertansatz nach Edwards, Thurstone (1952) basieren. Diese werden im Rahmen dieser Arbeit empirisch mit Prognosenverfahren, die das ordinale Skalenniveau vernachlässigen, verglichen (vgl. Abschnitte 5.10, 6.2, 8.4 und 9.3).

Kapitel 3

Klassische Ansätze für den Umgang mit „unvollständigen“ Datenmatrizen

Im folgenden Kapitel wird sich die Darstellung der klassischen Ansätze für den Umgang mit „unvollständigen“ Daten auf Verfahren beschränken, die „unvollständige“ Daten in Datenmatrizen verarbeiten. In marketingrelevanten Problemstellungen liegen viele Daten in Form von Datenmatrizen vor. Eine Stichprobe von (I) Versuchspersonen, die zu ihren (J) Konsumgewohnheiten befragt wurde, läßt sich durch eine Datenmatrix der Dimension $I \times J$ beschreiben. Ebenso kann ein Großteil der Daten, die Recommender-Systeme als Datenbasis zum Generieren personalisierter Empfehlungen verwenden, als Datenmatrizen $y = (y_{ij})$ interpretiert werden. (Die entsprechende Zufallszahl sei $Y = (Y_{ij})$ und wird als ebenfalls als Datenmatrix bezeichnet.) Die Zeilenindizes $i = 1, ..., I$, der Matrix entsprechen den sogenannten Elementen erster Modalität der Matrix. (Das sind die Nutzer des Recommender-Systems.) Entsprechend bezeichnen die Spaltenindizes $j = 1, ..., J$, die Elemente zweiter Modalität von Y. (Im Kontext eines Recommender-Systems werden letztere als Items bezeichnet.) Der Eintrag y_{ij} in der online-generierten Datenmatrix y eines Recommender-Systems ist die Bewertung des Items j durch die i-te Person.

Schnell (1986) hat Missing-Data Probleme im Rahmen geplanter Studien untersucht und sieht den Grund für das Fehlen von Daten in geplanten Studien als Fehler im Analysedesign und im Ablauf der Untersuchung. Zunächst einmal kann durchaus auch ein fehler- oder mangelhaftes Untersuchungsdesign selbst zur

Quelle fehlender Daten werden. Dann ist im Rahmen von Untersuchungen, die die Befragung von Personen erforderlich machen, immer damit zu rechnen, daß einige Probanden die Antwort verweigern. Ursachen hierfür können zwischenmenschliche Spannungen zwischen Befrager und Befragtem oder Zeitmangel in Verbindung mit mangelnder Motivation des Probanden sein. Insbesondere, wenn der Proband einen Fragebogen unbeobachtet und anonym ausfüllen soll, kann der Proband sich durch das Auslassen einiger Fragen Zeit und Mühe ersparen. Manche Fragen können auch dem Befragten zu persönlich erscheinen, was ebenfalls eine Antwortverweigerung nach sich ziehen kann. Dann ist es natürlich auch möglich, daß ein Befragter zu einer Frage keine Auskunft geben kann, weil er über den Sachverhalt nichts weiß. Als weitere Usachen für das Fehlen von Daten werden Unaufmerksamkeit eines Beobachters, Unvollständigkeit von Sekundärdaten und Fehler bei der Codierung und Übertragung von Daten genannt.

Im Gegensatz zu den klassischen im Rahmen einer geplanten Studie erhobenen Datenmatrizen im Marketing weisen die den Recommender-Systemen zugrundeliegenden Datensätze einige Besonderheiten auf. Die fehlenden Werte in der Datenmatrix eines Recommender-Systems sind eine von vornherein systembedingte Eigenschaft und keineswegs als Fehler anzusehen. Im Rahmen von Recommender-Systemen entsprechen die fehlenden Daten genau dem, was vorhergesagt werden soll (um auf dieser Basis personalisierte Empfehlungen zu generieren). Schließlich kann von keinem Benutzer erwartet werden, daß er Items bewertet, die er nicht kennt und es ist sinnlos, einem Nutzer Items zu empfehlen, die er bereits bewertet hat.

Weiterhin hatte man es bei den bisher in der Missing-Data Literatur betrachteten Problemen mit vergleichsweise vollständigen Datenmatrizen zu tun. Der Anteil der fehlenden Daten liegt bei den bisher überwiegend betrachteten Datensätzen zwischen 5 und 10 % (Bankhofer (1995)). Dagegen hat man es bei Recommender-Daten typischerweise mit nur bruchstückhaft vorhandenen Daten zu tun. Beim MovieLens-Datensatz „fehlen“ ca. 95 % der Bewertungs-Daten. Der online-Retailer Amazon.com vertrieb 2003 ca. 2,3 Millionen Produkte (Linden et. al. (2003)). Da kaum anzunehmen ist, daß die Amazon-Kunden in der Lage und bereit waren mehr als jeweils ein paar 100 davon zu bewerten, ist davon auszugehen, daß die Fehlendanteile bei von derart großen online-Shops betriebenen Recommender-Systemen sogar noch erheblich höher ausfallen dürften.

Die Bedingungen, die zum Fehlen von Daten führen, werden als Ausfallme-

chanismus bezeichnet. Man unterscheidet systematische Ausfallmechanismen, bei denen die Daten nicht zufällig fehlen, und unsystematische Ausfallmechanismen, bei denen zufällig Werte fehlen.

Traditionell unterteilt man Y in den beobachteten Anteil Y_{obs} und den nicht beobachteten Anteil Y_{mis}, so daß $Y = (Y_{obs}, Y_{mis})$ gilt. Sei V eine Matrix aus Indikatorvariablen, so daß gilt

$$V_{ij} = \begin{cases} 1, & \text{falls } Y_{ij} \text{ vorhanden ist} \\ 0, & \text{andernfalls} \end{cases} .$$

Rubin (1976) bezeichnet das Fehlen von Daten als zufällig, „missing at random“ (MAR), wenn die Wahrscheinlichkeitsverteilung von V unabhängig von den fehlenden Daten Y_{mis} ist:

$$P(V|Y_{obs}, Y_{mis}, \xi) = P(V|Y_{obs}, \xi). \tag{3.1}$$

v sei die Realisation von V. Hierbei ist ξ ein unbekannter Parameter. Bei einem MAR-Prozeß ist die Antwortrate demnach unabhängig von der Ausprägung der fehlenden Daten.

Beispiel 3.1:

Als Beispiel für den beschriebenen Sachverhalt kann eine Umfrage dienen, bei der nach dem Schulabschluß (Abitur, Realschulabschluß, Hauptschulabschluß, kein Schulabschluß) und dem Bundesland gefragt wird und einige Personen die Antwort in Bezug auf den Schulabschluß verweigern. Wenn die Antwortverweigerungsrate hinsichtlich der Frage nach dem höchsten erreichten Schulabschluß unabhängig von diesem wäre, läge ein MAR-Prozeß vor. Wenn aber Personen mit weniger hohen Schulabschluß eher dazu neigen würden, ihren Schulabschluß nicht anzugeben, dann wäre die MAR-Eigenschaft verletzt und es läge kein MAR-Prozeß vor. Der reinen Datenmatrix kann nicht angesehen werden, ob der Ausfallmechanismus bezüglich eines bestimmten Elements zweiter Modalität die MAR-Eigenschaft besitzt. Hierzu ist über die bloße Datenmatrix hinausgehende Information unabdingbar. Im allgemeinen führen Fragen nach Merkmalen, bei denen bestimmte Ausprägungen von den Befragten als peinlich empfunden werden, zur

Verletzung der MAR-Eigenschaft.

Ist dagegen das Fehlen der Daten nur von den Ausprägungen der fehlenden Daten selbst abhängig, gilt

$$P(V|Y_{obs}, Y_{mis}, \xi) = P(V|Y_{mis}, \xi) \tag{3.2}$$

und der zugrundeliegende Prozeß heißt „observed at random“ (OAR). Offenbar ist die Antwortrate bei einem OAR-Prozeß unabhängig von der Ausprägung der vorhandenen Daten.

Beispiel 3.2:

Die oben dargestellte Befragung wäre beispielhaft für einen OAR-Prozeß, da kein Zusammenhang zwischen der Antwortverweigerungstendenz bezüglich des Merkmals Schulabschluß und dem Bundesland bestehen dürfte. Sei

$$Y_{ij_S} = \begin{cases} 0, & \text{falls } i \text{ keinen Schulabschluß hat} \\ 1, & \text{falls } i \text{ nur den Hauptschulabschluß erworben hat} \\ 2, & \text{falls } i \text{ nur den Realschulabschluß erlangt hat} \\ 3, & \text{falls } i \text{ das Abitur hat} \end{cases}$$

und

$$Y_{ij_B} = \begin{cases} 1, & \text{falls } i \text{ im Bundesland } j_B \text{ lebt} \\ 0, & \text{sonst} \end{cases}, \; j_B \in \{1, ..., 16\}.$$

Es ergibt sich in diesem Zusammenhang als Korrelation

$$\tilde{r}^{VY}_{j_S, j_B} = \frac{\sum_{i=1}^{I} (v_{ij_S} - \bar{v}_{\cdot j_S})(y_{ij_B} - \bar{y}_{\cdot j_B})}{\sqrt{\sum_{i=1}^{I} (v_{ij_S} - \bar{v}_{\cdot j_S})^2} \sqrt{\sum_{i=1}^{I} (y_{ij_B} - \bar{y}_{\cdot j_B})^2}}.$$

Da die Korrelation zwischen dem Fehlen von Daten bei einem Merkmal (Indikatormatrix V) und den Ausprägungen der vorhandenen Daten bei allen ande-

ren Merkmalen (Datenmatrix Y_{obs}) berechenbar (s.o.) und mit Hilfe statistischer Tests auf Signifikanz überprüft werden können, benötigt man keine zusätzliche Information, um die OAR-Eigenschaft zu überprüfen.

Ein Ausfallmechanismus wird „missing completely at random“ (MCAR) genannt, wenn er sowohl zur Klasse der MAR-Prozesse als auch zur Klasse der OAR-Prozesse gehört.

Beispiel 3.3:

Wenn im Beispiel 3.1 nur Daten in Bezug auf das Bundesland fehlen würden, könnte nach der allgemeinen Lebenserfahrung davon ausgegangen werden, daß es sich dann um einen MCAR-Prozeß handeln würde.

Unsystematische Ausfallmechanismen gehen auf Ursachen zurück, die vom Gegenstand der Untersuchung unabhängig sind und nicht mit den Elementen der ersten und zweiten Modalität zusammenhängen. Sie sind allenfalls geeignet, das Untersuchungsergebnis geringfügig zu verzerren. Dagegen liegt es auf der Hand, daß eine Vernachlässigung systematischer Ausfallmechanismen zu erheblichen Verzerrungen des Analyseergebnisses führen kann. Wenn zum Beispiel die Wahrscheinlichkeit, daß eine Frau ihr Alter angibt, mit zunehmendem Alter fällt, sollte eine Vernachlässigung dieses Ausfallmechanismusses zu einer deutlichen Unterschätzung des Durchschnittsalters bei befragten Frauen führen. Daher kann eine Verletzung der MAR-Eigenschaft zu verzerrten Ergebnissen führen, wenn die fehlenden Daten ignoriert werden und nur die vorhandenen Daten zur Schätzung der Modellparameter herangezogen werden. Haben die Daten zwar die MAR-Eigenschaft aber nicht die OAR-Eigenschaft, so kann ein Ignorieren fehlender Werte ebenfalls zu verzerrten Schätzern führen.

Beispiel 3.4:

Ein Beispiel für einen solchen Prozeß wäre eine Untersuchung, in der die Probanden nach ihrer Schulbildung und zu ihren Ansichten zur Einwanderung von

Muslimen nach Europa (vor dem Hintergrund des 11. September 2001) befragt werden und nur im Hinblick auf letztere Frage Antworten verweigert werden. Personen mit höherer Schulbildung werden im allgemeinen sensibeler dafür sein, daß sie bereits durch die Äußerung von leichter Besorgnis als xenophob betrachtet werden könnten. Daher könnte die Antwortverweigerungsrate beim Thema Einwanderung von der Schulbildung aber nicht von der Einstellung zum Thema Einwanderung abhängen.

Auch hier kann ein Ignorieren der fehlenden Werte zu einer Verzerrung der Ergebnisse führen. Allerdings kann hier die beobachtete Information genutzt werden, um die Befragten so in verschiedene Klassen einzuordnen, daß innerhalb dieser Klassen auch die OAR-Eigenschaft gegeben ist (Brick, Kalton (1996)). Im betrachteten Beispiel könnten die Befragten nach ihrer Schulbildung in Klassen aufgeteilt werden. Innerhalb dieser Klassen wäre dann die Antwortverweigerungstendenz beim Thema Einwanderung zufällig und folglich könnten fehlende Werte innerhalb der Klassen ignoriert werden. Solche Gruppierungen können beispielsweise benutzt werden, um die vorhandenen Daten der verschiedenen Klassen unterschiedlich zu gewichten und so der Verzerrung entgegenzuwirken (Brick, Kalton (1996)).

In Bezug auf die meisten Datensätze ist das Erfülltsein der MCAR-Eigenschaft eine unrealistische Voraussetzung. Ein Ignorieren fehlender Daten setzt implizit das Vorliegen der MCAR-Eigenschaft voraus und kann zu Verzerrungen führen (Little, Rubin (2002)). Nach Möglichkeit sollte jeder systematische Ausfallmechanismus bei der Datenanalyse berücksichtigt werden. Die systematischen Ausfallmechanismen sind jedoch selten bekannt. Daher ist eine Berücksichtigung dieser Ausfallmechanismen nur in den seltensten Fällen ohne weitgehende Annahmen möglich und somit ebenfalls problematisch.

3.1 Strukturanalyse

Keine Untersuchungssituation der empirischen Marketingforschung ist in allen Einzelheiten kontrollierbar. Daher sind fehlende Daten niemals a priori auszuschließen. Darüberhinaus können in den wenigsten Untersuchungssituationen die Ursachen für das Fehlen von Daten eindeutig bestimmt werden.

Gleichzeitig ist es aber erforderlich, systematische Ausfallmechanismen für das Fehlen von Daten zu erkennen, da systematischen Ausfallmechanismen bei der Analyse der Daten Rechnung getragen werden muß. Die Strukturanalyse besteht aus einer Vielzahl von Verfahren, die geeignet sind, bestimmte systematische Ausfallmechanismen zu erkennen, deren Nichtberücksichtigung zu erheblichen Verzerrungen bei der Auswertung des unvollständigen Datenmaterials führen kann. Es kann aber nicht jeder systematische Ausfallmechanismus durch Methoden der Strukturanalyse entdeckt werden. Beispielsweise ist es i.d.R. nicht möglich, den Zusammenhang zwischen dem Fehlen von Daten und den Ausprägungen fehlender Werte zu untersuchen. Daher können die Ergebnisse einer Strukturanalyse nur notwendige Bedingung für die Annahme eines systematischen Ausfallmechanismuses aber niemals hinreichende Bedingung für die Annahme eines unsystematischen Ausfallmechanismuses sein. Nur wenn keines der Verfahren der Strukturanalyse zur Annahme eines systematischen Ausfallmechanismuses führt oder sich keine dieser Annahmen statistisch bestätigen läßt, ist es gerechtfertigt, von einem unsystematischen Ausfallmechanismus auszugehen.

Bankhofer (1995) unterteilt die Verfahren zur Strukturanalyse in drei verschiedenen Stufen, die Deskriptive Analyse, die Explorative Analyse und Anwendung statistischer Testverfahren.

Im Rahmen der Deskriptiven Analyse werden Kennzahlen berechnet, die über das Verhältnis von fehlenden zu vorhandenen Werten sowie über das Vorhandensein bestimmter Konzentrationstendenzen der fehlenden Daten in der Datenmatrix Auskunft geben sollen. Es handelt sich hierbei jedoch nur um das Sammeln erster Eindrücke; die Kennzahlen haben nur geringe Aussagefähigkeit über den Ausfallmechanismus.

Mit den Methoden der Explorativen Analyse sollen Abhängigkeiten in der Datenmatrix entdeckt werden. Hierbei wird nach Zusammenhängen innerhalb der Datenmatrix gesucht, die geeignet sind, das Vorliegen eines systematischen Ausfallmechanismuses zu belegen.

Zuletzt werden statistische Tests auf Konzentration der fehlenden Daten in bestimmten Bereichen der Datenmatrix sowie auf systematische Ausfallmechanismen (H_0: Daten fehlen zufällig, H_1: Daten fehlen systematisch) durchgeführt.

Durch die Anwendung statistischer Testverfahren können die mit den Mitteln der Deskriptiven und Explorativen Analyse gewonnenen Erkenntnisse und Verdachtsmomente weiter untersucht werden.

3.1.1 Deskriptive Analyse

Im Rahmen der Deskriptiven Analyse besteht zum einen die Möglichkeit, das Ausmaß sowie eventuell vorhandene Konzentrationstendenzen fehlender Werte in der Datenmatrix durch Kennzahlen auszudrücken. Zum anderen kann die Struktur des unvollständigen Datenmaterials auch grafisch veranschaulicht werden. Beide Alternativen sollen erste Anhaltspunkte für umfassendere Untersuchungen liefern.

Missing-Data-Maße (MD-Maße) sind Kennzahlen, die das Ausmaß des Vorhandenseins fehlender Daten sowie gegebenfalls Konzentrationstendenzen der fehlenden Werte in der Datenmatrix beschreiben. Ausgangspunkt für die nachfolgenden Definitionen für Missing-Data-Maße ist die Indikatormatrix V.

Rummel (1970) führte die Anzahl der fehlenden Daten beim i-ten Element der ersten Modalität ein:

$$V_{i.}^{mis} = J - \sum_{j=1}^{J} V_{ij}.$$

Analog läßt sich auch die Anzahl der fehlenden Daten beim j-ten Element der zweiten Modalität einführen:

$$V_{.j}^{mis} = I - \sum_{i=1}^{I} V_{ij}.$$

Dagegen zieht Schnell (1986) es vor, die Anzahl der vorhandenen Werte zu betrachten. Die Anzahl der vorhandene Daten beim i-ten Element der ersten Modalität ist

$$V_{i.}^{obs} = \sum_{j=1}^{J} V_{ij}.$$

Analog ist die Anzahl der vorhandenen Daten beim j-ten Element der zweiten Modalität definiert:

$$V_{.j}^{obs} = \sum_{i=1}^{I} V_{ij}.$$

Insbesondere bei einer großen Anzahl von Elementen der ersten und zweiten Modalität kann es sinnvoll sein, die Anzahl der insgesamt fehlenden Daten zu errechnen:

$$V^{mis} = \sum_{i=1}^{I} V_{i.}^{mis} = \sum_{j=1}^{J} V_{.j}^{mis} = I \cdot J - \sum_{i=1}^{I} \sum_{j=1}^{J} V_{ij}.$$

Ebenso läßt sich auch die Anzahl der vorhandenen Daten berechnen:

$$V^{obs} = \sum_{i=1}^{I} V_{i.}^{obs} = \sum_{j=1}^{J} V_{.j}^{obs} = \sum_{i=1}^{I} \sum_{j=1}^{J} V_{ij}.$$

Der Missing-Data-Indikator für das i-te Element der ersten Modalität wurde von Cohen, Cohen (1975) definiert als:

$$V_{i.}^{ind} = \begin{cases} 1, & \text{falls } V_{ij} = 1 \quad \forall j \in \{1, ..., J\} \\ 0, & \text{sonst} \end{cases}$$

Entsprechend ist

$$V_{.j}^{ind} = \begin{cases} 1, & \text{falls } V_{ij} = 1 \quad \forall i \in \{1, ..., I\} \\ 0, & \text{sonst} \end{cases}$$

der Missing-Data-Indikator des j-ten Elements der zweiten Modalität.

Insgesamt gilt $V^{obs} + V^{mis} = IJ$. Zudem sind auch relative Kennzahlen wie der Anteil von fehlenden (oder vorhandenen) Daten beim i-ten (j-ten) Element erster (zweiter) Modalität oder der Anteil der fehlenden (oder fehlenden) Daten in der Datenmatrix berechenbar.

Diese trivialen Indikatoren dienen dazu, um erste Eindrücke über die in der Datenmatrix enthaltenen Konzentrationstendenzen der fehlenden Werte zu quantifizieren. Zudem sind sie dazu geeignet, den Ausgangspunkt weiterer Untersuchungen zu bilden. (Z.B. können auf ihrer Basis statistische Tests durchgeführt werden.)

Neben dem bloßen Abzählen fehlender oder vorhandener Daten bei einzelnen Elementen der ersten und zweiten Modalität lassen sich auch die Interrelationen zwischen dem Fehlen oder Vorhandensein eines einzelnen Elementes der er-

sten oder zweiten Modalität und dem Fehlen oder Vorhandensein aller übrigen Elemente der ersten oder zweiten Modalität untersuchen. Zu diesem Zweck hat Brown (1983) das folgende Missing-Data-Maß vorgeschlagen:

$$\tilde{\xi}_j = 1 - \frac{\sum_{j' \neq j} \sum_{i=1}^{I} max\{0, V_{ij} + V_{ij'} - 1\}}{\sum_{j' \neq j} V_{.j'}^{obs}}.$$

$\tilde{\xi}_j$ mißt den Verlust, der dadurch entsteht, daß nur paarweise vollständig vorhandene Ausprägungen zwischen dem j-ten Element der zweiten Modalität und allen übrigen Elementen der zweiten Modalität $j' \in \{1, ..., J\} \setminus \{j\}$ berücksichtigt werden. Durch dieses Missing-Data-Maß und das analog zu bildende Missing-Data-Maß ξ_i lassen sich Isolationstendenzen einzelner Elemente der ersten oder zweiten Modalität erkennen und quantifizieren. Beide Kennzahlen sind Elemente des Intervalls [0,1]. Ein Wert von $\tilde{\xi}_j = 0$ drückt aus, daß es durch die Beschränkung auf paarweise vollständige Daten hinsichtlich des j-ten Elements der zweiten Modalität und allen Elementen der ersten Modalität zu keinem Datenverlust kommt. Das ist nur dann möglich, wenn es keine fehlenden Werte in der Datenmatrix gibt. Dagegen würde $\tilde{\xi}_j = 1$ bedeuten, daß das j-te Element der zweiten Modalität bei keinem Element der ersten Modalität mit irgendeinem anderen Element der zweiten Modalität gemeinsam vorhanden ist. Dies bedeutet die vollständige Isolation des j-ten Elements der zweiten Modalität in der gesamten Datenmatrix.

Zusätzlich werden grafische Verfahren eingesetzt, um Ausmaß und Konzentrationstendenzen der fehlenden Daten in der Datenmatrix erkennbar zu machen (Schnell (1986), Little, Smith (1987)). Diese Grafischen Verfahren können auf die komplette Datenmatrix oder einen geeignet reduzierten Teil der Datenmatrix angewendet werden. Durch Sortieren der Elemente der ersten (oder zweiten) Modalität kann eine übersichtliche Darstellung der Missing-Data-Muster erster (oder zweiter) Modalität erreicht werden. Hierzu existieren auch numerische Vorschriften wie die von Schnell (1986) vorgeschlagene Patternvariable, die zur Bestimmung der geeigneten Reihenfolge der Zeilen in der Indikatormatrix V eingesetzt werden kann.

Nicht alle im Datenmaterial durch andere Methoden erkennbaren Abhängigkeiten der fehlenden Werte sind durch grafische Verfahren abbildbar. Zudem sind grafische Verfahren nicht für hochdimensionale Datenmatrizen geeignet.

3.1.2 Explorative Analyse

Die explorative Analyse dient zur Erkennung möglicherweise vorhandener Abhängigkeitsbeziehungen der fehlenden Werte.

Grundsätzlich können im Rahmen der explorativen Analyse zwei verschiedene Ziele verfolgt werden.

Zum einen kann die Abhängigkeit des Fehlens der Daten eines Elements erster (zweiter) Modalität vom Fehlen von Daten bei anderen Elementen erster (zweiter) Modalität Gegenstand der Untersuchung sein. Da das häufige gleichzeitige Fehlen von Daten zu Elementen derselben Modalität ein Hinweis auf einen Zusammenhang zwischen diesen beiden Elementen ist, ist es naheliegend aber keineswegs notwendig, daß die Daten dieser beiden Elemente aus dem gleichen Grund fehlen. Dieser Grund kann sowohl in der Ausprägung der fehlenden Daten, die sich auf die betreffenden Elemente beziehen, als auch in den Ausprägungen im Hinblick auf diese Elemente vorhandener Daten liegen. Weil der zugrundeliegende Prozeß kein MAR-Prozeß wäre, wenn die Ausprägung der fehlenden Daten selbst ursächlich für ihr Fehlen wären, kann das gehäuft auftretende gleichzeitige Fehlen von Daten bezüglich bestimmter Elemente derselben Modalität als indirekter Hinweis auf die Verletzung der MAR-Eigenschaft interpretiert werden. Da dieses Phänomen jedoch auch durch die Ausprägungen von in Bezug zu anderen (vorhandenen) Elementen stehenden Daten verursacht sein könnte, ist es mit gleicher Berechtigung möglich, das simultane Fehlen von Daten zu bestimmten Elementen derselben Modalität als Indiz, das gegen einen OAR-Prozeß spricht, zu werten.

Beispiel 3.4:

Im Rahmen einer Umfrage sind Personen unter anderem nach ihrem Geschlecht, ihrem Alter und ihrem Einkommen gefragt worden. Bei der Auswertung wird beobachtet, daß das Fehlen der Altersangabe stark mit dem Fehlen der Variable Einkommen korreliert ist. Einerseits kann man versuchen, sich diesen Sachverhalt dadurch zu erklären, daß man vermutet, daß ältere Leute meistens konservativer sind und deshalb die Frage nach ihrem Einkommen und Alter beide als zu persönlich empfinden. Dies wäre eine Interpretation, die gegen einen MAR-Prozeß spricht, da nach dieser Interpretation die Ausprägung der Variable Alter

korreliert mit dem Fehlen derselben Variable wäre. Andererseits kann man aber auch vermuten, daß die Tendenz eines Menschen, persönliche Information über sich preiszugeben, nichts mit dem Lebensalter zu tun hat. Wenn man zudem unterstellt, daß Personen mit einem höheren Einkommen geneigter sind, Auskünft über ihr Einkommen zu geben, daß überdies Frauen ihr Alter lieber verschweigen als Männer und daß Männer mehr Geld als Frauen verdienen, würde man sich den obigen Sachverhalt dadurch erklären, daß besonders viele Personen weiblichen Geschlechts Auskünfte hinsichtlich ihres Alters und Einkommens verweigert haben. Damit wäre die Ausprägung der Variable Geschlecht korreliert mit dem Fehlen der Variablen Alter und Einkommen, was einen Hinweis auf das Nichtvorliegen eines OAR-Prozesses bedeuten würde. In beiden Fällen, kann vermutet werden, daß die MCAR-Eigenschaft nicht vorliegt.

Zur Bestimmung der Abhängigkeit des Fehlens der Daten eines Elements erster (zweiter) Modalität vom Fehlen von Daten bei anderen Elementen erster (zweiter) Modalität benötigt man nur die Indikatormatrix v. Mögliche Analysemethoden sind die Korrelationsanalyse, die Faktorenanalyse, die Clusteranalyse und die Dependenzanalyse.

Es ist auch möglich, den Focus der explorativen Analyse auf die Abhängigkeit der fehlenden Werte eines Elements erster (zweiter) Modalität von den vorhandenen Ausprägungen bei anderen Elementen erster (zweiter) Modalität zu richten. Hierdurch beabsichtigt man, direkte Anhaltspunkte für das Nichtvorliegen eines OAR-Prozesses zu erhalten. Man benötigt zu diesem Zweck sowohl die Datenmatrix y als auch die Indikatormatrix v. In diesem Zusammenhang lassen sich eher die Korrelationsanalyse und die Dependenzanalyse einsetzen, da die übrigen Verfahren zwar zur getrennten Analyse der Matrizen y und v aber weniger gut zur Untersuchung eventueller Abhängigkeiten von Einträgen der Matrix v von (vorhandenen) Einträgen der Datenmatrix y geeignet sind.

Auch mittels der explorativen Analyse können Abhängigkeiten der fehlenden Daten von ihren (unbekannten) Ausprägungen nicht erfaßt werden. Das liegt daran, daß die Verfahren der explorativen Analyse nur die Indikatormatrix und in einigen Fällen sowohl die Daten- als auch die Indikatormatrix nutzen. Dadurch können Zusatzinformationen wie die statistische Verteilung der Grundgesamtheit oder der betrachteten Stichprobe nicht verwendet werden. Hierdurch sind auch die Möglichkeiten der explorativen Analyse zur Erkennung von Abhängigkeits-

beziehungen bei fehlenden Daten eingeschränkt. Da nicht alle systematischen Ausfallmechanismen durch das Instrumentarium der explorativen Analyse erfaßt werden können, ist es unzulässig, aus einer ergebnislos verlaufenen explorativen Analyse auf das Fehlen eines systematischen Ausfallmechanismus zu schließen. Desweiteren können mit Hilfe der explorativen Analyse lediglich begründete Vermutungen über den zugrundeliegenden Ausfallmechanismus angestellt werden. Diese Vermutungen können im Rahmen der induktiven Analyse in Hypothesen umgeformt und mit Hilfe statistischer Tests überprüft werden.

3.1.2.1 Korrelationsanalyse

Die Idee, die Korrelationen zwischen dem Fehlen zweier verschiedener Elemente der zweiten (ersten) Modalität zur Analyse der MD-Struktur zu benutzen, geht auf Lösel, Wüstendörfer (1974) zurück.

Der Bravais-Pearson-Korrelationskoeffizient hinsichtlich zweier Elemente der ersten Modalität errechnet sich gemäß

$$r^v_{ii'} = \frac{\sum_{j=1}^{J}(v_{ij} - \bar{v}_{i.})(v_{i'j} - \bar{v}_{i'.})}{\sqrt{\sum_{j=1}^{J}(v_{ij} - \bar{v}_{i.})^2 \sum_{j=1}^{J}(v_{i'j} - \bar{v}_{i'.})^2}}.$$

wobei

$$\bar{v}_{i.} = \frac{1}{J}\sum_{j=1}^{J} v_{ij}.$$

gesetzt wurde. Hierbei bedeutet der Index v im Ausdruck $r^v_{ii'}$, daß die Korrelation anhand der Indikatormatrix v berechnet wird. Anzumerken ist, daß der Bravais-Pearson-Korrelationskoeffizient grundsätzlich nur auf kardinal skalierte Daten angewendet werden darf. Der hier verwendete Bravais-Parson-Korrelationskoeffizient führt wegen der binären Ausprägungen der Indikatormatrix v zu denselben Ergebnissen wie der Spearmanschen Rangkorrelationskoeffizient. Hierdurch ist abgesichert, daß der Bravais-Pearson-Korrelatinskoeffizient trotz des nominalen Skalenniveaus der vorliegenden Indikatormatrix ein geeignetes Maß der Korrelation der Elemente erster Modalität von v ist.

Nach Lösel, Wüstendörfer (1974) spricht es gegen das zufällige Fehlen von Daten, wenn jenseits der Hauptdiagonalen der Korrelationsmatrix, deren Einträge

nach der oben angegebenen Formel berechnet werden, Komponenten existieren, die deutlich von Null verschieden sind. Darüberhinaus bedeutet ein positives Vorzeichen eines signifikant von Null verschiedenen Koeffizienten $r^v_{ii'}$, daß Daten bezüglich i besonders häufig fehlen, wenn auch die Daten hinsichtlich i' nicht vorhanden sind. Dagegen bedeutet ein negatives Vorzeichen eines deutlich von Null abweichenden Werts für $r^v_{ii'}$, daß Information über i gerade dann tendenziell nicht verfügbar ist, wenn Daten bezüglich i' vorliegen. Im Rahmen der Anwendung statistischer Testverfahren werden statistische Testansätze vorgestellt, die dazu geeignet sind, zu überprüfen, wann ein bestimmter Korrelationskoeffizient signifikant von Null verschieden ist.

Zusätzlich läßt sich auch die Korrelation zwischen dem durch die Indikatormatrix v ausgedrückten Fehlen von Daten und den durch die Datenmatrix y selbst beschriebenen Ausprägungen berechnen. Problematisch ist hierbei das uneinheitliche Skalenniveau von v und y. Werden nur die vorhandenen Daten zur Berechnung des Korrelationskoeffizienten verwendet, so kann man nur unter Annahme eines MCAR-Prozesses unverzerrte Korrelationskoeffizienten erwarten. Wenn die MCAR-Annahme gerechtfertigt wäre, bräuchte man gar keine Strukturanalyse mehr vornehmen, da dies bereits die Eigenschaften voraussetzt, deren Verletzung gegebenenfalls mittels der Strukturanalyse nachgewiesen werden soll.

Anstelle der Berechnung einzelner Korrelationskoeffizienten wird in der Literatur auch eine kanonische Korrelationsanalyse von v und y vorgeschlagen (Frane (1978)). Hier stellt sich jedoch wieder das Problem fehlender Daten in der Matrix y, da bei der kanonischen Korrelationsanalyse die Vollständigkeit der Matrix y vorausgesetzt wird. Da die Strukturanalyse das Ziel hat, die Struktur des Ausfallmechanismus zu untersuchen, um ihm danach in geeigneter Weise Rechnung tragen zu können, ist die Anwendung von Verfahren zur Berücksichtigung der fehlenden Daten im Rahmen der Strukturanalyse nicht sinnvoll, da eine angemessene Berücksichtigung der fehlenden Daten erst durch Ergebnisse der Strukturanalyse möglich wird.

3.1.2.2 Faktorenanalyse

Lösel, Wüstendörfer (1974) empfehlen, die Faktorenanalyse auf die Indikatormatrix v anzuwenden. Unter Verwendung der im Rahmen der Korrelationsanalyse berechneten Korrelationsmatrizen $(r^v_{ii'})$ bzw. $(\tilde{r}^v_{jj'})$ läßt sich eine Hauptkomponentenanalyse durchführen.

Hierbei berechnet man die Kovarianzmatrix von v und bestimmt deren Eigenwerte und Eigenvektoren. Die Eigenvektoren bilden die Spalten der sogenannten Ladungsmatrix F. Hierbei ist der zum größten Eigenwert gehörige Eigenvektor die erste Spalte, der zum zweitgrößten Eigenwert gehörige Eigenvektor die zweite Spalte, usw. Zwischen den Matrizen v (v^T) und F besteht der Zusammenhang $vF = X$, wobei X die Faktorwertematrix bezeichnet. Somit gibt F an, in welcher Weise die ursprünglichen Spalten von v mit den Spalten der Faktorwertematrix X (den Faktoren) korreliert sind.

Es kann gezeigt werden, daß die Eigenwerte die zum jeweiligen Faktor gehörigen Varianzen sind. Daher läßt sich aus jedem Eigenvektor der Erklärungsanteil des zugehörigen Faktors berechnen.

Laden mehrere Spalten von v mittels F besonders stark auf einen bestimmten Faktor hoch, so kann davon ausgegangen werden, daß das Fehlen der zugehörigen Elemente zweiter Modalität korreliert ist. Dies widerspricht der Annahme zufällig fehlender Daten und kann somit als Indiz gegen einen MAR- oder OAR-Prozeß gedeutet werden. Dieses Indiz ist umso erster zu nehmen, je größer der Erklärungsanteil des zugehörigen Faktors ist.

3.1.2.3 Dependenzanalyse

Untersuchungsgegenstand der Dependenzanalyse ist der Einfluß einer unabhängigen Variable auf eine abhängige Variable. Im Kontext der Strukturanalyse einer unvollständigen Datenmatrix sind die folgende drei Ansätze relevant.

Möntmann et. al. (1983) beschreiben die Benutzung eines iterativen logistischen Regressionsverfahrens, bei dem die Indikatorvariable abhängige Variable ist und die Elemente der Datenmatrix Y als unabhängigige Variablen fungieren. Die zugehörige Regressionsgleichung ist:

$$V_{ij} = g_1(Y_{ix(j,min)}, ..., Y_{ix(j,max)}) + \tilde{\epsilon}_{ij}, E(\tilde{\epsilon}_{ij}) = 0, \quad \forall i \in \{1, ..., I\} \tag{3.3}$$

Hierbei ist $\{x(j, min), ..., x(j, max)\} \subseteq \{1, ..., J\} \setminus \{j\}$. Es ist zu beachten, daß die Anzahl der Beobachtungen groß im Verhältnis zu $|\{x(j, min), ..., x(j, max)\}|$ sein muß. Ebensogut könnte man die Regressionsgleichung

$$V_{ij} = \tilde{g}_1(Y_{y(i,min)j}, ..., Y_{y(i,max)j}) + \tilde{\tilde{\epsilon}}_{ij}, E(\tilde{\tilde{\epsilon}}_{ij}) = 0 \quad \forall j \in \{1, ..., J\}$$

betrachten, wobei $\{y(i,min),...,y(i,max)\} \subseteq \{1,...,I\} \setminus \{i\}$ ist.

Desweiteren findet sich bei Bankhofer (1995) in diesem Zusammenhang auch der nachfolgende Regressionsansatz

$$Y_{ij} = g_2(V_{i\tilde{x}^*(j,min)},...,V_{i\tilde{x}^*(j,max)}) + \epsilon^*_{ij}, E(\epsilon^*_{ij}) = 0 \qquad \forall i \in \{1,...,I\}. \tag{3.4}$$

Es gilt $\{\tilde{x}^*(j,min),...,\tilde{x}^*(j,max)\} \subseteq \{1,...,J\} \setminus \{j\}$. Die in den Formeln 3.3 und 3.4 wiedergegebenen Ansätze sind problematisch, da die Spalten der Datenmatrix Y fehlende Werte enthalten können. Beide Formeln dienen dazu, gegebenenfalls eine Verletzung der OAR-Eigenschaft zu belegen. Daher dürfen zur Behandlung fehlender Werte in den Spalten von Y auf keinen Fall Methoden eingesetzt werden, die bereits die OAR-Eigenschaft (bzw. die MCAR-Eigenschaft) voraussetzen. (In diesem Zusammenhang ist allenfalls die MAR-Annahme vor dem Hintergrund des Untersuchungszwecks gerechtfertigt.)

Weitere dependenzanalytische Ansätze zur Strukturanalyse verwenden Regressionsgleichungen, in denen sowohl die endogene als auch die exogenen Variablen aus der Indikatormatrix V stammen (Bankhofer (1995)):

$$V_{ij} = g_3(V_{i\tilde{x}(j,min)},...,V_{i\tilde{x}(j,max)}) + \tilde{\epsilon}^{**}_{ij}, E(\tilde{\epsilon}^{**}_{ij}) = 0 \qquad \forall i \in \{1,...,I\}. \tag{3.5}$$

Hier gilt $\{\tilde{x}(j,min),...,\tilde{x}(j,max)\}$. Sofern sich eine Formel 3.5 entsprechende Abhängigkeitsbeziehung belegen läßt, darf davon ausgegangen werden, daß entweder die MAR- oder die OAR-Eigenschaft nicht gegeben ist oder sogar beide Eigenschaften verletzt sind.

Zur Untersuchung der Abhängigkeitsbeziehung 3.3 eignet sich die logistische Regression. Es empfiehlt sich, Gleichung 3.4 mittels der Varianzanalyse zu untersuchen. Beim Analyseansatz 3.5 kann eine Diskriminanzanalyse durchgeführt werden.

3.1.2.4 Clusteranalyse

Die zeilen- bzw. spaltenweise Ähnlichkeit der Missing-Data-Muster in der Indikatormatrix V läßt sich mit Methoden der Clusteranalyse feststellen (Frane (1978)). Es lassen sich aus Elementen der ersten (zweiten) Modalität bestehende Cluster

bilden, die sich hinsichtlich der bei ihnen fehlenden Daten ähneln. Alle Elemente erster (zweiter) Modalität ohne fehlende Information hinsichtlich der zugehörigen Elemente zweiter (erster) Modalität gehören ins selbe Cluster und können von der Analyse ausgeschlossen werden. Falls die Daten zufällig fehlen, sollte kein besonders häufiges Missing-Data-Muster auffindbar sein, weshalb im Hinblick auf alle verbleibenden Elemente entweder keine Klassenstrukturen erkennbar sein sollten oder sich eine hohe Anzahl kleiner Cluster herausbildet. Es sollten immer mehrere Verfahren der Clusteranalyse benutzt werden.

Neben den von Frane (1978) nahegelegten einmodalen Verfahren wäre auch die Anwendung von zweimodalen Klassifikationsverfahren zur Auffindung von häufigen Missing-Data-Strukturen denkbar.

3.1.3 Statistische Testverfahren

Die Aufgabe von statistischen Testverfahren ist die Überprüfung der im Rahmen der deskriptiven und explorativen Analyse gewonnenen Eindrücke. Die in diesem Zusammenhang relevanten Tests lassen sich unterteilen in Tests auf Häufungen fehlender Daten und Tests auf systematische Ausfallmechanismen.

3.1.3.1 Statistische Tests auf Häufungen fehlender Daten

Das Testen auf Häufungen fehlender Daten läßt sich in zwei weitere Unterpunkte aufteilen. Es ist zum einen festzustellen, ob fehlenden Werte noch als seltenes Ereignis aufgefaßt werden können, und zum anderen ist zu analysieren, ob Häufungen fehlender Daten in Bezug auf bestimmte Gruppen von Elementen der ersten (zweiten) Modalität vorliegen.

Wenn die fehlenden Werte als seltenes Ereignis behandelt werden könnten, müßten sie Poisson-verteilt sein (Lösel, Wüstendörfer (1974)). Deshalb wäre zu überprüfen, ob die Werte $V_{i.}^{mis}$, $i \in \{1, ..., I\}$, bzw. $V_{.j}^{mis}$, $j \in \{1, ..., J\}$, einer Poisson-Verteilung entstammen oder nicht:

$$\begin{array}{ll} H_0: & \text{Es gilt } V_{i.}^{mis} \sim f_P(\bullet|\lambda_Z), i = 1, ..., I. \\ H_1: & \text{Die Bedingung unter } H_0 \text{ gilt nicht.} \end{array} \qquad \text{(T3.1)}$$

bzw.

$$\begin{aligned} H_0: &\quad \text{Es gilt } V_{.j}^{mis} \sim f_P(\bullet|\lambda_S), j = 1, ..., J. \\ H_1: &\quad \text{Die Nullhypothese } H_0 \text{ ist falsch.} \end{aligned}$$

Zum Testen der oben aufgestellten Hypothesen ist vorab die Bestimmung des Maximum-Likelihood Schätzers für λ_Z (bzw. λ_S) erforderlich. Auf Basis der Likelihood-Funktion

$$L(\lambda_Z|v_{1.}^{mis}, ..., v_{I.}^{mis}) = \prod_{i=1}^{I} f_P(v_{i.}^{mis}|\lambda_Z) = \prod_{i=1}^{I} \frac{\lambda_Z^{v_{i.}^{mis}}}{V_{i.}^{mis}!} \exp(-\lambda_Z)$$

ergibt sich (durch Maximieren der zugehörigen Log-Likelihood Funktion) der Maximum-Likelihood Schätzer

$$\hat{\lambda}_Z^{ML} = \frac{1}{I} \sum_{i=1}^{I} v_{i.}^{mis}.$$

$A^V(V_{i.}^{mis} = x_S, i = 1, ..., I)$ sei die Anzahl der $V_{i.}^{mis}, i = 1, ..., I$, die den Wert $x_S \in \{0, 1, 2, ...\}$ annehmen.

Unter H_0 (T3.1) müßte die Wahrscheinlichkeit, daß ein $V_{i.}^{mis}, i = 1, ..., I$, den Wert x_S annimmt, durch den Ausdruck

$$f_P(V_{i.}^{mis} = x_S|\hat{\lambda}_Z^{ML}) = \frac{(\hat{\lambda}_Z^{ML})^{x_S}}{x_S!} \exp(-\hat{\lambda}_Z^{ML}).$$

approximiert werden können.

Man unterteilt den Wertebereich $\{0, 1, 2, ...\}$ von x_S in I_{int} Mengen $Be(x_{int})$, $x_{int} = 1, ..., I_{int}$, aus benachbarten Elementen von $\{0, 1, 2, ...\}$, so daß

$$\{Be(x_1), ..., Be(x_{I_{int}})\} = \{0, 1, 2, ...\}$$

gilt. Für die Anzahl der Werte in der Menge $Be(x_{int}), x_{int} = 1, ..., I_{int}$, gilt

$$ABe(x_{int}) = \sum_{x_S \in Be(x_{int})} A^V(V_{i.}^{mis} = x_S, i = 1, ..., I).$$

Die Wahrscheinlichkeit, mit der $V_{i.}^{mis}$ im Intervall $Be(x_{int}), x_{int} = 1, ..., I_{int}$, liegt, ist unter H_0 (T3.1) gegeben durch

$$f_{Be(x_{int})}(V_{i.}^{mis} \in Be(x_{int})|\hat{\lambda}_Z^{ML}) = \sum_{x_S \in Be(x_{int})} f_{Be(x_{int})}(V_{i.}^{mis} = x_S|\hat{\lambda}_Z^{ML}).$$

Die Intervalle $Be(x_{int}), x_{int} = 1, ..., I_{int}$, und ihre Anzahl I_{int} sind so zu wählen, daß $ABe(x_{int}) \geq 5$ oder $If_{Be(x_{int})}(V_{i.}^{mis} \in Be(x_{int})|\hat{\lambda}_Z^{ML}) \geq 5 \; \forall \; x_{int} \in \{1, ..., I_{int}\}$ gilt. Das $(1-\alpha)$-Fraktil der Grenzverteilung der Testfunktion

$$\sum_{x_{int}=1}^{I_{int}} \frac{(ABe(x_{int}) - If_{Be(x_{int})}(V_{i.}^{mis} \in Be(x_{int})|\hat{\lambda}_Z^{ML}))^2}{If_{Be(x_{int})}(V_{i.}^{mis} \in Be(x_{int})|\hat{\lambda}_Z^{ML})}$$

liegt zwischen den Fraktilen $\chi^2_{1-\alpha,I_{int}-1}$ und $\chi^2_{1-\alpha,I_{int}-2}$ der χ^2-Verteilung (vgl. Albrecht (1980)). H_0 ist daher abzulehnen, wenn die Testfunktion größer als das $(1-\alpha)$-Fraktil (bzw. $\chi^2_{1-\alpha,I_{int}-1}$) ist. Bezüglich $V_{.j}^{mis}$ kann analog vorgegangen werden.

Bei der Untersuchung der Frage, ob Häufungen fehlender Daten in Bezug auf bestimmte Klassen von Elementen der ersten (zweiten) Modalität auftreten, kann ebenfalls auf den χ^2-Anpassungstest zurückgegriffen werden. Die Ausgangsüberlegung ist hierbei folgende: Falls die Daten nicht bei einzelnen Elementen signifikant öfter fehlen, sollte das Auftreten der fehlenden Werte im Bereich zwischen dem Minimum und Maximum von $v_{i.}^{mis}$, $i \in \{1, ..., I\}$, ($v_{.j}^{mis}$, $j \in \{1, ..., J\}$) gleichwahrscheinlich sein. Hierbei muß bedacht werden, daß das Minimum von $v_{i.}^{mis}$, $i \in \{1, ..., I\}$, ($v_{.j}^{mis}$, $j \in \{1, ..., J\}$) auch Null sein könnte. Dieser Fall ist aber nicht Gegenstand der Untersuchung, da in diesem Fall keine Daten fehlen. Daher muß die Null als Minimum von $v_{i.}^{mis}$, $i \in \{1, ..., I\}$, ($v_{.j}^{mis}$, $j \in \{1, ..., J\}$) ausgeschlossen werden. Aus diesen Überlegungen ergeben sich die folgenden Testhypothesen für die Elemente der ersten Modalität:

H_0 : Für alle von Null verschiedenen Realisationen von $V_{i.}^{mis}$ gilt $V_{i.}^{mis} \sim f_G(\bullet), i = 1, ..., I.$ (T3.2)

H_1 : Die Bedingung unter H_0 gilt nicht.

Hierbei verwendet man die Wahrscheinlichkeitsfunktion

$$f_G(V_{i.}^{mis}) = \frac{1}{1 + \max\limits_{\iota \in \{1,...,I\}} V_{\iota.}^{mis} - \min\limits_{\substack{\iota \in \{1,...,I\} \\ V_{\iota.}^{mis} \neq 0}} V_{\iota.}^{mis}},$$

wobei zu beachten ist, daß alle $i \in \{1, ..., I\}$ für die $V_{i.}^{mis} = 0$ gilt von der Untersuchung auszuschließen sind, da die Nullhypothese H_0 (T3.2) auf diese Werte nicht anwendbar ist.
Die entsprechende Hypothesen für die Elemente der zweiten Modalität sind:

H_0 : Für alle von Null verschiedenen Werte von $V_{.j}^{mis}$ gilt $V_{.j}^{mis} \sim f_G(\bullet), j = 1, ..., J.$

H_1 : Die Bedingung unter H_0 ist nicht erfüllt.

Die Wahrscheinlichkeitsfunktion ist in diesem Fall:

$$f_G(V_{.j}^{mis}) = \frac{1}{1 + \max\limits_{\iota \in \{1,...,J\}} V_{.\iota}^{mis} - \min\limits_{\substack{\iota \in \{1,...,J\} \\ V_{.\iota}^{mis} \neq 0}} V_{.\iota}^{mis}},$$

wobei wieder zu beachten ist, daß alle $j \in \{1, ..., J\}$ für die $V_{.j}^{mis}$ den Wert Null annimmt nicht Gegenstand der Nullhypothese H_0 (T3.2) sind.

3.1.3.2 Statistische Testverfahren auf systematische Ausfallmechanismen

An erster Stelle ist zu betonen, daß es nicht möglich ist, den statistischen Beweis für die Zufälligkeit des Fehlens von Daten zu erbringen. Es können lediglich bestimmte systematische Ausfallmechanismen ausgeschlossen werden.

Die mit Mitteln der Statistik belegbaren systematischen Ausfallmechanismen lassen sich in drei Arten unterteilen: die Abhängigkeit des Fehlens von Daten vom Fehlen anderer Daten, die Abhängigkeit des Fehlens von Daten von ihren eigenen Realisierungen und die Abhängigkeit des Fehlens von Daten von den Ausprägungen anderer Merkmale.

Da die Indikatormatrix v selbst vollständig ist, ist das Testen der Abhängigkeit des Fehlens von Daten vom Fehlen anderer Daten verhältnismäßig unproblematisch. In diesem Fall ergibt sich die folgende Nullhypothese:

$$H_0: \quad \text{Das Fehlen der Daten hängt nicht vom Fehlen der Daten bei anderen Elementen ab.} \tag{T3.3}$$

Beim Test nach Kim, Curry (1977) wird benutzt, daß

$$X_{KC} = \frac{(h^n - \tilde{h}^n)^2}{\tilde{h}^n} + \frac{(h^m - \tilde{h}^m)^2}{\tilde{h}^m} + \sum_{j \in J^{mis}} \frac{(h_j - \tilde{h}_j)^2}{\tilde{h}_j} \sim \chi^2_{|J^{mis}|+1},$$

wobei $J^{mis} = \{j \in \{1, ..., J\} | V^{ind}_{.j} = 0\}$ und h_j die Anzahl der Elemente erster Modalität bezeichnet, die beim Element j genau eine fehlende Ausprägung aufweisen. h^n ist die Anzahl der Elemente erster Modalität, bei der keine Ausprägung fehlt; h^m bezeichnet die Anzahl der Elemente erster Modalität, bei denen mehr als eine Ausprägung fehlt:

$$h^n = \sum_{i=1}^{I} V^{ind}_{i.} \quad \text{und} \quad h^m = I - h^n - \sum_{j \in J^{mis}} h_j.$$

Alle mit einer Schlange versehenen Größen bezeichen Werte, die man für dieselbe Variable ohne Schlangen-Symbol unter der Voraussetzung erwarten dürfte, daß das Fehlen der Daten unabhängig vom Fehlen anderer Daten ist.
Im einzelnen gilt:

$$\tilde{h}_j = I\tilde{V}^{mis}_{.j} \prod_{j' \in J^{mis}, j' \neq j} \tilde{V}^{obs}_{.j'}, \qquad \tilde{h}^n = I \prod_{j' \in J^{mis}} \tilde{V}^{obs}_{.j'}, \qquad \tilde{h}^m = I - \tilde{h}^n - \sum_{j' \in J^{mis}} \tilde{h}_{j'}.$$

Bei Ablehnung der Nullhypothese H_0 (T3.3) ist die MCAR-Annahme nicht aufrechtzuerhalten. Es kann unter dieser Voraussetzung davon ausgegangen werden, daß die MAR-Eigenschaft oder die OAR-Eigenschaft oder sogar beide Eigenschaften nicht erfüllt sind.

Alternativ kann die Nullhypothese H_0 (T3.3) auch überprüft werden, indem man testet, ob die im Rahmen der Explorativen Analyse auf Basis der Indikatormatrix errechneten Korrelationskoeffizienten signifikant von Null abweichen.

Es besteht hierbei sowohl die Möglichkeit, einzelne Korrelationskoeffizienten einer Korrelationsanalyse zu unterziehen, als auch alle Korrelationskoeffizienten gleichzeitig zu testen (Hartung, Elpelt (1995)).

Die Untersuchung der Abhängigkeit des Fehlens von Daten von ihren eigenen Realisationen ist nur bei über die bloße Datenmatrix hinausgehender Information (wie zum Beispiel die Kenntnis der der Ausgangsstichprobe zugrundeliegenden Wahrscheinlichkeitsverteilung) oder der nachträglichen („Follow-up“) Erhebung weiterer Daten möglich (vgl. Schafer (1997)). Damit ist es mit der unvollständigen Datenmatrix als alleiniger Datengrundlage nicht möglich, die MAR-Eigenschaft direkt zu falsifizieren.

Unter der Voraussetzung einer bekannten Wahrscheinlichkeitsverteilung für die vorhandenen Daten kann das Fehlen der Daten jedenfalls dann nicht dem Zufall zugeschrieben werden, wenn die vorhandenen Daten nicht dieser bekannten Wahrscheinlichkeitsverteilung genügen. Hierzu kann man den χ^2-Anpassungstest (vgl. Snedecor, Cochran (1989)) oder den Kolmogorov-Smirnov-Test (vgl. Rao, Toutenburg (1999)) benutzen. Alternativ kann man auch einzelne Verteilungsparameter überprüfen. Bei dieser Art von Tests vergleicht man Verteilungsparameter der bekannten Wahrscheinlichkeitsverteilung mit ihren Stichprobenäquivalenten. Beispielsweise läßt sich bei bekannter Varianz der Grundgesamtheit die Stichprobenvarianz im Rahmen eines χ^2-Tests für die Varianz mit der Varianz der Grundgesamtheit vergleichen. Ist die Grundgesamtheit normalverteilt, so kann man bei bekanntem Erwartungswert und bekannter Varianz das Abweichen des Stichprobenmittelwerts vom Erwartungwert durch den Einstichproben Gauß-Test überprüfen. Bei unbekannter Varianz und ansonsten gleichen Voraussetzungen kann man den Einstichproben t-Test benutzen. Für eine beliebige Wahrscheinlichkeitsverteilung kann der Mittelwert ab einem Stichprobenumfang von 30 durch den asymptotischen Gauß-Test überprüft werden.

Die dritte Klasse von Testverfahren analysiert die Abhängigkeit des Fehlens von Daten von den Ausprägungen anderer Elemente bei den vorhandenen Daten. Zu diesem Zweck werden die Tests auf Lokationsunterschiede bzw. Unabhängigkeit (Möntmann et. al. (1983)) und der Little-Test (Little (1988)) eingesetzt.

Bei den Tests auf Lokationsunterschiede bzw. Unabhängigkeit wählt man zu Anfang ein Element i (j) erster (zweiter) Modalität mit fehlenden Daten aus. Gibt es viele Elemente i mit $v_{i.}^{ind} \neq 1$, so ist es zweckmäßig, nur die Elemente zu betrachten, bei denen die meisten Daten fehlen (Frane (1978)). Bei allen übrigen

Elementen $i' \neq i$ werden alle vorhandenen Daten gemäß der Indikatorvariablen beim Element i in zwei Gruppen unterteilt, so daß alle $j \in \{1, ..., J\}$ für die $v_{ij} = 1$ gilt einer Gruppe angehören und die Menge

$$\{j \in \{1, ..., J\} | v_{ij} = 0\}$$

die Referenzgruppe bildet. Danach soll getestet werden, ob sich signifikante Unterschiede zwischen den beiden Gruppen hinsichtlich der Ausprägungen der zu den übrigen Elementen $i' \neq i$ vorhandenen Daten feststellen lassen. Falls beide Gruppen nachweislich verschieden sind, hat man gezeigt, daß ein Zusammenhang zwischen dem Fehlen des i-ten Elements und den Ausprägungen zumindest eines anderen Elements derselben Modalität besteht. Dadurch wäre bewiesen, daß die Daten nicht die OAR-Eigenschaft aufweisen.

Bei kardinalem Skalenniveau kommen hierzu unter der Normalverteilungsannahme bei bekannter Varianz der Zweistichproben-Gauß-Test und bei unbekannter Varianz der Zweistichproben-t-Test in Frage. Sofern zwar kardinales Skalenniveau vorausgesetzt werden darf, jedoch keine Normalverteilung vorliegt, kann man bei einem die Anzahl 30 übersteigenden Stichprobenumfang den asymptotischen Zweistichproben-Gauß-Test anwenden. Falls die Daten lediglich ordinal skaliert sind, eignet sich der Zweistichproben-Vorzeichentest (Hartung (1989)).

Der Little-Test (Little (1988)) dient ebenfalls zur Falsifikation der OAR-Annahme. Anders als die bereits vorgestellten Ansätze erspart dieser Test das Überprüfen aller einzelnen Elemente, da er geeignet ist, die OAR-Eigenschaft für die gesamte Datenmatrix zu überprüfen. Er ist allerdings nur in den Fällen anwendbar, in denen die MAR-Eigenschaft als gegeben betrachtet werden kann. Vor diesem Hintergrund wird im Rahmen dieses Testverfahrens versucht, die Nullhypothese

$$H_0: \quad \text{Bestimmte Muster fehlender Daten sind nicht mit besonderen Ausprägungen der vorhandenen Werte verbunden.} \tag{T3.4}$$

abzulehnen. Sofern dies gelingt, ist gezeigt, daß die OAR-Eigentschaft nicht erfüllt ist. Beim Little-Test (Little (1988)) bestimmt man zunächst aufgrund der Indikatormatrix die $N^Z_{pattern}$ ($N^S_{pattern}$) verschiedenen MD-Muster der Zeilen (Spalten) in der Indikatormatrix, so daß alle zur selben Klasse gehörenden Elemente erster (zweiter) Modalität dasselbe MD-Muster $n^Z_p \in \{1, ..., N^Z_{pattern}\}$ (bzw.

$n_p^S \in \{1, ..., N_{pattern}^S\}$) aufweisen. So erhält man $N_{pattern}^Z$ ($N_{pattern}^S$) verschiedene Klassen von Elementen der ersten (zweiten) Modalität. $J_{PZ}(n_p^Z)$ sei die Anzahl der Elemente zweiter Modalität, die in der Klasse $n_p^Z \in \{1, ..., N_{pattern}^Z\}$ vorhanden sind, und J sei die Gesamtzahl der Elemente zweiter Modalität. Man bezeichnet die Menge aller Elemente erster Modalität, die zur n_p^Z-ten Klasse gehören, als $C(n_p^Z)$. Die Matrix $\mathcal{T}(n_p^Z) \in \mathbb{R}^{J,J_{PZ}(n_p^Z)}$ bildet die Datenmatrix $Y \in \mathbb{R}^{I,J}$ auf eine Matrix $\tilde{Y}(n_p^Z) \in \mathbb{R}^{I,J_{PZ}(n_p^Z)}$ ab, die nur die $J_{PZ}(n_p^Z)$ Komponenten zweiter Modalität von Y enthält, die im n_p^Z-ten MD-Muster vorhanden sind: $\tilde{Y}(n_p^Z) = Y\mathcal{T}(n_p^Z)$. Der Zeilenvektor $\mathcal{R}ow(Y,i) \equiv (Y_{i1}, ..., Y_{iJ}) \in \mathbb{R}^{1,J}, i = 1, ..., I$, sei i-te Zeile der Datenmatrix Y.
Auf Basis dieser Definitionen ergeben sich die Vektoren

$$M_p^{Z,obs} = \sum_{i \in C(n_p^Z)} (\mathcal{R}ow(Y,i)\mathcal{T}(n_p^Z))' = \sum_{i \in C(n_p^Z)} \mathcal{T}(n_p^Z)'(\mathcal{R}ow(Y,i))' \in \mathbb{R}^{J_{PZ}(n_p^Z)}$$

und

$$M_p^Z = \sum_{i=1}^{J} \mathcal{T}(n_p^Z)'(\mathcal{R}ow(Y,i))' \in \mathbb{R}^{J_{PZ}(n_p^Z)}.$$

Beides sind Vektoren, deren Komponenten jeweils genau einem Element zweiter Modalität entsprechen, hinsichtlich dessen beim n_p^Z-ten MD-Muster alle Werte vorhanden sind. Es gilt $M_p^{Z,obs}, M_p^Z \in \mathbb{R}^{J_{PZ}(n_p^Z)} \subset \mathbb{R}^J$. Jede Komponente der Vektoren $M_p^{Z,obs}$ und M_p^Z bezieht sich auf ein unterschiedliches Element zweiter Modalität, welches in der Datenmatrix Y hinsichtlich aller zur Klasse $n_p^Z \in \{1, ..., J_{PZ}(n_p^Z)\}$ $M_p^{Z,obs}$ gehörenden Elemente erster Modalität vorhanden ist. $M_p^{Z,obs}$ ist dann der Durchschnittsvektor aller Elemente erster Modalität, die das n_p^Z-te MD-Muster aufweisen. M_p^Z ist der Durchschnittsvektor aller Elemente erster Modalität.

$$S^{n_p^Z} = \sum_{i=1}^{J} (\mathcal{T}(n_p^Z)'(\mathcal{R}ow(Y,i))' - M_p^Z)(\mathcal{T}(n_p^Z)'(\mathcal{R}ow(Y,i))' - M_p^Z)'$$

ist die Varianz-Kovarianz Matrix aller Elemente zweiter Modalität, die beim n_P^Z-ten MD-Muster gegeben sind. Es gilt $S^{n_p^Z} \in \mathbb{R}^{J_{PZ}(n_p^Z),J_{PZ}(n_p^Z)}$.

Unter der Voraussetzung, daß die Bewertungsdaten kardinales Skalenniveau aufweisen und daß die Zeilenvektoren der Datenmatrix Y unabhängig identisch normalverteilt sind, gilt nach Little (1988) näherungsweise

$$d^2 = \sum_{n_p^Z=1}^{N_{pattern}^Z} |C(n_p^Z)|(M_p^{Z,obs} - M_p^Z)'(S^{n_p^Z})^{-1}(M_p^{Z,obs} - M_p^Z) \sim \chi^2_{-J+\sum_{n_p^Z=1}^{N_{pattern}^Z} J_{PZ}(n_p^Z)}.$$

Sofern die Mittelwertsvektoren $M_p^{Z,obs}, n_p^Z = 1, ..., N_{pattern}^Z$, der einzelnen MD-Klassen signifikant von den auf Basis des gesamten Datensatzes gebildeten Mittelwertsvektoren M_p^Z abweichen, ist bewiesen, daß die OAR-Eigenschaft nicht vorliegt.

Problematisch ist, daß auch hier bereits im Rahmen der Strukturanalyse Verfahren zur Berücksichtigung der fehlenden Daten eingesetzt werden müssen, um M_p^Z und $S^{n_p^Z}$ zu berechnen. Da es zu diesem Zweck auch Verfahren gibt, die nicht die MCAR-Eigenschaft (in der die OAR-Eigenschaft enthalten ist), sondern nur die MAR-Eigenschaft voraussetzen, kann der Little-Test zumindest dann verwendet werden, wenn gegen die Annahme eines MAR-Prozesses keine Einwände bestehen.

3.2 Missing-Data Verfahren

Abhängig von der im Rahmen der Strukturanalyse untersuchten Struktur des Ausfallmechanismus muß eine Entscheidung hinsichtlich der Behandlung der fehlenden Daten zu Analysezwecken getroffen werden.

In vielen Fällen ist es naheliegend, die fehlenden Daten zu ignorieren und die Analyse nur auf Basis der vorhandenen Daten zu erstellen. Die entprechenden Verfahren werden in der deutschsprachigen Literatur als Eliminierungsverfahren bezeichnet (Schwab (1991)). Unter Voraussetzung der MCAR-Eigenschaft sind die Ergebnisse einer solchen Analyse unverzerrt (Rubin (1976)). Ist hingegen nur die MAR-Eigenschaft gegeben, so können unter der Voraussetzung einer weiteren noch näher zu erläuternden Eigenschaft Verfahren zur Behandlung der fehlenden Werte eingesetzt werden, die allerdings die Bekanntheit einer Likelihood-Funktion erfordern. Diese sogenannten Parameterschätzverfahren, die auf der Maximum-Likelihood-Theorie oder der Bayes'schen Statistik basieren, führen dann ebenfalls

zu unverzerrten Ergebnissen (Rubin (1976)). Im Abschnitt 3.2.3 wird genauer auf diese Methoden eingegangen.

In allen übrigen Fällen müssen die fehlenden Daten mit Verfahren behandelt werden, die ein Modell des zugrundeliegenden Ausfallmechanismus benutzen, da ansonsten mit verzerrten Ergebnissen gerechnet werden muß (Greenless et. al. (1982)).

3.2.1 Eliminierungsverfahren

Die im folgenden dargestellten Eliminierungsverfahren sind bereits Bestandteil statistischer Softwarepakete und werden häufig in Anwendungen verwendet. Da die MCAR-Annahme nur in den wenigsten Fällen erfüllt ist, ist der Einsatz von Elimierungsverfahren nur dann ratsam, wenn nur ein sehr kleiner Anteil der Daten fehlt (Little, Rubin (2002)). Im Hinblick auf Datenmatrizen unterscheidet man den grundsätzlichen Ausschluß aller Zeilen (Spalten), die fehlende Daten enthalten, und der Verwendung aller vorhandenen Information bei gleichzeitiger Vernachlässigung der fehlenden Daten.

Der Vorteil der Eliminierung aller Zeilen mit fehlenden Einträgen („complete-case analysis“) bzw. aller Spalten mit fehlenden Einträgen („complete-variable case“) liegt in der Tatsache begründet, daß nach Eliminierung aller Zeilen (Spalten) mit unvollständiger Information alle zur Behandlung vollständiger Datensätze geeigneten Verfahren problemlos angewandt werden können. Außerdem ist der Stichprobenumfang bezüglich jedes Elements zweiter Modalität gleich groß. Darum ist die Vergleichbarkeit univariater Statistiken gewährleistet. Nachteilig ist hingegen, daß selbst unter der Voraussetzung eines MCAR-Prozesses durch die Eliminierung Information verlorengeht, was die Varianz vergrößert. Weiterhin kann beim Fehlen sehr vieler Daten die verbleibende Information so dürftig sein, daß auf ihrer Grundlage gar keine Analyse mehr durchführbar ist. Dies ist beispielsweise bei für Recommender-Systemen typischen Datenmatrizen i.d.R. der Fall.

Um die bei den Eliminierungsverfahren beim Fehlen der MCAR-Eigenschaft auftretenden Verzerrungen zu reduzieren, sind eine Reihe von Verfahren entwickelt worden, die ebenfalls unbekannte Information eliminieren, aber dafür die vorhandene Information unterschiedlich gewichten (z.B. Little (1993)). Problematisch ist bei diesen Ansätzen, daß die Varianz in diesen Modellen schwer handhabbar ist (Little, Rubin (2002)). Sie sind daher nicht zu empfehlen, wenn die

Anzahl der vollständigen Zeilen (Spalten) in der Datenmatrix gering sind.

Die Verwendung aller gegebener Information bei gleichzeitiger Vernachlässigung fehlender Daten („available-case analysis“) hat den Vorteil, daß die gesamte verfügbare Information in die Analyse miteinbezogen wird. Das bringt den Nachteil mit sich, daß der Sichprobenumfang hinsichtlich der Elemente gleicher Modalität variiert. Letzteres führt dazu, daß zum einen die Vergleichbarkeit univariater Statistiken nicht gewährleistet ist und zum anderen die Kovarianzen und Korrelationen zwischen verschiedenen Elementen zweiter Modalität neu definiert werden müssen.

Hierzu existieren in der Literatur verschiedene Ansätze. Der erste Vorschlag geht auf Wilks (1932) zurück. Wilks schlug vor, die Mittelwerte aus den zum betrachteten Element zweiter Modalität vorhandenen Elementen erster Modalität zu berechnen:

$$\bar{y}_{.j} = \frac{1}{|I_j|} \sum_{i \in I_j} y_{ij}, \quad \text{wobei} \quad I_j = \{i' \in \{1, ..., I\} | v_{i'j} = 1\}.$$

Weiter gilt $I_{j_1 j_2} = \{i' \in \{1, ..., I\} | v_{i'j_1} = 1 \wedge v_{i'j_2} = 1\}$. Damit ergibt sich folgender Ausdruck für die approximierte Kovarianz nach Wilks (1932):

$$s_{j_1 j_2}^{Wilks2} = \frac{1}{(|I_{j_1 j_2}| - 1)} \sum_{i \in I_{j_1 j_2}} (y_{ij_1} - \bar{y}_{.j_1})(y_{ij_2} - \bar{y}_{.j_2})$$

und

$$s_j^{Wilks2} = \frac{1}{(|I_j| - 1)} \sum_{i \in I_j} (y_{ij} - \bar{y}_{.j})^2.$$

Daraus erhält man die angenäherte Korrelation

$$\tilde{r}_{j_1 j_2}^{Wilks} = \frac{s_{j_1 j_2}^{Wilks2}}{\sqrt{s_{j_1}^{Wilks2} s_{J_2}^{Wilks2}}}.$$

Bei dieser Näherung an die Korrelation wird alle zur Verfügung stehende Information benutzt, um die Mittelwerte bezüglich der Items j_1 und j_2 zu berechnen. Allerdings kann $\tilde{r}_{j_1 j_2}^{Wilks}$ außerhalb des Intervalles [-1,+1] liegen.

Daher wird von Matthai (1951) vorgeschlagen, auch die Mittelwerte und die Varianzterme $s_j^{Matthai2}$ nur aus den für beide Elemente der zweiten Modalität zur Verfügung stehenden Daten zu berechnen:

$$\bar{y}_{.j_1}^{j_1 j_2} = \frac{1}{|I_{j_1 j_2}|} \sum_{i \in I_{j_1 j_2}} y_{ij_1}.$$

Es gilt $\bar{y}_{.j_1}^{j_1 j_2} = \bar{y}_{.j_2}^{j_2 j_1}$. Das führt bei der Berechnung der Korrelation zwischen dem j_1-ten und dem j_2-ten Element zweiter Modalität zu

$$s_{j_1 j_2}^{Matthai2} = \frac{1}{(|I_{j_1 j_2}| - 1)} \sum_{i \in I_{j_1 j_2}} (y_{ij_1} - \bar{y}_{.j_1}^{j_1 j_2})(y_{ij_2} - \bar{y}_{.j_2}^{j_2 j_1})$$

und

$$s_{j_1}^{Matthai,j_2 2} = \frac{1}{(|I_{j_1 j_2}| - 1)} \sum_{i \in I_{j_1 j_2}} (y_{ij_1} - \bar{y}_{.j_1}^{j_1 j_2})^2.$$

Damit ergibt sich die Korrelation

$$\tilde{r}_{j_1 j_2}^{Matthai} = \frac{s_{j_1 j_2}^{Matthai2}}{\sqrt{s_{j_1}^{Matthai,j_2 2} s_{j_2}^{Matthai,j_1 2}}}.$$

Hier wird zwar weniger als die zur Verfügung stehende Information zur Berechnung der Mittelwerte herangezogen, dafür liegt $\tilde{r}_{j_1 j_2}^{Matthai}$ aber im Intervall [-1,+1].

Beispiel 3.5:

In der hypothetische Datenmatrix

$$(y_{ij}^h) = \begin{pmatrix} 5 & 3 & 3 & \mathrm{X} & 4 & 5 & \mathrm{X} & 4 \\ \mathrm{X} & \mathrm{X} & \mathrm{X} & 5 & \mathrm{X} & 3 & 4 & 4 \end{pmatrix}'$$

sind alle fehlende Daten durch den Eintrag „X“ gekennzeichnet. (Man beachte außerdem das Transpositionszeichen. Es gilt $Y^h \in \mathbb{R}^{8,2}$.) Auf Basis der Datenmatrix y^h ergibt sich $\tilde{r}_{12}^{Wilks} = -\sqrt{\frac{15}{8}} < -1$. Für die Korrelation nach Matthai erhält man $\tilde{r}_{12}^{Matthai} = -1$.

Eine Kombination beider Ansätze für die Kovarianzterme, $(s_{j_1j_2}^{Wilks2})$ und $(s_{j_1j_2}^{Matthai2})$, findet sich in der in der moderneren Literatur unter dem Namen „Pairwise Deletion“ wieder (Wothke (1993)). Nach der „Pairwise Deletion“-Methode berechnet man die Korrelation $\tilde{r}_{j_1j_2}^{PD} = \frac{s_{j_1j_2}^{Matthai2}}{\sqrt{s_{j_1}^{Wilks2} s_{j_2}^{Wilks2}}}$.
Nicht nur die Formel für die Korrelation nach Wilks $\tilde{r}_{j_1j_2}^{Wilks}$ kann wie anhand von Beispiel 3.5 demonstriert zu inkonsistenten Werten für die Korrelation führen. Das (nachfolgende) Beispiel 3.6 belegt, daß sich sowohl durch die Berechnung der Korrelation nach Matthai $\tilde{r}_{j_1j_2}^{Matthai}$ als auch durch die Vorschrift zur Berechnung der Korrelation nach dem „Pairwise Deletion“-Verfahren ebenfalls Werte jenseits des Intervalls $[-1, +1]$ ergeben können.

Beispiel 3.6:

Auf Basis der Datenmatrix

$$(y_{ij}^{h2}) = \begin{pmatrix} 2 & 1 & 3 & \mathrm{X} & \mathrm{X} & \mathrm{X} & 1 & 2 & 3 \\ 2 & 1 & 3 & 1 & 2 & 3 & \mathrm{X} & \mathrm{X} & \mathrm{X} \\ \mathrm{X} & \mathrm{X} & \mathrm{X} & 1 & 2 & 3 & 3 & 2 & 1 \end{pmatrix}' \in \mathbb{R}^{9,3}$$

erhält man:

$$\begin{aligned} \tilde{r}_{12}^{Wilks} = \tilde{r}_{12}^{PD} = +\frac{5}{4} \quad \tilde{r}_{12}^{Matthai} = +1 \\ \tilde{r}_{13}^{Wilks} = \tilde{r}_{13}^{PD} = -\frac{5}{4} \quad \tilde{r}_{13}^{Matthai} = -1 \\ \tilde{r}_{23}^{Wilks} = \tilde{r}_{23}^{PD} = +\frac{5}{4} \quad \tilde{r}_{23}^{Matthai} = +1. \end{aligned}$$

Man sieht, daß die nach der „Pairwise Deletion“-Methode berechneten Korrelationen $\tilde{r}_{j_1j_2}^{PD}$ (Beispiel 3.6) ebenso wie die auf Grundlage des Verfahrens von Wilks berechneten Korrelationen $\tilde{r}_{j_1j_2}^{Wilks}$ (Beispiel 3.5) außerhalb des Intervalls [-1,+1] liegen können. Zudem zeigt Beispiel 3.6, daß auch die Korrelationsberechnungsmethode nach Matthai zu inkonsistenten Ergebnissen führen kann, da Item 1 perfekt positiv mit Item 2 und perfekt negativ mit Item 3 korreliert ist aber trotzdem Item 2 perfekt positiv mit Item 3 korreliert ist.

Damit kann die Strategie der Verwendung der gesamten vorhandenen Information bei gleichzeitiger Vernachlässigung aller fehlenden Werte (die sogenannte „available case analysis“) zu inkonsistenten Ergebnissen führen.

Ein weiteres Argument gegen alle genannten Verfahren ist, daß die Kovarianzmatrizen $\tilde{r}_{j_1j_2}^{Wilks}$, $\tilde{r}_{j_1j_2}^{Matthai}$ und $\tilde{r}_{j_1j_2}^{PD}$ nicht notwendig positiv definit sind.

Beispiel 3.7:

Auch dies läßt sich anhand der Datenmatrix Y^{h2} aus Beispiel 3.6 zeigen. So ergibt sich für die Varianz-Kovarianz-Matrix nach Matthai, Wilks und der „Pairwise Deletion“-Methode auf Basis der Datenmatrix Y^{h2}:

$$(s_{j_1j_2}^2) = \begin{pmatrix} \kappa & 1 & \text{-1} \\ 1 & \kappa & 1 \\ \text{-1} & 1 & \kappa \end{pmatrix}.$$

Hier ist

$$\kappa = \begin{cases} \frac{4}{5} & , \quad \text{für } (s_{j_1j_2}^2) = (s_{j_1j_2}^{Wilks2}) \text{ und } (s_{j_1j_2}^2) = (s_{j_1j_2}^{PD2}) \\ 1 & , \quad \text{für } (s_{j_1j_2}^2) = (s_{j_1j_2}^{Matthai2}) \end{cases}$$

Die zugehörige quadratische Form $\vec{q}'(s_{j_1j_2})\vec{q}$ ist für $(s_{j_1j_2}^2) \in \{(s_{j_1j_2}^{Wilks2}), (s_{j_1j_2}^{PD2})\}$ nicht für jede Wahl für den Vektor $\vec{q} = (q_1, q_2, q_3)' \neq \vec{0}$ größer Null. Für $q_2 = 0$ und $q_1 = q_3 \geq 1$ ist die quadratische Form negativ. Daher wäre die zugehörige Kovarianzmatrix weder positiv definit noch positiv semidefinit. Auch im Fall $(s_{j_1j_2}^2) = (s_{j_1j_2}^{Matthai2})$ ist die zugehörige quadratische Form nicht für alle $\vec{q} \neq \vec{0}$ größer Null. Beispielsweise ist dies nicht erfüllt für die Wahl $q_2 = 0 \wedge q_1 = q_3$.

Allgemein ist eine Kovarianzmatrix $S = (S_{j_1j_2}^2)$ von (Y_{ij}) immer dann positiv definit, wenn die Datenmatrix Y_{ij} nicht singulär ist. Nur dann, wenn $Y = (Y_{ij})$ singulär ist, ist S wegen der Beziehung $var(\vec{q}'Y) = \vec{q}'S\vec{q}$ positiv semidefinit aber nicht positiv definit.

Unter der Voraussetzung, daß (Y_{ij}) in vollständigem Zustand nicht singulär ist, können somit alle Verfahren zur Berechnung der Kovarianz und Korrelation zu inkosistenten Ergebnissen führen.

Auch die praktische Schwierigkeiten bei der Analyse der auf Grundlage der beschriebenen Methoden errechneten Varianz-Kovarianzmatrizen $S = (s^2_{j_1 j_2})$ sollen nicht unerwähnt bleiben.

Da nur positiv oder negativ definite Matrizen invertierbar sind und der Weighted-Least-Squares-Schätzer (WLS-Schätzer) die Inversion der Varianz-Kovarianzmatrix erfordert, existiert nicht notwendig ein WLS-Schätzer.

Im Fall einer nur positiv semidefiniten Kovarianzmatrix kann y nicht als Design-Matrix zur Berechnung des OLS-Schätzers herangezogen werden, da $y'y$ nicht invertierbar ist, wenn y nicht vollen Rang hat.

Da bei nicht positiv semidefiniten Kovarianzmatrizen negative Eigenwerte autfreten können, kann es bei verschiedenen multivariaten Verfahren, wie zum Beispiel der Faktoren- oder Diskriminanzanalyse, zu Problemen kommen. Diese Probleme können umgangen werden, indem man auf Verfahren rekurriert, mittels derer man eine nicht semidefinite Matrix zu einer semidefiniten Matrix machen kann (vgl. Schnell (1986)). Die hierzu verwendeten Verfahren verändern die ursprünglich auf Grundlage der gegebenen Daten errechnete Kovarianzmatrix ohne irgendwelche Informationen oder Kenntnisse über die fehlenden Daten zu nutzen.

Wenn die Daten die MCAR-Eigenschaft besitzen und die Korrelationen gering sind, sind die Verfahren, die alle verfügbare Information zu Analysezwecken einbeziehen, besser geeignet als jene Verfahren, die nur den vollständigen Teil der Daten berücksichtigen. Dies ist das Ergebnis einer Simulationsstudie von Kim, Curry (1977). Andere Simulationsstudien belegen, daß bei großen Korrelationen die Verfahren, die die gesamte vorhandene Information verwenden, unterlegen sind (Haitovsky (1968), Azen, van Guilder (1981)).

Arbuckle (1996) konnte durch Simulationsstudien belegen, daß sowohl für Daten, die die MAR-Eigenschaft erfüllen, als auch Daten, welche die MCAR-Eigenschaft aufweisen, die „Pairwise Deletion“-Methode und ein Verfahren, das alle Zeilen, die fehlende Daten enthalten, gründsätzlich ausschließt, zu schlechteren Resultaten als ein Maximum-Likelihood-basiertes Verfahren führen.

3.2.2 Imputationsverfahren

Imputationsverfahren ergänzen die bereits vorhandenen Daten durch Schätzungen aller fehlenden Werte, so daß eine vollständige Datenmatrix gleicher Dimension resultiert. Diese Datenmatrix kann dann genau wie die durch die verschiedenen

Eliminierungsverfahren erzeugten reduzierten Datenmatrizen oder eine von vornherein vollständige Datenmatrix zum Ausgangspunkt aller quantitativen Verfahren werden, die das vollständige Vorliegen aller Einträge in der Datenmatrix erfordern. Speziell gegenüber den Eliminierungsverfahren haben die Imputationstechniken den Vorteil, daß es nicht zum Ausschluß eines unter Umständen nicht mehr tolerierbaren Teils der Daten aus der Analyse kommen kann (Schnell (1985)). Insbesondere bei den in jüngster Zeit vor dem Hintergrund möglicher Direkt-Marketing Maßnahmen und/oder zum Zweck der Erhöhung der Kundenzufriedenheit und Kundenbindung interessierenden Datenmatrizen aus dem Online- oder Kundenkarten-Bereich (als Beispiel seien hier nur die im Zusammenhang mit dem Recommender-System eines Online-Shops generierten Bewertungsdaten angeführt) ist dies ein erheblicher Vorteil. So würde eine konsequente Anwendung eines jeden Eliminierungsverfahrens auf einen solchen Datensatz (bzw. auf einen vergleichbaren Datensatz wie z.B. die MovieLens-Daten) dazu führen, daß keine Daten mehr zur Analyse zur Verfügung stehen, da weder ein Item (bzw. Film) existiert, der von allen Nutzern bewertet wurde, noch ein Nutzer, der alle Items (bzw. Filme) bewertet hat.

Alle Werte, die durch die Imputationsverfahren an die Stelle der fehlenden Daten gesetzt werden, sind entweder die Lageparameter einer bestimmten Wahrscheinlichkeitsverteilung oder Ziehungen aus einer solchen oder aber Daten, die man für hinreichend ähnlich wie die zu ersetzenden Daten verteilt hält.

Innerhalb der Imputationsverfahren kann zwischen algorithmisch und statistisch orientierten Ansätzen unterschieden werden. Während den statistisch orientierten Ansätzen immer ein formales statistisches Modell zugrundeliegt, weshalb die gemachten Verteilungsannahmen explizit sind, tritt bei den algorithmisch orientierten Ansätzen der Algorithmus selber in den Vordergrund. Die dem Algorithmus zugrundeliegenden Verteilungsannahmen basieren nicht auf einem vollständigen formalen statistischen Modell für die Daten. Dennoch liegen den Algorithmen immer Annahmen hinsichtlich der Verteilung der fehlenden Daten zugrunde, die sich jedoch durchaus darauf beschränken können, daß ein bestimmter Teil der fehlenden Daten hinreichend ähnlich wie ein bestimmter Teil der vorhandenen Daten derselben oder einer anderen Datenmatrix verteilt ist.

3.2.2.1 Statistisch orientierte Imputationsverfahren

Zu den statistisch orientierten Imputationsverfahren gehören die Imputation durch den Lageparameter (Wilks (1932)), die Imputation durch Regression (Federspiel et. al. (1959)), Buck (1960), Jackson (1968), Chan, Dunn (1972), Gleason, Staelin (1975), Anderson et. al. (1983)) sowie die Stochastische Imputation durch Regression (Herzog, Rubin (1983)).

Bei der Imputation durch den Lageparameter wird der Schätzer für den Lageparameter an Stelle der fehlenden Daten verwendet. Falls die vorliegenden Daten nominal skaliert sind, benutzt man den Modus. Liegt hingegen ordinales Skalenniveau vor, so ist der Median zu nehmen, während bei kardinal skalierten Werten das arithmetische Mittel zu verwenden ist (Wilks (1932)).
Als Schätzer für Y_{ij} kann dann entweder der Spaltenmittelwert (aller in einer Spalte j vorhandenen Werte) $\bar{Y}_{.j}$ oder der Zeilenmittelwert (aller in einer Zeile i gegebenen Einträge) $\bar{Y}_{i.}$ verwendet werden, wobei sich im ersten Fall

$$\hat{Y}_{ij} = \bar{y}_{.j} = \frac{1}{|I_j|} \sum_{i' \in I_j} y_{i'j} \text{ mit } I_j = \{i' \in \{1, ..., I\} | v_{i'j} = 1\}$$

und im zweiten Fall

$$\hat{Y}_{ij} = \bar{y}_{i.} = \frac{1}{|J_i|} \sum_{j' \in J_i} y_{ij'} \text{ mit } J_i = \{j' \in \{1, ..., J\} | v_{ij'} = 1\}$$

ergibt. Dieses Imputationsverfahren setzt die MCAR-Annahme voraus (Bankhofer (1995)). In der Praxis werden häufig Mittelwert-Imputationen verwendet (z.B. Sarwar et. al. (2000), Mild, Natter (2002)). Typischerweise werden in einer Spalte (Zeile) alle fehlenden Werte durch den Mittelwert aller in dieser Spalte (Zeile) vorhandenen Daten ersetzt. Seltener werden alle fehlenden Werte durch den Mittelwert aller vorhandenen Daten imputiert. Dies gibt Anlaß dazu, sich näher mit den Eigenschaften dieser Mittelwerte zu befassen.

In der Literatur werden zu diesem Zweck Klassen $k' \in \{1, ..., K'\}$ für die Elemente erster Modalität, sogenannte „adjustment cells“ betrachtet (Oh, Scheuren (1983), Little (1986)). Diese Klassen sind so zu wählen, daß die Elemente innerhalb einer Klasse möglichst homogen sind, während die Elemente verschiedener Zellen heterogen sein sollen. Jedes Element erster Modalität wird hierbei genau

einer Klasse zugeordnet. Die Menge aller Elemente der k'-ten Zelle sei $C'(k')$. Es gilt für alle $i \in \{1, ..., I\}$ und alle $k' \in \{1, ..., K'\}$:

$$\tilde{p}_{ik'} = \begin{cases} 1, & \text{falls } i \in C'(k') \\ 0, & \text{sonst} \end{cases} .$$

Der Mittelwert bezüglich der j-ten Spalte innerhalb der k'-ten Zelle ergibt sich durch

$$\bar{y}_{.j}^{C'(k')} = \frac{\sum\limits_{i \in I_j} \tilde{p}_{ik'} y_{ij}}{\sum\limits_{i \in I_j} \tilde{p}_{ik'}}.$$

Sofern $Y \in \mathbb{R}^{I,J}$ die Bewertungen von I Nutzern in Bezug auf J Items enthält, ist $\bar{y}_{.j}^{C'(k')}$ die durchschnittliche Bewertung der Items j, die von Nutzern aus der Klasse k' abgegeben wurden. Ist y dagegen das Ergebnis einer Befragung von I Personen im Hinblick auf J kardinale Merkmale (wie z.B. Einkommen), muß $\bar{y}_{.j}^{C'(k')}$ als der Durchschnittswert des j-ten Merkmals bezüglich aller Befragten aus Klasse k' interpretiert werden.
Mit dem Zellengewichtungsfaktor

$$N_j^{C'(k')} = \frac{\sum\limits_{i=1}^{I} \tilde{p}_{ik'} v_{ij}}{\sum\limits_{i=1}^{I} v_{ij}} = \frac{\sum\limits_{i \in I_j} \tilde{p}_{ik'}}{\sum\limits_{i=1}^{I} v_{ij}}$$

ergibt sich die folgende Darstellung für $\bar{Y}_{.j}$:

$$\bar{y}_{.j} = \sum_{k'=1}^{K'} \frac{\sum\limits_{i \in I_j} \tilde{p}_{ik'} y_{ij}}{\sum\limits_{i=1}^{I} V_{ij}} = \sum_{k'=1}^{K'} N_j^{C'(k')} \bar{y}_{.j}^{C'(k')}.$$

Die Zellenmittelwerte über die innerhalb der Zelle bezüglich j vorhandenen Bewertungen bzw. Merkmale werden mit dem Anteil an in der jeweiligen Zelle vorhandenen Daten über j an der Menge der insgesamt hinsichtlich j vorhandenen

Daten gewichtet. Zellen, in denen mehr Daten fehlen werden somit stark untergewichtet. Sind die Daten nicht MCAR, kann es sein, daß sich die Elemente einer Zelle sowohl im Hinblick auf die Antwortrate $N_j^{C'(k')}$ als auch in Bezug auf die Bewertungen $\bar{y}_{.j}^{C'(k')}$ wesentlich von den anderen Zellen unterscheidet. Falls mindestens eine Zelle k' in $\bar{y}_{.j}$ unter- bzw. übergewichtet wird und sich das zugehörige $\bar{y}_{.j}^{C'(k')}$ deutlich von den entsprechenden Werten anderer Zellen unterscheidet, ist $\bar{y}_{.j}$ dann i.d.R. ein verzerrter Schätzer für den Populationsmittelwert, weil bei seiner Berechnung fehlende Daten ignoriert werden. In diesen Fällen ist der angepaßte Mittelwert („weighting class estimator“, Oh, Scheuren (1983))

$$\bar{y}_{.j}^{A} = \sum_{k'=1}^{K'} \frac{|C'(k')|}{I} \bar{y}_{.j}^{C'(k')}$$

der geeignetere Schätzer für den Populationsmittelwert. Die Verwendung des angepaßten Mittelwerts ist äquivalent zur Gewichtung aller bezüglich j vorhandenen Daten aus Zelle k' mit $|C'(k')|/(N_j^{C'(k')}|I|)$ oder zur Imputation von $\bar{Y}_{.j}^{C'(k')}$ für alle im Hinblick auf j fehlenden Daten aus der k'-ten Zelle ($k' = 1, ..., K'$). Little (1986) konnte durch eine Simulationsstudie nachweisen, daß die Verwendung von $\bar{Y}_{.j}$ zu starken Verzerrungen führen kann, die durch die Verwendung des angepaßten Maßes $\bar{Y}_{.j}^{A}$ erheblich reduziert werden können.

Beispiel 3.8:

Von Little (1986) stammt ein interessantes Beispiel, indem es um Einkommen ($j = j_E$) geht. Hier ist Y_{ij_E} das Jahreseinkommen der Person i in US Dollar. Abhängig von ihrem Wohnsitz konnten die Befragten in 3 verschiedene Zellen $k' \in \{1, ..., K'\}$ eingeteilt werden:

	$k'=1$	$k'=2$	$k'=3$
$\lvert C'(k')\rvert$	100	100	100
$\sum_{i=1}^{I}\tilde{p}_{ik'}v_{ij_E}$	80	70	50
$\bar{y}_{.j_E}^{C'(k')}$	9800	11600	13600
$\bar{y}_{.j_E}^{C'(k')}\sum_{i=1}^{I}\tilde{p}_{ik'}v_{ij_E}$	780	815	680

Tabelle 3.1: Einkommensmittelwerte in verschiedenen Regionen (Beispiel 3.8)

In diesem Beispiel fehlen die meisten Bewertungen in der Zelle $k'=3$. Dies ist auch die Zelle mit dem höchsten Durchschnittseinkommen. Damit unterschätzt

$$\bar{y}_{.j_E}=\left(\frac{80}{200}\right)9800\,\$+\left(\frac{70}{200}\right)11600\,\$+\left(\frac{50}{200}\right)13600\,\$=11400\,\$$$

sehr wahrscheinlich den Populationsmittelwert. Der angepaßte Spaltenmittelwert

$$\bar{y}_{.j_E}^{A}=\left(\frac{100}{300}\right)9800\,\$+\left(\frac{100}{300}\right)11600\,\$+\left(\frac{100}{300}\right)13600\,\$=11700\,\$$$

scheint hier angebrachter zu sein. Dieses Beispiel illustriert, daß ein Ignorieren fehlender Werte keineswegs unbedenklich ist und z.B. Verzerrungen der Spaltenmittelwerte bewirken kann. Gleichzeitig sieht man, daß unter Voraussetzung einer geschickten Wahl für die Imputationsklassen („adjustment cells“) die zellenweise Ersetzung der fehlenden Werte aller Elemente erster Modalität einer Zelle $C'(k')$ durch die zugehörigen Mittelwerte $\bar{y}_{.j_E}^{C'(k')}$ der jeweiligen Zelle $C'(k')$ zu plausibeleren Ergebnissen führen kann.

Die in der Praxis häufig verwendete Imputation des Spaltenmittelwerts (bzw. Zeilenmittelwerts) ignoriert die fehlenden Daten bei der Berechnung desselben. Ist die MCAR-Eigenschaft nicht erfüllt, so kann dies zu verzerrten Mittelwerten führen. Vor diesem Hintergrund ist es nicht verwunderlich, daß die Verwendung dieser Mittelwerte anstelle der fehlenden Werte bereits bei einem Fehlendanteil von 41 % zu starken Verzerrungen führen kann (Myrtveit et. al. (2001)). Anhand ihres empirischen Vergleichs verschiedener Strategien für den Umgang mit fehlenden Daten kommen Myrtveit et. al. (2001) zu dem Ergebnis, daß die Imputation

von Mittelwerten nur dann erfolgen sollte, wenn die MCAR-Eigenschaft gegeben ist. Problematisch ist außerdem, daß die systematische Ersetzung fehlender Werte erster Modalität beim j-ten Element zweiter Modalität durch das arithmetische Mittel der vorhandenen Werte erster Modalität zum j-ten Element zweiter Modalität zu einer systematischen Unterschätzung der Varianz führt. Sofern die MCAR-Annahme erfüllt ist, ist die Stichprobenvarianz

$$s_j^{Wilks2} = \frac{1}{|I_j| - 1} \sum_{i \in I_j} (y_{ij} - \bar{y}_{.j})^2$$

erwartungstreuer Schätzer für die Varianz s_j^2. Sofern alle fehlenden Werte für Y_{ij} durch $\bar{Y}_{.j}$ ersetzt werden, gilt

$$\sum_{i=1}^{I} (y_{ij} - \bar{y}_{.j})^2 = \sum_{i \in I_j} (y_{ij} - \bar{y}_{.j})^2,$$

weshalb die Varianz s_j^2 durch die Stichprobenvarianz auf Basis der (durch Imputation ihrer Spaltenmittelwerte) vervollständigten Datenmatrix

$$s_j^{Imp,S2} = \frac{1}{I - 1} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{.j})^2$$

um den Faktor

$$\frac{|I_j| - 1}{I - 1}$$

unterschätzt wird. Durch die Imputation des arithmetischen Mittels wird die empirische Verteilung selbst unter der MCAR-Annahme verzerrt.

Um dieser systematischen Unterschätzung entgegenzuwirken, bietet es sich an, alle Varianzterme $s_j^{Imp,S2}$ mit dem Faktor $(I-1)/(|I_j|-1)$ und alle Kovarianzterme $s_{j_1 j_2}^{Imp,S2}$ entsprechend mit dem Faktor $(I-1)/(|I_{j_1 j_2}|-1)$ zu multiplizieren. Auf diese Weise erhält man korrigierte Werte $\tilde{s}_j^2$ und $\tilde{s}_{j_1 j_2}^2$ für die Stichprobenvarianzen und -kovarianzen.

Es fällt sofort auf, daß auf diese Weise berechneten Stichprobenvarianzen $\tilde{s}_j^2$ genau den Varianztermen nach Wilks s_j^{Wilks2} entsprechen:

$$\tilde{s}_j^2 = \frac{I-1}{|I_j|-1}\frac{1}{I-1}\sum_{i=1}^{I}(y_{ij}-\bar{y}_j^j)^2 = \frac{1}{|I_j|-1}\sum_{i\in I_j}(y_{ij}-\bar{y}_j^j)^2 = s_j^{Wilks2}.$$

Analog gilt für die Kovarianzen

$$\begin{aligned}\tilde{s}_{j_1j_2}^2 &= \frac{I-1}{|I_{j_1j_2}|-1}\frac{1}{I-1}\sum_{i=1}^{I}(y_{ij_1}-\bar{y}_{.j_1})(y_{ij_2}-\bar{y}_{.j_2})\\ &= \frac{1}{|I_{j_1j_2}|-1}\sum_{i\in I_{j_1j_2}}(y_{ij_1}-\bar{y}_{.j_1})(y_{ij_2}-\bar{y}_{.j_2}) = s_{j_1j_2}^{Wilks2}.\end{aligned}$$

Die auf diese Weise erhaltene Näherung für die Matrix $(\tilde{s}_{j_1j_2}^2)$ ist daher ebenso wie $(s_{j_1j_2}^{Wilks2})$ weder positiv definit noch positiv semidefinit. Deshalb kann die Imputation durch das arithmetische Mittel vor dem Hintergrund der resultierenden Stichprobenkovarianzmatrizen genauso zu Inkonsistenzen führen, wie die Berechnung der Kovarianzmatrix nach Wilks.

Außerdem kann die Imputation von Zeilen- bzw. Spaltenmittelwerten zu verzerrten (auf Basis der vervollständigten Datenmatrix berechneten) Stichproben-Äquivalenten der Zeilen- bzw. Spaltenerwartungswerte führen, sofern die MCAR-Annahme nicht erfüllt ist (vgl. Beispiel 3.8). Daher ist die Imputation von Zeilen- oder Spaltenmittelwerten als problematisch einzustufen.

Ein weiteres statistisch orientiertes Imputationsverfahren ist die sogenannte Imputation durch den Verhältnisschätzer. Auch diese Methode erfordert die MCAR-Annahme (Little, Rubin (2002)). Bei diesem Verfahren sucht man zu jedem Element zweiter Modalität j_1, bei dem fehlende Werte hinsichtlich der zugehörigen Elemente erster Modalität zu finden sind, nach einem weiteren Element zweiter Modalität j_2, das mit dem Element j_1 stark korreliert sein soll. Außerdem soll $I_{j_1} \subset I_{j_2}$ gelten. Als Formel für den Schätzer für Y_{ij_1} ergibt sich (Ford (1976), Platek, Gray (1983)):

$$\hat{Y}_{ij_1} = \frac{\sum_{i'\in I_{j_1}} y_{i'j_1}}{\sum_{i'\in I_{j_1}} y_{i'j_2}}\frac{1}{|I_{j_2}|}\sum_{i'\in I_{j_2}} y_{i'j_2}.$$

Der erste Faktor dieser Formel ist das Verhältnis des hinsichtlich der Menge I_{j_1} ermittelten Mittelwertes des j_1-ten Merkmals zum Mittelwert des j_2-ten Merk-

mals in Bezug auf I_{j_1}. Der restliche Faktor ist nichts anderes als das arithmetische Mittel aller vorhandenen Werte erster Modalität beim j_2-ten Element. Hierdurch wird versucht, einen besseren Schätzer für das arithmetische Mittel zu finden, was sinnvoll ist, wenn $|I_{j_1}| << |I_{j_2}|$.

Kritisch ist anzumerken, daß nicht zu jedem Element zweiter Modalität j_1 ein weiteres Element zweiter Modalität j_2 existieren muß, so daß die Bedingung $I_{j_1} \subset I_{j_2}$ erfüllt ist. Deshalb ist diese Strategie als Imputationsstrategie für eine gesamte Datenmatrix nur bei niedrigen Fehlendanteilen geeignet.

Bei online-generierten hochdimensionalen Datenmatrizen mit einer extrem großen Anzahl von fehlenden Daten wie z.B. dem MovieLens-Datensatz dürfte es sogar eher die Ausnahme sein, daß zu einem Element j_1 ein weiteres Element j_2 existiert, so daß $I_{j_1} \subset I_{j_2}$ gilt. Somit kann die Imputation des Verhältnisschätzers für diese Datenmatrizen nicht verwendet werden.

Da letztlich nur das arithmetische Mittel geschätzt wird, kann es hier genau wie bei der Imputation durch das arithmetische Mittel zur Unterschätzung der Varianzen und Kovarianzen und den daraus resultierenden Problemen kommen. Bei der Imputation durch Regression wird für jedes Element zweiter Modalität j_1, bei dem die zugehörige Spalte in der Datenmatrix fehlende Werte aufweist, ein Regressionsmodell verwendet, in dem das j_1-te Element als endogene Variable auf andere Elemente zweiter Modalität $j_2 \in \tilde{J}[j_1]$ regressiert wird. Hier ist $\tilde{J}[j_1] \subseteq \{1, ..., J\} \setminus \{j_1\}$ die Menge aller Spalten-Indizes, für die gilt, daß alle mit diesen Indizes verknüpften Spalten der Datenmatrix Y als Realisationen einer Zufallsvariable aufgefaßt werden können, die sich als exogene Variable in einem Regressionsmodell eignet, welches die mit der j_1-ten Spalte assoziierte Zufallsvariable als endogene Variable enthält. Der zugehörige Regressionsansatz ist:

$$Y_{ij_1} = \beta_0^{j_1} + \sum_{j_2 \in \tilde{J}[j_1]} \beta_{j_2}^{j_1} y_{ij_2} + \tilde{\epsilon}_i, E(\tilde{\epsilon}_i) = 0 \quad i = 1, ..., I.$$

Bei einer Variante dieser Methode werden nur die Elemente zweiter Modalität, bei denen alle Elemente erster Modalität vorhanden sind, als Regressoren verwendet (Buck (1960)). Dann wird der OLS-Schätzer für $\vec{\beta}$ berechnet. Alle fehlenden Einträge in der j_1-ten Spalte der Datenmatrix werden dann durch

$$\hat{Y}_{ij_1} = \hat{\beta}_0^{j_1} + \sum_{j_2 \in \tilde{J}[j_1]} \hat{\beta}_{j_2}^{j_1} y_{ij_2}$$

ersetzt. Diese Variante ist auf Datenmatrizen mit extrem vielen fehlenden Werten nicht anwendbar.

Federspiel et. al. (1959) empfehlen eine iterative Kombination aus der Imputation des arithmetischen Mittels und der Imputation durch Regression. Hierbei werden alle in der Datenmatrix (y_{ij}) fehlenden Werte zunächst durch das arithmetische Mittel $y_{ij}^0 = \bar{y}_{.j}$ für $n = 0$ ersetzt. Dann wird n um 1 inkrementiert (Schritt 1) und es werden für alle Spalten j in der Datenmatrix, in der Werte ergänzt werden mußten, Regressionen wie beim Verfahren nach Buck durchgeführt. Hierbei werden OLS-Schätzer

$$\vec{\hat{\beta}}^{j,n} = (\hat{\beta}_0^{j,n}, ..., \hat{\beta}_{|\tilde{J}[j]|})' \text{ für } \vec{\beta}^j = (\beta_0^j, ..., \beta_{|\tilde{J}[j]|}^j)' \ \forall j \in \{j' \in \{1, ..., J\} | V_{.j'}^{ind} \neq 1\}$$

auf Basis von y^{n-1} bestimmt (Schritt 2).
Zu diesem Zweck benutzt man

$$Y_{ij} = \beta_0^{j,n} + \sum_{j' \in \tilde{J}[j]} \beta_{j'}^{j,n} y_{ij'}^{n-1} + \epsilon_i, i = 1, ..., I, j \in J_i. \tag{3.6}$$

Danach werden für alle fehlenden Werte neue Schätzer

$$\hat{Y}_{ij}^n = \hat{\beta}_0^{j,n} + \sum_{j' \in \tilde{J}[j]} \hat{\beta}_{j'}^{j,n} y_{ij'}^{n-1}, i = 1, ..., I, j \in \{1, ..., J\} \setminus J_i$$

berechnet. Auf dieser Grundlage bestimmte man die neue vervollständigte Datenmatrix

$$y_{ij}^n = \begin{cases} \hat{Y}_{ij}^n, & \text{falls } v_{ij} = 0 \text{ ist} \\ y_{ij}, & \text{sonst} \end{cases}$$

(Schritt 3). Schritt 1 bis 3 werden solange iteriert, bis die Differenz der Imputationswerte $|\hat{Y}_{ij}^n - \hat{Y}_{ij}^{n-1}|$ für alle $i = 1, ..., I, j \in J_i$ eine vorgegebene untere Schranke unterschreitet (Jackson (1968), Anderson et. al. (1983)). Dies kann in der Praxis schon nach wenigen Iterationen der Fall sein (Jackson (1968)).

Ein gravierender Nachteil dieser Imputationsstrategie ist, daß für jedes Element zweiter Modalität, das fehlende Werte aufweist, ein geeignetes Regressionsmodell gefunden werden muß.

Insbesondere bei hochdimensionalen Datenmatrizen ($J >> 10^3$), bei denen hinsichtlich jedes Elements zweiter Modalität Daten fehlen, ist die Gesamtzahl der möglichen Regressionsmodelle

$$J \sum_{j=1}^{J-1} \left(\begin{array}{c} \text{J-1} \\ \text{j} \end{array} \right) = J(2^{J-1} - 1).$$

Daher ist es wenig zweckmäßig, das Verfahren nach Federspiel (1959) oder weitere Varianten des Regressionsansatzes (Chan, Dunn (1972), Gleason, Staelin (1975)) auf hochdimensionalen Datenmatrizen anzuwenden.

Im Rahmen der Stochastischen Imputation durch Regression wird zum (beispielsweise durch lineare Regression gewonnenen) Schätzer $\hat{Y}_{ij}$ eine Zufallszahl ϵ_{ij} addiert, die genauso wie die empirischen Residuen verteilt sein soll (Herzog, Rubin (1983)). Da jedoch die Anwendung von Regressionsverfahren auf hochdimensionale Datenmatrizen mit extrem vielen fehlenden Werten unzweckmäßig ist, braucht dieser Ansatz hier nicht weiterverfolgt werden.

3.2.2.2 Algorithmisch orientierte Imputationsverfahren

Ein Beispiel für ein Algorithmisch orientiertes Imputationsverfahren stellt die sogenannte Hot-Deck Imputation dar. Bei dieser Klasse von Verfahren wird versucht, die in einer Zeile fehlenden Werte durch Werte zu ersetzen, die in einer oder mehreren anderen Zeilen der Datenmatrix gegeben sind, wobei die Zeilen, aus denen die fehlenden Einträge entnommen werden, der Zeile, in der die fehlenden Einträge ergänzt werden, möglichst ähnlich sein müssen (David et. al. (1986)). Die Zeilen, aus denen die Imputationswerte stammen, heißen Donoren.

Bei der „Hot-Deck Imputation within Adjustment Cells“ werden alle Zeilen bestimmten Gruppen, eben den sogenannten „Adjustment Cells“ zugeordnet. Diese „Adjustment Cells“ werden so gebildet, daß die Elemente erster Modalität innerhalb derselben „Adjustment Cell“ hinsichtlich der bezüglich aller Elemente der jeweiligen „Adjustment Cell“ oder „Adjustment Class“ vorhandenen Daten möglichst homogen sind. Bei diesem Verfahren werden Schätzwerte für die fehlenden Werte in einer Zeile aus einer anderen Zeile entnommen, die aus derselben „Adjustment Class“ stammt (Hanson (1978)). Im Kontext der „Hot-Deck Imputation within Adjustment Cells“ werden zwei Zeilen als ähnlich zueinander bewertet, wenn sie derselben „Adjustment Cell“ angehören.

Allgemein existieren unterschiedliche Metriken zur Quantifizierung der Unähnlichkeit zweier Zeilen von Y. Beispielsweise werden die Maximale Abweichung

$$dist_1(i,i') = \max_{j \in J_{ii'}} |y_{ij} - y_{i'j}|, \quad J_{ii'} = J_i \cap J_{i'},$$

und die L_r-Norm

$$dist_2(i,i') = \left(\sum_{j \in J_{ii'}} (y_{ij} - y_{i'j})^r \right)^{\frac{1}{r}}$$

als Maße für die Verschiedenheit zweier Zeilen i und i' verwendet. Das sogenannte „Nearest Neighbour Hot-Deck“-Verfahren geht bei der Bestimmung der i-ten Zeile aus der gesamten unvollständigen Datenmatrix (y_{ij}) in drei Stufen vor. Zuerst werden allen Zeilen gesucht, die zumindest alle in der i-ten Zeile vorhandenen Einträge ebenfalls enthalten. Die Menge dieser Zeilen wird als

$$MZ(Y,i) = \{i_1 \in \{1,...,I\} \setminus \{i\} | v_{i_1 j} = 1 \; \forall \; j \in J_i\}$$

bezeichnet. Danach werden aus der Menge dieser Zeilen $MZ(Y,i)$ alle Zeilen ausgewählt, die zusätzlich noch alle zu imputierenden Werte enthalten. Aus dieser Menge $\widetilde{MZ}(Y,i) = \{i_1 \in MZ(Y,i) | v_{i_1 j} = 1 \; \forall \; j \in \{1,...,J\} \setminus J_i\}$ wird die Zeile i^* bestimmt, für die $dist_\hbar(i,i^*) < dist_S, \hbar = 1,2$, für einen fest gewählten Wert von $dist_S$ gilt oder welche die gewählte Distanz minimiert:

$$dist_\hbar(i,i^*) = \min_{i' \in \widetilde{MZ}(Y,i)} dist_\hbar(i,i'), \quad \hbar = 1,2.$$

Für schwach besetzte Datenmatrizen y wie die MovieLens-Daten ist dieses Verfahren völlig ungeeignet, weil diese Datenmatrizen nicht für jede i-te Zeile eine weitere Zeile i' enthalten, die neben allen in der i-ten Zeile von Y enthaltenen Elementen noch weitere Einträge aufweist.

Bei den sequentiellen Hot-Deck Techniken wird in jedem Schritt nur ein Teil der Elemente zweiter Modalität betrachtet (Colledge et. al.). Alle hinsichtlich dieser Elemente zweiter Modalität vollständigen Elemente erster Modalität werden als Donoren für die übrigen Elemente hinsichtlich der in dem jeweiligen Schritt bearbeiteten Elemente zweiter Modalität verwendet. Dieses Verfahren basiert auf

der Annahme, daß die Gruppen aus Elementen zweiter Modalität voneinander unabhängig sind.

In Bezug auf alle genannten Hot-Deck Verfahren führt die Analyse der durch sie vervollständigten Datenmatrix zu einer systematischen Unterschätzung der Standardabweichungen, da der Imputationsunsicherheit selbst nicht Rechnung getragen wird (Allison (2001)).

Diesen Problemen versucht man durch die Multiple Imputation zu begegnen. Bei der Multiplen Imputation wird zu jedem in der Datenmatrix (y_{ij}) fehlenden Eintrag ein ℓ_{imp}-dimensionaler Vektor aus Imputationswerten gebildet (Rubin (1978), Rubin (1987), Rubin (1996)). Auf diese Weise erhält man ℓ_{imp} verschiedene durch Imputation vervollständigte Datensätze. Die ℓ_{imp} verschiedenen Datensätze können aus ℓ_{imp} unterschiedlichen Ziehungen von Zufallszahlen auf der Basis eines einzigen Modells stammen oder aber auf der Grundlage mehrerer verschiedener Modelle generiert worden sein. Oft werden die Stochastische Imputation durch Regression oder verschiedene Hot-Deck Prozeduren verwendet. Im Rahmen der Multiplen Imputation legt man sich also zuerst auf ein oder mehrere Verfahren zur Bestimmung einzelner Imputationswerte fest. Danach werden auf jeden einzelnen der ℓ_{imp} vervollständigten Datensätze dieselbe Analyse-Methode angewandt, die man auch benutzt hätte, wenn die Daten von Anfang an vollständig gewesen wären. Mit Hilfe dieser Analyse-Methode berechnet man für jede der ℓ_{imp} vervollständigten Datenmatrizen den zugehörigen Schätzer für die interessierende Größe $\mathcal{G}^*$, $\hat{\mathcal{G}}^*_\ell$, und die Varianz desselben $var(\hat{\mathcal{G}}^*_\ell)$ für alle $\ell = 1, ..., \ell_{imp}$. Daraus ergibt sich der kombinierte Schätzer

$$\hat{\mathcal{G}}^* = \frac{1}{\ell_{imp}} \sum_{\ell=1}^{\ell_{imp}} \hat{\mathcal{G}}^*_\ell.$$

Wenn es sich bei den einzelnen Imputationswerten um auf der Basis eines stochastischen Modells gebildete Zufallszahlen handelt und das zugrundegelegte stochastische Modell gut ist, sollten die Unterschiede zwischen den ℓ_{imp} vervollständigten Datensätzen in ihren zugehörigen Schätzern zum Ausdruck kommen. Werden Imputationswerte auf der Grundlage mehrerer verschiedener stochastischer Modelle gebildet, so werden auch die Unterschiede hinsichtlich der gewählten Modelle durch die Schätzer $\hat{\mathcal{G}}^*_\ell, \ell = 1, ..., \ell_{imp}$, reflektiert.

Sofern die Imputationswerte selber das Ziel sind und keine weitere Analyse

der vollständigen Datenmatrix vorgesehen ist, ist das Verfahren der Multiplen Imputation nutzlos.

3.2.3 Parameterschätzverfahren

Als Parameterschätzverfahren werden all jene Methoden bezeichnet, die darauf abzielen, die gewünschten Parameter auf Basis der unvollständigen Datenmatrix zu schätzen (Schwab (1991)).

Die Parameterschätzverfahen lassen sich unterteilen in Maximum-Likelihood basierte und Bayes'sche Parameterschätzverfahren. Ein prominentes Beispiel für ein Maximum-Likelihood basiertes Parameterschätzverfahren ist der Expectation Maximization (EM) Algorithmus. Der bereits genannte empirische Vergleich verschiedener Strategien zum Umgang mit fehlenden Daten von Myrtveit et. al. (2001) legt nahe, solche modellbasierten Maximum-Likelihhood Ansätze anstelle des Ignorierens der fehlenden Werte oder der Mittelwertimputation einzusetzen.

3.2.3.1 Maximum-Likelihood basierte Parameterschätzverfahren

Bei den ersten Maximum-Likelihood basierten Verfahren von Wilks (1932) und Matthai (1951) setzte eine Schätzung der Mittelwerte, Varianzen und Kovarianzen jedoch voraus, daß entweder die Mittelwerte oder die Varianzen und Kovarianzen gegeben waren. Die Ansätze von Lord (1955) und Edgett (1956) erfordern, daß in der aus den Daten resultierenden Indikatormatrix spezielle Muster erkennbar sind. Falls kein besonderes („nested“) Muster in den Daten vorliegt, ist die Maximierung der Log-Likelihood-Funktion bis heute nur mit Rekurs auf iterative Verfahren durchführbar.

Y setzt sich aus einem Teil von Daten zusammen, der gegeben ist (Y_{obs}), und einem anderen Teil (Y_{mis}), welcher fehlt: $Y = (Y_{ij}) = (Y_{obs}, Y_{mis})$. Unter der Annahme, daß Y die MAR-Eigenschaft aufweist, läßt sich die Likelihood-Funktion $L(\theta|Y_{obs})$ aus der gemeinsamen Wahrscheinlichkeitsdichte $f(Y_{obs}, Y_{mis}|\theta)$ von Y_{obs} und Y_{mis} gegeben θ berechnen:

$$L(\theta|Y_{obs}) = f(Y_{obs}|\theta) = \int f(Y_{obs}, Y_{mis}|\theta) dY_{mis}.$$

Hierbei wird der Parameter θ im allgemeinen ein Vektor sein. θ bezieht sich auf

beide Teile der Datenmatrix Y.

Ein einfaches iteratives Verfahren zur Maximierung von $\log L(\theta|Y_{obs})$ ist der Newton-Raphson Algorithmus, der durch die Iterationsgleichung

$$\theta^{n+1} = \theta^n + \mathcal{I}_F^{-1}(\theta^n|y_{obs})\frac{\partial \log L(\theta^n|y_{obs})}{\partial \theta}$$

mit

$$\mathcal{I}_F(\theta|y_{obs}) = -\frac{\partial^2 \log L(\theta|y_{obs})}{\partial \theta \partial \theta'}$$

definiert ist.

Alternativ hierzu kann auch der Expectation Maximization (EM) Algorithmus oder eine seiner Abwandlungen verwendet werden.

Der EM Algorithmus iteriert zwei Stufen, den sogenannten Expectation-Schritt (E-Schritt) und dem Maximization-Schritt (M-Schritt).

Im Rahmen des E-Schritts wird der Erwartungswert der Log-Likelihood-Funktion $\log L(\theta|Y)$ für Y_{obs} und den Wert von θ aus dem vorherigen Iterationsschritt (für n=1: den Startwert θ^1) berechnet:

$$Q(\theta|\theta^n) = \int \log L(\theta|y_{obs}, \tilde{y}_{mis}) f(\tilde{y}_{mis}|y_{obs}, \theta = \theta^n) d\tilde{y}_{mis}.$$

Der M-Schritt findet einen neuen Wert θ^{n+1} indem er die im E-Schritt bestimmte Funktion $Q(\theta|\theta^n)$ maximiert, so daß für alle Werte von θ die Ungleichung

$$Q(\theta^{n+1}|\theta^n) \geq Q(\theta|\theta^n)$$

erfüllt ist.

Wenn die Log-Likelihood-Funktion linear in den erschöpfenden Schätzfunktionen („sufficient statistics“) ist, werden im E-Schritt die Erwartungswerte aller erschöpfenden Schätzfunktionen („sufficient statistics“) für die gesamten Daten Y auf Grundlage der verfügbaren Daten y_{obs} und des zuletzt berechneten Werts für den Parameter θ bestimmt (Dempster et. al. (1977)). Die Erwartungswerte der erschöpfenden Schätzfunktionen für die gesamten Daten Y werden wiederum verwendet, um im Rahmen des M-Schritts einen neuen Wert für den Parameter θ zu bestimmen.

Wegen $f(Y|\theta) = f(Y_{mis}|Y_{obs}, \theta) f(Y_{obs}|\theta)$ gilt die Größe, welche durch den EM-Algorithmus maximiert werden soll:

$$\log L(\theta|y_{obs}) = \log L(\theta|y_{obs}, Y_{mis}) - \log f(Y_{mis}|y_{obs}, \theta).$$

Bildet man den Erwartungswert dieser Größe hinsichtlich der Wahrscheinlichkeitsdichte $f(Y_{mis}|y_{obs}, \theta^n)$, so erhält man

$$\log L(\theta|y_{obs}) = Q(\theta|\theta^n) - H(\theta|\theta^n)$$

mit

$$Q(\theta|\theta^n) = \int \log L(\theta|\tilde{Y}_{mis}, y_{obs}) f(\tilde{Y}_{mis}|y_{obs}, \theta^n) d\tilde{Y}_{mis}$$

und

$$H(\theta|\theta^n) = \int \log f(\tilde{Y}_{mis}|y_{obs}, \theta) f(\tilde{Y}_{mis}|y_{obs}, \theta) d\tilde{Y}_{mis}.$$

Betrachtet man nun die Differenz der Log-Likelihood-Funktionen

$$\log L(\theta^{n+1}|y_{obs}) - \log L(\theta^n|y_{obs}) = Q(\theta^{n+1}|\theta^n) - Q(\theta^n|\theta^n) + H(\theta^n|\theta^n) - H(\theta^{n+1}|\theta^n),$$

so fällt unmittelbar auf, daß der Term $Q(\theta^{n+1}|\theta^n) - Q(\theta^n|\theta^n)$ durch die Wahl von θ^{n+1} im Rahmen des M-Schritts nichtnegativ sein muß und im allgemeinen positiv sein dürfte. Für den verbleibenden Term gilt

$$H(\theta^n|\theta^n) - H(\theta^{n+1}|\theta^n) = \int \log \left(\frac{f(\tilde{Y}_{mis}|y_{obs}, \theta^n)}{f(\tilde{Y}_{mis}|y_{obs}, \theta^{n+1})} \right) f(\tilde{Y}_{mis}|y_{obs}, \theta^n) d\tilde{Y}_{mis}.$$

Wegen der Konvexität der Funktion $g(x) = x \log x$ folgt aus Jensens Ungleichung

$$\int \left(\frac{f(\tilde{Y}_{mis}|y_{obs}, \theta^n)}{f(\tilde{Y}_{mis}|y_{obs}, \theta^{n+1})} \right) \log \left(\frac{f(\tilde{Y}_{mis}|y_{obs}, \theta^n)}{f(\tilde{Y}_{mis}|y_{obs}, \theta^{n+1})} \right) f(\tilde{Y}_{mis}|y_{obs}, \theta^{n+1}) d\tilde{Y}_{mis} \geq 0.$$

Das Gleicheitszeichen gilt genau dann, wenn

$$f(Y_{mis}|y_{obs}, \theta^{n+1}) = f(Y_{mis}|y_{obs}, \theta^{n})$$

gilt.

Damit ist gezeigt, daß der EM-Algorithmus durch den Schritt von θ^n auf θ^{n+1} die Log-Likelihood-Funktion vergrößert.

Der EM-Algorithmus basiert auf der Idee, die gesuchte Größe anhand einer vorherigen Schätzung dieser Größe und der gegebenen Daten zu bestimmen, bis die Schätzungen der gesuchten Größe konvergieren. Diese Idee ist prinzipiell auch ohne Rekurs auf eine Likelihood-Funktion anwendbar - sogar, wenn die gesuchte Größe die fehlenden Daten selbst sind (vgl. Abschnitt 5.3.4).

Nachteilig ist, daß die Konvergenzgeschwindigkeit des EM-Algorithmus vom Anteil fehlender Daten in der Matrix Y abhängt: Je größer das Verhältnis der fehlenden zu den beobachteten Werten in der zugrundeliegenden Datenmatrix ist, desto langsamer konvergiert der EM-Algorithmus (Dempster et. al. (1977). Gerade bei online generierten Datensätzen wie beispielsweise beim MovieLens-Datensatz erreichen die Anteile fehlender Daten bisher nicht gekannte Größen.

3.2.3.2 Bayes'sche Parameterschätzverfahren

Die „Data Augmentation"-Methode (Tanner, Wong (1987)) kann als Kombination des EM-Algorithmus mit dem Verfahren der Multiplen Imputation aufgefaßt werden.

Es soll hier nicht unerwähnt bleiben, daß das „Data Augmentation"-Verfahren auf der Bayes'schen Statistik basiert. Im Rahmen der Bayes'schen Statistik wird auch θ als Zufallsvariable behandelt. Grundlegend für die Bayes'sche Statistik ist der Satz von Bayes, der folgende Beziehung zwischen θ und Y_{obs} beschreibt:

$$P(\theta|Y_{obs}) \propto P(Y_{obs}|\theta)P(\theta).$$

Hier ist $P(\theta)$ die sogenannte Prior von θ, in der das gesamte von Y_{obs} unabhängige Vorwissen im Hinblick auf θ enthalten sein sollte. Dieses Vorwissen kann aus sehr unterschiedlichen Quellen stammen. Diese Quellen dürfen sich durchaus strukturell von den erhobenen Daten Y_{obs} unterscheiden. Beispielsweise können die

Vorkenntnisse aus völlig anders durchgeführten Versuchen stammen, aus Expertenbefragungen oder auch bloß die Ansicht des Experimentators vor der Erhebung der Daten Y_{obs} wiederspiegeln (Gelman et. al. (1995)). $p(Y_{obs}|\theta)$ ist die Likelihood und $P(\theta|Y_{obs})$ wird als Posterior von θ bezeichnet. Die Posterior integriert sämtliches Vorwissen aus der Prior mit der in den Daten Y_{obs} enthaltenen Information. Eine weitergehende Einführung in die Bayes'sche Statistik wird für das Verständnis des „Data Augmentation“-Verfahrens nicht benötigt und soll daher an dieser Stelle unterbleiben. Eine kurze systematische Einführung in die Bayes'sche Statistik erfolgt in Abschnitt 8.1.

Genau wie der EM-Algorithmus iteriert die „Data Augmentation“-Methode nacheinander zwei Schritte. Im sogenannten Imputationsschritt (I-Schritt), wird ein neuer Wert für die fehlenden Daten y_{mis}^{n+1} aus der Posterior $P(Y_{mis}|y_{obs}, \theta^n)$ gezogen:

$$y_{mis}^{n+1} \sim P(Y_{mis}|y_{obs}, \theta^n).$$

Danach erhält man den neuen Wert θ^{n+1} für θ durch Ziehung aus der Posterior $P(\theta|y_{obs}, y_{mis}^{n+1})$:

$$\theta^{n+1} \sim P(\theta|y_{obs}, y_{mis}^{n+1}).$$

Dieses Vorgehen entspricht dem Gibbs-Sampling. Deshalb ist die „Data Augmentation“-Methode ein Spezialfall des Metropolis-Hastings Algorithmus. Tierney (1994) konnte zeigen, daß die durch den Metropolis-Hastings Algorithmen erzeugten Folgen geometrisch gegen die Posterior-Verteilung konvergieren, sofern diese im gesamten Parameterbereich positiv ist. Daher konvergieren die durch die „Data Augmentation“-Methode erzeugte Folgen $\{y_{mis}^n\}$ und $\{\theta^n\}$ gegen eine Ziehung aus der Posterior $P(Y_{mis}, \theta|y_{obs})$. Da die Konvergenz erst nach einer bestimmten Anzahl von Iterationen eintritt, dürfen nur die Glieder der Folge Y_{mis}^n, θ^n berücksichtigt werden, die gezogen wurden, nachdem die Folge schon gegen Ziehungen aus der Gleichgewichtsverteilung $P(Y_{mis}, \theta|y_{obs})$ konvergiert ist. Dafür, ab wievielen Iterationen für praktische Zwecke davon ausgegangen werden kann, daß die Kongergenz hinreichend gut angenähert ist, existieren quantitative Maße (siehe Anhang E). Sofern die Konvergenz nicht schnell genug eintritt, empfiehlt es sich, den verwendeten Algorithmus zu modifizieren, um die Konvergenz

zu beschleunigen (Liu, Rubin (1996), (2002)). Im Unterschied zu allen übrigen Bayesischen Verfahren ermittelt die „Data Augmentation“-Methode neben der Verteilung für die Schätzer θ auch die Verteilung der fehlenden Daten Y_{mis} und kann somit auch benutzt werden, wenn die Vervollständigung der Datenmatrix selbst das Ziel der Analyse ist.

3.3 Praxisrelevante Schlußfolgerung

Durch eine Strukturanalyse kann gegebenenfalls belegt werden, daß eine MCAR-, MAR- oder OAR-Annahme nicht aufrechtzuerhalten ist. Dagegen ist die Strukturanalyse nicht dazu geeignet, den direkten Nachweis der MCAR-, MAR- oder OAR-Eigenschaft zu erbringen. Daher müssen immer mehrere Verfahren zur Strukturanalyse verwendet werden. Nur wenn keines der Ergebnisse dieser Methoden gegen die zu testende Annahme(n) spricht, darf (dürfen) diese aufrechterhalten werden.

Sofern bereits aus anderen Gründen davon auszugehen ist, daß bestimmte Annahmen nicht erfüllt sind, erscheint der mit der Strukturanalyse großer Datenmatrizen verbundene hohe numerische Aufwand in vielen Fällen nicht gerechtfertigt.

Gerade die zum Zweck der Empfehlung durch Recommender-Systeme zu berechnenden Prognosen müssen möglichst schnell durchgeführt werden. Nur so können neue Bewertungen der Nutzer zeitnah zur Verbesserung der (auf den vorhergesagten Bewertungen basierenden) Empfehlungen benutzt werden. Insbesondere bei neuen Nutzern ist dies wichtig, da neue Nutzer i.d.R. nur dann bereit sein werden, weitere Zeit und Mühe in die Bewertung von Items zu investieren, wenn sie im Anschluß daran als hilfreich empfundene Empfehlungen erhalten. Je weniger Bewertungen man zur Prognose verwendet, umso schlechter sind i.d.R. die resultierenden Empfehlungen. Daher ist es insbesondere im Hinblick auf die Qualität der Empfehlungen für neue Nutzer entscheidend, daß die Berechnung der Prognosen zeitnah erfolgt. Falls die verwendeten Verfahren nicht schnell genug sind, bleiben neue Bewertungen lange ungenutzt, was insbesondere bei neuen Nutzern zu schlechten Empfehlungen führen kann, da hier ohnehin nur wenige Bewertungen vorliegen. Schlechte Erfahrungen zu Beginn können dazu führen, daß der Nutzer das Recommender-System in Zukunft nicht mehr verwendet. Hört ein neuer Nutzer auf, das Recommender-System eines Online-Shops in Anspruch zu nehmen, so verfehlt das Recommender-System in Bezug auf die be-

treffende Person ihren Zweck, da es dann künftig nicht mehr in der Lage ist, einen Beitrag zu seiner Kundenzufriedenheit und Kundenbindung zu leisten. Darüberhinaus besteht sogar die Gefahr, daß die Unzufriedenheit mit dem Recommender-System eines Online-Shops die Zufriedenheit des betreffenden Kunden mit dem Online-Shop vermindert. Außerdem ist auch in Bezug auf Nutzer, die bereits viele Bewertungen abgegeben haben, zu berücksichtigen, daß eine Empfehlung umso hilfreicher ist, je aktueller sie ist. Daher ist die Schnelligkeit der zur Vorhersage verwendeten Verfahren auch im Hinblick auf alle Nutzer dieses Systems von ausschlaggebender Bedeutung. Deshalb erscheint es in diesem Zusammenhang wenig zweckdienlich, vor dem eigentlichen Prognose-Verfahren eine hohe Anzahl zeitintensiver Verfahren zur Strukturanalyse auszuführen. Vor diesem Hintergrund ist eher von vornherein die Verwendung einer Methode zu empfehlen, die möglichst wenig Annahmen erfordert.

Die meisten der bekannten Verfahren zur Behandlung fehlender Werte erfordern jedoch sehr weitgehende Annahmen und viele davon sind (wie dargelegt) zur Anwendung auf große Datenmatrizen mit extrem hohem Fehlendanteil völlig ungeeignet.

Das häufig in der Praxis zu findende Ignorieren fehlender Daten setzt implizit die MCAR-Eigenschaft voraus und kann - sofern diese nicht gegeben ist - zu Verzerrungen führen (Little, Rubin (2002)).

Auch die oft verwendete spalten- oder zeilenweise Mittelwertimputation basiert auf der MCAR-Annahme und kann bei deren Nichterfülltsein zu Verzerrungen führen (Little (1986)). Außerdem unterschätzen die resultierenden Stichprobenvarianzen die Varianz der Grundgesamtheit (vgl. Abschnitt 3.2.2.1). Außerdem kann es zu Inkonsistenzen kommen, da auf dieser Basis berechneten Varianz-Kovarianzmatrizen nicht positiv definit sein müssen.

Das „Data Augmentation“-Verfahren macht die MAR-Eigenschaft erforderlich. Außerdem setzt sie voraus, daß zwei bedingte Posteriors bekannt sind und ist vergleichsweise zeitaufwendig.

Gerade die MAR- (und damit zwangsläufig auch die MCAR-) Eigenschaft kann jedoch insbesondere im Zusammenhang mit online-generierten Bewertungsdaten nicht unbedingt vorausgesetzt werden. Schließlich bewerten die Nutzer i.d.R. nur Items, mit denen sie Erfahrungen haben - und dies sind hauptsächlich die Items, von denen sie sich zu irgend einem Zeitpunkt in der Vergangenheit etwas versprochen haben. Daher ist anzunehmen, daß in vielen Fällen die Tat-

sache, daß ein Nutzer ein Item nicht bewertet hat, ein Indiz für eine weniger positive Einstellung des betreffenden Nutzers gegenüber dem Item ist. Somit ist das Vorhandensein einer Bewertung nicht unabhängig von ihrer Ausprägung, weshalb die MAR- und MCAR-Annahmen nicht erfüllt sind. Daher ist im Kontext mit online-generierten Bewertungsdaten sowohl das Ignorieren fehlender Werte als auch die Imputation von Mittelwerten, die auf der Basis aller vorhandenen Daten bezüglich einer Modalität gebildet werden, problematisch.

Kapitel 4

Nicht-Bayes'sche kontentbasierte Verfahren

Die Verfahren, die man zur Schätzung der fehlenden Daten einer Zeile aus der „unvollständigen“ ordinalen Datenmatrix (y_{ij}) benutzt, werden klassisch danach unterteilt, welche Daten zur Schätzung dieser fehlenden Daten verwendet werden. Man unterscheidet sogenannte kontenentbasierte, kollaborative und hybride Verfahren (Adomavicius, Tuzhilin (2005)).

Im Hinblick auf jeden Nutzer $i \in \{1, ..., I\}$ werden im Rahmen der kontentbasierten Verfahren die Eigenschaften aller Items $a_j, j \in J_i$, für die er Bewertungen abgegeben hat, zur Schätzung der noch fehlenden Werte in der i-ten Zeile der Datenmatrix verwendet. (J_i bezeichnet die Menge aller Items, die der jeweilige Nutzer i bewertet hat: $J_i = \{j \in \{1, ..., J\} | v_{ij} = 1\}$.) Zu diesem Zweck ist erforderlich, daß die Eigenschaften a_j aller Items $j \in \{1, ..., J\}$ bekannt sind. Die kontentbasierten Verfahren benutzen die Ähnlichkeit der Items $j \in \{1, ..., J\} \backslash J_i$, die der betrachtete Nutzer i nicht bewertet hat, mit den Items aus der Menge J_i, die der Nutzer i bewertet hat, um herauszufinden, welche Items aus der Menge $\{1, ..., J\} \backslash J_i$ besonders interessant für den Nutzer i sein könnten.

Dagegen benutzen die Kollaborativen Verfahren die vorhandenen Einträge in den restlichen Zeilen der Datenmatrix $y = (y_{ij})$ um die fehlenden Werte in der i-ten Zeile derselben Matrix zu approximieren. Hierbei bedient man sich der Ähnlichkeit des i-ten Nutzers zu anderen Nutzern in der Datenmatrix um Näherungen für die fehlenden Werte in der i-ten Zeile von Y zu finden.

Hybride Verfahren sind solche, die sich zur Schätung der (fehlenden) Bewertungen jeder einzelnen Zeile von (y_{ij}) sowohl der Eigenschaften der Items

$a_j, j \in \{1, ..., J\}$ als auch aller in den übrigen Zeilen von (y_{ij}) vorhandenen Werte bedienen.

Alternativ könnten die verschiedenen Verfahren auch anhand ihres theoretischen Hintergrunds klassifiziert werden. Dies ist besonders deshalb naheliegend, da insbesondere in letzter Zeit immer mehr Verfahren zum Einsatz kommen, deren theoretischer Hintergrund erklärungsbedürftig ist. Es handelt sich hierbei um die im Rahmen von Kapitel 9 dargestelten Bayes'schen Verfahren. In diesem Kapitel werden nur Verfahren behandelt, die ohne Kenntnisse der Bayes'schen Statistik verständlich sind.

Eine weitere Möglichkeit wäre, die Verfahren danach zu unterscheiden, ob die fehlenden Werte der Datenmatrix $y = (y_{ij})$ oder nur die fehlenden Werte einer Transformation dieser Datenmatrix $T(\mathrm{y})$ geschätzt werden. Denn manchmal wird anstelle der Matrix y die zu einer binären Matrix transformierte Matrix $T(\mathrm{y})$ betrachtet (z.B. Billsus, Pazzani (1998)). Hierbei werden alle Werte y_{ij}, die nach Ansicht der betreffenden Autoren auf ein verstärktes Interesse der Person i am Item j schließen lassen in der transformierten Matrix $T(\mathrm{y})$ durch 1 ersetzt und alle übrigen durch 0.

Ebenfalls interessante Unterscheidungsmerkmale der verschiedenen Ansätze ist der Umgang mit fehlenden Werten in y. Manche Verfahren erfordern, daß die Datenmatrix zuerst durch plausible Werte vervollständigt wird. Die so vervollständigte Matrix wird dann benutzt, um Schätzer für die fehlenden Werte zu bestimmen. Nicht selten handelt es sich hierbei um einfache Mittelwert-Imputationen. In anderen Fällen werden die fehlenden Einträge in der Datenmatrix wie im Rahmen der kontentbasierten Methoden einfach ignoriert.

Weiterhin findet sich in der Literatur insbesondere für die kollaborativen Ansätze sehr oft die Unterteilung in speicherbasierte (heuristische) und modellbasierte Verfahren (Breese et. al. (1998)). Speicherbasierte Verfahren sind dabei alle Verfahren, die im Rahmen von Heuristiken alle vorhandenen Bewertungen zur Schätzung der fehlenden Werte heranziehen. Dagegen benutzen die modellbasierten Verfahren die Daten zuerst, um die Parameter und zum Teil auch die Struktur eines geeigneten Modells zu schätzen. Die auf diese Weise bestimmten Schätzer werden dann eingesetzt, um Approximationen für die fehlenden Werte zu berechnen.

Es ist außerdem nicht uninteressant, die modellbasierten Verfahren in lineare und nichtlineare Ansätze aufzuteilen.

Zudem können die Verfahren danach unterschieden werden, welches Skalenniveau sie voraussetzen. Hierbei berücksichtigen die wenigsten Verfahren das ordinale Skalenniveau der Bewertungsdaten. Die meisten verwendeten Verfahren sind eigentlich nur für kardinale Daten geeignet. Daneben werden auch Verfahren für nominale Daten verwendet.

Pazzani, Billsus (1997) vergleichen eine Reihe kontentbasierter Heuristiken und Modelle (darunter die Entscheidungsbäume C4.5 und CART) und zeigen empirisch, daß die sogenannten Naiven Bayes'schen Klassifikatoren, die TF-IDF Profil-Heuristik und ein Neuronales Netz den anderen anderen Verfahren im Hinblick auf das Schätzen der fehlenden Werte deutlich überlegen sind. Deshalb werden in diesem Abschnitt nur die drei genannten Verfahren behandelt, die in der Studie von Pazzani, Billsus (1997) vergleichbar gute Ergebnisse lieferten.

Die TF-IDF Heuristik und die Naiven Bayes'schen Klassifkatoren wurden speziell zur Empfehlung von Dokumenten entwickelt.

4.1 Die TF-IDF Profil-Heuristik

Die TF-IDF („term frequency inverse-document frequency“) Profil-Heuristik ist ein Information Retrieval Verfahren, das vor allem zur Empfehlung von Texten (Lang (1995)) und Webseiten (Balabanovic, Shoham (1997), Pazzani, Billsus (1997)) eingesetzt wurde.

Sei $a(t, d)$ die Anzahl, die angibt, wie oft ein bestimmtes Wort t im Dokument d vorkommt. Die Häufigkeit mit der ein bestimmtes Wort (Term) t in einem Dokument d vorkommt („term frequency“) kann als erster Anhaltspunkt dafür dienen, wie charakteristisch der Term t aus der Menge aller verwendeten Terme T für das Text-Dokument d aus der Dokumentenmenge D ist. Sind die Texte unterschiedlich lang, so ist die Termhäufigkeit allerdings erst nach geeigneter Normierung aussagekräftig. In langen Dokumenten werden oft dieselben Ausdrücke wiederholt benutzt. Ohne geeignete Normierung wären darum die Termhäufigkeiten hinsichtlich langer Dokumente größer (Singhal et. al. (1996)). Es existiert in der Literatur eine Vielzahl verschiedener Normierungsvarianten (vgl. Salton, Buckley (1988), Robertson et. al. (1996)). Häufig verwendete Ansätz sind die Maximumsnormalisierung (Salton, Buckley (1988), Lang (1995))

$$TF(t, d) = \frac{a(t, d)}{\max_{\underline{\tau} \in T} a(\underline{\tau}, d)} \tag{4.1}$$

und

$$TF(t,d) = \begin{cases} 0 & , \text{ falls } a(t,d) = 0 \\ \log(a(t,d)) & , \text{ sonst} \end{cases} . \tag{4.2}$$

D sei die Menge aller betrachteten Texte (Dokumente). $a_D(t)$ bezeichnet die Anzahl der Dokumente aus D, in denen der Term t vorkommt. Je mehr Dokumente $d \in D$ einen bestimmten Term t enthalten, umso ungeeigneter ist dieser Term um die Texte aus der Menge D voneinander zu unterscheiden. Deshalb wird die Termhäufigkeit noch mit dem Logarithmus der Größe

$$IDF(t) = \frac{|D|}{a_D(t)}$$

(„inverse document frequency") multipliziert, wobei $|D|$ die Anzahl der betrachteten Dokumente ist. Auf dieser Basis kann man unterschiedliche Versionen des Gewichtungsvektors $g^D(d) = (g_1^D(d), ..., g_{|T|}^D(d))'$ für jedes einzelne Dokument $d \in D$ erhalten. Sofern die Maximumsnormalisierung (4.1) verwendet wird, ist $g_t^D(d)$ gleich dem dem Produkt von $TF(t,d)$ und $\log(IDF(t)), t \in T$. Unter Verwendung von Formel (4.2) benutzt man dagegen üblicherweise die Cornell-Gewichtung (Buckley et. al. (1993),(1994), Ittner et. al. (1995)). Insgesamt ergeben sich die beiden häufig verwendeten Ansätze:

$$g_t^D(d) = \begin{cases} TF(t,d)\log(IDF(t)), & \text{für Formel (4.1)} \\ \dfrac{TF(t,d)\log(IDF(t))}{\sqrt{\sum_{\tau \in T} TF(\tau,d)\log(IDF(\tau))}}, & \text{für Formel (4.2)} \end{cases} , t \in T.$$

D_i ist die Menge der Texte, die der Nutzer i bewertet hat. Falls sich die Bewertungen auf Textdokumente $d \in D$ beziehen, gilt $J_i = D_i, J = |D|$. Für jeden Nutzer $i \in \{1, ..., I\}$ benutzt man die i-te Zeile der Matrix (Y_{id}) und alle zugehörigen Gewichte $g^D(d), d \in D_i$, um ein (vektorielles) Nutzer-Profil zu erstellen. Sei $B \subset \{1, ..., C\}$ eine Teilmenge aller Bewertungsmöglichkeiten $\{1, ..., C\}$, dann ist $R_{i1} = \{d \in D_i | y_{id} \in B\}$ die Menge aller Dokumente, hinsichtlich derer die Bewertungen des Nutzers i im Bereich B liegen. Analog ist $R_{i0} = \{d \in D_i | y_{id} \notin B\}$ die Menge aller Beurteilungen, die von i vorgenommen wurden und nicht in B

liegen. Auf der Basis dieser Definitionen gibt es zwei bekannte Möglichkeiten, ein Nutzer-Profil $\tilde{g}^U(i) = (\tilde{g}_1^U(i), ..., \tilde{g}_{|T|}^U(i))', i \in \{1, ..., I\}$ zu erstellen:

$$\tilde{g}_t^U(i) = \begin{cases} \frac{1}{|R_{i1}|} \sum_{\underline{d} \in R_{i1}} g_t^D(\underline{d}) & , \quad \text{Variante 1 (Lang (1995))} \\ \\ \tilde{g}_t^{U**}(i) & , \quad \text{Variante 2 (Billsus, Pazzani (1997))} \end{cases},$$

mit

$$\tilde{g}_t^{U**}(i) = \begin{cases} \tilde{g}_t^{U*}(i) & , \quad \text{für } \tilde{g}_t^{U*}(i) > 0 \\ 0 & , \quad \text{für } \tilde{g}_t^{U*}(i) \leq 0 \end{cases},$$

und

$$\tilde{g}_t^{U*}(i) = \frac{\alpha_U^1}{|R_{i1}|} \sum_{\underline{d} \in R_{i1}} g_t^D(\underline{d}) - \frac{\alpha_U^0}{|R_{i0}|} \sum_{\underline{d} \in R_{i0}} g_t^D(\underline{d})$$

wobei $\alpha_U^1 = 4\alpha_U^0$ ist. Hierbei wird B so gewählt, daß R_{i1} die Menge der Dokumente ist, hinsichtlich derer man von einem erhöhten Interesse des Nutzers i ausgeht. Auf diese Weise werden die Bewertungdaten Y nach dem Schema

$$T_B(Y)_{id} = \begin{cases} 1 & , \quad \text{falls } Y_{id} \in B \\ 0 & , \quad \text{sonst} \end{cases}$$

transformiert. Durch diese Transformation wird die vorhandene Information vergröbert. Dierdurch wird das vorgestellte Verfahren gleichermaßen auf ordinale, kardinale und binäre nominale Daten anwendbar.

Sowohl Lang (1995) als auch Pazzani, Billsus (1997) benutzen den Kosinus

$$\cos(\sphericalangle(g^U(i), g^D(d))) = \frac{g^U(i) \cdot g^D(d)}{\|g^U(i)\| \, \|g^D(d)\|} = \frac{\sum_{\underline{\tau} \in T} g_{\underline{\tau}}^U(i) g_{\underline{\tau}}^D(d)}{\sqrt{\sum_{\underline{\tau} \in T} \left(g_{\underline{\tau}}^U(i)\right)^2} \sqrt{\sum_{\underline{\tau} \in T} T \left(g_{\underline{\tau}}^D(d)\right)^2}}$$

des Winkels zwischen den Vektoren $g^D(d)$ und $g^U(i)$ um abzuschätzen, wie interessant das Dokument d für den Nutzer i sein dürfte. Je größer $\sphericalangle(\cos(g^U(i), g^D(d)))$ ausfällt, umso mehr scheint Dokument d der Person i zu nutzen. Daher empfiehlt

man einem Nutzer i die Dokumente $d \in D$, hinsichtlich derer $\cos(\sphericalangle(g^U(i), g^D(d)))$ besonders groß ausfällt.

Die TF-IDF Profil-Heuristik eignet sich zur Klassifikation von Texten. Eine Möglichkeit, dieses Verfahren z.B. zur Klassifikation der Produkte eines Online-Shops einzusetzen, wäre, alle Produkte anhand des zugehörigen Beschreibungstexts auf der Webseite zu klassifizieren.

4.2 Neuronale Netze

Ein weiteres von Pazzani, Billsus (1997) verwendetes kontentbasiertes Verfahren sind die Künstlichen Neuronalen Netze. Pazzani, Billsus (1997) verwenden ein sogenanntes Mehrschichtige Netzwerk mit Rückpropagation (Rumelhart et. al. (1986)). Die Abbildung 4.1 zeigt die am häufigsten verwendete Variante der Mehrschichtigen Netzwerke, das sogenannte Einschichtperzeptron („single hidden layer feed-forward neural network"). Diese Variante wurde auch von Billsus, Pazzani (1997) eingesetzt. Ein solches Netzwerk wird für jeden einzelnen Nutzer $i \in \{1, ..., I\}$ erstellt.

Der Vektor $a_j = (a_{j1}, ..., a_{j\kappa_A})', j \in J_i$, ist eine vollständige vektorielle Beschreibung des j-ten Items aus J_i. (Sind die Items Text-Dokumente, so ist $J_i = D_i$, $P = |T|$ und $a_j = g^D(j)$.)

Zwischen den Inputs $a_{j\kappa}, j \in J_i, \kappa = 1, ..., \kappa_A$, und den internen Einheiten des „hidden layers" $\tilde{Z}_{mj}, j \in J_i, m = 1, ..., M$, wird mittels der sogenannten Aktivierungsfunktion A_{akt} die nichtlineare Beziehung $\tilde{Z}^i_{mj} = A_{akt}(\underline{\alpha}^i_{m0} + \underline{\alpha}^{i'}_m a_j)$, hergestellt, wobei gilt $a_j = (a_{j1}, ..., a_{j\kappa_A})', j \in J_i, \underline{\alpha}^i_m = (\underline{\alpha}_{m1}, ..., \underline{\alpha}_{m\kappa_A})', m = 1, ..., M$. Dabei gibt M die Anzahl der inneren Einheiten an. Pazzani, Billsus (1997) verwenden $M = 12$. Die häufigste Wahl für die Aktivierungsfunktion ist die Sigmoidfunktion

$$A_{akt}(x) = \frac{1}{1 + \exp(-x)}.$$

Neuronale Netzwerke können sowohl zu Regressionszwecken als auch für Klassifikationsaufgaben eingesetzt werden. Im Hinblick auf Klassifikationsprobleme entspricht die Anzahl der verwendeten Klassen $\bar{z}$ der Anzahl der Output-Einheiten $g_{\bar{z}}(\mathbf{T}^i_{\bar{z}j})$. $g_{\bar{z}}(\mathbf{T}^i_{\bar{z}j})$ gibt an, wie wahrscheinlich ein Item j im Hinblick auf die Person i einer Klasse $\bar{z} \in \{1, 0\}$ zugeordnet werden kann. Mit $\tilde{Z}^i_j = (\tilde{Z}^i_{1j}, ..., \tilde{Z}^i_{Mj})'$ und

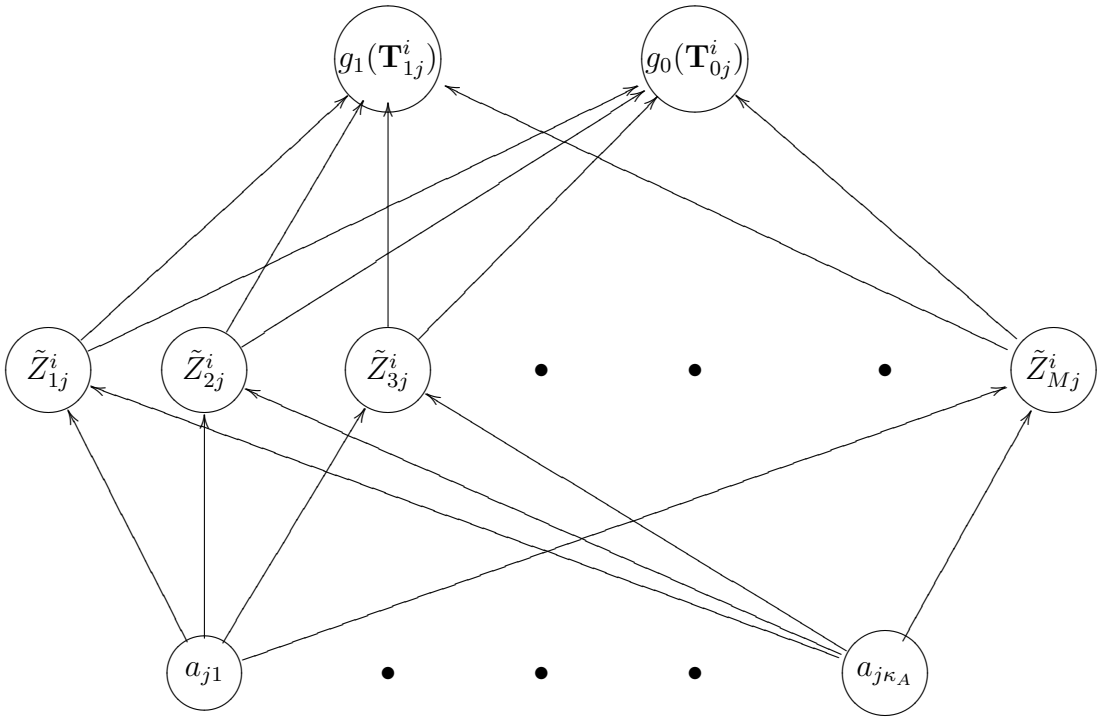

Abbildung 4.1: Mehrschichtiges Netzwerk zur Klassifikation im Zwei-Klassen Fall

$\mathbf{T}^i_{\bar{z}j} = \underline{\beta}^i_{\bar{z}0} + \underline{\beta}^{i\prime}_{\bar{z}} \tilde{Z}^i_j, \underline{\beta}^i_{\bar{z}} = (\underline{\beta}^i_{\bar{z}1}, ..., \underline{\beta}^i_{\bar{z}M})'$, werden diese Output-Einheiten durch die Formel

$$g_{\bar{z}}(\mathbf{T}^i_{\bar{z}j}) = \frac{\exp(\mathbf{T}^i_{\bar{z}j})}{\exp(\mathbf{T}^i_{1j}) + \exp(\mathbf{T}^i_{0j})}, \quad \bar{z} \in \{1, 0\},$$

berechnet.

Die Beziehung der Inputs zu den inneren Einheiten wird für jede Person durch $M(\kappa_A + 1)$ Gewichte $\underline{\alpha}^i_{m0}, \underline{\alpha}^i_m, m = 1, ..., M$, bestimmt. Um die nachfolgende Diskussion zu vereinfachen, werden hier keine Achsenabschnitte betrachtet.

In manchen Lehrbüchern (z.B. Rojas (1993)) wird dieses Diagramm mit einer 1 als zusätzlichen Input-Einheit dargestellt, die dem Einfluß des Achsenabschnitts $\underline{\alpha}^i_{m0}$ Rechnung tragen soll und mit allen inneren Elementen verknüpft ist. Ebenso wird machmal der Einfluß des Achsenabschnittsparameters $\underline{\beta}^i_{\bar{z}0}$ als zusätzliches inneres Element mit Verbindungen zu allen Output-Einheiten aber ohne Einflüsse der Inputs eingezeichnet.

Um die Parameter eines Neuronalen Netzes zu bestimmen, verwendet man

eine Fehlerfunktion $\mathcal{F}_i^{NN}$. Pazzani, Billsus (1997) verwenden zur Berechnung der Parameter das Rückpropagationsverfahren nach Rumelhart et. al. (1986), das zu diesem Zweck die quadratische Fehlerfunktion

$$\mathcal{F}_i^{NN} = \sum_{\bar{z}' \in \{1,0\}} \sum_{j' \in J_i} (T_{\bar{z}'}(y_{ij'}) - g_{\bar{z}}(\mathbf{T}^i_{\bar{z}'j'}))^2 = \sum_{j' \in J_i} \mathcal{F}_{ij'}^{NN}, \quad i = 1, ..., I,$$

mit

$$T_{\bar{z}}(y_{ij}) = \begin{cases} 1 & , \text{ falls } y_{ij} \geq \gamma_{\mathcal{R}} \wedge \bar{z} = 1 \\ 1 & , \text{ falls } y_{ij} < \gamma_{\mathcal{R}} \wedge \bar{z} = 0 \\ 0 & , \text{ sonst} \end{cases}$$

einsetzt. Die partiellen Ableitungen von $\mathcal{F}_{ij}^{NN}, j \in J_i, i = 1, ..., I$, für $\bar{z} \in \{1, 0\}$

$$\frac{\partial \mathcal{F}_{ij}^{NN}}{\partial \underline{\beta}^i_{\bar{z}m}} = \underbrace{-2(T_{\bar{z}}(Y_{ij}) - g_{\bar{z}}(\mathbf{T}^i_{\bar{z}j})) \frac{\partial g_{\bar{z}}(\mathbf{T}^i_{\bar{z}j})}{\partial \mathbf{T}^i_{\bar{z}j}}}_{= \underline{\delta}^i_{\bar{z}j}} \tilde{Z}^i_{mj} = \underline{\delta}^i_{\bar{z}j} \tilde{Z}^i_{mj}, \quad m = 1, ..., M,$$

und für $\kappa = 1, ..., \kappa_A$ und $m = 1, ..., M$,

$$\frac{\partial \mathcal{F}_{ij}^{NN}}{\partial \underline{\alpha}^i_{m\kappa}} = \underbrace{\sum_{\bar{z}' \in \{1,0\}} \underline{\delta}^i_{\bar{z}'j} \underline{\beta}^i_{\bar{z}'m} \frac{dA_{akt}(\underline{\alpha}^i_{m0} + \underline{\alpha}^{i'}_m a_j)}{d(\underline{\alpha}^i_{m0} + \underline{\alpha}^{i'}_m a_j)}}_{\underline{s}^i_{mj}} a_{j\kappa} = s^i_{mj} a_{j\kappa},$$

führen auf die Rückpropagationsgleichung

$$\underline{s}^i_{mj} = \frac{dA_{akt}(\underline{\alpha}^i_{m0} + \underline{\alpha}^{i'}_m a_j)}{d(\underline{\alpha}^i_{m0} + \underline{\alpha}^{i'}_m a_j)} \sum_{\bar{z}' \in \{1,0\}} \beta^i_{\bar{z}'m} \underline{\delta}^i_{\bar{z}'j}, \tag{4.3}$$

$j \in J_i, i = 1, ..., I, m = 1, ..., M$. Die entsprechenden Beziehungen bezüglich $\underline{\alpha}^i_{m0}$ und $\underline{\beta}^i_{\bar{z}0}$ lassen sich analog bestimmen. Das Rückpropagationsverfahren, das auch als Fehlerrückführungsmethode bezeichnet wird, beginnt mit Anfangswerten $\underline{\alpha}^{i[0]}_{m\kappa}$ und $\underline{\beta}^{i[0]}_{\bar{z}m}$ für $\underline{\alpha}^i_{m\kappa}$ und $\underline{\beta}^i_{\bar{z}m}, m = 0, ..., M, \kappa = 1, ..., \kappa_A, \bar{z} \in \{1, 0\}$ und alterniert zwischen Vorwärts- und Rückwärtsschritten. Im $(n+1)$-ten Schritt werden die im n-ten Schritt bestimmten Werte $\underline{\beta}^{i[n]}_{\bar{z}m}, \bar{z} \in \{1, 0\}, m = 1, ..., M$, und $\underline{\alpha}^{i[n]}_{m\kappa}$, mit $m = 0, ..., M, \kappa = 1, ..., \kappa_A$, verwendet, um daraus $\mathbf{T}^{i[n]}_{\bar{z}j} = \underline{\beta}^{i[n]}_{\bar{z}0} +$

$\underline{\beta}_{\bar{z}}^{i[n]'} A_{akt}(\underline{\alpha}_{m0}^{i[n]} + \underline{\alpha}_{m}^{i[n]'} a_j)$, mit $\bar{z} \in \{1,0\}, j \in J_i, i = 1,...,I$, zu berechnen und dadurch $g_{\bar{z}}(\mathbf{T}_{\bar{z}j}^{i[n]}), j \in J_i, i = 1,...,I$, $\bar{z} \in \{1,0\}$, zu bestimmen (n-ter Vortwärtsschritt). Im n-ten Rückwärtsschritt ermittelt man damit

$$\underline{\delta}_{\bar{z}j}^{i[n]} = -(T_{\bar{z}}(y_{ij}) - g_{\bar{z}}(\mathbf{T}_{\bar{z}j}^{i[n]})) \frac{\partial g_{\bar{z}}(\mathbf{T}_{\bar{z}j}^{i[n]})}{\partial \mathbf{T}_{\bar{z}j}^{i}}$$

und benutzt die Rückpropagationsgleichung (4.3), um hieraus $\underline{s}_{mj}^{i[n]}$ zu berechnen, $\bar{z} \in \{1,0\}, j \in J_i, i = 1,...,I$. Hiermit sind alle partiellen Ableitungen

$$\frac{\partial \mathcal{F}_{ij}^{NN,[n]}}{\partial \underline{\beta}_{\bar{z}m}^{i}} = \underline{\delta}_{\bar{z}j}^{i[n]} \tilde{Z}_{mj}^{i}, \bar{z} \in \{1,0\}, \;\; m = 1,...,M, j \in J_i, i = 1,...,I,$$

und

$$\frac{\partial \mathcal{F}_{ij}^{NN,[n]}}{\partial \underline{\alpha}_{m\kappa}^{i}} = \underline{s}_{mj}^{i[n]} a_{j\kappa}, \;\; \kappa = 1,...,\kappa_A, m = 1,...,M, j \in J_i, i = 1,...,I,$$

für $m = 1,...,M$ bestimmt. (Die entsprechenden Beziehungen bezüglich $\underline{\alpha}_{m0}^{i}$ und $\underline{\beta}_{\bar{z}0}^{i}$ werden analog bestimmt.) Auf Basis aller notwendigen partiellen Ableitungen können die Formeln

$$\underline{\alpha}_{m\kappa}^{i[n+1]} = \underline{\alpha}_{m\kappa}^{i[n]} - \gamma_{\mathcal{R}} \sum_{j' \in J_i} \frac{\partial \mathcal{F}_{ij'}^{NN,[n]}}{\partial \underline{\alpha}_{m\kappa}^{i}} \quad \text{und} \quad \underline{\beta}_{\bar{z}m}^{i[n+1]} = \underline{\beta}_{\bar{z}m}^{i[n]} - \gamma_{\mathcal{R}} \sum_{j' \in J_i} \frac{\partial \mathcal{F}_{ij'}^{NN,[n]}}{\partial \underline{\beta}_{\bar{z}m}^{i}}$$

dazu verwendet werden, neue Schätzer zu bestimmen. $\gamma_{\mathcal{R}}$ ist die Lernrate. Die neuen Schätzer können dann im $(n+1)$-ten Vorwärtsschritt wieder verwendet werden.

Dieses Modell liefert keinen Schätzer für die Bewertungsmatrix y. Aber es berechnet die Wahrscheinlichkeit der Zugehörigkeit zu einer Klasse von Items für eine bestimmte Person. Damit lassen sich Schätzer für die transformierte Datenmatrix $T(y)$ bestimmen.

Im Gegensatz zur TF-IDF Profil-Heuristik sind Neuronale Netze ein modellbasiertes Verfahren. Pazzani, Billsus (1997) nehmen eine Datentransformation vor. Diese Transformation führt zu einer starken Vergröberung der Daten, kann

j	Titel	a_{j1}	a_{j2}	a_{j3}	a_{j4}	a_{j5}	a_{j6}	Bewertung Y_{1j}
1	Armageddon	6	9	6	7	7	4	1
2	Magnolia	3	2	5	5	5	8	5
3	Casino	9	5	3	4	4	6	4
4	Taxi Driver	9	4	2	1	5	10	4
5	Platoon	9	9	3	0	9	9	3
6	Dangerous Liaisons	3	2	4	4	2	8	4
7	Barry Lyndon	5	4	2	4	3	8	-
8	Twister	6	8	4	4	4	3	-

Tabelle 4.1: Beispiel 4.1: Filmeigenschaften und Bewertungen

aber das ordinale Skalenniveau der Bewertungsdaten berücksichtigen. Daher ist das Modell auf ordinal- und kardinalskalierte Daten und auch auf binäre nominale Daten anwendbar. Die beschriebene Datentransformation kann zu Verzerrungen führen, da das Neuronale Netz nicht berücksichtigen kann, welche Items innerhalb der Menge B höher bzw. niedriger bewertet wurden. Falls B nur die beiden höchstmöglichen Bewertungen umfaßt ($B = \{C-1, C\}$) und deutlich mehr Items von einer Person mit $C-1$ als mit C bewertet wurden, lernt das Neuronale Netz im wesentlichen zwischen den mit $C-1$ bewerteten Items und den restlichen Items zu unterscheiden.

Prinzipiell wäre es möglich, die Bewertungsdaten untransformiert zu verwenden und einfach soviele Klassen zu schätzen wie es unterschiedliche Bewertungen gibt. Das vorgestellte Modell würde auch dann dann das ordinale Skalenniveau nicht berücksichtigen.

Beispiel 4.1:

Das Maß an Gewalt (a_{j1}), Action (a_{j2}), Humor (a_{j3}), Romantik (a_{j4}), Spannung (a_{j5}) und Charakterentwicklung (a_{j6}) seien Eigenschaften der Filme Armageddon ($j = 1$), Magnolia ($j = 2$), Casino ($j = 3$), Taxi Driver ($j = 4$), Platoon ($j = 5$), Dangerous Liaisons ($j = 6$), Barry Lyndon ($j = 7$) und Twister ($j = 8$) wie unten dargestellt ($http : //reel.com$). Die Eigenschaften werden auf einer diskreten Skala von 0 bis 10 gemessen. Ein bestimmter Nutzer ($i = 1$) hat für die Filme $j = 1, ..., 6$ die in der letzten Spalte der

Tabelle angegebenen Bewertungen auf einer diskreten Skala von 1 bis 5 abgegeben. Das Ziel ist jetzt, herauszufinden, ob diesem Nutzer eher der 7. Film (“Barry Lyndon“) oder der 8. („Twister“) empfohlen werden sollte. Für den ersten Nutzer ergeben sich aus den Bewertungen $y_{1j}, j = 1, ..., 6$, (Tabelle 4.1) deren Transformationen $T_1(y_{12}) = T_1(y_{13}) = T_1(y_{14}) = T_1(y_{16}) = 1$ (also $T_0(y_{12}) = T_0(y_{13}) = T_0(y_{14}) = T_0(y_{16}) = 0$) und $T_1(y_{11}) = T_1(y_{15}) = 1$ (also $T_0(y_{11}) = T_0(y_{15}) = 0$). Man verwendet nun die Daten der ersten 6 Personen um das Netzwerk zu trainieren. Die so bestimmten Parameter des Netzwerks können dann dazu verwendet werden, die letzten beiden Filme zu klassifizieren.

Für $M = 2$ und $\gamma_{\mathcal{R}} = 0,3$ erhält man die Gewichtungen zwischen den Input-Einheiten und den inneren Einheiten:

$\underline{\alpha}^1_{m\kappa}$	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$	$\kappa = 6$
$m = 1$	0,742	-1,782	0,049	0,816	-0,410	0,820
$m = 2$	2,274	-1,105	-1,159	-0,203	-0,585	-0,687

Tabelle 4.2: Beispiel 4.1: Gewichtungen zwischen den Input-Einheiten und den inneren Einheiten

Die Achsenabschnitte $\underline{\alpha}^1_{10} = 0,115$ und $\underline{\alpha}^1_{20} = -0,100$ können als das Negative der Schwellenwerte der Sigmoidfunktion interpretiert werden. Man erhält für die inneren Neuronen für Barry Lyndon $\tilde{Z}^1_{17} = 0,995$ und $\tilde{Z}^1_{27} = 0,028$ und für Twister $\tilde{Z}^1_{18} = 0,004$ und $\tilde{Z}^1_{28} = 0,006$.

Wie bei der logistischen Regression kann nur die Differenz zwischen den Vektoren $(\underline{\beta}^1_{\tilde{z}0}, ..., \underline{\beta}^1_{\tilde{z}M})'$ bestimmt werden. Daher kann man der allgemeinen Konvention folgen und $\underline{\beta}^1_0 = (\underline{\beta}^1_{00}, \underline{\beta}^1_{02})'$ auf Null setzen. Es ergibt sich für den Film Barry Lyndon mit $\underline{\beta}^1_{10} = -2,817$ und $\underline{\beta}^1_1 = (2,778, 3,140)'$ die geschätzte Wahrscheinlichkeit $g_1(\mathbf{T}_{17}) = 0,509$. Für Twister folgt $g_1(\mathbf{T}_{18}) = 0,058$. Man sollte daher der fiktiven Person $i = 1$ eher Barry Lyndon als Twister empfehlen.

4.3 Der genäherte „Naive Bayes“ Klassifikator

Der genäherte „Naive Bayes“ Klassifikator weist in der Form in der er als Kontentbasiertes Verfahren eingesetzt wurde (Pazzani, Billsus (1997), Mooney, Roy

(2000)) keine der Besonderheiten auf, die die Bayes'schen Verfahren im allgemeinen auszeichnen. Außerdem sind Kenntnisse der Bayes'schen Statistik nicht erforderlich, um diesen Ansatz zu verstehen.

Der vorgestellte Ansatz wurde von Mooney, Roy (2000) zur Empfehlung von Text-Dokumenten eingesetzt. Sei D_i die Menge der Dokumente, die der i-te Nutzer bewertet hat und sei T_d die Menge der Terme im Dokument $d \in D_i$. Zudem sei $y_{id} \in \{1, ..., C\}$.

Mooney, Roy (2000) verwenden zwei disjunkte Klassen $\bar{z} \in \{1, 0\}$ und führen für jede der beiden Klassen Gewichte ein:

$$\mathcal{G}_1(i, d) = \frac{y_{id} - 1}{C - 1} \quad \text{und} \quad \mathcal{G}_0(i, d) = 1 - \mathcal{G}_1(i, d).$$

$\mathcal{G}_1(i, d)$ ist groß, wenn der Nutzer i den Text $d \in D_i$ positiv bewertet. Die Klasse $\bar{z} = 1$ entspricht daher der Klasse der positiv bewerteten Dokumente.

$$\mathcal{E}_i(\bar{z}) = \frac{1}{|D_i|} \sum_{\underline{d} \in D_i} \mathcal{G}_{\bar{z}}(i, \underline{d}), \quad \bar{z} \in \{1, 0\}$$

ist ein heuristisches Maß dafür, wie stark der Nutzer i dazu tendiert, positiv ($\bar{z} = 1$) oder negativ ($\bar{z} = 0$) zu bewerten. Dokumente, die einen bestimmten Term t enthalten sind für einen Nutzer i dann von Interesse, wenn die Gewichte $\mathcal{G}_1(i, d)$ für die Dokumente groß sind, die diesen Term häufig enthalten. die Länge des Dokuments d sei

$$\mathcal{L}(d) = \sum_{\underline{\tau} \in T_d} a(\underline{\tau}, d).$$

Es ist zu bedenken, daß sich Terme häufiger wiederholen, je länger ein Dokument ist. Daher wird die Summe der gewichteten Anzahl $a(t, d)$ noch durch die gewichtete Summe der Text-Längen dividiert:

$$\mathcal{E}_i(t|\bar{z}) = \frac{\sum_{\underline{d} \in D_i} \mathcal{G}_{\bar{z}}(i, \underline{d}) a(t, \underline{d})}{\sum_{\underline{d} \in D_i} \mathcal{G}_{\bar{z}}(i, \underline{d}) \mathcal{L}(\underline{d})}, \bar{z} \in \{1, 0\}.$$

Sei T_d die Menge von Termen, die ein bestimmtes Dokument d enthält. Dann ist

$$\mathcal{E}_i(\bar{z}|d) \propto \mathcal{E}_i(\bar{z}) \prod_{\underline{\tau} \in T_d} \mathcal{E}_i(\underline{\tau}|\bar{z}), \bar{z} \in \{1, 0\}$$

ein heuristisches Maß dafür wie deutlich ein bestimmtes Dokument d vom Nutzer i einer Klasse $\bar{z}$ zugeordnet wird. Die Struktur dieser Heuristik erinnert an den Satz von Bayes, welcher in Kapitel 9 eingehend behandelt wird.

Auf dieser Basis können einem Nutzer $i \in \{1, ..., I\}$ die Dokumente $d \in D$ empfohlen werden, deren zugehörige Werte $\mathcal{E}_i(\bar{z}|d)$ besonders hoch sind.

Interessant ist, daß hier die untransformierten Daten benutzt werden, um das Ausmaß der Zugehörigkeit zu einer von zwei Klassen zu approximieren.

Dadurch, daß die Daten untransformiert verwendet werden, werden Verzerrungen durch die Transformation ausgeschlossen.

Wie in den beiden anderen vorgestellten kontentbasierten Verfahren werden fehlende Bewertungsdaten ignoriert.

4.4 Eigenschaften der kontentbasierten Verfahren

Nicht bei allen Items wäre denkbar, die relevanten Eigenschaften ohne großen Aufwand maschinell zu ermitteln (Shardanand, Maes (1995)). Da sich Termhäufigkeiten leicht computergestützt ermitteln lassen, ist es kein Zufall, daß sich die meisten kontentbasierten Verfahren mit der Empfehlung von Texten befassen. Bei anderen Items wie zum Beispiel Filmen oder CDs ist eine so einfache Bestimmung der relevanten Eigenschaften nicht immer möglich. Falls Texte existieren, die charakteristisch genug für die jeweiligen Items sind (wie beispielsweise die Produktbeschreibungen bei *Amazon.de*) ist es möglich, diese Texte zur Klassifikation der dazugehörigen Items zu verwenden. Ansonsten müssen die Attribute per Hand erfaßt werden. Hierbei ist zu beachten, daß der Aufwand, z.B. hinsichtlich der gesamten Produkte eines Online-Shops die Attribute zuverlässig von Mitarbeitern erfassen zu lassen, immens wäre.

Ein weiterer Kritikpunkt ist, daß fast immer nur sehr oberflächliche Eigenschaften bestimmt werden können (Balabanovic, Shoham (1997)). Im Falle der Textklassifikation führt die Beschränkung auf gewichtete Termhäufigkeiten da-

zu, daß Eigenschaften wie Qualität und Stil der Texte nicht beachtet werden. So kann beispielsweise nicht zwischen einem gut und einem schlecht geschriebenen Text unterschieden werden, falls beide dieselben Terme benutzen (Shardanand, Maes (1995)). Hieraus folgt, daß nicht immer alle wichtigen Merkmale gegeben sind oder für eine große Anzahl von Items einfach bestimmt werden können.

Insbesondere unter Marketing-Gesichtspunkten ist es zudem problematisch, daß den Nutzern nur solche Items empfohlen werden können, die Items ähneln, die sie vorher gesehen (und positiv bewertet) haben. Hierdurch werden dem Nutzer unter Umständen eine sehr homogene Menge von Items empfohlen, während es für ihn abwechselungsreicher wäre, verläßliche Empfehlungen für sehr unterschiedliche Items zu bekommen. Cross-Selling Chancen gehen auf diese Weise verloren.

Weiterhin ist es erforderlich, daß der Nutzer eine genügend große Anzahl an Items bewertet hat.

Da nur vorhandene Einträge in der Datenmatrix zur Erstellung des Nutzerprofils verwendet werden, ignorieren die Kontentbasierten Verfahren bei der Erstellung eines Profils für den i-ten Nutzer die fehlenden Werte in der i-ten Zeile der Datenmatrix. Elemente aus anderen Zeilen der Datenmatrix sind nicht erforderlich um die fehlenden Bewertungen des i-ten Nutzers zu schätzen. Da es aber meist nur möglich ist, recht oberflächliche Eigenschaften der Items quantitativ zu ermitteln, ist es durchaus möglich, daß auch wichtige Eigenschaften der Items nicht zur Verfügung stehen. Diese werden dann ebenfalls ignoriert.

Kapitel 5

Nicht-Bayes'sche Kollaborative Verfahren

Im Gegensatz zu den Kontentbasierten Verfahren benutzen die Kollaborativen Verfahren die Bewertungen anderer Nutzer um die Bewertung des i-ten Nutzers vorherzusagen. Mittlerweile wird die Gesamtheit dieser Verfahren in der Literatur als Kollaboratives Filtern bezeichnet. Früher bezog sich dieser Begriff hauptsächlich auf die einfachste und bekannteste Variante dieser Klasse von Verfahren, das Ähnlichkeitsverfahren.

5.1 Das Ähnlichkeitsverfahren

Alle der zahlreichen Varianten des Ähnlichkeitsverfahrens beruhen auf der Idee, zunächst ein quantitatives Maß für die Ähnlichkeit zwischen zwei Objekten zu berechnen, um dann für jedes einzelne dieser Objekte eine Menge möglichst ähnlicher Objekte zu bestimmen. Die Menge aller möglichst ähnlichen Objekte eines bestimmen Objekts wird als Nachbarschaft des betreffenden Objekts bezeichnet.

Je nachdem, ob als Objekte die Nutzer oder die Items verwendet werden, kann man von Nutzer-basierten beziehungsweise von Item-basierten Ähnlichkeitsverfahren sprechen.

Insbesondere die früheren Varianten des Ähnlichkeitsverfahrens verwenden Ähnlichkeitsbeziehungen zwischen den Nutzern und können daher als Nutzer-basierte Ähnlichkeitsverfahren bezeichnet werden. Diese Nutzer-basierten Methoden waren die ersten Kollaborativen Verfahren und wurden daher früher als

Kollaboratives Filtern bezeichnet. Sie basieren auf der Annahme, daß der Geschmack jedes Nutzers eine Struktur aufweist, die er mit einer Gruppe von anderen Nutzern gemeinsam hat (Shardanand, Maes (1994), Hill et. al. (1995)). Hat man für einen Nutzer eine Gruppe von ähnlichen Personen identifiziert, so sollte man ihm Items vorschlagen, die in dieser Gruppe (seiner Nachbarschaft) beliebt sind. Hierbei sollten jeder Nutzer aus der Nachbarschaft ein umso höheres Gewicht haben, je ähnlicher er dem betrachteten („aktiven") Nutzer ist.

Hinsichtlich der Nutzer-basierten Ähnlichkeitsverfahren ist der Bravais-Pearson Korrelationskoeffizient nach Matthai $r_{i\iota}^{Matthai}$ für Elemente der ersten Modalität (siehe Abschnitt 3.2.1)

$$r_{i\iota}^{Matthai} = \frac{\sum\limits_{j' \in J_{i\iota}} (y_{ij'} - \bar{y}_{i.}^{i\iota})(y_{\iota j'} - \bar{y}_{\iota.}^{\iota i})}{\sqrt{\sum\limits_{j' \in J_{i\iota}} (y_{ij'} - \bar{y}_{i.}^{i\iota})^2} \sqrt{\sum\limits_{j' \in J_{i\iota}} (y_{\iota j'} - \bar{y}_{\iota.}^{\iota i})^2}},$$

mit $J_{i\iota} = J_i \cap J_\iota$ und $\bar{y}_{i.}^{i\iota} = 1/|J_{i\iota}| \sum_{j' \in J_{i\iota}} y_{ij'}$ das meistverwendete Ähnlichkeitsmaß.

In der Literatur wird auch häufig die sogenannte Vektor-Ähnlichkeit $\mathcal{VS}_{i\iota}$ („vector similarity") genannt. Bei der Vektor-Ähnlichkeit wird für jeden Benutzer $i, \iota \in I$ ein J-komponentiger Vektor $\tilde{y}^i = (\tilde{y}_{i1}, ..., \tilde{y}_{iJ})'$ mit Komponenten

$$\tilde{y}_{ij} = \begin{cases} y_{ij}, & \text{falls } j \in J_i \\ & \forall i \in \{1, ..., I\}, j \in \{1, ..., J\} \\ 0, & \text{sonst} \end{cases}$$

erstellt. Die Vektor-Ähnlichkeit zwischen zwei Nutzern i und ι ist dann einfach der Kosinus des von den beiden Vektoren $\tilde{y}^i$ und $\tilde{y}^\iota$ eingeschlossenen Winkels:

$$\mathcal{VS}_{i\iota} = \frac{\tilde{y}^i \cdot \tilde{y}^\iota}{\|\tilde{y}^i\| \, \|\tilde{y}^\iota\|}.$$

Breese et. al. (1998) haben die mit diesen beiden beiden Ähnlichkeitsmaßen berechneten Ergebnisse miteinander verglichen. Bei diesem Vergleich führte der Bravais-Pearson Korrelationskoeffizient nach Matthai $r_{i\iota}^{Matthai}$ zu den besseren Ergebnissen. Die meisten Verfahren verwenden daher heute Korrelationen als Ähnlichkeitsmaße.

Da die Bewertungsdaten ordinales Skalenniveau haben, wäre es ratsam, auch Korrelationskoeffizienten zu untersuchen, die für ordinale Daten geeignet sind (siehe Anhang F). Herlocker et. al. (1999) haben den Bravais-Pearson'schen Korrelationskoeffizient nach Matthai mit seinem ranggeordneten Äquivalent, dem Spearman'schen Korrelationskoeffizient nach Matthai, empirisch verglichen. Beide Korrelationen führen in sehr guter Näherung zu denselben Ergebnissen (Herlocker et. al. (1999)). Es wäre naheliegend, noch weitere Korrelationsmaße mit der Bravais-Pearson'schen Korrelation zu vergleichen. Insbesondere wäre es interessant, Somers D zu verwenden.

Zur Bestimmung der Nachbarschaft des Nutzers i existieren unterschiedliche Ansätze. Der früheste Ansatz geht auf Shardanand, Maes (1995) zurück. Shardanand, Maes (1995) zufolge bilden für jeden Nutzer i alle anderen Personen $\iota \neq i$, deren Korrelation mit i größer als ein bestimmter positiver Schwellenwert ist, dessen Nachbarschaft. Infolgedessen erhält diese Version der Nachbarschaft nur mit dem betrachtenen Nutzer i positiv korrelierte Nutzer und wird daher als $\mathcal{IN}^+(i)$ bezeichnet. Häufiger wird jedoch eine Nachbarschaft $\mathcal{IN}(i)$ verwendet, die alle Nutzer $\iota \neq i$ enthält, deren Korrelation mit i betragsmäßig größer als ein bestimmter Schwellenwert ist. (Auf diese Weise gelingt es, auch Information zu benutzen, die von mit i stark antikorrelierten Nutzern stammt.) Hinsichtlich beider Ansätze ist problematisch, daß für manche Nutzer nur kleine (absolute) Korrelationen existieren. Um für diese Nutzer Nachbarschaften zu bestimmen muß ein niedriger Schwellenwert gewählt werden. Das hat wiederum zur Folge, daß die Nachbarschaften für Nutzer, deren Korrelationen größer ausfallen, sehr groß und damit ineffektiv werden.

Deshalb bestimmen Herlocker et. al. (1999) die Nachbarschaft jedes Nutzers i indem sie die $\mathcal{N}$ Nutzer $\iota \neq i$ bestimmen, deren absolute Korrelation mit i am größten ausfällt. Dieses Verfahren wird auch als $\mathcal{N}$-Nachbarn Ansatz bezeichnet und führt im Vergleich zu dem Schwellenwertansatz von Shardanand, Maes (1995) zu deutlich verbesserten Ergebnissen.

Sarwar et. al. (2000a) kommen ebenfalls zu dem Schluß, daß das $\mathcal{N}$-Nachbarn Verfahren alternativen Ansätzen überlegen ist.

Weiterhin legen die Ergebnisse von Herlocker et. al. (1999) nahe, für die von ihnen analysierten Filmbewertungsdaten $\mathcal{N}$ aus dem Bereich zwischen 10 und 80 zu wählen. Für $\mathcal{N} < 50$ erhält man im Mittel geringfügig bessere Ergebnisse. Dafür sind die Ergebnisse für $\mathcal{N} > 50$ deutlich stabiler. Diese Ergebnisse sind je-

doch abhängig von Struktur und Größe des verwendeten Datensatzes. Je größer die Anzahl der betrachteten Nutzer ist, umso größer ist im allgemeinen der Anteil der Nutzer, für die zumindest eine kleine Gruppe sehr ähnlicher Nutzer identifiziert werden kann. Daher ist anzunehmen, daß die optimale Wahl für $\mathcal{N}$ mit steigender Nutzerzahl fällt.

Zur Bestimmung der Schätzer finden sich in der Literatur unterschiedliche Ansätze. Shardanand, Maes (1995) verwenden für die Schätzung der Bewertung y_{ij} die folgende Gewichtung der Bewertungen für j:

$$\hat{Y}_{ij} = \frac{\sum_{\iota \in \mathcal{IN}_j^+(i)} r_{i\iota}^{Matthai} y_{\iota j}}{\sum_{\iota \in \mathcal{IN}_j^+(i)} |r_{i\iota}^{Matthai}|}.$$

Dabei bezeichnet $\mathcal{IN}_j^+(i) \subseteq \mathcal{IN}^+(i)$ die Menge der Nutzer aus der Nachbarschaft $\mathcal{IN}^+(i)$ von i, die Item j bewertet haben. Bei diesem Verfahren muß die Nachbarschaft $\mathcal{IN}^+(i)$ so bestimmt werden, daß sie möglichst stark positiv korrelierte Werte enthält. Zudem verwenden Shardanand, Maes (1995) eine Modifikation des Bravais-Pearson Korrelationskoeffizienten nach Matthai. Shardanand, Maes (1995) verwenden statt der Mittelwerte in $r_{i\iota}^{Matthai}$ den konstanten Wert 4. Diese Modifikation konnte sich jedoch nicht durchsetzen.

Diese Gewichtung der Bewertungen durch Nutzer aus der Menge $\mathcal{IN}_j^+(i)$ berücksichtigt nicht die Unterschiede im Bewertungsverhalten der Nutzer. Einige Nutzer tendieren dazu, hohe Bewertungen anzugeben, andere beschränken sich auf das untere Ende der Bewertungsskala. Auf diese Weise kann beispielsweise die Bewertung eines Items auf einer 5-Punkte Skala mit 4 für zwei verschiedene Benutzer unterschiedliche Bedeutung haben. Während es vernünftig wäre, anzunehmen, daß einem Nutzer, dessen durchschnittliche Filmbewertung bei 2,5 liegt, ein Film besonders gefallen hat, wenn er ihn mit 4 bewertet hat, ist diese Annahme bei einem Nutzer mit einer Durchschnittsbewertung von 4,5 Punkten weniger plausibel. Außerdem beschränkt man sich hierbei unnötigerweise auf die Information, die durch Betrachtung positiv korrelierter Nutzer gewonnen werden kann. Deshalb wird vorwiegend (Resnick et. al. (1994), Konstan et. al. (1997), Breese et. al. (1998)) der folgende Ansatz verwendet:

$$\hat{Y}_{ij} = \bar{y}_{i\cdot} + \frac{\sum_{\iota \in \mathcal{IN}_j(i)} r_{i\iota}^{Matthai} (y_{\iota j} - \bar{y}_{\iota\cdot})}{\sum_{\iota \in \mathcal{IN}_j(i)} |r_{i\iota}^{Matthai}|}.$$

Dabei ist die Nachbarschaft $\mathcal{IN}(i)$ nicht auf positiv korrelierte Nutzer beschränkt. $\mathcal{IN}_j(i) \subseteq \mathcal{IN}(i)$ ist die Menge der Nutzer aus der Nachbarschaft $\mathcal{IN}(i)$ von i, die Item j bewertet haben. Die Durchschnittsbewertung $\bar{Y}_{i\cdot}$ berücksichtigt das Bewertungsverhalten des i-ten Benutzers. Falls ein Nutzer $\iota \in \mathcal{IN}_j(i)$ positiv (negativ) mit i korreliert ist, wird durch seine Bewertung für Item j $y_{\iota j}$ genau dann ein Betrag zu hinzuaddiert (subtrahiert), wenn die Differenz $y_{\iota j} - \bar{y}_{\iota\cdot} > 0$ ($y_{\iota j} - \bar{y}_{\iota\cdot} < 0$) ist. Dabei ist dieser Betrag umso größer, je größer die Beträge der entsprechenden Korrelation und der jeweiligen Differenz ausfallen.

Dieses Verfahren ist eine lineare speicherbasierte Heuristik. Die ordinalen Bewertungsdaten werden hierbei behandelt als hätten sie kardinales Skalenniveau. Hieran würde sich auch dann nichts ändern, wenn man anstelle der Bravais-Pearson Korrelation den Spearman'schen Rangkorrelationskoeffizient oder auch Somers D benutzen würde, da auch dann noch die Differenzen $y_{\iota j} - \bar{y}_{\iota\cdot}$ verwendet würden.

Beispiel 5.1:

Es geht wieder um die Filme $j = 1, ..., 8$ aus Beispiel 4.1. Zusätzlich zu den Bewertungen der ersten Person sind Bewertungen von fünf weiteren Nutzern gegeben:

y_{ij}	Nutzer	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$i=1$	Bernd	1	5	4	4	3	4	-	-
$i=2$	Karin	-	4	2	2	2	5	-	3
$i=3$	Lars	-	4	4	4	3	-	5	1
$i=4$	Nadine	5	2	2	-	2	3	-	4
$i=5$	Viktor	4	-	2	3	4	4	-	4
$i=6$	Volker	1	5	4	4	4	-	4	-

Tabelle 5.1: Datenmatrix (Beispiel 5.1)

Es geht nun darum, den Schätzer $\hat{Y}_{17}$ zu berechnen. Da Lars und Volker die einzigen sind, die Bewertungen für Barry Lyndon ($j = 7$) abgegeben haben, reicht es aus, die Bravais-Pearson Korrelationen mit dem Verfahren nach Matthai $r_{13}^{Matthai}$ und $r_{16}^{Matthai}$ zu berechnen.

Mit $\bar{y}_{1.}^{13} = 4$ und $\bar{y}_{3.}^{31} = 3,75$ ergibt sich

$$r_{13}^{Matthai} = \frac{1\frac{1}{4} + 0 + 0 + (-1)\left(-\frac{3}{4}\right)}{\sqrt{1+1}\sqrt{\left(\frac{1}{4}\right)^2 3 + \left(\frac{3}{4}\right)^2}} = 0,816.$$

Wegen der niedrigen Anzahl an Nutzers ist es nicht zweckmäßig eine Nachbarschaft zu berechnen. Mit $r_{16}^{Matthai} = 0,902$ und $\bar{y}_{1.} = 3,5$ erhält man

$$\hat{Y}_{17} = 3,5 + \frac{0,816(5-3,5) + 0,902(4-3,5)}{|0,816| + |0,902|} = 4,47.$$

In vielen marketingrelevanten Anwendungen ist die Anzahl der Items sehr groß. So umfaßt beispielsweise das Sortiment des Online-Retailers Amazon.com mehrere Millionen Produkte (Linden et. al. (2003)). Daher kann jeder einzelne Nutzer dieser Systeme nur einen sehr geringen Teil der zur Verfügung stehenden Items kennen. Da es mit Zeitaufwand und etwas Mühe verbunden ist, Items zu bewerten, wird im allgemeinen die Anzahl der von einem spezifischen Nutzer bewerteten Items sogar noch deutlich geringer sein. Somit ist die Schnittmenge der gemeinsam bewerteten Items $J_i \cap J_\iota$ in vielen Fällen sehr klein. In zahlreichen Anwendungen existieren für zwei verschiedene Nutzer oft nur drei bis fünf gemeinsam bewertete Items (Herlocker et. al. (1999)). Hierunter leidet die Genauigkeit der Schätzer und damit die Verläßlichkeit der Empfehlungen. Trotzdem können zwischen fast allen Nutzern Korrelationen ausgerechnet werden, da eine Menge von sehr bekannten und häufig bewerteten Items existiert aus der mindestens ein gemeinsames Item von fast jedem Nutzerpaar bewertet wurde (Karypis (2000)). Da die Anzahl der Nutzer I sehr hoch ist, ist der mit der Berechnung von fast $I(I-1)/2$ Korrelationen verbundene Rechenaufwand immens. (Amazon.com hatte z.B. bereits im Jahr 2003 $I = 2,9 \cdot 10^7$ Kunden (Linden et. al. (2003)).)

Zu diesem Problem existieren in der Literatur verschiedene Lösungsansätze. Breese et. al. (1998) schlagen vor, anstelle des Durchschnitts $J_i \cap J_\iota$ die Vereinigung $J_i \cup J_\iota$ zu verwenden und alle fehlenden Werte durch einen konstanten Wert d_B zu ersetzen. Hierbei sollte d_B so gewählt werden, daß es einer neutralen oder geringfügig negativen Bewertung entspricht. Diese Imputationsstrategie kann zu starken Verzerrungen führen. Sie konnte sich weder in Literatur noch in

der Praxis durchsetzen.

Herlocker et. al. (1999) schlagen vor, Korrelationen, die auf weniger als 50 gemeinsamen Bewertungen existieren, schwächer zu gewichten. Hierzu wird jede Korrelation mit dem Gewichtungsfaktor

$$G(i,\iota) = \begin{cases} \dfrac{|J_i \cap J_\iota|}{50}, & \text{für } |J_i \cap J_\iota| < 50 \\ 1, & \text{sonst} \end{cases}$$

versehen. Dieses Verfahren wird als Signifikanzgewichtung bezeichnet und es wurde empirisch belegt, daß es zu etwas besseren Ergebnissen führen kann (Herlocker et. al. (1999)).

Auch wenn die Signifikanzgewichtung den Einfluß von Korrelationen, die nur auf wenigen Daten beruhen, auf die Schätzer abschwächt, ändert dies nichts daran, daß es problematisch ist, daß die meisten Korrelationen nur auf wenigen Bewertungen beruhen.

Diesem Problem wird durch die Item-basierte Variationen des Ähnlichkeitsverfahrens begegnet (Karypis (2000), Sarwar et. al. (2001)). Während das Ähnlichkeitsverfahren nach Shardanand, Maes (1995) versucht, einem Nutzer das zu empfehlen, was von möglichst ähnlichen Nutzers gut bewertet wurde, versuchen seine Item-basierten Variationen dem Nutzer Items zu empfehlen, die den Items möglichst ähnlich sind, für die er hohe Bewertungen abgegeben hat. Hierzu berechnet man die Ähnlichkeit zwischen den Items und nicht zwischen den Nutzern. Die Bravais-Pearson Korrelation mit der Modifikation nach Matthai für zwei Items j_1 und j_2 entspricht genau der Darstellung in Kapitel 3 und ist:

$$\tilde{r}^{Matthai}_{j_1 j_2} = \frac{\sum\limits_{i \in I_{j_1} \cap I_{j_2}} (y_{ij_1} - \bar{y}^{j_1 j_2}_{.j_1})(y_{ij_2} - \bar{y}^{j_2 j_1}_{.j_2})}{\sqrt{\sum\limits_{i \in I_{j_1} \cap I_{j_2}} (y_{ij_1} - \bar{y}^{j_1 j_2}_{.j_1})^2} \sqrt{\sum\limits_{i \in I_{j_1} \cap I_{j_2}} (y_{ij_2} - \bar{y}^{j_2 j_1}_{.j_2})^2}}.$$

Indem man anstelle der modifizierten Zeilenvektoren von y, $\tilde{y}^i, i = 1, ..., I$, einfach die analog modifizierten Spaltenvektoren von y, $\breve{y}^j, j = 1, ..., J$, verwendet, kann man die Vektor-Ähnlichkeit der Items j_1 und j_2 als den Kosinus des zwischen $\breve{y}^{j_1}$ und $\breve{y}^{j_2}$ eingeschlossenen Winkels berechnen:

$$\breve{\mathcal{VS}}_{j_1 j_2} = \frac{\breve{y}^{j_1} \cdot \breve{y}^{j_2}}{\|\breve{y}^{j_1}\| \, \|\breve{y}^{j_2}\|}.$$

Die auf diese Weise ermittelten Ähnlichkeiten zwischen dem betrachteten Item j_1 und und allen übrigen Items $j_2 \in \{1, ..., J\} \setminus \{j_1\}$ kann dazu benutzt werden, eine Nachbarschaft $\mathcal{JN}^+(j_1)$ für das betreffende Item j_1 zu erstellen. Hierbei werden die $\mathcal{N}$ anderen Items $j_2 \in \{1, ..., J\} \setminus \{j_1\}$ ausgewählt, deren positive Korrelation bzw. deren positive Kosinus-Ähnlichkeit mit j_1 am größten ist.
Die Schätzer werden nach der Formel

$$\hat{Y}_{ij_1} = \frac{\sum\limits_{j_2 \in \mathcal{JN}^+(j_1)} \tilde{R}_{j_1 j_2} y_{ij_2}}{\sum\limits_{j_2 \in \mathcal{JN}^+(j_1)} |\tilde{R}_{j_1 j_2}|}$$

berechnet, wobei an Stelle von $\tilde{R}_{j_1 j_2}$ entweder die Korrelation $\tilde{r}_{j_1 j_2}^{Matthai}$ auch der Kosinus des zwischen den modifizierten j_1- und j_2-ten Spaltenvektoren eingeschlossenen Winkels stehen kann. Mild, Natter (2002) konnten anhand einer umfassenden Studie empirisch belegen, daß die optimale Anzahl von Item-Nachbarn $\mathcal{N}^{\mathcal{J}}$ (genau wie die optimale Anzahl der Nutzer-Nachbarn $\mathcal{N}$) umso kleiner wird, je mehr Nutzer betrachtet werden. Steigende Nutzerzahlen verbessern die Qualität der Korrelationen und Kosinus-Ähnlichkeiten. Zudem erhöhen sich durch die steigende Nutzerzahl auch die Anzahl der Item-Paare, für die Nutzer existieren, die beide Items bewertet haben. Daher ist es bei hohen Nutzerzahlen wahrscheinlicher, zu einem Item eine kleine Menge sehr ähnlicher Items zu finden.

Die Item-basierte Version des Ähnlichkeitsverfahrens wurde mit dem klassischen Ähnlichkeitsverfahren empirisch verglichen. Karypis (2000) kann zeigen, daß die Item-basierten Ansätze bei einigen Datensätzen zu deutlich genaueren Schätzern für die Bewertungen und bei anderen zu nur geringfügigen Verschlechterungen der Genauigkeit führen. Bei den von Sarwar et. al. (2001) betrachteten Datensätzen führten beide vorgestellten Item-basierten Ähnlichkeitsverfahren zu zumindest geringfügig genaueren Schätzern als die Nutzer-basierten Verfahrensvarianten. Zudem führt hier der Kosinus-Ansatz gegenüber der Korrelation zu geringfügigen Verbesserungen der Genauigkeit.

Mit beiden Methoden bleiben Unterschiede im Bewertungsverhalten unterschiedlicher Nutzer unberücksichtigt. Deshalb schlagen Sarwar et. al. (2001) die Verwendung eines angepaßten Kosinus-Ähnlichkeitsmaßes vor, das nur Bewertungen von Nutzern verwendet, die beide betrachteten Items bewertet haben und von jeder Bewertung y_{ij} den Mittelwert der Bewertungen $\bar{y}_{i.}$ des jeweiligen Nutzers

subtrahiert:

$$c_a(j_1, j_2) = \frac{\sum\limits_{i \in I_{j_1} \cap I_{j_2}} (y_{ij_1} - \bar{y}_{i.})(y_{ij_2} - \bar{y}_{i.})}{\sqrt{\sum\limits_{i \in I_{j_1} \cap I_{j_2}} (y_{ij_1} - \bar{y}_{i.})^2} \sqrt{\sum\limits_{i \in I_{j_1} \cap I_{j_2}} (y_{ij_2} - \bar{y}_{i.})^2}}.$$

Dies kann zu einer drastischen Erhöhung der Schätzgenauigkeit führen (vgl. Sarwar et. al. (2001)). Die angepaßte Kosinus-Ähnlichkeit ist genau wie $\tilde{r}_{j_1 j_2}^{Matthai}$ und $\breve{\mathcal{VS}}_{j_1 j_2}$ nicht definiert, wenn $I_{j_1} \cap I_{j_2}$ die leere Menge ist.

Die meisten Nutzer beschränken sich auf mehrere Gruppen von Items (Karypis (2000)). Eine dieser Gruppen besteht aus den sehr bekannten und darum häufig bewerteten Items. Während jedes Paar dieser sehr bekannten Items von vielen Nutzern bewertet wurde, gibt es sehr viele Item-Paare aus anderen Gruppen, für die kein Nutzer existiert, der diese beiden Items beide bewertet hat. Im von Karypis (2000) verwendeten Datensatz existierten für über 99% der Item-Paare keine Nutzer, die Bewertungen für beide Items abgegeben hatten. Daher genügt es, die Ähnlichkeiten zwischen Item-Paaren (j_1, j_2), deren Durchschnitt $I_{j_1} \cap I_{j_2} \neq \emptyset$ ist, zu berechnen. Der hierdurch gegenüber den Nutzer-basierten Ähnlichkeitsverfahren gesparte Rechenaufwand ist sehr beachtlich und verkürzt die Rechenzeiten drastisch (Karypis (2000), Deshpande, Karypis (2004)).

Wegen der erhöhten Genauigkeit und dem drastisch reduzierten Rechenaufwand benutzt beispielsweise Amazon.com bereits Item-basierte Ähnlichkeitsverfahren (Linden et. al. (2003)).

Beispiel 5.2:

In diesem Beispiel wird wieder der Schätzer $\hat{Y}_{17}$ auf Grundlage der Datenmatrix aus Beispiel 5.1 berechnet. Dazu wird hier das angepaßte Kosinus-Ähnlichkeitsmaß $c_a(j_1, j_2)$ verwendet. Mit $I_7 \cap I_1 = \{6\}$ ergibt sich

$$c_a(7,1) = \frac{(y_{67} - \bar{y}_{6.})(y_{61} - \bar{y}_{6.})}{\sqrt{(y_{67} - \bar{y}_{6.})^2}\sqrt{(y_{61} - \bar{y}_{6.})^2}} = \frac{(4 - 3,5)(1 - 3,5)}{\sqrt{(4 - 3,5)^2}\sqrt{(1 - 3,5)^2}} = -1.$$

Wegen der niedrigen Dimensionen werden als Nachbarschaft $\mathcal{JN}^+(7)$ alle mit $j_1 = 7$ positiv korrelierten Items $j_2 \in \{1, ..., 8\} \setminus \{7\}$ herangezogen, für die mindestens ein Nutzer existiert, der sowohl Item $j_1 = 7$ als auch Item j_2 bewertet

hat. Für das Item $j_2 = 6$ existieren keine Bewertungen von Nutzern aus I_7. Für die übrigen angepaßten Kosinus-Korrelationen erhält man:

	$j_2 = 1$	$j_2 = 2$	$j_2 = 3$	$j_2 = 4$	$j_2 = 5$	$j_2 = 8$
$c_a(7, j_2)$	-1.000	0.600	0.894	0.894	-0.447	-1.000

Tabelle 5.2: Angepaßte Kosinus-Ähnlichkeiten (Beispiel 5.2)

Daher ergibt sich $\mathcal{JN}^+(7) = \{2, 3, 4\}$ und damit $\hat{Y}_{17} = 4,25$.

Stehen zusätzlich Daten über das Kaufverhalten der Nutzer zur Verfügung, so können diese ebenfalls dazu benutzt werden, um Ähnlichkeitsmaße zu berechnen. Diese Ähnlichkeitsmaße sind antisymmetrisch in j_1 und j_2.

Sei I_j^K die Menge der Nutzer, die Item j gekauft haben. Die bedingte Sample-Wahrscheinlichkeit dafür, daß j_2 gekauft wird, wenn j_1 bereits gekauft wurde ist $|I_{j_1}^K \cap I_{j_2}^K|/|I_{j_1}^K|$. Diese ist umso höher, je öfter j_2 gekauft wurde. Da es nicht darum geht, die häufigsten Items zu identifizieren, wird die bedingte Sample-Wahrscheinlichkeit noch mit einem Korrekturfaktor versehen und es ergibt sich:

$$\tilde{R}_{j_1 j_2}^{K,1} = \frac{|I_{j_1}^K \cap I_{j_2}^K|}{|I_{j_1}^K|(|I_{j_2}|)^{\alpha_0}},$$

wobei α_0 ein Parameter aus dem Intervall $[0, 1]$ ist. $\alpha_0 = 0,5$ ist ein typischer Wert. Mit $\tilde{R}_{j_1 j_2}^{K,1}$ lassen sich in der Regel etwas bessere Ergebnisse erzielen als mittels der angepaßten Kosinus-Ähnlichkeit (Deshpande, Karypis (2004)).
Alternativ kann auch das antisymmetrische Ähnlichkeitsmaß

$$\tilde{R}_{j_1 j_2}^{K,2} = \frac{\sum\limits_{i \in I_{j_2}} \tilde{y}_{ij_2}}{|I_{j_1}^K|(|I_{j_2}|)^{\alpha_0}},$$

mit

$$\tilde{y}_{ij_2} = \begin{cases} \dfrac{y_{ij_2}}{\sum\limits_{j_2' \in J_i} y_{ij_2'}}, & \text{für } v_{ij_2} = 1 \\ 0, & \text{für } v_{ij_2} = 0 \end{cases} \qquad \forall\ i \in I_{j_2}$$

verwendet werden. Mittels $\tilde{R}^{K,2}_{j_1 j_2}$ können im allgemeinen sogar noch geringfügig bessere Resultate als unter Verwendung von $\tilde{R}^{K,1}_{j_1 j_2}$ erreicht werden (Deshpande, Karypis (2004)).

5.2 Regressionsansätze

Die Regressionsansätze von Mild, Natter (2002) verwenden alle Bewertungen zu einem Item j_1 als abhängige Variable und wählen die unabhängigen Variablen aus der Menge der übrigen Items $j_2 \in \{1, ..., J\} \setminus \{j_1\}$ aus. Die Auswahl der unabhängigen Variablen erfolgt mittels einer Heuristik. Es werden zwischen dem betrachteten Item j_1 und den restlichen Items $j_2 \in \{1, ..., J\} \setminus \{j_1\}$ die Produkte $|\tilde{r}^{Matthai}_{j_1 j_2}| s_{j_2} |\hat{\beta}^{j_1}_{j_2}|$ berechnet. Hier ist $\tilde{r}^{Matthai}_{j_1 j_2}$ angepaßten Bravais-Pearson Korrelationen nach Matthai für Items, s_{j_2} ist die Standardabweichung des Items j_2

$$s_{j_2} = \sqrt{\frac{1}{|I_{j_2}| - 1} \sum_{i \in I_{j_2}} (y_{ij_2} - \bar{y}_{.j_2})^2}$$

und $\hat{\beta}^{j_1}_{j_2}$ sind die Schätzer einer vorherigen Regression mit j_1 als endogener Variable und allen verbleibenden Items $j_2 \in \{1, ..., J\} \setminus \{j_1\}$ als exogene Variablen. Im Rahmen dieser Heuristik sind Items $j_2 \in \{1, ..., J\} \setminus \{j_1\}$ als exogene Variable umso wichtiger, je größer der vorher berechnete Betrag des Schätzers $\hat{\beta}^{j_1}_{j_2}$ ausfällt, je stärker j_2 mit j_1 absolut korreliert ist und je unterschiedlicher das Item j_2 bewertet wurde. Fehlende Werte in den jeweiligen exogenen Variablen y_{ij_2} werden bei der Regression durch Item-Mittelwerte $\bar{y}_{.j_2}$ ersetzt. Dagegen werden fehlende Werte in der entsprechenden endogenen Variable y_{ij_1} nicht zur Berechnung der Schätzer verwendet.

Betrachtet wurden ein lineares Regressionsmodell und ein logistischer Regressionsansatz. Das lineare Regressionsmodell führte von allen verwendeten Regressionsansätzen zu den besten Ergebnissen, die Ergebnisse der Logistischen Regression waren erheblich schlechter (Mild, Natter (2002)).

Für große Anzahlen von Nutzern und einer im Verhältnis dazu kleinen Anzahl an Items (419) liefert die lineare Regression deutlich bessere Ergebnisse als das Item-basierte Ähnlichkeitsverfahren (Mild, Natter (2002)). Für noch kleinere

Anzahlen von Items verschlechtern sich die Ergebnisse der linearen Regression im Mittel (Mild, Natter (2002)).

Im Gegensatz zu den Ähnlichkeitsverfahren ist dieser Ansatz modellbasiert. Es werden lineare (lineare Regression) und nichtlineare Modelle (logistische Regression) verwendet. Fehlende Werte in den unabhängigen Variablen y_{ij_2} werden durch Spalten-Mittelwerte $\bar{y}_{.j_2}$ imputiert, während fehlende Werte für die jeweilige abhängige Variable y_{ij_1} ignoriert werden. Im Rahmen der linearen Regression werden die Bewertungsdaten wie Daten mit kardinalem Skalenniveau behandelt. Bei der logistischen Regression werden die Daten verwendet als hätten sie nominales Skalenniveau, wobei die Rangordnung verloren geht. Die Vernachlässigung der Rangordnung zwischen den Bewertungsdaten ist Grund für das verhältnismäßig schlechte Abschneiden der logistischen Regression.

Der vorgestellte Regressionsansatz berücksichtigt die Heterogenität der Items und vernachlässigt die Heterogenität der Nutzer, indem für jedes Item j der Schätzer $\hat{\beta}^j$ für alle Nutzer berechnet wird.

Beispiel 5.3:

In diesem Beispiel wird noch einmal der Schätzer $\hat{Y}_{17}$ auf Grundlage der Datenmatrix aus Beispiel 5.1 berechnet. Diesmal wird der lineare Regressionsansatz von Mild, Natter (2002) verwendet. Zunächst sind alle fehlenden Werte y_{ij_2} werden ihre zugehörigen Spalten-Mittelwerte $\bar{y}_{.j_2}$ zu ersetzen:

y_{ij}	Nutzer	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$i=1$	Bernd	1	5	4	4	3	4	$4,5^*$	3^*
$i=2$	Karin	$2,75^*$	4	2	2	2	5	$4,5^*$	3
$i=3$	Lars	$2,75^*$	4	4	4	3	4^*	5	1
$i=4$	Nadine	5	2	2	3.4^*	2	3	$4,5^*$	4
$i=5$	Viktor	4	4^*	2	3	4	4	$4,5^*$	4
$i=6$	Volker	1	5	4	4	4	4^*	4	3^*

Tabelle 5.3: Imputation der fehlenden Werte (Beispiel 5.3)

Die imputierten Werte wurden durch $*$-Zeichen markiert. Da nur Lars und Volker

Barry Lyndon ($j = 7$) bewertet haben, hat man nur die endogenen Größen y_{37} und y_{67} zur Verfügung, weshalb die zugehörige Design-Matrix nur zwei Zeilen hat. Daher kann man hier nur einen Regressor verwenden, da sonst $X'X$ nicht vollen Rang hat. Deshalb ist es nicht möglich, ein Modell mit allen $\{1, ..., 8\} \setminus \{7\}$ zu schätzen.

Zur Auswahl des einen Regressors empfiehlt es sich, das Produkt der Standardabweichung und des Betrages der angepaßten Kosinus-Korrelation $s_j|c_a(7, j)|$, $j \in \{1, ..., 8\} \setminus \{7\}$ zu betrachten (Tabelle 4.5).

$c_a(7, 6)$ kann nicht berechnet werden, da die einzigen beiden Personen $i = 3, 6$, die Barry Lyndon ($j = 7$) bewertet haben, beide keine Aussagen zu Dangerous Liaisons ($j = 6$) gemacht haben.

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=8$
$c_a(7,j)$	-1,000	0,600	0,894	0,894	-0,447	-	-1,000
s_j	2,061	1,225	1,095	0,894	0,894	0,816	1,414
$s_j\|c_a(7,j)\|$	2,061	0,735	0,980	0,800	-0,400	-	1,414

Tabelle 5.4: Standardabweichungen und angepaßte Kosinus-Ähnlichkeiten (Beispiel 5.3)

Da $\arg\max_{j'} s_{j'}|c_a(7, j')| = 1$ ist, empfiehlt es sich, den Film $j = 1$, Armageddon, als Regressor zu verwenden. Als zugehörige Design-Matrix X und Vektor y ergeben sich

$$X = \begin{pmatrix} 1 & 2{,}75 \\ 1 & 1 \end{pmatrix} \quad \text{und} \quad y = \begin{pmatrix} 5 \\ 4 \end{pmatrix}.$$

Damit erhält man

$$(X'X)^{-1} = \begin{pmatrix} 2{,}796 & -1{,}224 \\ -1{,}224 & 0{,}653 \end{pmatrix}, \quad X'y = \begin{pmatrix} 9 \\ 17{,}75 \end{pmatrix}$$

und somit

$$\hat{\beta}^7 = (X'X)^{-1}X'y = \begin{pmatrix} 3{,}43 \\ 0{,}57 \end{pmatrix}.$$

Als Schätzer ergibt sich $\hat{Y}_{17} = \hat{\beta}_0^7 + \hat{\beta}_1^7 y_{11} = 4,0$. Daß $\hat{\beta}_1^7$ positiv ist, obwohl $c_a(7,6)$ negativ ist, ist eine Folge der Mittelwert-Imputation. Da Lars Armageddon (Film $j = 1$) nicht bewertet hat, wird anstelle einer Bewertung einfach der Spaltenmittelwert $\hat{Y}_{\cdot 1} = 2,75$ verwendet. Somit scheint es, als habe Lars sowohl Armageddon als auch Barry Lyndon besser bewertet als Volker, weshalb $\hat{\beta}_1^7$ positiv ist.

An diesem Beispiel sieht man einige Probleme des vorgestellten Verfahrens. Offenbar kann die Mittelwert-Imputation zu Verzerrungen führen. Darüberhinaus ist dieses Verfahren ungeeignet, um Bewertungen von Items zu schätzen, hinsichtlich derer wenige oder keine Bewertungen abgegeben wurden.

5.3 Singulärwertzerlegung-basierte Verfahren

Die Motivation für die auf der Singulärwertzerlegung (SVD) basierenden Verfahren ist, durch eine Dimensionsreduktion der riesigen Datenmatrix y die generellen Strukturen der Daten zu erfassen und gleichzeitig unwichtige Information wegzulassen. Ihre gemeinsame Grundlage, die Singulärwertzerlegung, ist ein Verfahren aus der Linearen Algebra, dessen Geschichte ins 19. Jahrhundert zurückreicht (vgl. Stewart (1993)). Die Singulärwertzerlegung wird in der Literatur auch als „two-mode factor analysis“ bezeichnet (Deerwester et. al. (1990)). Klassische Verfahren zur Faktorenanalyse wie beispielsweise die Hauptkomponentenanalyse (Pearson (1901)) stellen eine gegebene Datenmatrix als Produkt zweier Matrizen, der Faktorwertematrix und der Ladungsmatrix, dar (vgl. Abschnitt 5.3.3). Durch eine Dimensionsreduktion der beiden Matrizen (Weglassen einer gleichen Anzahl hinterster Spalten), die als Vernachlässigung weniger wichtiger Faktoren interpretiert werden kann, kann eine Approximation der ursprünglichen Datenmatrix gefunden werden, die nur den wichtigsten Faktoren Rechnung trägt.

Da die Singulärwertzerlegung nur für Matritzen ohne fehlende Werte definiert ist, muß man an Stelle der Datenmatrix Y die vervollständigte $I \times J$ Datenmatrix A_Y betrachten.

Da A_Y eine reelle Matrix ist, gilt $A_Y = U\Sigma\mathcal{V}'$ mit orthogonalen Matrizen $U \in \mathbb{R}^{I,r}$, $\mathcal{V} \in \mathbb{R}^{J,r}$ und einer Diagonalmatrix $\Sigma \in \mathbb{R}^{r,r}$. Sei ohne Beschränkung der Allgemeinheit $I \geq J$.

${A_Y}'A_Y$ ist wegen $x'{A_Y}'A_Y x = (A_Y x)'A_Y x \geq 0$ positiv semidefinit, weshalb

alle Eigenwerte $\lambda_\mu, \mu = 1, ..., J$, von ${A_Y}'A_Y$ größer oder gleich Null sind. Für die zu ${A_Y}'A_Y$ gehörenden normierten Eigenvektoren $v_\mu = (v_{1\mu}, ..., v_{J\mu})'$ gilt dann ${A_Y}'A_Y v_\mu = \lambda_\mu v_\mu$ weshalb $v'_\mu {A_Y}'A_Y v_\mu = \|A_Y v_\mu\|^2 = \lambda_\mu v'_\mu v_\mu = \lambda_\mu$ und somit $\sqrt{\lambda_\mu} = \|A_Y v_\mu\|$. Zusätzlich gilt, daß wegen $A_Y {A_Y}' A_Y v_\mu = \lambda_\mu A_Y v_\mu$ der Vektor $\tilde{u}_\mu = A_Y v_\mu$ Eigenvektor der Matrix $A_Y {A_Y}'$ ist. Daher folgt für die zu den der positiven Eigenwerte $\lambda^p_\mu, \mu = 1, ..., r$, gehörenden normierten Eigenvektoren

$$u_\mu = \frac{A_Y v_\mu}{\|A_Y v_\mu\|} = \frac{A_Y v_\mu}{\sqrt{\lambda^p_\mu}} \quad \text{und somit} \quad A_Y v_\mu = \sqrt{\lambda^p_\mu} u_\mu, \quad \mu = 1, ..., r.$$

Die Anzahl der positiven Eigenwerte r ist gleich dem Rang der Matrix A_Y. (Entartete Eigenwerte, d.h. Eigenwerte zu denen mehrere Eigenvektoren existieren, werden mehrfach gezählt.) Dabei seien die die positiven Eigenwerte $\lambda^p_\mu, \mu = 1, ..., r$, nach ihrer Größe geordnet, so daß $\lambda^p_1 \geq \lambda^p_2 \geq ... \geq \lambda^p_r$ gilt. Letzteres ist eine Konvention, die sich als zweckmäßig erwiesen hat. Faßt man die Vektoren $v_\mu, \mu = 1, ..., J$, als Spalten einer Matrix $\mathcal{V}$ auf, so ergibt sich:

$$A_Y \underbrace{\begin{pmatrix} v_{11} & \cdots & v_{1r} \\ \vdots & \ddots & \vdots \\ v_{J1} & \cdots & v_{Jr} \end{pmatrix}}_{=\mathcal{V}} = (A_Y v_1 \cdots A_Y v_r) = \left(\sqrt{\lambda^p_1} u_1 \cdots \sqrt{\lambda^p_r} u_r\right)$$

$$= \underbrace{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \vdots \\ u_{I1} & \vdots & u_{Ir} \end{pmatrix}}_{=U} \underbrace{\begin{pmatrix} \sqrt{\lambda^p_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda^p_r} \end{pmatrix}}_{=\Sigma}.$$

$A'_Y A_Y$ ist symmetrisch und $v_\mu, \mu = 1, ..., r$, sind normierte Eigenvektoren zu $A'_Y A_Y$. Da die Eigenvektoren zu verschiedenen Eigenwerten orthogonal sind, sind die normierten Eigenvektoren $v_1, ..., v_r$ jedenfalls dann paarweise orthonormal, wenn die Eigenwerte nicht entartet sind. Unter dieser Voraussetzung gilt $\mathcal{V}\mathcal{V}' = E$, wobei E wie gewohnt die Einheitsmatrix bezeichnet. Wegen $\mathcal{V}\mathcal{V}' = E$ und $A_Y \mathcal{V} = U\Sigma$ folgt unter der Annahme, daß die Eigenwerte nicht entartet sind, die Beziehung $A_Y = U\Sigma\mathcal{V}'$.

Sofern $I > J$ gilt, bestimmt man die Singulärwertzerlegung, indem man die Eigenwerte λ_μ und Eigenvektoren v_μ zur Matrix $A_Y' A_Y$ bestimmt und diese normiert. (Wenn Eigenwerte entartet sind, erzeugt man vorher mittels des Gram-Schmidt Verfahrens aus den Eigenvektoren ein Orthogonalsystem.) Schließlich ordnet man alle positiven Eigenwerte ihrer Größe nach. Die orthonormierten Eigenvektoren, die zu positiven Eigenwerten λ_μ^p gehören, werden zu Spalten der Matrix $\mathcal{V}$, wobei die Eigenvektoren umso weiter links in der Matrix stehen, je größer ihr zugehöriger Eigenwert ist. Mittels der Beziehung $u_\mu = \frac{A_Y v_\mu}{\sqrt{\lambda_\mu^p}}$ kann man hieraus die Spalten von U bestimmen und es gilt $\Sigma = diag\{\sqrt{\lambda_1^p}, ..., \sqrt{\lambda_r^p}\}$.

Im Fall $J > I$ ist es zweckmäßig an Stelle der Eigenvektoren der Matrix $A_Y{}' A_Y$ die Eigenvektoren zur Matrix $A_Y A_Y{}'$ zu bestimmen und dann analog vorzugehen.

Wegen $u_\mu = A_Y \frac{v_\mu}{\sqrt{\lambda_\mu^p}}, \mu = 1, ..., r$, ist die Menge paarweise orthonormaler Vektoren $\{u_1, ..., u_r\}$ eine Orthogonalbasis des Spaltenraums von A_Y. Daher enthält jede Zeile von U alle Information über den jeweiligen Nutzer als r-dimensionalen Vektor. Weil $\{v_1, ..., v_r\}$ eine Orthogonalbasis des Zeilenraums von A_Y ist, kann jede Zeile von $\mathcal{V}$ als alle verfügbare Information hinsichtlich des zugehörigen Items in r Dimensionen aufgefaßt werden.

Die Hauptdiagonalelemente der Matrix Σ, $\sqrt{\lambda_\mu^p}$, $\mu = 1, ..., r$, werden als Singulärwerte bezeichnet. Die Singulärwerte $\sqrt{\lambda_\mu^p}, \mu = 1, ..., r$, können wegen $A_{Y,ij} = \sum_{\mu=1}^r u_{i\mu} v_{j\mu} \sqrt{\lambda_\mu^p}$ als Maß für die Stärke der Interaktion zwischen den μ-ten Spalten von U und $\mathcal{V}$ interpretiert werden. Daher kann jede Dimension $\mu \in \{1, ..., r\}$ als eine latente Größe interpretiert werden, die der Wechselwirkung zwischen Elementen der ersten Modalität und Elementen der zweiten Modalität zugrundeliegt.

Für eine gegebene Anzahl $R < r$ der ersten und daher per Konvention größten Singulärwerte definiert man $\Sigma_R = diag\{\sqrt{\lambda_1^p}, ..., \sqrt{\lambda_R^p}\}$. Die entsprechenden zu den R größten Singulärwerten gehörenden R ersten Spalten von U und $\mathcal{V}$ macht man in der gleichen Reihenfolge zu Spalten der Matrizen $U_R \in \mathbb{R}^{I,R}$ und $\mathcal{V}_R \in \mathbb{R}^{J,R}$. Hiedurch ergibt sich als Rang-R Approximation $\hat{A}_{Y,R}$ für A_Y:

$$\hat{A}_{Y,R} = U_R \Sigma_R \mathcal{V}_R'.$$

Sei $M_{\hat{A}_Y}^R = \{\hat{A}_Y | Rang(\hat{A}_Y) = R\}$ die Menge aller Approximationen für A_Y vom Rang R. Das Theorem von Eckart, Young (1936) besagt dann, daß die Rang-R Approximation $\hat{A}_{Y,R} = U_R \Sigma_R \mathcal{V}_R'$ die beste Approximationsmatrix vom Rang R

für A_Y hinsichtlich der Frobenius-Norm ist:

$$\hat{A}_{Y,R} = \arg \min_{A \in M^R_{\hat{A}_Y}} \sum_{i=1}^{I} \sum_{j=1}^{J} (A_{Y,ij} - A_{ij})^2.$$

Das Ausmaß der Dimensionsreduktion durch die Approximation wird durch die Wahl von R bestimmt. Einerseits sollte R groß genug sein, um die gesamte der Datenmatrix zugrundeliegende Struktur erfassen zu können. Andererseits darf $R < r$ auch nicht zu groß gewählt werden, da sonst unwichtige Details, die mehr mit dem verwendeten Datensatz als mit der den Daten allgemein zugrundeliegenden Struktur zu tun haben, berücksichtigt werden.

Durch die Singulärwertzerlegung wird eine Dimensionsreduktion erreicht, die es erlaubt, die allgemeine Struktur der Interaktion zwischen Nutzern und Items zu betrachten.

5.3.1 Das SVD-basierte Verfahren nach Billsus, Pazzani

Zu Beginn der Verfahrensvariante nach Billsus, Pazzani (1998) wird die Datenmatrix für jeden einzelnen Nutzer $i \in \{1, ..., I\}$ zu einer binären Datenmatrix $A_Y^{B,i} \in \mathbb{R}^{2(I-1),\alpha_i}$, mit $\alpha_i = |J_i|$, transformiert. Hierzu werden jeweils nur die Items aus J_i verwendet. Daher wird in der neuen Matrix an Stelle des entspechenden Index j aus der Datenmatrix Y der transformierte Index $\mathcal{TR}(j, J_i)$ verwendet, der jedem Wert j aus J_i seinen Rang in der Menge J_i zuordnet. Für jeden einzelnen der übrigen Nutzer $\iota \in \{1, ..., I\} \setminus \{i\}$ werden in der Matrix $A_Y^{B,i}$ zwei benachbarte Zeilen eingeführt. Weil i zuvor aus der Matrix entfernt wurde, müssen auch die Indizes ι der ursprünglichen Datenmatrix Y transformiert werden:

$$\mathcal{TO}(\iota, i) = \begin{cases} 2\iota - 1, & \text{für } i > \iota \\ 2\iota - 3, & \text{für } i < \iota \end{cases}$$

Durch diese Transformationsvorschrift erhält man den Index der ersten Zeile in $A_Y^{B,i}$, die zum Nutzer ι gehörende Werte enthält. Dabei gilt für alle Elemente der jeweils ersten Zeile $A^{B,i}_{Y,\mathcal{TO}(\iota,i)\mathcal{TR}(j,J_i)}$:

$$A^{B,i}_{Y,\mathcal{TO}(\iota,i)\mathcal{TR}(j,J_i)} = \begin{cases} 1, & \text{für } y_{\iota j} \geq \gamma_{\mathcal{S}} \\ 0, & \text{für } y_{\iota j} < \gamma_{\mathcal{S}} \\ 0, & \text{für } v_{\iota j} = 0 \end{cases}.$$

Die Elemente der ersten Zeile sind nur dann von Null verschieden, wenn die zugehörige Bewertung $Y_{\iota j}$ existiert und größer oder gleich einem Schwellenwert $\gamma_{\mathcal{S}}$ sind. Dagegen sind die Elemente der jeweils zweiten Zeile nur dann eins, falls $Y_{\iota j}$ existiert und unterhalb des Schwellenwerts $\gamma_{\mathcal{S}}$ liegen:

$$A^{B,i}_{Y,(\mathcal{TO}(\iota,i)+1)\mathcal{TR}(j,J_i)} = \begin{cases} 1, & \text{für } y_{\iota j} < \gamma_{\mathcal{S}} \\ 0, & \text{für } y_{\iota j} \geq \gamma_{\mathcal{S}} \\ 0, & \text{für } v_{\iota j} = 0 \end{cases}$$

Beispiel 5.4:

Ein Beispiel für diese Datentransformation ist die Bestimmung der Matrix $A^{B,1}_{Y}$ aus der Datenmatrix Y aus Beispiel 5.1. Als Schwellenwert wird $\gamma_{\mathcal{S}} = 4$ verwendet. Man erhält:

$A^{B,1}_{Y,xz}$	Nutzer	$z=1$	$z=2$	$z=3$	$z=4$	$z=5$	$z=6$
$x=1$	Karin	0	1	0	0	0	1
$x=2$	Karin	0	0	1	1	1	0
$x=3$	Lars	0	1	1	1	0	0
$x=4$	Lars	0	0	0	0	1	0
$x=5$	Nadine	1	0	0	0	0	0
$x=6$	Nadine	0	1	1	0	1	1
$x=7$	Viktor	1	0	0	0	1	1
$x=8$	Viktor	0	0	1	1	0	0
$x=9$	Volker	0	1	1	1	1	0
$x=10$	Volker	1	0	0	0	0	0

Tabelle 5.5: Datentransformation nach Billsus, Pazzani (1998) (Beispiel 5.4)

Nachdem $A^{B,i}_{Y}$ berechnet ist, wird die Singulärwertzerlegung von $A^{B,i}_{Y}$ berechnet. Die auf diese Weise erhaltenen Matrizen $U^{B,i} \in \mathbb{R}^{2(I-1),\alpha_i}$, $\mathcal{V}^{B,i} \in \mathbb{R}^{\alpha_i,r}$ und $\Sigma^{B,i} \in \mathbb{R}^{r,r}$ (mit $r = \text{Rang}(A^{B,i}_{Y})$) werden verwendet, um die Matrizen der entsprechenden Rang-R Approximation $U^{B,i}_{R} \in \mathbb{R}^{2(I-1),R}$, $\mathcal{V}^{B,i}_{R} \in \mathbb{R}^{\alpha_i,R}$ und $\Sigma^{B,i}_{R} \in \mathbb{R}^{R,R}$

zu berechnen. Dabei ist R für alle $i \in \{1, ..., I\}$ vorab fest gewählt. Billsus, Pazzani (1998) wählen $R = 0,9 \cdot r$. Dabei gilt hier für den Rang $r \leq \min\{2(I-1), \alpha_i\}$. Daher sind die Ränge r und R von i abhängig. Um die Notation nicht weiter zu komplizieren, werden trotzdem weiter r und R anstelle von $r(i)$ und $R(i)$ verwendet.

Durch $\mathcal{V}_R^{B,i} = \hat{A}_{Y,R}^{B,i'} U_R^{B,i} (\Sigma_R^{B,i})^{-1}$ wird die jeweils j-te Spalte der Rang-R Approximation $\hat{A}_Y^{B,i}$ mittels $U_R^{B,i}$ und $(\Sigma_R^{B,i})^{-1}$ auf die j-te Zeile von $\mathcal{V}_R^{B,i}$ abgebildet. Letztere kann als dimensionsreduzierte Beschreibung des j-ten Items interpretiert werden.

Billsus, Pazzani (1998) repräsentieren jedes Item $j \in \{1, ..., \alpha_i\}$ auf Basis der jeweiligen Spalte $a_j^{B,i}$ von $A_Y^{B,i} = (a_1^{B,i}, ..., a_{\alpha_i}^{B,i})$ durch $\hat{\nu}_{R,j}^{B,i'} = a_j^{B,i'} U_R^{B,i} (\Sigma_R^{B,i})^{-1}$.

Die entsprechende Rang-R Darstellung aller Items $j \in \{1, ..., J\}$ kann auf der Grundlage der analog zu $A_Y^{B,i}$ zu bildenden Matrix $A_Y^{U,i}$ erstellt werden. Es gelte für die jeweils erste zu einem Nutzer $\iota \in \{1, ..., I\} \setminus \{i\}$ gehörige Zeile

$$A_{Y,\mathcal{TO}(\iota,i)j}^{U,i} = \begin{cases} 1, & \text{für } y_{\iota j} \geq \gamma_{\mathcal{S}} \\ 0, & \text{für } y_{\iota j} < \gamma_{\mathcal{S}} \\ 0, & \text{für } v_{\iota j} = 0 \end{cases}, \quad \forall\ j \in \{1, ..., J\}.$$

In Bezug auf die zweite Zeile zum Nutzer $\iota \in \{1, ..., I\} \setminus \{i\}$ gilt

$$A_{Y,(\mathcal{TO}(\iota,i)+1)j}^{U,i} = \begin{cases} 1, & \text{für } y_{\iota j} < \gamma_{\mathcal{S}} \\ 0, & \text{für } y_{\iota j} \geq \gamma_{\mathcal{S}} \\ 0, & \text{für } v_{\iota j} = 0 \end{cases}, \quad \forall\ j \in \{1, ..., J\}.$$

Daher repräsentieren Billsus, Pazzani (1998) jedes Item $j \in \{1, ..., J\}$ auf Basis der Spalten der eben definierten Matrix $A_Y^{U,i} = (a_1^{U,i}, ..., a_J^{U,i})$ durch

$$\hat{\nu}_{R,j}^{i'} = a_j^{U,i'} U_R^{B,i} (\Sigma_R^{B,i})^{-1}.$$

Zu jedem Element $\hat{\nu}_{R,\mathcal{TR}(j,J_i)}^{B,i}$ gehört eine Bewertung des i-ten Nutzers $y_{ij}, j \in J_i$. Diese Bewertungen werden wieder nach dem Schema

$$T_{\bar{z}}(y_{ij}) = \begin{cases} 1 & , \text{ falls } y_{ij} \geq \gamma_{\mathcal{S}} \wedge \bar{z} = 1 \\ 1 & , \text{ falls } y_{ij} < \gamma_{\mathcal{S}} \wedge \bar{z} = 0 \\ 0 & , \text{ sonst} \end{cases}$$

transformiert. Billsus, Pazzani (1998) benutzen für jeden Nutzer $i \in \{1, ..., I\}$, der eine ausreichende Anzahl von Items bewertet hat, $\hat{\nu}^{B,i}_{R,\mathcal{TR}(j,J_i)}$ und $T_{\bar{z}}(y_{ij})$, mit $\bar{z} = 0, 1, j \in J_i$, um ein mehrschichtiges Neuronales Netzwerk mit Rückpropagation zu trainieren. Die so erhaltenen Gewichte $\alpha_{0m}, \alpha_m, m = 1, ..., M$, und $\beta_{0,\bar{z}}, \beta_{\bar{z}}, \bar{z} = 0, 1$, können dann dazu verwendet werden, für die Repräsentation $\hat{\nu}^{i'}_{R,j} = a^{U,i'}_j U^{B,i}_R (\sigma^{B,i}_R)^{-1}$ eines von i nicht bewerteten Items $j \in \{1, ..., J\} \setminus \{J_i\}$ die Wahrscheinlichkeit der Zugehörigkeit zur Klasse $\bar{z} = 1$ zu schätzen.

Beispiel 5.5:

In diesem Beispiel werden die Wahrscheinlichkeiten der Zugehörigkeit zur Klasse $\bar{z} = 1$ (also für eine Bewertung größer oder gleich $\gamma_{\mathcal{S}} = 4$) für die Filme Barry Lyndon ($j = 7$) und Twister ($j = 8$) in Bezug auf den ersten Nutzer (Bernd), $g_1(\mathbf{T}^1_{17})$ und $g_1(\mathbf{T}^1_{18})$, mit dem Singulärwertzerlegung-basierten Ansatz von Billsus, Pazzani (1998) geschätzt. Dazu ist zunächst die Singulärwertzerlegung für die transformierte Datenmatrix $A^{B,1}_Y$ aus Beispiel 5.4 zu berechnen.
Im ersten Schritt müssen die Eigenwerte der Matrix

$$A^{B,1'}_Y A^{B,1}_Y = \begin{pmatrix} 3 & 0 & 0 & 0 & 1 & 1 \\ 0 & 4 & 3 & 2 & 2 & 2 \\ 0 & 3 & 5 & 4 & 3 & 1 \\ 0 & 2 & 4 & 4 & 2 & 0 \\ 1 & 2 & 3 & 2 & 5 & 2 \\ 1 & 2 & 1 & 0 & 2 & 3 \end{pmatrix}$$

mittels der Formel

$$\det\left(A^{B,1'}_Y A^{B,1}_Y - \lambda E\right) = 0$$

berechnet werden. (det bezeichnet die Determinante und E ist die Einheitsmatrix.) Auf diese Weise ergeben sich die Eigenwerte

$$\lambda_1 = 13,321, \;\; \lambda_2 = 4,844, \;\; \lambda_3 = 2,917, \;\; \lambda_4 = 2,028, \;\; \lambda_5 = 0,619, \;\; \lambda_6 = 0,271.$$

Da alle Eigenwerte positiv sind, hat $A_Y^{B,1'} A_Y^{B,1}$ vollen Rang ($r = 6$). Die zugehörigen Eigenvektoren $v_1, ..., v_6$ können mit Hilfe der Formel

$$\left(A_Y^{B,1'} A_Y^{B,1} - \lambda_\mu E \right) v_\mu = 0, \ \mu = 1, ..., 6,$$

bestimmt werden. Die zugehörigen Eigenvektoren bilden die Spalten der Matrix

$$\mathcal{V}^{B,1} = \begin{pmatrix} -0,070 & 0,503 & -0,539 & -0,644 & -1,161 & -0,102 \\ -0,433 & 0,059 & 0,618 & -0,283 & -0,589 & 0,033 \\ -0,570 & -0,305 & -0,071 & -0,089 & 0,318 & -0,684 \\ -0,441 & -0,446 & -0,274 & -0,214 & 0,134 & 0,684 \\ -0,481 & 0,363 & -0,329 & 0,671 & -0,274 & 0,059 \\ -0,239 & 0,565 & 0,373 & -0,045 & 0,658 & 0,221 \end{pmatrix}.$$

Aus den Eigenwerten ergibt sich die Diagonalmatrix $\Sigma = diag\{\sqrt{\lambda_1}, ..., \sqrt{\lambda_6}\}$. Wegen $U^{B,1} = A_Y^{B,1} \mathcal{V}^{B,1} (\Sigma^{B,1})^{-1}$ ergibt sich daraus

$$U^{B,1} = \begin{pmatrix} -0,184 & 0,283 & 0,581 & -0,230 & 0,089 & 0,488 \\ -0,409 & -0,176 & -0,394 & 0,259 & 0,227 & 0,113 \\ -0,395 & -0,315 & 0,160 & -0,411 & -0,174 & 0,065 \\ -0,132 & 0,165 & -0,193 & 0,471 & -0,348 & 0,113 \\ -0,019 & 0,229 & -0,316 & -0,452 & -0,205 & -0,197 \\ -0,472 & 0,310 & 0,347 & 0,179 & 0,145 & -0,713 \\ -0,216 & 0,650 & -0,289 & -0,013 & 0,284 & 0,340 \\ -0,277 & -0,341 & -0,202 & -0,212 & 0,575 & 0,001 \\ -0,527 & -0,149 & -0,032 & 0,060 & -0,522 & 0,177 \\ -0,019 & 0,229 & -0,316 & -0,452 & -0,205 & -0,197 \end{pmatrix}.$$

Die Regel $R = 0,9 \cdot r = 5,4 \rightarrow 5$ erscheint wegen der niedrigen Dimensionen hier nicht zweckmäßig. Deshalb wird statt dessen $R = 2$ gewählt.

So ergeben sich die Matrizen der Rang-R Approximation

$$\Sigma_2^{B,1} = \begin{pmatrix} 3,650 & 0 \\ 0 & 2,201 \end{pmatrix},$$

$$\mathcal{V}_2^{B,1'} = \begin{pmatrix} \text{-0,070} & \text{-0,433} & \text{-0,570} & \text{-0,441} & \text{-0,481} & \text{-0,239} \\ \text{0,503} & \text{0,059} & \text{-0,305} & \text{-0,446} & \text{0,363} & \text{0,565} \end{pmatrix}$$

und die Matrix $U_2^{B,1} \in \mathbb{R}^{10,2}$, die aus den ersten beiden Spalten von $U^{B,1}$ besteht. Man erhält daher

$$U_2^{B,1}(\Sigma_2^{B,1})^{-1} = \begin{pmatrix} \text{-0,050} & \text{0,129} \\ \text{-0,112} & \text{-0,080} \\ \text{-0,108} & \text{-0,143} \\ \text{-0,036} & \text{0,075} \\ \text{-0,005} & \text{0,104} \\ \text{-0,129} & \text{0,141} \\ \text{-0,059} & \text{0,296} \\ \text{-0,076} & \text{-0,155} \\ \text{-0,144} & \text{-0,068} \\ \text{-0,005} & \text{0,104} \end{pmatrix}$$

und kann damit $\hat{\nu}_{2,j}^{B,1'} = a_j^{B,1'} U_2^{B,1}(\Sigma_2^{B,1})^{-1}, j = 1, ..., 6$, ausrechnen. Es ergeben sich

$$\hat{\nu}_{2,1}^{B,1} = \begin{pmatrix} \text{-0,070} \\ \text{0,503} \end{pmatrix}, \ \hat{\nu}_{2,2}^{B,1} = \begin{pmatrix} \text{-0,433} \\ \text{0,059} \end{pmatrix}, \ \hat{\nu}_{2,3}^{B,1} = \begin{pmatrix} \text{-0,570} \\ \text{-0,305} \end{pmatrix},$$

$$\hat{\nu}_{2,4}^{B,1} = \begin{pmatrix} \text{-0,441} \\ \text{-0.446} \end{pmatrix}, \ \hat{\nu}_{2,5}^{B,1} = \begin{pmatrix} \text{-0,481} \\ \text{0,363} \end{pmatrix}, \ \hat{\nu}_{2,6}^{B,1} = \begin{pmatrix} \text{-0,239} \\ \text{0,565} \end{pmatrix}.$$

Zusammen mit

$$T_1(y_{11}) = T_1(y_{15}) = 0, \ \ T_1(y_{12}) = T_1(y_{13}) = T_1(y_{14}) = T_1(y_{16}) = 1$$

und $M = 2$ erhält man als Parameter des Mehrschichtigen Neuronalen Netzwerks mit Rückpropagation

$$\alpha_{01}^1 = 0,139, \ \ \alpha_{02}^1 = -0,460, \ \ \alpha_1^1 = \begin{pmatrix} \text{-0,135} \\ \text{-5,506} \end{pmatrix}, \ \ \alpha_2^1 = \begin{pmatrix} \text{-0,041} \\ \text{-4,166} \end{pmatrix},$$

$$\beta_{10}^1 = -0,753 \quad \text{und} \quad \beta_1^1 = \begin{pmatrix} 3{,}233 \\ 2{,}178 \end{pmatrix}.$$

Wegen

$$a_7^{U,1'} = (0,0,1,0,0,0,0,0,1,0)$$
$$a_8^{U,1'} = (0,1,0,1,1,0,1,0,0,0)$$

folgen die Rang-R Repräsentationen

$$\hat{\nu}_{27}^{U,1} = \begin{pmatrix} \text{-}0{,}252 \\ \text{-}0{,}211 \end{pmatrix} \quad \text{und} \quad \hat{\nu}_{28}^{U,1} = \begin{pmatrix} \text{-}0{,}212 \\ 0{,}395 \end{pmatrix}$$

Zusammen mit den Parametern des Mehrschichtigen Neuronalen Netzwerks mit Rückpropagation lassen sich hieraus die Wahrscheinlichkeiten für die Zugehörigkeit zur Klasse $\bar{z} = 1$ der mindestens mit $\gamma_{\mathcal{R}} = 4$ bewerteten Items in Bezug auf Bernd für die Filme Barry Lyndon und Twister wie in Beispiel 4.1 zu berechnen. Man erhält $g_1(\mathbf{T}_{17}^1) = 0,958$ und $g_1(\mathbf{T}_{18}^1) = 0,467$. Das Verfahren nach Billsus, Pazzani (1998) ergibt, daß der Film Barry Lyndon Bernd empfohlen werden sollte.

Billsus, Pazzani (1998) belegen empirisch, daß ihr Singulärwertzerlegung-basiertes Verfahren dem (Nutzer-basierten) Ähnlichkeitsverfahren hinsichtlich der Klassifikationsgenauigkeit deutlich überlegen ist. Gleichzeitig konzedieren die Autoren den immensen Rechenaufwand des von ihnen vorgeschlagenen Verfahrens. Zu diesem numerischen Aufwand trägt auch bei, daß für jeden einzelnen Nutzer i eine Matrix $A_Y^{B,i}$ erstellt wird. Überdies ist es vor allem vor dem Hintergrund der typischen Größe der Benutzeranzahl I von online-Recommender-Systemen im Hinblick auf die Laufzeit nachteilig, daß diese Matrizen alle $2(I-1)$ Zeilen haben. Durch die Erstellung eines individuellen Modells für jeden Nutzer wird der Heterogenität der Nutzer Rechnung getragen. Die Besonderheiten der Items werden durch die Rang-R Projektion erfaßt.

Ebenso wie die Ähnlichkeitsverfahren und die Regressionsansätze ist das Verfahren nach Billsus, Pazzani (1998) nicht für selten bewertete Items geeignet.

Durch die Transformation der Daten geht Information verloren. Dieser Verlust an Information kann bei Bewertungsdaten zu Verzerrungen führen. Diese

Methode kann auf kardinale, ordinale und binäre nominale Daten angewendet werden. Durch die Verwendung von zwei Zeilen für jeden einzelnen Benutzer ist es gelungen, eine Darstellung zu finden, die es erlaubt, die fehlenden Daten zu ignorieren.

Nach der Datentransformation und der anschließenden Dimensionsreduktion durch die Rang-R Approximation der Singulärwertzerlegung wird ein Künstliches Neuronales Netz benutzt, um die Klassenzugehörigkeit der Items zu schätzen. Somit ist dieses Verfahren modellbasiert.

Die Dimensionsreduktion wird hier nur benutzt, um eine Rang-R Approximation der Itemvektoren zu finden.

5.3.2 Das SVD-basierte Verfahren nach Sarwar et. al.

Sarwar et. al. (2000b) erstellen eine einzige Matrix $A_Y \in \mathbb{R}^{I,J}$ für alle Nutzer. Dazu werden zunächst alle in der Ausgangsdatenmatrix nicht vorhandenen Daten durch die zugehörigen Spaltenmittelwerte $\bar{y}_{.j}$ ersetzt. Damit verwendet das Verfahren dieselbe Methode zur Imputation fehlender Daten wie beim Regressionsansatz von Mild, Natter (2002). Auf diese Weise entsteht die Matrix $\tilde{A}$:

$$\tilde{A}_{ij} = \begin{cases} y_{ij}, & \text{für } v_{ij} = 1 \\ \bar{y}_{.j}, & \text{für } v_{ij} = 0 \end{cases}, \ \forall i \in \{1, ..., I\}, \forall j \in \{1, ..., J\}.$$

Die Matrix A_Y entsteht aus der Matrix $\tilde{A}$ durch

$$A_{Y,ij} = \tilde{A}_{ij} - \bar{y}_{i.}, \ \forall i \in \{1, ..., I\}, \forall j \in \{1, ..., J\}.$$

Berechnet wird dann die Singulärwertzerlegung $A_Y = U\Sigma\mathcal{V}'$ von A_Y und daraus nach Wahl von R die Näherung $\hat{A}_{Y,R} = U_R\Sigma_R\mathcal{V}'_R$. Mit

$$\tilde{\Sigma}_R = \begin{pmatrix} (\lambda_1^p)^{1/4} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (\lambda_R^p)^{1/4} \end{pmatrix}$$

kann man nun die Matrizen $U_R\tilde{\Sigma}_R \in \mathbb{R}^{I,R}$ und $\tilde{\Sigma}_R\mathcal{V}'_R \in \mathbb{R}^{R,J}$ benutzen, um y_{ij} zu

schätzen. Sei $\Psi(i) \in \mathbb{R}^{R,1}$ die i-te Spalte von $(U_R\tilde{\Sigma}_R)'$ und $\psi(j) \in \mathbb{R}^{R,1}$ die j-te Spalte von $\tilde{\Sigma}_R\mathcal{V}_R'$. Dann berechnet sich der Schätzer im Ansatz von Sarwar et. al. (2000b) gemäß

$$\hat{Y}_{ij} = \bar{y}_{i.} + \Psi(i)'\psi(j), \;\; i \in \{1, ..., I\}, \;\; j \in \{1, ..., J\}.$$

Dieser Ansatz berücksichtigt das individuelle Bewertungsverhalten der Nutzer, indem zuerst von jedem Element jeder Zeile in der vervollständigten Matrix $\tilde{A}$ der Mittelwert der Bewertungen des entsprechenden Nutzers subtrahiert wird und man dann zu den auf diese Weise um die Mittelwerte bereinigten Interaktionsschätzern $\Psi(i)'\psi(j)$ den Bewertungsmittelwert des betrachteten Nutzers hinzuaddiert.

Beispiel 5.6:

Das Verfahren von Sarwar et. al. (2000b) gestattet die Berechnung von $\hat{Y}_{17}$. Dazu muß zunächst die Matrix $\tilde{A}$ bestimmt werden. Man kann hier auf die vervollständigte Matrix aus Beispiel 5.3 zurückgreifen. Durch Subtraktion von $\bar{Y}_{i.}$ von $\tilde{A}_{ij}$ erhält man A_Y.

$$A_Y = \begin{pmatrix} \text{-2,50} & \text{1,50} & \text{0,50} & \text{0,50} & \text{-0,50} & \text{0,50} & \text{1,00} & \text{-0,50} \\ \text{-0,25} & \text{1,00} & \text{-1,00} & \text{-1,00} & \text{-1,00} & \text{2,00} & \text{1,50} & \text{0,00} \\ \text{-0,75} & \text{0,50} & \text{0,50} & \text{0,50} & \text{-0,50} & \text{0,50} & \text{1,50} & \text{-2,50} \\ \text{2,00} & \text{-1,00} & \text{-1,00} & \text{0.40} & \text{-1,00} & \text{0,00} & \text{1,50} & \text{1,00} \\ \text{0,50} & \text{0,50} & \text{-1,50} & \text{-0,50} & \text{0,50} & \text{0,50} & \text{1,00} & \text{0,50} \\ \text{-2,67} & \text{1,33} & \text{0,33} & \text{0,33} & \text{0,33} & \text{0,33} & \text{0,33} & \text{-0,67} \end{pmatrix}$$

Durch Lösen von $\det(A_Y A_Y' - \lambda E) = 0$ erhält man die Eigenwerte

$$\lambda_1^p = 29,865, \; \lambda_2^p = 15,855, \; \lambda_3^p = 6,489, \; \lambda_4^p = 2,517, \; \lambda_5^p = 1,645, \; \lambda_6^p = 0,053.$$

Mittels $(A_Y A_Y' - \lambda_\mu E)\, u_\mu = 0, \;\; \mu = 1, ..., 6$, lassen sich die zugehörigen Eigenvektoren und damit $U = (u_1, ..., u_6)$ bestimmen.

Es ergibt sich:

$$U = \begin{pmatrix} 0{,}571 & \text{-}0{,}096 & \text{-}0{,}097 & 0{,}555 & \text{-}0{,}108 & 0{,}580 \\ 0{,}190 & \text{-}0{,}713 & \text{-}0{,}271 & \text{-}0{,}326 & \text{-}0{,}508 & \text{-}0{,}134 \\ 0{,}411 & \text{-}0{,}218 & 0{,}808 & \text{-}0{,}293 & 0{,}213 & 0{,}015 \\ \text{-}0{,}398 & \text{-}0{,}504 & 0{,}296 & 0{,}656 & 0{,}049 & \text{-}0{,}260 \\ \text{-}0{,}073 & \text{-}0{,}415 & \text{-}0{,}347 & \text{-}0{,}166 & 0{,}782 & 0{,}250 \\ 0{,}553 & 0{,}089 & \text{-}0{,}239 & 0{,}207 & 0{,}266 & \text{-}0{,}718 \end{pmatrix}.$$

Daraus berechnet man

$$\mathcal{V} = A_Y' U \Sigma^{-1} = \begin{pmatrix} \text{-}0{,}748 & \text{-}0{,}219 & 0{,}299 & \text{-}0{,}257 & 0{,}015 & 0{,}396 \\ 0{,}430 & \text{-}0{,}138 & \text{-}0{,}314 & \text{-}0{,}066 & 0{,}103 & 0{,}743 \\ 0{,}181 & 0{,}430 & 0{,}303 & 0{,}075 & \text{-}0{,}448 & 0{,}335 \\ 0{,}066 & 0{,}149 & 0{,}329 & 0{,}549 & 0{,}216 & \text{-}0{,}161 \\ \text{-}0{,}024 & 0{,}300 & \text{-}0{,}249 & \text{-}0{,}299 & 0{,}691 & \text{-}0{,}078 \\ 0{,}186 & \text{-}0{,}442 & \text{-}0{,}172 & \text{-}0{,}338 & \text{-}0{,}377 & \text{-}0{,}363 \\ 0{,}181 & \text{-}0{,}662 & 0{,}285 & 0{,}323 & 0{,}308 & 0{,}098 \\ \text{-}0{,}387 & \text{-}0{,}045 & \text{-}0{,}663 & 0{,}561 & \text{-}0{,}168 & 0{,}065 \end{pmatrix}.$$

Mit der Wahl $R = 3$ folgt

$$\tilde{\Sigma}_3 = \begin{pmatrix} 2{,}338 & 0 & 0 \\ 0 & 1{,}995 & 0 \\ 0 & 0 & 1{,}596 \end{pmatrix}$$

und damit

$$\Psi(1) = \begin{pmatrix} 1{,}334 \\ \text{-}0{,}191 \\ \text{-}0{,}155 \end{pmatrix} \quad \text{und} \quad \psi(7) = \begin{pmatrix} 0{,}422 \\ \text{-}1{,}320 \\ 0{,}455 \end{pmatrix} \quad \left(\psi(8) = \begin{pmatrix} \text{-}0{,}905 \\ \text{-}0{,}089 \\ \text{-}1{,}058 \end{pmatrix} \right).$$

Insgesamt folgt

$$\begin{aligned} \hat{Y}_{17} &= \bar{y}_{1\cdot} + \Psi(1)'\psi(7) \\ &= 3,5 + ((1,334 \cdot 0,422) + (-0,191 \cdot (-1,320)) + (-0,155 \cdot 0,455)) \\ &= 4,25 \end{aligned}$$

für Barry Lyndon (und $\hat{y}_{18} = 2,47$ für Twister). Wenn nun Bewertungen von 4 größer wieder als Hinweis auf eine positive Einstellung hinsichtlich des betrachteten Items aufgefaßt werden und Bewertungen unter 4 als negative Bewertungen interpetiert werden, sollte Bernd Barry Lyndon aber nicht Twister empfohlen werden. Damit kommt der SVD-basierte Ansatz von Sarwar et. al. (2000b) zum selben Ergebnis wie das Verfahren nach Billsus, Pazzani (1998).

Das Verfahren nach Sarwar et. al. (2000b) ist ein heuristikbasierter linearer Ansatz. Da Differenzen und Summen gebildet werden, werden die ordinalen Bewertungsdaten wie Daten mit kardinalem Skalenniveau behandelt. Die fehlenden Daten werden durch Item-Mittelwerte ersetzt. Die Daten werden untransformiert verwendet. (Die Subtraktion der jeweiligen Nutzer-Mittelwerte vor der Singulärwertzerlegung kann als Bestandteil der Heuristik aufgefaßt werden.)

Sarwar et. al. (2000b) haben empirisch belegt, daß dieser SVD-basierte Ansatz bei eher kleinen prozentualen Anteilen des Trainingsdatensatzes an der gesamten verwendeten Datenmenge zu deutlich besseren Resultaten führen kann als das Nutzer-basierte Ähnlichkeitsverfahren. Bereits bei einem mittleren prozentualen Anteil des Trainingsdatensatzes an der Gesamtdatenmenge führt das Nutzerbasierte Ähnlichkeitsverfahren zu geringfügig besseren Ergebnissen. (Es wurden nur Ergebnisse für Trainingsdatensätze, die weniger als 50 % der gesamten Daten enthalten, angegeben. Sarwar et. al. betrachten eigentlich das Größenverhältnis zwischen Trainings- und Testdatensatz.) Gute Resultate bei verhältnismäßig kleinen Trainingsmengen sind ein Hinweis auf die gute Eignung eines Verfahrens zur Approximation von Datenmatrizen mit hohem Fehlendanteil. Da die Fehlendanteile der marketingrelevanten Datensätze in der Praxis sehr hoch sein können, ist das bessere Abschneiden bei verhältnismäßig hohem Fehlendanteil ein wichtiger Vorteil. Bei ungefähr demselben prozentualen Anteilen des Trainingsdatensatzes an der gesamten verwendeten Datenmenge, bei dem das SVD-basierte Verfahren nach Sarwar et. al. (2000b) dem Nutzer-basierten Verfahren am deutlichsten überlegen ist, sind die Unterschiede zwischen Item-basiertem und Nutzerbasiertem Ähnlichkeitsverfahren nicht groß (Sarwar et. al. (2001)). Dies kann als Hinweis dafür interpretiert werden, daß das SVD-basierte Verfahren auch dem Item-basierten Ähnlichkeitsverfahren bei vergleichsweise hohen Fehlendanteilen überlegen sein könnte. Da die Ergebnisse jedoch von Datensatz zu Datensatz variieren, müßte eigentlich ein direkter Vergleich zwischen Item-basiertem Ähnlich-

keitsverfahren und dem SVD-basierten Ansatz auf der Grundlage verschiedener Datensätze angestellt werden. Deshalb werden in Abschnitt 5.10 der SVD-basierte Ansatz nach Sarwar et. al. (2000b) und das Item-basierte Ähnlichkeitsverfahren für hohe Fehlendanteile empirisch miteinander verglichen.

5.3.3 Die Hauptkomponentenmethode

Die Hauptkomponentenmethode nach Goldberg et. al. (2001) bedient sich der bereits erwähnten Hauptkomponentenanalyse. Die Autoren verwenden eine Eichmenge von Items, die von jedem Nutzer bewertet werden müssen. Es wird in diesem Ansatz davon ausgegangen, daß jeder Nutzer jedes Item aus der Eichmenge G bewertet hat. Statt der Datenmatrix $y \in \mathbb{R}^{I,J}$ wird nun die stark reduzierte Datenmatrix $y_G \in \mathbb{R}^{I,|G|}$ betrachtet, die nur Bewertungen für Items aus G aber dafür auch keine fehlenden Werte enthält. Aus dieser Datenmatrix y_G wird dann die spaltenzentrierte Matrix $A^Z \in \mathbb{R}^{I,|G|}$ erzeugt, deren Komponenten durch

$$A^Z_{ij} = \frac{y_{G,ij} - \bar{y}_{G,.j}}{\sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(y_{G,ij} - \bar{y}_{G,.j})^2}}, \text{ mit } \bar{y}_{G,.j} = \frac{1}{I}\sum_{i=1}^{I} y_{G,ij},$$

$\forall i \in \{1, ..., I\}, \forall j \in G$, gegeben sind. Goldberg et. al. (2001) verwenden die Pearson'sche Kovarianzmatrix (Pearson (1901))

$$C_P(A^Z) = \frac{1}{I-1} A^{Z'} A^Z$$

um die Hauptkomponentenzerlegung $A^Z = X^Z_A F'_A$ zu bestimmen. F_A heißt Ladungsmatrix und besteht aus orthonormierten Eigenvektoren von $C_P(A^Z)$. Jedem dieser Eigenvektoren entspricht ein sogenannter Faktor. X^Z_A wird als Faktorwertematrix bezeichnet. F_A diagonalisiert $C_P(A^Z)$:

$$F'_A C_P(A^Z) F_A = C_P(X^Z_A).$$

$C_P(X^Z_A)$ ist eine Diagonalmatrix, auf deren Hauptdiagonale die Eigenwerte der Matrix $C_P(A^Z)$ stehen. Weil (wegen $A^Z = U^Z \Sigma^Z \mathcal{V}^Z$) $A^{Z'} A^Z = \mathcal{V}^Z \Sigma^Z \Sigma^Z \mathcal{V}^{Z'}$ durch $\mathcal{V}^Z$ diagonaliert wird, gilt $F_A = \mathcal{V}^Z$. Außerdem ist die Diagonalmatrix

$\Sigma^Z\Sigma^Z$ proportional zu $C_P(X_A^Z)$. X_A^Z ergibt sich durch $X_A^Z = A^Z F_A = A^Z\mathcal{V}^Z$. Daher kann von der Singulärwertzerlegung von A^Z auf F_A und X_A^Z geschlossen werden.

Die Eigenwerte von $C_P(A^Z)$ sind wegen $F_A' C_P(A^Z)F_A = C_P(X_A^Z)$ die mit den jeweiligen Faktoren verbundenen Varianzen nach der Drehung von A^Z durch $\mathcal{V}^Z = F_A$. Die Bedeutung eines Faktors kann somit anhand der Größe des zugehörigen Eigenwerts bestimmt werden. Wieder werden nur die R größten Eigenwerte betrachtet. Die Projektion einer Zeile $y_{G,i}' = (y_{G,i1}, ..., y_{G,i|G|})$ von y_G auf den von den Spalten der Matrix $\mathcal{V}_R^Z = (v_{R,1}^Z, ..., v_{R,R}^Z)$ aufgespannten Raum ist

$$Proj_{\mathcal{V}_R^Z}(y_{G,i}) = \sum_{\tilde{r}=1}^{R} \frac{(Y_{G,i}' v_{R,\tilde{r}}^Z)}{(v_{R,\tilde{r}}^{Z'} v_{R,\tilde{r}}^Z)} v_{R,\tilde{r}}^Z = \sum_{\tilde{r}=1}^{R} (y_{G,i}' v_{R,\tilde{r}}^Z) v_{R,\tilde{r}}^Z.$$

Daher enthält die i-te Zeile von $y_G\mathcal{V}_R^Z$ die Koordinaten des i-ten Nutzers in dem von den Spalten von $\mathcal{V}_R^Z = (v_{R,1}^Z, ..., v_{R,R}^Z)$ aufgespannten Raum. Goldberg et. al. (2001) wählen $R = 2$ und nennen den von den zugehörigen Eigenvektoren aufgespannten Raum die Eigenebene. Auf diese Weise werden die Bewertungen jedes Nutzers für Items aus der Eichmenge auf die Eigenebene projiziert. Die Nutzer werden dann anhand ihrer Position in der Eigenebene klassifiziert. Hierzu bestimmt man vorab die minimale rechteckige Zelle, die alle Projektionen auf die Eigenebene enthält. Die zugehörige Heuristik basiert auf der Beobachtung, daß die Dichte der „user projections“ zum Mittelpunkt der Zelle hin zunimmt. Im ersten Schritt wird die Eigenebene durch Schnitte parallel zur Ordinate und Abszisse der Eigenebene in 4 Zellen mit näherungsweise gleichvielen Elementen aufgeteilt. Bis man die gewünschte Anzahl von Clustern erhalten hat, teilt man jede neue Unterzelle, die direkt am Mittelpunkt liegt, in vier weitere Unterzellen auf, deren Begrenzungslinien wieder parallel zu Ordinate und Abszisse verlaufen. Nachdem alle Nutzer klassifiziert sind, wird die Klassifikation

$$\bar{p}_{i\bar{k}} = \begin{cases} 1, & \text{falls } i \text{ als dem } \bar{k}\text{-ten Cluster ist} \\ 0, & \text{sonst} \end{cases}$$

dazu verwendet, die Bewertungen für neue Items zu schätzen. In Bezug auf jedes Cluster $\bar{k} \in \{1, ..., \bar{K}\}$ berechnet man für jedes einzelne Item $j \in \{1, ..., J\}$ die mittlere Bewertung, die von Nutzern aus dem Cluster $\bar{k}$ hinsichtlich des jeweiligen Items j abgegeben wurde. Bezüglich eines bestimmten Nutzers i berechnet man

auf dieser Grundlage den jeweiligen Schätzer, indem man hinsichtlich jedes Items j den Mittelwert aller Nutzer verwendet, die im selben Cluster wie i sind:

$$\hat{Y}_{ij} = \begin{cases} \dfrac{\sum_{\bar{k}'=1}^{\bar{K}} \sum_{x \in I_j} \delta(\bar{p}_{i\bar{k}'}, \bar{p}_{x\bar{k}'}) y_{xj}}{\sum_{\bar{k}'=1}^{\bar{K}} \sum_{x \in I_j} \delta(\bar{p}_{i\bar{k}'}, \bar{p}_{x\bar{k}'})}, & \text{falls } \sum_{\bar{k}'=1}^{\bar{K}} \sum_{x \in I_j} \delta(\bar{p}_{i\bar{k}'}, \bar{p}_{x\bar{k}'}) > 0 \\ \bar{y}_{.j}, & \text{sonst} \end{cases}$$

mit

$$\delta(\bar{p}_{i\bar{k}'}, \bar{p}_{x\bar{k}'}) = \begin{cases} 1, & \text{für } \bar{p}_{i\bar{k}'} = \bar{p}_{x\bar{k}'} = 1 \\ 0, & \text{sonst} \end{cases} .$$

Die Wahl von $R = 2$ ist willkürlich und klein. Für größere R könnte es zu Problemen führen, daß man in jedem Schritt 2^R neue Unterzellen erhält. Außerem ist fraglich, ob bei größeren R die Punktdichte immer noch eine so einfache Struktur aufweist. Es kann daher sein, daß dieses Verfahren nur für kleine Werte von R funktioniert. Durch die sehr kleine Wahl von R wird man auf eine grobe Beschreibung der Nutzer festgelegt.

Außerdem ist problematisch, wie man eine Menge von Items bestimmt, für die jeder Nutzer Bewertungen abgeben kann. In der Praxis gibt es zu fast jedem noch so bekannten Item Personen, die sich mit diesem Item nicht auskennen. Manche Nutzer könnten abgeschreckt werden, wenn man von ihnen vorab die Bewertung einer Liste von Items verlangen würde. Insbesondere eine lange Liste könnte bei vielen Nutzern dazu führen, daß sie sich von dem System abwenden. Hieraus folgt, daß die Eichmenge klein sein sollte. Daher müßten die Elemente des sogenannten Eichsets zum einen allgemein bekannt und zum anderen besonders informativ sein. Diese Anforderungen werden in der Praxis schwer zu erfüllen sein. Nur unter der Voraussetzung, daß sich eine solche Menge hinsichtlich aller am Recommender-System (potentiell) Interessierten bestimmen läßt, kann die Hauptkomponentenmethode zur Prognose von Bewertungsdaten herangezogen werden.

Die Beschränkung auf hinsichtlich der Items aus der Eichmenge G abgegebenen Bewertungen in Bezug auf die Bestimmung der Klassenzugehörigkeiten im

Rahmen des Hauptkomponentenverfahrens erinnert konzeptionell an die Eliminierungsverfahren (Abschnitt 3.2.1). Beide Ansätze berücksichtigen absichtlich ganze Spalten (und Zeilen), in denen Einträge fehlen, nicht, um die Bestimmung auf der Basis einer zwar erheblich verkleinerten aber dafür vollständigen Datenmenge durchführen zu können. Die in Abschnitt 3.2.1 vorgestellten Eliminierungsverfahren schließen üblicherweise nach der Erhebung unvollständige Spalten (bzw. Zeilen) von der Berechnung aus. Sofern die Daten nicht die MCAR-Eigenschaft aufweisen, kann dies erhebliche Verzerrungen nach sich ziehen. Diese Gefahr wird im Rahmen der Hauptkomponentenanalyse dadurch umgangen, indem a priori eine Menge von Items bestimmt wird, die von allen Nutzern zu bewerten sind. (Allerdings ist dieses Vorgehen mit erheblichen Schwierigkeiten verbunden. In diesem Zusammenhang ist problematisch, daß dieses Vorgehen in der Praxis dazu führen kann, daß am Recommender-System interessierte Personen im Anfangsstadium abgeschreckt oder sogar abgewiesen werden.)

Die anhand der Position im Eigenraum bestimmten Cluster können als eine moderne Version der Zellen („adjustment cells“) von Oh, Scheuren (1983) und Little (1986) betrachtet werden. Die Verwendung von Klassenmittelwerten als Schätzer für die fehlenden Werte ist äquivalent zur Imputation des Zellen-Mittelwerts bei Little (1986). Unter der in der Praxis wohl kaum erfüllbaren Bedingung, daß sich eine Menge allgemein bekannter Items bestimmen läßt, die von jedem am Recommender-System Interessierten sowohl problemlos bewertet werden kann als auch bereitwillig bewertet werden und die zudem auch aussagekräftig genug zu einer Klassifikation der Nutzer anhand ihres Geschmacks ist, könnten so (unverzerrte) Gruppen ähnlicher Personen identifiziert werden. Es ist jedoch im Hinblick auf von Recommender-Systemen erhobenen Bewertungsdaten i.d.R. nicht davon auszugehen, daß innerhalb dieser Nutzer-Klassen die MAR-Annahme erfüllt ist. (Auch sehr ähnliche Nutzer beschäftigen sich bevorzugt mit Items, von denen sie sich versprechen, daß sie ihnen gefallen könnten. Daher ist auch innerhalb dieser Nutzer-Klassen das Fehlen einer Bewertung eines Items durch einen Nutzer ein Indiz dafür, daß dem betreffenden Nutzer das jeweilige Item weniger gut gefallen könnte.)

Ein Nachteil dieses heuristikbasierten, linearen Verfahrens ist, daß nur Bewertungen für Items aus der Eichmenge zur Klassifikation der Personen benutzt werden. In Bezug auf die unter Marketing-Gesichtspunkten interessanten Anwendungen im Recommender-Bereich hat dieses Vorgehen typischerweise zur Folge,

daß der größte Teil der verfügbaren Daten nicht bei der Bestimmung der Klassenzugehörigkeiten berücksichtigt wird.

Beispiel 5.7:

$\hat{Y}_{17}$ soll mit Hilfe der Hauptkomponentenmethode bestimmt werden. Die Eichmenge besteht aus den Filmen $j = 1, ..., 6$. Innerhalb dieses Beispiels nehmen wir ausnahmsweise an, daß uns die Bewertungen für alle Filme aus der Eichmenge gegeben sind:

y_{ij}	Nutzer	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$i=1$	Bernd	1	5	4	4	3	4	-	-
$i=2$	Karin	2	4	2	2	2	5	-	3
$i=3$	Lars	2	4	4	4	3	4	5	1
$i=4$	Nadine	5	2	2	2	2	3	-	4
$i=5$	Viktor	4	2	2	3	4	4	-	4
$i=6$	Volker	1	5	4	4	4	4	4	-

Tabelle 5.6: Vervollständigte Datenmatrix y_G ($G = \{1, ..., 6\}$ wurde als Eichmenge verwendet)

$y_G \in \mathbb{R}^{6,6}$ enthält alle Nutzer und die Filme $j \in \{1, ..., 6\}$ aus der Eichmenge G. Als z-transformierte Matrix erhält man

$$A^Z = \begin{pmatrix} \text{-0,913} & \text{0,976} & \text{0,913} & \text{0,848} & \text{0,000} & \text{0,000} \\ \text{-0,304} & \text{0,244} & \text{-0,913} & \text{-1,187} & \text{-1,118} & \text{1,581} \\ \text{-0,304} & \text{0,244} & \text{0,913} & \text{0,848} & \text{0,000} & \text{0,000} \\ \text{1,521} & \text{-1,220} & \text{-0,913} & \text{-1,187} & \text{-1,118} & \text{-1,581} \\ \text{0,913} & \text{-1,220} & \text{-0,913} & \text{-0,170} & \text{1,118} & \text{0,000} \\ \text{-0,913} & \text{0,976} & \text{0,913} & \text{0,848} & \text{1,118} & \text{0,000} \end{pmatrix}.$$

Mit Hilfe der Formel

$$\det\left(A^{Z'}A^Z - \lambda E\right) = 0$$

ergeben sich die Eigenwerte

$$\lambda_1 = 18,608,\ \lambda_2 = 7,141,\ \lambda_3 = 3,749,\ \lambda_4 = 0,471,\ \lambda_5 = 0,030,\ \lambda_6 = 0,000.$$

Wegen

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6} = 0,86$$

werden in diesem niedrigdimensionalen Beispiel durch die Wahl $R = 2$ keine wesentlichen Faktoren weggelassen. Es reicht aus, die Eigenvektoren zu den beiden größten Eigenwerten zu bestimmen:

$$\left(A^{Z'}A^{Z} - \lambda_\mu E\right) v_\mu = 0,\ \ \mu = 1, 2.$$

Damit ergibt sich

$$y_G\mathcal{V}_2 = \begin{pmatrix} \text{-7,140} & \text{0,079} \\ \text{-4,230} & \text{1,798} \\ \text{-6,183} & \text{-0,472} \\ \text{-1,462} & \text{-0,932} \\ \text{-3,121} & \text{-1,288} \\ \text{-7,404} & \text{-0,399} \end{pmatrix}.$$

Die Zeilen dieser Matrix können als Koordinaten in der Eigenebene aufgefaßt werden.

Man erhält auf diese Weise die in Abbildung 5.1 wiedergegebene Darstellung der Nutzer anhand ihrer Projektionen auf die Eigenebene. Die vertikale und horizontale Line dient der Unterteilung der Nutzer in Klassen.

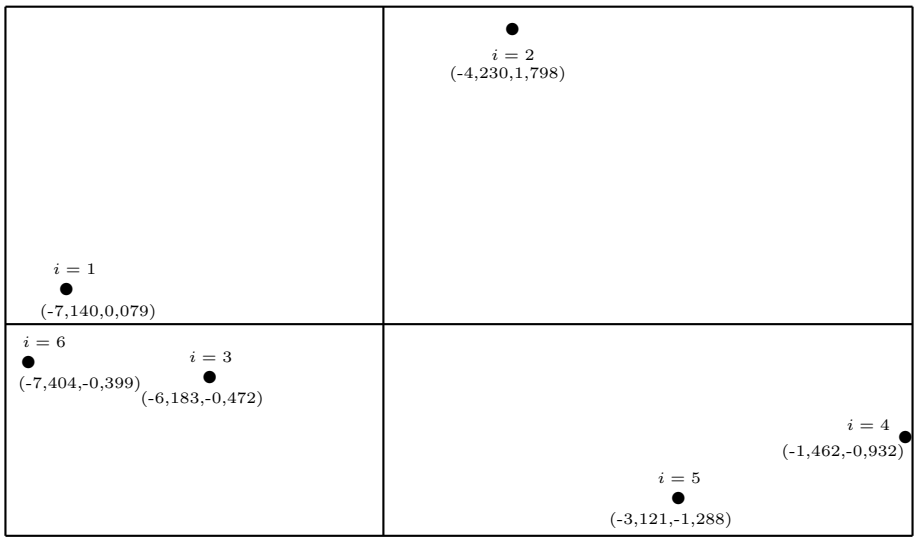

Abbildung 5.1: Eigenebene (Beispiel 5.7)

In diesem Beispiel würde ein Vorgehen nach dem Verfahren von Goldberg et. al. (2001) zu einelementigen Klassen führen. Um dies zu vermeiden wird hier der horizontale Schnitt ignoriert. So erhält man zwei verschiedene Klassen. $i = 1, 3$ und 6 gehören zur ersten Klasse und alle übrigen Nutzer sind Elemente der zweiten Klasse:

$$(\bar{p}_{i\bar{k}}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Damit kann das Ergebnis bestimmt werden:

$$\hat{Y}_{17} = \frac{5+4}{1+1} = 4,5.$$

Die Daten werden zwar nicht transformiert, aber je nach Größe der Eigenwerte kann die Projektion auf die Eigenebene zu erheblichen Informationsverlusten führen. Da Differenzen gebildet werden, werden die Bewertungsdaten auch hier wie kardinale Daten benutzt.

Insbesondere das praktische Problem, das dadurch entsteht, daß man den Nutzern vorab eine Liste von Bewertungen für vorgegebene Items abverlangt, spricht gegen dieses Verfahren, da es immer möglich ist, daß einer Person, die das System in Zukunkt nutzen möchte, ein oder mehrere Items aus der Menge G nicht bekannt sind. Das Verfahren schneidet im Hinblick auf die Genauigkeit der geschätzten Werte für die fehlenden Daten ähnlich wie das Nutzer-basierte Ähnlichkeitsverfahren ab.

5.3.4 Das EM-SVD Verfahren

Das EM-SVD Verfahren nach Srebro, Jaakkola (2003) kann als eine nicht auf der Bayes'schen Statistik basierende Version der im Abschnitt 3.2.3.2 dargestellten Data Augmentation Methode aufgefaßt werden. Wie beim Data Augmentation Verfahren wird ein konzeptionell an den EM-Algorithmus (siehe Abschnitt 3.2.3.1) erinnerndes Verfahren verwendet, bei dem jede Iteration aus zwei Schritten besteht. Zu Beginn des Verfahrens werden Anfangswerte für alle fehlenden Daten gewählt. In der jeweils n-ten Iteration werden zunächst alle in der Datenmatrix fehlenden Werte durch die in der $(n-1)$-ten Iteration bestimmten Schätzer ersetzt (erster Schritt). Auf dieser Datenbasis berechnet man die Singulärwertzerlegung. Deren Rang-$R(n)$ Approximation ist der Schätzer in der n-ten Iteration (zweiter Schritt).

Zuerst verwendet man die Startwerte $n = 2, \hat{Y}^0_{R(0)} = O_E^{I,J}, \hat{Y}^1_{R(1)} = O_N^{I,J}$ und $R(0) = R(1) = r$. Dabei bezeichnet $O_E^{I,J} \in \mathbb{R}^{I,J}$ eine Matrix, deren Elemente alle 1 sind. Alle Elemente von $O_N^{I,J} \in \mathbb{R}^{I,J}$ gleich einem konstanten Wert, der einer neutralen Bewertung entspricht und ungleich 1 ist. Im ersten Schritt werden die Matrizen v und $\bar{v}$, für deren Elemente für alle $i \in \{1, ..., I\}$ und $j \in \{1, ..., J\}$ gilt $v_{ij} = 1 - \bar{v}_{ij}$, dazu benutzt, die fehlenden Werte in y durch die Schätzer der $(n-1)$-ten Iteration $\hat{Y}^{n-1}_{R(n-1)}$ zu ersetzen. Dabei ist $\hat{Y}^{n-1}_{R(n-1)}$ die Rang-$R(n-1)$ Approximation von y in $(n-1)$-ten Schritt. Das Symbol $\odot$ steht für elementweise Multiplikation:

$$v \odot y = \begin{pmatrix} v_{11}y_{11} & \cdots & v_{1J}y_{1J} \\ \vdots & \ddots & \vdots \\ v_{I1}y_{I1} & \cdots & v_{IJ}y_{ij} \end{pmatrix}.$$

Startwerte: $n = 2, \hat{Y}^0_{R(0)} = O_E^{I,J} \in \mathbb{R}^{I,J}, \hat{Y}^1_{R(1)} = O_N^{I,J} \in \mathbb{R}^{I,J}, R(0) = R(1) = r$

Solange$\Big(\exists\ i \in \{1, ..., I\},\ j \in \{1, ..., J\} : \left\|\hat{Y}^{n-1}_{R(n-1),ij} - \hat{Y}^{n-2}_{R(n-2),ij}\right\| > \epsilon_e\Big)$ {

1. Schritt: $\hat{Y}^n = v \odot y + \bar{v} \odot \hat{Y}^{n-1}_{R(n-1)}$

2. Schritt: Singulärwertzerlegung von $\hat{Y}^n$ $(\hat{Y}^n = U^n \Sigma^n \mathcal{V}^{n\prime})$
Falls$(R(n-1) > R_{min})\,\{R(n) = R(n-1) - 1\}$
$U^n_{R(n)} \leftarrow U^n$, $\Sigma^{n-1}_{R(n)} \leftarrow \Sigma^n$, $\mathcal{V}^n_{R(n)} \leftarrow \mathcal{V}^n$
$\hat{Y}^n_{R(n)} = U^n_{R(n)} \Sigma^n_{R(n)} \mathcal{V}^n_{R(n)}$

$n \leftarrow n + 1$
}

Abbildung 5.2: EM-SVD Algorithmus

Im darauffolgenden Schritt wird die Singulärwertzerlegung für die auf diese Weise bestimmte Matrix $\hat{Y}^n$ berechnet. Es ergeben sich die Matrizen U^n, Σ^n und $\mathcal{V}^n$, die zu den Matrizen $U^n_{R(n)}, \Sigma^n_{R(n)}$ und $\mathcal{V}^n_{R(n)}$ gemacht werden müssen. Dabei besteht $U^n_{R(n)}$ aus den ersten $R(n)$ Spalten von U^n und $\mathcal{V}^n_{R(n)}$ aus den ersten $R(n)$ Spalten von $\mathcal{V}^n$. $\Sigma^n_{R(n)}$ ist eine Diagonalmatrix, auf deren Hauptdiagonale die ersten $R(n)$ Hauptdiagonalelemente von Σ^n stehen. Solange $R(n-1)$ größer ist als ein vorgegebener Wert R_{min} wird im jeweils nächsten ersten Schritt eine Approximation für die fehlenden Werte verwendet, deren Rang um eins geringer ist.

Auf diese Weise alterniert der Algorithmus zwischen den beiden dargestellten Schritten, bis die Schätzer konvergieren ($\epsilon_e > 0$ ist entsprechend zu wählen).

Durch die Verwendung von konstanten Werten, die einer mittelmäßigen Bewertung entsprechen, werden zu Beginn des Algorithmus Annahmen über die fehlenden Daten gemacht. Im weiteren Verlauf des Algorithmus, werden diese Annahmen jedoch fortwährend korrigiert.

Beispiel 5.8:

In diesem Beispiel wird der Schätzer für Bernds Bewertung von Barry Lyndon $\hat{Y}_{17}$ mit Hilfe des EM-SVM Verfahrens von Srebro, Jaakkola (2003) berechnet. Es

werden $R_{min} = 2$ und $\epsilon = 0,001$ verwendet. Als neutraler Wert für die Elemente der Matrix O_N wird 3 verwendet. In der ersten Iteration ergibt sich mit dem Startwert $n = 2$ im ersten Schritt:

$$\hat{Y}^2 = \begin{pmatrix} 1 & 5 & 4 & 4 & 3 & 4 & 3 & 3 \\ 3 & 4 & 2 & 2 & 2 & 5 & 3 & 3 \\ 3 & 4 & 4 & 4 & 3 & 3 & 5 & 1 \\ 5 & 2 & 2 & 3 & 2 & 3 & 3 & 4 \\ 4 & 3 & 2 & 3 & 4 & 4 & 3 & 4 \\ 1 & 5 & 4 & 4 & 4 & 3 & 4 & 3 \end{pmatrix}.$$

Im zweiten Schritt muß die Singulärwertzerlegung dieser Matrix berechnet werden. Man bestimmt zunächst die Eigenwerte von $\hat{Y}^2(\hat{Y}^2)'$:

$$\lambda_1 = 521,536,\ \lambda_2 = 26,890,\ \lambda_3 = 9,234,\ \lambda_4 = 5,028,\ \lambda_5 = 1,886,\ \lambda_6 = 0,426$$

(Diese Eigenwerte sind nicht entartet.) Die Matrix U^2 besteht aus den zugehörigen Eigenvektoren. Σ^2 ist eine Diagonalmatrix, auf deren Hauptdiagonalen die Wurzeln dieser Eigenwerte stehen. Die Beziehung $\mathcal{V}^2 = (\hat{Y}^2)'U^2(\Sigma^2)^{-1}$ ermöglicht die Berechnung von $\mathcal{V}^2$. Da alle Eigenwerte positiv sind, hat $\hat{Y}^2(\hat{Y}^2)'$ vollen Rang, weshalb $r = 6$ gilt. Daher ist $R(1) = 6 > R_{min}$. Hieraus folgt $R(2) = 5$. Die Rang-5 Approximation ist

$$\hat{Y}_5^2 = U_5^2\Sigma_5^2(\mathcal{V}_5^2)' = \begin{pmatrix} 0{,}96 & 5{,}10 & 3{,}94 & 3{,}84 & 2{,}93 & 3{,}86 & 3{,}21 & 3{,}13 \\ 3{,}02 & 3{,}95 & 2{,}03 & 2{,}08 & 2{,}04 & 5{,}07 & 2{,}89 & 2{,}92 \\ 2{,}99 & 4{,}03 & 3{,}98 & 3{,}95 & 2{,}98 & 2{,}96 & 5{,}07 & 1{,}05 \\ 5{,}01 & 1{,}97 & 2{,}02 & 3{,}06 & 2{,}02 & 3{,}05 & 2{,}93 & 3{,}95 \\ 3{,}97 & 3{,}07 & 1{,}96 & 2{,}89 & 3{,}96 & 3{,}91 & 3{,}14 & 4{,}09 \\ 1{,}05 & 4{,}88 & 4{,}07 & 4{,}19 & 4{,}08 & 3{,}16 & 3{,}76 & 2{,}83 \end{pmatrix}.$$

$\hat{Y}_5^2$ wird dann im ersten Schritt der darauffolgenden Iteration ($n = 3$) benutzt, um die fehlenden Werte zu imputieren. Man erhält $\hat{Y}^3$, indem man alle in y fehlenden Einträge durch die entsprechenden Einträge aus $\hat{Y}_5^2$ ersetzt.

Hinsichtlich der resultierenden Matrix

$$\hat{Y}^3 = \begin{pmatrix} 1 & 5 & 4 & 4 & 3 & 4 & 3{,}21 & 3{,}13 \\ 3{,}02 & 4 & 2 & 2 & 2 & 5 & 2{,}89 & 3 \\ 2{,}99 & 4 & 4 & 4 & 3 & 2{,}96 & 5 & 1 \\ 5 & 2 & 2 & 3{,}06 & 2 & 3 & 2{,}93 & 4 \\ 4 & 3{,}07 & 2 & 3 & 4 & 4 & 3{,}14 & 4 \\ 1 & 5 & 4 & 4 & 4 & 3{,}16 & 4 & 2{,}83 \end{pmatrix}$$

wird dann wieder die Singulärwertzerlegung bestimmt. Bei dem betrachteten Beispiel waren 370 Iterationen auszuführen. Als Endergebnis für den Schätzer ergibt sich $\hat{Y}_{17} = 3,86$.

Die Verwendung der Singulärwertzerlegung würde eigentlich kardinales Skalenniveau voraussetzen. Das Verfahren ist linear und die Daten werden nicht transfomiert. Um der Unvollständigkeit der Daten Rechnung zu tragen wird ein zum Data Augmentation Verfahren analoger Algorithmus verwendet. (Die Data Augmentation Methode ist nur dann unbedenklich, wenn die MAR-Eigenschaft erfüllt ist.) Wegen den in der Praxis zu erwartenden hohen Fehlendanteilen ist der mit diesem Verfahren verbundene Rechenaufwand nicht zu unterschätzen.

5.4 Das Gradientenverfahren zur Matrixfaktorisation

Das Gradientenverfahren zur Matrixfaktorisation nach Srebro, Jaakkola (2003) ist eng mit den Singulärwertzerlegung-basierten Verfahren verwandt. Im Gegensatz zu den Singulärwertzerlegung-basierten Methoden bestimmt das Gradientenverfahren seine Schätzer nicht durch indirektes Minimieren der Frobenius-Norm für eine vervollständigte Datenmatrix A_Y sondern durch Minimieren der gewichteten Frobenius Norm

$$\breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}}) = \sum_{i=1}^{I} \sum_{j=1}^{J} v_{ij}(y_{ij} - (\breve{U}\breve{\mathcal{V}}')_{ij})^2,$$

wobei $\breve{U} \in \mathbb{R}^{I,R}$ und $\breve{\mathcal{V}} \in \mathbb{R}^{J,R}$ nach geeigneter Wahl für $R \leq r$. Falls in y keine

fehlenden Werte vorkommen, besteht eine Beziehung zwischen $\breve{U}$ und $\breve{\mathcal{V}}$ und den Matrizen U_R, Σ_R und $\mathcal{V}_R$ der Singulärwertzerlegung. Dann gilt $\breve{U} = U_R \Sigma_R \Xi_U$ und $\breve{\mathcal{V}} = \mathcal{V}_R \Xi_V$, für geeignete Matrizen $\Xi_U \in \mathbb{R}^{R,R}$ und $\Xi_V \in \mathbb{R}^{R,R}$, für die $\Xi_U \Xi_V' = E$ gilt, wobei E wie gewohnt die Einheitsmatrix bezeichnet.
Mit der Definition

$$\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{U}'} = \begin{pmatrix} \frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{U}_{11}} & \cdots & \frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{U}_{1R}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{U}_{I1}} & \cdots & \frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{U}_{IR}} \end{pmatrix}$$

ergeben sich die folgenden partiellen Ableitungen:

$$\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{U}'} = 2(v \odot ((\breve{U}\breve{\mathcal{V}}') - y))\breve{\mathcal{V}}$$

$$\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{\mathcal{V}}'} = 2(v' \odot ((\breve{\mathcal{V}}\breve{U}') - y'))\breve{U}$$

Das Symbol $\odot$ bedeutet auch an dieser Stelle, daß die betroffenen Faktoren elementweise miteinander zu multiplizieren sind.
Die i-te Zeile von $\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{U}'}$ ist

$$\left(\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{U}'}\right)_i = \begin{pmatrix} 2\sum_{j'=1}^{J}(v_{ij'}((\sum_{x=1}^{R}\breve{U}_{ix}\breve{\mathcal{V}}_{j'x}) - y_{ij'}))\breve{\mathcal{V}}_{j'1} \\ \vdots \\ 2\sum_{j'=1}^{J}(v_{ij'}((\sum_{x=1}^{R}\breve{U}_{ix}\breve{\mathcal{V}}_{j'x}) - y_{ij'}))\breve{\mathcal{V}}_{j'R} \end{pmatrix}'.$$

Mit den Zeilenvektoren $V_i' = (v_{i1}, ..., v_{iJ})$, $\breve{U}_i' = (\breve{U}_{i1}, ..., \breve{U}_{iR})$, $\breve{\mathcal{V}}_j' = (\breve{\mathcal{V}}_{j1}, ..., \breve{\mathcal{V}}_{jR})$ und $y_i' = (y_{i1}, ..., y_{iJ})$ und den Diagonalmatrizen $\underline{V}_i = diag(v_{i1}, ..., v_{iJ})$ kann die Frobenius-Norm zeilenweise in Bezug auf $\breve{U}$ für vorgegebene Werte von $\breve{\mathcal{V}}$ minimiert werden:

$$\left(\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}})}{\partial \breve{U}'}\right)_i = 2(V_i' \odot ((\breve{U}_i'\breve{\mathcal{V}}') - y_i'))\breve{\mathcal{V}} \equiv 0 \Leftrightarrow (V_i' \odot (\breve{U}_i'\breve{\mathcal{V}}'))\breve{\mathcal{V}} = (V_i' \odot y_i')\breve{\mathcal{V}}.$$

Wegen der Beziehungen $V_i' \odot (\breve{U}_i'\breve{\mathcal{V}}') = \breve{U}_i'\breve{\mathcal{V}}'\underline{V}_i$ und $V_i' \odot y_i' = y_i'\underline{V}_i$ ergibt sich $(\breve{U}_i'\breve{\mathcal{V}}'\underline{V}_i)\breve{\mathcal{V}} = y_i'\underline{V}_i\breve{\mathcal{V}}$. Durch Transposition erhält man daraus $\breve{\mathcal{V}}'\underline{V}_i\breve{\mathcal{V}}\breve{U}_i = \breve{\mathcal{V}}'\underline{V}_i Y_i$. Daher ist für gegebenes $\breve{\mathcal{V}} = \breve{\mathcal{V}}^*$

$$\breve{U}_i^* = (\breve{\mathcal{V}}^{*\prime}\underline{V}_i\breve{\mathcal{V}}^*)^{-1}(\breve{\mathcal{V}}^{*\prime}\underline{V}_i Y_i)$$

die bezüglich der Frobenius-Norm optimale Lösung für die i-te Zeile. Für $\breve{\mathcal{V}} = \breve{\mathcal{V}}^*$ ist $\breve{U}^* = (\breve{U}_1^*, ..., \breve{U}_I^*)'$ die Lösung für $\breve{U}$, so daß

$$\left.\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}}^*)}{\partial \breve{U}'}\right|_{\breve{U}=\breve{U}^*} = 0.$$

und

$$\left.\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}}^*)}{\partial \breve{\mathcal{V}}'}\right|_{\breve{U}=\breve{U}^*} = 2(v' \odot ((\breve{\mathcal{V}}^*\breve{U}^{*\prime}) - y'))\breve{U}^*.$$

Diese Beziehung zwischen den partiellen Ableitungen wird im Gradientenverfahren zur Matrixfaktorisation nach Srebro, Jaakkola (2003) benutzt, um $\breve{U}$ und $\breve{\mathcal{V}}$ iterativ zu bestimmen. Ausgehend von einem Startwert $\breve{\mathcal{V}}^1$ für $\breve{\mathcal{V}}$ berechnet das Verfahren in jedem n-ten Schritt die Gleichungen (1) bis (5). So ergibt sich der folgende Algorithmus:

Startwerte: $n = 2, \breve{\mathcal{V}}^1, \hat{Y}^0 = O_N^{I,J}, \hat{Y}^1 = O_E^{I,J}$

Solange$\Big(\exists\ i \in \{1, ..., I\},\ j \in \{1, ..., J\} : \left\|\hat{Y}_{ij}^{n-1} - \hat{Y}_{ij}^{n-2}\right\| > \epsilon\Big)$ {

(1) $\breve{U}_i^n = ((\breve{\mathcal{V}}^{n-1})'\underline{V}_i\breve{\mathcal{V}}^{n-1})^{-1}((\breve{\mathcal{V}}^{n-1})'\underline{V}_i y_i)\ \forall\ i \in \{1, ..., I\}$

(2) $\breve{U}^n = (\breve{U}_1^n, ..., \breve{U}_I^n)'$

(3) $\breve{\mathcal{V}}^n = \breve{\mathcal{V}}^{n-1} - \lambda_G \left.\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}}^{n-1})}{\partial \breve{\mathcal{V}}'}\right|_{\tilde{U}=\tilde{U}^n}$

(4) $\hat{Y}^n = \breve{U}^n(\breve{\mathcal{V}}^n)'$

(5) $n \leftarrow n + 1$

}

Abbildung 5.3: Algorithmus des Gradientenverfahrens zur Matrixfaktorisation

λ_G bezeichnet hier die Schrittweite. Auf diese Weise werden lokale Minima der (nicht konvexen) gewichteten Frobenius-Norm $\mathcal{F}_G(\breve{U}, \breve{\mathcal{V}})$ gefunden (Srebro, Jaakkola (2003)). Diese sind nicht notwendigerweise auch global.

Beispiel 5.9:

In diesem Beispiel wird der Schätzer für Bernds Bewertung von Barry Lyndon $\hat{Y}_{17}$ mit Hilfe des Gradientenverfahrens zur Matrixfaktorisation nach Srebro, Jaakkola (2003) berechnet. Es werden $R = 2$ und $\epsilon = 0,001$ verwendet. Für die Schrittweite wird $\lambda_G = 0,0005$ gewählt. Als Startwert $\breve{\mathcal{V}}^1$ wird die Rang-R Approximation von $\mathcal{V}^2$ aus Beispiel 5.8 verwendet. Daher basiert $\breve{\mathcal{V}}^1$ auf der Singulärwertzerlegung der vervollständigtenm Matrix $\hat{Y}^2$ aus Beispiel 5.8. Mit

$$\breve{\mathcal{V}}^1 = \mathcal{V}_2^2 = \begin{pmatrix} \text{-0,295} & \text{0,719} \\ \text{-0,416} & \text{-0,364} \\ \text{-0,327} & \text{-0,365} \\ \text{-0,361} & \text{-0,169} \\ \text{-0,326} & \text{-0,087} \\ \text{-0,391} & \text{0,172} \\ \text{-0,377} & \text{-0,102} \\ \text{-0,318} & \text{0,376} \end{pmatrix}.$$

und

$$\underline{V}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

ergibt sich mit $y_1 = (1, 5, 4, 4, 3, 4, 0, 0)'$

$$\begin{aligned}\breve{U}_1^2 &= ((\breve{\mathcal{V}}^1)'\underline{V}_1\breve{\mathcal{V}}^1)^{-1}((\breve{\mathcal{V}}^1)'\underline{V}_1 y_1) \\ &= \begin{pmatrix} 1{,}336 & -0{,}128 \\ -1{,}127 & 1{,}991 \end{pmatrix}\begin{pmatrix} -7{,}670 \\ -2{,}811 \end{pmatrix} = \begin{pmatrix} -9{,}884 \\ -2{,}369 \end{pmatrix}.\end{aligned}$$

Die Vektoren $\breve{U}_2^2, ..., \breve{U}_6^2$ werden durch analoge Rechnungen bestimmt. Insgesamt erhält man

$$\breve{U}^2 = \begin{pmatrix} \breve{U}_1^{2'} \\ \breve{U}_2^{2'} \\ \breve{U}_3^{2'} \\ \breve{U}_4^{2'} \\ \breve{U}_5^{2'} \\ \breve{U}_6^{2'} \end{pmatrix} = \begin{pmatrix} -9{,}884 & -2{,}369 \\ -8{,}968 & 1{,}789 \\ -8{,}745 & -3{,}407 \\ -7{,}872 & 3{,}330 \\ -9{,}784 & 1{,}895 \\ -10{,}002 & -2{,}563 \end{pmatrix}.$$

Daher folgt

$$\left.\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}}^1)}{\partial \breve{\mathcal{V}}'}\right|_{\tilde{U}=\tilde{U}^2} = 2(v' \odot ((\breve{\mathcal{V}}^1\breve{U}^{2'}) - y'))\breve{U}^2 = \begin{pmatrix} -6{,}632 & -2{,}509 \\ -1{,}361 & -9{,}211 \\ -12{,}762 & -3{,}553 \\ -16{,}557 & 5{,}866 \\ 0{,}208 & 0{,}510 \\ 18{,}908 & 3{,}208 \\ 22{,}943 & 9{,}035 \\ -10{,}830 & -3{,}811 \end{pmatrix}.$$

Damit ergibt sich gemäß Formel (3) des Gradientenverfahren-Algorithmus

$$\breve{\mathcal{V}}^2 = \breve{\mathcal{V}}^1 - \lambda_G \left.\frac{\partial \breve{\mathcal{F}}_G(\breve{U}, \breve{\mathcal{V}}^1)}{\partial \breve{\mathcal{V}}'}\right|_{\tilde{U}=\tilde{U}^2} = \begin{pmatrix} -0{,}291 & -0{,}720 \\ -0{,}416 & -0{,}359 \\ -0{,}321 & -0{,}363 \\ -0{,}353 & -0{,}172 \\ -0{,}326 & -0{,}087 \\ -0{,}400 & 0{,}170 \\ -0{,}389 & -0{,}106 \\ -0{,}313 & 0{,}378 \end{pmatrix}.$$

Wegen $\hat{Y}^2 = \breve{U}^2(\breve{\mathcal{V}}^2)'$ ergibt sich im ersten Iterationsschritt für $n = 2$ $\hat{Y}_{17} = 4,10$. Nach 797 Schritten ($n = 798$) bricht das Verfahren ab. Man erhält somit als Endergebnis $\hat{Y}_{17} = 3,86$. Dieses Ergebnis ähnelt dem Ergebnis des EM-SVD Algorithmus, der die gleiche Startlösung verwendet. Ändert man die Startlösung $\breve{\mathcal{V}}^1$, indem man die fehlenden Werte in y nicht durch 3 sondern durch 2 ersetzt, sodann die Singulärwertzerlegung der auf diese Weise vervollständigten Matrix $\hat{\tilde{Y}} = \tilde{U}\tilde{\Sigma}\tilde{\mathcal{V}}'$ bestimmt und die Rang-R Approximation von $\tilde{\mathcal{V}}$ zur Startlösung $\breve{\mathcal{V}}^1$ macht, so erhält man ceteris paribus nach 584 Iterationen ebenfalls $\hat{Y}_{17} = 3,86$.

Die Verwendung der gewichteten Frobenius-Norm würde eigentlich kardinales Datenniveau voraussetzen. Das Gradientenverfahren zur Matrixfaktorisation nach Srebro, Jaakkola (2003) ist ein lineares heuristikbasiertes Verfahren. Durch die Verwendung der gewichteten Frobenius-Norm anstelle der einfachen Frobenius-Norm gelingt es Srebro, Jaakkola (2003) die fehlenden Daten zu ignorieren. Das Ignorieren fehlender Daten ist nicht unproblematisch, da es implizit die MCAR-Eigenschaft voraussetzt. Eine Schwierigkeit ist das Bestimmen einer guten Startlösung $\breve{\mathcal{V}}^1$ für $\breve{\mathcal{V}}$. Verwendet man wie in Beispiel 5.8 die Singulärwertzerlegung zur Ermittlung von $\breve{\mathcal{V}}^1$, so müssen zumindest zur Ermittlung der Startlösung Annahmen über die fehlenden Werte gemacht werden. Für große I, J und R kann der rechnerische Aufwand des Verfahrens beachtlich sein (Srebro, Jaakkola (2003)). Die Daten werden untransformiert verwendet.

5.5 Ordinale Matrixfaktorisation

Alle auf der Singulärwertzerlegung basierenden Ansätze haben für eine vorgegebene Datenmatrix eine Approximation niedrigeren Ranges bestimmt, indem eine bestimmte Anzahl kleinerer Singulärwerte weggelassen wurde. Analog wurde beim Gradientenverfahren zur Matrixfaktorisation die gewichtete Frobenius-Norm des Abstandes zwischen den Daten und einer niedrigrangigeren Approximation der Daten minimiert.

Sei wieder A_Y die vervollständigte Variante der Datenmatrix y und sei $A_Y = U\Sigma\mathcal{V}'$ die zugehörige Singulärwertzerlegung und $A_Y = \breve{U}\breve{\mathcal{V}}'$. r sei der Rang von

A_Y, $\breve{U}$ und $\breve{\mathcal{V}}$. Dann ist die Spur-Norm von A_Y gegeben durch

$$\|A_Y\|_\Sigma = Spur(\Sigma) = \sum_{x=1}^{r} \sqrt{\lambda_x^p}.$$

Statt die Anzahl der Singulärwerte zu beschränken, könnte man die Summe der Singulärwerte $\|A_Y\|_\Sigma$ auch auf andere Weise verkleinern.
Für die Frobenius-Normen der Matrizen $\breve{U}$ und $\breve{\mathcal{V}}$

$$\left\|\breve{U}\right\|_F = \sqrt{\sum_{i'=1}^{I}\sum_{x=1}^{r} \breve{U}_{i'x}^2} \quad \text{und} \quad \left\|\breve{\mathcal{V}}\right\|_F = \sqrt{\sum_{j'=1}^{J}\sum_{x=1}^{r} \breve{\mathcal{V}}_{j'x}^2}.$$

gilt nach Srebro et. al. (2005) die Beziehung

$$\min_{\substack{\breve{U},\,\breve{\mathcal{V}} \\ A_Y = \breve{U}\breve{\mathcal{V}}'}} \frac{1}{2}\left(\left\|\breve{U}\right\|_F^2 + \left\|\breve{\mathcal{V}}\right\|_F^2\right) = \|A_Y\|_\Sigma .$$

Deshalb ist

$$\frac{1}{2}\left(\left\|\breve{U}\right\|_F^2 + \left\|\breve{\mathcal{V}}\right\|_F^2\right)$$

eine obere Schranke für $\|A_Y\|_\Sigma$.

Die ordinale Matrixfaktorisation nach Rennie, Srebro (2005) beruht auf der Idee, Matrizen $\breve{U}$ und $\breve{\mathcal{V}}$ zu finden, die

$$\frac{1}{2}\left(\left\|\breve{U}\right\|_F^2 + \left\|\breve{\mathcal{V}}\right\|_F^2\right)$$

anstelle von $\|A_Y\|_\Sigma$ verkleinern und gleichzeitig Abweichungen von $A_Y = \breve{U}\breve{\mathcal{V}}'$ durch einen zusätzlichen Strafkostenanteil der Zielfunktion entgegenwirken.

Dieser Strafkostenterm berücksichtigt außerdem das ordinale Skalenniveau der Bewertungsdaten. Im Rahmen des Strafkostenansatzes wird der Schätzer $\hat{A}_Y = \breve{U}\breve{\mathcal{V}}'$ als latente kontinuierliche Größe aufgefaßt, die den ordinalen Bewertungsdaten zugrundeliegt aber selbst nicht in Erscheinung tritt. Wie bei den in Kapitel 2 vorgestellten klassischen Verfahren für ordinalskalierte Daten werden

daher Schwellenwerte benötigt, mittels derer diese latente Größe einem (ordinalen) Schätzer zugeordnet werden kann. Zu den verschiedenen möglichen Bewertungen $c \in \{1, ..., C\}$ gehören für jeden Nutzer i Schwellenwerte $\gamma_{i1} \leq ... \leq \gamma_{iC-1}$. (Außerdem gilt $\gamma_{i0} = -\infty$.)
Mit der Fallunterscheidungsvariable

$$\mathcal{X}_{ij}^{c} = \begin{cases} +1, & \text{für } c \geq y_{ij} \\ -1, & \text{für } c < y_{ij} \end{cases}$$

und

$$\gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1(C-1)} \\ \vdots & \ddots & \vdots \\ \gamma_{I1} & \cdots & \gamma_{I(C-1)} \end{pmatrix}$$

ergibt sich als Strafkostenanteil der Zielfunktion der Term

$$\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} \sum_{c'=1}^{C-1} h(\mathcal{X}_{i'j'}^{c'}(\gamma_{i'c'} - \hat{A}_{Y,i'j'})).$$

Hier gilt

$$h(z) = \begin{cases} \frac{1}{2} - z, & \text{für } z < 0 \\ \frac{1}{2}(1-z)^2, & \text{für } 0 \leq z \leq 1 \\ 0, & \text{für } z > 1 \end{cases} .$$

Es können die folgenden vier Fälle auftreten:

$$\begin{array}{lll} \text{1. Fall:} & \gamma_{ic} \geq \hat{A}_{Y,ij} \wedge c \geq y_{ij} & (\Rightarrow \mathcal{X}_{ij}^{c}(\gamma_{ic} - \hat{A}_{Y,ij}^{x}) \geq 0) \\ \text{2. Fall:} & \gamma_{ic} < \hat{A}_{Y,ij}^{x} \wedge c \geq y_{ij} & (\Rightarrow \mathcal{X}_{ij}^{c}(\gamma_{ic} - \hat{A}_{Y,ij}^{x}) < 0) \\ \text{3. Fall:} & \gamma_{ic} \geq \hat{A}_{Y,ij}^{x} \wedge c < y_{ij} & (\Rightarrow \mathcal{X}_{ij}^{c}(\gamma_{ic} - \hat{A}_{Y,ij}^{x}) \leq 0) \\ \text{4. Fall:} & \gamma_{ic} < \hat{A}_{Y,ij}^{x} \wedge c < y_{ij} & (\Rightarrow \mathcal{X}_{ij}^{c}(\gamma_{ic} - \hat{A}_{Y,ij}^{x}) > 0) \end{array}$$

Im ersten und vierten Fall ist die geschätzte latente Größe in Bezug auf die verwendeten Schwellenwerte in guter Übereinstimmung mit der ordinalen Größe. Nur wenn $\mathcal{X}_{ij}^{c}(\gamma_{ic} - \hat{A}_{Y,ij}^{x}) < 1$ ist kann es in diesen Fällen sein, daß geringfügige Strafkosten entstehen. Die beiden übrigen Fälle führen meist zu deutlich größeren

Strafkosten, da in diesen Fällen der Schätzer für die latente Größe im Widerspruch zur ordinalen Größe steht. Dabei fallen die Strafkosten umso höher aus, je weiter der Schätzer $\hat{A}^x_{Y,ij}$ über den Schwellenwert γ_{ic} hinausgeht obwohl $c < y_{ij}$ ist (zweiter Fall) bzw. je weiter der Schätzer $\hat{A}^x_{Y,ij}$ hinter dem Schwellenwert γ_{ic} zurückbleibt obgleich $c > y_{ij}$ ist.
Insgesamt folgt die um den Strafkostenterm vervollständigte Zielfunktion

$$\mathcal{Z}_{MF}(\breve{U},\breve{\mathcal{V}},\gamma) = \frac{1}{2}\left(\left\|\breve{U}\right\|_F^2 + \left\|\breve{\mathcal{V}}\right\|_F^2\right) + \mathcal{C}_{MF}\sum_{i'=1}^{I}\sum_{j'\in J_{i'}}\sum_{c'=1}^{C-1} h(\mathcal{X}^{c'}_{i'j'}(\gamma_{i'c'} - (\breve{U}\breve{\mathcal{V}}')_{i'j'}))$$

Rennie, Srebro (2005) verwenden

$$h'(z) = \begin{cases} -1, & \text{für } z < 0 \\ -1 + z, & \text{für } 0 \le z \le 1 \\ 0, & \text{für } z > 1 \end{cases}$$

als Ableitung der Funktion $h(z)$. Damit ergeben sich die partiellen Ableitungen

$$\frac{\partial \mathcal{Z}_{MF}(\breve{U},\breve{\mathcal{V}},\gamma)}{\partial \breve{U}_{ia}} = \breve{U}_{ia} - \mathcal{C}_{MF}\sum_{c'=1}^{C-1}\sum_{j'\in J_i} \mathcal{X}^{c'}_{ij'} h'(\mathcal{X}^{c'}_{ij'}(\gamma_{ic'} - (\breve{U}\breve{\mathcal{V}}')_{ij'}))\breve{\mathcal{V}}_{j'a}$$

$$\frac{\partial \mathcal{Z}_{MF}(\breve{U},\breve{\mathcal{V}},\gamma)}{\partial \breve{\mathcal{V}}_{ja}} = \breve{\mathcal{V}}_{ja} - \mathcal{C}_{MF}\sum_{c'=1}^{C-1}\sum_{i'\in I_j} \mathcal{X}^{c'}_{i'j} h'(\mathcal{X}^{c}_{i'j}(\gamma_{i'c'} - (\breve{U}\breve{\mathcal{V}}')_{i'j}))\breve{U}_{i'a} \ .$$

$$\frac{\partial \mathcal{Z}_{MF}(\breve{U},\breve{\mathcal{V}},\gamma)}{\partial \gamma_{ic}} = \mathcal{C}_{MF}\sum_{j'\in J_i} \mathcal{X}^{c}_{ij'} h'(\mathcal{X}^{c}_{ij'}(\gamma_{ic} - (\breve{U}\breve{\mathcal{V}}')_{ij'}))$$

Diese ermöglichen die iterative Berechnung der Matrizen $\breve{U}$ und $\breve{\mathcal{V}}$ sowie der Schwellenwerte γ. $\mathcal{Z}_{MF}(\breve{U},\breve{\mathcal{V}},\gamma)$ ist nicht konvex in $\breve{U}$ und $\breve{\mathcal{V}}$. Experimentell konnte an einem niedrigdimensionalen Problem gezeigt werden, daß alle berechneten Minima dieser Zielfunktion auch globale Minima waren (Rennie, Srebro (2005)).

Die Zielfunktion mit dem Strafkosten-Ansatz $\mathcal{Z}_{MF}(\breve{U},\breve{\mathcal{V}},\gamma)$ von Rennie, Srebro (2005) wird dem ordinalen Skalenniveau der Daten besser gerecht als beispielsweise die gewichtete Frobenius-Norm $\mathcal{F}_G(\breve{U},\breve{\mathcal{V}})$ nach Srebro, Jaakkola (2003). Weil die gewichtete Frobenius-Norm die Summe der quadratischen Abstände zwischen y und $\hat{Y}$ minimiert, trugen für alle Werte $y_{ij} = C$ die Schätzer $\hat{Y}_{ij} = C + \Delta_C$

für $\Delta_C > 0$ gleich viel zur Zielfunktion bei wie $\hat{Y}_{ij} = C - \Delta_C$. Durch die Verwendung von Schwellenwerten $\gamma_{i1}, ..., \gamma_{i(C-1)}, i = 1, ..., I$, werden positive Abweichungen von $y_{ij} = C$ nicht in der Zielfunktion $\mathcal{Z}_{MF}(\breve{U}, \breve{\mathcal{V}}, \gamma)$ berücksichtigt. Ebenso leisten negative Abweichungen von $y_{ij} = 1$ keinen Beitag zu dieser Zielfunktion. Außerdem vermeidet der Ansatz erfolgreich das Bilden von Differenzen zwischen den ordinalen Daten y_{ij} und ihren reellen Schätzern. Die Verwendung reeller (latenter) Werte und die Verwendung von $C-1$ Schranken für diese Werte ist typisch für ordinale Ansätze (vgl. Kapitel 2).

Fehlende Daten werden ignoriert, was implizit die MCAR-Eigenschaft voraussetzen würde. Der lineare heuristikbasierte Ansatz von Rennie, Srebro (2005) verwendet die Daten untransformiert. Zur Implementation wurde das Konjugierte Gradientenverfahren mit der Polak-Ribière Modifikation benutzt (siehe Anhang D). Nach Rennie, Srebro (2005) führt die Verwendung des ordinalen Matrixfaktorisationsansatzes zu besseren Ergebnissen als die Verwendung des Item-basierten Ähnlichkeitsverfahrens, des Gradientenverfahrens nach Srebro, Jaakkola (2003), der SVD-EM Methode und des Hauptkomponentenverfahrens (Goldberg et. al. (2001)).

Beispiel 5.10:

Es wird der Schätzer für Bernds Bewertung von Barry Lyndon $\hat{Y}_{17}$ mit Hilfe des Matrixfaktorisationsverfahrens für ordinale Datenmatrizen nach Rennie, Srebro (2005) berechnet. Zu Beginn muß für alle Nutzer, Items und Bewertungsgruppen c $\mathcal{X}_{ij}^c$ bestimmt werden. Es gilt:

$\mathcal{X}_{1j}^c$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$c=1$	$+1$	-1	-1	-1	-1	-1	-	-
$c=2$	$+1$	-1	-1	-1	-1	-1	-	-
$c=3$	$+1$	-1	-1	-1	$+1$	-1	-	-
$c=4$	$+1$	-1	$+1$	$+1$	$+1$	$+1$	-	-

Tabelle 5.7: Beispiel 5.10: Bewertungstensor $\mathcal{X}_{1j}^c$

Da der erste Nutzer keine Bewertungen für die Items $j = 1$ und $j = 2$ abgegeben

hat, fehlen die letzten beiden Spalten. Die entsprechenden Einträge werden zur Berechnung der Gradienten nicht benötigt. Im ersten Schritt werden Startwerte für $\breve{U}$, $\breve{\mathcal{V}}$ und γ benutzt. Als Startwerte $\breve{U}^1$ und $\breve{\mathcal{V}}^1$ (für $\breve{U}$ und $\breve{\mathcal{V}}$) werden $\breve{U}^1 = U\Sigma^{1/2}$ und $\breve{\mathcal{V}}^1 = \mathcal{V}\Sigma^{1/2}$ verwendet, wobei U, Σ und $\mathcal{V}$ Matrizen der Singulärwertzerlegung $A_Y = U\Sigma\mathcal{V}$ sind. Für die Berechnung der Startwerte wird hier $A_Y = \hat{Y}^2$ mit $\hat{Y}^2$ aus Beispiel 5.8 gewählt.

Nach dieser Initialisierung werden in jedem weiteren Schritt die drei Gradienten benutzt, um neue Werte für $\breve{U}$, $\breve{\mathcal{V}}$ und γ aus den im vorherigen Schritt bestimmten Werten zu berechnen. Es wird die gleiche Abbruchbedingung wie in Beispiel 5.8 und 5.9 verwendet. Nach 704 Iterationen bricht das Verfahren ab und man erhält die Matrix der latenten reellen Werte $A_Y^{705} = \breve{U}^{705}(\breve{\mathcal{V}}^{705})'$. Auf diese Weise erhält man $A_{Y,17}^{705} = 3,43$. Die Schwellenwerte $\gamma_{1c}, c \in \{1, ..., C-1\}$ für den ersten Nutzer sind $\gamma_{11} = 1,62$, $\gamma_{12} = 1,88$, $\gamma_{13} = 3,03$, $\gamma_{14} = 4,75$. Daher ergibt sich aufgrund des latenten Werts $A_{Y,17}^{705} = 3,43$ wegen der Beziehung $\gamma_{13} < A_{Y,17}^{705} < \gamma_{14}$ der Schätzer $\hat{Y}_{17} = 4$.

5.6 Zweimodale Clusterverfahren

Zweimodale Clusterverfahren benutzen die zweimodale Datenmatrix y um simultan die Elemente der ersten Modalität und die Elemente der zweiten Modalität zu clustern. Die Matrix $P = (p_{ik}) \in \mathbb{R}^{I,K}$ enthält alle Information über die Cluster-Zugehörigkeit der Elemente der ersten Modalität:

$$p_{ik} = \begin{cases} 1, & \text{falls } i \in \{1, ..., I\} \text{ zum Cluster} \\ & k \in \{1, ..., K\} \text{ gehört} \\ 0, & \text{sonst} \end{cases} . \tag{5.1}$$

K ist die Anzahl der Cluster für die Elemente erster Modalität. Für die Elemente zweiter Modalität beschreibt die Matrix $Q = (q_{jl}) \in \mathbb{R}^{J,L}$ das Verhältnis der Elemente der zweiten Modalität $j \in \{1, ..., J\}$ zu den für sie vorgesehenen Clustern $l = 1, ..., L$:

$$q_{jl} = \begin{cases} 1, & \text{falls } j \in \{1, ..., J\} \text{ zum Cluster} \\ & l \in \{1, ..., L\} \text{ gehört} \\ 0, & \text{sonst} \end{cases} . \tag{5.2}$$

Dabei ist L die Anzahl der Cluster zweiter Modalität. Die Anzahl der Cluster K und L müssen gewählt werden. Hinsichtlich der Matrizen P und Q können unterschiedliche Nebenbedingungen verwendet werden. Beispielsweise kann gefordert werden, daß jedes Element zu genau einem Cluster gehört:

$$p_{ik} \in \{0,1\}, i = 1,...,I, k = 1,...,K, \quad \wedge \quad \sum_{k'=1}^{K} p_{ik'} = 1, i = 1,...,I, \quad (N1)$$

$$q_{jl} \in \{0,1\}, j = 1,...,J, l = 1,...,L, \quad \wedge \quad \sum_{l'=1}^{L} q_{jl'} = 1, j = 1,...,J, \quad (N2)$$

Hier handelt es sich um nicht-überlappende Cluster. Falls ein Element zu mehreren Clustern gehören kann, spricht man von überlappenden Clustern. Dann gilt:

$$p_{ik} \in \{0,1\}, i = 1,...,I, k = 1,...,K, \quad \wedge \quad \sum_{k'=1}^{K} p_{ik'} \geq 1, i = 1,...,I, \quad (N3)$$

$$q_{jl} \in \{0,1\}, j = 1,...,J, l = 1,...,L, \quad \wedge \quad \sum_{l'=1}^{L} q_{jl'} \geq 1, j = 1,...,J, \quad (N4)$$

Zudem existieren auch Fuzzy-Ansätze für Zweimodales Clustering (Schlecht, Gaul (2004)). Bei diesen ist $p_{ik}, q_{jl} \notin \{0,1\}$ sondern $p_{ik}, q_{jl} \in [0,1]$. In der Literatur gibt es ein- und mehrmodale Clusterverfahren, die auch für unvollständige Datensätze geeignet sind (z.B. Espejo, Gaul (1986), Gaul, Schader (1994), Schlecht, Gaul (2004)).

Alle zur Vorhersage online-generierter ranggeordneter Bewertungsdaten verwendeten zweimodalen Clusterverfahren (Schlecht, Gaul (2004), George, Merugu (2005)) bestimmen Schätzer $\hat{S}_Y^x, x = 1,2$, für die Daten y, indem sie die gewichtete Frobenius-Norm

$$\mathcal{F}_G(P,Q) = \sum_{i'=1}^{I} \sum_{j' \in J_{i'}} v_{i'j'}(y_{i'j'} - \hat{S}_{Y,i'j'}^x(P,Q))^2, \ x = 1,2,$$

unter den Nebenbedingungen $(N1)$ und $(N2)$ minimieren. Für den Schätzer existieren unterschiedliche Ansätze $\hat{S}_Y^x, x = 1,2$. Sofern y keine fehlenden Einträge enthält, ist die gewichtete Frobenius-Norm $\mathcal{F}_G$ mit der Frobenius-Norm $\mathcal{F}$ äquivalent. Für vollständige Datenmatrizen y kann die Frobenius-Norm durch die

ursprüngliche Version des AE-Algorithmus nach Gaul, Schader (1996) minimiert werden:

1. P und Q werden den Nebenbedingungen entsprechend gewählt.
2. W wird aufgrund von P und Q berechnet. Falls $x = 2$ ist, berechne $\hat{y}_{i.}$ und $\hat{y}_{.j}$.
3. Die folgenden Schritte werden wiederholt bis sich keine Änderungen in P und Q mehr ergeben:
   a) $\forall i \in \{1, ..., I\}$:

      Versuche das betrachtete Element einem anderen Cluster erster Modalität unter Beachtung der Nebenbedingungen zuzuordnen. Berechne nach jedem Austausch W auf Basis der durch den Austausch geänderten Matrix P. Falls $x = 2$ ist, müssen außerdem noch $\tilde{w}_{k.}, k = 1, ..., K$, und $\tilde{w}_{.l}, l = 1, ..., L$, berechnet werden. Bestimme $\mathcal{F}$ für die neuen Werte von P und W neu. Akzeptiere die Änderung, falls $\mathcal{F}$ sich dadurch verkleinert. Ansonsten wird die Änderung rückgängig gemacht.

   b) $\forall j \in \{1, ..., J\}$:

      Ändere versuchsweise die Clusterzugehörigkeit des betrachteten Elements zweiter Modalität gemäß den Nebenbedingungen. Ermittle W neu auf Basis der durch jeweiligen Austausch geänderten Matrix Q. Sofern $x = 2$ ist, müssen außerdem noch $\tilde{w}_{k.}, k = 1, ..., K$, und $\tilde{w}_{.l}, l = 1, ..., L$, ermittelt werden. Berechne $\mathcal{F}$ für die geänderten Matrizen Q und W neu. Falls $\mathcal{F}$ sich durch eine Änderung verkleinert führe diese durch. Andernfalls mache die Änderung rückgängig.

Abbildung 5.4: AE-Algorithmus

Der AE-Algorithmus versucht, die Frobenius-Norm $\mathcal{F}$ zu minimieren, indem immer wieder Elemente erster oder zweiter Modalität versuchsweise einem anderen Cluster zugeordnet werden. Verbessert sich auf Weise nach Anpassung der Gewichtematrix W die Zielfunktion, wird die Änderung übernommen. Dieser Algorithmus kann unabhängig von der Wahl des Schätzers $\hat{S}_Y^x, x = 1, 2$, verwendet werden.

Der bekannteste Ansatz für den Schätzer $\hat{S}^1_{Y,ij}, x = 1$, wird bei einer Gruppe von zweimodalen Clustering-Ansätzen verwendet, die auf einer Verallgemeinerung des ADCLUS-Modells von Shepard, Arabie (1979) basieren. Prominente Mitglieder dieser Verfahrensklasse sind das GENNCLUS-Verfahren (Desarbo (1982)), das PENCLUS-Verfahren (Both, Gaul (1987)) und das AE-Verfahren nach Gaul, Schader (1996). Diese Klasse von zweimodalen Clusterverfahren, benutzt den Schätzer $\hat{S}_Y = PWQ' + \tilde{C}$, wobei alle Elemente von $\tilde{C} \in \mathbb{R}^{I,J}$ gleich und konstant sind. Bei nicht-überlappenden Ansätzen kann $\tilde{C}$ weggelassen werden (Baier et. al. (1997)).

Die Algorithmen der meisten ADCLUS-Verallgemeinerungen verwenden als Zielfunktion die Frobenius-Norm $\mathcal{F}$ und können fehlende Daten nicht berücksichtigen. Daher können sie zur Minimierung der gewichteten Frobenius-Norm $\mathcal{F}_G$ nur eingesetzt werden, wenn alle Einträge der Matrix V gleich 1 sind. Im Gegensatz zu den meisten Verfahren seiner Klasse wurde der Alternating-Exchanges (AE) Algorithmus (Gaul, Schader (1996)) so verallgemeinert, daß er auf Datenmatrizen mit fehlenden Werten anwendbar ist (Gaul et. al. (2006)).

Die Matrix $W \in \mathbb{R}^{K,L}$ wird als Gewichtungsmatrix bezeichnet. Im Fall eines nicht-überlappenden nicht-fuzzy zweimodalen Ansatzes gelten die Nebenbedingungen $(N1)$ und $(N2)$ und daher folgt

$$\begin{aligned}\mathcal{F}_G &= \sum_{i'=1}^{I}\sum_{j'=1}^{J} v_{i'j'}\left(y_{i'j'} - \sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{i'k'}w_{k'l'}q_{j'l'}\right)^2 \\ &= \sum_{i'=1}^{I}\sum_{j'=1}^{J} v_{i'j'}\left(\sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{i'k'}(y_{i'j'} - w_{k'l'})q_{j'l'}\right)^2.\end{aligned}$$

$p_{ik'}(y_{ij} - w_{k'l'})q_{jl'}$ ist für jedes Paar (i,j) nur dann von Null verschieden, wenn $p_{ik'} = q_{jl'} = 1$ gilt. Wegen $\sum_{k'=1}^{K} p_{ik'} = 1$ und $\sum_{l'=1}^{L} q_{jl'} = 1$ ist dies für jedes Paar (i,j) nur genau für ein bestimmtes Paar (k,l) der Fall. Man erhält deshalb

$$\begin{aligned}\mathcal{F}_G &= \sum_{i'=1}^{I}\sum_{j'=1}^{J}\sum_{k'=1}^{K}\sum_{l'=1}^{L} v_{i'j'}p^2_{i'k'}(y_{i'j'} - w_{k'l'})^2 q^2_{j'l'} \\ &= \sum_{i'=1}^{I}\sum_{j'=1}^{J}\sum_{k'=1}^{K}\sum_{l'=1}^{L} V_{i'j'}p_{i'k'}(y_{i'j'} - w_{k'l'})^2 q_{j'l'}.\end{aligned}$$

Aus

$$\frac{\partial \mathcal{F}_G}{\partial w_{kl}} = -2 \sum_{i'=1}^{I} \sum_{j'=1}^{J} v_{i'j'} p_{i'k} (y_{i'j'} - w_{kl}) q_{j'l} \equiv 0$$

ergibt sich

$$w_{kl} = \frac{\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} p_{i'k} y_{i'j'} q_{j'l}}{\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} p_{i'k} q_{j'l}} \tag{5.3}$$

Diese Beziehung benutzen Gaul et. al. (2006) um den AE-Algorithmus (Gaul, Schader (1996)) so zu verallgemeinern, daß er auch auf unvollständige Datenmatrizen anwendbar ist. Gaul et. al. (2006) definieren zu diesem Zweck:

$$w_{kl} = \begin{cases} \dfrac{\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} p_{i'k} y_{i'j'} q_{j'l}}{\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} p_{i'k} q_{j'l}}, & \text{falls } \sum_{i'=1}^{I} \sum_{j' \in J_{i'}} p_{i'k} q_{j'l} > 0 \\ 0, & \text{sonst} \end{cases} \tag{5.4}$$

In Bezug auf die von Recommender-Systemen erhobenen Bewertungsdaten sind die typischen Nutzer-Zahlen I und Item-Zahlen J so hoch, daß es durch die Wahl eher kleiner Cluster-Anzahlen K und L selbst bei sehr hohen Fehlendanteilen praktisch nie vorkommt, daß die Bedingung

$$\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} p_{i'k} q_{j'l} > 0 \quad \text{für irgendeine Kombination } (k, l)$$

nicht erfüllt ist. Anstelle der Null könnte auch eine andere Konstante (z.B. eine typische neutrale Bewertung) oder der Mittelwert $\bar{y}_{\cdot\cdot}$ verwendet werden. Die von Gaul et. al. (2006) getroffene Wahl hat bei positiven Bewertungsdaten den erheblichen Vorteil, daß man dem jeweiligen $\hat{S}^1$-Schätzer sofort ansehen kann, falls er nicht auf Basis mindestens einer gegebenen Bewertung berechnet wurde.

Nachdem alle Elemente erster und zweiter Modalität Clustern zugeordnet sind, kann man solange Zeilen und Spalten miteinander vertauscht, bis alle Elemente, die derselben Klasse angehören, nebeneinander in der Matrix stehen. Daher wird die Datenmatrix durch das zweimodale Clusterverfahren in Partitionen unterteilt. Zu jeder dieser Partitionen $C(k,l)$ gehören alle Daten aller Elemente des Clusters erster Modalität $C_1(k)$, die in Bezug auf Elemente des Clusters zweiter Modalität $C_2(l)$ stehen ($y_{ij} \in C(k,l) \Leftrightarrow (i \in C_1(k) \wedge j \in C_2(l))$. Jedes Element w_{kl} der Matrix W ist somit der Durchschnitt aller vorhandenen Einträge y_{ij}, deren Element erster Modalität i zum Cluster erster Modalität k gehört und deren Element zweiter Modalität j Bestandteil des Clusters zweiter Modalität $C_2(l)$ ist. w_{kl} ist somit der Durchschnitt aller gegebenen Werte $y_{ij} \in C(k,l)$. Der Schätzer $\hat{S}_Y^1$ entspricht somit der Imputation des Partition-Mittelwerts.

Die Formel $\hat{S}_Y^1 = PWQ'$ erinnert bei oberflächlicher Betrachtung an die Singulärwertzerlegung. Im Gegensatz zur Singulärwertzerlegung ist W keine Diagonalmatrix und die Elemente der Matrizen P und Q unterliegen den angegebenen Nebenbedingungen.

Genau wie bei der Wahl des verwendeten Ranges R bei der Rang-R Approximation der Singulärwertzerlegung müssen K und L groß genug gewählt werden, um alle wesentlichen allgemeinen Tendenzen zwischen Nutzern und Items abzubilden. Gleichzeitig müssen K und L genau wie der gewählte Rang R klein genug sein, damit von Besonderheiten des spezifischen Trainingsdatensatzes abstrahiert werden kann. Die Wahl $K = I$ und $L = J$ würde zwar genau wie die Wahl $R = r$ zu einer perfekten Anpassung des Datensatzes führen, die sich ergebende Klassifikation wäre aber vollkommen nutzlos, da dann jedes Cluster nur aus einer einzigen Person bestünde. Da das Ziel ist, für jeden Nutzer und jedes Item möglichst ähnliche Nutzer und Items zu ermitteln und die Bewertungen dieser Nutzer bezüglich der ähnlichen Items zur Schätzung fehlender Werte einzusetzen, sind zu kleine Cluster unvorteilhaft, weil auf diese Weise die Vorhersagen auf sehr wenigen Daten basieren. Zu große Cluster können dazu führen, daß Nutzer und Items zur Vorhersage der Bewertung eines Nutzers für ein Item benutzt werden, die den betrachteten Elementen nicht ähnlich genug sind.

Beispiel 5.11:

Es wird der Schätzer für Bernds Bewertung von Barry Lyndon $\hat{Y}_{17}$ mit Hilfe des

$\hat{S}_y^1$-Schätzers der nicht-überlappenden ADCLUS-Verallgemeinerungen bestimmt. Hier wird $K = 2$ und $L = 3$ benutzt. Zu Beginn des Verfahrens werden durch Zufallszahlen Matrizen P^1 und Q^2 erzeugt, die den oben angegebenen Nebenbedingungen gerecht werden:

$$P^1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \quad \text{und} \quad Q^1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}'.$$

Auf Basis von P^1 und Q^1 kann W^1 berechnet werden:

$$W^1 = \begin{pmatrix} 4{,}00 & 3{,}17 & 2{,}83 \\ 3{,}00 & 3{,}00 & 3{,}75 \end{pmatrix}.$$

Damit ergibt sich ein Zielfunktionswert $\mathcal{F}_G^1 = 38,75$. Die erste Iteration ($n = 2$) beginnt damit, daß die Clusterzugehörigkeit des ersten Nutzers versuchsweise geändert wird. Da es nur zwei Nutzer-Cluster gibt, ändert sich durch $p_{11}^2 = 0$ und $p_{12}^2 = 1$ die gesamte Gewichtematrix.
So erhält man $\mathcal{F}_G^2 = 43,8$, weshalb wegen $\mathcal{F}_G^1 < \mathcal{F}_G^2$ die Änderung wieder rückgängig gemacht wird. Da keine weiteren Nutzer-Cluster mehr vorhanden sind, versucht man nun durch die Änderung der Clusterzugehörigkeit des zweiten Nutzers den Zielfunktionswert zu verringern. Auf diese Weise werden alle Nutzer und Items nacheinander betrachtet bis sich keine Änderungen mehr ergeben. Als Resultat ergeben sich die zweimodalen Klassenzugehörigkeiten

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{und} \quad Q = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}'.$$

Mit diesen Clusterzugehörigkeiten ergibt sich nach entsprechender Umordnung der Zeilen und Spalten die folgende Darstellung für die Daten:

		$l=1$			$l=2$		$l=3$		
		$j=3$	$j=4$	$j=5$	$j=1$	$j=8$	$j=2$	$j=6$	$j=7$
$k=1$	$i=1$	4	4	3	1	-	5	4	-
	$i=3$	4	4	3	-	1	4	-	5
	$i=6$	4	4	4	1	-	5	-	4
$k=2$	$i=2$	2	2	2	-	3	4	5	-
	$i=4$	2	-	2	5	4	2	3	-
	$i=5$	2	3	4	4	4	-	4	-

w_{kl}	$l=1$ $j=3,4,5$	$l=2$ $j=1,8$	$l=3$ $j=2,6,7$
$k=1$ $i=1,3,6$	$w_{11}=\bar{y}_{\cdot\cdot}^{C(1,1)}=3,78$	$w_{12}=\bar{y}_{\cdot\cdot}^{C(1,2)}=1,0$	$w_{13}=\bar{y}_{\cdot\cdot}^{C(1,3)}=4,5$
$k=2$ $i=2,4,5$	$w_{21}=\bar{y}_{\cdot\cdot}^{C(2,1)}=2,38$	$w_{22}=\bar{y}_{\cdot\cdot}^{C(2,2)}=3,6$	$w_{23}=\bar{y}_{\cdot\cdot}^{C(2,3)}=4,0$

Abbildung 5.5: Partitionen $C(k,l)$ von y und ihre Beziehung zu W

Man erkennt, daß die die Elemente w_{kl} von W einfach die Partitionsmittelwerte $\bar{y}_{\cdot\cdot}^{C(k,l)}$ der Partitionenen $C(k,l)$ sind. Da Bernd ($i=1$) Element des ersten Nutzer-Clusters ist und Barry Lyndon ($j=7$) zum dritten Item-Cluster gehört ergibt sich der Schätzer $\hat{Y}_{17}=\hat{S}^1_{Y,17}=w_{13}=4,5$.

Die Gewichtematrix W ermöglicht eine Interpretation der Cluster erster Modalität anhand ihrer Beziehung zu den Clustern zweiter Modalität und umgekehrt. Dies wird im folgenden genauer erläutert. Im Beispiel 5.11 sind die Nutzer aus dem ersten Cluster erster Modalität Nutzer, die eine deutliche Abneigung gegenüber den Filmen aus dem zweiten Item-Cluster aufweisen. Das zweite Cluster zweiter Modalität besteht aus den Filmen Twister und Armageddon. Beides sind Action-Filme. Zudem haben die Nutzer aus dem ersten Nutzer-Cluster eine deutliche Präferenz für Filme aus dem dritten Item-Cluster, die auch von den Personen

aus dem zweiten Nutzer-Cluster tendenziell bevorzugt werden. Es handelt sich hierbei um die Filme Magnolia, Dangerous Liaisons und Barry Lyndon. Alle drei sind Dramen. Das erste Nutzer-Cluster hat eine positivere Einstellung als das zweite in Bezug auf das erste Item-Cluster, das aus den Filmen Casino, Taxi Driver und Platoon besteht. Diese Filme setzen sich auf unbequeme Weise mit Themen wie Macht, Sinnsuche und Krieg auseinander. Zudem kommt in diesen Filmen deutlich mehr Gewalt vor als in den übrigen Filmen. Insgesamt wissen wir daher über die Nutzer des ersten Clusters, daß sie Action-Filme ablehnen, Dramen bevorzugen und keine Abneigung gegen Gewalt oder die Behandlung problematischer Inhalte in Filmen haben.

Umgekehrt lassen sich auch die einzelnen Item-Cluster durch ihre in der Gewichtematrix W enthaltene Beziehung zu den Clustern erster Modalität interpretieren. Beispielsweise werden die Dramen von den Nutzern aus beiden Clustern erster Modalität bevorzugt und sind somit Filme, die im betrachteten Beispiel von allen Nutzern gemocht werden. Somit könnte man das Ergebnis selbst dann interpretieren, wenn man nicht über über die Datenmatrix y hinausgehende Zusatzinformation in Bezug auf die Elemente der ersten bzw. der zweiten Modalität verfügt. Solange man die Items kennt, sind Hintergrundinformationen über die Nutzer nicht erforderlich, um die Nutzer-Cluster zu interpretieren.

Die zweimodalen Clusterverfahren sind daher nicht nur eine Möglichkeit, Schätzer für die in einer Datenmatrix fehlenden Bewertungen zu berechnen, sondern darüberhinaus dazu geeignet, generelle Tendenzen der Daten zu extrahieren.

Der Schätzer $\hat{S}_Y^1$ erhält für gegebene Clusterzugehörigkeiten (P und Q) die Mittelwerte der zweimodalen Partitionen $C(k,l)$ von y. Sei

$$\bar{y}_{\cdot\cdot}^{C(k,l)} = \frac{\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} p_{i'k} y_{i'j'} q_{j'l}}{\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} p_{i'k} q_{j'l}}, \text{ mit } \sum_{i'=1}^{I} \sum_{j' \in J_{i'}} p_{i'k} q_{j'l} > 0,$$

dann ergibt sich $\bar{y}_{\cdot\cdot}^{C(k,l)} = w_{kl}$ unmittelbar aus den Definitionen. Weiter gilt

$$\overline{\hat{S}^1_{Y_{\cdot\cdot}}}^{C(k,l)} = \frac{\sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \hat{S}^1_{Y,i'j'} q_{j'l}}{\sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} q_{j'l}} = \frac{\sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} w_{kl} q_{j'l}}{\sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} q_{j'l}} = w_{kl}. \qquad (5.5)$$

Damit erhält $\hat{S}_Y^1$ sowohl für vollständige wie für unvollständige Datenmatrizen die Mittelwerte der Partitionen für gegebene P und Q. (Dies würde auch gelten,

wenn man hinsichtlich Formel (5.5) innerhalb der Partition $C(k,l)$ in Bezug auf die Schätzer $\hat{S}_Y^1$ nur die Einträge $\hat{S}_{Y,i'j'}^1$ berücksichtigt, für die $v_{i'j'} = 1$ ist.)

Falls die Daten innerhalb der betrachteten Partitionen die MCAR-Eigenschaft aufweisen, sind die Partitionsmittelwerte unverzerrt. Daher ist ihre Erhaltung durch den Schätzer $\hat{S}_Y^1$ vorteilhaft.

Da Nutzer im allgemeinen mehr Erfahrung und Information in Bezug auf Items haben, die sie präferieren, werden die Nutzer dazu tendieren, eher Bewertungen für von ihnen präferierte Items abzugeben. Dementsprechend sind die Daten nicht MCAR und der Mittelwert aller Bewertungen dürfte nach oben verzerrt sein. Werden die Bewertungen jedoch in Partitionen $C(k,l)$ unterteilt, die nur Bewertungen ähnlicher Nutzer (Nutzer-Cluster k) bezüglich ähnlicher Items (Item-Cluster l) enthalten, so ist es durchaus plausibel, daß das Fehlen von Werten innerhalb der einzelnen Partitionen sowohl unabhängig vom Wert der gegebenen Daten ist als auch nichts mit den unbekannten Ausprägungen der fehlenden Werte zu tun hat. Sofern die durch den Algorithmus identifizierten Nutzer- und Item-Cluster hinreichend homogen sind, wäre das Vorliegen der der MCAR-Eigenschaft innerhalb der Partitionen plausibel aber nicht zwingend. Es ist somit denkbar, daß die Partitionsmittelwerte zumindest weniger starken Verzerrungen als die globalen Mittelwerte unterliegen. Aus diesem Grund ist die Erhaltung der Partitionsmittelwerte eine wesentlicher Vorteil des Schätzers $\hat{S}_Y^1$.

Die Bewertungsdaten werden wie kardinale Daten behandelt. Fehlende Werte werden bei der Klassifikation nicht berücksichtigt. Da die Bewertungsdaten im allgemeinen nicht die MCAR-Eigenschaft aufweisen, kann dies die Partition insbesondere bei hohen Fehlendanteilen verzerren. Es ist fraglich, ob innerhalb dieser verzerrten Partitionen die MCAR-Eigenschaft zumindest näherungsweise erfüllt ist.

Der Schätzer $\hat{S}_Y^2$ (Banerjee et. al. (2004), George, Merugu (2005)) kann als eine Erweiterung des $\hat{S}_Y^1$-Schätzers aufgefaßt werden, die den $\hat{S}_Y^1$-Schätzer beinhaltet. Die Heterogenität der Nutzer wird durch den zusätzlichen Term

$$\begin{aligned} \bar{y}_{i\cdot} - \sum_{k'=1}^{K} p_{ik'} \tilde{w}_{k'\cdot} &= \bar{y}_{i\cdot} - \sum_{k'=1}^{K} p_{ik'} \frac{\sum_{l'=1}^{L} \left(\sum_{\iota=1}^{I} \sum_{j'=1}^{J} p_{\iota k'} v_{\iota j'} q_{j'l'} \right) w_{k'l'}}{\sum_{l'=1}^{L} \left(\sum_{\iota=1}^{I} \sum_{j'=1}^{J} p_{\iota k'} v_{\iota j'} q_{j'l'} \right)} \\ &= \bar{y}_{i\cdot} - \sum_{k'=1}^{K} p_{ik'} \frac{\sum_{l'=1}^{L} a_{k'l'} w_{k'l'}}{\sum_{l'=1}^{L} a_{k'l'}} \end{aligned} \tag{5.6}$$

mit

$$a_{kl} = \sum_{\iota=1}^{I} \sum_{j'=1}^{J} p_{\iota k} v_{\iota j'} q_{j'l}$$

berücksichtigt. Dieser Term (5.6) ist die Differenz zwischen der durchschnittlichen Bewertung des i-ten Nutzers und dem Durchschnitt aller Bewertungen $\tilde{w}_{k.}$, die von Nutzern abgegeben wurden, die zu demselben Nutzer-Cluster k wie i gehören. Dadurch wird dem individuellen Bewertungsverhalten der Nutzer Rechnung getragen. Sofern der betrachtete Nutzer i dazu tendiert, großzügiger zu bewerten als allgemein in seinem Nutzer-Cluster bewertet wird, fällt der obige Term positiv aus. Sind die Bewertungen von Nutzer i hingegen im Durchschnitt niedriger als die Durchschnittsbewertung von Personen aus dem Nutzer-Cluster, zu dem er gehört, ist der obige Term negativ. $a_{kl} = |C(k,l)|$ ist die Anzahl der in der Partition $C(k,l)$ vorhandenen Bewertungen. Analog wird auch der Heterogenität der Items ein Anteil am Schätzer $\hat{S}_Y^2$ zugemessen. Dies geschieht durch Addition von

$$\bar{y}_{.j} - \sum_{l'=1}^{L} q_{jl'} \tilde{w}_{.l'} = \bar{y}_{.j} - \sum_{l'=1}^{L} q_{jl'} \left(\sum_{k'=1}^{K} a_{k'l'} \right)^{-1} \sum_{k'=1}^{K} a_{k'l'} w_{k'l'}. \tag{5.7}$$

Hierdurch wird berücksichtigt, ob das Item j im Durchschnitt höher oder niedriger bewertet wurde als der Durchschnitt aller Bewertungen für Items aus demselben Item-Cluster wie j.
Insgesamt ergibt sich

$$\hat{S}_{Y,ij}^2 = \sum_{k'=1}^{K} \sum_{l'=1}^{L} p_{ik'} w_{k'l'} q_{jl'} + \bar{y}_{i.} - \sum_{k'=1}^{K} p_{ik'} \tilde{w}_{k'.} + \bar{y}_{.j} - \sum_{l'=1}^{L} q_{jl'} \tilde{w}_{.l'}. \tag{5.8}$$

Zur Berechnung wird ein Algorithmus verwendet, der in jeder Iteration zuerst W für gegebene P und Q, danach P für gegebene Q und W und schließlich Q für gegebene P und W optimiert (Banerjee et. al. (2004), George, Merugu (2005)). Die zur Bestimmung von $\hat{S}_Y$ verwendete Zielfunktion ist die gewichtete Frobenius-Norm. W und die auf W basierenden Größen $\tilde{w}_{k.}, k \in \{1, ..., K\}$, und $\tilde{w}_{.l}, l \in \{1, ..., L\}$, werden pro Iteration nur einmal ausgerechnet. Dadurch ist

dieses Verfahren deutlich schneller. Da sich W bei diesem Algorithmus während der Optimierung von P (Q) nicht ändert, genügt es außerdem für jeden Nutzer i (jedes Item j) anstelle der Frobenius-Norm $\mathcal{F}_G$ nur

$$\mathcal{F}_G^{P,x}(p^i, W, \tilde{W}_1, \tilde{W}_2, Q) = \sum_{j'=1}^{J} v_{ij'}(y_{ij'} - \hat{S}^x_{Y,ij'}(p^i, W, \tilde{W}_1, \tilde{W}_2, Q))^2, \qquad x = 1, 2,$$

$$\left(\mathcal{F}_G^{Q,x}(q^j, W, \tilde{W}_1, \tilde{W}_2, P) = \sum_{i'=1}^{I} v_{i'j}(y_{i'j} - \hat{S}^x_{Y,i'j}(q^j, W, \tilde{W}_1, \tilde{W}_2, P)^2, \;\; x = 1, 2, \right)$$

zu betrachten. Hier ist $p^i = (p_{i1}, ..., p_{iK})$ die i-te Zeile von P und $q^j = (q_{j1}, ..., q_{jL})$ ist die j-te Zeile von Q. Weiter gelten $\tilde{W}_1 = (\tilde{w}_{1.}, ..., \tilde{w}_{K.})$ und $\tilde{W}_2 = (\tilde{w}_{.1}, ..., \tilde{w}_{.L})$. Damit erhält man den folgenden Algorithmus:

1. Bestimme P und Q gemäß den Nebenbedingungen.
2. Setze $n = 2$, $P^1 = P$ und $Q^1 = Q$.
3. Wähle $P^0 \neq P^1$ und $Q^0 \neq Q^1$.
4. Falls $x = 2$ ist, berechne $\bar{Y}_{i.}, i = 1, ..., I$, und $\bar{Y}_{.j}, j = 1, ..., J$.
5. Bis $P^{n-1} = P^{n-2}$ und $Q^{n-1} = Q^{n-2}$:
   a) Berechne W^n, $\tilde{W}_1^n$ und $\tilde{W}_2^n$.
   b) $\forall i \in \{1, ..., I\} : p^{i,n} = \arg\min_{p^i \text{u.d.N.}} \mathcal{F}_G^{P,x}(p^i, W^n, \tilde{W}_1^n, \tilde{W}_2^n, Q^{n-1})$
   c) $\forall j \in \{1, ..., J\} : q^{j,n} = \arg\min_{q^j \text{u.d.N.}} \mathcal{F}_G^{Q,x}(q^j, W^n, \tilde{W}_1^n, \tilde{W}_2^n, P^n)$
   d) $n \leftarrow n + 1$

Abbildung 5.6: Variante des AE-Algorithmus nach George, Merugu (2005)

Die Abkürzung „u.d.N.“ weist darauf hin, daß die Zielfunktion unter den in den Formeln ($N1$) und ($N2$) angegebenen Nebenbedingungen zu minimieren ist.

Der Algorithmus von George, Merugu (2005) kann als einfachere Variante des AE-Algorithmus aufgefaßt werden. Da die Clusterzugehörgkeiten aller Nutzer und Items in Schritt 5b) und 5c) auf Basis derselben in Schritt 5a) bestimmten Matrix W berechnet werden, ist diese Variante des AE-Algorithmus deutlich schneller. Es existiert bereits eine hocheffiziente parallele Version dieses Algorithmus (George,

Merugu (2005)). Fehlende Werte werden ignoriert. Die Bewertungsdaten werden wie Daten mit kardinalem Skalenniveau behandelt.

Beispiel 5.12:

Das Ziel ist die Berechnung von $\hat{Y}_{17}$ mittels $\hat{S}_Y^2$. Für die Anzahlen der Cluster erster und zweiter Modalität werden wieder $K = 2$ und $L = 3$ gewählt. Zu Beginn des Verfahrens werden durch Zufallszahlen Matrizen P^1 und Q^1 erzeugt, die den oben angegebenen Nebenbedingungen gerecht werden:

$$P^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{und} \quad Q^1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}'.$$

Auf Basis von P^1 und Q^1 kann W^1 berechnet werden:

$$W^1 = \begin{pmatrix} 3{,}13 & 2{,}20 & 4{,}20 \\ 3{,}57 & 4{,}13 & 2{,}00 \end{pmatrix}.$$

Als Zielfunktionswert ergibt sich $\mathcal{F}_G = 29,06$. Man geht nun genauso vor wie in Beispiel 5.11, verwendet dabei aber statt $\hat{S}_Y^1$ den Schätzer $\hat{S}_Y^2$. Auf diese Weise erhält man Cluster-Zugehörigkeitsmatrizen für Elemente erster und zweiter Modalität

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{und} \quad Q = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}'.$$

Nach entsprechender Umordnung der Zeilen und Spalten erhält man damit die folgenden Partitionen:

		$l=1$		$l=2$			$l=3$				
k	i	$j=1$	$j=8$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$\bar{y}_{i.}$	$\tilde{w}_{k.}$
	2	-	3	4	2	2	2	5	-	3	
1	4	5	4	2	2	-	2	3	-	3	3,17
	5	4	4	-	2	3	4	4	-	$\frac{21}{6}$	
	1	1	-	5	4	4	3	4	-	$\frac{21}{6}$	
2	3	-	1	4	4	4	3	-	5	$\frac{21}{6}$	3,56
	6	1	-	5	4	4	4	-	4	$\frac{22}{6}$	
$\bar{y}_{.j}$		2,75	3	4	3	3,4	3	4	4,5		
$\tilde{w}_{.l}$		$\frac{23}{8}$			3,44			3,58			

Abbildung 5.7: Zweimodale Cluster (Beispiel 5.12)

Mit

$$W = \begin{pmatrix} 4{,}00 & 2{,}49 & 3\frac{1}{3} \\ 1{,}00 & 4{,}22 & 3{,}83 \end{pmatrix}$$

erhält man

$$\begin{aligned} \hat{Y}_{17} = \hat{S}^2_{Y,17} &= w_{23} + \bar{y}_{1.} + \bar{y}_{.7} - \tilde{w}_{2.} - \tilde{w}_{.3} \\ &= 3,83 + 3,5 + 4,5 - 3,56 - 3,58 = 4,70. \end{aligned}$$

Für dieselben Anfangsmatrizen P^1 und Q^1 kommen sowohl der ursprüngliche AE-Algorithmus als auch seine Variation nach George, Merugu (2005) unter Verwendung des $\hat{S}^2_Y$-Schätzers zu diesem Ergebnis. Die Tatsache, daß beide Algoritmen hier gleiche Ergebnisse produzieren, legt nahe, daß der mögliche Genauigkeitsverlust durch den schnelleren Algorithmus zumindest akzeptabel sein könnte.

Die Verwendung des Schätzers $\hat{S}^2_Y$ führt bei einem bestimmten Datensatz zu kleineren AAD-Werten als die Anwendung des verallgemeinerten ADCLUS-Schätzers

$\hat{S}_Y^1$ (Banerjee et. al. (2004)). Da der verwendete Datensatz recht klein war (500 Nutzer, 200 Items) und zudem die AAD-Werte innerhalb gewisser Schranken von Datensatz zu Datensatz variieren, sollte hieraus noch nicht die Überlegenheit des $\hat{S}_Y^2$-Schätzers gefolgert werden. Vielmehr sind weitere empirische Vergleiche der beiden Schätzer erforderlich. George, Merugu (2005) konnten anhand eines etwas größeren Datensatzes zeigen, daß das zweimodale $\hat{S}_Y^2$-Clusterverfahren in Bezug auf die Genauigkeit zu sehr ähnlichen Ergebnissen wie ein Singulärwertzerlegung-basierter Ansatz (Sarwar et. al. (2000b)) und das Gradientenverfahren nach Srebro, Jaakkola (2003) führt. George, Merugu (2005) weisen überdies darauf hin, daß der zweimodale Clusteransatz in ihrer Implementation weniger Rechenzeit benötigt als die genannten Verfahren.

Man kann zeigen, daß der Schätzer $\hat{S}_Y^2$ bei vollständigen Datenmatrizen Y nicht nur die Partitionsmittelwerte sondern auch die Zeilen- und Spaltenmittelwerte erhält. Leider gilt das nicht für unvollständige Datematrizen.

Um zu zeigen, daß der Schätzer $\hat{S}_Y^2$ die Partitionsmittelwerte einer vollständigen Matrix Y erhält, müssen die letzten vier Terme von

$$\overline{\hat{S}_{Y_{..}}^{2}}^{C(k,l)} \left(\sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} q_{j'l} \right) = \sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} w_{kl} q_{j'l} + \sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \bar{y}_{i'.} q_{j'l}$$

$$- \sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \tilde{w}_{k.} q_{j'l} + \sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \bar{y}_{.j'} q_{j'l} - \sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \tilde{w}_{.l} q_{jl'}$$

wegfallen. Falls in der Matrix Y keine Einträge fehlen, gilt

$$\sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \bar{y}_{i'.} q_{j'l} = \left(\sum_{j'=1}^{J} q_{j'l} \right) \sum_{i'=1}^{I} p_{i'k} \bar{y}_{i'.} = \left(\sum_{j'=1}^{J} q_{j'l} \right) \tilde{w}_{k.} \sum_{i'=1}^{I} p_{i'k}.$$

Unter der Bedingung, daß keine Einträge in Y fehlen, ist dies genau gleich dem dritten Term. Analog zeigt man die Äquivalenz des vierten und fünften Terms. Somit erhält $\hat{S}_Y^2$ die Partitionsmittelwerte vollständiger Datenmatrizen.

Das folgende Beispiel zeigt, daß diese Beziehung nicht gelten muß, falls Werte von y fehlen.

Beispiel 5.13:

Die Schätzermatrix $\hat{S}_Y^2$ aus Beispiel 5.12 ist:

		$l=1$		$l=2$			$l=3$				
k	i	$j=1$	$j=8$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$\bar{y}_{i.}$	$\tilde{w}_{k.}$
	2	3,708	3,958	2,824	1,824	2,224	2,583	3,583	4,083	3	
1	4	3,708	3,958	2,824	1,824	2,224	2,583	3,583	4,083	3	3,17
	5	4,208	4,458	3,324	2,324	2,724	3,083	4,083	4,583	$\frac{21}{6}$	
	1	0,819	1,069	4,729	3,729	4,129	3,194	4,194	4,694	$\frac{21}{6}$	
2	3	0,819	1,069	4,729	3,729	4,129	3,194	4,194	4,694	$\frac{21}{6}$	3,56
	6	0,986	1,236	4,896	3,896	4,296	3,361	4,361	4,861	$\frac{22}{6}$	
$\bar{y}_{.j}$		2,75	3	4	3	3,4	3	4	4,5		
$\tilde{w}_{.l}$		$\frac{23}{8}$			3,44			3,58			

Abbildung 5.8: $\hat{S}_Y^2$-Schätzer und Partitionen aus Beispiel 5.13

Als Schätzermittelwert der Partition $C(k,l)$ ergibt sich $\overline{\hat{S}_{Y_{..}}^{2\,C(1,3)}} = 3,58 \neq w_{13}$. Falls man die Erhaltung der Partitionsmittelwerte nur für Schätzer $\hat{S}_{Y,ij}^2$ fordert, für die $v_{ij} = 1$ gilt, würde man als Partitionsmittelwert den Schätzer

$$\overline{\hat{S}_{Y_{..}}^{2\,C(1,3),V}} = \frac{1}{6}\sum_{i'=1}^{I}\sum_{j'=1}^{J} p_{i'1}\hat{S}_{Y,i'j'}^2 v_{i'j'}q_{j'3} = 3,25 \neq w_{13}$$

erhalten. Damit ist gezeigt, daß der Schätzer $\hat{S}_Y^2$ nicht alle Partitionsmittelwerte der Ausgangsmatrix y erhält, sofern Einträge in der Ausgangsmatrix fehlen. Alternativ kann man die letzten vier Terme von $\overline{\hat{S}_{Y_{..}}^{2\,C(1,3),V}}$ betrachten.

Wegen

$$\begin{aligned}&\textstyle\sum_{i',j'} p_{i'1}\bar{y}_{i'.}v_{i'j'}q_{j'3} - \sum_{i',j'} p_{i'1}\tilde{w}_{1.}V_{i'j'}q_{j'3} + \sum_{i',j'} p_{i'1}\bar{y}_{.j'}v_{i'j'}q_{i'3}\\ &\textstyle -\tilde{w}_{.3}\sum_{i',j'} p_{i'1}v_{i'j'}q_{j'3} = 19 - 19 + 21 - 21,48 = -0,48 \neq 0\end{aligned}$$

folgt ebenfalls, daß $\overline{\hat{S}_{Y_{..}}^{C(1,3),V}} \neq w_{13}$ gilt.

Außerdem erhält $\hat{S}_Y^2$ die Zeilen- und Spaltensummen vollständiger Datenmatrizen Y. Ohne fehlende Werte gelten

$$\tilde{w}_{.l} = \frac{\sum_{i'=1}^{I}\sum_{j'=1}^{J}\sum_{k'=1}^{K} p_{i'k'} y_{i'j'} q_{j'l}}{\sum_{i'=1}^{I}\sum_{j'=1}^{J}\sum_{k'=1}^{K} p_{i'k'} q_{j'l}} \quad \text{und} \quad \bar{y}_{..} = \frac{\sum_{i'=1}^{I}\sum_{j'=1}^{J}\sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{i'k'} y_{i'j'} q_{j'l'}}{\sum_{i'=1}^{I}\sum_{j'=1}^{J}\sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{i'k'} q_{j'l'}},$$

woraus

$$J\bar{y}_{..} = \sum_{j'=1}^{J}\sum_{l'=1}^{L} q_{j'l'}\tilde{w}_{.l'}$$

folgt. Zudem gilt für vollständige Datenmatrizen Y

$$\tilde{w}_{k.} = \frac{\sum_{i'=1}^{I}\sum_{j'=1}^{J}\sum_{l'=1}^{L} w_{kl'} p_{i'k} q_{j'l'}}{\sum_{i'=1}^{I}\sum_{j'=1}^{J}\sum_{l'=1}^{L} p_{i'k} q_{j'l'}} = \frac{\sum_{i'=1}^{I} p_{i'k} \sum_{j'=1}^{J}\sum_{l'=1}^{L} w_{kl'} q_{j'l'}}{\sum_{i'=1}^{I} p_{i'k} J} = \frac{1}{J}\sum_{j'=1}^{J}\sum_{l'=1}^{L} w_{kl'} q_{j'l'}.$$

Daraus folgt für Matrizen ohne fehlende Einträge

$$\sum_{j'=1}^{J}\sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{ik'} w_{k'l'} q_{j'l'} = \sum_{k'=1}^{K} p_{ik'} \sum_{j'=1}^{J}\sum_{l'=1}^{L} w_{k'l'} q_{j'l'} = J\sum_{k'=1}^{K} p_{ik'}\tilde{w}_{k'.}.$$

Insgesamt ergibt sich deshalb, daß der Schätzer $\hat{S}_Y^2$ die Zeilenmittelwerte vollständiger Matrizen Y erhält, da sich unter der Voraussetzung, daß in Y keine Einträge fehlen, alle Terme außer $J\bar{y}_{i.}$ in der Summe

$$\begin{aligned}\sum_{j'=1}^{J} \hat{S}_{Y,ij'}^2 &= \sum_{j'=1}^{J}\sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{ik'} w_{k'l'} q_{j'l'} + J\bar{y}_{i.} - J\sum_{k'=1}^{K} p_{ik'}\tilde{w}_{k'.} + J\bar{y}_{..} - \sum_{j'=1}^{J}\sum_{l'=1}^{L} q_{j'l'}\tilde{w}_{.l'} \\ &= J\bar{y}_{i.}.\end{aligned}$$

gegenseitig aufheben. Analog kann gezeigt werden, daß $\hat{S}_Y^2$ die Spaltenmittelwerte von Datenmatrizen erhält, sofern keine Einträge dieser Matrizen fehlen. Beides gilt nicht für unvollständige Matrizen.

Beispiel 5.14:

Mit der Schätzermatrix $\hat{S}_Y^2$ aus Beispiel 5.12 erhält man für die dritte Zeile der Matrix y:

$$\frac{1}{8}\sum_{j'=1}^{8}\hat{S}^2_{Y,3j'} = 3,32 \neq \bar{y}_{3.} \quad \text{und} \quad \left(\sum_{j'=1}^{8} v_{3j'}\right)^{-1}\sum_{j'=1}^{8} v_{3j'}\hat{S}^2_{Y,3j'} = 3,59 \neq \bar{y}_{3.}.$$

Für die erste Spalte von Y gilt

$$\frac{1}{6}\sum_{i'=1}^{6}\hat{S}^2_{Y,i'1} = 2,38 \neq \bar{y}_{.1} \quad \text{und} \quad \left(\sum_{i'=1}^{6} v_{i'1}\right)^{-1}\sum_{i'=1}^{6} v_{i1}\hat{S}^2_{Y,i'1} = 2,43 \neq \bar{y}_{.1}.$$

Aus Beispiel 5.14 folgt, daß $\hat{S}_Y^2$ die Zeilen- und Spaltensummen unvollständiger Matrizen y nicht generell erhält. Daher erhält der komplizierte Schätzer $\hat{S}_Y^2$ für Datenmatrizen y, in denen Einträge fehlen, sogar weniger Information aus y als $\hat{S}_Y^1$. Dafür ist $\hat{S}_Y^2$ in der Lage, die Heterogenität von Nutzern und Items zu berücksichtigen.

Die zweimodalen Clusterverfahren sind ein Spezialfall der Imputation innerhalb von Zellen („adjustment cells“).

Der Schätzer $\hat{S}_Y^1$ ist ein Spezialfall der Mittelwertimputation innerhalb von Zellen. Wie in Kapitel 3 erörtert, setzt die Mittelwert-Imputation die MCAR-Eigenschaft voraus. Sofern die zur betrachteten Partition gehörenden Cluster erster und zweiter Modalität hinreichend homogen sind, ist es zwar nicht zwingend aber doch plausibel, daß die MCAR-Eigenschaft innerhalb von Klassen erfüllt ist. Daher ist der $\hat{S}_Y^1$-Schätzer auch vor dem Hintergrund der in Kapitel 3 dargestellten Überlegungen ein sinnvoller Schätzer. Problematisch ist hierbei nur, daß die Partitionen nur aufgrund der vorhandenen Werte gebildet werden können. Bei hohen Fehlendstrukturen könnte dies die reale Homogenität der Nutzer- und Item-Cluster beeinträchtigen und dadurch zu (stärker) verzerrten Schätzern führen.

Im Schätzer $\hat{S}_Y^2$ werden gewichtete Mittelwerte der Partitionsmittelwerte sowie die Zeilen- und Spaltenmittelwerte verwendet. Dabei wurde mit der Anzahl von in der jeweiligen Partition $C(k,l)$ vorhandenen Bewertungen

$$a_{kl} = \sum_{\iota=1}^{I} \sum_{h \in J_\iota} p_{\iota k} q_{hl}$$

gewichtet:

$$\tilde{w}_{k.} = \left(\sum_{l'=1}^{L} a_{kl'} \right)^{-1} \sum_{l'=1}^{L} a_{kl'} w_{kl'} \quad \text{und} \quad \tilde{w}_{.l} = \left(\sum_{k'=1}^{K} a_{k'l} \right)^{-1} \sum_{k'=1}^{K} a_{k'l} w_{k'l}.$$

Auch $\bar{y}_{i.}$ und $\bar{y}_{.j}$ können als auf ähnliche Weise gewichtete Mittelwerte aufgefaßt werden. Mit

$$\zeta_{il} = \sum_{h=1}^{J} v_{ih} q_{hl}, \quad \xi_{jk} = \sum_{z=1}^{I} v_{zj} p_{zk},$$

und

$$\bar{y}_{.j}^{C_1(k)} = \frac{\sum_{\iota=1}^{I} y_{\iota j} v_{\iota j} p_{\iota k}}{\sum_{\iota=1}^{I} v_{\iota j} p_{\iota k}}$$

gilt

$$\bar{y}_{i.} = \frac{\sum_{l'=1}^{L} \zeta_{il'} \bar{y}_{i.}^{C_2(l')}}{\sum_{l'=1}^{L} \zeta_{il'}} \quad \text{und} \quad \bar{y}_{.j} = \frac{\sum_{k'=1}^{K} \xi_{jk'} \bar{y}_{.j}^{C_1(k')}}{\sum_{k'=1}^{K} \xi_{jk'}}.$$

Wie an Beispiel 3.8 illustriert wurde, können solche mit den Anzahlen der in der betrachteten Zelle (bzw. Partition) vorhandenen Daten gewichtete Mittelwerte verzerrt sein, falls nicht für alle fehlenden Werte in der gesamten Datenmatrix

y die MCAR-Eigenschaft gilt. Der Grund hierfür ist, daß sich die Fehlendstrukturen verschiedener Zellen (bzw. Partitionen) stark unterscheiden können, sofern die Daten nicht zufällig fehlen. Hierdurch kann es zur Übergewichtung von Zellen (Partitionen) mit vergleichsweise niedriger Fehlendstruktur bei gleichzeitiger Untergewichtung von Zellen (Partititionen) mit überdurchschnittlichem Fehlendanteil kommen. Es ist in solchen Fällen empfehlenswert, mit der Anzahl von Zellenelementen (Partitionselementen) zu gewichten, die vorliegen würde, wenn die Datenmatrix vollständig wäre (Little (1986)). Somit sollten anstelle von a_{kl}, ζ_{il} und ξ_{jk} die Gewichtungen

$$a^*_{kl} = \sum_{\iota=1}^{I}\sum_{h=1}^{J} p_{\iota k} q_{hl}, \quad \zeta^*_l = \sum_{h=1}^{J} q_{hl} \quad \text{und} \quad \xi^*_k = \sum_{z=1}^{I} p_{zk}$$

verwendet werden. Damit ergeben sich

$$\bar{y}^*_{i.} = \frac{\sum_{l'=1}^{L} \zeta^*_{l'} \bar{y}^{C_2(l')}_{i.}}{\sum_{l'=1}^{L} \zeta^*_{l'}} \quad \text{und} \quad \bar{y}^*_{.j} = \frac{\sum_{k'=1}^{K} \xi^*_{k'} \bar{y}^{C_1(k')}_{.j}}{\sum_{k'=1}^{K} \xi^*_{k'}},$$

sowie

$$\tilde{w}^*_{k.} = \left(\sum_{l'=1}^{L} a^*_{kl'}\right)^{-1} \sum_{l'=1}^{L} a^*_{kl'} w_{kl'} \quad \text{und} \quad \tilde{w}^*_{.l} = \left(\sum_{k'=1}^{K} a^*_{k'l}\right)^{-1} \sum_{k'=1}^{K} a^*_{k'l} w_{k'l}.$$

Zur Berechnung der entsprechenden Schätzer

$$\hat{S}^{2*}_{Y,ij} = \sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{ik'} w_{k'l'} q_{jl'} + \bar{y}^*_{i.} - \sum_{k'=1}^{K} p_{ik'} \tilde{w}^*_{k'.} + \bar{y}^*_{.j} - \sum_{l'=1}^{L} q_{jl'} \tilde{w}^*_{.l'} \tag{5.9}$$

kann die in Abbildung 5.9 dargestellte neue Variante des AE-Algorithmus benutzt werden. Dabei werden die Werte $q^{j,n-1} = (q^{n-1}_{j1}, ..., q^{n-1}_{jL}), j \in \{1, ..., J\}$ aus der $(n-1)$-ten Iteration benutzt, um

$$\zeta^{*n}_l = \sum_{h=1}^{J} q^{n-1}_{hl} \quad \text{und} \quad \bar{y}^{C_2(l),n}_{i.} = \frac{\sum_{h=1}^{J} y_{ih} v_{ih} q^{n-1}_{hl}}{\sum_{h=1}^{J} q^{n-1}_{hl} v_{ih}}$$

1. Bestimme P und Q gemäß den Nebenbedingungen.
2. Setze $n = 2$, $P^1 = P$ und $Q^1 = Q$.
3. Wähle $P^0 \neq P^1$ und $Q^0 \neq Q^1$.
4. Berechne $\forall j \in \{1, ..., J\} : \bar{y}_{.j}^{*1}$.
5. Bis $P^{n-1} = P^{n-2}$ und $Q^{n-1} = Q^{n-2}$:
   a) Berechne W^n, $\tilde{W}_1^{*n}$ und $\tilde{W}_2^{*n}$ auf Basis von P^{n-1} und Q^{n-1}.
   b) $\forall l \in \{1, ..., L\} : \zeta_l^{*n} = \sum_{h=1}^{J} q_{hl}^{n-1}$
   c) $\forall i \in \{1, ..., I\} :$
      (i) Berechne $\bar{y}_{i.}^{*n}$.
      (ii) Bestimme
      $p^{i,n} = \arg\min_{p^i \text{u.d.N.}} \mathcal{F}_G^{P,x}(p^i, W^n, \tilde{W}_1^n, \tilde{W}_2^n, Q^{n-1}, \bar{y}_{i.}^{*n}, \bar{y}_{col}^{*n-1})$.
   d) $\forall k \in \{1, ..., K\} : \xi_k^{*n} = \sum_{z=1}^{I} p_{zk}^n$
   e) $\forall j \in \{1, ..., J\} :$
      (i) Berechne $\bar{y}_{.j}^{*n}$.
      (ii) Bestimme
      $q^{j,n} = \arg\min_{q^j \text{u.d.N.}} \mathcal{F}_G^{Q,x}(q^j, W^n, \tilde{W}_1^n, \tilde{W}_2^n, P^n, \bar{y}_{.j}^{*n}, \bar{y}_{row}^{*n})$.
   f) $n \leftarrow n + 1$

Abbildung 5.9: Variante des AE-Algorithmus zur Vermeidung von Verzerrungen

zu bestimmen. Beides fließt in die Berechnung von

$$\bar{y}_{i.}^{*n} = \frac{\sum_{l'=1}^{L} \zeta_{l'}^{*n} \bar{y}_{i.}^{C_2(l'),n}}{\sum_{l'=1}^{L} \zeta_{l'}^{*n}} \quad \text{und} \quad \bar{y}_{row}^{*n} = (\bar{y}_{1.}^{*n}, ..., \bar{y}_{I.}^{*n})$$

ein und wird dazu benutzt, neue Werte $p^{i,n} = (p_{i1}^n, ..., p_{iK}^n), i = 1, ..., I$, zu ermitteln. Diese werden dann wiederum in den Formeln

$$\xi_k^n = \sum_{z=1}^{I} p_{zk}^n \quad \text{und} \quad \bar{y}_{.j}^{C_1(k),n} = \frac{\sum_{z=1}^{I} p_{zk}^n y_{zj} v_{zj}}{\sum_{z=1}^{I} p_{zk}^n v_{zj}}$$

verwendet, woraus sich schließlich

$$\bar{y}_{.j}^{*n} = \frac{\sum_{k'=1}^{K} \xi_{k'}^{*n} \bar{y}_{.j}^{C_1(k'),n}}{\sum_{k'=1}^{K} \xi_{k'}^{*n}} \quad \text{und} \quad \bar{y}_{col}^{*n} = (\bar{y}_{.1}^{*n}, ..., \bar{y}_{.J}^{*n})$$

ergibt. In den Schritten 5 c) (ii) und 5 e) (ii) ist die Zielfunktion unter den Nebenbedingungen (u.d.N.) $(N1)$ und $(N2)$ zu minimieren.

Da in der Literatur die Unterteilung in Test- und Trainingsdatensatz per Zufallszahlen vorgenommen wird (z.B. George, Merugu (2005)), ist die MCAR-Eigenschaft dort tatsächlich erfüllt. Daher führt die Verwendung von $\hat{S}_Y^2$ bei den dargestellten Evaluationen nicht zu verzerrten Schätzern. In der Praxis ist jedoch davon auszugehen, daß die MCAR-Eigenschaft nicht vorliegt. Daher ist fraglich, ob man nicht $\hat{S}_Y^{2*}$ anstelle von $\hat{S}_Y^2$ verwenden sollte. Außerdem sollten allgemein alle Verfahren auch anhand von Test- und Trainingsdatensätzen evaluiert werden, für die die MCAR-Annahme nicht erfüllt ist. Es ist zu beachten, daß die Gefahr möglicher Verzerrungen durch Ignorieren der fehlenden Daten keineswegs eine Besonderheit des $\hat{S}_Y^2$-Ansatzes ist. Es ist vielmehr eine Besonderheit des $\hat{S}_Y^{2*}$-Ansatzes, daß er versucht, der Gefahr möglicher Verzerrungen entgegenzuwirken.

Um zu sehen, ob der Schätzer $\hat{S}_Y^{2*}$ (5.9) die Partitionsmittelwerte der unvollständigen Datenmatrix näherungsweise erhält, betrachtet man wieder die Summe

$$\sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \hat{S}_{Y,i'j'}^{2*} q_{j'l}.$$

Diese Summe zerfällt in fünf Terme. Der zweite und dritte Term dieser Summe sind

$$\sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \bar{y}_{i'.}^{*} q_{j'l} = \zeta_l^* \sum_{i'=1}^{I} p_{i'k} \bar{y}_{i'.}^{*} \quad \text{und} \quad \sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \sum_{h=1}^{K} p_{i'h} \tilde{w}_{h.}^{*} q_{j'l} = \zeta_l^* \sum_{i'=1}^{I} p_{i'k} \tilde{w}_{k.}^{*}.$$

Unter Voraussetzung der MCAR-Eigenschaft innerhalb der Partitionen gilt

$$\sum_{i'=1}^{I} p_{i'k} \bar{y}_{i'.}^{*} \approx \sum_{i'=1}^{I} p_{i'k} \tilde{w}_{k.}^{*}.$$

Analog kann man zeigen, daß sich der vierte und fünfte Term der Summe gegenseitig näherungsweise aufheben, sofern die fehlenden Daten innerhalb der Partitionen zufällig fehlen. Daher erhält der Schätzer $\hat{S}_Y^{2*}$ unter Voraussetzung der MCAR-Eigenschaft innerhalb der Partitionen die Partitionsmittelwerte in guter Näherung.

Eine weitere Möglichkeit wäre die Verwendung des zweimodalen Schätzers $\hat{S}_Y^3$ nach Cheng, Church (2000):

$$\hat{S}_{Y,ij}^3 = \sum_{k'=1}^{K} \sum_{l'=1}^{L} p_{ik'} (\bar{y}_{i.}^{C_2(l')} + \bar{y}_{.j}^{C_1(k')} - w_{k'l'}) q_{jl'}$$

Dieser Schätzer wurde bisher gar nicht zur Schätzung von Bewertungsdaten benutzt. Da dieser Schätzer nur Mittelwerte aus Elementen der betrachteten Partitionen verwendet, verringert er genau wie $\hat{S}_Y^1$ die Gefahr der Verzerrung durch nicht zufällig fehlende Daten. Gleichzeitig berücksichtigt er wie die Schätzer $\hat{S}_Y^2$ und $\hat{S}_Y^{2*}$ die Heterogenität der Nutzer und Items. Sofern das Fehlen von Daten innerhalb der Partitionen zufällig (MCAR) ist, gilt

$$\sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} \hat{S}_{Y,i'j'}^3 q_{j'l} = \sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} q_{j'l} \left(\bar{y}_{i'.}^{C_2(l)} + \bar{y}_{.j'}^{C_1(k)} - w_{kl} \right) \approx \sum_{i'=1}^{I} \sum_{j'=1}^{J} p_{i'k} w_{kl} q_{j'l}.$$

Somit erhält $\hat{S}_Y^3$ unter Voraussetzung der MCAR-Eigenschaft innerhalb der Partitionen die Partitionsmittelwerte vollständiger und unvollständiger Datenmatrizen. Weil unter der Bedingung, daß das Fehlen innerhalb der Partitionen zufällig ist,

$$\sum_{i'=1}^{I} \hat{S}_{Y,i'j}^3 p_{i'k} = \sum_{i'=1}^{I} \sum_{l'=1}^{L} p_{i'k} q_{jl'} \bar{y}_{.j}^{C_1(k)} + \underbrace{\sum_{i'=1}^{I} p_{i'k} \left(\sum_{l'=1}^{L} q_{jl'} \bar{y}_{i'.}^{C_2(l')} - \sum_{l'=1}^{L} q_{jl'} w_{kl'} \right)}_{\approx 0}$$

gilt, sind auch die Spaltenmittelwerte innerhalb der Partitionen unter der genannten Voraussetzung in guter Näherung erhalten. Dasselbe gilt für die Zeilenmittelwerte innerhalb der Partitionen.

Die dargestellten Eigenschaften des Schätzers von Cheng, Church (2000) legen seine Verwendung zum Schätzen der fehlenden Werte unvollständiger Datenmatrizen nahe. Zu diesem Zweck kann wieder eine Variante des AE-Algorithmus benutzt werden, bei der nacheinander W, P und Q optimiert werden.

1. Bestimme P und Q gemäß den Nebenbedingungen.
2. Setze $n = 2$, $P^1 = P$ und $Q^1 = Q$.
3. Wähle $P^0 \neq P^1$ und $Q^0 \neq Q^1$.
4. $\forall j \in \{1, ..., J\}, \forall k \in \{1, ..., K\} : \bar{y}_{.j}^{C_1(k),1} = (\sum_{i \in I_j} p_{ik}^1)^{-1} \sum_{i \in I_j} p_{ik}^1 y_{ij}$
5. Bis $P^{n-1} = P^{n-2}$ und $Q^{n-1} = Q^{n-2}$:
   a) Berechne W^n.
   b) $\forall i \in \{1, ..., I\}, \forall l \in \{1, ..., L\} : \bar{y}_{i.}^{C_2(l),n} = (\sum_{j \in J_i} q_{jl}^{n-1})^{-1} \sum_{j \in J_i} q_{jl}^{n-1} y_{ij}$
   c) $\forall i \in \{1, ..., I\} : p^{i,n} = \arg\min_{p^i \text{u.d.N.}} \mathcal{F}_G^{P,3}(p^i, W^n, \bar{y}_1^n, \bar{y}_2^{n-1}, Q^{n-1})$
   d) $\forall j \in \{1, ..., J\}, \forall k \in \{1, ..., K\} : \bar{y}_{.j}^{C_1(k),n} = (\sum_{i \in I_j} p_{ik}^n)^{-1} \sum_{i \in I_j} p_{ik}^n y_{ij}$
   e) $\forall j \in \{1, ..., J\} : q^{j,n} = \arg\min_{q^j \text{u.d.N.}} \mathcal{F}_G^{Q,3}(q^j, W^n, \bar{y}_1^n, \bar{y}_2^n, P^n)$
   f) $n \leftarrow n + 1$

Abbildung 5.10: AE-Variante zur Berechnung von $\hat{S}_Y^3$

Dabei sind

$$\bar{y}_1^n = \begin{pmatrix} \bar{y}_{1.}^{C_2(1),n} & \cdots & \bar{y}_{1.}^{C_2(L),n} \\ \vdots & \ddots & \vdots \\ \bar{y}_{I.}^{C_2(1),n} & \cdots & \bar{y}_{I.}^{C_2(L),n} \end{pmatrix} \quad \text{und} \quad \bar{y}_2^n = \begin{pmatrix} \bar{y}_{.1}^{C_1(1),n} & \cdots & \bar{y}_{.1}^{C_1(K),n} \\ \vdots & \ddots & \vdots \\ \bar{y}_{.J}^{C_1(1),n} & \cdots & \bar{y}_{.J}^{C_1(K),n} \end{pmatrix}.$$

In den Schritten 5 c) und 5 e) müssen bei der Optimierung der Zielfunktion wieder die Nebenbedingungen $(N1)$ und $(N2)$ beachtet werden. Dazu verwendet man

$$\mathcal{F}_G^{P,3}(p^i, W^n, \bar{y}_1^n, \bar{y}_2^{n-1}, Q^{n-1}) = \sum_{j' \in J_i} q_{j'l}^{n-1} \left(y_{ij'} - \hat{S}_{Y,ij'}^{3,P,n} \right)^2$$

mit

$$\hat{S}_{Y,ij}^{3,P,n} = \sum_{k'=1}^{K} \sum_{l'=1}^{L} p_{ik'} \left(\bar{y}_{i.}^{C_2(l'),n} + \bar{y}_{.j}^{C_1(k'),n-1} - w_{k'l'}^n \right) q_{jl'}^{n-1}$$

und

$$\mathcal{F}_G^{Q,3}(q^j, W^n, \bar{y}_1^n, \bar{y}_2^n, P^n) = \sum_{i' \in I_j} p_{i'k}^n \left(y_{i'j} - \hat{S}_{Y,i'j}^{3,Q,n} \right)^2$$

mit

$$\hat{S}_{Y,ij}^{3,Q,n} = \sum_{k'=1}^{K} \sum_{l'=1}^{L} q_{jl'} \left(\bar{y}_{i.}^{C_2(l'),n} + \bar{y}_{.j}^{C_1(k'),n} - w_{k'l'}^n \right) p_{ik'}^n.$$

Wie üblich bricht das Verfahren ab, sobald sich die Clusterzugehörigkeiten nicht mehr ändern.

5.7 Imputation innerhalb zweimodaler Partitionen

Der $\hat{S}_Y^1$-Schätzer ist ein Spezialfall der Imputation innerhalb von Klassen. Neben der klasseninternen Mittelwertimputation gibt es in der Literatur wie in Kapitel 3 bereits erwähnt sogenannte klasseninterne Hot-Deck Verfahren (Hanson (1978), Brick, Kalton (1996)). Dabei wird meist eine innerhalb der betrachteten Klasse vollständige Zeile oder Spalte benutzt, um die innerhalb der betrachteten Klasse fehlenden Werte zu approximieren.

Zweimodale Clusterverfahren sind unter Umständen dazu geeignet, zweimodale Daten in sinnvolle Klassen (nämlich die Partitionen) zu unterteilen. Zudem sind zweimodale Clusterverfahren auch auf sehr große Datensätze anwendbar.

Um alle Schätzer für die gesamte Datenmatrix y zu berechnen, müssen alle Partitionen der Matrix nacheinander abgearbeitet werden, da jede einzelne Partition nur einen Teil der zweimodalen Matrix enthält.

Neben der Methode der Imputation des Partitionsmittelwerts, gibt es auch andere einfache Möglichkeiten, um die Partitionen zur Imputation der fehlenden Werte zu benutzen.

Denkbar wäre eine randomisierte Hot-Deck Imputation immerhalb der Partitionen (RHDP). Wegen der hohen Fehlendstrukturen sind die in Kapitel 3 vorgestellten Hot-Deck Verfahren innerhalb von „adjustment cells“ nicht übertragbar. Es wird sich nur selten eine Partition finden lassen, in der überhaupt zu einem

Element erster oder zweiter Modalität alle Werte in der Partition gegeben sind. Statt dessen könnte für jeden fehlenden Wert y_{ij} per Zufallsziehung ein Wert y_{iz} aus der i-ten Zeile und derselben Partition $C(k,l)$ wie y_{ij} imputiert werden $(y_{ij}, y_{iz} \in C(k,l))$. Sofern kein Wert in der i-ten Zeile der Partition $C(k,l)$ gegeben ist, wird zufällig ein Wert $y_{xj} \in C(k,l)$ gewählt, der in derselben Spalte wie $y_{ij} (\in C(k,l))$ steht.

Alternativ hierzu wäre auch eine korrelationsbasierte Hot-Deck Imputation innerhalb von Partionen möglich (KHDP). Für jeden fehlenden Wert $y_{ij} \in C(k,l)$ wird der Wert $y_{ij_{max}} \in C(k,l)$ imputiert, wobei $j_{max} = \arg\max_{j_1 \in C_2(l)} \tilde{r}_{jj_1}^{Matthai}$ gilt. $\tilde{r}_{jj_1}^{Matthai}$ ist der Bravais-Pearson'sche Korrelationskoeffizient für Items mit der Modifikation nach Matthai. Falls kein Wert in derselben Zeile i aus der Partition $C(k,l)$ vorhanden ist, verwendet man $y_{i_{max}j} \in C(k,l)$ anstelle des fehlenden Werts $y_{ij} \in C(k,l)$, wobei $i_{max} = \arg\max_{z \in C_1(k)} r_{iz}^{Matthai}$ ist, wobei diesmal der Bravais-Pearson'sche Korrelationskoeffizient für Nutzer nach Matthai $r_{iz}^{Matthai}$ verwendet wird.

Bei nicht zu großen Fehlendanteil könnte man auch ein Ähnlichkeitsverfahren innerhalb von Partitionen (SP) einsetzen. Da aber durch die Beschränkung auf Partitionen die einzelnen Korrelationen auf noch wenigen Daten basieren würden, ist es nicht empfehlenswert, die Korrelationen beschränkt auf die betrachtete Partition $C(k,l)$ zu berechnen. Dafür bietet es sich an, anstelle der Zeilen- bzw. Spaltenmittelwerte $\bar{y}_{i.}$ und $\bar{y}_{.j}$ für die Schätzung des fehlenden Werts $y_{ij} \in C(k,l)$ die Werte

$$\bar{y}_{i.}^{C_2(l)} = \begin{cases} \dfrac{\sum\limits_{j_1 \in J_i} q_{j_1 l} y_{ij_1}}{\sum\limits_{j_1 \in J_i} q_{j_1 l}}, & \text{falls } \sum\limits_{j_1 \in J_i} q_{j_1 l} > 0 \\ \bar{y}_{i.}, & \text{sonst} \end{cases}$$

und

$$\bar{y}_{.j}^{C_1(k)} = \begin{cases} \dfrac{\sum\limits_{i_1 \in I_j} p_{i_1 k} y_{i_1 j}}{\sum\limits_{i_1 \in I_j} p_{i_1 k}}, & \text{falls } \sum\limits_{i_1 \in I_j} p_{i_1 k} > 0 \\ \bar{y}_{.j}, & \text{sonst} \end{cases}$$

zu verwenden. Damit ergeben sich für $y_{ij} \in C(k,l)$ die Schätzer

$$\hat{Y}_{ij}^{SP1} = \begin{cases} \bar{y}_{i.}^{C_2(l)} + \dfrac{\sum\limits_{\iota \in I_j} r_{i\iota}^{Matthai} p_{\iota k} (y_{\iota j} - \bar{y}_{\iota.}^{C_2(l)})}{\sum\limits_{\iota \in I_j} p_{\iota k} |r_{i\iota}^{Matthai}|}, & \text{falls } \sum\limits_{\iota \in I_j} p_{\iota k} > 0 \\ \bar{y}_{i.}^{C_2(l)} + \dfrac{\sum\limits_{\iota \in I_j} r_{i\iota}^{Matthai} (y_{\iota j} - \bar{y}_{\iota.}^{C_2(l)})}{\sum\limits_{\iota \in I_j} |r_{i\iota}^{Matthai}|}, & \text{sonst} \end{cases}$$

bzw.

$$\hat{Y}_{ij}^{SP2} = \begin{cases} \dfrac{\sum\limits_{\iota \in I_j} r_{i\iota}^{Matthai} p_{\iota k} y_{\iota j}}{\sum\limits_{\iota \in I_j} p_{\iota k} |r_{i\iota}^{Matthai}|}, & \text{falls } \sum\limits_{\iota \in I_j} p_{\iota k} > 0 \\ \dfrac{\sum\limits_{\iota \in I_j} r_{i\iota}^{Matthai} y_{\iota j}}{\sum\limits_{\iota \in I_j} |r_{i\iota}^{Matthai}|}, & \text{sonst} \end{cases}$$

für die Nutzer-basierten Ähnlichkeitsverfahren innerhalb zweimodaler Partitionen. Für die Item-basierten Ähnlichkeitsverfahren innerhalb von Partitionen verwendet man

$$\hat{Y}_{ij}^{SP3} = \begin{cases} \dfrac{\sum\limits_{j_1 \in J_i} q_{j_1 l} \tilde{r}_{jj_1}^{Matthai} y_{ij_1}}{\sum\limits_{j_1 \in J_i} q_{j_1 l} |\tilde{r}_{jj_1}^{Matthai}|}, & \text{falls } \sum\limits_{j_1 \in J_i} q_{j_1 l} > 0 \\ \dfrac{\sum\limits_{h \in J_i} \tilde{r}_{jh}^{Matthai} y_{ih}}{\sum\limits_{h \in J_i} |\tilde{r}_{jh}^{Matthai}|}, & \text{sonst} \end{cases}$$

Da in den Korrelationen $r_{i\iota}^{Matthai}$ Mittelwerte $\bar{Y}_{i.}$ verwendet werden, besteht in der Praxis auch bei den auf Partitionen beschränkten Ähnlichkeitsverfahren die Gefahr der Verzerrung durch Ignorieren fehlender Werte. Daher ist möglicherweise die Verwendung der gewichteten Korrelationen

$$r_{i\iota}^{N*} = \frac{\sum\limits_{j' \in J_i \cap J_\iota} \sum\limits_{l'=1}^{L} q_{j'l'} \zeta_{l'}^{*} (y_{ij'} - \bar{y}_{i.}^{*})(y_{ij'} - \bar{y}_{\iota.}^{*})}{\sqrt{\sum\limits_{j' \in J_i \cap J_\iota} \sum\limits_{l'=1}^{L} q_{jl'} \zeta_{l'}^{*} (y_{ij'} - \bar{y}_{i.}^{*})^2 \sum\limits_{j' \in J_i \cap J_\iota} \sum\limits_{l'=1}^{L} q_{j'l'} \zeta_{l'}^{*} (y_{ij'} - \bar{y}_{\iota.}^{*})^2}}$$

und

$$\tilde{r}_{j\underline{h}}^{I*} = \frac{\sum\limits_{i' \in I_j \cap I_{\underline{h}}} \sum\limits_{k'=1}^{K} p_{i'k'} \xi_{k'}^{*} (y_{i'j} - \bar{y}_{.j}^{*})(y_{i'\underline{h}} - \bar{y}_{.\underline{h}}^{*})}{\sqrt{\sum\limits_{i' \in I_j \cap I_{\underline{h}}} \sum\limits_{k'=1}^{K} p_{i'k'} \xi_{k'}^{*} (y_{i'j} - \bar{y}_{.j}^{*})^2 \sum\limits_{i' \in I_j \cap I_{\underline{h}}} \sum\limits_{k'=1}^{K} p_{i'k'} \xi_{k'}^{*} (y_{i'\underline{h}} - \bar{y}_{.\underline{h}}^{*})^2}}.$$

vorteilhaft. Zudem können auch die jeweils nicht zur Bestimmung der Partitionen benutzten Schätzer $\hat{S}_Y^x, x = 1, 2, 3$, als Imputations-Schätzer verwendet werden.

Ausschlaggebend für den Erfolg der Imputation innerhalb zweimodaler Partitionen ist, ob das verwendete Verfahren zur zweimodalen Klassifikation in der Lage ist, Partitionen zu erzeugen, die bezüglich der zugehörigen Nutzer und Items so homogen sind, daß innerhalb dieser Partitionen die MCAR-Eigenschaft in guter Näherung erfüllt ist. Alle vorgestellten zweimodalen Clusterverfahren ignorieren bei der Bestimmung der Cluster-Zugehörigkeit die fehlenden Werte. Hierdurch werden nicht zwangsläufig Partitionen erzeugt, die mit den Partitionen vergleichbar wären, die man erhalten würde, wenn alle Daten gegeben wären. Daher wäre durchaus denkbar, daß bei hohen Fehlendanteilen die fehlenden Daten sich von den innerhalb derselben Partition vorhandenen Daten deutlich unterscheiden. In diesem Fall wäre die MCAR-Annahme nicht gerechtfertigt.

5.8 Ordinales zweimodales Clusterverfahren

Alle bisher betrachteten zweimodalen Clusterverfahren setzen kardinales Skalenniveau voraus. Analog zur ordinalen Matrixfakorisation läßt sich für die zweimodalen Clusterverfahren eine Zielfunktion finden, die das ordinale Skalenniveau der Bewertungsdaten berücksichtigt:

$$\mathcal{Z}_O^x(\gamma) = \sum_{i'=1}^{I} \sum_{j' \in J_{i'}} \sum_{c'=1}^{C-1} h(\chi_{i'j'}^{c'}(\gamma_{i'c'} - \hat{S}_{Y,i'j'}^x)), \quad x = 1, 2.$$

$\mathcal{Z}_O^x$ entspricht bis auf den verwendeten Schätzer $\hat{S}_{Y,ij}^x$ dem Strafkostenterm in der zur ordinalen Matrixfaktorisation verwendeten Zielfunktion. Es gilt wieder

$$h(z) = \begin{cases} \frac{1}{2} - z, & \text{für } z < 0 \\ \frac{1}{2}(1-z)^2, & \text{für } 0 \leq z \leq 1 \\ 0, & \text{für } z > 1 \end{cases} .$$

Genau wie der Schätzer $\hat{A}_{Y,ij} = (\breve{U}\breve{\mathcal{V}}')_{ij}$ werden die Schätzer $\hat{S}_{Y,ij}^x, x = 1, 2$, als Schätzer einer latenten kontinuierlichen Variable aufgefaßt, die den ordinalen Bewertungsdaten zugrundeliegt aber selbst nicht in Erscheinung tritt.

Sowohl der Schätzer $(\breve{U}\breve{\mathcal{V}}')_{ij}$ im Ansatz von Rennie, Srebro (2005) als auch die Schätzer $\hat{S}_{Y,ij}^x, x = 1, 2$, für die kontinuierliche latente Größe im eigenen Ansatz haben kardinales Skalenniveau.

Man könnte in den Formeln der Schätzer $\hat{S}_{Y,ij}^x, x = 1, 2$, die verwendeten Mittelwerte durch den jeweils entsprechenden Median ersetzen. Da dann die Zahlenwerte der größten und kleinsten Bewertungen einer Partition keine Rolle mehr spielten, könnten hierdurch weniger homogene Partitionen erzeugt werden. Zudem würde der Rechenaufwand erheblich erhöht.

Auch hier empfiehlt es sich, die Zielfunktion $Z_O^x, x = 1, 2$, in verschiedene Anteile

$$\mathcal{Z}_O^{P,x}(p^i, \gamma, W, \tilde{W}_1, \tilde{W}_2, Q) = \sum_{j' \in J_{i'}} \sum_{c'=1}^{C-1} h(\chi_{ij'}^{c'}(\gamma_{ic'} - \hat{S}_{Y,ij'}^x)), \quad x = 1, 2,$$

und

$$\mathcal{Z}_O^{Q,x}(q^j,\gamma,W,\tilde{W}_1,\tilde{W}_2,P)=\sum_{i'\in I_j}\sum_{c'=1}^{C-1}h(\chi_{i'j}^{c'}(\gamma_{i'c'}-\hat{S}_{Y,i'j}^{x})),\quad x=1,2,$$

zu zerlegen. Es ergibt sich der folgende Abwandelung der AE-Variante nach George, Merugu (2005):

1. Bestimme P und Q gemäß den Nebenbedingungen.
2. Setze $n=2$, $P^1=P$ und $Q^1=Q$.
3. Wähle $P^0\neq P^1$ und $Q^0\neq Q^1$.
4. Falls $x=2$ ist, berechne $\bar{y}_{i.}, i=1,...,I$, und $\bar{y}_{.j}, j=1,...,J$.
5. Wähle Startwerte γ^1 für γ.
6. Bis $P^{n-1}=P^{n-2}$ und $Q^{n-1}=Q^{n-2}$:
   a) Berechne W^n, $\tilde{W}_1^n=(\tilde{w}_{1.}^n,...,\tilde{w}_{K.}^n)$ und $\tilde{W}_2^n=(\tilde{w}_{.1}^n,...,\tilde{w}_{.L}^n)$.
   b) Bestimme γ^n mittels des Konjugierten Gradientenverfahrens.
   c) $\forall i\in\{1,...,I\}: p^{i,n}=\arg\min_{p^i\text{u.d.N.}}\mathcal{Z}_O^{P,x}(p^i,\gamma,W^n,\tilde{W}_1^n,\tilde{W}_2^n,Q^{n-1})$
   d) $\forall j\in\{1,...,J\}: q^{j,n}=\arg\min_{q^j\text{u.d.N.}}\mathcal{Z}_O^{Q,x}(q^j,\gamma,W^n,\tilde{W}_1^n,\tilde{W}_2^n,P^n)$
   e) $n\leftarrow n+1$

Abbildung 5.11: Algorithmus zur Bestimmung der zweimodalen Clusterzugehörigkeiten für ordinale Daten

Hierbei ist $\gamma=(\gamma_{11},...,\gamma_{1C-1},...,\gamma_{I1},...,\gamma_{IC-1})'$. Zweckmäßige Startwerte wären zum Beispiel $\gamma_{ic}^1=c+\frac{1}{2}, i=1,...,I$. Im Schritt 6b) wird der Vektor γ mit Hilfe der Polak-Ribière'schen Variante des Konjugierten Gradientenverfahrens (Nocedal, Wright (1999), Press et. al. (2002)) ermittelt (siehe Anhang D). Es wird der Gradientenunabhängigkeitstest (Nocedal, Wright (1999)) verwendet.

In den Schritten 6c) und 6d) sind wieder die bekannten Nebenbedingungen $(N1)$ und $(N2)$ zu beachten.

Man benutzt den Gradient

$$\frac{\partial \mathcal{Z}_O^x}{\partial \gamma'} = \begin{pmatrix} \sum_{j' \in J_1} \chi_{1j'}^1 h'(\chi_{1j'}^1(\gamma_{11} - \hat{S}_{Y,1j'}^x)) \\ \vdots \\ \sum_{j' \in J_1} \chi_{1j'}^{C-1} h'(\chi_{1j'}^{C-1}(\gamma_{1C-1} - \hat{S}_{Y,1j'}^x)) \\ \vdots \\ \sum_{j' \in J_I} \chi_{Ij'}^1 h'(\chi_{Ij'}^1(\gamma_{I1} - \hat{S}_{Y,Ij'}^x)) \\ \vdots \\ \sum_{j' \in J_I} \chi_{Ij'}^{C-1} h'(\chi_{Ij'}^{C-1}(\gamma_{IC-1} - \hat{S}_{Y,Ij'}^x)) \end{pmatrix}.$$

Hierbei verwendet man dieselbe Funktion h' wie beim Verfahren nach Rennie, Srebro (2005). Auch dieser zweimodale Clusteransatz ist eine lineare Heuristik, die die fehlenden Werte ignoriert. Im Unterschied zu den übrigen vorgestellten zweimodalen Clusteransätzen trägt dieser Ansatz dem ordinalen Skalenniveau der Bewertungsdaten Rechnung.

Beispiel 5.15:

Das Ziel ist die Berechnung des Schätzers für Bernds Bewertung von Barry Lyndon $\hat{Y}_{17}$ mit Hilfe des zweimodalen Clusteransatzes für ordinale Daten. Wie in Beispiel 5.10 wird für alle Nutzer, Items und Bewertungsgruppen c der Bewertungstensor $\mathcal{X}_{ij}^c$ bestimmt. Es wird $x = 2$ gewählt. (Das ordinale zweimodale Clusterverfahren, das statt $\hat{S}_{Y,ij}^2$ den Schätzer $\hat{S}_{Y,ij}^1$ verwendet, verläuft analog und ist etwas weniger kompliziert.)
Mit den Startwerten

$$P^1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \quad \text{und} \quad Q^1 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix}'$$

ergeben sich

$$W^2 = \begin{pmatrix} 3{,}27 & 3{,}38 & 3{,}64 \\ 3{,}50 & 3{,}50 & 2{,}00 \end{pmatrix}.$$

und die folgenden Anfangsbedingungen:

		$l=1$			$l=2$		$l=3$				
k	i	$j=1$	$j=2$	$j=8$	$j=3$	$j=6$	$j=4$	$j=5$	$j=7$	$\bar{y}_{i.}$	$\tilde{w}^2_{k.}$
	1	1	5	-	4	4	4	3	-	$\frac{21}{6}$	
	3	-	4	1	4	-	4	3	5	$\frac{21}{6}$	
1	4	5	2	4	2	3	-	2	-	3	$\frac{103}{30}$
	5	4	-	4	2	4	3	4	-	$\frac{21}{6}$	
	6	1	5	-	4	-	4	4	4	$\frac{22}{6}$	
2	2	-	4	3	2	5	2	2	-	3	3
	$\bar{y}_{.j}$	2,75	4	3	3	4	3,4	3	4,5		
	$\tilde{w}^2_{.l}$		$\frac{43}{13}$		3,4			$\frac{44}{13}$			

Abbildung 5.12: Zweimodale Cluster zu Beginn des ordinalen zweimodalen Clusterverfahrens (Beispiel 5.15)

Mit den Anfangswerten $\gamma^1_{ic} = c + \frac{1}{2}, i = 1, ..., I, c = 1, ..., C-1$, berechnet zunächst man mit Hilfe des Konjugierten Gradientenverfahrens γ^2. Die ersten 4 Komponenten dieses Vektors sind

$$\gamma^2_{11} = 2,68, \quad \gamma^2_{12} = 2,68, \quad \gamma^2_{13} = 3,14, \quad \gamma^2_{14} = 5,07.$$

Es ergibt sich mit $p^{1,1} = (1,0)'$ der Zielfunktionswert

$$\mathcal{Z}^{P,2}_O(p^{1,1}, \gamma^2, W^2, \tilde{W}^2_1, \tilde{W}^2_2, Q^1) =$$
$$\sum_{j' \in J_1} \sum_{c'=1}^{C-1} h(\chi^{c'}_{1j'}(\gamma^2_{ic'} - \hat{S}^2_{Y,ij'}(p^{1,1}, W^2, \tilde{W}^2_1, \tilde{W}^2_2, Q^1))) = 3{,}32,$$

Verändert man die Clusterzugehörigkeit des ersten Nutzers so ergibt sich mit

$\tilde{p}^1 = (0,1)'$ der Zielfunktionswert:

$$\mathcal{Z}_O^{P,2}(\tilde{p}^1, \gamma^2, W^2, \tilde{W}_1^2, \tilde{W}_2^2, Q^1) =$$
$$\sum_{j' \in J_1} \sum_{c'=1}^{C-1} h(\chi_{1j}^{c'}(\gamma_{ic'}^2 - \hat{S}_{Y,1j'}^2(\tilde{p}^1, W^2, \tilde{W}_1^2, \tilde{W}_2^2, Q^1))) = 8{,}21,$$

weshalb die alte Clusterzugehörigkeit beibehalten wird. Auf diese Weise wird die Clusterzugehörigkeit aller Nutzer (P^2) für vorgegebene Werte $W^2, \tilde{W}_1^2, \tilde{W}_2^2, \gamma^2$ und Q^1 bestimmt. Ganz analog optimiert man die Clusterzugehörigkeit der Items für gegebene Werte $W^2, \tilde{W}_1, \tilde{W}_2, \gamma^2$ und P^2. Danach inkrementiert man n und wiederholt dasselbe bis sich keine Änderungen der Clusterzugehörigkeiten mehr ergeben.
Insgesamt erhält man dieselbe Klassifikation wie in Beispiel 5.12:

		$l=1$		$l=2$			$l=3$				
k	i	$j=1$	$j=8$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$\bar{y}_{i.}$	$\tilde{w}_{k.}$
	2	-	3	4	2	2	2	5	-	3	
1	4	5	4	2	2	-	2	3	-	3	3,17
	5	4	4	-	2	3	4	4	-	$\frac{21}{6}$	
	1	1	-	5	4	4	3	4	-	$\frac{21}{6}$	
2	3	-	1	4	4	4	3	-	5	$\frac{21}{6}$	3,56
	6	1	-	5	4	4	4	-	4	$\frac{22}{6}$	
$\bar{y}_{.j}$		2,75	3	4	3	3,4	3	4	4,5		
$\tilde{w}_{.l}$		$\frac{23}{8}$			3,44			3,58			

Abbildung 5.13: Zweimodale Cluster (Beispiel 5.15)

Daher ergibt sich wieder $\hat{S}_{Y,17}^2 = 4,69$. Zudem erhält man die Schwellenwerte $\gamma_{11} = \gamma_{12} = 2,19$, $\gamma_{13} = 3,31$ und $\gamma_{14} = 5,20$. Wegen $\gamma_{13} < \hat{S}_{Y,17}^2 < \gamma_{14}$ ergibt sich der Schätzwert $\hat{Y}_{17}^{O,2} = 4$.

Durch die Schätzung individueller Schwellenwerte kann das individuelle Bewertungsverhalten besser berücksichtigt werden. Ein Vorteil der vorgestellten Methode gegenüber dem Verfahren zur ordinalen Matrixfaktorisation liegt in der besse-

ren Interpretierbarkeit zweimodaler Clusterverfahren. In diesem Zusammenhang ist vor allem die kompakte Zusammenfassung der Interdependenzen zwischen Nutzer-Clustern und Item-Clustern durch die Gewichtematrix W zu nennen.

5.9 Gütemaße

Um der Frage nachzugehen, ob ein Modell zur Vorhersage von Daten geeignet ist, welche zur Schätzung der Modellparameter nicht verwendet wurden, ist es zwechmäßig, die vorhandenen Daten in zwei disjunkte Datensätze, den Trainings- und den Testdatensatz, zu unterteilen. Der Trainingsdatensatz wird dann dazu verwendet, das Modell und seine Parameter zu bestimmen. Dagegen fungiert der Testdatensatz als alleinige Datengrundlage zur Bestimmung der Güte der Vorhersage. Hierbei werden die Vorhersagen für den Testdatensatz durch Rekurs auf das mittels der Trainingsdaten ermittelte Modell und seine mit Hilfe der Trainingsdaten geschätzten Parameter generiert. Alle zur Evaluation der Vorhersagegüte verwendeten Maße vergleichen die Testdaten mit den auf Basis der Trainingsdaten berechneten Prognosen für die Testdaten.

Die Menge aller Items, bezüglich derer Bewertungen eines Nutzers $i \in \{1, ..., I\}$ im Testdatensatz vorhanden sind, wird als J_i^{test} bezeichnet.

Zur Bestimmung der Güte der Vorhersage können dieselben Gütemaße verwendet werden, die auch zur Beurteilung der Güte der Anpassung verwendet werden. Zu nennen wären in diesem Zusammenhang R^2 und AAD. Letzteres ist durch

$$AAD = \sum_{i'=1}^{I} \sum_{j' \in J_{i'}^{test}} |y_{i'j'} - \hat{y}_{i'j'}|$$

definiert. Während das in der Statistik gebräuchliche R^2 in der Literatur, die sich mit der Prognose ranggeordneter Bewertungsdaten vor dem Hintergrund abzugebender Empfehlungen auseinandersetzt, praktisch nicht vorkommt, erfreut sich der AAD-Wert in dieser Literatur allgemeiner Beliebtheit. Es existiert eine Vielzahl empirischer Studien, in denen das AAD als alleiniges Gütemaß zum Vergeich verschiedener Verfahren verwendet wird. (Abschnitt 5.10 belegt, daß dies zumin-

dest aus Marketing-Gesichtspunkten zu falschen Schlußfolgerungen verleiten kann und daher als äußerst problematisch einzustufen ist.)

Präzision (*Prec.*) und Recall (*Rec.*) sind die meistbenutzten Metriken zur Beurteilung einer Suchmaschine (z.B. van Rijsbergen (1979)) und werden in der Literatur häufig mit dem Namen Cleverdon und dem Cranfield Research Project in Verbindung gebracht. Beide Metriken lassen sich zur Evaluation der Vorhersagegüte benutzen und werden auf Basis der folgenden 2*2 Kontingenztafel berechnet:

Testdaten	Prognose		$\sum$
	Erfolg (e)	Mißerfolg (m)	
Erfolg (e)	$\widetilde{a_{ee}}$	$\widetilde{a_{em}}$	$\widetilde{a_{e.}}$
Mißerfolg (m)	$\widetilde{a_{me}}$	$\widetilde{a_{mm}}$	$\widetilde{a_{m.}}$
$\sum$	$\widetilde{a_{.e}}$	$\widetilde{a_{.m}}$	

Tabelle 5.8: Kontingenztafel (zur Berechnung von Precision und Recall)

Hier bezeichnet $\widetilde{a_{ee}}$ ($\widetilde{a_{mm}}$) die Anzahl der realen Erfolge (Mißerfolge) im Testdatensatz, die durch das verwendete Modell korrekt vorausgesagt wurden. Entsprechen gibt $\widetilde{a_{em}}$ ($\widetilde{a_{me}}$) an, wie häufig durch das Modell ein Mißerfolg (Erfolg) prognostiziert wurde, während in Wirklichkeit ein Erfolg (Mißerfolg) vorlag. Um diese Kontingenztafel zu erstellen, muß die Bewertungsskala (sofern sie nicht bereits binär ist) geeignet transformiert werden. Dies kann geschehen, indem man eine geeignete Teilmenge $M^T(\bar{z} = 1) \subset \{1, ..., C\}$ der möglichen Bewertungen $\{1, ..., C\}$ als Erfolg und alle übrigen Bewertungen als Mißerfolg definiert. Bei Modellen, in denen nicht eine Bewertungsprognose berechnet, sondern nur die Wahrscheinlichkeit für eine bestimmte Prognose geschätzt wird (wie zum Beispiel bei den Kumulativen Modellen oder beim Intervalldaten-Ansatz), muß außerdem definiert werden, wie groß die geschätzte Wahrscheinlichkeit für die gewählte Bewertungsteilmenge $M^T(\bar{z} = 1)$ sein muß, damit ein Erfolg ($\bar{z} = 1$) prognostiziert werden kann. Im Kontext von Recommender-Systemen wird Erfolg hinsichtlich der meisten Modelle genau dann prognostiziert, wenn die prognostizierte Bewertung $\hat{Y}_{ij}, j \in J_i^{test}$ einen bestimmten Schwellenwert γ_{PR} erreicht und außerdem davon auszugehen ist, daß dem Nutzer das betreffende Item noch nicht bekannt ist ($v_{ij}, j \in J_i^{test}$).

Dementsprechend ist

$$\begin{aligned}
\widetilde{a_{ee}} &= \sum_{i'=1}^{I} |\{j' \in J_{i'}^{test} | \hat{Y}_{i'j'} \geq \gamma_{PR} \wedge y_{i'j'} \geq \gamma_{PR} \wedge v_{ij} = 0\}| \\
\widetilde{a_{mm}} &= \sum_{i'=1}^{I} |\{j' \in J_{i'}^{test} | \hat{Y}_{i'j'} < \gamma_{PR} \wedge y_{i'j'} < \gamma_{PR} \wedge v_{i'j'} = 0\}| \\
\widetilde{a_{me}} &= \sum_{i'=1}^{I} |\{j' \in J_{i'}^{test} | \hat{Y}_{i'j'} \geq \gamma_{PR} \wedge y_{i'j'} < \gamma_{PR} \wedge v_{i'j'} = 0\}| \\
\widetilde{a_{em}} &= \sum_{i'=1}^{I} |\{j' \in J_{i'}^{test} | \hat{Y}_{i'j'} < \gamma_{PR} \wedge y_{i'j'} \geq \gamma_{PR} \wedge v_{i'j'} = 0\}|
\end{aligned}$$

Die Präzision gibt an, wie häufig eine Erfolgsprognose richtig ist:

$$Prec. = \frac{\widetilde{a_{ee}}}{\widetilde{a_{.e}}}.$$

Sie ist ein Maß dafür, wie sicher es ist, daß auch ein Erfolg vorliegt, wenn das Modell einen solchen voraussagt.

Durch den Recall erhält man eine Aussage darüber, wie häufig die Vorhersage stimmt, wenn ein Erfolg vorliegt:

$$Rec. = \frac{\widetilde{a_{ee}}}{\widetilde{a_{e.}}}.$$

Hierdurch kann man erkennen, wie sicher durch das Verfahren ein Erfolg als solcher erkannt wird.

Ein weiteres Gütemaß ist der sogenannte F_1-Wert, der auf den Definitionen von Präzision und Recall basiert:

$$F_1 = 2\frac{Prec. \cdot Rec.}{Prec. + Rec.} \quad \left(\text{allgemein: } F_\alpha = (1+\alpha)\frac{Prec. \cdot Rec.}{\alpha \cdot Prec. + Rec.}\right).$$

Die Verwendung dieses Werts ist vor allem dann sinnvoll, wenn es gleichermaßen auf Präzision und Recall ankommt. In machen Anwendungen steht aber eine dieser beiden Größen im Vordergrund des Interesses. Geht es um die Empfehlung wissenschaftlicher Texte, ist es sehr viel problematischer, wenn eine Empfehlung für einen relevanten Text unterbleibt, als wenn einige Texte vorgeschlagen werden, obwohl sie irrelevant sind. Ist das Ziel dagegen die Empfehlung von Büchern, Tonträgern oder Filmen für den privaten Gebrauch, so ist es weniger wichtig, daß einem Nutzer alle Items empfohlen werden, die ihm gefallen. Da die meisten Nutzer nur Zeit haben, einen kleinen Teil dieser Produkte zu konsumieren, sind

hier nur wenige Empfehlungen erforderlich. Wichtiger ist, daß die Empfehlungen, die gemacht werden, möglichst oft stimmen. Daher steht bei der Empfehlung dieser Items die Präzision in Vordergrund. Da sich durch Wahl eines höheren (niedrigeren) Schwellenwerts γ_{PR} im allgemeinen die Präzision (der Recall) auf Kosten des Recalls (der Präzision) erhöhen läßt, sollte auch dann, wenn eine der beiden Gütemaße wichtiger ist, das andere mit aufgeführt werden.

Vor dem Hintergrund, daß es in vielen marketingrelevanten Anwendungen erfolgsentscheidend ist, wie hilfreich dem Nutzer die Empfehlungen erscheinen, ist es vorteilhaft, im Hinblick auf Präzision und Recall einen möglichst hohen Schwellenwert γ_{PR} zu wählen.

Ein Nachteil von Präzision und Recall ist, daß nur zwischen sogenannten Erfolgen und Mißerfolgen (meist in Abhängigkeit vom verwendeten Schwellenwert) unterschieden wird. Hierbei berücksichtigen Präzision und Recall nicht, wie weit eine (nach Maßgabe des gewählten Schwellenwerts) fehlerhafte Prognose von der Realität entfernt ist. In der Praxis kann es in Bezug auf die Kundenzufriedenheit und die Kundenbindung durchaus einen erheblichen Unterschied machen, ob einem Nutzer ein Item empfohlen wurde, das er für mittelmäßig hält, oder ob er dieses Item für unbrauchbar hält. Während im ersten Fall die Empfehlung nicht unbedingt erheblich zur Erhöhung der Kundenzufriedenheit beiträgt, ist im zweiten Fall sogar vorstellbar, daß der Nutzer durch diese Empfehlung verärgert wird. Präzision und Recall können diesen erheblichen Unterschied nicht berücksichtigen.

Neben diesen oft in der Literatur verwendeten Gütekriterien wird hier auch der seltener verwendete Breese-Wert (Breese et. al. (1998)) benutzt. Letzterer wird auch als gewichteter Recall bezeichnet (Ziegler et. al. (2005). Dieses Gütemaß berücksichtigt stärker den von Recommender-Systemen beabsichtigten Zweck der Prognose, der darin besteht auf Basis aller Prognosen der von einem Nutzer nicht bewerteten Items eine Liste von Empfehlungen für diesen Nutzer zu generieren.

Der Breese-Wert $R_{B,i}$ für den i-ten Nutzer soll den (zu erwartenden) Nutzen abbilden, den eine bestimmte geordnete Liste von Empfehlungen für den Nutzer i hat. Hierbei steht das Item aus dem Testdatensatz des Nutzers i an höchster Position, das den höchsten Schätzer aufweist (bzw. dessen Wahrscheinlichkeit für eine Höchstbewertung maximal ist). Je höher der Schätzer für ein Item aus dem sich auf den betreffenden Nutzer i beziehenden Teil des Testdatensatzes J_i^{test} ist, umso höher ist dieses Item in der (geordneten) Empfehlungsliste $\{1, ..., J_{i,list}\}$

anzusiedeln. Es gilt $j_{i,list} \in \{1, ..., J_{i,list}\}$.

Je größer die positive Differenz zwischen der im Testdatensatz vorhandenen tatsächlichen Bewertung des betrachteten Nutzers und der Durchschnittsbewertung des Nutzers (oder einer neutralen Bewertung) ausfällt, umso hilfreicher ist diese Empfehlung für Nutzer i.

Die Breese'sche Bestimmung des Nutzens basiert auf der Annahme, daß einer Empfehlung umso eher gefolgt wird, je höher ihre Position in der Liste ist. Daher werden alle positiven Differenzen noch mit einem Gewichtungsfaktor multipliziert, der umso größer ausfällt, je tiefer die Position des Items in der betrachteten Liste ist:

$$R_{B,i} = \sum_{j'_{i,list}=1}^{J_{i,list}} \frac{\max(y_{ij'_{i,list}} - \bar{d}_i^*, 0)}{2^{(j'_{i,list}-1)/(\alpha_c-1)}}.$$

Man wählt in der Regel $\alpha_c = 5$. $\bar{d}_i^*, i = 1, ..., I$, übernimmt eine vergleichbare Funktion, die der Schwellenwert bei Präzision und Recall hat. Hier wird $\bar{d}_i^* = \bar{y}_{i.}$ gewählt. Auf diese Weise können auch Bewertungen, die kleiner als die Höchstbewertung sind, in die Berechnung des Breese-Werts miteinfließen.

$R_{B,i}^{max}$ sei der maximal erreichbare Nutzen, falls alle Items der Reihenfolge ihrer Bewertung nach auf der Empfehlungsliste angeordnet worden wären. Sowohl $R_{B,i}$ als auch $R_{B,i}^{max}$ müssen für alle Nutzer $i \in \{1, ..., I\}$ berechnet werden. Insgesamt ergibt sich damit der Breese-Wert

$$R_B = 100 \frac{\sum_{i'=1}^{I} R_{B,i'}}{\sum_{i'=1}^{I} R_{B,i'}^{max}}.$$

Für kleine Testdatensätze besteht die Gefahr, daß sich in Bezug auf einige Nutzer einelementige Empfehlungslisten bilden. Daher fällt der Breese-Wert tendenziell besser aus, falls der Anteil des Testdatensatzes an der insgesamt betrachteten Datenmenge gering ist. Daher ist der Breese-Wert nicht unabhängig von der Größe des verwendeten Testdatensatzes. Mittels des Breese-Werts sollten daher nur Ergebnisse verglichen werden, die sich auf gleich große Testdatensätze beziehen.

Unter Marketing-Gesichtspunkten ist der Breese-Wert trotz dieser praktischen Schwierigkeiten das sinnvollste Gütemaß, da er direkt den (zu erwartenden) Nutzen der aus den Prognosen resultierenden Empfehlungen heuristisch quantifiziert.

5.10 Empirischer Vergleich

Im folgenden Abschnitt werden in der bisherigen Literatur nicht angestellte empirische Vergleiche der wichtigsten bekannten Verfahren dargestellt. Zudem wird das neue zweimodale Clusterverfahren für ordinalskalierte Daten mit den bisherigen Verfahren empirisch verglichen. Alle Verfahren werden zunächst für per Zufallszahlen erzeugten Test- und Trainigsdatensätzen bei variierenden Anteilen von Test- und Trainigsdatensatz am gesamten Datensatz miteinander verglichen. Danach erfolgt auch ein Vergleich der wichtigsten Methoden für unterschiedlich stark verzerrte Test- und Trainingsdatensätze.

Als Datengrundlage der ersten Vergleiche wird ein Teil des MovieLens-Datensatzes, D1, verwendet. Dieser Teil umfaßt die Bewertungen von 1067 Nutzern für 418 Filme. Hierbei wurden Nutzer und Filme so ausgewählt, daß die Fehlendstruktur des Datensatzes nicht zu hoch ist. Die Fehlendstruktur liegt bei 79 %. Dies ermöglicht den Vergleich bei hohen prozentualen Anteilen des Testdatensatzes am gesamten Datensatz. Hierdurch wird eine Situation simuliert, in der nur ein kleiner Teil der relevanten Informationen zur Berechnung des Schätzer zur Verfügung steht. Auf diese Weise kann auch die Abhängigkeit der Güte der verschiedenen Verfahren von der Menge an verfügbarer Information untersucht werden. Unter Annahme der MCAR-Eigenschaft simuliert der Anteil des Testdatensatzes am gesamten Datensatz die Fehlendstruktur.

Sarwar et. al. (2000b) konnten zeigen, daß ihr SVD-basierter Ansatz dem Nutzer-basierten Ähnlichkeitsverfahren bei hohen Anteilen des Testdatensatzes am Gesamtdatensatz überlegen ist. Weil die Schätzer des Nutzer-basierten und des Item-basierten Ähnlichkeitsverfahrens bei hohen Anteilen des Testdatensatzes am Gesamtdatensatz weniger stark voneinander abweichen als die Schätzer des Nutzer-basierten Ähnlichkeitsansatzes und des genannten SVD-basierten Ansatzes (Sarwar et. al. (2000b), Sarwar et. al. (2001)), wurde bereits vermutet, daß der SVD-basierte Ansatz auch dem Item-basierten Ähnlichkeitsverfahren überlegen sein könnte. Diese Vermutung soll an dieser Stelle anhand des eben beschriebenen Datensatzes empirisch überprüft werden. Dazu wird der *AAD*-Wert im Testdatensatz für verschiedene verhältnismäßig hohe Anteile des Testdatensatzes am Gesamtdatensatz (10 bis 90 %) berechnet. Eine Variante des Item-basierten Ähnlichkeitsverfahrens benutzt die angepaßte Kosinus-Ähnlichkeit (IBK), die andere (IBBPM) verwendet den nach Matthai modifizierten Bravais-Pearson Korrelationskoeffizient. Stellvertretend für alle SVD-basierten Ansätze wird das SVD-

basierte Verfahren nach Sarwar et. al. (2000b) verwendet, da alle übrigen Ansätze entweder deutlich mehr Rechenaufwand verursachen (EM-SVD nach Srebro, Jaakkola (2003), Verfahren nach Billsus, Pazzani (1998)), oder problematische Voraussetzungen an die Daten stellen (Hauptkomponentenmethode nach Goldberg et. al. (2001)). Zunächst werden alle unterschiedlichen Methoden wie in der Literatur üblich anhand der jeweiligen *AAD*-Werte verglichen.

Methode:	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
IBK	0,790	0,788	0,782	0,790	0,797	0,813	0,823	0,862	0,973
IBBPM	0,723	0,727	0,721	0,730	0,741	0,760	0,781	0,830	0,938
SVD	0,726	0,733	0,740	0,748	0,757	0,768	0,775	0,782	0,801

Tabelle 5.9: AAD der Item-basierten Ähnlichkeitsverfahren IBK und IBBPM und des SVD-basierten Ansatzes nach Sarwar et. al. (2000b) unter Verwendung von $R = 14$

Die Ergebnisse des Ähnlichkeitsansatzes lassen sich bei diesem Datensatz nicht durch die Verwendung einer aus einer festen Anzahl ähnlichster Items bestehenden Nachbarschaft für jedes Item verbessern. Man erkennt, daß der SVD-basierte Ansatz nach Sarwar et. al. (2000a) den Item-basierten Ähnlichkeitsverfahren bei hohen prozentualen Anteilen des Testdatensatzes am Gesamtdatensatz in Bezug auf die *AAD*-Werte deutlich überlegen ist. Da in der Praxis genaue Schätzer für Daten mit extrem hohen Fehlendtrukturen gefragt sind, ist dies ein wichtiger Vorteil des SVD-basierten Ansatzes.

Bisher wurde nur an einem verhältnismäßig kleinen Datensatz gezeigt, daß das zweimodale Clusterverfahren, das den $\hat{S}_Y^2$-Schätzer benutzt, im Mittel über alle unterschiedlichen Anteile des Testdatensatzes am gesamten Datensatz zu genaueren Schätzern führt als das entsprechende Verfahren, das sich des bekannteren $\hat{S}_Y^1$-Schätzers bedient (Banerjee et. al. (2004)). Deshalb wurde das AAD für diese beiden Verfahren für unterschiedliche Anteile des Testdatensatzes am eben verwendeten Datensatz D1 und unterschiedliche Clustergrößen K und L berechnet.

Cluster:		Anteil des Testdatensatzes am gesamten Datensatz:								
K	L	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
10	5	0,745	0,744	0,734	0,737	0,751	0,752	0,766	0,787	0,844
10	10	0,723	0,725	0,736	0,739	0,743	0,755	0,768	0,795	0,899
10	15	0,717	0,726	0,723	0,728	0,743	0,755	0,772	0,795	0,892
15	5	0,724	0,731	0,743	0,734	0,751	0,765	0,776	0,798	0,866
15	10	0,737	0,733	0,735	0,747	0,744	0,755	0,779	0,815	0,909
15	15	0,722	0,734	0,731	0,743	0,754	0,766	0,774	0,817	0,907
20	5	0,722	0,730	0,731	0,750	0,752	0,753	0,785	0,824	0,903
20	10	0,730	0,733	0,728	0,742	0,756	0,759	0,779	0,825	0,931
20	15	0,719	0,727	0,739	0,745	0,751	0,765	0,787	0,826	0,940

Tabelle 5.10: AAD des zweimodalen Clusterverfahrens ($\hat{S}_Y^1$-Schätzer)

Cluster:		Anteil des Testdatensatzes am gesamten Datensatz:								
K	L	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
10	5	0,715	0,712	0,718	0,728	0,742	0,757	0,770	0,792	0,847
10	10	0,717	0,721	0,728	0,727	0,730	0,734	0,738	0,754	0,789
10	15	0,729	0,730	0,723	0,725	0,726	0,736	0,740	0,756	0,787
15	5	0,722	0,721	0,728	0,730	0,731	0,733	0,740	0,757	0,789
15	10	0,731	0,723	0,724	0,730	0,730	0,733	0,738	0,749	0,788
15	15	0,730	0,730	0,731	0,728	0,730	0,739	0,742	0,758	0,799
20	5	0,721	0,726	0,724	0,726	0,731	0,734	0,738	0,753	0,793
20	10	0,717	0,734	0,729	0,722	0,729	0,736	0,743	0,755	0,792
20	15	0,722	0,729	0,728	0,728	0,734	0,735	0,747	0,758	0,801

Tabelle 5.11: AAD des zweimodalen Clusterverfahrens ($\hat{S}_Y^2$-Schätzer)

Für vergleichsweise niedrige Clustergrößen läßt sich mit dem herkömmlichen $\hat{S}_Y^1$-Schätzer für große Anteile des Testdatensatzes am gesamten Datensatz sogar im Hinblick auf die durch den *AAD*-Wert gemessene Schätzgenauigkeit etwas bessere Ergebnisse erzielen als mittels des $\hat{S}_Y^2$-Schätzers. Die besten Ergebnisse erhält man allerdings für etwas größer gewählten Clustergrößen unter Verwendung des

$\hat{S}_Y^2$-Schätzers. Hier kommen die Clustergrößen K=10-15 und L=10-15 in Frage. Diese Ergebnisse sind deutlich besser als als jene des SVD-basierten Ansatzes und der Item-basierten Ähnlichkeitsverfahren, während die $\hat{S}_Y^1$-Resultate nur besser als die AAD-Schätzgenauigkeit der Item-basierten Ähnlichkeitsverfahren sind. Da durch die Wahl kleiner Clustergrößen der Rechenaufwand verringert werden kann, scheint die Wahl $K = L = 10$ optimal zu sein, sofern der $\hat{S}_Y^2$-Schätzer benutzt wird. Im Hinblick auf den bekannteren $\hat{S}_Y^1$-Schätzer können kleinere AAD-Werte durch eher kleine Werte für K und L erreicht werden.

Man erkennt, daß die Berücksichtigung der Heterogenität der Nutzer und Items durch den $\hat{S}_Y^2$-Schätzer zu einer Erhöhung der Schätzgenauigkeit führen kann.

Bei zufällig fehlenden Daten ergeben sich für das zweimodale $\hat{S}_Y^{2*}$-Verfahren die folgenden Ergebnisse:

Cluster:		Anteil des Testdatensatzes am gesamten Datensatz:								
K	L	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
10	5	0,705	0,720	0,722	0,727	0,734	0,754	0,779	0,804	0,853
10	10	0,707	0,720	0,719	0,729	0,730	0,744	0,759	0,763	0,804
10	15	0,707	0,724	0,724	0,726	0,727	0,738	0,740	0,765	0,796

Tabelle 5.12: AAD des zweimodalen Clusterverfahrens ($\hat{S}_Y^{2*}$-Schätzer)

Bei zufällig fehlenden Daten ergeben sich bei eher kleinen Clustergrößen sehr ähnliche Werte. Je höher die Anzahl der Cluster ist und je größer der prozentuale Anteil des Testatensatzes am gesamten Datensatz ausfällt, umso weniger Werte stehen zur Schätzung der einzelnen $\bar{y}_{i.}^{C_2(l)}$ und $\bar{y}_{.j}^{C_1(k)}$ zur Verfügung. Daher ist dieses Verfahren nur für Nutzer $i \in \{1, ..., I\}$ und Items $j \in \{1, ..., J\}$ geeignet, bei denen J_i/L und I_j/K verhältnismäßig groß ausfallen. Es empfiehlt sich bei Verwendung dieses Verfahrens die Clusteranzahlen K und L eher klein zu wählen.

Für die $\hat{S}_Y^3$-Schätzer nach Cheng, Church (2000) ergibt sich eine beeindruckende Anpassung im Trainingdatensatz. Dagegen ist die Anpassung im Testdatensatz weit weniger überzeugend. Insbesondere ergeben sich ungewöhnlich ungenaue Vorhersagen, wenn nur wenige Werte zur Berechnung der Schätzer zur Verfügung stehen und die Anzahlen der Cluster K und L groß sind. Grund hierfür ist, daß

dann $\bar{y}_{i\cdot}^{C_2(l)}$ und $\bar{y}_{\cdot j}^{C_1(k)}$ auf Basis sehr weniger Werte berechnet werden müssen. Somit ist der Schätzer von Cheng, Church (2000) nicht zur Schätzung fehlender Werte geeignet:

Cluster:		Anteil des Testdatensatzes am gesamten Datensatz:								
K	L	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
10	5	0,755	0,767	0,788	0,804	0,824	0,858	0,894	0,955	1,026
10	10	0,716	0,809	0,826	0,848	0,869	0,906	0,944	0,993	1,025

Tabelle 5.13: AAD des zweimodalen Clusterverfahrens ($\hat{S}_Y^3$-Schätzer)

Als vorhergesagter Erfolg im Sinne der Definitionen von Präzision und Recall werden hier alle Schätzer gewertet, die größer als 4,5 sind. Als tatsächlicher Erfolg werden nur die höchstmögliche Bewertung $C = 5$ angesehen. Da das MovieLens-System ganzzahlige Bewertungen von 1 bis 5 zuläßt zuläßt, wäre es ebenso möglich Bewertungen von 4 oder 5 als tatsächlichen Erfolg zu interpretieren und im Hinblick auf den vorhergesagten Erfolg einen kleineren Schwellenwert zu wählen. Hierdurch ließen sich leichter höhere Präzisions- und Recall-Werte erreichen. Es gibt aber sehr viele Filme und auch recht viele gute Filme. Daher hat niemand Zeit alle Filme zu sehen und die meisten Nutzer werden weder Zeit noch Lust haben, alle Filme, die sie besser als bloß mittelmäßig finden würden, anzuschauen. Wegen der Vielzahl an Filmen ist es im Interesse des Nutzers, daß das Recommender-System nicht bloß Filme identifizieren kann, die er besser als mittelmäßig findet, sondern in der Lage ist, einem bestimmten Nutzer Filme zu empfehlen, die ihm ganz besonders gut gefallen. Ähnlich verhält es sich mit den meisten anderen Items, weshalb es ratsam ist, vorhergesagten und realen Erfolg eher streng zu definieren.

Mit den obigen Definitionen ergeben sich in Bezug auf das zweimodale $\hat{S}_Y^1$-Clusterverfahren im Hinblick auf den 1067 Nutzer und 418 Filme umfassenden Ausschnitt aus dem MovieLens-Datensatz, D1, die in Tabelle 5.15 aufgelistete Werte für Präzision und Recall (in eckigen Klammern).

Cluster:		Anteil des Testdatensatzes am gesamten Datensatz:								
K	L	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
10	5	0,295 [0,057]	0,122 [0,027]	0,295 [0,060]	0,421 [0,095]	0,286 [0,056]	0,445 [0,077]	0,385 [0,075]	0,198 [0,042]	0,411 [0,102]
10	10	0,433 [0,098]	0,445 [0,097]	0,414 [0,093]	0,365 [0,072]	0,463 [0,078]	0,522 [0,109]	0,509 [0,111]	0,493 [0,127]	0,429 [0,148]
10	15	0,497 [0,091]	0,493 [0,118]	0,545 [0,115]	0,459 [0,086]	0,393 [0,078]	0,493 [0,103]	0,525 [0,093]	0,517 [0,089]	0,427 [0,158]
15	10	0,611 [0,115]	0,500 [0,121]	0,610 [0,126]	0,603 [0,124]	0,598 [0,134]	0,546 [0,107]	0,561 [0,099]	0,502 [0,122]	0,367 [0,151]
20	15	0,608 [0,108]	0,548 [0,144]	0,544 [0,115]	0,604 [0,130]	0,595 [0,130]	0,590 [0,133]	0,563 [0,141]	0,503 [0,138]	0,362 [0,159]

Tabelle 5.14: $PREC$ und REC (in eckigen Klammern) des zweimodalen Clusterverfahrens ($\hat{S}_Y^1$-Schätzer)

Cluster:		Anteil des Testdatensatzes am gesamten Datensatz:								
K	L	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
10	5	84,28	74,09	71,02	66,54	61,50	56,11	56,43	56,94	46,02
10	10	85,87	74,37	71,76	67,37	66,01	64,22	57,98	54,58	48,56
10	15	84,81	74,29	70,07	63,89	65,89	59,16	59,67	53,21	45,57
15	10	85,22	75,71	67,15	68,19	64,11	62,79	56,97	53,61	47,32
20	15	86,29	76,98	71,59	66,35	65,50	63,18	59,25	53,36	47,29

Tabelle 5.15: Breese-Werte R_B des zweimodalen Clusterverfahrens ($\hat{S}_Y^1$-Schätzer)

Auffallend sind die sehr niedrigen und stark variierenden Präzisions- und Recall-Werte bei kleinen Clusteranzahlen K und L. Trotz verhältnismäßig guter AAD-Anpassung bei kleinen K und L scheint der $\hat{S}_Y^1$-Schätzer insbesondere bei kleinen Werten für die Anzahl der Cluster ungeeignet zur Empfehlung von besonders interessanten Items zu sein. Das liegt daran, daß die Partitionen umso größer werden, je kleiner K und L sind. Große Partitionen neigen eher dazu, eine hohe Anzahl heterognener Elemente zu umfassen. Daher fallen die Partitionsmittelwerte bei kleinen Clusteranzahlen ($K < 10, L < 10$) meist weniger extrem aus, weshalb es nur wenige Schätzer gibt, die die Schwelle 4,5 überschreiten. Daher

Cluster:		Anteil des Testdatensatzes am gesamten Datensatz:								
K	L	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
10	5	0,650 [0,227]	0,622 [0,227]	0,601 [0,240]	0,589 [0,240]	0,547 [0,229]	0,545 [0,232]	0,505 [0,236]	0,489 [0,253]	0,444 [0,276]
10	10	0,626 [0,187]	0,621 [0,204]	0,611 [0,193]	0,594 [0,192]	0,596 [0,201]	0,590 [0,200]	0,563 [0,218]	0,545 [0,219]	0,495 [0,231]
10	15	0,619 [0,178]	0,624 [0,208]	0,617 [0,200]	0,600 [0,198]	0,594 [0,196]	0,585 [0,201]	0,572 [0,218]	0,534 [0,205]	0,497 [0,241]
15	10	0,599 [0,190]	0,633 [0,197]	0,604 [0,191]	0,591 [0,200]	0,579 [0,197]	0,596 [0,202]	0,571 [0,206]	0,533 [0,226]	0,497 [0,237]
20	15	0,591 [0,196]	0,615 [0,194]	0,593 [0,202]	0,599 [0,186]	0,589 [0,200]	0,584 [0,206]	0,559 [0,212]	0,536 [0,222]	0,480 [0,238]

Tabelle 5.16: $PREC$ und REC (in eckigen Klammern) des zweimodalen Clusterverfahrens ($\hat{S}_Y^2$-Schätzer)

Cluster:		Anteil des Testdatensatzes am gesamten Datensatz:								
K	L	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
10	5	86,71	77,76	72,52	69,35	65,86	64,39	63,22	62,15	65,94
10	10	86,84	77,62	72,27	68,25	66,86	63,51	63,17	61,67	65,34
10	15	86,51	77,76	72,71	68,96	65,36	64,16	62,79	60,59	65,62
15	10	86,47	77,36	71,78	68,78	65,86	63,19	62,43	65,52	65,30
20	15	86,29	77,63	72,56	69,74	68,66	67,01	66,77	64,88	63,08

Tabelle 5.17: Breese-Werte R_B des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens

fällt insbesondere der Recall bei kleinen Clusteranzahlen extrem niedrig aus. Insgesamt erweist sich die Wahl $K = L = 10$ vor allem aufgrund der Breese-Werte als zufriedenstellende Wahl.

Hinsichtlich der auf Basis des zweimodalen $\hat{S}_Y^2$-Schätzers berechneten Werte fällt insbesondere der Recall für alle Clusteranzahlen sehr viel größer aus beim klassischen Ansatz für den Schätzer. Als Grund hierfür kann angesehen werden, daß der $\hat{S}_Y^2$-Schätzer sowohl der Heterogenität der Items als auch der Heterogenität der Nutzer Rechnung trägt und somit für alle Clusteranzahlen stärker variiert als der $\hat{S}_Y^1$-Schätzer. Sowohl anhand von Präzision und Recall (Tabellen 5.14 und 5.16) als auch auf Basis der Breese-Werte (Tabellen 5.15 und 5.17) kann der auf-

grund der *AAD*-Werte (Tabellen 5.11 und 5.12) gewonnene Eindruck, daß der $\hat{S}_Y^2$-Schätzer in Bezug auf bekannte Items zu besseren Ergebnissen führen kann, bestätigt werden. In dieser Hinsicht führen alle verwendeten Gütemaße zu derselben Schlußfolgerung.
Die Ergebnisse der Imputation innerhalb zweimodaler Partitionen, deren Bildung auf den Ergebnissen des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens basiert, sind in Tabelle 5.18 wiedergegeben:

Ansatz:	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
SP	0,728	0,737	0,748	0,759	0,779	0,803	0,835	0,894	0,961
	0,589	0,573	0,551	0,526	0,494	0,467	0,436	0,386	0,355
	0,227	0,227	0,230	0,225	0,229	0,236	0,250	0,268	0,243
	86,02	75,72	68,95	64,75	60,90	55,59	52,10	49,11	47,14
RHDP	0,888	0,894	0,897	0,907	0,908	0,915	0,921	0,934	0,978
	0,371	0,367	0,362	0,358	0,363	0,358	0,358	0,351	0,329
	0,353	0,356	0,342	0,347	0,357	0,347	0,355	0,355	0,322
	79,86	65,64	59,47	53,78	48,64	47,44	46,27	46,95	44,59
KHDP	0,848	0,852	0,859	0,871	0,879	0,890	0,904	0,928	0,976
	0,410	0,403	0,400	0,388	0,385	0,383	0,374	0,357	0,332
	0,432	0,426	0,421	0,407	0,397	0,386	0,380	0,355	0,328
	86,50	76,83	71,86	68,55	65,40	62,15	58,28	53,36	50,00
$\hat{S}_Y^2$	0,715	0,718	0,719	0,725	0,731	0,741	0,755	0,787	0,881
	0,623	0,623	0,609	0,605	0,600	0,586	0,561	0,506	0,419
	0,177	0,185	0,184	0,186	0,187	0,191	0,200	0,229	0,279
	86,57	77,11	72,24	70,07	68,00	66,89	65,79	62,99	57,90
$\hat{S}_Y^3$	0,725	0,733	0,740	0,745	0,756	0,771	0,789	0,823	0,885
	0,599	0,601	0,578	0,576	0,556	0,543	0,561	0,458	0,371
	0,219	0,224	0,226	0,220	0,218	0,211	0,200	0,129	0,121
	85,19	75,28	69,48	66,25	63,71	60,72	58,18	48,45	43,82

Tabelle 5.18: *AAD*, Präzision, Recall und Breese-Wert R_B (untereinander) der Imputation innerhalb der mittels $\hat{S}_Y^1$ erzeugten Partitionen (für $K = L = 10$)

Alle auf Basis des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens berechneten Imputations-

verfahren führen im Hinblick auf AAD, Präzision und Recall zu schlechteren Resultaten ab als der $\hat{S}_Y^2$-Schätzer selbst. Innerhalb der mittels des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens (für $K = L = 10$) bestimmten Partitionen werden der Ähnlichkeitsansatz innerhalb von zweimodalen Partitionen (SP), die randomisierte Hot-Deck Imputation innerhalb von zweimodalen Partitionen (RHDP), die korrelationsbasierte Hot-deck Imputation innerhalb von zweimodalen Partitionen (KHDP), der $\hat{S}_Y^2$-Schätzer und der Schätzer $\hat{S}_Y^3$ nach Cheng, Church (2000) zur Imputation fehlender Werte eingesetzt.

Im Hinblick auf den AAD-Wert wird das klassische zweimodale Clusterverfahren nur vom $\hat{S}_Y^2$-Imputationsansatz übertroffen. Der $\hat{S}_Y^3$-Imputationsansatz führt zu vergleichbaren bzw. teilweise etwas schlechteren AAD-Werten als der bekannte $\hat{S}_Y^1$-Schätzer. Alle übrigen Verfahren führen zu einer schlechteren AAD-Genauigkeit.

Solange der Anteil des Testdatensatzes am gesamten Datensatz 60 % nicht überschreitet, ist der auf Basis der mittels des klassischen zweimodalen Clusterverfahrens bestimmten Partitionen berechnete $\hat{S}_Y^2$-Schätzer bezüglich AAD, Präzision und Recall genauso gut wie der $\hat{S}_Y^2$-Schätzer, der auf Basis des zweimodalen $\hat{S}_Y^2$-Clusteransatzes bestimmt wird. Sofern ein verhältnismäßig hoher Anteil an Information verfügbar ist, genügt es daher, die Partitionen mit Hilfe der klassischen zweimodalen Clusterverfahrens zu bestimmen und auf dieser Grundlage den $\hat{S}_Y^2$-Schätzer zu berechnen. Auf diese Weise läßt sich etwas Rechenaufwand sparen. Bei allen auf den zweimodalen Partitionen basierenden Imputationsverfahren verschlechtert sich die Genauigkeit stark bei sehr hohen Anteilen des Testdatensatzes am Gesamtdatensatz. Bezüglich des Breese-Wertes führt der auf der Grundlage der mittels des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens bestimmten Partitionen berechnete $\hat{S}_Y^2$-Schätzer bis zu einem Testdatenanteil von 80 % sogar zu etwas besseren Ergebnissen als der auf Basis des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens bestimmte $\hat{S}_Y^2$-Schätzer (vgl. Tabellen 5.15 und 5.17). Bei einem Testdatenanteil von 90 % lassen sich mittels der letzteren Methode zwar deutlich bessere Ergebnis als unter Verwendung des betrachteten Imputationsansatzes erzielen. Dennoch ergibt ein Vergleich mit Tabelle 5.15, daß sich im Hinblich auf den geschätzten Nutzen der resultierenden Empfehlungslisten durch Imputation des $\hat{S}_Y^2$-Schätzers auf Basis der mittels des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens berechnenten Partitionen wesentlich Verbesserungen gegenüber der üblichen Imputation des $\hat{S}_Y^1$-Schätzers erzielen lassen. Dies spricht dafür, bei nicht zu hohen

		Anteil des Testdatensatzes am gesamten Datensatz:								
C_{MF}	R	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
2	14	0,727	0,744	0,765	0,800	0,859	0,897	0,938	0,912	0,881
		0,497	0,492	0,476	0,455	0,419	0,388	0,370	0,370	0,389
		0,234	0,234	0,232	0,220	0,220	0,221	0,187	0,149	0,145
		80,57	73,86	70,07	68,36	65,94	63,97	62,20	60,14	53,47
5	100	0,719	0,707	0,705	0,696	0,697	0,703	0,723	0,754	0,832
		0,516	0,557	0,554	0,568	0,591	0,596	0,576	0,528	0,414
		0,241	0,261	0,252	0,249	0,251	0,239	0,205	0,154	0,150
		82,90	75,56	72,05	69,37	67,33	65,88	64,04	61,91	55,64
10	100	0,683	0,667	0,673	0,676	0,680	0,695	0,713	0,747	0,827
		0,610	0,618	0,621	0,617	0,610	0,603	0,571	0,521	0,406
		0,277	0,273	0,256	0,254	0,238	0,211	0,195	0,161	0,167
		82,89	76,50	72,11	69,55	67,30	65,84	63,90	61,87	56,14
15	100	0,691	0,692	0,697	0,705	0,718	0,721	0,729	0,756	0,831
		0,558	0,566	0,567	0,554	0,593	0,601	0,566	0,547	0,408
		0,255	0,258	0,261	0,259	0,237	0,235	0,210	0,154	0,147
		82,84	76,51	72,10	69,55	67,30	65,85	63,34	61,63	56,10

Tabelle 5.19: *AAD*, Präzision, Recall und Breese-Wert R_B (untereinander) für die Ordinalen Matrixfaktorisation (Rennie, Srebro (2005))

Fehlendanteilen das etwas schnellere zweimodale $\hat{S}_Y^1$-Clusterverfahren zu Bestimmung der Clusterzugehörigkeiten und der daraus resultierenden Partitionen zu verwenden und auf Basis der letzteren den $\hat{S}_Y^2$-Schätzer zur Berechnung der Prognosen einzusetzen.

Tabelle 5.19 zeigt die Ergebnisse des Verfahrens nach Rennie, Srebro (2005) bezüglich D1. Man sieht, daß man sich durch die Berücksichtigung des ordinalen Skalenniveaus im Hinblick auf das *AAD* auch gegenüber dem zweimodalen $\hat{S}_Y^2$-Verfahren noch erheblich verbessern kann - solange der Anteil des Testdatensatzes an der gesamten Datenmenge nicht höher als 80 % ist. In Bezug auf die Präzision ergeben sich im Vergleich zum zweimodalen $\hat{S}_Y^2$-Clusteransatz keine nennenswerten Verbesserungen. Allerdings fällt auch die Präzision für das ordinale Verfahren nach Rennie, Srebro (2005) bei einem Anteil des Testdatensatzes

von 90% am Gesamtdatensatz deutlich schlechter aus als die des zweimodalen $\hat{S}_Y^2$-Clusteransatzes. Außerdem fallen die Breese-Werte erheblich schlechter aus als beim zweimodalen $\hat{S}_Y^2$-Clusterverfahren. Während im Hinblick auf die D1-Daten die *AAD*-Werte eindeutig für das Verfahren zur ordinalen Matrixfaktorisation sprechen, ist hinsichtlich der Breese-Werte das zweimodale $\hat{S}_Y^2$-Clusterverfahren der Methode nach Rennie, Srebro (2005) vorzuziehen. Da der Breese-Wert den Nutzen der resultierenden Empfehlungslisten anzunähern versucht, erscheint das zweimodale $\hat{S}_Y^2$-Clusterverfahren im Hinblick auf Anwendungen im Zusammenhang mit Recommender-Systemen geeigneter als das Verfahren zur ordinalen Matrixfaktorisation zu sein. Wären die beiden Verfahren bezüglich D1 nur anhand ihrer *AAD*-Werte einander gegenübergestellt worden, hätte die Evaluation beider Methoden zu unter Marketing-Gesichtspunkten falschen Schlußfolgerungen verleiten können. Daher ist die in der Literatur häufig zu findende Praxis problematisch, Verfahren ausschließlich auf Basis der *AAD*-Werte empirisch zu vergleichen. Es ist in diesem Zusammenhang beklagenswert, daß gerade in der neueren in diesem Zusammenhang relevanten Literatur praktisch keine Breese-Werte angegeben werden.

In Tabelle 5.20 sind die Resultate des vorgestellten zweimodalen Clusterverfahrens für ordinalskalierte Daten aufgelistet.

Bei geringen Anteilen des Testdatensatzes am gesamten Datensatz (D1) fällt das *AAD* sogar noch kleiner aus als beim Verfahren nach Rennie, Srebro (2005) und ist somit in dieser Beziehung allen vorgestellten Verfahren überlegen. Bei höheren Anteilen des Testdatensatzes am gesamten Datensatz ist das zum zweimodalen Clusterverfahren für ordinalskalierte Daten gehörende *AAD* dafür etwas größer. Beide ordinale Verfahren produzieren für verhältnismäßig große Testdatensätze größere *AAD*-Werte als das einfache zweimodale $\hat{S}_Y^2$-Clusterverfahren. Im Hinblick auf die Präzision führen die beiden ordinalen Verfahren zu ähnlichen Ergebnissen bezüglich D1 wie das klassische zweimodale Clusterverfahren.

Im Hinblick auf kleine Anteile des Testdatensatzes am D1-Datensatz lassen sich mit dem ordinalen zweimodalen Custerverfahren Breese-Werte erzielen, die vergleichbar mit den auf Basis des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens berechneten Breese-Werten sind. Dagegen fallen die Breese-Werte bei kleinen Testdatenanteilen teilweise etwas schlechter aus als im Hinblick auf das Verfahren zur ordinalen Matrixfaktorisation (OMF).

Cluster:		Anteil des Testdatensatzes am gesamten Datensatz:								
K	L	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
10	5	0,651	0,665	0,668	0,684	0,695	0,712	0,726	0,757	0,831
		0,597	0,598	0,601	0,574	0,566	0,524	0,512	0,546	0,421
		0,264	0,272	0,272	0,260	0,273	0,270	0,283	0,279	0,298
		86,97	77,73	72,48	69,33	65,54	62,38	62,34	62,02	58,29
10	10	0,664	0,670	0,672	0,692	0,705	0,717	0,738	0,771	0,843
		0,607	0,610	0,594	0,562	0,552	0,525	0,513	0,462	0,408
		0,269	0,273	0,284	0,262	0,270	0,272	0,276	0,292	0,287
		87,05	77,91	71,45	68,51	65,55	63,31	61,98	60,65	58,30
10	15	0,671	0,670	0,686	0,698	0,703	0,7209	0,742	0,773	0,855
		0,584	0,614	0,545	0,559	0,548	0,530	0,507	0,466	0,422
		0,281	0,282	0,261	0,265	0,274	0,275	0,264	0,280	0,294
		86,74	77,88	72,52	68,01	65,62	64,05	62,22	60,26	56,31
15	5	0,667	0,672	0,674	0,690	0,700	0,716	0,743	0,774	0,834
		0,612	0,613	0,591	0,563	0,563	0,512	0,509	0,461	0,401
		0,284	0,274	0,281	0,267	0,275	0,273	0,271	0,284	0,302
		86,56	77,68	71,67	67,10	65,50	62,62	61,28	59,93	57,32
15	10	0,669	0,670	0,691	0,699	0,709	0,725	0,743	0,777	0,851
		0,607	0,619	0,565	0,566	0,540	0,523	0,509	0,457	0,398
		0,265	0,280	0,266	0,273	0,271	0,274	0,279	0,279	0,299
		86,52	75,83	72,11	67,36	65,97	63,23	61,94	60,31	56,12

Tabelle 5.20: AAD, Präzision, Recall und Breese-Wert R_B (untereinander) für das ordinale zweimodale Clusterverfahren (eigener Ansatz)

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
IBBPM	87,12	75,87	68,15	62,48	59,13	52,79	46,04	38,86	33,41
SVD	86,05	78,17	73,66	71,10	69,37	68,32	67,61	66,78	66,51
$\hat{S}_Y^1$	85,87	74,37	71,76	67,37	61,01	64,22	57,98	54,58	48,56
$\hat{S}_Y^2$	86,84	77,62	72,27	68,25	66,86	63,51	63,17	61,67	65,34
OZC	87,05	77,91	71,45	68,51	65,55	63,31	61,98	60,65	58,30
OMF	82,89	76,50	72,11	69,55	67,30	65,84	63,90	61,87	56,14

Tabelle 5.21: Breese-Werte R_B verschiedener Nicht-Bayes'scher kollaborativer Verfahren bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz

Tabelle 5.21 enthält einen Vergleich der Breese-Werte der wichtigsten (Nicht-Bayes'schen) kollaborativen Verfahren auf Basis der resultierenden Breese-Werte. Verglichen werden die Version des Item-basierten Ähnlichkeitsverfahrens, bei welcher der nach Matthai modifizierte Bravais-Pearson'sche Korrelationskoeffizient benutzt wird (IBBPM), der SVD-basierte Ansatz nach Sarwar et.al. (2000a), die zweimodalen $\hat{S}_Y^1$- und $\hat{S}_Y^2$-Clusterverfahren, das Verfahren zur ordinalen Matrixfaktorisation nach Rennie, Srebro (2005) und das ordinale zweimodale Clusterverfahren. Insbesondere in Bezug auf große Testdatenanteile führt die Verwendung des SVD-basierten Verfahrens hinsichtlich des D1-Datensatzes zu den größten Breese-Werten.

Zusätzlich wurden auch die Ergebnisse der wichtigsten Verfahren in Bezug auf einen etwas größerer Ausschnitt aus dem MovieLens-Datensatz, D2, berechnet.

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,725	0,728	0,732	0,735	0,739	0,746	0,760	0,784	0,850
Prec.	0,668	0,577	0,564	0,640	0,627	0,623	0,592	0,573	0,517
Rec.	0,127	0,132	0,136	0,144	0,152	0,157	0,152	0,154	0,151
R_B	74,90	69,60	65,57	62,87	63,30	57,66	58,52	54,03	55,28

Tabelle 5.22: Ergebnisse des zweimodalen $\hat{S}^1$-Clusterverfahrens (Datensatz D2)

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,711	0,736	0,737	0,736	0,738	0,741	0,748	0,756	0,800
Prec.	0,618	0,571	0,580	0,580	0,567	0,557	0,553	0,537	0,483
Rec.	0,261	0,265	0,263	0,265	0,272	0,269	0,275	0,279	0,309
R_B	75,63	69,37	67,21	65,13	65,28	64,72	63,79	63,37	61,67

Tabelle 5.23: Ergebnisse des zweimodales $\hat{S}^2$-Clusterverfahren (Datensatz D2)

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,703	0,718	0,721	0,722	0,723	0,725	0,726	0,728	0,771
Prec.	0,552	0,571	0,573	0,568	0,577	0,573	0,571	0,572	0,510
Rec.	0,389	0,274	0,253	0,278	0,241	0,254	0,237	0,226	0,190
R_B	72,48	67,50	66,02	65,94	63,79	63,06	60,92	59,23	55,06

Tabelle 5.24: Ergebnisse der ordinalen Matrixfaktorisation (Datensatz D2)

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,720	0,721	0,725	0,729	0,736	0,738	0,740	0,750	0,789
Prec.	0,607	0,595	0,588	0,581	0,568	0,575	0,566	0,549	0,506
Rec.	0,267	0,297	0,272	0,275	0,273	0,272	0,281	0,282	0,295
R_B	78,42	71,89	68,81	66,13	64,40	62,73	61,83	60,72	60,52

Tabelle 5.25: Ergebnisse des zweimodales Clusterverfahren für ordinale Daten (Datensatz D2)

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,735	0,742	0,748	0,752	0,761	0,772	0,781	0,794	0,826
Prec.	0,739	0,738	0,740	0,708	0,666	0,674	0,625	0,566	0,521
Rec.	0,092	0,083	0,069	0,061	0,059	0,045	0,041	0,044	0,072
R_B	76,16	71,42	69,20	67,66	65,81	65,23	64,06	62,81	60,85

Tabelle 5.26: Ergebnisse des SVD-basierten Verfahrens (Datensatz D2)

Die zugehörigen Ergebnisse sind in den Tabellen 5.22 bis 5.26 enthalten. Der

verwendete Ausschnitt D2 umfaßt 2000 Nutzer und 3043 Items und ist somit wesentlich größer als D1. Der Fehlendanteil beträgt 95% und ist daher deutlich höher als im bisher verwendeten D1-Datensatz.

Auch bezüglich der D2-Daten führt der SVD-basierte Ansatz zu vergleichsweise großen *AAD*-Werten. Mittels des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens lassen sich im Hinblick auf hohe Anteile des Testdatensatzes an der Datenmenge D2 sogar größere Breese-Werte erzielen unter Verwendung der SVD-basierten Methode. Dennoch erweist sich der SVD-basierte Ansatz hinsichtlich kleiner und mittlerer Testdatenanteile als das beste der hier betrachteten (Nicht-Bayes'schen) kollaborativen Verfahren bezüglich D2 im Hinblick auf den geschätzten Nutzen der resultierenden Empfehlungslisten.

In der Praxis der Recommender-Systeme ist es von entscheidender Bedeutung, ob ein Verfahren auch auf einer schmalen Datengrundlage Prognosen abgeben kann, die zu nützlichen Empfehlungen führen. Insbesondere im Hinblick auf neue Nutzer liegen meist nur wenige Bewertungen vor. Wenn die ersten Empfehlungen einem neuen Nutzer nicht hilfreich erscheinen, besteht die Gefahr, daß er sich von dem Recommender-System abwendet. Unter dieser Voraussetzung kann ein Recommender-System eines Online-Shops in Bezug auf diesen (ehemaligen) Nutzer seinen Zweck, die Kundenbindung und die Kundenzufriedenheit zu erhöhen, nicht mehr erfüllen. Darüberhinaus besteht auch die Gefahr, daß ein durch eine oder mehrere schlechte Empfehlungen verärgerter neuer Nutzer sich von dem das Recommender-System betreibenden Online-Shop getäuscht fühlt. In diesem Fall sind negative Einflüsse auf die Kundenzufriedenheit und die Kundenbindung nicht auszuschließen. Daher sprechen die vergleichsweise hohen Breese-Werte, die das zweimodale $\hat{S}_Y^2$-Clusterverfahren bezüglich hoher Testdatenanteile an den D2-Daten erzielen konnte, für eine bessere Eignung dieses Verfahrens.

Interessant wäre in diesem Zusammenhang eine Untersuchung des Einflusses von Datensatzgröße, Fehlendanteil und Dimension der zugrundeliegenden Datenmatrix (I, J) auf die Resultate der Verfahren. Auch anhand des größeren D2-Datensatzes mit der realistischeren Fehlendstruktur bestätigt sich die Beobachtung, daß vergleichsweise gute *AAD*-Werte nicht in allen Fällen mit vergeichsweise guten Breese-Werten verbunden sind (und umgekehrt).

Sowohl im vorangegangenen Teil der Arbeit als auch in der Literatur wird die Einteilung in Test- und Trainingsdatensatz mittels Zufallszahlen vorgenommen.

Daher sollten sich die Verteilungen von Test- und Trainingsdatensatz bei großen Datenmengen im allgemeinen gleichen.

In der Praxis ist aber damit zu rechnen, daß die Nutzer eher über Erfahrungen und Kenntnisse im Hinblick auf Items verfügen, die sie tendenziell bevorzugen. Beispielsweise sehen die meisten Menschen nur dann einen Film an, wenn sie erwarten, daß er ihnen gefallen könnte. Gleiches gilt für Produkte wie CDs und Bücher. Daher haben die Nutzer vorwiegend Erfahrungen mit Items, die sie präferieren, weshalb die Menge ihrer Bewertungen niemals zufällig ausgewählt ist und ihr Durchschnitt deutlich über der durchschnittlichen Bewertung liegen dürfte, die der Nutzer für alle übrigen Items abgeben würde, sofern diese kennen würde. Deshalb ist davon auszugehen, daß in der Praxis für jeden Nutzer eine Menge von tendenziell besser bewerteten Items verwendet werden muß, um seine Bewertungen im Hinblick auf unbekannte bzw. nicht bewertete Items zu bestimmen.

Es könnte auch vermutet werden, daß möglicherweise noch ein weiterer Effekt die Daten verzerrt. Sofern die Erwartungshaltung der Nutzer gegenüber den Items hoch ist, könnte es dazu kommen, daß Nutzer mit übertrieben schlechten Bewertungen reagieren, falls das Item ihren hohen Erwartungen nicht vollständig entspricht. Beispielsweise wäre vorstellbar, daß jemand ein mittelmäßiges Item als schlecht einstuft, weil er dafür einen hohen Preis bezahlen mußte. Dagegen ist einzuwenden, daß die Gefahr von Überreaktionen jedenfalls dann gering ist, wenn keine monitären Kosten oder Opportunitätskosten mit dem Konsum des Items verbunden sind oder wenn diese Kosten als vernachlässigbar oder als Teil eines gewählten Lebensstils (wie z.B. Ausgehen am Samstagabend) empfunden werden. Bei der Erhebung der hier betrachteten MovieLens-Daten wurde den Nutzern die Gelegenheit gegeben, Qualitätsurteile im Hinblick auf bestimmte Filme abzugeben - und zwar unabhängig davon, ob monitäre Kosten oder Opportunitätskosten entstanden sind oder wann sie den Film gesehen haben. Somit ist es verhältnismäßig unwahrscheinlich, daß den Bewertenden (Opportunitäts-) kosten entstanden sind und sie außerdem die Bewertung direkt im Anschluß an den Konsum des Films vornehmen. Da im MovieLens-Datensatz niedrige Bewertungen im Vergleich zu mittelmäßigen und hohen Bewertungen selten vorkommen, kann davon ausgegangen werden, daß die Verzerrungseffekte durch Verärgerung oder Enttäuschung vernachlässigbar gegenüber der zuvor genannten Ursache von Verzerrungen sein dürften.

Folglich würde eine Erprobung der Verfahren unter praxisnahen Bedingun-

gen erfordern, daß der Traingsdatensatz tendenziell höhere Bewertungen als der Testdatensatz enthält. Als Datengrundlage dient in diesem Zusammenang D1.

Zur Erzeugung realistischer Test- und Trainingsdatensätze wurde nur ein Teil des Testdatensatzes per Zufallsziehung aus dem gesamten zur Verfügung stehenden Datensatz (D1) bestimmt. Der restliche Teil der Testdaten wurde aus der Menge an Bewertungen, die kleiner oder gleich 3 sind, ausgewählt. Der prozentuale Anteil des Testdatensatzes, der nur aus einer Menge von Items bestehen darf, die kleiner oder gleich 3 sind, kann als Grad der Verzerrung interpretiert werden. Um auch sehr hohe Verzerrungsgrade simulieren zu können, umfaßt der Testdatensatz insgesamt nur 30% des gesamten Datensatzes. Hierdurch werden sowohl Test- als auch Trainingsdatensatz verzerrt.
Tabelle 5.27 zeigt die durchschnittlichen Bewertungen in Test- und Trainingsdatensatz für unterschiedlich Verzerrungsgrade:

	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
Trainingsdaten	3,52	3,60	3,68	3,77	3,83
Testdaten	3,32	3,15	2,96	2,74	2,61

Tabelle 5.27: Durchschnittsbewertung der verzerrten Test- und Trainingsdaten

Die Histogramme (Abbildungen 4.15 bis 4.19) verdeutlichen den Effekt des jeweiligen Verzerrungsgrads. Rechts neben den Histogrammen der Daten finden sich die Histogramme der auf Grundlage der Trainingsdaten berechneten $\hat{S}_Y^2$-Schätzer. Auch bei stark verzerrten Daten ähneln die Histgramme der Schätzer für die Testdaten den jeweils zugehörigen Histgrammen für die Trainingsdaten auffallend. Daher sind die $\hat{S}_Y^2$-Schätzer verzerrt und überschätzen wie erwartet die Bewertungen des Testdatensatzes. Wie man Tabelle 5.28 entnehmen kann, trifft dies sogar in stärkerem Maße auch auf die Schätzer der übrigen Verfahren zu. Auch durch den $\hat{S}_Y^{2*}$-Schätzer lassen sich nur geringfügige Verbesserungen erreichen.

Abbildung 5.16:
Histogramm der verzerrten Test- und Trainingdaten bei 60% Verzerrung(links)
Histogramm der Schätzer für Test- und Trainingdaten bei 60% Verzerrung(rechts)

Abbildung 5.14:
Histogramm der verzerrten Test- und Trainingdaten bei 20% Verzerrung(links)
Histogramm der Schätzer für Test- und Trainingdaten bei 20% Verzerrung(rechts)

Abbildung 5.15:
Histogramm der verzerrten Test- und Trainingdaten bei 40% Verzerrung(links)
Histogramm der Schätzer für Test- und Trainingdaten bei 40% Verzerrung(rechts)

Die systematischen Überschätzung der Testdaten macht sich durch stark erhöhte *AAD*-Werte bemerkbar. Da der Recall meist verhältnismäßig hoch ist, läßt sich durch eine Erhöhung des Präzision-und-Recall Schwellenwertes von 4,5 (bzw. durch eine nachträgliche Erhöhung des Schwellenwerts $\gamma_{i4}, i = 1, ..., I$, bei den Verfahren für ordinale Daten) eine immerhin höhere Präzision auf Kosten des Recalls erreichen. Um die Präzisionswerte der unverzerrten Datensätzen auch nur annähernd zu erreichen, müssen allerdings in Bezug auf den Recall zum Teil Werte von 0,05 bis 0,01 in Kauf genommen werden. Auch die Breese-Werte (siehe Tabelle 5.29 und Anhang B) belegen, daß sich die Qualität der auf Basis der Schätzer erstellten Empfehlungen durch verzerrte Trainingsdaten merklich verschlechtert.

Während die *AAD*-Werte unter realistischen Bedingungen erheblich größer ausfallen dürften, als die Ergebnisse der meisten veröffentlichten Studien suggerieren, geht der durch die Breese-Werte geschätzte Nutzen der auf Basis der Schätzer generierten Empfehlungen zumindest in Bezug auf die hier betrachteten Verfahren nicht vollständig verloren.

Abbildung 5.17:
Histogramm der verzerrten Test- und Trainingdaten bei 80% Verzerrung(links)
Histogramm der Schätzer für Test- und Trainingdaten bei 80% Verzerrung(rechts)

Abbildung 5.18:
Histogramm der verzerrten Test- und Trainingdaten bei 90% Verzerrung(links)
Histogramm der Schätzer für Test- und Trainingdaten bei 90% Verzerrung(rechts)

Verfahren	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
zweimodales	0,737	0,779	0,814	0,912	0,974
$\hat{S}_Y^1$-Clusterverfahren	0,617	0,450	0,485	0,355	0,268
mit $K = 10, L = 10$	0,163	0,057	0,182	0,140	0,190
zweimodales	0,732	0,760	0,807	0,884	0,940
$\hat{S}_Y^2$-Clusterverfahren	0,566	0,478	0,443	0,359	0,209
mit $K = 10, L = 10$	0,243	0,274	0,271	0,307	0,310
zweimodales	0,729	0,756	0,806	0,884	0,942
$\hat{S}_Y^{2*}$-Clusterverfahren	0,566	0,533	0,455	0,369	0,212
mit $K = 10, L = 10$	0,242	0,253	0,268	0,304	0,297
zweimodales	0,834	0,866	0,902	0,942	0,987
$\hat{S}_Y^3$-Clusterverfahren	0,419	0,358	0,303	0,269	0,194
mit $K = 10, L = 10$	0,295	0,303	0,323	0,318	0,299
zweimodales ordinales	0,702	0,736	0,803	0,889	0,954
Clusterverfahren	0,571	0,533	0,483	0,365	0,220
mit $K = 10, L = 10$	0,266	0,273	0,296	0,304	0,298
ordinale	0,713	0,749	0,811	0,916	0,980
Matrixfaktorisation	0,541	0,505	0,424	0,296	0,186
mit $\mathcal{C}_{MF} = 5, R = 100$	0,259	0,288	0,294	0,310	0,333
ordinale	0,690	0,735	0,806	0,930	1,010
Matrixfaktorisation	0,587	0,555	0,483	0,353	0,219
mit $\mathcal{C}_{MF} = 10, R = 100$	0,264	0,295	0,300	0,316	0,342
SVD-basiertes	0,751	0,788	0,848	0,948	1,021
Verfahren	0,704	0,641	0,552	0,427	0,261
nach Sarwar et. al.	0,157	0,133	0,119	0,223	0,245
Imputation von $\hat{S}_Y^2$ innerhalb	0,727	0,761	0,799	0,898	0,936
zweimodaler Partitionen auf	0,589	0,485	0,457	0,351	0,260
Basis des $\hat{S}_Y^1$-Clusterverfahrens	0,235	0,254	0,266	0,287	0,259

Verfahren	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
$\hat{S}_Y^1, K = 10, L = 10$	68,90	67,09	64,93	62,28	65,74
$\hat{S}_Y^2, K = 10, L = 10$	69,69	68,03	65,34	63,93	65,49
$\hat{S}_Y^{2*}, K = 10, L = 10$	69,77	68,24	65,39	63,90	64,87
$\hat{S}_Y^3, K = 10, L = 10$	63,52	62,97	60,22	56,47	50,61
OZC, $K = 10, L = 10$	70,05	67,36	65,42	64,34	65,86
OMF, $C_{MF} = 5, R = 100$	68,56	67,00	65,22	63,71	64,96
OMF, $\mathcal{C}_{MF} = 10, R = 100$	68,64	67,23	65,44	63,70	65,04
SVD	71,86	68,89	66,28	62,93	60,44
Imputation (S_Y^2)	69,77	68,57	64,98	64,11	65,88

Tabelle 5.29: Breese-Werte R_B (untereinander) bei unterschiedlichen Graden der Verzerrung

Die Imputation des $\hat{S}_Y^2$-Schätzers innerhalb der mittels des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens bestimmten zweimodalen Partitionen wird in Tabelle 5.29 als Imputation ($\hat{S}_Y^2$) bezeichnet.

Bei hohen Verzerrungsgraden führt das auf der Singulärwertzerlegung basierende Verfahren zu erheblich schlechteren Breese-Werten als das zweimodale $\hat{S}_Y^2$-Clusterverfahren, das ordinale zweimodale Clusterverfahren, das Verfahren zur ordinalen Matrixfaktorisation und das Imputationsverfahren. Der reale Verzerrungsgrad ist nicht bekannt. Es wäre aber durchaus naheliegend, daß er in vielen Anwendungen im Zusammenhang mit Recommender-Systemen hoch ist. Daher sprechen die deutlich kleineren Breese-Werte bei hohen Verzerrungsgraden gegen das SVD-basierte Verfahren und für das zweimodale $\hat{S}_Y^2$-Clusterverfahren.

In dieser Arbeit wurden hauptsächlich Verfahren betrachtet, die zu deutlich besseren Ergebnissen als das Nutzer-basierte Ähnlichkeitsverfahren führen. Der empirische Vergleich von Calderón-Benavides et. al. (2004) belegt, daß für eine eine Reihe von hier nicht beschriebener Verfahren (Support Vector Maschines, Dependenz-Netzwerke, Weighted Majority Prediction (WMP) und speicherbasierte WMP) die AAD-Werte in den meisten Fällen sogar schlechter ausfallen als die des Nutzer-basierten Ähnlichkeitsverfahrens.

Insgesamt erweist sich das zweimodale $\hat{S}_Y^2$-Clusterverfahren in Bezug auf die Vorhersage von Bewertungen bekannter Nutzer und bekannter Items als das beste Verfahren. Beide betrachteten ordinalen Verfahren berücksichtigen ebenso wie

das zweimodale $\hat{S}^2$-Clusterverfahren die Heterogenität der Nutzer und Items. Im Rahmen des Verfahrens zur ordinalen Matrixfaktorisation wird der Heterogenität der Nutzer durch die Matrix $\breve{U}$ und der Heterogenität der Items mit Hilfe von $\breve{\mathcal{V}}$ Rechnung getragen. Das ordinale zweimodale Clusterverfahren kann die Heterogenität durch Verwendung des $\hat{S}_Y^2$-Schätzers abbilden. Ein Vergleich zwischen den Ergebnissen des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens und den Resultaten der beiden ordinalen Ansätze belegt, daß die zusätzliche Berücksichtigung des ordinalen Skalenniveaus nicht zwangsläufig zu Ergebnisverbesserungen führen muß. Die Berücksichtigung der Heterogenität scheint wichtiger zu sein als die Berücksichtigung des ordinalen Skalenniveaus.

5.11 Eigenschaften der kollaborativen Verfahren

Genau wie die kontentbasierten Verfahren haben auch die kollaborativen Ansätze Probleme, gute Schätzer für die Bewertungen von Nutzern zu berechnen, die erst wenige Bewertungen abgegeben haben. Im Gegensatz zu den kontentbasierten Methoden können die kollaborativen Verfahren nur sinnvolle Schätzer für Items berechnen, die bereits von genügend Nutzern bewertet wurden. Eine Folge hiervon ist, daß neue oder unbekannte Items nicht empfohlen werden können.

Unabhängig davon, ob es ihm gefällt oder nicht, erhält jeder Nutzer ständig Empfehlungen von Freunden, Kollegen, Verwandten, Nachbarn und einer Menge anderer Personen für bereits bekannte Items. Je nachdem, wie gut der Empfehlende den Geschmack und die Bedürfnisse des Nutzers einschätzen kann, können diese persönlichen Empfehlungen sogar deutlich passender sein als die Empfehlungen eines Recommender-Systems. Daher ist jede Empfehlung für ein allgemein bekanntes Item sehr wahrscheinlich bloß die Wiederholung einer Empfehlung, die dem Nutzer bereits von einer ihm bekannten Person gemacht wurde. Dies mag als weitere Bestätigung der ursprünglichen Empfehlung(en) oder zur Erinnerung dienen. Interessanter wären für den Nutzer allerdings Empfehlungen für neue oder weniger bekannte Items, da solche Items deutlich seltener von anderen Menschen empfohlen werden und der Nutzer in diesem Fall möglicherweise erst durch das Recommender-System von der Existenz dieses Items erfährt. Außerdem wären verläßliche Schätzer für neue und bisher nicht bewertete Items auch

für Marktforscher oder die Betreiber eines Online-Geschäfts interessant. Diese Schätzer könnten beispielsweise dazu benutzt werden, abzuschätzen, welche und wieviele Kunden eines Online-Retailers Interesse an einem bestimmten Produkt haben, bevor der betreffende Artikel überhaupt in das Sortiment aufgenommen wurde. Marktforscher könnten mit den Schätzern die Beliebteit eines Produkts in einem bestimmten Segment abschätzen, bevor dieses Produkt überhaupt hergestellt wurde. Wegen der dargelegten wirtschaftlichen Bedeutung der Schätzer für neue und unbekannte Items, werden in Kapitel 9 verschiedene bekannte und eigene Ansätze zur Berechnung solcher Schätzer vorgestellt und verglichen.

Da die kollaboratischen Verfahren die Eigenschaften der Items nicht benutzen können die kollaborativen Verfahren im Gegensatz zu den kontent-basierten Verfahren auf Items jeder Art angewandt werden. Das Problematik der Erhebung der relevanten Eigenschaften beziehungsweise das Problem der Bestimmung der relevanten Eigenschaften selbst kann hierdurch umgangen werden.

Ein weiterer Vorteil der kollaborativen Verfahren ist, daß sie anders als die kontentbasierte Verfahren auch Empfehlungen für Items abgeben können, die völlig verschieden von den Items sind, die der Nutzer vorher bewertet oder auch nur gesehen hat. Auf diese Weise werden die Empfehlungen abwechselungsreicher. So kann beispielsweise den Nutzern eines Online-Stores eine eher vielfältige Menge unterschiedlicher Wahlmöglichkeiten als eine homogene Menge (homogene Mengen) von Produkten empfohlen werden. Außerdem werden so Empfehlungen von Items möglich, die einer völlig anderen Kategorie als die bisher betrachteten Items angehören. Diese Empfehlungen könnten theoretisch dazu führen, daß der erst auf diese Kategorie von Items aufmerksam gemacht wird. Da aber nur Empfehlungen für eher bekannte Items abgegeben werden können, wird es eher so sein, daß der Nutzer durch diese Empfehlung einmal mehr auf diese Kategorie von Items aufmerksam gemacht wird. Zum Beispiel kann eine solche Empfehlung bei dem Besucher eines Online-Geschäfts bewirken, daß dieser bemerkt, daß er Items einer bestimmten Produkt-Kategorie unter anderem auch in dem betreffenden Online-Geschäft erwerben kann.

Falls für einen Nutzer keine genügend große Menge an ähnlichen Nutzern existiert, kann ein kollaboratives Verfahren für diesen Nutzer keine guten Schätzer berechnen. Dagegen ist der Erfolg kontent-basierter Verfahren unabhängig von jeder (Ähnlichkeits-)Beziehung zwischen den Nutzern. In der folgenden Tabelle werden die wesentlichen Eigenschaften der kontent-basierten und kollaborativen

Verfahren zusammenfassend gegenübergestellt:

Eigenschaften der kontent-basierten und kollaborativen Verfahren	
kontent-basierte Verfahren	kollaborative Verfahen
Es lassen sich nur gute Empfehlungen für Nutzer bestimmen, die eine ausreichend große Anzahl von Items bewertet haben.	
Die relevanten Eigenschaften der Items müssen quantitativ erhebbar sein und erhoben werden.	Die Eigenschaften der Items werden nicht benutzt.
Es können einem Nutzer nur Items empfohlen werden, die so ähnlich sind wie die Items, die er zuvor positiv bewertet hat.	Auch Items, die völlig anders sind, als alle Items, die der Nutzer vorher positiv bewertet hat, können ihm empfohlen werden.
Es spielt keine Rolle, ob ein Nutzer anderen Nutzern ähnlich ist oder nicht.	Es lassen sich nur gute Empfehlungen für Nutzer bestimmen, für die eine ausreichende Menge ähnlicher Nutzer vorhanden ist.
Selbst Items, die noch von keinem Nutzer bewertet wurden, können empfohlen werden.	Nur Items, die von einer hinreichenden Anzahl von Nutzern bewertet wurden, können empfohlen werden.

Abbildung 5.19: Vergleich zwischen kontent-basierten und kollaborativen Verfahren

Die hybriden Verfahren können als Versuch verstanden werden, durch die Verbindung kontent-basierter und kollaborativer Ansätze die Nachteile beider Verfahren abzuschwächen oder vollständig zu beseitigen.

5.12 Zusammenfassung

In den letzten Jahren sind in der Literatur eine Vielzahl kollaborativer Verfahren vorgeschlagen worden.

Problematisch ist, daß in der Literatur empirische Verfahrensvergleiche vor allem auf Basis des *AAD*-Werts angestellt werden. Der Nutzen der aus den Prognosen resultierenden Empfehlungslisten, welcher durch den Breese-Wert heuristisch quantifiziert werden kann, bleibt hierdurch weitgehend unberücksichtigt. Es konnte in Abschnitt 5.10 gezeigt werden, daß Verfahrensvergleiche auf Basis des Breese-Werts zu unterschiedlichen Ergebnissen führen können als Gegenüberstellungen unterschiedlicher Methoden anhand der resultierenden *AAD*-Werte.

Ein weiterer Fehler der vorherrschenden Vergleichsmethodik ist, daß die Evaluation und Gegenüberstellung der Verfahren unter idealisierten Annahmen vorgenommen wird, die in der Praxis nicht erfüllt sein dürften. Insbesondere ist davon auszugehen, daß sich die Menge der abgegebenen Bewertungen strukturell von den fehlenden Bewertungen unterscheidet, da die Nutzer im allgemeinen eher Items bewerten können, von denen sie sich zu irgendeinem Zeitpunkt etwas versprochen haben. Dieser Umstand wird in der vorliegenden Arbeit erstmals berücksichtigt. Im Rahmen von Verfahrensvergleichen muß dieser Problematik Rechnung getragen werden.

Bezüglich des unter Marketing-Gesichtspunkten wichtigsten Gütemaßes, des Breese-Werts, erzielen das SVD-basierte Verfahren und das zweimodale $\hat{S}^2_Y$- Clusterverfahren bei allen Testdatenanteilen in Bezug auf die unverzerrten D1-Daten Ergebnisse, die denen der übrigen Nicht-Bayes'schen kollaborativen Verfahren i.d.R. deutlich überlegen sind. Dabei ist das SVD-basierte Verfahren meist signifikant besser als das zweimodale $\hat{S}^2_Y$-Clusterverfahren. Gleiches gilt in Bezug auf Präzision und Recall. Trotzdem weisen das zweimodale $\hat{S}^2_Y$-Clusterverfahren, das Verfahren zur ordinalen Matrixfaktorisation und das ordinale zweimodale Clusterverfahren bezüglich desselben Datensatzes bei der gleichen traditionellen Unterteilung in Test- und Trainingsdatensatz bei allen Testdatenanteilen meist erheblich kleinere (und darum bessere) *AAD*-Werte als das SVD-basierte Verfahren auf.

Der durch den Breese-Wert heuristisch quantifizierte Nutzen der resultierenden Empfehlungslisten ist im Hinblick auf Kundenzufriedenheit und Kundenbindung wichtiger als die allgemeine Genauigkeit der den Empfehlungen zugrundeliegenden Prognosen (welche die Nutzer i.d.R. nicht kennen). Daher folgt, daß sich das SVD-basierte Verfahren im Hinblick auf die unverzerrten D1-Daten als das für CRM-Zwecke am besten geeignete Nicht-Bayes'sche kollaborative Verfahren erwiesen hat.

Hinsichtlich des größten Datensatzes, D2, lassen sich bei hohen Testdatenanteilen unter Verwendung des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens höhere Breese-Werte erzeugen, als mittels des SVD-basierten Ansatzes. Zudem führt das zweimodale $\hat{S}_Y^2$-Clusterverfahren bezüglich der unverzerrten D2-Daten insgesamt zu den besten Ergebnissen. Dies kann als ein erheblicher Vorteil des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens betrachtet werden, da D2 der größere Datensatz mit der höheren Fehlendstruktur ist.

Tabelle 5.29 belegt, daß das SVD-basierte Verfahren bei hohen Verzerrungsgraden zu vergleichsweise schlechten Ergebnissen führen kann. In diesem Kontext erweisen sich das zweimodale $\hat{S}_Y^2$-Clusterverfahren, das ordinale zweimodale Clusterverfahren und die Imputation innerhalb zweimodaler Partitionen als geeigneter. Problematisch ist in disem Zusammenhang, daß der reale Verzerrungsgrad unbekannt ist.

Insgesamt sprechen die Ergebnisse für die Verwendung des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens. Anstelle des letzteren können außerdem bei nicht zu hohen Fehlendanteilen mittels des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens zweimodale Partitionen erzeugt werden, um dann innerhalb dieser den $\hat{S}_Y^2$-Schätzer zu imputieren.

Kapitel 6

Nicht-Bayes'sche Hybride Verfahren

Als hybride Verfahren werden alle Ansätze bezeichnet, die sowohl die vorhandenen Bewertungen als auch die relevanten Eigenschaften der vorhandenen Items verwenden. Die Nicht-Bayes'schen hybriden Verfahren zerfallen in drei unterschiedliche Klassen.

6.1 Kombinationsansätze

Kombinationsansätze sind alle hybriden Verfahren, die sowohl auf einem kontentbasierten als auch auf einem kollaborativen Verfahren basieren, die unabhängig voneinander berechnet werden.

Claypool et. al. (1999) schlagen eine gewichtete Addition der mittels des kontentbasierten und des kollaborativen Verfahrens berechneten Schätzer vor. Der Schätzer des kollaborativen Verfahrens wird stärker gewichtet, wenn die verfügbare Information über die Nutzer groß ist und die kollaborativen Verfahren daher bekanntermaßen zu besseren Ergebnissen als die kontentbasierten Verfahren führen. Umgekehrt wird unter den Bedingungen, unter denen die kontentbasierten Verfahren allgemein etwas besser als die kollaborativen Verfahren abschneiden, der Schätzer des kontentbasierten Verfahrens stärker gewichtet.

Good et. al. (1999) verwenden zuerst getrennt ein kontentbasiertes und ein

kollaboratives Verfahren, um Empfehlungen zu generiereren. Für jeden Nutzer werden dann die erfolgversprechendsten auf den kontentbasierten und auf den kollaborativen Verfahren basierenden Empfehlungen miteinander kombiniert. Auf diese Weise können die Vorteile beider Verfahrensklassen miteinander kombiniert werden. Die auf dem kollaborativen Ansatz beruhenden Empfehlungen erhöhen die Vielfalt des Angebots und ermöglichen in stärkerem Maße Cross-Selling, während die mit dem kontentbasierten Verfahren generierten Empfehlungen zielsicher bekannte Interessen des Nutzers berücksichtigen. Tran, Cohen (2000) schlagen vor, für die Nutzer, die weniger Bewertungen als ein vorher festgelegter Schwellenwert abgegeben haben, die mittels des kontentbasierten Verfahren generierten Empfehlungen zu benutzen, und für alle übrigen Nutzer nur die Empfehlungen des kollaborativen Verfahrens zu verwenden.

6.2 Kontentbasierte Ergänzungen kollaborativer Verfahren

Die zweite Unterklasse innerhalb der Hybriden Verfahren bilden die kontentbasierten Ergänzungen kollaborativer Verfahren. Kontentbasierte Ergänzungen kollaborativer Verfahren sind Abwandelungen ursprünglich kollaborativer Verfahren, die Information über die Eigenschaften der Items benutzen, um Schwachstellen des kollaborativen Verfahrens auszugleichen.

Melville et. al. (2002) schlagen vor, alle fehlenden Werte durch mittels eines kontentbasierten Verfahrens berechnete Schätzwerte zu ersetzen. Auf dieser Grundlage sollen dann mit Hilfe der kollaborativen Methode die endgültigen Schätzer berechnet werden.

Basu (1998) benutzt die Genre-Zugehörigkeit von Filmen, um eine Nachbarschaft für den betrachteten Nutzer zu bestimmen. Die Nachbarschaft eines Nutzers besteht dann aus den Nutzern, die ähnlich Präferenzen hinsichtlich des Film-Genres wie der betrachtete Nutzer haben.

Das Verfahren nach Pazzani (1999) stellt eine kontentbasierte Abwandelung des Nutzer-basierten Ähnlichkeitsverfahrens dar. Zunächst müssen die relevanten Eigenschaften der Items von den irrelevanten unterschieden werden. Hierzu ver-

wendet Pazzani (1999) den Winnow-Algorithmus (Littlestone, Warmuth (1994), Blum et. al. (1995)). Jedes Item habe κ_{MAX} Eigenschaften, deren jeweilige Ausprägungen die Komponenten des Vektors

$$\underline{a}_j = \begin{pmatrix} a_{j1} \\ \vdots \\ a_{j\kappa_{MAX}} \end{pmatrix}$$

seien. Es gilt $\underline{a}_{j\kappa} \geq 0, j = 1, ..., J, \kappa = 1, ..., \kappa_{MAX}$.

Der Winnow-Algorithmus ist eine Heuristik zur Ermittlung einer Gewichtungsmatrix $\mathcal{W} \in \mathbb{R}^{I,\kappa_{MAX}}$ zu einem bestimmten konstanten Schwellenwert τ_S.

Die Kontante τ_B ist die kleinste Bewertung $c \in \{1, ..., C\}$ ab der davon ausgegangen werden kann, daß das betrachtete Item dem Nutzer, der die Bewertung abgegeben hat, gefallen hat. τ_{AS} bezeichnet den Schwellenwert, der überschritten werden muß, damit die Ausprägung $\underline{a}_{j\kappa}$ einer bestimmten Eigenschaft κ als stärker ausgeprägt angesehen wird. Je nachdem, ob der Schwellenwert τ_{AS} überschritten wird oder nicht, ist $\mathcal{A}_{j\kappa}$ 1 oder 0:

$$\mathcal{A}_{j\kappa} = \begin{cases} 1, & \text{falls } \underline{a}_{j\kappa} > \tau_{AS} \\ 0, & \text{sonst} \end{cases}$$

und sei weiter $\mathcal{A}_j = (\mathcal{A}_{j1}, ..., \mathcal{A}_{j\kappa_{MAX}})'$. Der binäre Vektor $\mathcal{A}_j$ gibt an, welche Eigenschaften bei dem Item j stärker ausgeprägt sind. Zu Beginn werden die Gewichtungen $\mathcal{W}_{i\kappa}$ aller Nutzer $i \in \{1, ..., I\}$ und Eigenschaften $\kappa = 1, ..., \kappa_A$ auf 1 gesetzt. Es gilt $\mathcal{W}_i = (\mathcal{W}_{i1}, ..., \mathcal{W}_{i\kappa_{MAX}})'$. Dann wird für jede gegebene Bewertung zunächst $\mathcal{S}_{ij} = \mathcal{W}_i'\mathcal{A}_j$ von neuem berechnet. $\mathcal{S}_{ij}$ ist ein grobes Maß dafür, wie sehr der Nutzer i das Item j mag. Falls $\mathcal{S}_{ij}$ kleiner als der Schwellenwert τ_S ist und der betrachtete Nutzer in Bezug auf das j-te Item eine Bewertung abgegeben hat, die darauf schließen läßt, daß ihm das Item gefallen hat, werden die Gewichte $\mathcal{W}_{i\kappa}$ aller Eigenschaften κ, für die $\mathcal{A}_{j\kappa} = 1$ gilt, mit 2 multipliziert. Wenn $\mathcal{S}_{ij} > \tau_S$ gilt und der jeweilige Nutzer das betrachtete Item als weniger gut empfunden zu haben zu haben scheint, werden die Gewichte $\mathcal{W}_{i\kappa}$ aller Eigenschaften κ, für die $\mathcal{A}_{j\kappa} = 1$ gilt, durch 2 dividiert. Dies wird für alle $j \in J_i$ solange wiederholt, bis das Verfahren konvergiert. Auf diese Weise wird die Matrix $\mathcal{W}$ nach und nach zeilenweise bestimmt. Es werden für jeden Nutzer i die Eigenschaften $\kappa \in \{1, ..., \kappa_A\}$

```
Für alle (i ∈ {1, ..., I}) {
        E_{W_i} = 1
        W_i = (1, ..., 1)'
          Solange (E_{W_i} ≠ 0) {
            Für alle (κ ∈ {1, ..., κ_MAX}) {W_{iκ}^{alt} ← W_{iκ}}
              Für alle (j ∈ J_i) {
                S_ij ← W_i' A_j
                 Falls (S_ij < τ_S ∧ y_ij ≥ τ_B) {
                    Für alle (κ ∈ {1, ..., κ_MAX}) {
                       Falls (A_{jκ} = 1) {W_{iκ} ← 2W_{iκ}}
                    }
                  }
                  Falls (S_ij > τ_S ∧ y_ij < τ_B) {
                    Für alle (κ ∈ {1, ..., κ_MAX}) {
                       Falls (A_{jκ} = 1) {W_{iκ} ← (1/2)W_{iκ}}
                    }
                  }
              }
            E_{W_i} = Σ_{κ'=1}^{κ_MAX} (W_{iκ'} − W_{iκ'}^{alt})^2
          }
}
```

Abbildung 6.1: Winnow-Algorithmus

mit den größten Gewichten $\mathcal{W}_{i\kappa}$ ausgewählt. Die Nutzer-Profile werden ähnlich wie bei der TF-IDF Profil-Heuristik mit Hilfe eines Rochio-Ansatzes erzeugt. Mit $\tilde{R}_{i\geq} = \{j \in J_i | y_{ij} \geq \tau_B\}$ und $\tilde{R}_{i<} = \{j \in J_i | y_{ij} < \tau_B\}$ ergibt sich

$$\tilde{g}^*_\kappa(i) = \frac{\alpha_\geq}{|\tilde{R}_{i\geq}|} \sum_{j' \in \tilde{R}_{i\geq}} \underline{a}^*_{j'\kappa} - \frac{\alpha_<}{|\tilde{R}_{i<}|} \sum_{j' \in \tilde{R}_{i<}} \underline{a}^*_{j'\kappa}, \kappa \in \{1, ..., \kappa_{MAX}\},$$

wobei $\alpha_\geq = 4\alpha_<$ gewählt wird. Man erhält das vektorielle Nutzer-Profil

$$g^*_\kappa(i) = \begin{cases} \tilde{g}^*_\kappa(i), & \text{für } \tilde{g}^*_\kappa(i) > 0 \\ 0, & \text{sonst} \end{cases} \qquad \forall\, i \in \{1, ..., I\}, \kappa \in \{1, ..., \kappa_{MAX}\}.$$

Sei $\tau_{\mathcal{W}}$ der Schwellenwert, den ein Element $\mathcal{W}_{i\kappa}$ der Gewichtematrix $\mathcal{W}$ überschreiten muß, damit die zugehörige Eigenschaft κ im Hinblick auf den Nutzer i als relevant eingestuft wird. Weiter sei

$$\delta(\mathcal{W}_{i\kappa} > \tau_{\mathcal{W}}) = \begin{cases} 1, & \text{falls } \mathcal{W}_{i\kappa} > \tau_{\mathcal{W}} \\ 0, & \text{sonst} \end{cases} .$$

Damit erhält man den Mittelwert $\overline{g^*}(i) \in \mathbb{R}$ aller in Bezug auf i relevanten Eigenschaften:

$$\overline{g^*}(i) = \frac{\sum_{\kappa'=1}^{P} \delta(\mathcal{W}_{i\kappa'} > \tau_{\mathcal{W}}) g^*_{\kappa'}(i)}{\sum_{\kappa'=1}^{P} \delta(\mathcal{W}_{i\kappa'} > \tau_{\mathcal{W}})}.$$

Zur Berechnung des modifizierten Bravais-Pearson Korrelationskoeffizienten werden die Nutzer-Profile nach der Vorschrift

$$g^*_{\kappa}(i_1; i_2) = \begin{cases} g^*_{\kappa}(i_1), & \text{falls } \mathcal{W}_{i_1\kappa} > \tau_{\mathcal{W}} \wedge \mathcal{W}_{i_2\kappa} > \tau_{\mathcal{W}} \\ 0, & \text{falls } \mathcal{W}_{i_1\kappa} \leq \tau_{\mathcal{W}} \wedge \mathcal{W}_{i_2\kappa} > \tau_{\mathcal{W}} \\ g^*_{\kappa}(i_1), & \text{falls } \mathcal{W}_{i_1\kappa} > \tau_{\mathcal{W}} \wedge \mathcal{W}_{i_2\kappa} \leq \tau_{\mathcal{W}} \\ \overline{g^*}(i_1), & \text{falls } \mathcal{W}_{i_1\kappa} \leq \tau_{\mathcal{W}} \wedge \mathcal{W}_{i_2\kappa} \leq \tau_{\mathcal{W}} \end{cases}$$

transformiert. Ist eine Eigenschaft nur für genau einen der beiden Nutzer relevant, so setzt die Transformation $g^*_{\kappa}(i_1; i_2)$ die zugehörige Eigenschaft im Profil eines der beiden betrachteten Nutzer auf 0. Damit erhält man den modifizierten Bravais-Person Korrelationskoeffizient

$$\tilde{r}^{g*}_{i_1 i_2} = \frac{\sum_{\kappa'=1}^{\kappa_{MAX}} (g^*_p(i_1; i_2) - \overline{g^*}(i_1))(g^*_{\kappa'}(i_2; i_1) - \overline{g^*}(i_2))}{\sqrt{\sum_{\kappa'=1}^{\kappa_{MAX}} (g^*_{\kappa'}(i_1; i_2) - \overline{g^*}(i_1))^2} \sqrt{\sum_{\kappa'=1}^{\kappa_{MAX}} (g^*_{\kappa'}(i_2; i_1) - \overline{g^*}(i_2))^2}}.$$

Dieser Korrelationskoeffizient berücksichtigt Eigenschaften nicht, die für sowohl i_1 als auch i_2 irrelevant sind. Ist dasselbe Merkmal genau für einen der beiden Nutzer nicht (hinreichend) relevant, dann sind beide Nutzer umso stärker negativ

miteinander korreliert, je höher das zu dem betreffenden Merkmal gehörende Gewicht im Nutzer-Profil des jeweils anderen Nutzers ist.

Mittels der so bestimmten Korrelationen $\tilde{r}_{i\iota}^{g*}, i, \iota \in \{1, ..., I\}, i \neq \iota$ werden dann die Schätzer eines Nutzer-basierten Ähnlichkeitsansatzes ausgerechnet. Es ergibt sich

$$\hat{Y}_{ij} = \bar{y}_{i.} + \frac{\sum_{\iota \in I_j} \tilde{r}_{i\iota}^{g*}(y_{\iota j} - \bar{y}_{\iota.})}{\sum_{\iota \in I_j} |\tilde{r}_{i\iota}^{g*}|}.$$

Beispiel 6.1:

Es geht wieder um die Schätzung von Bernds Bewertung von Barry Lyndon. Als Attribute der Items verwenden wir die Eigenschaften aus Beispiel 4.1, deren Ausprägungen in Tabelle 4.1 wiederzufinden sind.

Zunächst wird die Gewichtematrix $\mathcal{W}$ mit dem Winnow-Algorithmus bestimmt. Mit den Parametern $\tau_S = 10$, $\tau_{AS} = 4$ und $\tau_B = 4$ erhält man die folgende Gewichtematrix:

$$\mathcal{W} = \begin{pmatrix} 2 & 1 & 2 & 2 & 2 & 8 \\ 1 & 1 & 2 & 2 & 2 & 4 \\ 4 & 1 & 2 & 2 & 2 & 8 \\ 4 & 4 & 2 & 2 & 2 & 1 \\ 8 & 8 & 2 & 2 & 4 & 4 \\ 4 & 2 & 2 & 2 & 4 & 8 \end{pmatrix}.$$

Um den Schätzer $\hat{Y}_{17}$ zu berechnen, benötigt man nur zwei Korrelationen, da nur der dritte und der sechste Nutzer Bewertungen für den Film Barry Lyndon abgegeben haben. Mit $\tilde{r}_{13}^{g*} = 0,990$ und $\tilde{r}_{16}^{g*} = 0,968$ und der Konstante $\tau_{\mathcal{W}} = 1$ erhält man

$$\hat{Y}_{17} = 3,5 + \frac{0,990(5 - 3,5) + 0,968(4 - 3,667)}{0,990 + 0,968} = 4,42.$$

Das Verfahren nach Pazzani (1999) wurde auch auf den kleineren Datensatz D1 angewendet. Die Eigenschaften aller 418 Filme wurden der Website *reel.com*

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,713	0,721	0,721	0,724	0,729	0,727	0,735	0,747	0,785
Prec.	0,645	0,596	0,621	0,631	0,608	0,601	0,585	0,571	0,509
Rec.	0,201	0,178	0,182	0,186	0,181	0,180	0,171	0,199	0,202
R_B	85,28	77,19	73,06	70,10	68,59	68,41	68,03	67,17	66,88

Tabelle 6.1: AAD, Präzision, Recall und Breese-Wert (R_B) des Verfahrens nach Pazzani (1999) bezüglich D1

entnommen. (Ursprünglich wurden die Items nach der Verfügbarkeit ihrer Eigenschaften ausgewählt.) Für alle 418 Filme sind 14 verschiedene Eigenschaften gegeben. Die Ausprägungen der Eigenschaften sind ganzzahlige Werte von 0 bis 10.

Wieder wurden die Parameter $\tau_B = 4$, $\tau_S = 10$ und $\tau_{AS} = 4$ verwendet. Da die Elemente der Gewichtematrix fast immer deutlich kleiner als 1 sind, empfiehlt es sich $\tau_{\mathcal{W}}$ ebenfalls deutlich kleiner 1 zu wählen. Optimale Ergebnisse erhält man für die Wahl $\tau_{\mathcal{W}} = 0,03$. AAD, Präzision und Recall des Schätzungen nach der hybriden Methode von Pazzani (1999) für $\tau_{\mathcal{W}} = 0,03$ sind (untereinander) in Tabelle 6.1 aufgelistet.

Bei einem Anteil des Testdatensatzes am gesamten Datensatz von 70% sind die Ergebnisse des Verfahrens nach Pazzani (1999) hinsichtlich AAD, Präzision und Recall vergleichbar mit den Ergebnissen des Verfahrens zur ordinalen Matrixfaktorisation ($\mathcal{C}_{MF} = 10, R = 100$, vgl. Tabelle 5.19) und den Ergebnissen des ordinalen zweimodalen Clusterverfahrens ($K = L = 10$, vgl. Tabelle 5.20). Bei einem Anteilen des Testdatensatzes von $80-90$% der gesamten Datenmenge sind die Ergebnisse des hybriden Ansatzes den Resultaten dieser beiden kollaborativen Verfahren überlegen, bei kleineren Anteilen sind die genannten kollaborativen Verfahren dem hybriden Verfahren überlegen. Ist der Anteil des Testdatensatz an der verfügbaren Datenmage sehr hoch, so stehen nur sehr wenige Bewertungen zur Berechnung der Schätzer zur Verfügung. Dies kann der hybride Ansatz durch die Verwendung von Information über die Items kompensieren. Die Ergebnisse des hybriden Ansatzes übertreffen die des zweimodalen $\hat{S}^1_Y$-Clusterverfahrens (vgl. Tabellen 5.10 und 5.15). Für kleine Anteile des Testdatensatzes an der insge-

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,730	0,719	0,730	0,731	0,733	0,736	0,743	0,762	0,802
Prec.	0,630	0,609	0,622	0,610	0,589	0,594	0,591	0,548	0,499
Rec.	0,205	0,192	0,197	0,192	0,196	0,201	0,192	0,199	0,180
R_B	86,66	79,67	75,78	73,38	71,11	70,33	69,33	68,65	67,61

Tabelle 6.2: AAD, Präzision, Recall und Breese-Wert (R_B) des Verfahrens nach Pazzani (1999) in Bezug auf den 2020 Nutzer und 418 Items umfassenden Ausschnitt aus dem MovieLens-Datensatz D3

samt verwendeten Datenmenge sind die AAD-Werte des hybriden Verfahrens fast deckungsgleich mit den entsprechenden AAD-Ergebnissen des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens. Auch für große Testdatenanteile gilt, daß das zweimodale $\hat{S}_Y^2$-Clusterverfahren sogar zu Ergebnissen führt, die mit denen des hybriden Ansatzes, der zusätzliche Information verwendet, gut vergleichbar sind. Jedoch bleiben in diesem Bereich selbst die Breese-Werte des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens hinter den Breese-Werten des Pazzani-Verfahrens deutlich zurück.

Für hohe Anteile des Testdatensatzes am gesamten Datensatz unterscheiden sich die Resultate des Verfahrens nach Pazzani (1999) vor allem durch vergleichsweise hohe Bresse-Werte von den Ergebnissen der betrachteten kollaborativen Methoden. Der wesentliche Vorteil des hybriden Verfahrens nach Pazzani (1999) gegenüber den kollaborativen Ansätzen kann daher im Breese-Wert der Vorhersagen auf Basis weniger Bewertungen gesehen werden.

In Tabelle 6.2 sind die Ergebnisse des hybriden Verfahrens nach Pazzani (1999) in Bezug auf D3, den mittelgroßen Auschnitt aus dem MovieLens-Datensatz, aufgelistet. Dieser Ausschnitt umfaßt die Bewertungen von insgesamt 2020 Nutzern in Bezug auf dieselben 418 Filme und hat einen etwas höheren Fehlendanteil als der Datensatz, der lediglich Bewertungen von 1067 Nutzern enthält.

Mit Hilfe der Methode von Pazzani lassen sich zwar bei verhältnismäßig wenigen Bewertungen noch gute Schätzer berechnen. Dennoch ist es auch mit diesem Verfahren nicht möglich, Schätzer für Items zu berechnen, die bisher von keinem der Nutzer bewertet wurden. Ebenso ist es auch mit diesem Ansatz zumindest problematisch, Schätzer für Items zu berechnen, hinsichtlich derer nur von sehr

wenigen Nutzern Bewertungen existieren. In diesem Fall existiert in Bezug auf die sehr selten bewerteten Items nicht für jeden Nutzer ein anderer Nutzer, der dieses Item bewertet hat und zudem dem betrachten besonders ähnlich (oder unähnlich) ist.

Pazzani (1999) konnte empirisch belegen, daß sein hybrider Ansatz den damaligen reinen kontentbasierten und kollaborativen Verfahren überlegen ist. Insbesondere war das hybride Verfahren von Pazzani hinsichtlich aller betrachteten Szenarien vor allem dem reinen kontentbasierten Verfahren überlegen. Zur Bestimmung der kontentbasierten Empfehlungen geht Pazzani (1999) sehr ähnlich vor wie bei TF-IDF Profil-Heuristik. Zuerst werden die Nutzer-Profile mit demselben Rochio-Ansatz erstellt, wie im kontentbasierten Verfahren. Danach wird die Ähnlichkeit zwischen Nutzern und Items wie bei der herkömmlichen TF-IDF Profil-Heuristik berechnet. Jedem Nutzer werden die Items empfohlen, deren Ähnlichkeitsmaß (Korrelation, Kosinus-Ähnlichkeit) in Bezug auf sein vektorielles Nutzer-Profil am größten ausfallen.

Bislang war es die in der Literatur vorherrschende Tendenz, daß die hybriden Ansätze zu besseren Ergebnissen führen als reine kontentbasierte und reine kollaborative Ansätze (z.B. Balabanovic, Shoham (1997), Soboroff, Nicholas (1999), Melville et. al. (2002)). Deshalb ist es bemerkenswert, daß *AAD*, Präzision, Recall und Breese-Wert zweier modernerer kollaborativer Verfahren (des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens und des ordinalen zweimodaden Clusterverfahrens) zumindest bei geringen Anteilen des Testdatensatzes an der insgesamt verwendeten Datenmenge besser ausfallen als die entsprechenden Resultate des hybriden Ansatzes. Interessant ist insbesondere, daß sogar wenn verhältnismäßig wenig Daten zur Verfügung stehen, das zweimodale $\hat{S}_Y^2$-Clusterverfahren hinsichtlich *AAD*-Wert, Präzision und Recall fast zu mit den Ergebnissen des hybriden Verfahrens vergleichbaren Resultaten führt, obwohl dieses zusätzliche Information verwendet.

6.3 Selbstständige Hybride Ansätze

Als selbstständige hybride Ansätze werden alle hybriden Verfahren bezeichnet, denen ein Modell zugrundeliegt, das die Eigenschaften der Items mit ihren Bewertungen in eine direkte Beziehung zueinander setzt.

Würden im Regressionsansatz von Natter, Mild (2002) die Eigenschaften der

Items und nicht Bewertungen durch andere Nutzer als Regressoren verwendet, würde es sich um einen hybriden Ansatz handeln. Ein weiteres Beispiel für ein selbstständiges hybrides Verfahren ist der Ansatz zur Textklassifikation nach Soboroff, Nicholas (1999).

6.4 Eigenschaften der hybriden Verfahren

Auch die hybriden Methoden erzielen bessere Resultate, wenn mehr Bewertungen vorhanden sind. Genau wie die kontentbasierten Verfahren erfordern die hybriden Ansätze, daß die relevanten Eigenschaften der Items zum einen quantitativ erhebbar sind und zum anderen vorab erhoben wurden. Mittels aller hybrider Verfahren ist es möglich, Nutzern Items zu empfehlen, die völlig anders sind als jene Items, die sie vorher positiv bewertet haben. Dies macht die hybriden Ansätze gerade unter Marketing-Gesichtspunkten interessant, da auf diese Weise Cross-Selling Chancen besser genutzt werden können. Dagegen ermöglichen nicht alle hybriden Verfahren Empfehlungen für Items, die von noch keinem Nutzer bewertet worden sind. Auf Grundlage der hybriden Kombinationsansätze können Empfehlungen für noch unbekannte Items zwar generiert werden. Diese basieren dann aber in der Regel auf den Ergebnissen des reinen kontentbasierten Verfahrens. Dagegen ist man mit Hilfe der selbstständigen hybriden Verfahren in der Lage, auch Items zu empfehlen, die zuvor von keinem Nutzer bewertet wurden. Da die wichtigsten selbstständigen hybriden Ansätze auf der Bayes'schen Statistik beruhen, werden im Kapitel 9 einige selbstständige hybride Ansätze vorgestellt.

Kapitel 7

Resultate der Nicht-Bayes'schen Verfahren

Im Rahmen der Kapitel 4 bis 6 wurden die wichtigsten Nicht-Bayes'schen Ansätze zur Vorhersage „unvollständiger" Bewertungsdaten vorgestellt, analysiert und verglichen.

Aus der Literatur ist bekannt, daß die Ergebnisse der reinen kollaborativen Verfahren im allgemeinen deutlich besser sind als die der reinen kontentbasierten Verfahren (z.B. Pazzani (1999), Melville et. al. (2002)). Während auch in jüngster Zeit immer wieder neue kollaborative und hybride Verfahren veröffentlicht werden, die sich mit Bewertungsdaten in Bezug auf Produkte wie Filme oder CDs befassen, gibt es keine entsprechenden neueren reinen kontentbasierten Ansätze.

Abbildung 7.1 vergeicht alle verschiedenen Nicht-Bayes'schen Strategien zur Vorhersage ranggeordneter Bewertungsdaten. Die überwiegend verwendete Strategie zum Umgang mit fehlenden Werten ist, sie zu ignorieren. Einige der SVD-basierten Methoden verwenden statt dessen die Mittelwert-Imputation. Mild, Natter (2002) benutzen im Rahmen ihres Regressionsansatzes beide Strategien. Fehlende Realisationen der endogenen Variable werden ignoriert, während die nicht vorhandenen Realisationen der exogenen Variable durch Mittelwerte ersetzt werden. Da alle Variablen in diesem Ansatz die Bewertung bezüglich eines bestimmten Items sind, kann man diesen Ansatz als Verbindung beider beliebter Strategien zum Umgang mit fehlenden Werten interpretieren.

Überblick über die Nicht-Bayes'schen Verfahren (Teil I)

Verfahren	Datenbasis	Datentrafo	Behandlung fehlender Werte	Methode	Skalenniveau
TF-IDF-Profil Heuristik (Pazzani, Billsus (1997), Lang (1995))	kontent-basiert	Null-Eins-Trafo	Ignorieren	speicher-basierte lineare Heuristik	geeignet für binäre nominale, ordinale und kardinale Daten
Neuronale Netze (Pazzani, Billsus (1997))	kontent-basiert	Null-Eins-Trafo	Ignorieren	speicher-basiertes lineares Modell	geeignet für binäre nominale, ordinale und kardinale Daten
Naiver Klassifikator (Pazzani, Billsus (1997), Mooney, Roy (1995))	kontent-basiert	keine	Ignorieren	speicher-basierte lineare Heuristik	Daten werden behandelt als hätten sie kardinales Skalenniveau.
Ähnlichkeitsverfahren (Shardanand, Maes (1995), u.v.a.)	kollaborativ	keine	Ignorieren	speicher-basierte lineare Heuristik	Daten werden behandelt als hätten sie kardinales Skalenniveau.
Regression (Mild, Natter (2002))	kollaborativ	keine	Ignorieren bzw. Mittelwert-Imputation	lineare und nicht-lineare Modelle	Behandlung als kardinale (lineare) bzw. nominale Daten (logistische Regression)
SVD-basierte Verfahren (Sarvar et. al. (2000), Goldberg et. al. (2001),u.v.a.)	kollaborativ	keine Ausnahme: Null-Eins-Trafo bei Billsus und Pazzani (1998)	Ignorieren Ausnahmen: Mittelwert-Imputation (Sarwar et. al. (2000)), Data Augmentation (Srebro u. Jaakkola (2003))	lineare Heuristik	Behandlung als kardinale Daten (Ausnahme: Methode von Billsus und Pazzani (1998))
Gradientenverfahren (Srebro u. Jaakkola (2003))	kollaborativ	keine	Ignorieren	lineare Heuristik	Behandlung als kardinale Daten

Überblick über die Nicht-Bayes'schen Verfahren (Teil II)

Verfahren	Datenbasis	Datentrafo	Behandlung fehlender Werte	Methode	Skalenniveau
Ordinale Matrixfaktorisation (Rennie, Srebro (2003))	kollaborativ	keine	Ignorieren	lineare Heuristik	Strafkostenansatz zur Berücksichtigung des ordinalen Skalenniveaus
Zweimodale Clusterverfahren (Banerjee et. al. (2004), u.v.a.)	kontentbasiert	keine	Ignorieren	lineare Heuristik	Bewertungsdaten werden wie kardinale Daten behandelt.
Ordinales Zweimodales Clusterverfahren	kontentbasiert	keine	Ignorieren	nichtlineare Heuristik	Strafkostenansatz zur Berücksichtigung des ordinalen Skalenniveaus
Verfahren nach Pazzani (1999)	hybrid	Null-Eins Trafo im Rahmen des Winnow-Alg., sonst keine	Ignorieren	nichtlineare Heuristik	Bewertungsdaten werden wie kardinale Daten behandelt

Abbildung 7.1: Überblick über die Nicht-Bayes'schen Verfahren

Die meisten Verfahren können als lineare Heuristiken bezeichnet werden.
Es wurde bereits darauf hingewiesen, daß sowohl das Ignorieren fehlender Werte als auch die Mittelwert-Imputation eigentlich die MCAR-Eigenschaft erfordern. Da ein Zusammenhang besteht zwischen der Tatsache, ob ein Nutzer ein Item bewerten kann, und seiner positiven Einstellung in Bezug auf dieses Item in der Vergangenheit, ist das Fehlen einer Bewertung nicht unabhängig davon, welche Einstellung der Bewertende gegüber dem betreffenden Item hat. Deshalb ist kann nicht vorausgesetzt werden, daß die MAR-Annahme (und folglich die MCAR-Annahme) erfüllt ist. Die festgestellten Ergebnisverschlechterungen mit zunehmendem Verzerrungsgrad illustrieren die möglichen Folgen der gängigen Praxis, nicht vorhandene Daten bei der Berechnung der Schätzer zu ignorieren.

Da die Verfahren dazu dienen, hinsichtlich jedes Nutzers von diesem bisher unbewertete Items zu entdecken, die seinem Geschmack entsprechen, hilft die Information, daß die meisten unbewerteten Items schlechter bewertet werden würden als die bewerteten Gegenstände, nicht weiter. Auch die Idee von Little (1986), Teilmittelwerte geeignet zu gewichten, um hierdurch den zu erwartenden Verzerrungen der Schätzer entgegenzuwirken, führt auf Basis der bekannten zweimodalen Clusterverfahren nicht zu nützlicheren Vorhersagen. Beleg hierfür sind die nur geringfügigen Ergebnisverbesserungen, die sich durch Verwendung des zweimodalen $\hat{S}_Y^{2*}$-Clusterverfahrens gegenüber dem herkömmlichen zweimodalen $\hat{S}_Y^2$-Clusterverfahren erzielen lassen.

Aus der Sicht eines Betreibers eines Recommender-Systems ist es besonders wichtig, daß das System gerade auch auf Basis weniger Bewertungen gute Empfehlungen machen kann. Sind die ersten Empfehlungen für einen neuen Nutzer nicht zufriedenstellend, besteht die große Gefahr, daß die betreffende Person das Recommendersystem in Zukunft nicht mehr nutzt. Daher ist es besonders wichtig, daß ein Verfahren gerade auch auf der Basis von nur wenigen Bewertungen gute Schätzer berechnen kann. Aus diesem Grund kommt den Ergebnissen bei hohen Anteilen des Testdatensatzes an der verwendeten Datenmenge besondere Bedeutung zu.

Mittels des hybriden Verfahrens nach Pazzani (1999) lassen sich bei hohen Testdatenanteilen bezüglich des D1-Datensatzes die besten Breese-Werten erzielen. Deshalb scheint es in besonderem Maße dazu geeignet zu sein, für neu hinzukommende Nutzer, die erst wenige Bewertungen abgegeben haben, Prognosen zu berechnen, auf deren Basis hilfreiche Empfehlungen generiert werden können. Außerdem führt es bei hohen und mittleren Verzerrungsgraden zu den größten Breese-Werten. Deshalb dürfte die Qualität der resultierenden Empfehlungen im Hinblick auf dieses Verfahren weniger stark durch den Unterschied zwischen den (zur Berechnung verwendeten) gegebenen Bewertungen und den zu schätzenden Bewertungen beeinträchtigt werden. Bei kleinen und mittleren Anteilen des Testdatensatzes an der Datenmenge D1 führt das Verfahren nach Pazzani (1999) zu vergleichsweise guten, wenngleich nicht den besten Breese-Werten. Deshalb erweist sich dieses Verfahren in Bezug auf das ökonomisch sinnvollste Gütemaß als beste Option. Wie alle hybriden Verfahren erfordert das hybride Verfahren nach Pazzani (1999) quantitativ erfaßte Item-Eigenschaften und kann daher nicht auf die D2-Daten angewandt werden.

Das SVD-basierte Verfahren nach Sarwar et. al. (2000b) erreicht in Bezug auf die D1-Daten bei hohen Testdatensätzen beinahe ebenso überzeugende Breese-Werte wie das hybride Verfahren nach Pazzani (1999). Bei hohen Verzerrungsgraden führt es allerdings zu sehr niedrigen Breese-Werten. Bezüglich der D2-Daten erzielt man auf Basis dieser Methode nicht die besten Ergebnisse. Wegen der starken Überlegenheit des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens bezüglich des großen Datenteils D2 und hinsichtlich hoher Verzerrungsgrade scheint das zweimodale $\hat{S}_Y^2$-Clusterverfahren besser zur Generierung hilfreicher Prognosen in der Praxis geeignet zu sein.

Das hybride Verfahren nach Pazzani (1999) verwendet ebenso wie das Nutzerbasierte Ähnlichkeitsverfahren vor allem die (im Trainingsdatensatz vorhandenen) Bewertungen anderer Nutzer, die mit dem Nutzer, dessen Bewertungen vorherzusagen sind, stark (positiv oder negativ) korreliert sind, zur Bestimmung der Vorhersagen. Bezüglich eines selten bewerteten Items sind definitionsgemäß wenige Nutzer vorhanden, die dieses Item beurteilt haben. Somit sind unter diesen Voraussetzungen im Hinblick auf das betreffende Item nicht unbedingt Bewertungen von Nutzern verfügbar, die mit dem jeweiligen Nutzer, dessen Bewertung hinsichtlich des betrachteten Items zu prognostizieren ist, stark (positiv oder negativ) korreliert sind. Daher ist davon auszugehen, daß die Bewertungen hinsichtlich dieser Items im allgemeinen nicht besonders stark vom Mittelwert der Bewertungen des betrachteten Nutzers abweichen dürften. Somit ist es vergleichsweise unwahrscheinlich, daß auf Basis des hybriden Verfahrens ein weniger bekanntes Item empfohlen wird. Wegen der Vernachlässigung des Neuigkeitsgrads durch den Breese-Wert sind daher die auf Basis der Methode nach Pazzani (1999) resultierenden Breese-Werte eher ein Beleg dafür, daß im Rahmen dieses Verfahrens unzutreffende Bewertungen vermieden werden, indem überwiegend Empfehlungen für bekannte Items abgegeben werden. Vor diesem Hintergrund erscheinen die auf Basis dieses Verfahrens erreichten hohen Breese-Werte weniger beeindruckend.

Im Rahmen der in den in den Abschnitten 5.10 und 6.2 wiedergegebenen empirischen Untersuchungen wird der Neuigkeitsgrad nur durch den Fehlendanteil des zugrundeliegenden Datensatzes und die verschiedenen Anteile des Testdatensatzes am gesamten zugrundeliegenden Teildatensatz berücksichtigt. D1 enthält überwiegend bekannte Items. Da dieser Teildatensatz insgesamt 94200 Bewertungen enthält, stehen bei 418 Filmen pro Film durchschnittlich 225 Bewertungen zur Verfügung. Dieser Anzahl stehen in Bezug auf den betrachteten Datensatz

lediglich 1067 Nutzer gegenüber. Daher wird hier auch bei hohen Anteilen des Testdatensatzes am gesamten Datensatz eine Situation simuliert, in der die meisten Items als hinreichend bekannt bezeichnet werden können. Da D2 als 337720 Bewertungen bezüglich 3043 Items besteht, sind im Hinblich auf diesen Datensatz im Durchschnitt lediglich 110 Bewertungen pro Item gegeben. Außerdem sind unter den 3043 Items viele, die ohnehin schon deutlich seltener bewertet wurden. Daher entspricht der maximale Testdatenanteil bezüglich des Teildatensatzes D2 einer Situation, in der viele Items vorhanden sind, hinsichtlich derer weniger als 11 Bewertungen vorliegen. Dieser kleinen Anzahl an Bewertungen stehen immerhin 2000 Nutzer gegenüber. Somit entspricht Testdatenanteil von 90 % in Bezug auf D2 einer Situation, in der bezüglich einer hohen Anzahl von Items wenig Bewertungen verfügbar sind.

Vor diesem Hintergrund ist es ein enttäuschendes Ergebnis, daß ausgerechnet das Verfahren, das im Hinblick auf den D2-Datensatz die höchsten Breese-Werte erreicht, diese nur um den Preis erzielt, nicht als Basis zur Empfehlung weniger oft bewerteter Items dienen zu können.

Alle kollaborativen Verfahren und auch das hybride Verfahren nach Pazzani (1999) können nicht zu Vorhersagen in Bezug auf neue Items verwendet werden. Hinsichtlich der Items, die deutlich seltener als die übrigen im Trainingsdatensatz enthaltenen Items bewertet wurden, führen sie zu unzuverlässigen Prognosen. Prognosen bezüglich neuer Items sind ohnehin nicht mittels dieser Ansätze bestimmbar. Da unzutreffende Empfehlungen vermieden werden müssen, sollten auch zur Vorhersage weniger bekannter Items andere Verfahren eingesetzt werden. Das motiviert den Versuch, neue Verfahren zu entwickeln, die zur Empfehlung neuer Items einsetzbar sind.

Während in der Vergangenheit die hybriden Verfahren den reinen kollaborativen Ansätzen in den meisten Fällen in Bezug auf die Präzision deutlich überlegen war (z.B. Pazzani (1999), Melville et. al. (2002)), kann im Hinblick auf das hybride Verfahren nach Pazzani (1999) und das zweimodale $\hat{S}_Y^2$-Clusterverfahren selbst dann nicht von einer deutlichen Überlegenheit des hybriden Verfahrens hinsichtlich AAD, Präzision und Recall gesprochen werden, wenn durchschnittlich pro Nutzer nur ca. 11 Bewertungen im Testdatensatz verfügbar sind. Das illustriert den mit dem zweimodalen $\hat{S}_Y^2$-Clusterverfahren und vergleichbaren moderneren kollaborativen Verfahren einhergehenden Genauigkeitsgewinn. Vor diesem Hintergrund ist die Frage berechtigt, ob man nicht ob nicht durch eine kontent-

basierte Erweiterung des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens gute Bewertungen hinsichtlich neuer Items erreichen könnte.

Die Genauigkeit der mittels des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens berechenbaren Schätzer ist ein deutlicher Hinweis auf die besondere Bedeutung, die der Berücksichtigung der Heterogenität zukommt. Die Hierarchischen Bayes'schen Verfahren sind in besonderem Maße dazu geeignet, der Heterogenität Rechnung zu tragen. Daher bietet liegt der Versuch nahe, zu versuchen, ob man auf Basis eines hybriden Hierarchischen Bayes'schen Ansatzes hilfreiche Empfehlungen hinsichtlich neuer und weniger bekannter Items machen kann.

Kapitel 8

Bayes'sche Verfahren

In diesem Kapitel werden zunächst die Grundlagen der Bayes'schen Statistik erläutert, die für das Verständnis der Bayes'schen Verfahren erforderlich sind. Danach werden diese Grundlagen benutzt, um die Funktionsweise und die Vorzüge der Hierarchischen Ansätze zu erläutern. Sodann werden ein kollaboratives hierarchisches Bayes'sches Verfahren und zwei hybride hierarchische Bayes'sches Methoden empirisch miteinander verglichen. Zudem werden verschiedene Möglichkeiten zur Variablenselektion im Hinblick auf Hierarchische Ansätze diskutiert.

8.1 Grundlagen der Bayes'schen Statistik

Im Rahmen der Bayes'schen Statistik werden sowohl die Daten $\mathcal{D}$ als auch die Parameter θ eines Modells zur Auswertung der Daten als Zufallsgrößen behandelt. θ bezeichnet hier allgemein die Parameter. Je nach Zusammenhang kann es sich bei θ um einen Vektor oder um ein Skalar handeln. Nach dem Satz von Bayes besteht zwischen der bedingten Wahrscheinlichkeitsfunktion $P(\mathcal{D}|\theta)$ und der Wahrscheinlichkeitsfunktion $P(\theta)$ die Beziehung

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \propto P(\mathcal{D}|\theta)P(\theta).$$

Das Zeichen „$\propto$“ bedeutet „ist proportional zu“. $P(\mathcal{D}|\theta)$ wird auch Likelihood genannt. Die bedingte Wahrscheinlichkeitsfunktion $P(\theta|\mathcal{D})$ wird als Posterior von θ bezeichnet. $P(\theta)$ ist die Prior von θ und kann jede Art von verfügbarer Information über θ enthalten, die unabhängig von den Daten $\mathcal{D}$ ist. So können via der Prior auch Informationen aus vielerlei anderen Quellen explizit in die Analyse

der Daten $\mathcal{D}$ miteinbezogen werden. Bei den anderen Quellen kann es sich um die Ansichten von Experten (Sandor, Wedel (2001)), um Theorien (Montgomery, Rossi (1999)) oder einfach um andere Datensätze (Kamakura, Wedel (1997), Wedel, Pieters (2000), Hofstede et. al. (2002)) handeln. Wichtig ist, daß hierdurch zusätzliche Information explizit zum Bestandteil der Analyse der Daten $\mathcal{D}$ gemacht wird. Die Posterior $P(\theta|\mathcal{D})$ vereint die durch die Daten $\mathcal{D}$ repräsentierte und in der Likelihood $P(\mathcal{D},\theta)$ enthaltene Information mit der in der Prior $P(\theta)$ quantifizierten von $\mathcal{D}$ unabhängigen Ausgangsinformation, die aus der Analyse anderer (z.B. zuvor erhobener) Daten, Expertenbefragungen oder anderen Quellen (s.o.) stammen kann. Je nach Wahl der Prior kann der Einfluß der von $\mathcal{D}$ unabhängigen Information auf die Posterior gering oder auch hoch sein. Bei Wahl einer wenig informativen Prior enthält die Posterior nicht viel mehr Information, als aus den in der Likelihood $P(\mathcal{D}|\theta)$ verwendeten Daten $\mathcal{D}$ gewonnen werden kann. Z.B. ist die Wahl der Gleichverteilung als Prior möglich, die sich lediglich auf die Angabe eines Intervalls für den Wertebereich von θ beschränkt. Umgekehrt ist es auch möglich, die Prior so zu wählen, daß sie sehr genaue Aussagen in Bezug auf θ macht. Sofern die Likelihood nur wenige Daten enthält, wird die Posterior in diesem Fall sehr ähnlich wie die Prior ausfallen (siehe z.B. Zellner (1971)). Daher ist zu beachten, daß man durch jede Wahl einer Prior nicht nur eine Aussage über θ selbst macht, sondern darüberhinaus das Ausmaß der mit dieser Aussage verbundenen Sicherheit quantifiziert. Letzteres sollte immer bewußt und kritisch erfolgen.

Es existieren auch sogenannte nichtinformative Priors, die vollkommener Unwissenheit hinsichtlich θ Rechnung tragen sollen (z.B. Jeffreys (1961), Berger, Bernardo (1992)).

Es ist vorteilhaft, daß neu hinzukommende und von $\mathcal{D}$ unanhängige Daten $\mathcal{D}'$ sich im Rahmen der Bayes'schen Statistik immer mit den alten Erkenntnissen aus der vorherigen Posterior $P(\theta|\mathcal{D})$ in Einklang bringen lassen, indem man die alte Posterior zur neuen Prior macht:

$$P(\theta|\mathcal{D}',\mathcal{D}) \propto P(\mathcal{D}',\mathcal{D}|\theta)P(\theta) = P(\mathcal{D}'|\theta)P(\mathcal{D}|\theta)P(\theta) \propto P(\mathcal{D}'|\theta)P(\theta|\mathcal{D}).$$

Die so berechnete neue Posterior enthält dann sowohl die Vorinformation aus der vorherigen Prior $P(\theta)$, als auch die in den Daten $\mathcal{D}$ und $\mathcal{D}'$ enthaltene empirische Information.

Nicht für jede Likelihood und jede Prior kann die Posterior analytisch bestimmt werden. In vielen Fällen ist die analytische Bestimmung der Posterior sehr aufwendig (vgl. z.B. Zellner (1971)). Sofern aber die Formeln der Likelihood und der Wahrscheinlichkeitsdichte der Prior eine ähnlich Form haben, ist häufig der Verteilungstyp der Posterior derselbe wie der der Prior. In diesen Fällen läßt sich die Posterior analytisch bestimmen.

Beispiel 8.1:

In dem folgenden fiktiven Beispiel geht es um die Berechnung der Wahrscheinlichkeit, mit der ein Konsument aus dem Segment $\tilde{t}$ ein Produkt j_X kauft. Im Segment $\tilde{t}$ sind 10^4 Kunden. Leider sind nur die Kaufentscheidungen eines kleinen Teils dieser Personengruppe ($n_B^* = 10$) dokumentiert worden. Von diesen 10 Konsumenten, deren Kaufentscheidung bekannt ist, haben 5 das Produkt j_X erworben. Faßt man die Anzahl der Personen, die das Produkt j_X gekauft haben, als binomialverteilte Zufallsvariable N_B^* auf, die die Werte $0, 1, ..., n_B^* = 10$, annehmen kann, so ergibt sich die Likelihood

$$P(N_B^* = 5|\theta) = \left(\begin{array}{c} n_B^* \\ 5 \end{array} \right) \theta^5 (1-\theta)^{n_B^*-5} \propto \theta^5 (1-\theta)^5.$$

Der Parameter θ der Binomialverteilung, der die Kaufwahrscheinlichkeit angibt, wird als stochastische Größe aufgefaßt, deren Prior zu bestimmen ist. Das Management hat vorab die Verkaufsmitarbeiter nach ihren Erfahrungen befragt. Obwohl die meisten Mitarbeiter eher konservative Prognosen hinsichtlich der zu erwartenden Kaufwahrscheinlichkeit θ abgegeben haben, gab es sogar Mitarbeiter, die für θ Werte aus dem Intervall $[0,8, 0,9]$ angegeben haben. Eine Abverkaufswahrscheinlichkeit von mehr als 90% hielt jedoch keiner der Befragten für realistisch. Als Ergebnis dieser Befragung ergab sich das in Abbildung 8.1 dargestellte Histogramm, für das die Beta-Verteilung als Approximation herangezogen werden kann. Allgemein gilt für die Beta-Verteilung

$$Beta(\theta|\alpha_B, \beta_B) = \frac{\Gamma(\alpha_B + \beta_B)}{\Gamma(\alpha_B)\Gamma(\beta_B)} \theta^{\alpha_B - 1} (1-\theta)^{\beta_B - 1}.$$

Abbildung 8.2 verdeutlicht, wie unterschiedlich die Beta-Verteilung je nach Wahl

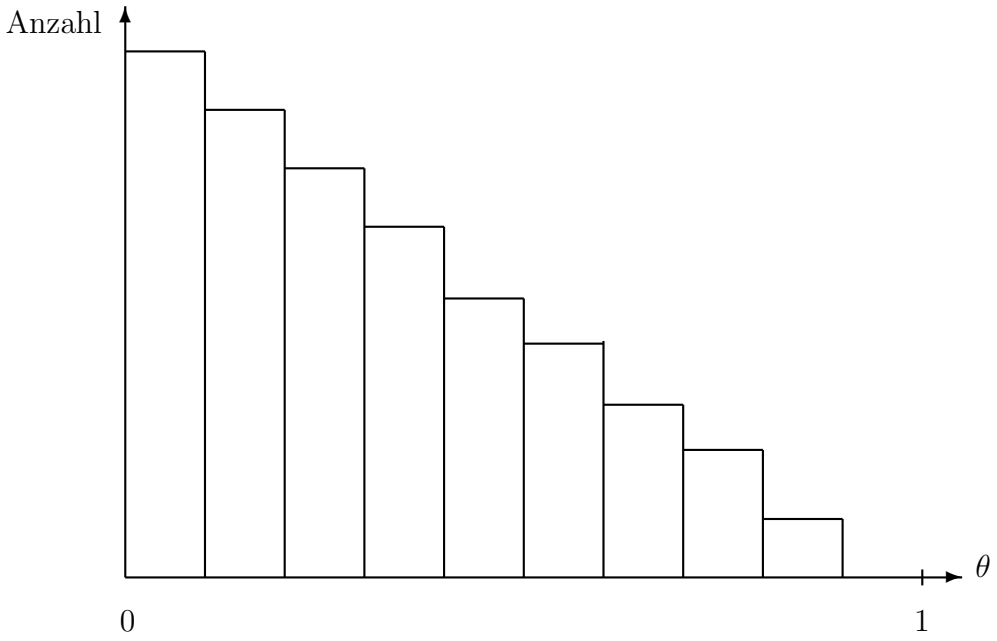

Abbildung 8.1: Vorinformation über θ aufgrund der Befragung der Verkaufsmitarbeiter der Unternehmung (Beispiel 8.1)

der Parameter α_B und β_B ausfallen kann. Die Vorinformation aus Abbildung 8.1 läßt sich durch die Beta-Verteilung mit den Parametern $\alpha_B = 1$ und $\beta_B = 2$ (die gepunktete Linie aus Abbildung 8.2) approximieren. Man erhält somit die Prior

$$P(\theta) = Beta(\theta|\alpha_B = 1, \beta_B = 2) = \frac{\Gamma(3)}{\Gamma(1)\Gamma(2)}(1-\theta) \propto (1-\theta)$$

Daraus ergibt sich in diesem Beispiel wegen

$$\begin{aligned} P(\theta|N_B^* = 5) &\propto P(N_B^* = 5|\theta)P(\theta) \\ &\propto \theta^5(1-\theta)^5\theta^{\alpha_B-1}(1-\theta)^{\beta_B-1} = \theta^{\alpha_B+5-1}(1-\theta)^{\beta_B+5-1} \end{aligned}$$

mit $\alpha_B = 1, \beta_B = 2$ die Beta-Verteilung $Beta(\theta|1+5, 2+5) = Beta(\theta|6, 7)$ als Posterior.
Der Erwartungswert der Beta-Verteilung $Beta(\theta|\alpha'_B, \beta'_B)$ ist

$$E(\theta) = \frac{\alpha'_B}{\alpha'_B + \beta'_B}.$$

Abbildung 8.2: Beta-Verteilung $Beta(\theta|\alpha_B, \beta_B)$ bei unterschiedlichen Parameterkombinationen (α_B, β_B)

In diesem Beispiel erhält man

$$E(\theta|N_B^* = 5) = \frac{\alpha_B + 5}{\alpha_B + 5 + \beta_B + 5} = \frac{6}{6 + 7} = 0,462.$$

Wenn keine Vorinformation vorläge, müßte eine nichtinformative Verteilung als Prior gewählt werden. In diesen Fall bietet sich $Beta(\theta|1, 1)$ an (siehe Abbildung 8.2). Damit ergibt sich für die Posterior, wenn keinerlei Vorinformation Bestandteil der Analyse werden soll $P(\theta|N_B^* = 5) = Beta(1 + 5, 1 + 5) = Beta(6, 6)$ mit Erwartungswert

$$E(\theta|N_B^* = 5) = \frac{\alpha_B + 5}{\alpha_B + 5 + \beta_B + 5} = \frac{6}{6 + 6} = 0,5.$$

Ohne Vorinformation ergibt sich mittels der Bayes'schen Statistik dasselbe Resultat, das man auch mit Hilfe der klassischen Statistik erhalten hätte.

Allgemein gilt: Wählt man die Prior $P(\theta)$ aus einer Familie von parametrischen Verteilungen, die aus sogenannten zur verwendeten Likelihood *konjugierten* Priors bestehen, dann gehört die Posterior $P(\theta|\mathcal{D})$ zur selben Familie wie die Prior. Damit sind dann die Momente der Posterior bekannt. Eine kurze Übersicht über die zu bestimmten bekannten Likelihoods konjugierten Verteilungen bietet Tabelle 8.1.

Likelihood	konjugierte Verteilung
Binomialverteilung	Beta-Verteilung
Multinomialverteilung	Dirichlet-Verteilung
Poisson-Verteilung	Gamma-Verteilung
Multivariate Normalverteilung mit bekannter Kovarianzmatrix	Multivariate Normalverteilung
Multivariate Normalverteilung mit unbekannter Kovarianzmatrix	Mischverteilung aus Multivariater Normalverteilung und inverser Wishart-Verteilung

Tabelle 8.1: Likelihoods und ihre konjugierten Verteilungen

Zumindest für in der Praxis häufig verwendeten Likelihoods aus der ersten Spalte von Tabelle 8.1 sind konjugierte Verteilungen bekannt. Manchmal ist dem Problem aber eine Prior aus der zur gegebenen Likelihood konjugierten Verteilungsfamilie nicht angemessen. In diesen Fällen muß die Posterior auf andere Weise bestimmt werden.

Wenn es möglich wäre, aus dieser Posterior Werte zu ziehen, könnte man die Posterior immerhin simulieren. Leider kann es extrem schwierig werden, Zufallszahlen aus einer beliebigen multivariaten Verteilung zu ziehen. Deshalb verwendet man andere Methoden.

8.1.1 Der Metropolis-Hastings Algorithmus

Eingesetzt werden sogenannte Markov Chain Monte Carlo (MCMC) Verfahren, wie z.B. der Metropolis-Hastings Algorithmus. Die diesen Ansätzen zugrundeliegende Idee ist, eine Markovkette auf dem Parameterraum zu erzeugen, so daß sich die Posterior als stationäre Verteilung der Markovkette ergibt und daher dazu verwendet werden kann, die Posterior zu simulieren.

Der Metropolis-Hastings Algorithmus kann dazu benutzt werden, um Ziehungen aus einer beliebigen Verteilung vorzunehmen, sofern die Abhängigkeit der zugehörigen Wahrscheinlichkeitsfunktion von θ bekannt ist. In der vorliegenden Arbeit wird der Metropolis-Hastings Algorithmus dazu verwendet, Ziehungen aus der Posterior zu erhalten.

Die folgenden Ausführungen sind für ein Verständnis des Metropolis-Hastings Algorithmus wichtig. Weil es die Notation erheblich vereinfacht, werden hier zunächst nur diskrete Zustände betrachtet. Per Konvention bezeichnet der obere Index im Ausdruck θ^r den Zustand, in dem sich die Zufallsvariable θ gerade befindet. Der untere Index im Ausdruck θ_n bezeichnet den betrachteten Zeitpunkt. Daher bedeutet die Gleichung $\theta_n = \theta^r$, daß der Prozeß zur Zeit n im Zustand r ist. Für eine Markovkette erster Ordnung gilt für die Übergangswahrscheinlichkeit vom n-ten in den $(n+1)$-ten Zustand:

$$\Pi_{rs}(n) = p(\theta_{n+1} = \theta^s | \theta_n = \theta^r, \theta_{n-1} = \theta^{r_{n-1}}, ..., \theta_0 = \theta^{r_0}) = p(\theta_{n+1} = \theta^s | \theta_n = \theta^r).$$

Ist die Markovkette homogen, kann n weggelassen werden, so daß

$$\Pi_{rs} = p(\theta_{n+1} = \theta^s | \theta_n = \theta^r).$$

gilt. Sei M_{state} die Menge aller möglichen Zustände ($r, s \in M_{state}, |M_{state}| = \varrho$). Da irgendeiner der möglichen Zustände beim Übergang in den Folgezustand $n+1$ angenommen werden muß, gilt $\sum_{s \in M_{state}} \Pi_{rs} = 1, \quad \forall\, r \in M_{state}$. Die aus den Übergangswahrscheinlichkeiten zusammengesetzte Matrix

$$P_T = \begin{pmatrix} \Pi_{00} & \cdots & \Pi_{0\varrho} \\ \vdots & \ddots & \vdots \\ \Pi_{\varrho 0} & \cdots & \Pi_{\varrho\varrho} \end{pmatrix}$$

heißt Übergangs- bzw. Transitionsmatrix.
Sei die Wahrscheinlichkeit $p(\theta_n = \theta^s) = \pi_{ns}$ und $\pi'_n = (\pi_{n1}, ..., \pi_{n\varrho})$ die zugehörige (diskrete) Wahrscheinlichkeitsverteilung mit

$$\sum_{s \in M_{state}} \pi_{ns} = 1 \quad .$$

Die Wahrscheinlichkeitsverteilung π_1 ergibt sich aus der vorherigen Wahrscheinlichkeitsverteilung π_0 durch die Vorschrift $\pi'_1 = \pi'_0 P_T$. Durch iteratives Einsetzen ergibt sich außerdem $\pi'_n = \pi'_0 P_T^n$. Man sagt, daß eine stationäre Verteilung π (Grenzverteilung) vorliegt, wenn $\pi' = \pi' P_T$ gilt. Z.B. ist der Vektor $(\frac{1}{3}, \frac{2}{3})'$ stationäre Verteilung zur Übergangsmatrix

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$

Um eine stationäre Verteilung zu erhalten, benutzt man den Begriff der Zeitumkehrbarkeit oder Zeitreversibilität. Zeitumkehrbarkeit bedeutet, daß sich das Übergangsverhalten der Markovkette nicht verändert, wenn man die Reihenfolge der Markovkette umkehrt. Für eine rückwärtslaufende Kette gilt

$$p(\theta_n = \theta^s | \theta_{n+1} = \theta^{r_1}, ..., \theta_{n+v} = \theta^{r_v}) = \frac{p(\theta_n = \theta^s, \theta_{n+1} = \theta^{r_1}, ..., \theta_{n+v} = \theta^{r_v})}{p(\theta_{n+1} = \theta^{r_1}, ..., \theta_{n+v} = \theta^{r_v})} =$$

$$\frac{p(\theta_n = \theta^s) p(\theta_{n+1} = \theta^{r_1} | \theta_n = \theta^s) p(\theta_{n+2} = \theta^{r_2}, ..., \theta_{n+v} = \theta^{r_v} | \theta_n = \theta^s, \theta_{n+1} = \theta^{r_1})}{p(\theta_{n+1} = \theta^{r_1}) p(\theta_{n+2} = \theta^{r_2}, ..., \theta_{n+v} = \theta^{r_v} | \theta_{n+1} = \theta^{r_1})}.$$

Die Markov-Eigenschaft der vorwärtslaufenden Kette impliziert, daß

$$\begin{aligned} &p(\theta_{n+2} = \theta^{r_2}, ..., \theta_{n+v} = \theta^{r_v} | \theta_n = \theta^s, \theta_{n+1} = \theta^{r_1}) \\ &= p(\theta_{n+2} = \theta^{r_2}, ..., \theta_{n+v} = \theta^{r_v} | \theta_{n+1} = \theta^{r_1}). \end{aligned}$$

Daher folgt:

$$p(\theta_n = \theta^s | \theta_{n+1} = \theta^{r_1}, ..., \theta_{n+v} = \theta^{r_v}) = \frac{p(\theta_n = \theta^s) p(\theta_{n+1} = \theta^{r_1} | \theta_n = \theta^s)}{p(\theta_{n+1} = \theta^{r_1})}.$$

Also hat die rückwärtslaufende Kette ebenfalls die Markov-Eigenschaft:

$$p(\theta_n = \theta^s | \theta_{n+1} = \theta^{r_1}, ..., \theta_{n+v} = \theta^{r_v}) = p(\theta_n = \theta^s | \theta_{n+1} = \theta^{r_1}).$$

Sei $r = r_1$ und $P_T^* = (\Pi_{rs}^*)$ die Übergangsmatrix der rückwärtslaufenden Kette. Wegen

$$p(\theta_{n+1} = \theta^r) p(\theta_n = \theta^s | \theta_{n+1} = \theta^r) = p(\theta_n = \theta^s) p(\theta_{n+1} = \theta^r | \theta_n = \theta^s)$$

folgt dann $\pi_{n+1,r}\Pi_{rs}^* = \pi_{ns}\Pi_{sr}$. Zeitreversibilität oder Zeitumkehrbarkeit bedeutet nun, daß $\Pi_{rs}^* = \Pi_{rs}$ für alle $r, s \in M_{state}$ gilt. Ist eine Markovkette im Hinblick auf eine bestimmte Wahrscheinlichkeitsverteilung ω zeitreversibel, so gilt die Gleichung $\omega_r \Pi_{rs} = \omega_s \Pi_{sr}$. Unter dieser Bedingung folgt

$$\sum_{r \in M_{state}} \omega_r \Pi_{rs} = \omega_s \sum_{r \in M_{state}} \Pi_{sr} = \omega_s$$

und damit $\omega' P_T = \omega'$. Falls eine Markovkette in Bezug auf eine Verteilung ω zeitreversibel ist, ist ω daher auch stationäre Verteilung der Kette.

Die Begingung $\pi_r \Pi_{rs} = \pi_s \Pi_{sr}, r, s \in M_{state}$, kann benutzt werden, um mittels eines MCMC Verfahrens und einer Übergangsvorschrift ϑ zur stationären Verteilung zu gelangen.

Der Metropolis-Hastings Algorithmus (Metropolis et. al. (1953), Hastings (1970)) erzeugt eine Markovkette mit π als stationärer Verteilung, indem eine zu Beginn benutzte Markovkette geeignet modifiziert wird. Hierzu beginnt man mit einer beliebigen Transformationsvorschrift ϑ.

Für diskrete Zustände handelt es sich hierbei um eine Übergangsmatrix

$$\vartheta = \begin{pmatrix} \vartheta_{00} & \cdots & \vartheta_{0\varrho} \\ \vdots & \ddots & \vdots \\ \vartheta_{\varrho 0} & \cdots & \vartheta_{\varrho\varrho} \end{pmatrix}.$$

Ausgehend vom Zustand θ^r zieht man einen Zustand θ^s mit der Wahrscheinlichkeit ϑ_{rs}. Auf Basis dieser beliebig bzw. nach praktischen Gesichtspunkten zu

wählenden Übergangsmatrix kann man nun die modifizierte Übergangsmatrix $(\tilde{\Pi}_{rs}) = (\vartheta_{rs}\alpha_P(\theta^r, \theta^s))$ mit

$$\alpha_P(\theta^r, \theta^s) = \min\left\{1, \frac{\pi_s \vartheta_{sr}}{\pi_r \vartheta_{rs}}\right\}$$

berechnen. Hierfür wird nur das Verhältnis π_s/π_r benötigt. Das erspart die Berechnung der Normierungskonstanten der Verteilung π.
Wegen

$$\pi_r \vartheta_{rs} \min\left\{1, \frac{\pi_s \vartheta_{sr}}{\pi_r \vartheta_{rs}}\right\} = \min\left\{\pi_r \vartheta_{rs}, \pi_s \vartheta_{sr}\right\} = \pi_s \vartheta_{sr} \min\left\{1, \frac{\pi_r \vartheta_{rs}}{\pi_s \vartheta_{sr}}\right\}$$

gilt $\pi_r \tilde{\Pi}_{rs} = \pi_s \tilde{\Pi}_{sr}$, weshalb die modifizierte Übergangsmatrix $(\tilde{\Pi}_{rs})$ die Grundlage für eine Markovkette bildet, deren stationäre Verteilung π ist.

Im folgenden wird die Posterior $P(\theta^\phi|\mathcal{D})$ zur Vereinfachung der Notation als $\mathcal{P}(\theta^\phi)$ bezeichnet. Für kontinuierliche Zustände θ^ϕ und θ^φ gilt dieselbe Argumentationskette wie für diskreten Zustände θ^r und θ^s. Wenn es darum geht, Werte aus der Posterior $\mathcal{P}(\theta^\phi)$ zu ziehen, ergibt sich für kontinuierliche Zustände und der Transformationsvorschrift für kontinuierliche Zustände $\vartheta(\theta^\phi, \theta^\varphi)$ der Metropolis-Hastings Algorithmus gemäß Abbildung 8.3:

Wähle einen Startwert $\theta_0 = \theta^\phi$, $n = 0$.
Solange $n < n_{MAX}$:

1. ziehe $\theta^\varphi \sim \vartheta(\theta_n, \bullet)$
2. berechne $\alpha_P(\theta_n, \theta^\varphi) = \min\left\{1, \dfrac{\mathcal{P}(\theta^\varphi)\vartheta(\theta^\varphi, \theta_n)}{\mathcal{P}(\theta_n)\vartheta(\theta_n, \theta^\varphi)}\right\}$
3. $\theta_{n+1} = \begin{cases} \theta^\varphi, & \text{mit Wahrscheinlichkeit } \alpha_P(\theta_n, \theta^\varphi) \\ \theta_n, & \text{mit Wahrscheinlichkeit } 1 - \alpha_P(\theta_n, \theta^\varphi) \end{cases}$
4. $n \leftarrow n + 1$

Abbildung 8.3: Metropolis-Hastings Algorithmus

θ_n steht für den im n-ten Schritt angenommenen Zustand. Falls die Posterior hinreichend wohldefiniert (d.h. z.B. im gesamten Parameterraum positiv) ist, konvergieren alle Metropolis-Hastings Algorithmen gegen die Posterior (Tierney (1994)).

Man geht davon aus, daß sich die Kette während der n_{BURN} ersten Iterationen in der sogenannten Einbrennphase befindet, während der die Markovkette gegen die Posterior konvergiert. Nach der Einbrennphase entsprechen die Zustände der Kette Ziehungen aus der Posterior. Die ersten $n_{BURN} < n_{MAX}$ Iterationen werden in der Praxis nicht zur Berechnung der Schätzer verwendet.

Zur Überprüfung, ob Konvergenz nach den ersten $n_{V,B}$ Iterationen eingetreten ist, existieren verschiedene Methoden (Gelfand, Smith (1990), Gelman, Rubin (1992a),(1992b), Geweke (1992), Johnson (1996)). Auf Basis dieser Methoden kann man n_{BURN} näherungsweise bestimmen (siehe Anhang E).

Mittels der $n_{MAX} - n_{BURN}$ verbleibenden Ziehungen lassen sich die Momente der Posterior empirisch bestimmen. Ist θ multivariat normalverteilt, so sind die Momente der Verteilung der Erwartungswert und die Varianz-Kovarianzmatrix. Der Mittelwert der $n_{MAX} - n_{BURN}$ Zufallszahl(vektoren) konvergiert gemäß dem starken Gesetz der großen Zahl für Markovketten (Breiman (1959)) im Limes $n_{MAX} - n_{BURN} \to \infty$ fast sicher (mit Wahrscheinlichkeit 1) gegen deren Erwartungswert auf Basis der Posterior. Daher ist dieser Mittelwert das Bayes'sche Analogon des Schätzers.

Beispiel 8.2:

Es geht erneut um die Wahrscheinlichkeit, mit der ein Kunde aus dem Segment $\tilde{t}$ ein Produkt j_X erwirbt. Wie bereits aus Beispiel 8.1 bekannt ist, ist die Posterior $\mathcal{P}(\theta) \propto \theta^5(1-\theta)^6$. Der Startwert sei $\theta_0 = \theta^\phi = 0,7$. Damit ergibt sich

$$\mathcal{P}(\theta_0) \propto (0,7)^5(0,3)^6 = 1,23 \cdot 10^{-4}.$$

Als Transformationsvorschrift ϑ wird die Normalverteilung mit $\sigma^2 = 1$ gewählt:

$$\vartheta(\theta_0, \theta^\varphi) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\theta^\varphi - \theta_0)^2}{2}\right\}.$$

Die Transformationsverschrift ϑ wird nun dazu benutzt, den Wert θ^φ zu ziehen:

$$\theta^\varphi \sim \vartheta(\theta_0 = 0,7, \bullet).$$

Angenommen, es ergibt sich auf diese Weise $\theta^\varphi = 0,66$. Man erhält

$$\mathcal{P}(\theta^{\varphi}) \propto (0,66)^5(0,34)^6 = 1,93 \cdot 10^{-4}.$$

Wegen der Symmetrie der Normalverteilung gilt $\vartheta(\theta_n, \theta^{\varphi}) = \vartheta(\theta^{\varphi}, \theta_0)$. Damit erhält man

$$\alpha_P(\theta_0, \theta^{\varphi}) = \min\left\{1, \frac{\mathcal{P}(\theta^{\varphi} = 0,66)}{\mathcal{P}(\theta_0 = 0,7)}\right\} = \min\left\{1, \frac{1,93 \cdot 10^{-4}}{1,23 \cdot 10^{-4}}\right\} = 1.$$

Daher ist $\theta^{\varphi} = 0,66$ mit Wahrscheinlichkeit 1 das nächste Glied der Folge

$$\theta_1 = \theta^{\varphi} = 0,66$$

und die erste Iteration ist beendet. Wäre $\alpha_P(\theta_0, \theta^{\varphi})$ kleiner als eins gewesen, so wäre eine weitere Zufallszahlziehung erforderlich gewesen. Falls man nach 1000 Iterationen dieser Art unterstellt, daß die Zeitreihe $\{\theta_n\}_{n=1001}^{1500}$ vom verwendeten Startwert unabhängig ist, erhält man mit $\widehat{E(\theta)} = \frac{1}{500}\sum_{n=1001}^{1500}\theta_n$ einen Schätzer für θ. Im vorliegenden Fall ergeben sich $\widehat{E(\theta)} = \frac{1}{500}\sum_{n=1001}^{1500}\theta_n = 0,462$ und damit dasselbe Ergebnis wie in Beispiel 8.1.

Der in Beispiel 8.2 vorgestellte Metropolis-Hastings Algorithmus entspricht wegen der Symmetrie der Übergangsfunktion ϑ dem ursprünglichen Algorithmus von Metropolis et. al. (1953).

8.1.2 Gibbs-Sampling

Ein Spezialfall des Metropolis-Hastings Algorithmus ist das Gibbs-Sampling (Geman, Geman (1984), Gelfand, Smith (1990)). Oft ist es möglich, den Parametervektor θ in disjunkte Teilmengen von Parametern $\theta^{[b]}, b = 1, ..., B$, aufzuspalten, so daß $\theta = (\theta^{[1]'}, ..., \theta^{[B]'})'$ gilt und bedingte Posteriors bestimmbar sind, die die selbe Form haben, wie die folgenden (zur Erklärung des Gibbs-Algorithmus verwendeten) Gibbs-Posteriors

$$G_b(\theta^{[b]}|\theta^{[-b]}) = \frac{\mathcal{P}(\theta)}{\int \mathcal{P}(\theta^{[b]}, \underline{\theta}^{[-b]})d\underline{\theta}^{[-b]}}, \quad b \in \{1, ..., B\},$$

wobei der Vektor

$$\theta^{[-b]} = (\theta^{[1]'}, ..., \theta^{[b-1]'}, \theta^{[b+1]'}, ..., \theta^{[B]'})'$$

alle Komponenten von θ bis auf $\theta^{[b]}$ enthält.
Dabei kann jede Parametergruppe $\theta^{[b]}, b = 1, ..., B$, einen oder mehrere Parameter beinhalten. Beim Gibbs-Sampling benutzt man sukzessive die B bedingten Verteilungen, um neue Werte für eine Parametergruppe $\theta^{[b]}, b = 1, ..., B$, zu ziehen. Hierbei werden immer die zuletzt gezogenen Parameter(gruppen) zur Ziehung des jeweils nächsten Parameters (der jeweils nächsten Parametergruppe) benutzt:

Wähle einen Startwert $\theta_0 = (\theta_0^{[1]}, ..., \theta_0^{[B]})'$, $n = 0$.
Solange $n < n_{MAX}$ ziehe rekursiv:

$$\begin{aligned}
\theta_{n+1}^{[1]} &\sim G_1(\bullet|\theta_n^{[2]}, ..., \theta_n^{[B]}) \\
\theta_{n+1}^{[2]} &\sim G_2(\bullet|\theta_{n+1}^{[1]}, \theta_n^{[3]}, ..., \theta_n^{[B]}) \\
\theta_{n+1}^{[3]} &\sim G_3(\bullet|\theta_{n+1}^{[1]}, \theta_{n+1}^{[2]}, \theta_n^{[4]}, ..., \theta_n^{[B]}) \\
&\vdots \\
\theta_{n+1}^{[B]} &\sim G_B(\bullet|\theta_{n+1}^{[1]}, ..., \theta_{n+1}^{[B-1]}) \\
n + 1 &\leftarrow n
\end{aligned}$$

Abbildung 8.4: Gibbs-Sampling Algorithmus

Das Gibbs-Sampling kann als ein Spezialfall eines Metropolis-Hastings Algorithmus aufgefaßt werden, der innerhalb jeder n-ten Iteration B Ziehungen mittels

$$\vartheta^{[b]}(\theta_n, \theta^{\varphi}) = \begin{cases} G_b(\theta^{\varphi[b]}|\theta_{n,n+1}^{[-b]}), & \text{falls } \theta^{\varphi[-b]} = \theta_{n,n+1}^{[-b]} \\ 0, & \text{sonst} \end{cases} \quad \text{für } b \in \{1, ..., B\},$$

vornimmt, aufgefaßt werden (Gelman et. al. (1995)). Hier wird die Definition $\theta_{n,n+1}^{[-b]} = (\theta_{n+1}^{[1]'}, ..., \theta_{n+1}^{[b-1]'}, \theta_n^{[b+1]'}, ..., \theta_n^{[B]'})'$ für $b \in \{1, ..., B\}$ verwendet. Weiter ist

$$\alpha_P^{[b]}(\theta_n, \theta^{\varphi}) = \min\left\{1, \frac{\mathcal{P}(\theta^{\varphi})\vartheta^{[b]}(\theta^{\varphi}, \theta_n)}{\mathcal{P}(\theta_n)\vartheta^{[b]}(\theta_n, \theta^{\varphi})}\right\} = \min\left\{1, \frac{\mathcal{P}(\theta^{\varphi})G_b(\theta_{n,n+1}^{[b]}|\theta^{\varphi[-b]})}{\mathcal{P}(\theta_n)G_b(\theta^{\varphi[b]}|\theta_{n,n+1}^{[-b]})}\right\}.$$

Wegen $\mathcal{P}(\theta^{\varphi}) = \mathcal{P}(\theta^{\varphi[-b]}, \theta^{\varphi[b]}) = \mathcal{P}(\theta^{[-b]}_{n,n+1}, \theta^{\varphi[b]}) = \mathcal{P}(\theta^{[-b]}_{n,n+1}) G_b(\theta^{\varphi[b]} | \theta^{[-b]}_{n,n+1})$, der Bedingung $\theta^{\varphi[-b]} = \theta^{[-b]}_{n,n+1}$ und $\mathcal{P}(\theta_n) = \mathcal{P}(\theta^{[b]}_n, \theta^{[-b]}_{n,n+1}) = \mathcal{P}(\theta^{[-b]}_{n,n+1}) G_b(\theta^{[b]}_n | \theta^{[-b]}_{n,n+1})$ ist

$$\alpha^{[b]}_P(\theta_n, \theta^{\varphi}) = \min \left\{ 1, \frac{\mathcal{P}(\theta^{[-b]}_{n,n+1}) G_b(\theta^{\varphi[b]} | \theta^{[-b]}_{n,n+1}) G_b(\theta^{[b]}_n | \theta^{\varphi[-b]})}{\mathcal{P}(\theta^{[-b]}_{n,n+1}) G_b(\theta^{[b]}_n | \theta^{[-b]}_{n,n+1}) G_b(\theta^{\varphi[b]} | \theta^{[-b]}_{n,n+1})} \right\} = 1.$$

Daher ist das Gibbs-Sampling eine Variante des Metropolis-Hastings Algorithmus, bei der jede Iteration B Metropolis-Hastings Ziehungen mit einem konstanten $\alpha^{[b]}_P(\theta_n, \theta^{\varphi}) = 1, b = 1, ..., B$, erfordert.

In der heutigen Bayes'schen Literatur findet man selten reine Metropolis-Hastings Algorithmen oder Gibbs-Sampler. Die meisten neueren Ansätze kombinieren in jeder Iteration Ziehungen nach dem Vorbild des Gibbs-Samplers mit Ziehungen nach dem Metropolis-Hastings Schema (z.B. Chien, George (1999)).

Der Einfluß der Likelihood auf die Posterior von θ steigt mit zunehmender Datenmenge und abnehmendem Informationsgehalt der Prior. Je informativer die Prior ist (je kleiner beispielsweise im Fall einer Normalverteilung die Varianz gewählt ist), umso stärker ist ihr Einfluß auf die resultierende Posterior von θ. Dies ist bei der Bestimmung der Prior zu beachten. Man sollte der Unsicherheit der Vorinformation durch die Prior kritisch Rechnung tragen (d.h. im Fall einer Normalverteilung sollte die Varianz eher größer als die vermutete Varianz gewählt werden), da sonst der Einfluß der in der Likelihood enthaltenen Daten $\mathcal{D}$ möglicherweise zu gering ist. Auf keinen Fall darf die Prior so gewählt werden, daß sie eine Sicherheit bezüglich der Vorinformation suggeriert, die gar nicht vorhanden ist, da sonst eine verzerrte Posterior zu erwarten ist. Im Hinblick auf die Parametrisierung der Prior ist eher zu einem gesunden Mißtrauen gegenüber der Vorinformation zu raten.

8.1.3 Hierarchische Verfahren

Da die Posterior nach der Erhebung der Daten $\mathcal{D}$ auf Grundlage der Prior berechnet wird, steht sie in der Hierarchie über der Prior. Als hierarchische Modelle werden diejenigen Modelle bezeichnet, die zusätzlich zu dieser Hierarchie eine weitere Hierarchieebene innerhalb der Prior benutzen. Dazu wird hier, um eine durch einen Parameter τ beschreibbare Verteilung P explizit zu bezeichnen, die

Notation P_τ verwendet. So wie die bislang vernachlässigten Parameter τ_P der Prior $P(\theta) = P_{\tau_P}(\theta)$ die Ziehungen von θ beeinflussen, bestimmen die sogenannten Hyperprior-Parameter τ_H im Rahmen der Hierarchischen Modelle die Verteilung der Prior-Parameter τ_P. Damit setzt sich die Prior der Hierarchischen Ansätze aus der Hyperprior $P_{\tau_H}(\tau_P)$ und der bisher verwendeten Prior $P_{\tau_P}(\theta)$ zusammen:

$$P(\theta, \tau_P) = P_{\tau_P}(\theta) P_{\tau_H}(\tau_P).$$

Damit ergibt sich für die Posterior statt $P(\theta|\mathcal{D})$ die explizite Bezeichnung

$$P(\theta, \tau_P|\mathcal{D}) \propto P(\mathcal{D}|\theta, \tau_P) P(\theta, \tau_P) = P(\mathcal{D}|\theta, \tau_P) P_{\tau_P}(\theta) P_{\tau_H}(\tau_P).$$

(In Beispiel 8.1 war $\tau_P = (\alpha_B, \beta_B)'$.) Ebenso wie τ_P die Gestalt der Prior im Rahmen der in Abschnitt 8.1 betrachteten Modelle explizit mitbestimmt und a priori gegeben war, legt τ_H explizit und a priori die zusammengesetzte Prior der Hierarchischen Ansätze fest. Im Kontext der Hierarchischen Modelle ist τ_P nicht konstant, sondern muß ebenso wie θ bestimmt werden. Da τ_P aber auch im Hierarchischen Ansatz die Grundlage für die Bestimmung von θ bildet, steht τ_P in der Hierarchie auf einer höheren Stufe als θ. Diese hierarchische Beziehung zwischen τ_P und θ bildet die Grundlage jedes Hierarchischen Modells. Während τ_P meist nur sehr allgemeine Tendenzen in den Daten $\mathcal{D} = (\mathcal{D}_1, ..., \mathcal{D}_Z)$ wiederspiegelt, kann θ für jede von Z Untereinheiten (wie z.B.: Daten des z-ten Individuums, Daten in Bezug auf das z-te Cluster von Individuen, Daten hinsichtlich der z-ten Geschäftsstelle) individuell - allerdings auf Basis von τ_P - bestimmt werden. Für ein solches Modell verwendet man die Prior

$$P(\theta^{(1)}, ..., \theta^{(Z)}, \tau_P) = P_{\tau_H}(\tau_P) \prod_{z=1}^{Z} P_{\tau_P}(\theta^{(z)}).$$

Dies führt auf die Posterior

$$P(\theta^{(1)}, ..., \theta^{(Z)}, \tau_P|\mathcal{D}) \propto P(\mathcal{D}|\theta^{(1)}, ..., \theta^{(Z)}, \tau_P) P_{\tau_H}(\tau_P) \prod_{z=1}^{Z} P_{\tau_P}(\theta^{(z)}).$$

Es wird im Rahmen dieser Modelle typischerweise vorausgesetzt, daß die Likelihood unabhängig von τ_P ist und alle Datenuntereinheiten $D_z, z = 1, ..., Z$, unabhängig identisch verteilt sind. Daher gilt für die Likelihood:

$$P(\mathcal{D}|\theta^{(1)}, ..., \theta^{(Z)}, \tau_P) = P(\mathcal{D}_1, ..., \mathcal{D}_Z|\theta^{(1)}, ..., \theta^{(Z)}) = \prod_{z=1}^{Z} P_z(\mathcal{D}_z|\theta^{(z)})$$

Zur Bestimmung geeigneter bedingten Posteriors, die die Funktion der Gibbs-Posteriors erfüllen können, benutzt man die Unabhängigkeit der Likelihood von τ_P und die Unabhängigkeit der Prior von den Daten. Meistens werden außerdem Beziehungen zwischen der Likelihood und der Prior und zwischen der Prior und der Hyperprior verwendet. Hierbei ist die folgende Überlegung hilfreich: Für festes τ_P wäre nur noch die Beziehung zwischen der Likelihood und der Prior zu berücksichtigen. Falls die Prior konjugierte Verteilung zur Likelihood ist, ist die zugehörige Posterior bekannt. Daher lassen sich für einen festen Wert von τ_P (und gegebene Daten D_z) die Parameter der Likelihood $\theta^{(z)}, z = 1, ..., Z$, bestimmen und es ist möglich, aus dieser Verteilung Werte Realisationen von $\{\theta^{(z)}\}_{z=1}^{Z}$ zu ziehen. Für gegebene Werte von $\{\theta^{(z)}\}_{z=1}^{Z}$, spielt nur noch die Beziehung zwischen der Prior und Hperprior eine Rolle. Falls die Hyperprior konjugierte Verteilung zur Prior ist, ist auf Basis fester Werte für $\{\theta^{(z)}\}_{z=1}^{Z}$ auch eine entsprechende Verteilung für den Prior-Parameter τ_P gefunden, aus der dieser gezogen werden kann. Auf diese Weise lassen sich die notwendigen bedingten Posteriors bestimmen, die für den Gibbs-Sampling Algorithmus benötigt werden. Es ergeben sich die Beziehungen:

$$\begin{array}{lll} \text{(i)} & \theta^{(z)} & |\tau_P, \mathcal{D}_z, \ \forall z \ \{1, ..., Z\} \\ \text{(ii)} & \tau_P & |\{\theta^{(z)}\}_{z=1}^{Z} \end{array}$$

Hier bedeutet $\theta^{(z)}| \quad \tau_P, \mathcal{D}_z$, daß für jeden Wert von $\theta^{(z)}$ auf Basis fester Werte für τ_P und $\mathcal{D}_z$ bestimmt bzw. gezogen wird.

Abhängig von der gewählten Likelihood, Prior und Hyperprior kann es sein, daß die einzelnen $\theta^{(z)}, z = 1, ..., Z$, und bzw. τ_P selbst Parametergruppen sind, die selbst geeignet zerlegt werden müssen, so daß sich ihrerseits für sie bedingte Verteilungen ergeben. Das hierzu notwendige Vorgehen wird am Beispiel des linearen hierarchischen Regressionsmodells nach Rossi et. al. (1996) genauer erklärt.

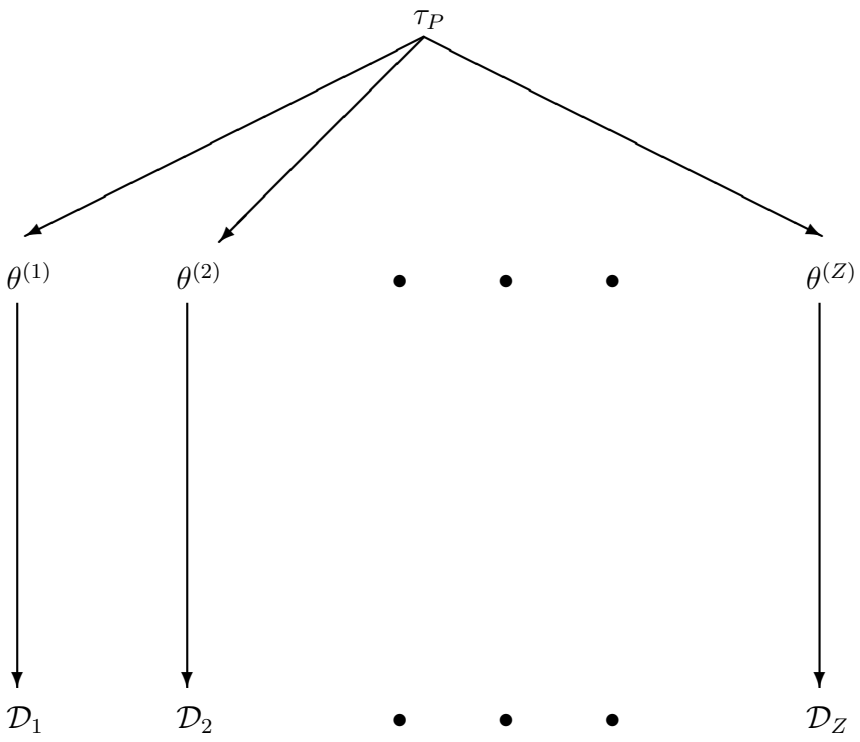

Abbildung 8.5: Struktur eines typischen Hierarchischen Modells (nach Gelman et. al. (1995))

In die Verteilung jedes der Zufallswerte $\theta^{(z)}, z = 1, ..., Z$, aus Teilschritt (i) fließen zum einen via τ_P allgemeine Tendenzen und zum anderen durch den z-ten Datenteil selbst die Besonderheiten der jeweils betrachteten Datenuntereinheit, $\mathcal{D}_z, z = 1, ..., Z$, mit ein.

Da die Verteilung jedes $\theta^{(z)}$ auf den Daten $\mathcal{D}_z$ (und τ_P selbst) basiert, werden auf diese Weise indirekt die gesamten Daten $\mathcal{D} = \{\mathcal{D}_z\}_{z=1}^{Z}$ zur Grundlage der Verteilung von τ_P. Dadurch, daß alle $\{\theta^{(z)}\}_{z=1}^{Z}$ (bzw. indirekt alle $\{\mathcal{D}_z\}_{z=1}^{Z}$) gemeinsam die Grundlage der Verteilung von τ_P bilden, werden die Unterschiede zwischen den verschiedenen $\theta^{(z)}$ in Bezug auf τ_P nivelliert. Insofern spiegelt die Verteilung von τ_P die allgemeinen Tendenzen der Daten wieder. Sind zu einer Untereinheit $\mathcal{D}_z, z = 1, ..., Z$, nur wenig Daten vorhanden, so kommt den durch den τ_P repräsentierten allgemeinen Tendenzen mehr Gewicht bei der Bestimmung von $\theta^{(z)}, z = 1, ..., Z$, zu. Auf diese Weise werden allgemeine Tendenzen in der gesamten Datenmenge $\mathcal{D}$ dazu benutzt, um fehlende Information hinsichtlich der z-ten

Untereinheit zu kompensieren. Ist dagegen die Datenmenge $\mathcal{D}_z$ vergleichsweise groß, steigt der Einfluß der Likelihood auf die Verteilung von $\theta^{(z)}, z = 1, ..., Z$. In diesem Fall treten die individuellen Tendenzen der z-ten Untereinheit stärker in den Vordergrund (Gelman et. al. (1995)). Mit dieser Methode ist es möglich, individuelle $\theta^{(z)}$-Verteilungen für individuelle Dateneinheiten $\mathcal{D}_z$ zu bestimmen, für die beispielsweise mit einer OLS-Regression keine individuellen Schätzer hätten bestimmt werden können. In den Fällen, in denen genügend Daten $\mathcal{D}_z$ vorhanden wären, um allein auf der Grundlage von $\mathcal{D}_z$ einen OLS-Schätzer zu berechnen, der genausoviele Komponenten wie $\theta^{(z)}$ hat, führen die Hierarchischen Bayes'schen Verfahren zu genaueren Ergebnissen als herkömmliche Verfahren (Gelman et. al. (1995)). Ein weiterer Vorteil der hierarchischen Bayes'schen Ansätze ist, daß durch die Hierarchie der Parameter Abhängigkeiten zwischen den Parametern erzeugt werden. Hierdurch ist es möglich, mit einer verhältnismäßig hohen Anzahl von Parametern zu arbeiten und dennoch eine Überanpassung an den betrachteten Datensatz (Overfitting) zu vermeiden (Gelman et. al. (1995)). Dies ist besonders in den Fällen vorteilhaft, in denen das Ziel die Vorhersage noch unbekannter Daten ist.

8.1.4 Bayes'sche Regressorenselektion

In diesem Abschnitt werden verschiedene Bayes'sche Vorgehensweisen dafür betrachtet, die relevanten Regressoren zu bestimmen. Verschiedene Kombinationen von Regressoren werden hier als unterschiedliche Modelle bezeichnet. Z.B. kann $\mathcal{M}^3$ das lineare Regressionsmodell mit der endogenen Größe Y_v und den Regressoren X_v^1 und X_v^2 sein und $\mathcal{M}^4$ kann das lineare Regressionsmodell mit der abhängigen Variable Y_v und den unabhängigen Variablen X_v^1, X_v^3 und X_v^4 sein.

Der traditionelle Bayes'sche Ansatz zum Vergleich zweier Modelle $\mathcal{M}^1$ und $\mathcal{M}^2$ ist der Bayes-Faktor

$$B(\mathcal{M}^2; \mathcal{M}^1|\mathcal{D}) = \frac{\left(\frac{P(\mathcal{M}^2|\mathcal{D})}{P(\mathcal{M}^1|\mathcal{D})}\right)}{\left(\frac{P(\mathcal{M}^2)}{P(\mathcal{M}^1)}\right)}.$$

Hier ist

$$P(\mathcal{D}|\mathcal{M}^\mu) = \int P(\mathcal{D}|\tilde{\theta}^\mu, \mathcal{M}^\mu) P(\tilde{\theta}^\mu|\mathcal{M}^\mu) d\tilde{\theta}^\mu, \;\; \mu = 1, 2,$$

wobei $P(\mathcal{D}|\theta^\mu, \mathcal{M}^\mu)$ die Likelihood eines Modells $\mathcal{M}^\mu$ mit Parametern θ^μ, $\mu = 1, 2$, und Daten $\mathcal{D}$ ist. $P(\mathcal{M}^\mu)$ kann als Prior des Modells $\mathcal{M}^\mu$, $\mu = 1, 2$, interpretiert werden. $P(\theta^\mu|\mathcal{M}^\mu)$ ist die Prior des Parameters θ^μ unter der Bedingung, daß das Modell $\mathcal{M}^\mu$ verwendet wird. Aufgrund des Satzes von Bayes gilt

$$B(\mathcal{M}^2; \mathcal{M}^1|\mathcal{D}) = \frac{\left(\frac{P(\mathcal{D}|\mathcal{M}^2)P(\mathcal{M}^2)}{P(\mathcal{D})}\right)\left(\frac{P(\mathcal{D})}{P(\mathcal{D}|\mathcal{M}^1)P(\mathcal{M}^1)}\right)}{\left(\frac{P(\mathcal{M}^2)}{P(\mathcal{M}^1)}\right)} = \frac{P(\mathcal{D}|\mathcal{M}^2)}{P(\mathcal{D}|\mathcal{M}^1)}.$$

Ist $B(\mathcal{M}^2; \mathcal{M}^1|\mathcal{D})$ deutlich größer als eins, so sollte $\mathcal{M}^2$ Modell $\mathcal{M}^1$ vorgezogen werden. Sofern der Bayes-Faktor viel kleiner als 1 ist, sprechen die Daten $\mathcal{D}$ eher für Modell $\mathcal{M}^1$. Da die Berechnung des Integrals

$$P(\mathcal{D}|\mathcal{M}^\mu) = \int P(\mathcal{D}|\tilde{\theta}^\mu, \mathcal{M}^\mu)P(\tilde{\theta}^\mu|\mathcal{M}^\mu)d\tilde{\theta}^\mu, \;\; \mu = 1, 2,$$

insbesondere für Modelle mit sehr vielen Parametern oft schwer und aufwendig ist, kommt man immer mehr von diesem Ansatz ab. Da nur jeweils zwei Modelle miteinander verglichen werden können, ist dieses Modell ungeeignet, um eine kleine Menge relevanter Regressoren aus einer Vielzahl möglicher Regressoren zu extrahieren.

Das häufigste verwendete Gütemaß ist das DIC (Spiegelhalter et. al. (2002)). Der Definition des DIC liegt die unskalierte Devianz $D_u(\mathcal{D}; \theta) = -2 \log L(\theta|\mathcal{D})$ zugrunde. Hier bezeichnet $L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)$ die Likelihood. Der Mittelwert der unskalierten Devianz auf Basis der Posterior

$$\overline{D_u(\mathcal{D}; \theta)} = \frac{1}{n_{MAX} - n_{BURN}} \sum_{n=n_{BURN}+1}^{n_{MAX}} D_u(\mathcal{D}; \theta_n)$$

wird in der Praxis oft dazu benutzt, um Modelle informell miteinander zu vergleichen (Zeger, Karim (1991), Gilks et. al. (1993), Richardson, Green (1997)). Je höher der Posterior-Mittelwert der unskalierten Devianz ausfällt, umso schlechter ist die Anpassung des Modells an die Daten $\mathcal{D}$. Es ist immer möglich, diese Größe durch die Aufnahme irrelevanter Regressoren in das Modell zu verkleinern. Deshalb wird zum Posterior-Mittelwert der unskalierten Devianz noch ein Strafkostenterm p_D hinzuaddiert:

$$DIC = \overline{D_u(\mathcal{D};\theta)} + p_D,$$

wobei $p_D = \overline{D_u(\mathcal{D};\theta)} - D_u(\mathcal{D};\bar{\theta})$, wobei meistens

$$\bar{\theta} = \frac{1}{n_{MAX} - n_{BURN}} \sum_{n=n_{BURN}+1}^{n_{MAX}} \theta_n$$

ist (seltener: Median oder Modalwert). p_D wird als effektive Anzahl der Parameter interpretiert (Spiegelhalter et. al. (2002)). (Es gilt $DIC = D_u(\mathcal{D};\bar{\theta}) + 2p_D$.)

Da man nach Schätzung des Bayes'schen Modells in der Praxis meist ohnehin $\theta_n, n = n_{BURN}+1, ..., n_{MAX}$, zur Verfügung hat, ist die Berechnung des DIC oft einfach.

Die Modellwahl unter Verwendung des DIC kann ähnlich wie im Rahmen der Nicht-Bayes'schen Statistik mittels der Steinmetz-Methode („backward selection", „top down approach"), der Maurer-Methode („forward selection","bottom up approach") oder einer Kombination aus beiden Verfahren erfolgen (van der Linde (2005)). Bei der Maurer-Methode wird mit einem sehr kleinen Modell, das zu Anfang noch keine irrelevanten Regressoren enthält, begonnen. Diesem Modell wird dann sukzessive solange jeweils eine weitere Variable neu hinzugefügt, bis keine im Modell noch nicht enthaltene Variable mehr existiert, durch deren Hinzunahme sich das DIC verringern ließe. Bedient man sich hingegen des Steinmetz-Verfahrens, so wird ein sehr großes Anfangsmodell gewählt, das alle auch nur möglicherweise relevanten Variablen umfaßt. Nach und nach wird dann ein weiterer Regressor aus dem Modell weggelassen, bis im Modell kein Regressor mehr vorhanden ist, durch dessen Fehlen sich das DIC verringern ließe.

Es ist bekannt, daß das DIC in manchen Fällen dazu neigt, komplexere Modelle zu bevorzugen (van der Linde (2005)). In der Literatur (Spiegelhalter et. al. (2002)) findet sich die Empfehlung, nicht immer das Modell mit dem kleinsten DIC zu wählen, sondern vielmehr mehrere der Modelle mit verhältnismäßig kleinem DIC in Betracht zu ziehen. Damit wird das DIC eines Modells nicht zur alleinigen Grundlage der Modellwahl, sondern hat vielmehr die Funktion eine begrenzte Zahl von etwas besser geeigneten Modellen auszuwählen, aus denen dann nach anderen Kriterien (beispielsweise ökonomischen) das sinnvollste Modell ausgewählt werden sollte.

Insbesondere für die Auswahl aus einer großen Menge zur Verfügung stehender Variablen ist die automatisierte MCMC Modellbestimmung für lineare Modelle (Madigan, York (1995), Hoeting et. al. (1996), Raftery et. al. (1997)) eine naheliegende Option. Im Rahmen dieser Verfahren wird eine Markovkette $\{\mathcal{M}_n, n = 1, 2, ...\}$ im diskreten Zustandsraum der Modelle mit Gleichgewichtsverteilung $\mathcal{P}(\mathcal{M}_n) = P(\mathcal{M}_n|\mathcal{D})$ erzeugt. Sei $\{\mathcal{M}^1, ..., \mathcal{M}^{\Re}\}$ die Menge aller betrachteten Modelle bzw. Variablenkombinationen. Dann ist

$$\mathcal{P}(\mathcal{M}^\mu) = P(\mathcal{M}^\mu|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M}^\mu)P(\mathcal{M}^\mu)}{\sum_{\aleph=1}^{\Re} P(\mathcal{D}|\mathcal{M}^\aleph)P(\mathcal{M}^\aleph)}$$

die Posterior von $\mathcal{M}^\mu$. Man verwendet einen Metropolis-Hastings Algorithmus mit einer symmetrischen Übergangsvorschrift für den Übergang von $\mathcal{M}$ zu $\mathcal{M}'$. Damit ergibt sich die zu verwendende Wahrscheinlichkeit für den Übergang von $\mathcal{M}$ nach $\mathcal{M}'$

$$\alpha_P(\mathcal{M}, \mathcal{M}') = \min\left\{1, \frac{\mathcal{P}(\mathcal{M}')}{\mathcal{P}(\mathcal{M})}\right\}.$$

In $\alpha_P(\mathcal{M}, \mathcal{M}')$ kürzt sich die Summe $\sum_{\aleph=1}^{\Re} P(\mathcal{D}|\mathcal{M}^\aleph)P(\mathcal{M}^\aleph)$ heraus. Mittels der Folge $\mathcal{M}_n, n = n_{BURN}+1, ..., n_{MAX}$, läßt sich die Posterior aller einzelnen Modelle $\mathcal{M}^\mu, \mu = 1, ..., \Re$, durch

$$\hat{\mathcal{P}}(\mathcal{M}^\mu) = \frac{1}{n_{MAX} - n_{BURN}} \sum_{n=n_{BURN}+1}^{n_{MAX}} \delta(\mathcal{M}_n = \mathcal{M}^\mu), \quad \mu = 1, ..., \Re,$$

mit

$$\delta(\mathcal{M}_n = \mathcal{M}^\mu) = \begin{cases} 1, & \text{falls } \mathcal{M}_n = \mathcal{M}^\mu \\ 0, & \text{sonst} \end{cases}$$

abschätzen. Es ist allerdings zu beachten, daß die Anzahl der verglichenen Modelle $\Re \ll n_{MAX} - n_{BURN}$ sein muß. Bei großen Anzahlen von Variablen A_{var}^{max} ist es wegen $\Re = 2^{A_{var}^{max}}$ nicht zweckmäßig, alle möglichen Variablenkombinationen zu verwenden. In diesen Fällen empfiehlt es sich, auf Methoden der Versuchsplanung zurückzugreifen.

Falls im Rahmen eines hierarchischen Ansatzes mit $\mathcal{D}_z = (Y_{\mathcal{D}}^{(z)}, X_{\mathcal{D}}^{(z)})$ und $\theta^{(z)} = (\theta_1^{(z)'}, \theta_2^{(z)'})', z = 1, ..., Z$, die Größen $X_{\mathcal{D}}^{(z)}$ und $\theta_1^{(z)}$ im allgemeinen für jede Untereinheit $\mathcal{D}_z$ verschieden ausfallen und zudem das Modell die lineare Beziehung $Y_{\mathcal{D}}^{(z)} = X_{\mathcal{D}}^{(z)}\theta_1^{(z)} + \epsilon^{(z)}, E(\epsilon^{(z)}) = 0, var(\epsilon^{(z)}) = \tilde{\sigma}_z, z = 1, ..., Z$, enthält, kann für jede einzelne dieser Gleichungen eine MCMC Modellbestimmung durchgeführt werden. Auf diese Weise wird die Heterogenität der $\mathcal{D}_z$, mit $z = 1, ..., Z$, berücksichtigt.

Dabei sollte die vorgegebene Anzahl maximaler Regressoren A_{reg}^{max} eher klein gewählt werden. Es kann dann ein balancierter unvollständiger Blockplan (BIBD) mit $a^V = A_{var}^{max}$ Behandlungsausprägungen und $n_a^V = A_{reg} = A_{reg}^{max} - 1$ Variablen in einem Modell (bzw. Block) bestimmt werden. Auf diese Weise tritt jede einzelne Variable in der gleichen Anzahl r^V von Modellen (Blöcken) auf und jedes Variablenpaar kommt genau in λ^V Modellen (Blöcken) vor. Zusätzlich können noch alle Modelle betrachtet werden, die durch Weglassen einer Variable oder Hinzufügen einer noch nicht im Modell enthaltenen Variable aus den mittels des balancierten unvollständigen Blockplans generierten Modellen (Variablenblöcken) erhalten werden können. So ist die Anzahl der Variablen in jedem einzelnen Modell zwar klein aber weniger starr.

Man verwendet dann für alle Z verschiedene Teildatensätze dieselben $b^V = \frac{r^V a^V}{n_a^V}$ Modelle und ihre durch Weglassen oder Hinzufügen eines Regressors aus ihnen erzeugbaren Varianten. Für jede der betrachteten A_{var}^{max} Variablen kann dann ermittelt werden, wie häufig sie insgesamt in einem der Z einzelnen Teilmodelle mit maximaler Posterior-Wahrscheinlichkeit enthalten ist. Variablen, die deutlich häufiger als andere in einem Modell mit maximaler Posterior-Wahrscheinlichkeit enthalten sind, könnten als Regressoren geeignet sein. Auf diese Weise läßt sich zumindest eine grobe Vorauswahl der Regressoren treffen. Dieses Verfahren kann mit anderen Methoden zur Variablenselektion kombiniert werden.

8.1.5 Der ADF-Algorithmus

Der ADF-Algorithmus zur Approximation der Posterior („Assumed-Density Filtering“) wurde in verschiedenen Forschungsdisziplinen unabhängig voneinander entwickelt und angewendet. Zu den unterschiedlichen Forschungsbereichen zählen neben der Statistik (Bernardo, Giron (1988), Lauritzen (1992), Stephens (1997))

auch die Künstliche Intelligenz (Boyer, Koller (1998)).

Der Grundgedanke des Verfahrens ist, die Likelihood $P(\mathcal{D}_1, ..., \mathcal{D}_{|D|}|\theta)$ für die voneinander unabhängigen Daten $\mathcal{D}_\rho, \rho = 1, ..., |D|$, in Faktoren aufzuspalten, so daß

$$P(\mathcal{D}|\theta) = P(\mathcal{D}_1, ..., \mathcal{D}_{|D|}|\theta) = \prod_{\rho=1}^{|D|} P_\rho(\mathcal{D}_\rho|\theta)$$

gilt und dann nacheinander alle Teile $P_\rho(\mathcal{D}_\rho|\theta)$ zusammen mit der vorher approximierten Posterior (bzw. im zweiten Schritt zusammen mit der im ersten Schritt genäherten Prior) durch eine einfachere Verteilung anzunähern. Dabei wird die im ρ-ten Schritt approximierte Posterior $\hat{P}(\theta|\phi^\rho)$ im $(\rho+1)$-ten Schritt zusammen mit $P_{\rho+1}(\mathcal{D}_{\rho+1}|\theta)$ dazu verwendet die Posterior $P(\theta|\mathcal{D}_{\rho+1}, ..., \mathcal{D}_1)$ durch $\hat{P}(\theta|\phi^{\rho+1})$ anzunähern.
In diesem Zusammenhang erweist sich die Definition

$$t_\rho(\mathcal{D}_\rho|\theta) = \begin{cases} P_\rho(\mathcal{D}_\rho|\theta), & \text{für } \rho = 1, ..., |D| \\ P(\theta), & \text{für } \rho = 0 \end{cases}$$

als hilfreich. Hier ist $\mathcal{D}_0 = \{\}$. Zur Approximation der Posterior $P(\mathcal{D}|\theta)$ werden $|D| + 1$ Schritte ausgeführt. Im ersten Schritt ist $\rho = 0$ und die Prior

$$P(\theta|\mathcal{D}_0, \phi^0) = P(\theta|\mathcal{D}_0) = t_0(\mathcal{D}_0|\theta) = P(\theta)$$

wird durch eine geschätzte Verteilung $\hat{P}^1(\theta|\phi^1)$ approximiert. ϕ^1 bezeichnet die zugehörigen Momente. Man bestimmt ϕ^1, indem man die Kullback-Leibler Divergenz zwischen $P(\theta)$ und der Approximation $\hat{P}^1(\theta|\phi)$ minimiert:

$$\begin{aligned} \phi^1 &= \arg\min_\phi \mathbf{KL}\left(P(\theta) \parallel \hat{P}^1(\theta|\phi)\right) = \arg\min_\phi \left(\int P(\tilde{\theta}) \log \frac{P(\tilde{\theta})}{\hat{P}^1(\tilde{\theta}|\phi)} d\tilde{\theta}\right) \\ &= \arg\min_\phi \left(\int P(\tilde{\theta}) \log \frac{P(\tilde{\theta})}{\hat{P}^1(\tilde{\theta}|\phi)} d\tilde{\theta} + \int (\hat{P}^1(\tilde{\theta}|\phi) - P(\tilde{\theta})) d\tilde{\theta}\right). \end{aligned}$$

Im ρ-ten Schritt $(\rho > 1)$ berechnet man rekursiv die Posterior

$$P^{\rho}(\theta|\mathcal{D}_{\rho},...,\mathcal{D}_1) = \frac{t_{\rho}(\mathcal{D}_{\rho}|\theta)P^{\rho-1}(\theta|\mathcal{D}_{\rho-1},...,\mathcal{D}_1)}{\int t_{\rho}(\mathcal{D}_{\rho}|\tilde{\theta})P^{\rho-1}(\tilde{\theta}|\mathcal{D}_{\rho-1},...,\mathcal{D}_1)d\tilde{\theta}}$$

$$\approx \frac{t_{\rho}(\mathcal{D}_{\rho}|\theta)\hat{P}^{\rho-1}(\theta|\phi^{\rho-1})}{\int t_{\rho}(\mathcal{D}_{\rho}|\tilde{\theta})\hat{P}^{\rho-1}(\tilde{\theta}|\phi^{\rho-1})d\tilde{\theta}} = \tilde{P}^{\rho}(\theta|\mathcal{D}_{\rho},\phi^{\rho-1})$$

Hier bezeichnet $\hat{P}^{\rho-1}(\theta|\phi^{\rho-1})$ die Approximation der Posterior auf Basis der Daten $\mathcal{D}_{\rho-1},...,\mathcal{D}_1$ im $(\rho-1)$-ten Schritt. Mittels

$$\phi^{\rho} = \arg\min_{\phi} \mathbf{KL}\left(\tilde{P}^{\rho}(\theta|\mathcal{D}_{\rho},\phi^{\rho-1}) \parallel \hat{P}^{\rho}(\theta|\phi)\right)$$

bestimmt man den neuen Parameter ϕ^{ρ} und damit die Gestalt der ρ-ten Approximation der Posterior auf der Datengrundlage $\mathcal{D}_{\rho},...,\mathcal{D}_1$.

Problematisch ist, daß dieses Verfahren zu unterschiedlichen Ergebnissen führt, wenn die Reihenfolge der Daten $\mathcal{D}_{|D|},...,\mathcal{D}_1$ vertauscht wird. Insbesondere, wenn die zu Anfang des Verfahrens verwendeten Daten untypisch sind, kann die ADF-Approximation der Posterior eine schlechte Näherung sein. In einigen Fällen läßt sich der numerische Aufwand durch die Benutzung analytischer Beziehungen verkleinern.

Beispiel 8.3:

Mit der Wahl

$$\hat{P}^{\rho}(\theta|\phi=\{m,v\}) = \frac{1}{(2\pi v)^{\mathcal{K}/2}}\exp\left(-\frac{1}{2v}(\theta-m)'(\theta-m)\right)$$

für die approximierte Posterior im ρ-ten Schritt und $\mathcal{K} = dim(\theta)$ ergibt sich mittels

$$\nabla_m \mathbf{KL}\left(\tilde{P}^{\rho}(\theta|\mathcal{D}_{\rho},\phi^{\rho-1}) \parallel \hat{P}^{\rho}(\theta|\phi)\right) \equiv 0$$

die Beziehung

$$E_{\tilde{P}^\rho}(\theta) = E_{\hat{P}^\rho}(\theta).$$

Durch Nullsetzen des Gradienten ∇_v der Kullback-Leibler Divergenz erhält man mit Hilfe dieser Beziehung die Gleichung

$$E_{\tilde{P}^\rho}(\theta'\theta) = E_{\hat{P}^\rho}(\theta'\theta).$$

Mit

$$\mathcal{Z}^\rho(\mathcal{D}_\rho, m, v) = \int \frac{t_\rho(\mathcal{D}_\rho|\tilde{\theta})}{(2\pi v)^{\mathcal{K}/2}} \exp\left(-\frac{1}{2v}(\tilde{\theta} - m)'(\tilde{\theta} - m)\right) d\tilde{\theta}$$

erhält man $E_{\tilde{P}^\rho}(\theta) = m^{\rho-1} + v^{\rho-1}\nabla_m \log Z^\rho(\mathcal{D}_\rho, m^{\rho-1}, v^{\rho-1})$ und

$$\begin{aligned}
&E_{\tilde{P}^\rho}(\theta)' E_{\tilde{P}^\rho}(\theta) - E_{\tilde{P}^\rho}(\theta'\theta) = \mathcal{K} v^{\rho-1} \\
&-(v^\rho)^2 \left[(\nabla_m \log Z^\rho(\mathcal{D}_\rho, m^{\rho-1}, v^{\rho-1}))'(\nabla_m \log Z^\rho(\mathcal{D}_\rho, m^{\rho-1}, v^{\rho-1}))\right. \\
&\left.-\nabla_v \log Z^\rho(\mathcal{D}_\rho, m^{\rho-1}, v^{\rho-1})\right].
\end{aligned}$$

Falls für die Faktoren der Likelihood

$$t_\rho(\mathcal{D}_\rho|\theta) = \frac{1}{(2\pi)^{\mathcal{K}/2}} \exp\left(-\frac{(\mathcal{D}_\rho - \theta)'(\mathcal{D}_\rho - \theta)}{2}\right)$$

gilt, ergibt sich

$$Z^\rho(\mathcal{D}_\rho, m^{\rho-1}, v^{\rho-1}) = \frac{\exp\left(-\frac{(\mathcal{D}_\rho - m^{\rho-1})'(\mathcal{D}_\rho - m^{\rho-1})}{2(v^{\rho-1}+1)}\right)}{(2\pi(v^{\rho-1}+1))^{\mathcal{K}}}.$$

Man erhält somit durch Einsetzen von $Z^\rho(\mathcal{D}_\rho, m^{\rho-1}, v^{\rho-1})$ in die oben angegebenen Formeln

$$E_{\tilde{P}^\rho}(\theta) = m^{\rho-1} + v^{\rho-1}\frac{\mathcal{D}_\rho - m^{\rho-1}}{v^{\rho-1}+1}$$

und (da die getroffene Wahl für $\hat{P}^{\rho}(\theta|\phi = \{m, v\})$ impliziert, daß alle Komponenten von θ die gleiche Varianz aufweisen)

$$E_{\tilde{P}^{\rho}}\left(\frac{(\theta - E_{\tilde{P}^{\rho}}(\theta))'(\theta - E_{\tilde{P}^{\rho}}(\theta))}{\mathcal{K}}\right) = \frac{E_{\tilde{P}^{\rho}}(\theta'\theta) - E_{\tilde{P}^{\rho}}(\theta)' E_{\tilde{P}^{\rho}}(\theta)}{\mathcal{K}}$$
$$= v^{\rho-1} - \frac{(v^{\rho-1})^2}{v^{\rho-1} + 1}.$$

Hieraus ergeben sich die neuen Parameterwerte

$$m^{\rho} = m^{\rho-1} + v^{\rho-1}\frac{\mathcal{D}_{\rho} - m^{\rho-1}}{v^{\rho-1} + 1}$$

und

$$v^{\rho} = v^{\rho-1} - \frac{(v^{\rho-1})^2}{v^{\rho-1} + 1}.$$

Auf diese Weise lassen sich in einigen Fällen die Beziehungen zwischen den alten und den neuen Parametern analytisch angeben. Hierdurch läßt sich der Rechenaufwand verringern.

8.1.6 Der EP-Algorithmus

Der EP-Algorithmus („Expectaion Propagation") zur Approximation der Posterior (Minka (2001a),(2001b)) ist eine Variante des ADF-Algorithmus, bei der in jedem Schritt ein einzelner Faktor $t_{\rho^*}(\mathcal{D}_{\rho^*}|\theta)$ durch seine Näherung

$$\hat{t}^{\rho^*}_{n_{\rho^*}}(\phi^{\rho^*}_{n_{\rho^*}+1}|\theta)$$

auf Basis aller bisherigen Näherungen $\hat{t}^{\rho}_{n_{\rho}}(\phi^{\rho}_{n_{\rho}}|\theta), \rho \in \{0, 1, ..., |D|\} \setminus \rho^*$, von neuem approximiert wird. Der Index n_{ρ} bezeichnet, wie häufig der Parameter ϕ^{ρ} im Verlauf des Verfahrens neu berechnet wurde. Verwendet wird immer das $\phi^{\rho}_{n_{\rho}}$ mit dem höchsten n_{ρ}. In Laufe des Verfahrens werden die einzelnen Terme mehrmals approximiert bis das Verfahren konvergiert. Es ist hierbei möglich, die Reihenfolge in der die Parameter $\phi^{|D|}, ..., \phi^0$ (von neuem) geschätzt werden beliebig zu

verändern. Deshalb spielt die Reihenfolge der Parameter beziehungsweise der zu den Parametern gehörenden Daten keine Rolle. Der Algorithmus konvergiert in den meisten Fällen.

Jede Näherung $\hat{t}^{\rho}_{n_\rho}(\phi^{\rho}_{n_\rho}|\theta)$ ist eindeutig durch den Parameter $\phi^{\rho}_{n_\rho}$ bestimmt. Es gilt $\phi^{\rho}_{n_\rho} = \{m(\rho, n_\rho), v(\rho, n_\rho), s(\rho, n_\rho)\}$. Eine häufige Wahl ist:

$$\hat{t}^{\rho}_{n_\rho}(\phi^{\rho}_{n_\rho}|\theta) = s(\rho, n_\rho) \exp\left(-\frac{(\theta - m(\rho, n_\rho))'(\theta - m(\rho, n_\rho))}{2v(\rho, n_\rho)}\right).$$

Im folgenden werden nur Normalverteilungs-Approximationen betrachtet. Der Parameter $s(\rho, n_\rho)$ ist zur Schätzung der Posterior nicht erforderlich und wird aus diesem Grund im weiteren Verlauf der Arbeit nicht berücksichtigt.
Wegen der Beziehung $P^{\rho}(\theta|\mathcal{D}_\rho, ..., \mathcal{D}_1) \propto t_\rho(\mathcal{D}_\rho|\theta) P^{\rho-1}(\theta|\mathcal{D}_{\rho-1}, ..., \mathcal{D}_1)$ kann die gesamte Posterior auch als

$$P^{|D|}(\theta|\mathcal{D}_{|D|}, ..., \mathcal{D}_1) = \frac{\prod_{\rho=0}^{|D|} t_\rho(\mathcal{D}_\rho|\theta)}{\int \prod_{\rho'=0}^{|D|} t_{\rho'}(\mathcal{D}_{\rho'}|\tilde{\theta}) d\tilde{\theta}}$$

dargestellt werden. Diese Darstellung macht man sich im Rahmen der EP-Variante des ADF-Algorithmus zunutze, um die (gesamte) Posterior zu approximieren:

$$P(\theta|\mathcal{D}_{|D|}, ..., \mathcal{D}_1) \approx \tilde{P}(\theta|\phi^{|D|}_{n_{|D|}}, ..., \phi^{0}_{n_0}) = \frac{\prod_{\rho=0}^{|D|} \hat{t}^{\rho}_{n_\rho}(\phi^{\rho}_{n_\rho}|\theta)}{\int \prod_{\rho'=0}^{|D|} \hat{t}^{\rho'}_{n_{\rho'}}(\phi^{\rho'}_{n_{\rho'}}|\tilde{\theta}) d\tilde{\theta}}.$$

Hieraus erhält man die Näherung

$$P^{-\rho}(\theta|\mathcal{D}_{|D|}, ..., \mathcal{D}_{\rho+1}, \mathcal{D}_{\rho-1}, ..., \mathcal{D}_1) \approx \wp^{-\rho}(\theta|\phi^{-\rho}) = \frac{\tilde{P}(\theta|\phi^{|D|}_{n_{|D|}}, ..., \phi^{0}_{n_0})}{\hat{t}^{\rho}_{n_\rho}(\phi^{\rho}_{n_\rho}|\theta)},$$

wobei $\phi^{-\rho} = \{\phi^{|D|}_{n_{|D|}}, ..., \phi^{\rho+1}_{n_{\rho+1}}, \phi^{\rho-1}_{n_{\rho-1}}, ..., \phi^{0}_{n_0}\}$ gilt. Im Unterschied zum klassischen ADF-Algorithmus müssen bei der EP-Variante zu Beginn des Verfahrens die

Setze $n_\rho = 1, \hat{t}^\rho_{n_\rho}(\phi^\rho_0|\theta) = 0, \hat{t}^\rho_{n_\rho}(\phi^\rho_1|\theta) = 1, \rho = 0, 1, ..., |D|$.
Solange $\Big(\exists \rho \in \{0, 1, ..., |D|\} : \phi^\rho_{n_\rho} \neq \phi^\rho_{n_\rho - 1}\Big)$:

- Wähle $\rho \in \{0, 1, ..., |D|\}$.
- Bestimme $\wp^{-\rho}(\phi^{-\rho}|\theta)$ neu.
- Berechne damit $\phi^\rho_{n_\rho+1}$.
- Ermittle $\hat{t}^\rho_{n_\rho}(\phi^\rho_{n_\rho+1}|\theta)$.
- Setze $n_\rho \leftarrow n_\rho + 1$.
- Bestimme damit $\tilde{P}(\theta|\phi^{|D|}_{n_{|D|}}, ..., \phi^1_{n_1})$.

Abbildung 8.6: EP-Algorithmus

$\hat{t}_\rho(\phi^\rho_{n_\rho}|\theta), \rho = 0, 1, ..., |D|$, initialisiert werden. ($\phi^0_{n_0}$ ist ein Prior-Parameter.) In jedem einzelnen Schritt wird zunächst ein beliebiges $\rho \in \{0, 1, ..., |D|\}$ gewählt und $\wp^{-\rho}(\theta|\phi^{-\rho})$ bestimmt. ($\phi^{-\rho}$ umfaßt immer die zuletzt berechneten Parameter $\phi^{|D|}_{n_{|D|}}, ..., \phi^{\rho+1}_{n_{\rho+1}}, \phi^{\rho-1}_{n_{\rho-1}}, ..., \phi^0_{n_0}$.) Den neuen Parameter $\phi^\rho_{n_\rho+1}$ erhält man dann mittels

$$\phi^\rho_{n_\rho+1} = \arg\min_{\phi^\rho} \mathbf{KL}\left(\wp^{-\rho}(\theta|\phi^{-\rho}) t_\rho(\mathcal{D}|\theta) \;\;\|\;\; \wp^{-\rho}(\theta|\phi^{-\rho}) \hat{t}^\rho_{n_\rho}(\phi^\rho|\theta)\right).$$

Hier ist $\wp^{-\rho}(\theta|\phi^{-\rho}) t_\rho(\mathcal{D}|\theta)$ das EP-Äquivalent von $\tilde{P}^\rho(\theta|\mathcal{D}_\rho, \phi^{\rho-1})$ und das Produkt $\wp^{-\rho}(\theta|\phi^{-\rho}) \hat{t}^\rho_{n_\rho}(\phi^\rho|\theta)$ entspricht $\hat{P}^\rho(\theta|\phi)$. Sofern man $\wp^{-\rho}(\theta|\phi^{-\rho})$ bereits bestimmt hat, kann $\phi^\rho_{n_\rho+1}$ genau wie ϕ^ρ im Rahmen des ADF-Algorithmus berechnet werden.

Auch dieses Verfahren kann durch die Benutzung analytischer Beziehungen wesentlich vereinfacht werden. Falls alle Faktoren $\hat{t}_\rho(\phi^\rho_{n_\rho}|\theta)$ Wahrscheinlichkeitsdichten der Normalverteilung sind, ist es möglich und empfehlenswert anstelle ihres (normierten) Produkts eine einfachere Darstellung der Gauß-Funktion zu verwenden.

Beispiel 8.4:

Für

$$\hat{t}^\rho_{n_\rho}(\phi^\rho_{n_\rho}|\theta) \propto \exp\left(-\frac{(\theta - m(\rho, n_\rho))'(\theta - m(\rho, n_\rho))}{2v(\rho, n_\rho)}\right), \rho = 0, ..., |D|,$$

ist beispielsweise die Verwendung von

$$\hat{P}^{|D|}(\theta|m,v) = \frac{1}{(2\pi v)^{\mathcal{K}/2}} \exp\left(-\frac{(\theta-m)'(\theta-m)}{2v}\right)$$

möglich.
Damit ergibt sich nach Wahl von $\rho \in \{0,1,...,|D|\}$

$$\wp^{-\rho}(\theta|\phi^{-\rho}) \propto \frac{\hat{P}^{|D|}(\theta|m,v)}{\hat{t}^{\rho}_{n_\rho}(\phi^{\rho}_{n_\rho}|\theta)} \propto \left(\frac{1}{v^{-\rho}}\right)^{\mathcal{K}/2} \exp\left(-\frac{(\theta-m^{-\rho})'(\theta-m^{-\rho})}{2v^{-\rho}}\right).$$

Deshalb erhält man die Beziehung

$$\hat{t}^{\rho}_{n_\rho}(\phi^{\rho}_{n_\rho}|\theta) \propto \exp\left(-\frac{(\theta-m)'(\theta-m)}{2v}\right) \exp\left(+\frac{(\theta-m^{-\rho})'(\theta-m^{-\rho})}{2v^{-\rho}}\right).$$

Andererseits ist (mit $\phi^{\rho}_{n_\rho} = \{m(\rho,n_\rho), v(\rho,n_\rho)\}$)

$$\hat{t}^{\rho}_{n_\rho}(\phi^{\rho}_{n_\rho}|\theta) \propto \exp\left(-\frac{(\theta-m(\rho,n_\rho))'(\theta-m(\rho,n_\rho))}{2v(\rho,n_\rho)}\right),$$

weshalb sich die Beziehungen $(v^{-\rho})^{-1} = (v)^{-1} - (v(\rho,n_\rho))^{-1}$ (B8.1) und

$$m^{-\rho} = \left(\frac{v^{-\rho}}{v}\right) m - \left(\frac{v^{-\rho}}{v(\rho,n_\rho)}\right) m(\rho,n_\rho) \quad \text{(B8.2)}$$

ergeben. Hiermit ist $\wp^{-\rho}(\theta|\phi^{-\rho})$ eindeutig bestimmt - sofern $\hat{P}^{|D|}(\theta|m,v)$ durch m und v und $\hat{t}^{\rho}_{n_\rho}(\phi^{\rho}_{n_\rho}|\theta)$ durch $\phi^{\rho}_{n_\rho} = (m(\rho,n_\rho), v(\rho,n_\rho))$ gegeben sind. Somit kann die neue Parametrisierung für die gesamte Wahrscheinlichkeitsdichte der Posterior genau wie im Rahmen des ADF-Algorithmus (siehe Beispiel 8.3) bestimmt werden:

$$m \leftarrow m^{-\rho} + v^{-\rho}\frac{\mathcal{D}_\rho - m^{-\rho}}{v^{-\rho}+1}$$

$$v \leftarrow v^{-\rho} - \frac{(v^{-\rho})^2}{v^{-\rho}+1}.$$

Die aktuelle Parametrisierung $\phi^{\rho}_{n_\rho+1}$ von $\hat{t}^{\rho}_{n_\rho+1}(\phi^{\rho}_{n_\rho+1}|\theta)$ erhält man mit den eben berechneten Werten für m und v via zu B8.1 und B8.2 analoger Beziehungen durch

$$\begin{aligned}(v(\rho, n_\rho+1))^{-1} \leftarrow &\quad (v)^{-1} - (v^{-\rho})^{-1} \\ m(\rho, n_\rho+1) \leftarrow &\quad m^{-\rho} + \frac{v(\rho, n_\rho+1)}{v}(m - m^{-\rho}) \\ &\quad = m^{-\rho} + (v(\rho, n_\rho+1) + v^{-\rho})\frac{\mathcal{D}_\rho - m^{-\rho}}{v^{-\rho}+1}\end{aligned}$$

Danach setzt man $n_\rho \leftarrow n_\rho + 1$ und wählt ein neues $\rho \in \{0, 1, ..., |D|\}$. Dies wiederholt man solange, bis das Verfahren konvergiert.

Während der Metropolis-Hastings Algorithmus längst zum Standard-Instrumentarium der Bayes'schen Statistik gehört, ist insbesondere der EP-Algorithmus eine bisher nur selten verwendete heuristische Methode, deren Eigenschaften nur unzureichend untersucht sind.

8.2 Kollaborative Verfahren

Das erste Bayes'sche Modell für im Zusammenhang mit Recommendersystemen generierte Bewertungsdaten geht auf Chien, George (1999) zurück. Dieses Modell benutzt die Likelihood der Multinomialverteilung. Daher werden die Daten in diesem Modell behandelt, als hätten sie nur nominales Skalenniveau. Infolgedessen geht relevante Information verloren und das Modell schneidet bezüglich der Schätzgenauigkeit nur geringfügig besser als das Nutzer-basierte Ähnlichkeitsverfahren nach Shardanand, Maes (1995) ab.

Zudem werden alle Nutzer einer bestimmten Nutzerklasse zugeordnet. Für alle Nutzer, die derselben Nutzerklasse angehören, wird derselbe Schätzer für die Wahrscheinlichkeit, daß ein bestimmtes Item mit $c \in \{1, ..., C\}$ bewertet wird, verwendet. Daher bleibt die Heterogenität der Nutzer innerhalb einer Nutzerklasse unberücksichtigt. Aus diesen Gründen wird das Modell von Chien, George (1999) hier nicht näher betrachtet.

Als weitaus effektiver haben sich zwei neuere hierarchische Ansätze erwiesen, die für jeden einzelnen Nutzer eine latente Variable mit Hilfe verschiedener Gauß'scher Prozesse (GP) modellieren und zur Schätzung ihrer Posterior den EP-Algorithmus verwenden.

8.2.1 Allgemeine Darstellung GP-basierter Verfahren

Die kollaborativen Verfahren, die auf dem Gauß'schen Prozeß beruhen (Yu et. al. (2006)), greifen wie die klassischen Verfahren zur Modellierung ordinalskalierter Daten auf das Konzept der latenten Variable zurück. Für jeden Nutzer $i \in \{1, ..., I\}$ wird ein ganzer Vektor $\mathbf{f}_i = (\mathbf{f}_{i1}, ..., \mathbf{f}_{iJ})' \in \mathbb{R}^J$ aus latenten Größen verwendet. Jede einzelne Komponente $\mathbf{f}_{ij}$ dieses Vektors beschreibt die den Bewertungen zugrundeliegende Beziehung des Nutzers i zum Item j. Es gilt $Y_{ij} = c \Leftrightarrow \mathbf{f}_{ij} \in (\gamma_{c-1}, \gamma_c]$. $\mathbf{f}_i(\alpha_i)$ besteht aus den Komponenten von $\mathbf{f}_i$, die der i-te Nutzer bewertet hat.

Da die latenten Variablen $\mathbf{f}_i$ keine Observablen sind, muß eine Prior für $\mathbf{f}_i$ gefunden werden, die es ermöglicht, $\mathbf{f}_i$ im Rahmen eines Bayes'schen Ansatzes auszuintegrieren. Zu diesem Zweck ist es sinnvoll, die $\mathbf{f}_i, i = 1, ..., I$, als Realisationen der Zufallsvariablen eines Gauß'schen Prozesses aufzufassen. Ein Gauß'scher Prozeß (GP) ist durch seinen Erwartungswert $\mathbf{h}$ und seine Kovarianz $\mathbf{K}$ definiert (Williams (1996)). Im Rahmen eines nicht-hierarchischen Ansatzes (ohne Hyperprior) ergibt sich daher für die Prior jedes einzelnen $\mathbf{f}_i$ die Darstellung

$$P(\mathbf{f}_i|\mathbf{h}, \mathbf{K}) = \frac{1}{(2\pi)^{J/2}|\mathbf{K}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{f}_i - \mathbf{h})'\mathbf{K}^{-1}(\mathbf{f}_i - \mathbf{h})\right).$$

Als Hyperprior für $\mathbf{h}$ und $\mathbf{K}$ verwendet man beim hierarchischen Ansatz die übliche Kombination einer multivariaten Normalverteilungsdichte mit der Wahrscheinlichkeitsdichte einer inversen Wishart-Verteilung:

$$P(\mathbf{h}, \mathbf{K}) \propto \frac{1}{|\mathbf{K}|^{\tau_h/2}} \exp\left(-\frac{1}{2}(\mathbf{h} - \mathbf{h}_0)'\left(\frac{\mathbf{K}}{\pi_h}\right)^{-1}(\mathbf{h} - \mathbf{h}_0) - \frac{tr(\mathbf{K}^{-1}(\tau_h\mathbf{K}_0))}{2}\right).$$

Da die Nutzer im allgemeinen nur einen kleinen Teil der Items $j \in \{1, ..., J\}$ bewertet haben, stehen für jeden Nutzer nicht die vollständigen Daten zur Verfügung. Sei J_i die Menge von Items, die der Nutzer i bewertet hat und sei $\alpha_i = |J_i|$. Der Vektor $y_i = (y_{i1}, ..., y_{i\alpha_i})' \in \mathbb{R}^{\alpha_i}$ enthält alle Bewertungen, die der Nutzer i abgegeben hat. Mittels der latenten Variablen ergibt sich für die Likelihood jedes Nutzers i die allgemeine Darstellung $P(Y_i|\mathbf{f}_i(\alpha_i), \Omega_i)$.

Die GP-basierten Methoden sind hierarchische Verfahren. Zunächst approximiert man für alle Nutzer $i = 1, ..., I$, getrennt für vorgegebene Prior-Parameter

$\mathbf{h}$ und $\mathbf{K}$ und konstante Ω_i

$$P(\mathbf{f}_i|y_i,\mathbf{h},\mathbf{K},\Omega_i) \propto P(y_i|\mathbf{f}_i(\alpha_i),\Omega_i)P(\mathbf{f}_i|\mathbf{h},\mathbf{K})$$

mit Hilfe des EP-Algorithmus durch

$$\hat{P}(\mathbf{f}_i) = \frac{1}{(2\pi)^{J/2}|\hat{\mathcal{K}}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{f}_i-\hat{\mathbf{f}}_i)'(\hat{\mathcal{K}}_i)^{-1}(\mathbf{f}_i-\hat{\mathbf{f}}_i)\right).$$

Es gilt $\hat{\mathbf{f}}_i \in \mathbb{R}^J$ und $\hat{\mathcal{K}}_i \in \mathbb{R}^{J,J}$. Dabei verwendet man den Produkt-Ansatz

$$\tilde{P}(\mathbf{f}_i|\phi^{i,J}_{n_{iJ}},...,\phi^{i,0}_{n_{i0}}=\{\mathbf{h},\mathbf{K}\}) \propto \prod_{j'\in J_i} \hat{t}_{i,j'}(\phi^{i,j'}_{n_{ij'}}|\mathbf{f}_{ij'})P(\mathbf{f}_i|\mathbf{h},\mathbf{K})$$

mit $\hat{P}(\mathbf{f}_i) = \tilde{P}(\mathbf{f}_i|\phi^{i,J}_{n_{iJ}},...,\phi^{i,0}_{n_{i0}})$ und

$$\hat{t}_{i,j}(\phi^{i,j}_{n_{ij}}|\mathbf{f}_{ij}) \propto \exp\left(-\frac{(\mathbf{f}_{ij}-m_i(j,n_{ij}))^2}{2v_i(j,n_{ij})}\right).$$

Den neuen Schätzer $\phi^{i,j}_{n_{ij}+1}$ auf Basis von $\phi^{i,0}_{n_{i0}},...,\phi^{i,J}_{n_{iJ}}$ erhält man durch

$$\phi^{i,j}_{n_{ij}+1} = \arg\min_{\phi^{i,j}} \mathbf{KL}\left(\frac{\hat{P}(\mathbf{f}_i)}{\hat{t}_{i,j}(\phi^{i,j}_{n_{ij}}|\mathbf{f}_{ij})}t_{i,j}(Y_{ij}|\mathbf{f}_{ij}) \;\;\|\;\; \frac{\hat{P}(\mathbf{f}_i)}{\hat{t}_{i,j}(\phi^{i,j}_{n_{ij}}|\mathbf{f}_{ij})}\hat{t}_{i,j}(\phi^{i,j}|\mathbf{f}_{ij})\right).$$

Sobald die Verteilung des betreffenden $\mathbf{f}_i$ (näherungsweise) bestimmt ist, berechnet man damit den neuen Wert für Ω_i, $\hat{\Omega}_i$. Nachdem auf diese Weise für alle $i \in \{1,...,I\}$ (getrennt) neue Werte für $\mathbf{f}_i$, $\hat{\mathcal{K}}_i$ und Ω_i auf Basis von $\mathbf{h}$ und $\mathbf{K}$ bestimmt worden sind, werden auf dieser Grundlage neue Werte für $\mathbf{h}$ und $\mathbf{K}$ berechnet.

Die Bestimmung von $\Omega_i, i = 1,...,I$, basiert auf Maximierung einer unteren Schranke von $\log P(y|\mathbf{h},\mathbf{K},\Omega) = \log P(y_1,...,y_I|\mathbf{h},\mathbf{K},\Omega_1,...,\Omega_I)$. Es gilt:

$$P(y|\mathbf{h},\mathbf{K},\Omega) = \prod_{i'=1}^{I} P(y_{i'}|\mathbf{h},\mathbf{K},\Omega_{i'}) = \prod_{i'=1}^{I} \int P(y_{i'}|\tilde{\mathbf{f}}_{i'}(\alpha_{i'}),\Omega_{i'})P(\tilde{\mathbf{f}}_{i'}|\mathbf{h},\mathbf{K})d\tilde{\mathbf{f}}_{i'}.$$

Daraus erhält man mittels der Jensen'schen Ungleichung eine untere Schranke für $\log P(y|\mathbf{h},\mathbf{K},\Omega)$, die anstelle von $\log P(y|\mathbf{h},\mathbf{K},\Omega)$ maximiert werden kann:

$$\begin{aligned}
\log P(y|\mathbf{h},\mathbf{K},\Omega) &= \sum_{i'=1}^{I} \log\left(\int \hat{P}(\tilde{\mathbf{f}}_{i'}) \frac{P(y_{i'}|\tilde{\mathbf{f}}_{i'}(\alpha_{i'}),\Omega_{i'})P(\tilde{\mathbf{f}}_{i'}|\mathbf{h},\mathbf{K})}{\hat{P}(\tilde{\mathbf{f}}_{i'})} d\tilde{\mathbf{f}}_{i'}\right) \\
&\geq \sum_{i'=1}^{I} \int \hat{P}(\tilde{\mathbf{f}}_{i'}) \log\left(\frac{P(y_{i'}|\tilde{\mathbf{f}}_{i'}(\alpha_{i'}),\Omega_{i'})P(\tilde{\mathbf{f}}_{i'}|\mathbf{h},\mathbf{K})}{\hat{P}(\tilde{\mathbf{f}}_{i'})}\right) d\tilde{\mathbf{f}}_{i'} \\
&= \sum_{i'=1}^{I} \Bigg[\int \hat{P}(\tilde{\mathbf{f}}_{i'}) \log P(y_{i'}|\tilde{\mathbf{f}}_{i'}(\alpha_{i'}),\Omega_{i'}) d\tilde{\mathbf{f}}_{i'} + \\
&\quad \int \hat{P}(\tilde{\mathbf{f}}_{i'}) \log P(\tilde{\mathbf{f}}_{i'}|\mathbf{h},\mathbf{K}) d\tilde{\mathbf{f}}_{i'} - \int \hat{P}(\tilde{\mathbf{f}}_{i'}) \log \hat{P}(\tilde{\mathbf{f}}_{i'}) d\tilde{\mathbf{f}}_{i'}\Bigg] \\
&= \mathfrak{S}_u \left[\log P(y|\mathbf{h},\mathbf{K},\Omega)\right].
\end{aligned}$$

Durch Maximieren der unteren Schranke $\mathfrak{S}_u \left[\log P(y|\mathbf{h},\mathbf{K},\Omega)\right]$ ergibt sich mit

$$\hat{P}(\mathbf{f}_i) = \frac{1}{(2\pi)^{J/2}|\hat{\mathcal{K}}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{f}_i - \hat{\mathbf{f}}_i)'(\hat{\mathcal{K}}_i)^{-1}(\mathbf{f}_i - \hat{\mathbf{f}}_i)\right)$$

den neuen Wert für Ω_i:

$$\hat{\Omega}_i = \arg\min_{\Omega_i} \int \hat{P}(\tilde{\mathbf{f}}_i) \log P(y_i|\tilde{\mathbf{f}}_i(\alpha_i),\Omega_i) d\tilde{\mathbf{f}}_i.$$

Nachdem alle $\hat{\mathbf{f}}_i$, $\hat{\mathcal{K}}_i$ und $\hat{\Omega}_i, i = 1, ..., I$, bestimmt sind, müssen auf dieser Grundlage die Prior-Parameter $\mathbf{h}$ und $\mathbf{K}$ neu berechnet werden. Hierdurch wird die individuelle Information aus den Daten y_i, die in $\hat{\mathbf{f}}_i$,$\hat{\mathcal{K}}_i$ und $\hat{\Omega}_i$ enthalten ist, gebündelt und mit der in den Hyperprior-Parametern $\mathbf{h}_0$ und $\mathbf{K}_0$ manifestierten Vorinformation in Zusammenhang gebracht. Auf diese Weise vereinen $\mathbf{h}$ und $\mathbf{K}$ die gesamte individuelle Information der einzelnen Daten mit der Vorinformation. Daher beschreiben $\mathbf{h}$ und $\mathbf{K}$ die allgemeinen Tendenzen aller latenten Variablen.

Diese allgemeinen Tendenzen fließen wiederum via

$$\tilde{P}(\mathbf{f}_i|\phi^{i,J}_{n_{iJ}},...,\phi^{i,1}_{n_{i1}},\phi^0_{n_0}=\{\mathbf{h},\mathbf{K}\})\propto\prod_{j'\in J_i}\hat{t}_{i,j'}(\phi^{i,j'}_{n_{ij'}}|\mathbf{f}_{ij'})P(\mathbf{f}_i|\mathbf{h},\mathbf{K})$$

in die Bestimmung (oder - je nach Wahl von $P(y_i,\mathbf{f}_i(\alpha_i),\Omega_i)$ - Näherung) der einzelnen Verteilungen der $\mathbf{f}_i$ durch den EP-Algorithmus mit ein. Daher enthalten $\hat{\mathbf{f}}_i$ und $\hat{\mathcal{K}}_i$ neben den individuellen Tendenzen der Daten Y_i immer auch allgemeine Tendenzen. Hierdurch wird dieser Ansatz kollaborativ. Indem man den Logarithmus von $P(\mathbf{h},\mathbf{K}|y,\Omega)\propto P(y|\mathbf{h},\mathbf{K},\Omega)P(\mathbf{h},\mathbf{K})'$ mit der unteren Schranke $\mathfrak{S}_u\left[\log P(y|\mathbf{h},\mathbf{K},\Omega)\right]$ von $\log P(y|\mathbf{h},\mathbf{K},\Omega)$ anstelle von $\log P(y|\mathbf{h},\mathbf{K},\Omega)$ für konstante $\Omega=\{\hat{\Omega}_1,...,\hat{\Omega}_I\}$ maximiert, können $\mathbf{h}$ und $\mathbf{K}$ näherungsweise bestimmt werden . Hierbei verwendet man die zuletzt bestimmten Werte für $\hat{\Omega}_i, i=1,...,I$. Auf diese Weise erhält man eine analytische Näherungslösung für $\mathbf{h}$ und $\mathbf{K}$ auf Basis ihrer gemeinsamen Posterior.
Aus der Bedingung erster Ordung

$$\nabla_{\mathbf{h}}\left(\mathfrak{S}_u\left[\log P(y|\mathbf{h},\mathbf{K},\Omega)\right]+\log P(\mathbf{h},\mathbf{K})\right)\equiv 0$$

mit $0\in\mathbb{R}^J$ und der Prior

$$P(\mathbf{h},\mathbf{K})\propto\frac{1}{|\mathbf{K}|^{\tau_h/2}}\exp\left(-\frac{1}{2}(\mathbf{h}-\mathbf{h}_0)'\left(\frac{\mathbf{K}}{\pi_h}\right)^{-1}(\mathbf{h}-\mathbf{h}_0)\right)\exp\left(-\frac{tr(\mathbf{K}^{-1}(\tau_h\mathbf{K}_0))}{2}\right)$$

ergibt sich die Gleichung

$$\begin{aligned}
&=\nabla_{\mathbf{h}}\left(\textstyle\sum_{i'=1}^I\int\hat{P}(\tilde{\mathbf{f}}_{i'})\log P(\tilde{\mathbf{f}}_{i'}|\mathbf{h},\mathbf{K})d\tilde{\mathbf{f}}_{i'}\right)+\nabla_{\mathbf{h}}\log P(\mathbf{h},\mathbf{K})\\
&=\textstyle\sum_{i'=1}^I\int\hat{P}(\tilde{\mathbf{f}}_{i'})\left(-\frac{1}{2}\right)\nabla_{\mathbf{h}}(-\tilde{\mathbf{f}}'_{i'}\mathbf{K}^{-1}\mathbf{h}-\mathbf{h}'\mathbf{K}^{-1}\tilde{\mathbf{f}}_{i'}+\mathbf{h}'\mathbf{K}^{-1}\mathbf{h})d\tilde{\mathbf{f}}_{i'}\\
&-\tfrac{\pi_h}{2}\nabla_{\mathbf{h}}(\mathbf{h}'\mathbf{K}^{-1}\mathbf{h}-\mathbf{h}'_0\mathbf{K}^{-1}\mathbf{h}-\mathbf{h}'\mathbf{K}^{-1}\mathbf{h}_0).\\
&=\textstyle\sum_{i'=1}^I\int\left(\hat{P}(\tilde{\mathbf{f}}_{i'})\left(-\frac{1}{2}\right)(-2)\mathbf{K}^{-1}(\tilde{\mathbf{f}}_{i'}-\mathbf{h})\right)d\tilde{\mathbf{f}}_{i'}-\tfrac{\pi_h}{2}2\mathbf{K}^{-1}(\mathbf{h}-\mathbf{h}_0)\\
&=\textstyle\sum_{i'=1}^I\mathbf{K}^{-1}E_{\hat{P}}(\mathbf{f}_{i'})-\sum_{i'=1}^I\mathbf{K}^{-1}\mathbf{h}-\pi_h\mathbf{K}^{-1}(\mathbf{h}-\mathbf{h}_0)\equiv 0
\end{aligned}$$

Daraus erhält man wegen $E_{\hat{P}}(\mathbf{f}_i)=\hat{\mathbf{f}}_i$ den neuen Wert für $\mathbf{h}$:

$$\hat{\mathbf{h}}=\frac{1}{I+\pi_h}\left(\sum_{i'=1}^I\hat{\mathbf{f}}_{i'}+\pi_h\mathbf{h}_0\right).$$

Die Herleitung der Formel für $\hat{\mathbf{K}}$ erfolgt analog. Es gilt

$$\frac{\partial}{\partial \mathbf{K}} \mathfrak{S}_u \left[\log P(y|\mathbf{h}, \mathbf{K}, \Omega)\right] =$$

$$\frac{\partial}{\partial \mathbf{K}} \sum_{i=1}^{I} \int \hat{P}(\tilde{\mathbf{f}}_i) \left(-\frac{1}{2} \log |\mathbf{K}| - \frac{1}{2} (\tilde{\mathbf{f}}_i - \mathbf{h})' \mathbf{K}^{-1} (\tilde{\mathbf{f}}_i - \mathbf{h}) \right) d\tilde{\mathbf{f}}_i.$$

Für symmetrische Matrizen $\mathbf{K}$ gelten die Beziehungen (vgl. Magnus, Neudecker (1988), S. 178 f.)

$$\frac{\partial |\mathbf{K}|}{\partial \mathbf{K}} = |\mathbf{K}| \mathbf{K}^{-1} \quad \text{und} \quad \frac{\partial A \mathbf{K}^{-1}}{\partial \mathbf{K}} = -(\mathbf{K}^{-1} A \mathbf{K}^{-1})'.$$

Da $\mathbf{K}_0$ ebenfalls eine symmetrische Matrix ist, ist $(\mathbf{K}^{-1} \mathbf{K}_0 \mathbf{K}^{-1})$ ebenfalls symmetrisch. Infolgedessen ergibt sich

$$\frac{\partial}{\partial \mathbf{K}} \mathfrak{S}_u \left[\log P(y|\mathbf{h}, \mathbf{K}, \Omega)\right] =$$

$$\sum_{i=1}^{I} \int \hat{P}(\tilde{\mathbf{f}}_i) \left(-\frac{\mathbf{K}^{-1}}{2} + \frac{1}{2} \mathbf{K}^{-1} (\tilde{\mathbf{f}}_i - \mathbf{h})(\tilde{\mathbf{f}}_i - \mathbf{h})' \mathbf{K}^{-1} \right) d\tilde{\mathbf{f}}_i =.$$

$$-\frac{I}{2} \mathbf{K}^{-1} + \frac{1}{2} \mathbf{K}^{-1} \left(\sum_{i=1}^{I} E_{\hat{P}}((\mathbf{f}_i - \mathbf{h})(\mathbf{f}_i - \mathbf{h})') \right) \mathbf{K}^{-1}.$$

Wegen

$$\begin{aligned} E_{\hat{P}}((\mathbf{f}_i - \mathbf{h})(\mathbf{f}_i - \mathbf{h})') &= E_{\hat{P}}((\mathbf{f}_i - \hat{\mathbf{f}}_i + \hat{\mathbf{f}}_i - \mathbf{h})(\mathbf{f}_i - \hat{\mathbf{f}}_i + \hat{\mathbf{f}}_i - \mathbf{h})') \\ &= E_{\hat{P}}((\mathbf{f}_i - \hat{\mathbf{f}}_i)(\mathbf{f}_i - \hat{\mathbf{f}}_i)') + E_{\hat{P}}((\hat{\mathbf{f}}_i - \mathbf{h})(\hat{\mathbf{f}}_i - \mathbf{h})') \\ &= \hat{\mathcal{K}}_i + (\hat{\mathbf{f}}_i - \mathbf{h})(\hat{\mathbf{f}}_i - \mathbf{h})'. \end{aligned}$$

gilt

$$\frac{\partial}{\partial \mathbf{K}} \mathfrak{S}_u \left[\log P(y|\mathbf{h}, \mathbf{K}, \Omega)\right] = -\frac{I}{2} \mathbf{K}^{-1} + \frac{1}{2} \sum_{i=1}^{I} \left(\mathbf{K}^{-1} \left[\hat{\mathcal{K}}_i + (\hat{\mathbf{f}}_i - \mathbf{h})(\hat{\mathbf{f}}_i - \mathbf{h})' \right] \mathbf{K}^{-1} \right).$$

Ganz analog ergibt sich

$$\frac{\partial}{\partial \mathbf{K}} \log P(\mathbf{h}, \mathbf{K}) = -\frac{\tau_h}{2}\mathbf{K}^{-1} + \frac{\pi_h}{2}\mathbf{K}^{-1}(\mathbf{h}-\mathbf{h}_0)(\mathbf{h}-\mathbf{h}_0)'\mathbf{K}^{-1} + \frac{\tau_h}{2}\mathbf{K}^{-1}\mathbf{K}_0\mathbf{K}^{-1}.$$

Aus der Bedingung erster Ordnung

$$\frac{\partial}{\partial \mathbf{K}} \left(\mathfrak{S}_u \left[\log P(y|\mathbf{h}, \mathbf{K}, \Omega) \right] + \log P(\mathbf{h}, \mathbf{K}) \right) \equiv 0$$

folgt somit für den neuen Wert für $\mathbf{K}$

$$\hat{\mathbf{K}} = \frac{1}{I + \tau_h} \left(\pi_h (\mathbf{h}-\mathbf{h}_0)(\mathbf{h}-\mathbf{h}_0)' + \tau_h \mathbf{K}_0 + \sum_{i=1}^{I} \left[(\hat{\mathbf{f}}_i - \mathbf{h})(\hat{\mathbf{f}}_i - \mathbf{h})' + \hat{\mathcal{K}}_i \right] \right).$$

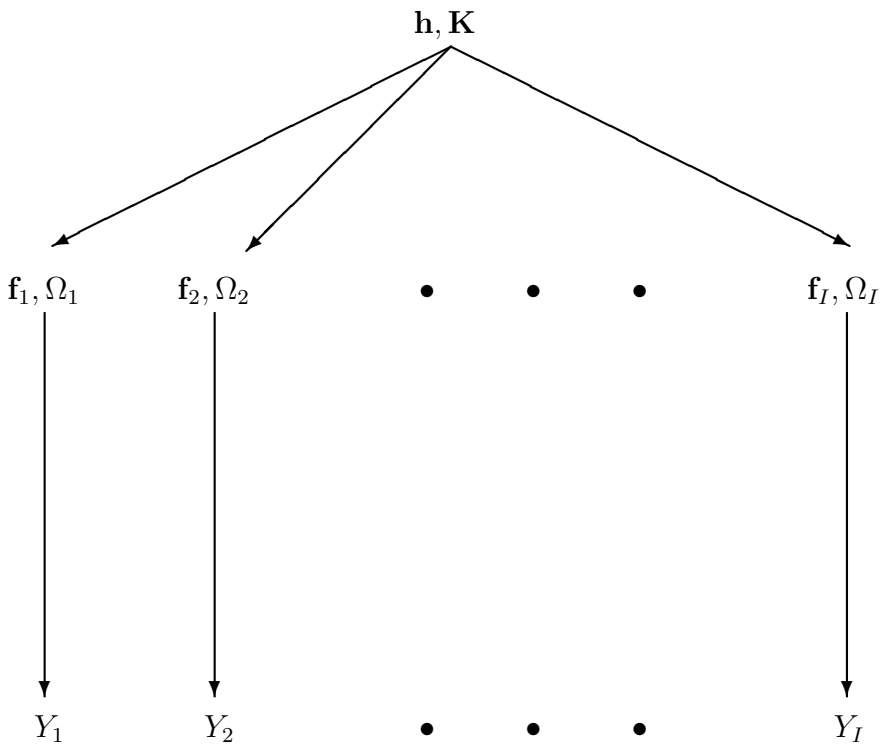

Abbildung 8.7: Struktur eines hierarchischen GP-Ansatzes

In Abbildung 8.7 ist die hierarchische Struktur des beschriebenen hierarchischen GP-Modells graphisch dargestellt. Ergänzend soll hier noch einmal daran erinnert werden, daß zur Bestimmung der Verteilung von $\mathbf{f}_i$ (durch $\hat{\mathbf{f}}_i$ und $\hat{\mathcal{K}}_i$) der letzte bestimmte Wert für Ω_i benötigt wird. Umgekehrt erfordert die Bestimmung des neuen Werts für Ω_i die zuvor bestimmten Werte für $\hat{\mathbf{f}}_i$ und $\hat{\mathcal{K}}_i$, $i = 1, ..., I$. Trotz dieser wechselseitigen Abhängigkeit stehen alle drei Prior-Parameter auf derselben hierarchischen Stufe.

Abschließend soll noch einmal ausdrücklich darauf hingewiesen werden, daß das im Rahmen der Bayes'schen Statistik unübliche Schätzer-Dach nur dazu da ist, den im Verlauf des Algorithmus neu berechneten Wert einer Variable (bzw. ihres Erwartungswerts) zu bezeichnen.

8.2.2 Varianten des GP-basierten Verfahrens

Die unterschiedlichen Varianten des GP-basierten Verfahrens unterscheiden sich voneinander durch unterschiedliche Ansätze für $P(y_i|\mathbf{f}_i, \Omega_i)$. Der einfachste Ansatz (Yu et. al. (2006)) wählt

$$P_{YYTK}(Y_i|\mathbf{f}_i, \Omega_i) = \frac{1}{(2\pi\Omega_i)^{\alpha_i/2}} \exp\left(-\frac{1}{2\Omega_i}\sum_{j\in J_i}(y_{ij} - \mathbf{f}_{ij})^2\right).$$

Hierbei werden die Bewertungen eines bestimmten i-ten Nutzers $y_{ij}, j \in J_i$ als ein Gauß'scher Prozeß aufgefaßt.

Da $P_{YYTK}(y_i|\mathbf{f}_i, \Omega_i)$ die Wahrscheinlichkeitsdichte der Normalverteilung ist, führt der EP-Algorithmus für diesen Ansatz zum gleichen Ergebnis für $\hat{\mathbf{f}}_i$ und $\hat{\mathcal{K}}_i$, das man auch analytisch aus der zugehörigen Posterior von $\mathbf{f}_i$ hätte ableiten können.

Sei $\mathbf{K}_{[\alpha_i][\alpha_i]} \in \mathbb{R}^{\alpha_i,\alpha_i}$ die Kovarianz zwischen allen Items, die der Nutzer i bewertet hat und sei $\mathbf{K}_{[J][\alpha_i]} \in \mathbb{R}^{J,\alpha_i}$ die Kovarianz zwischen allen Items und den Items, für die der i-te Nutzer Bewertungen abgegeben hat. $\mathbf{h}(\alpha_i) \in \mathbb{R}^{\alpha_i}$ sei der Vektor, der nur aus Komponenten von $\mathbf{h}$ besteht, die vom i-ten Nutzer beurteilt wurden. Analytisch erhält man für den Posterior-Erwartungswert von $\mathbf{f}_i$ auf Basis von Y_i (Williams, Rasmussen (1996), Rasmussen, Williams (2006), S. 27) die Formel

$$\hat{\mathbf{f}}_i = \mathbf{K}_{[J][\alpha_i]}(\mathbf{K}_{[\alpha_i][\alpha_i]} - \Omega_i E_{\alpha_i\times\alpha_i})^{-1}(y_i - \mathbf{h}(\alpha_i)) + \mathbf{h}.$$

Für den entsprechenden Posterior-Erwartungswert $\hat{\mathcal{K}}_i$ erhält man analytisch

$$\hat{\mathcal{K}}_i = \mathbf{K} - \mathbf{K}_{[J][\alpha_i]}(\mathbf{K}_{[\alpha_i][\alpha_i]} + \Omega_i E_{\alpha_i \times \alpha_i})^{-1}\mathbf{K}'_{[J][\alpha_i]}$$

(Williams, Rasmussen (1996)). Eine Herleitung dieser analytischen Ausdrücke findet sich bei Rasmussen (1996). Diese beiden analytischen Ausdrücke machen die Berechnung von $\hat{\mathbf{f}}_i$ und $\hat{\mathcal{K}}_i$ mittels des EP-Algorithmus, die in jedem einzelnen Schritt für alle Nutzer separat durchgeführt werden muß, überflüssig und führen somit zu einer drastischen Reduktion des Rechenaufwands.

Für den GP-Ansatz nach Yu et. al. (2006) läßt sich auch ein analytischer Ausdruck für Ω_i finden. Da Ω_i nur in $P(y_i|\mathbf{f}_i(\alpha_i), \Omega_i)$ vorkommt, reduziert sich die Bedingung erster Ordnung

$$\frac{\partial}{\partial \Omega_i}\left(\mathfrak{S}_u\left[\log P(y|\mathbf{h}, \mathbf{K}, \Omega)\right] + \log P(\mathbf{h}, \mathbf{K})\right) \equiv 0$$

auf

$$\frac{\partial}{\partial \Omega_i}\left(\sum_{i'=1}^{I}\int \hat{P}(\tilde{\mathbf{f}}_{i'}) \log P(y_{i'}|\tilde{\mathbf{f}}_{i'}(\alpha_{i'}), \Omega_{i'}) d\tilde{\mathbf{f}}_{i'}\right)$$

$$= \frac{1}{2\Omega_i}\left(-\alpha_i + \frac{1}{\Omega_i} E_{\hat{P}}\left(\sum_{j' \in J_i}(y_{ij'} - \mathbf{f}_{ij'})^2\right)\right)$$

$$= \frac{1}{2\Omega_i}\left(-\alpha_i + \frac{1}{\Omega_i} E_{\hat{P}}\left(\sum_{j' \in J_i}(y_{ij'} - \hat{\mathbf{f}}_{ij'} + \hat{\mathbf{f}}_{ij'} - \mathbf{f}_{ij'})^2\right)\right)$$

$$= \frac{1}{2\Omega_i}\left(-\alpha_i + \frac{1}{\Omega_i}\left(\sum_{j' \in J_i}(y_{ij'} - \hat{\mathbf{f}}_{ij'})^2 + E_{\hat{P}}\left(\sum_{j' \in J_i}(\hat{\mathbf{f}}_{ij'} - \mathbf{f}_{ij'})^2\right)\right)\right)$$

$$= \frac{1}{2\Omega_i}\left(-\alpha_i + \frac{1}{\Omega_i}\left(\sum_{j' \in J_i}\left((y_{ij'} - \hat{\mathbf{f}}_{ij'})^2 + \hat{\mathcal{K}}_{i,j'j'}\right)\right)\right) \equiv 0.$$

Hieraus folgt

$$\hat{\Omega}_i = \frac{1}{\alpha_i}\left(\sum_{j' \in J_i}\left((y_{ij'} - \hat{\mathbf{f}}_{ij'})^2 + \hat{\mathcal{K}}_{i,j'j'}\right)\right).$$

Hierdurch wird der Rechenaufwand nochmals verringert. Problematisch an diesem Ansatz ist, daß die Wahrscheinlichkeitsdichte der Normalverteilung mit der latenten Größe $\mathbf{f}_i$ als Verteilung der Bewertungsdaten benutzt wird. Dadurch werden die ordinalskalierten Daten einfach wie normalverteilte Daten behandelt. Zudem werden fehlende Werte ignoriert. Eine schematische Darstellung des hierarchischen GP-Verfahrens nach Yu et. al. (2006) findet sich in Abbildung 8.6.

Startwerte: $\mathbf{K}_0 = E_{J\times J}, \mathbf{h}_0 = 0 \in \mathbb{R}^J, \mathbf{f}_i = 0, \Omega_i, i = 1, ..., I,$
$\mathbf{h} = \mathbf{h}_0, \mathbf{K} = \mathbf{K}_0, n = 0, \mathbf{S_f} = \infty$

Solange $\mathbf{S_f} \geq \epsilon$:

$n \leftarrow n + 1$

$\hat{\mathcal{K}}_i = \mathbf{K} - \mathbf{K}_{[J][\alpha_i]}(\mathbf{K}_{[\alpha_i][\alpha_i]} + \Omega_i E_{\alpha_i\times\alpha_i})^{-1}\mathbf{K}'_{[J][\alpha_i]}, i = 1, ..., I$

$\hat{\mathbf{f}}_i = \mathbf{K}_{[J][\alpha_i]}(\mathbf{K}_{[\alpha_i][\alpha_i]} - \Omega_i E_{\alpha_i\times\alpha_i})^{-1}(y_i - \mathbf{h}(\alpha_i)) + \mathbf{h}, i = 1, ..., I$

$\hat{\Omega}_i = \frac{1}{\alpha_i}\left(\sum_{j'\in J_i}\left((y_{ij'} - \hat{\mathbf{f}}_{ij'})^2 + \hat{\mathcal{K}}_{i,j'j'}\right)\right), i = 1, ..., I$

$\hat{\mathbf{h}} = \frac{1}{I+\pi_h}\left(\sum_{i'=1}^{I}\hat{\mathbf{f}}_{i'} + \pi_h\mathbf{h}_0\right)$

$\hat{\mathbf{K}} = \frac{1}{I+\tau_h}\left(\pi_h(\hat{\mathbf{h}} - \mathbf{h}_0)(\hat{\mathbf{h}} - \mathbf{h}_0)' + \tau_h\mathbf{K}_0 + \sum_{i=1}^{I}\left[(\hat{\mathbf{f}}_i - \hat{\mathbf{h}})(\hat{\mathbf{f}}_i - \hat{\mathbf{h}})' + \hat{\mathcal{K}}_i\right]\right)$

$\mathbf{S_f} = \sum_{i'=1}^{I}(\hat{\mathbf{f}}_{i'} - \mathbf{f}_{i'})'(\hat{\mathbf{f}}_{i'} - \mathbf{f}_{i'})$

$\mathbf{f}_i \leftarrow \hat{\mathbf{f}}_i, \Omega_i \leftarrow \hat{\Omega}_i, i = 1, ..., I, \mathbf{h} \leftarrow \hat{\mathbf{h}}, \mathbf{K} \leftarrow \hat{\mathbf{K}}$

Abbildung 8.8: Schematische Darstellung des hierarchischen GP-Verfahrens nach Yu et. al. (2006)

Der Ansatz von Chu, Ghahramani (2005) stellt eine Alternative zu dem vorgestellten Ansatz dar. Es werden Schwellenwerte $\gamma_{i1} < ... < \gamma_{i(C-1)}, i \in \{1, ..., I\}$, benutzt und die Autoren postulieren für jede einzelne Komponente von Y_i die ideale Beziehung

$$P_{ideal}(y_{ij}|\mathbf{f}_{ij}, \Omega_i) = \begin{cases} 1, & \text{falls } \gamma_{i(y_{ij}-1)} < \mathbf{f}_{ij} \leq \gamma_{iy_{ij}} \\ 0, & \text{sonst} \end{cases}.$$

Hier werden $y_{ij} - 1$ und y_{ij} als Indizes verwendet. Als Likelihood verwenden Chu, Ghahramani (2005) eine abgeschwächte Variante dieser idealen Beziehung, die Abweichungen δ_{ij} von der Regel $\gamma_{y_{ij}-1} < \mathbf{f}_{ij} \leq \gamma_{y_{ij}}$ toleriert:

$$P^*_{CG}(y_{ij}|\mathbf{f}_{ij}, \Omega_i) = \int P_{ideal}(y_{ij}|\mathbf{f}_{ij} + \tilde{\delta}_{ij}, \Omega_i)\frac{1}{\sqrt{2\pi\Omega_i}}\exp\left(-\frac{\tilde{\delta}^2_{ij}}{2\Omega_i}\right)d\tilde{\delta}_{ij}.$$

Damit ergibt sich insgesamt

$$P_{CG}(Y_i|\mathbf{f}_i(\alpha_i), \Omega_i) = \prod_{j' \in J_i} P^*_{CG}(y_{ij'}|\mathbf{f}_{ij'}, \Omega_i).$$

In diesem Ansatz werden für jeden Nutzer i die Abweichungen vom Idealzustand $\delta_{ij}, j \in J_i$ als Gauß'scher Prozeß modelliert.

Das ursprüngliche Modell nach Chu, Ghahramani (2005) war kein hierarchischer Ansatz. In diesem Modell wurden für alle Nutzer separate Modelle berechnet. Erst nach der beschriebenen Modifikation dieses Ansatzes durch Yu et. al. (2006) wurde dieses Methode zu einem hierarchischen Ansatz.

In jedem einzelnen Schritt dieses Verfahrens muß für jeden der Nutzer getrennt $\hat{\mathbf{f}}_i$ und $\hat{\mathcal{K}}_i$ mit Hilfe des EP-Algorithmus berechnet werden. Zudem müssen in jedem Schritt für jeden einzelnen Nutzer die Schwellenwerte $\gamma_{i1}, ..., \gamma_{i(C-1)}$ mittels des konjugierten Gradientenverfahrens berechnet werden. Auch für $\Omega_i, i = 1, ..., I$, ist kein analytischer Ausdruck bekannt. Insgesamt ist der Rechenaufwand immens.

Für den Ansatz von Chu, Ghahramani spricht, daß dem ordinalen Skalenniveau der Bewertungsdaten Rechnung getragen wird. Auch dieses Modell ignoriert fehlende Daten.

Im Rahmen ihrer empirischen Untersuchungen vergleichen Yu et. al. (2006) beide Varianten der beschriebenen hierarchischen GP-Ansätze mit ihren nicht-hierarchischen Gegenstücken. Die hierarchischen Verfahren führen in allen Fällen zu erheblich besseren Ergebnissen als ihre nicht-hierarchischen Varianten. Das belegt eindrucksvoll den Nutzen der hierarchischen Ansätze.

Die hierarchischen Verfahren mit den Likelihood-Ansätzen nach Yu et. al. (2006) und nach Chu, Ghahramani (2005) wurden von Yu et. al. (2005) empirisch miteinander verglichen. Die schnellere GP-Variante nach Yu et. al. führte in den meisten Fällen zu deutlich genaueren Schätzern als der Likelihood-Ansatz nach Chu, Ghahramani (2005). Dies ist bemerkenswert, da die schnellere Variante das ordinale Skalenniveau der Bewertungsdaten vernachlässigt. Da die einfachere Variante nach Yu et. al. (2006) erstens genauer und zweitens mit einem deutlich geringerem Rechenaufwand verbunden ist, ist dieses Verfahren besser zur Beschreibung der Bewertungsdaten geeignet.

Es ist empfehlenswert, die Datenmatrix vorab einer Transformation zu unterziehen. Für jeden Nutzer sollte von allen Items, die der betreffende Nutzer

bewertet hat, die durchschnittliche Bewertung dieses Nutzers subtrahiert werden. Nachdem das beschriebene Verfahren auf die so transformierte Datenmatrix angewandt wurde und konvergiert ist, ergeben sich die Schätzer $\hat{Y}_{ij} = \mathbf{f}_{ij} + \bar{y}_{i.}, i = 1, ..., I, j = 1, ..., J$. Für $\mathbf{h}_0$ kann der Nullvektor gewählt werden. Als Startwerte für $\mathbf{h}$ empfiehlt sich $\mathbf{h}_0$, $\tau_h \mathbf{K}_0$ ist ein geeigneter Startwert für $\mathbf{K}$. Außerdem müssen Startwerte für $\Omega_i, i = 1, ..., I$, gewählt werden.

Beispiel 8.5:

In diesem Beispiel wird die Bewertung Y_{17} des ersten Nutzers hinsichtlich des bekannten Items $j = 7$ mit Hilfe des hierarchischen GP-Verfahrens nach Yu et. al. (2006) geschätzt. Als Datengrundlage wird wieder die Bewertungen aus Beispiel 5.1 verwendet. Es ist empfehlenswert, die Daten vor Beginn des Verfahrens einer Transformation zu unterziehen. Im Rahmen dieser Transformation muß von jeder Bewertung Y_{ij} die durchschnittliche Bewertung $\bar{Y}_{i.}$ des i-ten Nutzers subtrahiert werden. Auf diese Weise entsteht die transformierte Datenmatrix Y^t, die in Tabelle 8.2 wiedergegeben ist.

Als Hyperparameter π_h wählt man $\pi_h = 1$. $\mathbf{K}_0$ sollte dann die a priori bekannte oder vermutete Kovarianzstruktur der Items sein. Der Einfachheit halber wird hier $\mathbf{K}_0 = E_{J \times J}$ gewählt. Das beschriebene Verfahren nach Yu et. al. (2006) wird auf die transformierten Daten Y^t angewandt.

Y^t_{ij}	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$i=1$	-2,50	1,50	0,50	0,50	-0,50	0,50	-	-
$i=2$	-	1,00	-1,00	-1,00	-1,00	2,00	-	0,00
$i=3$	-	0,50	0,50	0,50	-0,50	-	1,50	-2,50
$i=4$	2,00	-1,00	-1,00	-	-1,00	0,00	-	1,00
$i=5$	0,50	-	-1,50	-0,50	0,50	0,50	-	0,50
$i=6$	-2,67	1,33	0,33	0,33	0,33	-	0,33	-

Tabelle 8.2: Datentransformation (Beispiel 8.5)

Als Startwert für $\mathbf{h}$ wird der Nullvektor verwendet. Zu Beginn des Verfahrens sei außerdem $\Omega_i = 0, i = 1, ..., I$, da diese Wahl die Berechnung erheblich vereinfacht.

Im ersten Schritt des Verfahrens nach Yu et. al. (2006) berechnet man bezüglich des ersten Nutzers den Vektor $\mathbf{f}_1$ wie folgt:

$$\hat{\mathbf{f}}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \text{-2,50} \\ \text{1,50} \\ \text{0,50} \\ \text{0,50} \\ \text{-0,50} \\ \text{0,50} \end{pmatrix} = \begin{pmatrix} \text{-2,50} \\ \text{1,50} \\ \text{0,50} \\ \text{0,50} \\ \text{-0,50} \\ \text{0,50} \\ \text{0,00} \\ \text{0,00} \end{pmatrix}$$

Außerdem erhält man

$$\hat{\mathcal{K}}_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Damit erhält man in der ersten Iteration $\Omega_1 = 1/3$. Für die übrigen Nutzer ergeben sich entsprechende Matrizen $\hat{\mathcal{K}}_i, i = 2, ..., I$. Als Posterior-Erwartungswerte der latenten Variable erhält man in der ersten Iteration die Vektoren

$$\underbrace{\begin{pmatrix} \text{0,00} \\ \text{1,00} \\ \text{-1,00} \\ \text{-1,00} \\ \text{-1,00} \\ \text{2,00} \\ \text{0,00} \\ \text{0,00} \end{pmatrix}}_{=\hat{\mathbf{f}}_2}, \underbrace{\begin{pmatrix} \text{0,00} \\ \text{0,50} \\ \text{0,50} \\ \text{0,50} \\ \text{-0,50} \\ \text{0,00} \\ \text{1,50} \\ \text{-2,50} \end{pmatrix}}_{=\hat{\mathbf{f}}_3}, \underbrace{\begin{pmatrix} \text{2,00} \\ \text{-1,00} \\ \text{-1,00} \\ \text{0,00} \\ \text{-1,00} \\ \text{0,00} \\ \text{0,00} \\ \text{1,00} \end{pmatrix}}_{=\hat{\mathbf{f}}_4}, \underbrace{\begin{pmatrix} \text{0,50} \\ \text{0,00} \\ \text{-1,50} \\ \text{-0,50} \\ \text{0,50} \\ \text{0,50} \\ \text{0,00} \\ \text{0,50} \end{pmatrix}}_{\hat{\mathbf{f}}_5}, \underbrace{\begin{pmatrix} \text{-2,67} \\ \text{1,33} \\ \text{0,33} \\ \text{0,33} \\ \text{0,33} \\ \text{0,00} \\ \text{0,33} \\ \text{0,00} \end{pmatrix}}_{=\hat{\mathbf{f}}_6}.$$

Hiermit erhält man in der ersten Iteration den neuen Wert $\hat{\mathbf{h}}$ für $\mathbf{h}$:

$$\hat{\mathbf{h}} = \begin{pmatrix} 0{,}38 \\ 0{,}48 \\ \text{-}0{,}31 \\ \text{-}0{,}02 \\ \text{-}0{,}31 \\ 0{,}43 \\ 0{,}26 \\ \text{-}0{,}14 \end{pmatrix}.$$

Außerdem erhält man in der ersten Interation

$$\hat{\mathbf{K}} = \begin{pmatrix} 2{,}76 & \text{-}1{,}13 & \text{-}0{,}80 & \text{-}0{,}34 & \text{-}0{,}31 & 0{,}02 & \text{-}0{,}03 & 0{,}26 \\ \text{-}1{,}13 & 0{,}95 & 0{,}35 & 0{,}07 & 0{,}07 & 0{,}19 & 0{,}05 & \text{-}0{,}25 \\ \text{-}0{,}80 & 0{,}35 & 0{,}74 & 0{,}32 & 0{,}03 & \text{-}0{,}22 & 0{,}20 & \text{-}0{,}46 \\ \text{-}0{,}34 & 0{,}07 & 0{,}32 & 0{,}55 & 0{,}04 & \text{-}0{,}27 & 0{,}13 & \text{-}0{,}21 \\ \text{-}0{,}31 & 0{,}07 & 0{,}03 & 0{,}04 & 0.46 & \text{-}0{,}15 & \text{-}0{,}01 & 0{,}03 \\ 0{,}02 & 0{,}19 & \text{-}0{,}22 & \text{-}0{,}27 & \text{-}0{,}15 & 0{,}88 & \text{-}0{,}11 & 0{,}10 \\ \text{-}0{,}03 & 0{,}05 & 0{,}20 & 0.13 & \text{-}0{,}01 & \text{-}0{,}11 & 0{,}98 & \text{-}0{,}49 \\ 0{,}26 & \text{-}0{,}25 & \text{-}0{,}46 & \text{-}0{,}21 & 0{,}03 & 0{,}10 & \text{-}0{,}49 & 1{,}46 \end{pmatrix}.$$

Ab der zweiten Iteration setzt man zu Beginn jeder neuen Iteration $\mathbf{h}_0 = \mathbf{h}$ und $\mathbf{K}_0 = \mathbf{K}$, wobei $\mathbf{h}$ und $\mathbf{K}$ die in der vorherigen Iteration berechneten neuen Werte bezeichnen.

Nach fünf Iterationen konvergiert das Verfahren und man erhält als Posterior-Erwartungswert der latenten Größe in Bezug auf den ersten Nutzer

$$\hat{\mathbf{f}}_1 = (-2,50, 1,49, 0,50, 0,49, -0,49, 0,50, 0,83, -1,62)'.$$

Damit ergibt sich als Schätzer für Bernds Bewertung des Films Barry Lyndon $\hat{Y}_{17} = \hat{\mathbf{f}}_{17} + \bar{y}_{1\cdot} = 4,33$.

8.3 Hybride Verfahren

8.3.1 Hybrides Hierarchisches Verfahren (HBLR)

Das hybride hierarchische Verfahren (HBLR) beruht auf einem linearen Regressionsansatz. Hier wird das hierarchische lineare Modell nach Rossi et. al. (1997) benutzt. Dieses Modell ist dem hybriden Regressionsansatz von Ansari et. al. (2000) sehr ähnlich. Dieser lineare Ansatz wurde zur Vorhersage von Film-Bewertungen eingesetzt. Im Rahmen dieses Ansatzes werden Genre-Zugehörigkeit und die Beurteilung des Films durch Experten als Item-Eigenschaften benutzt. Zudem werden auch demographische Daten der Nutzer (nämlich ihr Geschlecht und Alter) verwendet. Sowohl die demographischen Daten als auch die Item-Eigenschaften werden in dem Regressionsansatz als exogene Größen eingesetzt. Als endogene Variable dient die entsprechende Bewertung des betreffenden Nutzers in Bezug auf den jeweiligen Film.

Die Resultate von Ansari et. al. lassen darauf schließen, daß sich durch die zusätzliche Berücksichtigung von Geschlecht und Alter nur geringfügige Verbesserungen der Schätzgenauigkeit erreichen lassen. Außerdem ist zu bedenken, daß es Nutzer abschrecken könnte, wenn sie vorab Informationen über sich selbst preisgeben sollen. Ansari et. al. (2000) haben bei ihrer empirischen Studie nicht untersucht, wie stark eine Ergebnisverbesserung durch zwei weitere Bewertungen pro Nutzer ausfallen würde. Da die Erhöhung der Schätzgenauigkeit durch die Berücksichtigung von Geschlecht und Alter sehr klein ist, darf man annehmen, daß durch die Berücksichtigung von zwei weiteren Filmen pro Nutzer mindestens genauso große Ergebnisverbesserungen zu erzielen gewesen wären. Aus diesen Gründen werden Geschlecht und Alter hier nicht verwendet. Ohne diese beiden demographischen Größen entspricht das Modell von Ansari et. al. (2000) dem Ansatz von Rossi et. al. (1997).

Desweiteren werden hier an Stelle der Genre-Zugehörigkeit die bereits erwähnten Film-Eigenschaften wie z.B. das Ausmaß der im Film gezeigten Gewalt oder der Grad der Spannung, die durch den Film erzeugt wird, benutzt. Hierdurch wird die Beschreibung der Items wesentlich genauer. Der Vektor X_{ij} beinhaltet neben einer 1 für den Achsenabschnitt die Eigenschaften des j-ten Items und die durchschnittliche Bewertung des j Items durch professionelle Film-Kritiker. $\beta^i, i = 1, ..., I$, und $X_{ij}, j \in J_i, i = 1, ..., I$, haben beide $\kappa_A + 1$ Komponenten. Die Beziehung zwischen der Bewertung Y_{ij} und sowohl dem individuellen Geschmack

des Nutzers (β^i) als auch dem Vektor X_{ij} kann durch das lineare Modell

$$Y_{ij} = X'_{ij}\beta^i + \epsilon_{ij}$$

mit $\epsilon_{ij} \sim u.i.v.N(0, \sigma_i^2)$, mit $i = 1, ..., I$, $j = 1, ..., \alpha_i$ und $\alpha_i = |J_i|$ beschrieben werden. Auf diese Weise erhält man I verschiedene Modelle

$$Y_i = X_i\beta^i + \epsilon_i,$$

wobei $Y'_i = (Y_{i1}, ..., Y_{i\alpha_i})$, $X'_i = (X_{i1}, ..., X_{i\alpha_i})$, $\epsilon'_i = (\epsilon_{i1}, ..., \epsilon_{i\alpha_i})$ und

$$\epsilon_i \sim u.i.v.N(0, \sigma_i^2 I_{\alpha_i \times \alpha_i}), i = 1, ..., I,$$

ist. Zwischen den unerschiedlichen Geschmäckern der Nutzer und der allgemeinen geschmacklichen Tendenz wird der Zusammenhang

$$\beta^i = \Delta' z_i + v_i$$

postuliert. z'_i ist ein d-komponentiger Zeilenvektor, der für jeden Nutzer die Zeilen der Matrix Δ individuell gewichten kann. Auf diese Weise können die Zeilenvektoren $z'_i, i = 1, ..., I$, die jeweiligen Nutzer in Verbindung mit der Matrix Δ charakterisieren.
Die Matrizen

$$Z = \begin{pmatrix} z'_1 \\ \vdots \\ z'_I \end{pmatrix} \in \mathbb{R}^{I,d} \quad \text{und} \quad \Delta \in \mathbb{R}^{d,\kappa_A+1}$$

können daher benutzt werden, um alle Nutzer gemeinsam zu charakterisieren. $\Delta \in \mathbb{R}^{d,\kappa_A+1}$ ist eine Matrix aus Regressionskoeffizienten. Ohne jegliches Vorwissen über die Nutzer gilt $d = 1$. Dann ist Z ein I-dimensionaler Vektor aus lauter Einsen. Δ' ist in diesem Fall die Verallgemeinerung aller $\beta^i, i = 1, ..., I$.

Zudem wird angenommen, daß $v_i \sim u.i.v.N(0, V_\beta)$ ist. Falls Δ, V_β und σ_i^2 gegeben sind, ist die multivariate Normalverteilung $N(\Delta' z_i, V_\beta)$ für $\beta^i, i = 1, ..., I$, konjugierte Verteilung zur Likelihood

$$P(y_i|\beta^i, \sigma_i^2) \propto \frac{1}{|\sigma_i^2 I_{\alpha_i \times \alpha_i}|^{1/2}} \exp\left\{-\frac{(y_i - X_i\beta^i)'(y_i - X_i\beta^i)}{2\sigma_i^2}\right\}.$$

Daher ergibt sich als Posterior für β^i unter der Voraussetzung, daß σ_i^2 gegeben ist, ebenfalls die Normalverteilung:

$$\beta^i|\sigma_i^2, y_i, X_i \sim N\left(\bar{\beta}^i, \left(\frac{X_i'X_i}{\sigma_i^2} + V_\beta^{-1}\right)^{-1}\right),$$

wobei

$$\bar{\beta}^i = \left(\frac{X_i'X_i}{\sigma_i^2} + V_\beta^{-1}\right)^{-1}\left(\frac{X_i'y_i}{\sigma_i^2} + V_\beta^{-1}\Delta' z_i\right)$$

gilt. Andererseits ist für bekannte $\beta^i, i = 1, ..., I$, die zur obigen Likelihood konjugierte inverse Wishart-Verteilung $IW(\nu, V_i)$ eine mögliche Wahl für die Prior von $\sigma_i^2, i = 1, ..., I$. Somit ist die Posterior von σ_i^2 für gegebene $\beta^i, i = 1, ..., I$, auch eine inverse Wishart-Verteilung:

$$\sigma_i^2|\beta^i, X_i, y_i \sim IW\left(\nu + \alpha_i, \sqrt{\frac{\nu V_i^2 + \alpha_i \varphi_i^2}{\nu + \alpha_i}}\right),$$

wobei

$$\varphi_i^2 = \frac{1}{\alpha_i}(y_i - X_i\beta^i)'(y_i - X_i\beta^i).$$

Für gegebene Δ und V_β können diese beiden (bedingten) Posteriors dazu benutzt werden, um einen Gibbs-Sampler zu entwickeln, so daß abwechselnd σ_i^2 für die neuesten Werte von β^i für alle $i \in \{1, ..., I\}$ und dann β^i auf Grundlage der zuvor ermittelten $\sigma_i^2, i \in \{1, ..., I\}$ bestimmt werden kann.
Da es sich um einen hierarchischen Ansatz handelt, werden auch Priors für Δ und V_β verwendet. Die Likelihood des multivariaten Regressionsmodells $B = Z\Delta + V$ (mit $B' = (\beta^1, ..., \beta^I), V' = (v_1, ..., v_I) \in \mathbb{R}^{\kappa_A+1,I}$ und $Z' = (z_1, ..., z_I) \in \mathbb{R}^{d,I}$)

$$P(B|Z, \Delta, V_\beta) \propto \frac{1}{|V_\beta|^{I/2}} \exp\left(-\frac{1}{2} tr((B - Z\Delta)'(B - Z\Delta)V_\beta^{-1})\right)$$

kann wegen der Beziehung

$$Z'(B - \hat{B}) = Z'(B - Z(Z'Z)^{-1}Z'B) = (Z' - Z')B = 0,$$

mit $\hat{B} = Z\hat{\Delta}, \hat{\Delta} = (Z'Z)^{-1}Z'B$, und der hieraus folgenden Relation

$$(B - Z\Delta)'(B - Z\Delta) = \underbrace{(B - Z\hat{\Delta})'(B - Z\hat{\Delta})}_{=S_B} + (\Delta - \hat{\Delta})'Z'Z(\Delta - \hat{\Delta})$$

auch als das folgende Produkt dargestellt werden:

$$\begin{aligned} P(B|Z, \Delta, V_\beta) \quad &\propto \frac{1}{|V_\beta|^{(I-d)/2}} \exp\left(-\tfrac{1}{2} tr(S_B V_\beta^{-1})\right) \\ &\cdot \frac{1}{|V_\beta|^{d/2}} \exp\left(-\tfrac{1}{2} tr\left((\Delta - \hat{\Delta})'Z'Z(\Delta - \hat{\Delta})V_\beta^{-1}\right)\right). \end{aligned}$$

Im Hinblick auf weitere Umformungen der Likelihood ist es erforderlich zwei weitere Definitionen einzuführen. *vec* steht für den Vektor-Operator, der angewandt auf eine Matrix

$$\Delta = \begin{pmatrix} \delta_{11} & \cdots & \delta_{1(\kappa_A+1)} \\ \vdots & \ddots & \vdots \\ \delta_{d1} & \cdots & \delta_{d(\kappa_A+1)} \end{pmatrix} = (\Delta_1 \cdots \Delta_{\kappa_A+1})$$

deren Spalten $\Delta_1, ..., \Delta_{\kappa_A+1}$ nacheinander in einen $d\kappa$-dimensionalen Vektor einordnet:

$$vec(\Delta) = \begin{pmatrix} \Delta_1 \\ \vdots \\ \Delta_{\kappa_A+1} \end{pmatrix}.$$

$\otimes$ bezeichnet das Kronecker-Produkt, das zwischen zwei Matrizen $\tilde{A} \in \mathbb{R}^{m,n}$ und $\tilde{B} \in \mathbb{R}^{u,v}$ durch

$$\tilde{A} \otimes \tilde{B} = \begin{pmatrix} \tilde{a}_{11}\tilde{B} & \tilde{a}_{12}\tilde{B} & \cdots & \tilde{a}_{1n}\tilde{B} \\ \tilde{a}_{21}\tilde{B} & \tilde{a}_{22}\tilde{B} & \cdots & \tilde{a}_{2n}\tilde{B} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}_{m1}\tilde{B} & \tilde{a}_{m2}\tilde{B} & \cdots & \tilde{a}_{mn}\tilde{B} \end{pmatrix} \in \mathbb{R}^{mu,nv}$$

definiert ist. Nach Magnus, Neudecker (1988) gelten für geeignet dimensionierte Matrizen $\tilde{A}, \tilde{B}$ und $\tilde{C}$ die Beziehungen

$$\begin{aligned} tr(\tilde{A}'\tilde{B}) &= vec(\tilde{A})'vec(\tilde{B}) \\ vec(\tilde{A}\tilde{B}\tilde{C}) &= \tilde{C}' \otimes \tilde{A}vec(\tilde{B}) \end{aligned}$$

Wegen der obigen Beziehungen ergibt sich

$$\begin{aligned} tr((\Delta - \hat{\Delta})'Z'Z(\Delta - \hat{\Delta})V_\beta^{-1}) &= vec(\Delta - \hat{\Delta})'vec(Z'Z(\Delta - \hat{\Delta})V_\beta^{-1}) \\ &= vec(\Delta - \hat{\Delta})'V_\beta^{-1} \otimes Z'Zvec(\Delta - \hat{\Delta}). \end{aligned}$$

Somit erhält man

$$\begin{aligned} P(B|Z, \Delta, V_\beta) \quad \propto & \frac{1}{|V_\beta|^{(I-d)/2}} \exp\left(-\tfrac{1}{2} tr(S_B V_\beta^{-1})\right) \\ & \cdot \frac{1}{|V_\beta|^{d/2}} \exp\left(-\tfrac{1}{2} vec(\Delta - \hat{\Delta})'V_\beta^{-1} \otimes Z'Zvec(\Delta - \hat{\Delta})\right) . \end{aligned}$$

Aufgrund dieser Darstellung der Likelihood erkennt man, daß die konjugierte Prior ein Produkt $P(vec(\Delta), V_\beta) = P(vec(\Delta)|V_\beta)P(V_\beta)$ aus einer Normalverteilungsdichte $P(vec(\Delta)|V_\beta)$ für $vec(\Delta)$ und einer inversen Wishart Prior für V_β ist. Im einzelnen gilt:

$$\begin{aligned} V_\beta|\nu_0, V_0 &\sim IW(\nu_0, V_0) \\ vec(\Delta)|V_\beta, vec(\bar{\Delta}), A &\sim N(vec(\bar{\Delta}), V_\beta \otimes A^{-1}) \end{aligned} .$$

Hier sind $\nu_0 \in \mathbb{R}, V_0 \in \mathbb{R}^{\kappa_A+1,\kappa_A+1}, A \in \mathbb{R}^{d,d}$ und $\bar{\Delta} \in \mathbb{R}^{I,d}$ sind Hyperprior-Parameter.
Die zugehörigen bedingten Posteriors sind

$$\begin{aligned} V_\beta|\{\beta^i\}_{i=1}^I, \nu_0, V_0 &\sim IW\left(\nu_0 + I, V_0 + \sum_{i'=1}^{I} (\beta^{i'} - \bar{\beta}^{i'})(\beta^{i'} - \bar{\beta}^{i'})'\right) \\ vec(\Delta)|V_\beta, \bar{\Delta}, A &\sim N(vec(\bar{\Delta}), V_\beta \otimes (A + Z'Z)^{-1}). \end{aligned}$$

Insgesamt können alle bedingten Posteriors benutzt werden, um eine neue Realisation der jeweiligen Zufallszahl zu bestimmen, sofern alle übrigen Zufallsvariablen ihre (letzten) Realisationen annehmen.

Die angegebenen bedingten Posteriors lassen sich dazu benutzen, einen Gibbs-Sampling Algorithmus (Gibbs-Sampler) für das hierarchische lineare Modell zu konzipieren (Abbildung 8.9).

Startwerte: $\{\sigma^2_{i,0}\}^I_1$, Δ_0, $V_{\beta,0}$ und $n = 0$.
Solange $n \leq n_{MAX}$:

1. ziehe $\beta^i_{n+1} \sim N\left(\bar{\beta}^i_{n+1}, \left(\frac{X_i' X_i}{\sigma^2_{i,n}} + V^{-1}_{\beta,n}\right)^{-1}\right)$, $i = 1, ..., I$
   mit $\bar{\beta}^i_{n+1} = \left(\frac{X_i' X_i}{\sigma^2_{i,n}} + V^{-1}_{\beta,n}\right)^{-1} \left(\frac{X_i' y_i}{\sigma^2_{i,n}} + V^{-1}_{\beta,n}\Delta_n' z_i\right), i = 1, ..., I$

2. ziehe $\sigma^2_{i,n+1} \sim IW\left(\nu + \alpha_i, \sqrt{\frac{\nu V_i^2 + (y_i - X_i\beta^i_{n+1})'(y_i - X_i\beta^i_{n+1})}{\nu + \alpha_i}}\right)$,
   $i = 1, ..., I$

3. ziehe $V_{\beta,n+1} \sim IW\left(\nu_0 + I, V_0 + \sum_{i=1}^{I}(\beta^i_{n+1} - \bar{\beta}^i_{n+1})(\beta^i_{n+1} - \bar{\beta}^i_{n+1})'\right)$,
   mit $\bar{\beta}^i_{n+1} = \left(\frac{X_i' X_i}{\sigma^2_{i,n}} + V^{-1}_{\beta,n}\right)^{-1} \left(\frac{X_i' y_i}{\sigma^2_{i,n}} + V^{-1}_{\beta,n}\Delta_n' z_i\right), i = 1, ..., I$

4. ziehe $vec(\Delta_{n+1}) \sim N\left(vec(\bar{\Delta}), V_{\beta,n+1} \otimes (A + Z'Z)^{-1}\right)$

5. $n \leftarrow n + 1$

Abbildung 8.9: Gibbs-Sampler für das Hierarchische Lineare Modell

Die schnelle Konvergenz dieses Gibbs-Samplers wurde empirisch nachgewiesen (McCulloch, Rossi (1994)).

Die gesamte in allen y_i (β^i), $i = 1, ..., I$, enthaltene Information wird durch V_β and Δ gebündelt. Das auf diese Weise erhaltene Wissen um die allgemeinen Tendenzen der Daten wird dann wiederum zusammen mit den Daten y_i dazu eingesetzt, die individuellen Tendenzen der Dateneinheit Y_i, $\beta^i, \sigma^2_i, i = 1, ..., I$, zu schätzen. Deshalb ist es möglich, auch individuelle Präferenzen von Nutzern zu schätzen, die nur sehr wenige Bewertungen abgegeben haben. Bei diesen Nutzern erhalten die in der Prior enthaltenen allgemeinen Tendenzen einen stärkeren

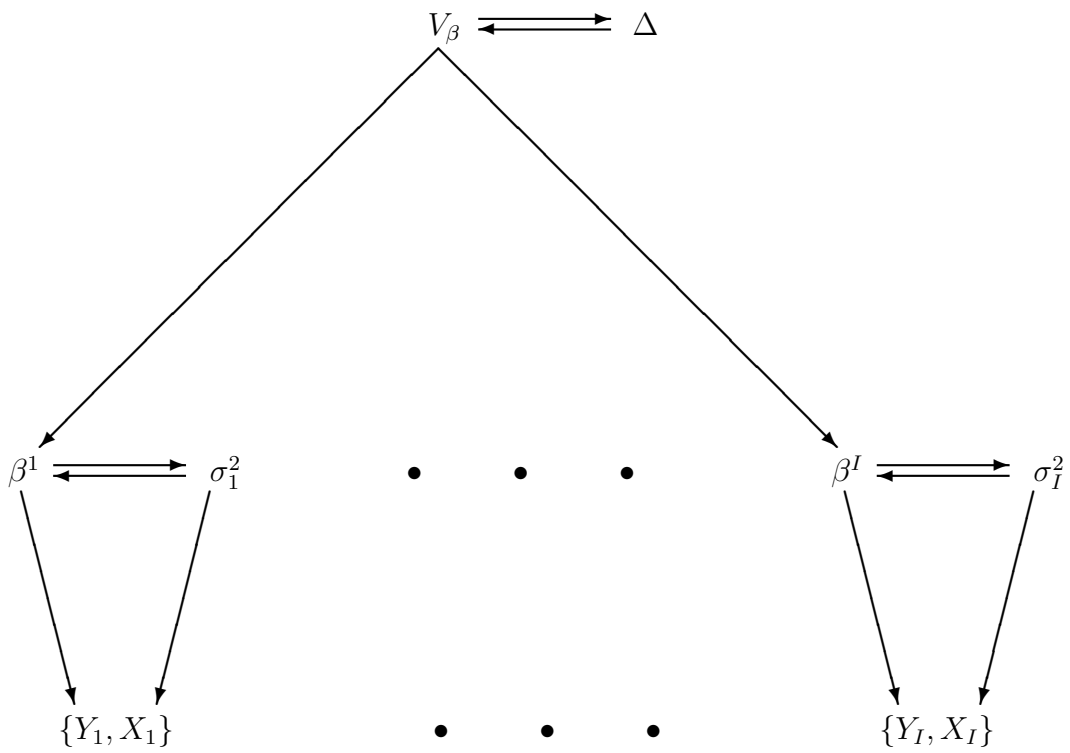

Abbildung 8.10: Struktur des von Rossi et. al. (1996) verwendeten Hierarchischen Linearen Modells

Einfluß auf die Posterior. Somit wird Information von allen verfügbaren Einheiten gebündelt und zur Ersetzung fehlender Information benutzt. Dagegen ist bei Nutzern, die verhältnismäßig viele Items bewertet haben, der Einfluß der Prior - und damit der allgemeinen Tendenzen - auf die Posterior deutlich geringer. Auf diese Weise erhalten die allgemeinen Tendenzen vor allem bei den Nutzern ein starkes Gewicht, für die mit anderen Methoden gar keine Nutzer-spezifischen Schätzer berechenbar gewesen wären. Wie bereits erwähnt sind in den Fällen, in denen die Anzahl der Bewertungen ausreichen würde, um auf die einzelnen Nutzer beschränkte individuelle OLS-Schätzer zu berechnen, die individuellen Schätzer, die sich mittels eines Hierarchischen Bayes-Ansatzes bestimmen lassen, genauer (Gelman et. al. (2004)).

Beispiel 8.6:

Es werden wieder die Daten aus Beispiel 4.1 und 5.1 zugrunde gelegt. Zusätzlich

zu diesen Eigenschafts- und Bewertungsdaten ist in diesem Beispiel für jeden der 8 Filme $j = 1, ..., 8$ eine fiktive durchschnittliche Kritikerbewertung $\bar{Y}_{\cdot j}^{C}$ (Kritik) gegeben:

j	Titel	a_{j1}	a_{j2}	a_{j3}	a_{j4}	a_{j5}	a_{j6}	Kritik $\bar{Y}_{\cdot j}^{C}$
1	Armageddon	6	9	6	7	7	4	2,33
2	Magnolia	3	2	5	5	5	8	4,25
3	Casino	9	5	3	4	4	6	4,50
4	Taxi Driver	9	4	2	1	5	10	4,33
5	Platoon	9	9	3	0	9	9	3,47
6	Dangerous Liaisons	3	2	4	4	2	8	3,67
7	Barry Lyndon	5	4	2	4	3	8	4,45
8	Twister	6	8	4	4	4	3	2,78

Tabelle 8.3: Beispiel 8.2: Filmeigenschaften und durchschnittliche Kritikerbewertung

Für $\nu_0 = 10, V_0 = 0,01 \cdot diag(400, 1, 1, 1, 1, 1, 1, 10), \bar{\Delta} = (0, 0, 0, 0, 0, 0, 0, 0)$ und $A = 0,01$ sowie $n_{MAX} = 25000$ und $n_{BURN} = 12500$ erhält man mit den Startwerten

$$\sigma_{i,0}^2 = \frac{1}{|J_i| - 1} \sum_{j \in J_i} (y_{ij} - \bar{y}_{i\cdot})^2, \;\; \Delta_0 = (0, 0, 0, 0, 0, 0, 0, 0) \;\text{ und }\; V_{\beta,0} = E_{8\times 8},$$

mit $E_{8\times 8}$ als Einheitsmatrix insgesamt $n_{MAX} - n_{BURN}$ verschiedene Ziehungen $\{\beta_n^i\}_1^I$, $\{\sigma_{i,n}^2\}_1^I$, Δ_n, $V_{\beta,n}, n = n_{BURN} + 1, ..., n_{MAX}$. Da es sich um kontinuierliche Werte handelt, kommt man durch Bildung der entsprechenden Mittelwerte zu guten Approximationen für die Erwartungswerte. Diese geschätzten Erwartungswerte sind das Bayes'sche Analogon der Schätzer. Man erhält

$$\widehat{E(\beta^i)} = \frac{1}{n_{MAX} - n_{BURN}} \sum_{n=n_{BURN}+1}^{n_{MAX}} \beta_n^i \;\text{ und }\; \widehat{E(Y_{ij})} = \widehat{E(X_j'\beta^i)} = X_j' \widehat{E(\beta^i)},$$

mit $X_j' = (1 a_j')$.

Auf diese Weise ergeben sich

$$\widehat{E(\beta^i)} = \begin{pmatrix} 3{,}276 \\ -0{,}230 \\ 0{,}037 \\ -0{,}085 \\ -0{,}018 \\ -0{,}040 \\ 0{,}120 \\ 0{,}322 \end{pmatrix} \quad \text{und} \quad \widehat{E(\Delta')} = \begin{pmatrix} 3{,}349 \\ -0{,}236 \\ 0{,}061 \\ -0{,}072 \\ -0{,}005 \\ -0{,}034 \\ 0{,}108 \\ 0{,}230 \end{pmatrix}.$$

Der ersten Komponente des Vektors $\widehat{E(\Delta')}$ entnimmt man, daß die Nutzer tendenziell positiv bewerten. Je mehr Gewalt in einem Film dargeboten wird, umso schlechter fällt in der Regel seine Bewertung aus (zweite Komponente). Dagegen fällt die Bewertung für einen Film umso besser aus, je höher die durchschnittliche Kritikerbewertung ausfiel (letzte Komponente) und je stärker sich der Film sich mit Charakterentwicklung auseinander setzt.

Der Nutzer Bernd bewertet Filme mit guten Kritikerbewertungen tendenziell sogar noch deutlich höher als die übrigen Nutzer. Er scheint auch noch etwas mehr Wert auf Charakterentwicklung zu legen.

Als Schätzer für Bernds Bewertung des Films Barry Lyndon ergibt sich auf diese Weise $\widehat{E(X'_{17}\beta^1)} = 4,30$.

8.3.2 Hybrides HB-Verfahren auf Basis zweimodaler Cluster

Ohne jegliche Vorinformation über die Nutzer ist es sinnvoll, $d = 1$ zu wählen. In diesem Fall ist Z ein I-dimensionaler Vektor, dessen Einträge alle 1 sind. Dann ist Δ' ein Vektor mit derselben Dimension wie β^i, $i = 1, ..., I$, und kann als gemeinsamer Prior-Erwartungswert aller β^i, $i = 1, ..., I$, aufgefaßt werden. Der Hyperprior-Parameter $\bar{\Delta}$ ist unter diesen Umständen gleich $vec(\bar{\Delta})$. Außerdem ist $\bar{\Delta}$ der Erwartungswert von Δ. Ohne Vorinformation sollte $\bar{\Delta}$ nur Nullen enthalten. Der zugehörige Prior-Typ wird im folgenden als Z_1-Prior bezeichnet.

Im Unterschied zur Z_1-Prior basiert die Z_2-Prior auf der Clusterzugehörigkeit der Nutzer. Da das Resultat eines zweimodalen Clusterverfahrens die Interaktion

zwischen den Zeilen (Nutzern) und Spalten (Items) der Datenmatrix in komprimierter Form wiedergibt, enthalten die entsprechenden Nutzer-Cluster Personen, deren Bewertungsverhalten hinsichtlich verschiedener Gruppen von Items ähnlich ist. Die Zuordnung zu einem bestimmten Nutzer-Cluster kann daher als Beschreibung des Bewertungsverhaltens oder Geschmacks des Nutzers interpretiert werden. Man kann daher die Auffassung vertreten, daß die Cluster-Zugehörigkeit der Nutzer ein besserer Hinweis auf deren Präferenzen ist, als ihr Alter und Geschlecht. Daher und aus weiteren bereits dargelegten Gründen werden hier abweichend vom Ansatz von Ansari et. al. (2000) nicht Alter und Geschlecht sondern die Cluster-Zugehörigkeit der Nutzer als Charakteristika der Nutzer verwendet. Ein einfacher Ansatz, um das Ergebnis eines zweimodalen Clusterverfahrens in das beschriebene hierarchische lineare Modell miteinfließen zu lassen, ist $Z = P$ und $d = K$ zu setzen. In diesem Fall wäre Δ eine $K \times (\kappa_A + 1)$-Matrix:

$$\Delta = \begin{pmatrix} \bar{\beta}^{\{1\}\prime} \\ \vdots \\ \bar{\beta}^{\{K\}\prime} \end{pmatrix},$$

wobei die einzelnen $\bar{\beta}^{\{k\}}$ die zum k-ten Nutzer-Cluster gehörenden Parameter sind, $k = 1, ..., K,$. Für die Hyperprior-Parameter Matrix $\bar{\Delta}$ gilt dann entsprechend $\bar{\Delta}' = (\bar{\beta}_H^{\{1\}}, ..., \bar{\beta}_H^{\{K\}})$. Auf diese Weise kann das a priori vorhandene Wissen um das ähnliche Bewertungsverhalten der Nutzer die Prior von $\beta^i, i = 1, ..., I,$ dominieren.

Beispiel 8.7:

In diesem Beispiel wird zusätzlich zu den bereits in Beispiel 5.1 verwendeten Daten die Matrix P aus Beispiel 5.11 benutzt.
Diese Vorinformation kann mittels einer Z_2-Prior in das hierarchische lineare Modell integriert werden:

$$Z = P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Damit ergeben sich mit $A = 0,01E_{2\times 2}$ ceteris paribus die Ergebnisse

$$\widehat{E(\beta^i)} = \begin{pmatrix} \text{-0,339} \\ \text{-0,107} \\ \text{-0,110} \\ \text{-0,105} \\ \text{0,004} \\ \text{0,126} \\ \text{0,093} \\ \text{1,085} \end{pmatrix} \quad \text{und} \quad \widehat{E(\Delta')} = \begin{pmatrix} \text{-0,378} & \text{5,762} \\ \text{-0,106} & \text{-0,124} \\ \text{-0,107} & \text{0,014} \\ \text{-0,107} & \text{0,090} \\ \text{0,003} & \text{0,013} \\ \text{0,125} & \text{-0,045} \\ \text{0,093} & \text{0,085} \\ \text{1,082} & \text{-0,722} \end{pmatrix}.$$

Durch die Trennung der Nutzer anhand ihrer Cluster-Zugehörigkeit gelingt eine weitergehende Berücksichtigung der Cluster-spezifischen Heterogenität, die ihren Ausdruck auch in veränderten Schätzern $\beta^i, i = 1, ..., I$, für alle einzelnen Nutzer findet.

Man erhält so einen etwas höheren geschätzten Erwartungswert für Bernds Bewertung des Films Barry Lyndon: $\widehat{E(X'_{17}\beta^1)} = 4,44$.

Sei $e_I \in \mathbb{R}^I$ ein I-dimensionaler Vektor, dessen Komponenten alle Eins sind. Durch die Wahl von $(e_I P)$ für Z kann für jeden Nutzer sowohl allgemeine als auch Cluster-spezifische Information gewonnen und verwendet werden (Rossi et. al. (1996)). In diesem Fall muß man auch der Matrix Δ eine weitere Zeile $\bar{\beta}^{\{0\}'}$ hinzufügen. $\bar{\beta}^{\{0\}}$ enthält dann die allgemeinen im gesamten Datensatz vorhandenen Tendenzen während die einzelnen $\bar{\beta}^{\{k\}}, k = 1, ..., K$, die Cluster-spezifischen Abweichungen dieser allgemeinen Tendenzen enthält. Die zugehörige Prior wird hier als Z_3-Prior bezeichnet.

Alternativ kann die dimensionsreduzierte Zusammenfassung der Datenmatrix y, W, dazu benutzt werden, ein separates Regressionsmodell für jedes einzelne Nutzer-Cluster $k \in \{1, ..., K\}$ zu erstellen.

Hinsichtlich jedes Nutzer-Clusters k lassen sich Item-Cluster identifizieren, die von den Nutzern des k-ten Clusters bevorzugt oder abgelehnt werden, indem man die Elemente der k-ten Zeile von W miteinander vergleicht. Die Item-Cluster können dann entweder mittels der Gemeinsamkeiten der in ihnen enthaltenen Items oder anhand von den Durchschnittwerten für die Eigenschaften der zum betreffenden Cluster gehörenden Items interpretiert werden. In den meisten

Fällen ist es möglich, ein paar der für das l-te Item-Cluster charakteristischen Eigenschaften zu identifizieren, indem man die Mittelwerte

$$\bar{a}_{.\kappa}^{\{l\}} = \frac{\sum_{j=1}^{J} q_{jl} a_{j\kappa}}{\sum_{j=1}^{J} q_{jl}}, \kappa = 1, ..., \kappa_A,$$

mit den Mittelwerten aller übrigen Item-Cluster $\bar{a}_{.\kappa}^{\{\ell\}}, \ell \in \{1, ..., L\} \backslash \{l\}$ vergleicht. Ein im Vergleich zu anderen Clustern hoher Wert von $\bar{a}_{.\kappa}^{\{l\}}$ ist ein Hinweis darauf, daß hohe Ausprägungen der Eigenschaft $\kappa \in \{1, ..., \kappa_A\}$ charakteristisch für die Items aus dem l-ten Cluster sein könnte. Eigenschaften, die gleichermaßen Charakteristika von durch die Personen aus dem betrachteten Nutzer-Cluster hoch wie niedrig bewerteten Item-Clustern sind, können als vernachlässigbar betrachtet werden. Dagegen sollten Eigenschaften, die bei Items besonders stark ausgeprägt sind, die von den Nutzern des betrachteten Nutzer-Clusters allgemein besonders hoch bewertet wurden, als mögliche Regressoren für das Modell hinsichtlich des betrachteten Nutzer-Clusters in Betracht gezogen werden. Auf diese Weise können möglicherweise in Bezug auf jedes Nutzer-Cluster ein paar Eigenschaften identifiziert werden, die als Regressoren geeignet sein könnten. Diese sollten auf ihre Eignung als unabhängige Variablen überprüft werden. Auf diese Weise lassen sich auch ohne eine numerisch aufwendige MCMC Modellbestimmung Hinweise auf einige möglicherweise relevante Variablen erhalten. Insbesondere, wenn die Anzahl der Eigenschaften, Items und Item-Cluster groß ist, kann sich dieses Vorgehen als sinnvoll erweisen.

Zudem kann auch die aus der Cluster-Zugehörigkeit der Items resultierende Information genutzt werden. Hierzu empfiehlt sich die Verwendung der Dummy-Variablen

$$\delta_{jl} = \begin{cases} 1, & \text{falls } j \text{ zum } l\text{-ten Item-Cluster gehört} \\ 0, & \text{sonst} \end{cases}$$

Die Cluster-Zugehörigkeit der Items kann - sofern sie signifikanten Einfluß auf die Bewertungen hat - direkt in den Regressionsansatz $Y_{ij} = X'_{ij}\beta^i + \epsilon_{ij}$ integriert werden, indem die Dummies einfach wie zusätzliche Regressoren verwendet werden. Auf diese Weise ließe sich die Zugehörigkeit eines Items zu einer bestimmten Gruppe von Items berücksichtigen. Die zu einem bestimmten Dummy gehörende Komponente des β^i-Vektors beschreibt dann die Reaktion des i-ten Nutzers auf

Items aus dem betreffenden Cluster. Die Dummies sollten hierbei genau wie alle übrigen Variablen behandelt werden. Nur relevante Dummies sind zu berücksichtigen.

Falls wenigstens für die meisten Nutzer $i \in \{1, ..., I\}$ $\alpha_i > \kappa_A L$ gilt, ist es möglicherweise vorteilhaft, eine andere Strategie zur Integration der Item-Cluster Zugehörigkeit in das hierarchische lineare Modell zu wählen. $\dot{X}_{ij} \in \mathbb{R}^{\kappa_A}$ enthält im Unterschied zu $X_{ij} \in \mathbb{R}^{\kappa_A+1}$ nur die Eigenschaften des von i bewerteten Items j und die durchschnittliche Kritikerbewertung dieses Items. Man definiert

$$\tilde{X}_i = \begin{pmatrix} 1 & \delta_{11}\dot{X}'_{i1} & \cdots & \delta_{1L}\dot{X}'_{i1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \delta_{\alpha_i 1}\dot{X}'_{i\alpha_i} & \cdots & \delta_{\alpha_i L}\dot{X}'_{i\alpha_i} \end{pmatrix} \in \mathbb{R}^{1+L\kappa_A}$$

und außerdem $\tilde{\beta}^i = (1, \tilde{\beta}^{i\{1\}'}, ..., \tilde{\beta}^{i\{L\}'})'$. Diese Definitionen benutzt man in der Regressionsgleichung:

$$Y_i = \tilde{X}_i\tilde{\beta}^i + \epsilon_i, i = 1, ..., I.$$

Die Vektoren $\tilde{\beta}^{i\{l\}} \in \mathbb{R}^{\kappa_A}$ beschreiben die Reaktion des i-ten Nutzers auf alle Regressoren des Modells im Hinblick auf Items aus dem l-ten Item-Cluster. Dieser Ansatz ist dann sinnvoll, wenn für unterschiedliche Gruppen von Filmen unterschiedliche Eigenschaften wichtig sind. So ist „Humor“ bei einer Komödie ein Qualitätsmerkmal, auf das nur wenige verzichten mögen. Dagegen erhoffen sich dieselben Konsumenten in Bezug auf Action- oder Horror-Filme in der Regel weniger zum Lachen gebracht zu werden und dürften hier eher an den Merkmalen „Spannung“ und „Spezialeffekte“ interessiert sein. Was von denselben Personen bei einer Gruppe von Filmen erhofft wird, kann hinsichtlich einer anderen Gruppe von Filmen sogar unerwünscht sein. Dieser Ansatz erlaubt, für jede Gruppe von Items ein anderes Modell zu verwenden.

8.3.3 Hybrides hierarchisches GP-basiertes Verfahren

Durch eine kleine Änderung der Likelihood läßt sich das kollaborative hierarchische GP-basierte Verfahren nach Yu et. al. (2006) zu einem hybriden Verfahren

abwandeln. Statt den bisher verwendeten Funktionen $P(y_i|\mathbf{f}_i(\alpha_i), \Omega_i)$ verwendet man einfach

$$P_{hy}(y_i|\mathbf{b}^i, \Omega_i) = \frac{1}{(2\pi\Omega_i)^{(\kappa_A+1)/2}} \exp\left(-\frac{(y_i - X_i\mathbf{b}^i)'(y_i - X_i\mathbf{b}^i)}{2\Omega_i}\right)$$

Hierbei ist X_i wieder die verwendete Design-Matrix für den i-ten Nutzer und die Zufallsvariable $\mathbf{b}^i \in \mathbb{R}^{\kappa_A+1}, i = 1, ..., I$, entspricht dem Regressionsparameter. Die Vektoren $\mathbf{b}^i \in \mathbb{R}^{\kappa_A+1}, i = 1, ..., I$, haben immer genau eine Komponente mehr als Regressoren im Modell verwendet werden. Man benutzt

$$P_{hy}(\mathbf{b}^i|\mathbf{h}_b, \mathbf{K}_b) = \frac{1}{(2\pi)^{(\kappa_A+1)/2}|\mathbf{K}_b|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{b}^i - \mathbf{h}_b)'\mathbf{K}_b^{-1}(\mathbf{b}^i - \mathbf{h}_b)\right)$$

und

$$P_{hy}(\mathbf{h}_b, \mathbf{K}_b) \propto |\mathbf{K}_b|^{-\tau_b/2} \exp\left(-\frac{\pi_b}{2}(\mathbf{h}_b - \mathbf{h}_{b,0})'\mathbf{K}_b^{-1}(\mathbf{h}_b - \mathbf{h}_{b,0}) - \frac{1}{2}tr((\tau_b\mathbf{K}_{b,0})\mathbf{K}_b^{-1})\right).$$

Zudem verwendet man

$$\hat{P}_{hy}(\mathbf{b}^i) = \frac{1}{(2\pi)^{(\kappa_A+1)/2}|\hat{\mathcal{K}}_{b,i}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{b}^i - \hat{\mathbf{b}}^i)'\hat{\mathcal{K}}_{b,i}^{-1}(\mathbf{b}^i - \hat{\mathbf{b}}^i)\right)$$

anstelle von $\hat{P}(\mathbf{f}_i)$. Hiermit erhält man die Likelihood

$$P_{hy}(y|\mathbf{h}_b, \mathbf{K}_b, \Omega) = \prod_{i=1}^{I} \int P_{hy}(y_i|\tilde{\mathbf{b}}^i, \Omega_i) P_{hy}(\tilde{\mathbf{b}}^i|\mathbf{h}_b, \mathbf{K}_b) d\tilde{\mathbf{b}}_i.$$

Hieraus erhält man unter Verwendung der Jensen'schen Ungleichung wieder die untere Schranke der Log-Likelihood

$$\mathfrak{S}_{u,hy}[\log P_{hy}(y|\mathbf{h}_b, \mathbf{K}_b, \Omega)] = \sum_{i'=1}^{I} \Bigg[\int \hat{P}_{hy}(\tilde{\mathbf{b}}^{i'}) \log P_{hy}(y_{i'}|\tilde{\mathbf{b}}^i, \Omega_{i'}) d\tilde{\mathbf{b}}^{i'}$$

$$+ \int \hat{P}_{hy}(\tilde{\mathbf{b}}^{i'}) \log P_{hy}(\tilde{\mathbf{b}}^{i'}|\mathbf{h}_b, \mathbf{K}_b) d\tilde{\mathbf{b}}^{i'} + \int \hat{P}(\tilde{\mathbf{b}}^{i'}) \log \hat{P}(\tilde{\mathbf{b}}^{i'}) d\tilde{\mathbf{b}}^{i'} \Bigg] \quad .$$

Startwerte: $\mathbf{K}_{b,0} = E_{(\kappa_A+1)\times(\kappa_A+1)}, \mathbf{h}_{b,0} = 0 \in \mathbb{R}^{\kappa_A+1}, \mathbf{h}_b \leftarrow \mathbf{h}_{b,0},$
$\mathbf{K}_b \leftarrow \mathbf{K}_{b,0}, \mathbf{b}^i = 0, \Omega_i, i = 1, ..., I, n = 0, \mathbf{S_b} = \infty$

Solange $\mathbf{S_b} \geq \epsilon$:

$n \leftarrow n + 1$

$\hat{\mathbf{b}}^i = \left(\mathbf{K}_b^{-1} + \frac{X_i' X_i}{\Omega_i}\right)^{-1} \left(\frac{X_i' y_i}{\Omega_i} + \mathbf{K}_b^{-1}\mathbf{h}_b\right), i = 1, ..., I$

$\hat{\mathcal{K}}_{b,i} = \left(\mathbf{K}_b^{-1} + \frac{X_i' X_i}{\Omega_i}\right)^{-1}, i = 1, ..., I$

$\hat{\Omega}_i = \frac{1}{\alpha_i}\left[(y_i - X_i\hat{\mathbf{b}}^i)'(y_i - X_i\hat{\mathbf{b}}^i) + tr(X_i' X_i \hat{\mathcal{K}}_{b,i})\right], i = 1, ..., I$

$\hat{\mathbf{h}}_b = \frac{1}{I+\pi_b}\left(\sum_{i=1}^I \hat{\mathbf{b}}^i + \pi_b \mathbf{h}_{b,0}\right)$

$\hat{\mathbf{K}}_b = \frac{1}{I+\tau_b}\Big(\pi_b(\hat{\mathbf{h}}_b - \mathbf{h}_{b,0})(\hat{\mathbf{h}}_b - \mathbf{h}_{b,0})' + \tau_b \mathbf{K}_{b,0}$
$+ \sum_{i=1}^I \left[(\hat{\mathbf{b}}^i - \hat{\mathbf{h}}_b)(\hat{\mathbf{b}}^i - \hat{\mathbf{h}}_b)' + \hat{\mathcal{K}}_{b,i}\right]\Big)$

$\mathbf{S_b} = \sum_{i=1}^I (\hat{\mathbf{b}}^i - \mathbf{b}^i)'(\hat{\mathbf{b}}^i - \mathbf{b}^i)$

$\mathbf{h}_b \leftarrow \hat{\mathbf{h}}_b, \mathbf{K}_b \leftarrow \hat{\mathbf{K}}_b, \mathbf{b}^i \leftarrow \hat{\mathbf{b}}^i, \Omega_i \leftarrow \hat{\Omega}_i, i = 1, ..., I$

Abbildung 8.11: Schematische Darstellung des hybriden hierarchischen GP-Verfahrens

Durch Maximieren der unteren Schranke

$$\mathfrak{S}_{u,hy}(\log P_{hy}(y|\mathbf{h}_b, \mathbf{K}_b, \Omega)) + \log P_{hy}(\mathbf{h}_b, \mathbf{K}_b)$$

der Posterior erhält man

$$\hat{\mathbf{h}}_b = \frac{1}{I+\pi_b}\left(\sum_{i=1}^I \hat{\mathbf{b}}^i + \pi_b \mathbf{h}_{b,0}\right)$$

$$\hat{\mathbf{K}}_b = \frac{1}{I+\tau_b}\Big(\pi_b(\mathbf{h}_b - \mathbf{h}_{b,0})(\mathbf{h}_b - \mathbf{h}_{b,0})' + \tau_b \mathbf{K}_{b,0} + \sum_{i=1}^I \left[(\hat{\mathbf{b}}^i - \mathbf{h}_b)(\hat{\mathbf{b}}^i - \mathbf{h}_b)' + \hat{\mathcal{K}}_{b,i}\right]\Big)$$

$$\hat{\Omega}_i = \frac{1}{\alpha_i}\left[(y_i - X_i\hat{\mathbf{b}}^i)'(y_i - X_i\hat{\mathbf{b}}^i) + tr(X_i' X_i \hat{\mathcal{K}}_{b,i})\right] \quad .$$

Die Herleitung dieser Größen verläuft völlig analog zu der Herleitung der entsprechenden Größen im Rahmen des kollaborativen GP-basierten Ansatzes nach

Yu et. al. (2006).

Für den Posterior-Erwartungswert $\hat{\mathbf{b}}^i$ für $\mathbf{b}^i$ und die zugehörige Posterior-Kovarianz $\hat{\mathcal{K}}_{b,i}, i = 1, ..., I$, erhält am die bekannten Formeln (z.B. O'Hagan (1978))

$$\hat{\mathbf{b}}^i = \left(\mathbf{K}_b^{-1} + \frac{X_i' X_i}{\Omega_i}\right)^{-1} \left(\frac{X_i' y_i}{\Omega_i} + \mathbf{K}_b^{-1}\mathbf{h}_b\right)$$
$$\hat{\mathcal{K}}^i = \left(\mathbf{K}_b^{-1} + \frac{X_i' X_i}{\Omega_i}\right)^{-1} \quad .$$

Mit diesen Formeln ergibt sich der in Abbildung 8.11 dargestellte Algorithmus. Auch dieses Verfahren kann sowohl auf die transformierten Daten Y^t als auch auf untransformierte Daten angewandt werden.

Beispiel 8.8:

Das hybride GP-Verfahren wird auf die transformierten Daten Y^t aus Beispiel 8.5 angewendet. Zusätzlich werden die Eigenschafts-Daten aus Beispiel 4.1 verwendet. Der Algorithmus verläuft analog zu Beispiel 8.5. Als Ergebnis erhält man

$$\mathbf{h}_b = \begin{pmatrix} 0{,}04 \\ -0{,}13 \\ 0{,}01 \\ -0{,}01 \\ 0{,}01 \\ -0{,}08 \\ 0{,}15 \end{pmatrix} \quad \text{und} \quad \mathbf{b}^1 = \begin{pmatrix} 0{,}05 \\ -0{,}04 \\ -0{,}20 \\ -0{,}04 \\ -0{,}05 \\ -0{,}06 \\ 0{,}25 \end{pmatrix}$$

Mit $X_j' = (1 a_j')$ und $\bar{y}_{1.} = 3,50$ ergibt sich

$$\begin{aligned} \hat{Y}_{17} &= X_7' \mathbf{b}^1 + \bar{y}_{1.} \\ &= 0,05 - 0,04 \cdot 5 - 0,20 \cdot 4 - 0,04 \cdot 2 - 0,05 \cdot 4 - 0,06 \cdot 3 + 0,25 \cdot 8 + 3,5 \\ &= 4,09. \end{aligned}$$

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,648	0,664	0,677	0,689	0,714	0,752	0,778	0,807	0,834
Prec.	0,687	0,697	0,720	0,721	0,708	0,658	0,567	0,463	0,383
Rec.	0,227	0,214	0,178	0,143	0,076	0,030	0,025	0,032	0,049
R_B	84,41	77,34	74,28	71,37	70,12	68,44	67,36	65,51	64,09

Tabelle 8.4: *AAD*, *Prec.* und *Rec.* für das hierarchische GP-Verfahren nach Yu et. al. (2006) bei unterschiedlichen Testdatenanteilen

8.4 Empirische Ergebnisse

In diesem Abschnitt werden die empirischen Resultate der behandelten Bayes'schen Verfahren dargestellt. Hierbei werden die Bayes'schen Verfahren sowohl untereinander als auch mit den Verfahren aus den Kapiteln 5 und 6 verglichen. Da beide hybriden Bayes'schen Verfahren die Auswahl bestimmter Regressoren erfordern, wird vor der Diskussion der Resultate beider hybrider Verfahren die Auswahl der Regressoren behandelt.

8.4.1 Resultate des kollaborativen GP-Ansatzes

In Bezug auf die bereits in den vorherigen empirischen Untersuchungen verwendeten MovieLens-Teildatensatz, welcher Bewertungen von 1067 Nutzern in Bezug auf 418 Filme umfaßt, ergeben sich für das hierarchische GP-Verfahren nach Yu et. al. (2006) die in Tabelle 8.4 aufgelisteten Ergebnisse.

Das hierarchische GP-Verfahren nach Yu et. al. (2006) liefert bei geringen Anteilen des Testdatensatzes an der insgesamt verfügbaren Datenmenge Ergebnisse, die den Ergebnissen der Nicht-Bayes'schen kollaborativen Verfahren in Bezug auf die Genauigkeit (*AAD*) überlegen sind. Allerdings kann in Bezug auf das Verfahren zur ordinalen Matrixfaktorisation nach Rennie, Srebro (2005) und das ordinale zweimodale Clusterverfahren hierbei von keiner starken Überlegenheit gesprochen werden. Erst bei etwas höheren Testdatenanteilen an der gesamten Dantenmenge weist das kollaborative Verfahren nach Yu et. al. (2006) im Vergleich zu den betrachteten Nicht-Bayes'schen Verfahren hinsichtlich des kleineren Datensatzes (1067 Nutzer, 418 Items) überwiegend größere Breese-Werte auf.

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,660	0,664	0,672	0,6832	0,696	0,709	0,731	0,767	0,806
Prec.	0,712	0,705	0,708	0,690	0,678	0,678	0,669	0,621	0,552
Rec.	0,233	0,228	0,213	0,195	0,180	0,148	0,109	0,072	0,071
R_B	77,07	73,06	70,90	68,77	66,89	65,20	63,38	62,12	60,72

Tabelle 8.5: *AAD*, *Prec.* und *Rec.* für das hierarchische GP-Verfahren nach Yu et. al. (2006) bei unterschiedlichen Testdatenanteilen im Hinblick auf den großen Datensatz (2000 Nutzer, 3043 Items)

Hinsichtlich des größeren Datensatzes, der die Bewertungen von 2000 Nutzer in Bezug auf 3043 Filme enthält, kommt es dagegen nur bei geringen Anteilen des Testdatensatzes an der Gesamtdatenmenge zu Ergebnissen, die jene der Nicht-Bayes'schen kollaborativen Verfahren merklich übertreffen. Dies legt die Vermutung nahe, daß das kollaborative Verfahren nach Yu et. al. (2006) zu schlechteren Ergebnissen führen könnte, wenn die Anzahl der Items im Vergleich zur Anzahl der durchschnittlich pro Nutzer verfügbaren Bewertungen sehr groß ist. Da in der Praxis davon auszugehen ist, daß die Anzahl der Items wesentlich höher ausfallen dürfte, könnte dies die praktische Relevanz des kollaborativen Verfahrens nach Yu et. al. (2006) erheblich verringern. Im Hinblick auf die verzerrten Datensätze liefert das kollaborative GP-Verfahren bei allen Verzerrungsgraden höhere Breese-Werte als alle Nicht-Bayes'schen kollaborativen Verfahren (vgl. Anhang B.3). Erwähnenswert ist auch, daß das kollaborative GP-Verfahren nach Yu et. al. (2006) bei geringen Verzerrungsgraden sogar zu besseren Ergebnissen führt als das kollaborative Verfahren nach Pazzani (1999). Diese Ergebnisse basieren jedoch auf dem Datensatz mit 418 Items. Da der Nutzen der ersten Empfehlungen meist ausschlaggebend die weitere Verwendung eines Empfehlungssystems ist, ist es in praxi besonders wichtig, daß ein solches System in der Lage ist, insbesondere auch auf Basis weniger Bewertungen sinnvolle Empfehlungen zu generieren. Daher ist kollaborative GP-Verfahren im Vergleich zu den behandelten Nicht-Bayes'schen kollaborativen Verfahren auf jeden Fall in den Anwendungen von erhöhter praktischer Bedeutung, wenn die Anzahl der betrachteten Items eher gering ist. Fraglich ist, ob das kollaborative GP-Verfahren nach Yu et. al. (2006) auch in den wesentlich häufigeren Fällen, wenn die Anzahl der Items sehr groß

κ	Eigenschaft
1	Grad der Freizügigkeit der Darstellung von Sex („sex“)
2	Grad der dargestellten Gewalt („violence“)
3	Grad der Familienfreundlichkeit („family appeal“)
4	Grad der Mainstream-Tendenzen („hollywood style“)
5	Actionfilmcharakter („action“)
6	Grad an Komik („humor“)
7	Romanzencharakter („romance“)
8	Grad der erzeugten Spannung („suspense“)
9	Dramencharakter („drama depth“)
10	Grad der dargestellten Charakterentwicklung („character development“)
11	Grad an Exzentrizität („offbeat energy“)
12	Grad der cinematographischen Perfektion („cinematography“)
13	Grad der allgemeinen Beliebtheit der Filmmusik („soundtrack“)
14	Grad der Verwendung von Spezialeffekten („special effects“)

Abbildung 8.12: Eigenschaften der MovieLens-Filme (*http* : *//www.reel.com* entnommen)

ist, zu guten Ergebnissen führt.

8.4.2 Auswahl der Regressoren

In Abschnitt 8.1.4 wurden verschiedene Ansätze zur Auswahl von Regressoren vorgestellt und kurz diskutiert. Es existiert bis heute kein allgemein anerkanntes oder unproblematisches Verfahren zur Regressorenselektion hinsichtlich hierarchischer Modelle.

Eine Möglichkeit, eine grobe Vorauswahl von möglicherweise geeigneten Variablen zu treffen, ist die bereits in Abschnitt 8.1.4 beschriebene Kombination der MCMC Modellbestimmung mit Methoden der Versuchsplanung.

Bereits in Beispiel 6.1 wurde erwähnt, daß zu 418 der im MovieLens-Datensatz vorkommenden Filme jeweils die Ausprägungen 14 verschiedener Eigenschaften dem Informationsportal *http* : *//www.reel.com* entnommen wurden. Die Ausprägungen dieser Eigenschaften wurden auf einer Skala von 0 bis 10 angegeben. Eine Auflistung der verschiedenen Eigenschaften findet sich in Abbildung 8.12.

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
M_0	227948	202342	177704	152869	126899	102591	76320	51818	33176
M_1	222192	197987	173983	149352	124448	99971	74916	50237	32006
M_2	226294	201113	176669	152136	126454	101599	76507	51557	32607

Tabelle 8.6: DIC der Modelle M_0, M_1 und M_3 für den Trainingsdatensatz bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz

Außerdem kommt die durchschnittliche Bewertung des Films (j) durch professionelle Kritiker $\bar{Y}_{.j}^{C}$ als Regressor in Frage. Der Variablen $\bar{Y}_{.j}^{C}$ kommt eine Sonderstellung zu. Eliashberg, Shugan (1997) argumentieren, daß sich Kritikerbewertungen als Vorhersagen für die Bewertungen durch die Zielgruppen der jeweiligen Kritiker eignen: „For example, Cosmopolitan magazine might hire a critic. This critic, over time, might learn to write reviews corresponding to the taste of Cosmopolitan magazine subscribers. If so, the critic survives, becoming an asset for the magazine. If not, the critic becomes less popular, less widely read, and eventually replaced. Critics who learn may change their opinions after a change in employment. ... Eventually, surviving critics are actually perfectly representative of the average Cosmopolitan reader. Motion pictures liked by Cosmopolitan magazine critics are also liked by Cosmopolitan magazine readers.“. Dieser Auffassung zufolge kann die durchschnittliche Kritikerbewertung $\bar{Y}_{.j}^{C}$ als allgemein wahrgenommene Qualität eines Films oder Items interpretiert werden. Daher wäre $\bar{Y}_{.j}^{C}$ nach Eliashbergs, Shugans (1997) These ein wichtiger Regressor, da er geeignet ist, zwischen einem „guten“ und einem „schlechten“ Film zu unterscheiden, selbst wenn beide Filme ansonsten die gleichen Merkmale aufweisen. Sei $M_1 = M(1, ..., 14, \bar{Y}^C)$ das Modell, das alle Eigenschaften $\kappa = 1, ..., 14$ und die durchschnittlichen Kritikerbewertungen $\bar{Y}^C$ als Regressoren enthält. $M_0 = M(\bar{Y}^C)$ sei das Modell, in dem die durchschnittlichen Kritikerbewertungen als alleiniger Regressor verwendet werden. Ein empirischer Vergleich der hierarchischen linearen Regressionsverfahren für $M_1 = M(1, ..., 14, \bar{Y}^C)$ und $M_0 = M(\bar{Y}^C)$ führt zu der Erkenntnis, daß zwar die zusätzliche Berücksichtigung aller Eigenschaften gegenüber der alleinigen Berücksichtigung der durchschnittlichen Kritikerbewertung im Trainingsdatensatz zu einer moderaten Verkleinerung des DIC führen kann, aber dennoch konstatiert werden muß, daß $M_0 = M(\bar{Y}^C)$ im Vergleich

zu $M_1 = M(1, ..., 14, \bar{Y}^C)$ erstaunlich gut ist (vgl. Tabellen 8.6). Beim Übergang vom Modell $M_2 = M(1, ..., 14)$ zum Modell $M_1 = M(1, ..., 14, \bar{Y}^C)$ durch die zusätzliche Aufnahme von $\bar{Y}^C$ ins Modell wird das DIC deutlich verkleinert (siehe Tabelle 8.6). Dies belegt empirisch, daß $\bar{Y}^C$ ein wichtiger Regressor ist.

Es sei an dieser Stelle darauf hingewiesen, daß das DIC nicht für die Fälle definiert ist, in denen sowieso ein Testdatensatz zur Validierung des Modells zur Verfügung steht (vgl. Vehtari (2001)). In diesen Fällen ist es ausreichend, herkömmliche Gütemaße für die reine Anpassung des Modells an die Testdaten zu verwenden. Zusätzliche Strafkosten für Modellkomplexität sind in diesen Fällen überflüssig. Die Validierung des Modells kann dann allein durch auf Basis der Testdaten berechneten reinen Anpassungsmaßen wie AAD oder R^2 erfolgen.

Die Unterteilung in Test- und Trainingsdatensatz im Rahmen dieser Arbeit bezieht sich allein auf die ex post Validierung der aufgrund eines Verfahrens berechneten Schätzer. Das Ziel ist, allein auf Basis des Trainingsdatensatzes Vorhersagen mittels verschiedener Verfahren zu machen, um diese Vorhersagen dann anhand der Testdaten zu überprüfen. Es wäre daher inkonsistent, für alle Verfahren, die eine Auswahl von Variablen erfordern, für diese Auswahl schon den Testdatensatz zu benutzen. Entweder muß der Trainingsdatensatz für diese Modelle weiter unterteilt werden, oder man muß sich geeigneter Gütemaße wie des DIC bedienen, die neben der reinen Anpassung an die Daten auch die Modellkomplexität berücksichtigen. Dessen unbeschadet können die Ergebnisse des Testdatensatzes auch dazu benutzt werden, die Modellwahlstrategien selbst zu validieren.

Empirische Hinweise auf weitere relevante Regressoren können mit der in Abschnitt 8.1.4 dargestellten Kombination aus MCMC Modellbestimmung und Methoden der Versuchsplanung gewonnen werden. Da bereits bekannt ist, daß die durchschnittliche Kritikerbewertung ein relevanter Regressor ist, sollte jedes Probemodell die durchschnittliche Kritikerbewertung als Regressor enthalten. Hinsichtlich der $a^V = 14$ Eigenschaften wird ein balancierter unvollständiger Blockplan für $n_a^V = 4$ Versuchelemente in einem Block erstellt. Sei b^V die Anzahl der Blöcke, r^V die Anzahl der Wiederholungen der einzelnen Behandlungsausprägungen im Gesamtversuch und sei λ^V die Anzahl des Auftretens der einzelnen Behandlungsausprägungspaare in der Versuchsanlage. Dann müssen die Bedingungen

$$r^V \cdot a^V = n_a^V \cdot b^V \quad \text{und} \quad \lambda^V = \frac{r^V(n_a^V - 1)}{a^V - 1}$$

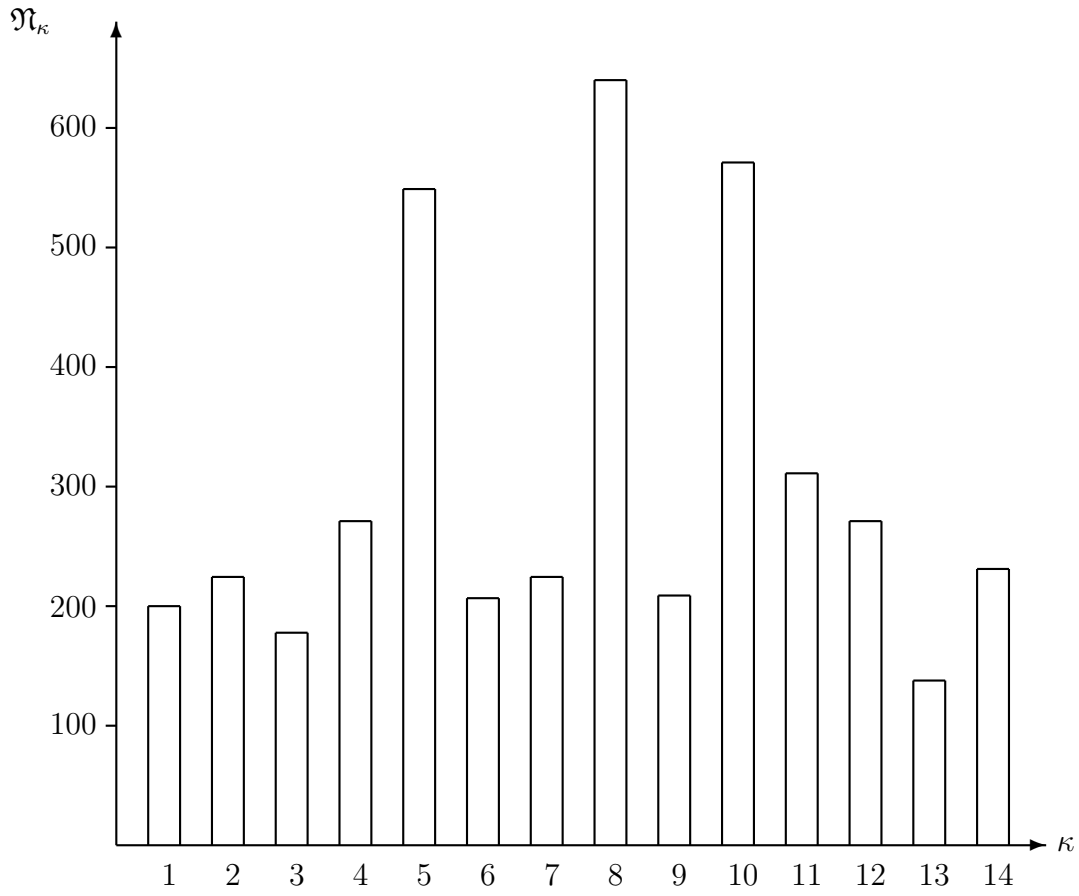

Abbildung 8.13: Häufigkeit $\mathfrak{N}_\kappa$, mit der die jeweilige Eigenschaft κ im Modell mit der höchsten Posterior-Wahrscheinlichkeit für einen einzelnen Nutzer enthalten ist, $\kappa = 1, ..., 14$

erfüllt sein (Rasch, Herrendörfer (1982)). Aus diesen Bedingungen folgt, daß für $a^V = 14$ und $n_a^V = 4$ der balancierte unvollständige Blockplan mit den übrigen Parametern $\lambda^V = 6, r^V = 26$ und $b^V = 91$ die kleinstmögliche balancierte unvollständige Versuchsanlage ist. Diese wurde mit der CADEMO-Software (Rasch, Kubinger (2006)) erstellt und ist in Anhang C wiedergegeben.

Für jedes der 91 in dem Blockplan enthaltenen Modelle wurde die Menge aller Modelle gebildet, die durch Hinzufügen oder Weglassen einer Variable aus dem betreffenden Modell hervorgehen können. Die Gesamtmenge dieser Modelle wurde als Ausgangspunkt von 1067 separaten MCMC Modellbestimmungen (Hoeting et. al. (1996), Raftery et. al. (1997)) verwendet. So wurde für jeden der

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
M_3	225701	200573	175231	150652	125520	101163	75380	49795	33900
M_4	224052	199655	174602	150231	125736	100604	75843	50496	33812
M_5	224111	199346	174614	149968	125441	100581	75779	50678	32678
M_6	226008	200713	175827	151658	125947	101377	76211	50620	34547
M_7	223510	198852	174032	149613	124947	100496	75450	50616	32465
M_8	223591	199045	174438	149596	124991	100584	75666	50694	32501

Tabelle 8.7: DIC auf Basis der Trainingsdaten für die Modelle M_3, M_4, M_5, M_6, M_7 und M_8 für verschiedene Anteile des Testdatensatzes an den insgesamt vorhandenen Daten

1067 Nutzer getrennt das Modell mit der höchsten Posterior-Wahrscheinlichkeit identifiziert. Da 1067 MCMC Modellbestimmungen für eine derart große Menge an Kandidatenmodellen mit erheblichem numerischem Aufwand verbunden ist, konnte nicht für jeden Trainingsdatensatz getrennt eine MCMC Modellbestimmung durchgeführt werden. Um Inkonsistenzen zu vermeiden wurde die MCMC Modellbestimmung nur auf Basis des kleinsten verwendeten Trainingsdatensatzes vorgenommen. Zudem wurde darauf geachtet, daß dieser Trainingsdatensatz in allen übrigen Trainingsdatensätzen enthalten ist. Abbildung 8.13 zeigt die aus den 1067 MCMC Modellbestimmungen resultierenden unterschiedlichen empirischen Häufigkeiten $\mathfrak{N}_\kappa$, mit der eine bestimmte Eigenschaft κ im Modell mit der höchsten Posterior-Wahrscheinlichkeit für einen der Nutzer enthalten ist. Als mit Abstand wichtigste Eigenschaften erweisen sich der Actionfilmcharakter ($\kappa = 5$), der Grad der erzeugten Spannung ($\kappa = 8$) und der Grad der dargestellten Charakterentwicklung ($\kappa = 10$). Weitere möglicherweise wichtige Variablen sind der Grad der Mainstream-Tendenz ($j = 4$), der Grad der Exzentrizität ($\kappa = 11$) und der Grad der cinematographischen Perfektion ($\kappa = 12$).
Dagegen, daß alle Eigenschaften allgemein relevante Regressoren sein könnten, spricht, daß

$$\sum_{\kappa=1}^{\kappa_A} \mathfrak{N}_\kappa < 4 \cdot I$$

gilt. Da jedes der 91 Modelle 4 Regressoren umfaßt und außerdem auch alle

Modelle, die durch Hinzufügen einer zusätzlichen Variable aus den 91 Modellen hervorgehen, zur Grundlage der MCMC Modellbestimmung wurden, würde die allgemeine Relevanz aller Regressoren eher dafür sprechen, daß die Modelle mit einer zusätzlichen Variable eine höhere Posterior-Wahrscheinlichkeit aufweisen düften. Das würde aber wiederum implizieren, daß die obige Summe größer als $4 \cdot I$ sein sollte. Daher ist die Annahme, daß alle Regressoren relevant sind, nicht plausibel. Folglich ist davon auszugehen, daß einer oder mehrere der im Rahmen des M_1-Modells verwendeten Regressoren allenfalls in Bezug auf eine Minderheit von Nutzern relevant sind.

Als Alternativen zur MCMC Modellbestimmung wurden bereits die Maurer- und die Steinmetz-Methode genannt. Beide Verfahren lassen sich automatisieren. Im Rahmen der Mauerer Methode werden in jedem Schritt ausgehend von einer Regressorenmenge alle Mengen, die aus der jeweiligen Regressorenmenge und einer nicht in ihr enthaltenen Variable bildbar sind, gebildet. Alle diese Kandidatenmengen werden als Regressionsmodell verwendet und die zugehörige hierarchischen Modelle werden berechnet. Für jedes dieser hierarchischen Modelle berechnet man dann das DIC. Die Kandidatenmenge mit dem kleinsten DIC wird dann im darauffolgenden Schritt zur Regressorenmenge. Nach drei Schritten der Maurer-Methode erhält man ausgehend von einer Regressorenmenge, die nur aus der durchschnittlichen Kritikerbewertung besteht, ebenfalls den Actionfilmcharakter ($\kappa = 5$), den Grad der erzeugten Spannung ($\kappa = 8$) und den Grad der dargestellten Charakterentwicklung ($\kappa = 10$) als erfolgversprechendste Regressoren. Hieraus resultiert das Modell $M_3 = M(5, 8, 10, \bar{Y}^C)$. Weitere möglicherweise geeignete Variablen sind der Grad der Mainstream-Tendenz ($j = 4$), der Grad der Exzentrizität ($\kappa = 11$) und der Grad der cinematographischen Perfektion ($\kappa = 12$). Ausgehend von dem Modell $M_3 = M(5, 8, 10, \bar{Y}^C)$ und der auf den Resultaten der MCMC Modellbestimmungen basierenden Vermutung, daß auch der Grad der Mainstream-Tendenz ($j = 4$), der Grad der Exzentrizität ($\kappa = 11$) und der Grad der cinematographischen Perfektion ($\kappa = 12$) geeignete Regressoren sein könnten, wurden eine Reihe verschiedener Modelle gebildet. Ein Modell $M_\mu = M(x, z, ...)$ verwenden die Eigenschaften $x, z, ...$ als Regressoren, die nach dem großem M innerhalb der runden Klammern stehen. Für die Modelle

$$\begin{array}{lll} M_3 = M(5, 8, 10, \bar{Y}^C) & & M_6 = M(5, 8, 10, 12, \bar{Y}^C) \\ M_4 = M(5, 8, 10, 11, \bar{Y}^C) & \text{und} & M_7 = M(4, 5, 8, 10, 11, 12, \bar{Y}^C) \\ M_5 = M(4, 5, 8, 10, \bar{Y}^C) & & M_8 = M(4, 5, 8, 10, 11, \bar{Y}^C) \end{array}$$

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
M_0	0,299	0,304	0,301	0,303	0,311	0,304	0,320	0,319	0,314
M_1	0,414	0,419	0,419	0,421	0,425	0,427	0,428	0,433	0,434
M_2	0,356	0,360	0,364	0,364	0,374	0,377	0,384	0,396	0,407
M_3	0,336	0,337	0,338	0,338	0,344	0,338	0,362	0,370	0,377
M_4	0,355	0,359	0,362	0,363	0,363	0,368	0,376	0,379	0,385
M_5	0,348	0,350	0,356	0,352	0,355	0,357	0,365	0,370	0,394
M_6	0,334	0,335	0,336	0,336	0,342	0,343	0,356	0,368	0,381
M_7	0,358	0,362	0,365	0,370	0,368	0,370	0,380	0,404	0,393
M_8	0,355	0,357	0,359	0,364	0,370	0,373	0,371	0,383	0,391

Tabelle 8.8: Bestimmtheitsmaß R^2 der Modelle $M_0, M_1, M_2, M_3, M_4, M_5, M_6, M_7$ und M_8 für den Trainingsdatensatz bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz

wurden auf Basis der Trainingsdatensätze die DIC-Werte berechnet.

Da das DIC tendenziell komplexere Modelle trotz Strafkostenterm bevorzugt (van der Linde (2005)), empfiehlt es sich, nicht immer das Modell mit dem kleinsten DIC zu wählen, sondern vielmehr mehrere der Modelle mit verhältnismäßig kleinem DIC in Betracht zu ziehen (Spiegelhalter et. al. (2002)). Es ist außerem ratsam, neben dem DIC auch andere Gütemaße und Kriterien zu benutzen.

Das Modell mit den meisten Regressoren, M_1, weist den kleinsten DIC-Wert auf (siehe Tabellen 8.6 und 8.7). Für die beiden deutlich einfacheren Modelle M_7 und M_8 ergeben sich nur verhältnismäßig geringe Ergebnisverschlechterungen gegenüber M_1. Da mit diesen minimalen Erhöhungen des DIC-Werts erhebliche Vereinfachungen des zugrundeliegenden Modells einhergehen, sollte man M_7 und M_8 dem Basismodell M_1 vorziehen. Auch die Erhöhung des Bestimmtheitsmaßes (Tabelle 5.9) durch die Hinzunahme von 8 (M_7) bzw. 9 (M_8) zusätzlichen Regressoren spricht gegen Modell M_1. Da sich zudem die beiden Modelle M_7 und M_8 bezüglich ihrer $DIC-$, R^2- und $AAD-$Werte kaum unterscheiden, sollte man sich auch hier für das einfachere Modell von beiden (M_8) entscheiden. Vergleicht man M_8 und M_3 anhand ihrer R^2- und AAD-Werte miteinander, so fällt auf, daß mit der Verwendung der zwei zusätzlichen exogenen Größen kein beson-

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
M_0	0,738	0,735	0,737	0,738	0,730	0,736	0,724	0,730	0,729
M_1	0,689	0,688	0,690	0,687	0,682	0,680	0,670	0,664	0,653
M_2	0,709	0,706	0,706	0,707	0,699	0,698	0,691	0,688	0,673
M_3	0,721	0,720	0,718	0,718	0,716	0,721	0,712	0,695	0,689
M_4	0,711	0,710	0,705	0,707	0,709	0,701	0,701	0,689	0,697
M_5	0,711	0,710	0,706	0,708	0,706	0,705	0,701	0,694	0,690
M_6	0,708	0,719	0,717	0,721	0,715	0,716	0,709	0,696	0,690
M_7	0,705	0,703	0,701	0,700	0,699	0,698	0,694	0,684	0,683
M_8	0,707	0,706	0,705	0,702	0,700	0,699	0,697	0,691	0,685

Tabelle 8.9: *AAD* der Modelle $M_0, M_1, M_2, M_3, M_4, M_5, M_6, M_7$ und M_8 im Hinblick auf den Trainingsdatensatz bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz

ders großer Genauigkeitsgewinn verbunden ist. Dies würde für die Wahl von M_3 sprechen.

Die Anpassung aller Modelle an den Trainingsdatensatz ist eher schwach. Dies spricht dafür, daß noch wesentliche Einflußgrößen fehlen. Hierbei muß es sich nicht um meßbare Größen handeln. Vor dem Hintergrund, daß die Anzahl der Regressoren zwischen vier und fünfzehn variiert, ist das Anpassungsverhalten der Modelle recht ähnlich. Das würde dafür sprechen, ein möglichst einfaches Modells zu verwenden.

8.4.3 Empirischer Vergleich der hybriden Verfahren

Ziel dieses Abschnitts ist ein empirischer Vergleich der hier vorgestellten hybriden Verfahren. Hierbei handelt es sich um das hierarchische lineare Regressionsmodell (HBLR-Modell) nach Rossi et. al. (1996) und das vorgestellte hybride GP-Modell in Anlehnung an das kollaborative Verfahren nach Yu et. al. (2006).

Zur Berechnung des HBLR-Modells nach Rossi et. al. (1996) wurden $\nu = 3, V_0 = diag(4, 0, 01, \cdots, 0, 01, 0, 1) \in \mathbb{R}^{\kappa_A+1,\kappa_A+1}, \nu_0 = 3 + \kappa_A$, $A = 0,01$ und $V_i = \widehat{var}(Y_i)$ verwendet, wobei $\widehat{var}(Y_i)$ die Stichprobenvarianz der Bewertungen des i-ten Nutzers bezeichnet. Als Wert für $\bar{\Delta} \in \mathbb{R}^{\kappa_A+1}$ wurde der Nullvektor

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
M_0	0,752	0,761	0,754	0,754	0,759	0,759	0,765	0,769	0,797
M_1	0,734	0,730	0,726	0,732	0,738	0,741	0,747	0,756	0,789
M_2	0,757	0,757	0,750	0,750	0,761	0,761	0,767	0,773	0,787
M_3	0,741	0,744	0,746	0,746	0,749	0,744	0,755	0,760	0,779
M_4	0,731	0,735	0,743	0,739	0,740	0,748	0,752	0,757	0,786
M_5	0,740	0,739	0,743	0,741	0,744	0,747	0,750	0,759	0,781
M_6	0,738	0,744	0,746	0,738	0,748	0,748	0,752	0,761	0,780
M_7	0,729	0,738	0,741	0,738	0,742	0,744	0,749	0,757	0,775
M_8	0,732	0,734	0,734	0,738	0,742	0,744	0,749	0,756	0,780

Tabelle 8.10: *AAD* des hierarchischen linearen Regressionsmodells für $M_0, M_1, M_2, M_3, M_4, M_5, M_6, M_7$ und M_8 hinsichtlich des Testdatensatzes bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz

gewählt. Durch die Wahl unterschiedlicher Diagonlelemente für V_0 ist es möglich, die Regressoren unterschiedlich zu gewichten. Je kleiner das κ-te Diagonalelement gewählt wird, umso wahrscheinlicher ist es, daß das entsprechende Diagonalelement in der Matrix V_β klein ausfällt. Hierdurch es wahrscheinlicher, daß auch das κ-te Diagonalelement der Varianz der Normalverteilung, aus der die Werte $\beta^i, i = 1, ..., I$, gezogen werden, klein ist. Hierdurch werden sich Abweichungen von der κ-ten Komponente des Prior-Erwartungswerts von $\beta^i, \Delta' z_i$, im Verlauf des HBLR-Algorithmus nur dann durchsetzen können, wenn die Daten y_i sehr deutlich für diese Abweichungen sprechen. Da Δ hier aus einer Normalverteilung um den Nullvektor $\bar{\Delta}$ gezogen wird, bewirkt die gewählte Parametrisierung, daß die κ-te Komponente jedes β^i nach Eintreten der Konvergenz des MCMC-Verfahrens nur dann erheblich von Null abweichen dürfte, wenn die Daten Y_i sehr deutlich für eine solche Abweichung sprechen. Indem man weniger wichtigen Regressoren kleinere Diagonalelemente in V_0 zuordnet, ist es möglich ihren Beitrag zur Schätzung zu vermindern ohne ganz auf sie zu verzichten. In allen Modellen wurde 0,1 als zur durchschnittlichen Kritikerbewertung $\bar{Y}^C$ gehörendes Diagonalelement gewählt. Der Wert 4 entspricht dem Achsenabschnitt. Für alle übrigen Regressoren werden die Diagonalelemente 0,01 verwendet. Hierdurch werden die Effekte aller Variablen gegenüber der durchschnittlichen Kritikerbewertung ab-

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
M_0	0,651	0,675	0,620	0,602	0,642	0,656	0,615	0,585	0,503
	0,048	0,042	0,037	0,041	0,042	0,038	0,031	0,44	0,035
M_1	0,612	0,694	0,639	0,647	0,642	0,663	0,643	0,605	0,509
	0,092	0,105	0,083	0,091	0,090	0,086	0,072	0,077	0,065
M_2	0,603	0,656	0,613	0,606	0,610	0,624	0,613	0,580	0,487
	0,070	0,075	0,059	0,063	0,062	0,061	0,055	0,053	0,048
M_3	0,682	0,674	0,652	0,628	0,656	0,659	0,650	0,614	0,538
	0,079	0,083	0,061	0,065	0,062	0,054	0,049	0,061	0,055
M_4	0,613	0,704	0,664	0,654	0,641	0,652	0,663	0,617	0,518
	0,079	0,085	0,071	0,079	0,070	0,073	0,060	0,064	0,055
M_5	0,605	0,701	0,651	0,643	0,643	0,648	0,638	0,607	0,526
	0,076	0,083	0,071	0,076	0,072	0,068	0,061	0,067	0,056
M_6	0,624	0,683	0,657	0,630	0,650	0,633	0,645	0,611	0,504
	0,063	0,071	0,061	0,065	0,064	0,055	0,051	0,060	0,054
M_7	0,638	0,697	0,659	0,660	0,651	0,663	0,644	0,613	0,512
	0,084	0,090	0,076	0,085	0,079	0,077	0,063	0,069	0,058
M_8	0,620	0,697	0,650	0,658	0,643	0,662	0,646	0,613	0,519
	0,084	0,089	0,075	0,084	0,077	0,077	0,064	0,068	0,061

Tabelle 8.11: Präzision (oben) und Recall (unten) des hierarchischen linearen Regressionsmodells (HBLR-Modell) für $M_0, M_1, M_2, M_3, M_4, M_5, M_6, M_7$ und M_8 bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz (verwendeter Schwellenwert: 4,5)

geschwächt.

Tabelle 8.10 gibt die *AAD*-Werte im Hinblick auf den Testdatensatz hinsichtlich des HBLR-Modells wieder. Bis auf die Modelle M_0 und M_2 führen alle vorgestellten Variablenkombinationen zu guten und bemerkenswerterweise recht ähnlichen Ergebnissen. Solange ein Modell wenigstens die wichtigsten Variablen als Regressoren verwendet und nicht zu viele weniger geeignete Variablen umfaßt scheint es im Hinblick auf die *AAD*-Werte nicht wichtig zu sein, welches Modell zur Vorhersage gewählt wird. Als Grund hierfür kann die gewählte konservative

Parametrisierung verstanden werden. Für niedrige Anteile des Testdatensatzes am gesamten Datensatz ergeben sich *AAD*-Werte, die etwas größer sind als die entsprechenden Werte der besten kollaborativen Verfahren. Dafür ergeben sich bei hohen Anteilen des Testdatensatzes an der betrachteten Datenmenge sogar kleinere *AAD*-Werte als in in allen in Kapitel 4 dargestellten kollaborativen Verfahren. In diesem für marketingrelevante Anwendungen besonders wichtigen Bereich sind die *AAD*-Werte gut vergleichbar mit den *AAD*-Werten, die man im mittels des hybriden Verfahrens nach Pazzani (1999) erhält.

Da die mit Hilfe des HBLR-Verfahrens generierten Schätzer aufgrund der gewählten Parametrisierung eher konservative Prognosen sind, fällt die Präzision im allgemeinen recht hoch aus, während der Recall nur niedrige Werte erreicht. Die in Tabelle 5.11 aufgelisteten Präzisions- und Recallwerte basieren auf dem Schwellenwert 4,5. Zum Schwellenwert 4,25 erhält man für die HBLR-Verfahren Werte für Präzision und Recall, die gut mit den auf Grundlage des Verfahrens von Pazzani (1999) berechneten entsprechenden Werten vergleichbar sind.

Die Ergebnisse des komplexesten Modells (M_1) sind im Hinblick auf den *AAD*-Wert allen übrigen Modellen im Rahmen des HBLR-Verfahrens zum Teil sogar geringfügig überlegen. Dieses Modell wies im Testdatensatz den kleinsten *DIC*-Wert auf. Diese Ergebnisse legen die Vermutung nahe, daß der absolute *DIC*-Wert durchaus zur Auswahl des besten hierarchischen Modells geeignet sein kann, sofern es um die Vorhersage von Bewertungen im Hinblick auf Items geht, hinsichtlich derer bereits Bewertungen im Trainingsdatensatz vorhanden sind und eine konservative Parametrisierung gewählt wird. Im Rahmen eines Bayes'schen Ansatzes, in dem für jeden Nutzer individuelle Schätzer bestimmt werden, kann auch die Verwendung von für die Gesamtheit der Nutzer weniger relevanten Regressoren vorteilhaft sein. Ein Vergleich der Breese-Werte führt dagegen eher zu der Vermutung, daß die Regressorenkombination M_3 optimal ist. Dies würde für die Wahl eines Modells mit möglichst niedrigem Komplexitätsgrad sprechen. Die Unterschiede zwischen den Modellen sind allerdings im Rahmen des konservativen HBLR-Ansatzes so gering, daß sich der mit einer MCMC Modellbestimmung verbundene Rechenaufwand nicht rechtfertigen läßt. Dies ist ein wichtiger Vorteil des konservativen HBLR-Ansatzes.

Tabelle 8.13 enthält die *AAD*-Werte des hybriden GP-Modells. Die *AAD*-Werte des hybriden GP-Verfahrens sind für verhältnismäßig kleine Testdatenmengen

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
M_0	90,65	84,09	74,68	67,34	62,54	60,38	56,22	53,40	49,92
M_1	86,84	83,48	75,28	66,64	63,18	60,73	57,47	55,15	51,24
M_2	86,56	81,54	72,78	64,52	61,17	57,18	53,90	52,20	48,91
M_3	89,91	82,73	73,37	66,32	62,86	60,27	56,55	53,57	50,95
M_4	85,29	80,97	71,58	63,80	59,72	57,36	53,68	51,33	49,88
M_5	86,27	82,98	72,93	65,37	61,33	58,15	55,62	52,52	50,77
M_6	86,87	80,88	72,10	63,69	59,61	56,59	53,72	50,82	48,95
M_7	85,10	82,63	72,30	64,50	60,92	59,04	55,76	52,70	49,96
M_8	85,68	83,07	72,81	65,25	61,36	59,14	55,79	53,20	49,97

Tabelle 8.12: Breese-Werte (R_B) des hierarchischen linearen Regressionsmodells für $M_0, M_1, M_2, M_3, M_4, M_5, M_6, M_7$ und M_8 hinsichtlich des Testdatensatzes bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz

kleiner als die *AAD*-Werte des HBLR-Ansatzes. Man erkennt, das die komplexeren Modelle M_1, M_2, M_7 und M_8 bei höheren Anteilen des Testdatensatzes an den betrachteten Daten zu starken Überanpassungseffekten führen können. Bei sehr geringen Anteilen des Testdatensatzes an der verwendeten Datenmenge (also wenn viele Bewertungen zur Berechnung der Schätzer zur Verfügung stehen) kann das komplexeste Modell (M_1) im Vergleich mit den anderen Modellen im Rahmen des GP-Verfahrens zu sehr guten Werten führen. Falls die Datenmenge ausreichend groß ist, scheint Bayes'schen Modellen die Verwendung von einigen weniger wichtigeren Regressoren nicht zu schaden. Für kleine Datenmengen ist es dagegen in diesem Ansatz notwendig, vorab die wichtigsten Variablen zu identifizieren. Falls nur wenig Daten zur Bestimmung der Schätzer verwendet werden können, erhält man im Rahmen des hybriden GP-Ansatzes erstaunlich gute Werte, wenn nur die durchschnittliche Kritikerbewertung (M_0) als Regressor verwendet wird.

Die Breese-Werte des GP-Verfahrens für die verschiedenen Regressorenkombinationen können Tabelle 8.14 entnommen werden. Vor allem, wenn viele Daten zur Berechnung der Schätzer verwendet werden, fallen die Breese-Werte für das GP-Verfahren erheblich kleiner aus als die entsprechenden Werte hinsichtlich des konservativen HBLR-Ansatzes. Daher ist davon auszugehen, daß in die-

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
M_0	0,713	0,726	0,723	0,720	0,725	0,727	0,732	0,742	0,768
M_1	0,695	0,722	0,741	0,769	0,785	0,796	0,797	0,820	0,969
M_2	0,719	0,736	0,761	0,780	0,784	0,811	0,821	0,837	0,979
M_3	0,704	0,715	0,720	0,723	0,724	0,734	0,734	0,767	0,795
M_4	0,694	0,720	0,711	0,721	0,718	0,736	0,734	0,776	0,814
M_5	0,694	0,710	0,717	0,722	0,714	0,731	0,735	0,771	0,806
M_6	0,709	0,716	0,719	0,723	0,723	0,738	0,736	0,771	0,816
M_7	0,694	0,721	0,713	0,723	0,718	0,742	0,737	0,789	0,839
M_8	0,709	0,713	0,718	0,727	0,723	0,741	0,743	0,797	0,873

Tabelle 8.13: *AAD* des hybriden GP-Modells für $M_0, M_1, M_2, M_3, M_4, M_5, M_6, M_7$ und M_8 hinsichtlich des Testdatensatzes bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz

sem Bereich das konservative HBLR-Verfahren trotz schwächerer *AAD*-Werte zu nützlicheren Empfehlungen führen kann. Bei hohen Anteilen des Testdatensatzes nivellieren sich die Unterschiede zwischen den Breese-Werten der beiden hybriden Ansätze. Auffallend sind die verhältnismäßig hohen Breese-Werte, die sich in diesem Bereich bei Wahl des GP-Verfahrens ergeben können, wenn das am wenigsten komplexe Modell (M_0) verwendet wird. Vor diesem Hintergrund scheinen im Hinblick auf das hierarchische GP-Verfahren Modelle mit geringem Komplexitätsgrad deutlich geeigneter für marketingrelevante Vorhersagezwecke als komplexere Ansätze zu sein. Da das GP-Verfahren im Gegensatz zum konservativen HBLR-Verfahren kein MCMC-Ansatz ist, kann im Hinblick auf das GP-Verfahren definitionsgemäß kein *DIC*-Wert berechnet werden. Anders als beim konservativen HBLR-Ansatz können hier auf keinen Fall weniger wichtige Variablen zusätzlich verwendet werden. Falls nicht die wichtisten Variablen wie die durchschnittliche Kritikerbewertung a priori bekannt sind, erscheint es im Hinblick auf das GP-Verfahren sinnvoll, wenige elementare Variablen zu identifizieren, die erheblich wichtiger als alle übrigen Variablen sind und diese als Regressoren zu verwenden.

Auch bei verzerrten Trainigsdatensätzen erweisen sich die mittels des HBLR-

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
M_0	79,63	70,84	65,79	63,08	60,94	58,79	57,75	57,01	56,29
M_1	81,04	71,01	64,31	59,76	59,04	55,27	53,61	53,53	46,75
M_2	79,72	67,65	62,66	56,80	54,37	53,97	50,77	49,36	45,81
M_3	80,87	71,37	65,86	63,19	61,36	57,75	57,31	54,03	51,82
M_4	80,96	71,24	66,50	62,71	62,31	58,38	57,19	53,94	51,19
M_5	81,29	71,55	66,33	62,54	61,75	58,27	57,23	54,23	51,29
M_6	81,06	69,49	65,47	61,11	59,92	56,86	56,67	53,67	50,61
M_7	80,92	71,32	66,22	62,68	62,31	58,04	57,01	53,33	49,46
M_8	80,68	70,23	66,09	61,94	61,07	57,38	56,16	52,25	48,19

Tabelle 8.14: Breese-Werte (R_B) des hierarchischen GP-Ansatzes für $M_0, M_1, M_2, M_3, M_4, M_5, M_6, M_7$ und M_8 hinsichtlich des Testdatensatzes bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz

Verfahrens berechneten Schätzer besser zur Generierung nützlicher Empfehlungslisten geeignet als die auf Basis des hybriden GP-Verfahrens ermittelten Schätzer (vgl. Anhang B.4). Die von der hybriden Abwandlung des kollaborativen Verfahrens von Yu et. al. (2006) erhofften höheren Breese-Werte bewahrheiten sich nicht.

Sofern ein größerer Ausschnitt aus dem MovieLens-Datensatz verwendet wird, der die Bewertungen von 2020 Nutzern hinsichtlich der 418 Filme beinhaltet, ist zu bedenken, daß in diesem Datensatz im Durchschnitt pro Nutzer ca. insgesamt 64 Bewertungen zur Verfügung stehen, während im kleineren Datensatz pro Nutzer insgesamt ca. 88 Bewertungen vorhanden sind. Sofern im großen Datensatz 10-50 % der gesamten Datenmenge im Testdatensatz enthalten ist bzw. im kleinen Datensatz 40-70 % der verfügbaren Daten zum Testdatensatz gehört, stehen im Trainingsdatensatz pro Nutzer im Durchschnitt ca. 30-60 Bewertungen als Datengrundlage zur Bestimmung der Schätzer zu Verfügung. In beiden Fällen erhält man mit dem hybriden GP-Verfahren Breese-Werte die insgesamt zumindest nicht viel schlechter als die entsprechenden Werte hinsichtlich des HBLR-Verfahrens sind. Da das hybride GP-Verfahren das mit Abstand schnellste behandelte Verfahren ist und das HBLR-Verfahren eher zu den langsameren Verfahren gehört, kann es sofern sehr viele Bewertungen zur Berechnung der Schätzer zur

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,736	0,747	0,745	0,747	0,745	0,749	0,753	0,760	0,786
Prec.	0,683	0,642	0,661	0,664	0,659	0,653	0,612	0,612	0,578
Rec.	0,075	0,071	0,073	0,069	0,078	0,082	0,073	0,074	0,076
R_B	84,57	76,43	71,56	67,98	66,62	64,40	63,07	62,74	61,58

Tabelle 8.15: *AAD*, Präzision, Recall und Breese-Werte (R_B) des hierarchischen linearen Regressionsansatzes für die Regressorenkombination M_8 hinsichtlich des Testdatensatzes bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz (2020 Nutzer, 418 Filme)

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
AAD	0,711	0,726	0,721	0,734	0,724	0,753	0,750	0,797	0,814
Prec.	0,576	0,537	0,543	0,527	0,543	0,495	0,502	0,440	0,422
Rec.	0,142	0,184	0,178	0,189	0,168	0,193	0,191	0,244	0,253
R_B	82,72	73,72	68,86	64,92	63,29	59,66	58,15	54,22	52,45

Tabelle 8.16: *AAD*, Präzision, Recall und Breese-Werte (R_B) des hybriden GP-Verfahrens für die Regressorenkombination M_8 hinsichtlich des Testdatensatzes bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz (2020 Nutzer, 418 Filme)

Verfügung stehen empfehlenswert sein, das hybride GP-Verfahren anstelle des HBLR-Verfahrens zu verwenden. Gerade in diesem Bereich führen die zweimodalen Clusterverfahren zu erheblich höheren Breese-Werten (siehe Tabelle 4.20). Daher scheinen zweimodale Clusterverfahren besser geeignet zur Vorhersagen auf der Basis von 30-60 Bewertungen zu sein. (Hierbei ist allerdings zu berücksichtigen, daß bei den zweimodalen Clusterverfahren im Gegensatz zu den hybriden hierarchischen Verfahren nicht nur die durchschnittliche Anzahl von Bewertungen pro Nutzer sondern auch die Anzahl der betrachteten Items J ausschlaggebend für die Qualität der Bewertungen ist. Es ist daher nicht davon auszugehen, daß die zweimodalen Clusterverfahren auch für sehr große Werte von J auf Basis von durchschnittlich 30-60 Bewertungen pro Nutzer zu besseren Ergebnissen führen

müssen als die hybriden hierarchischen Regressionsansätze.)

Falls hinsichtlich des kleinen Datensatzes 10-30 % der Daten dem Testdatensatz angehören, sind durchschnittlich mehr als 60 Bewertungen pro Nutzer zur Berechnung der Schätzer verfügbar. In diesem Bereich kann man mit Hilfe des konservativen HBLR-Verfahrens zum Teil erheblich höhere Breese-Werte als bei Verwendung aller übrigen Verfahren erzielen (vgl. Tabelle 5.12).

Kapitel 9

Neue Items

Insbesondere auf Basis der kollaborativen Ansätze und der bislang bekannten kontentbasierten Erweiterungen kollaborativer Verfahren lassen sich im allgemeinen keine sinnvollen Empfehlungen für weniger häufig bewertete Items generieren. Beispielsweise beruhen die im Rahmen eines Item-basierten Ähnlichkeitsansatzes berechneten Korrelationen für vergleichsweise selten beurteilte Items i.d.R. auf wenigen gemeinsamen Beobachtungen und sind im allgemeinen nicht besonders verläßlich. Beim Nutzer-basierten Ähnlichkeitsverfahren existieren im Hinblick auf nicht so oft bewertete Items meist nur wenige oder gar keine Nutzer in der Nachbarschaft der Person, deren Bewertung zu schätzen ist. Gleiches gilt für das hybride Verfahren nach Pazzani (1999).

Im Hinblick auf gar nicht beurteilte Items können weder mittels kollaborativer Verfahren noch mit Hilfe ihrer kontentbasierten Erweiterungen Schätzer berechnet werden.

Es wurde bereits darauf hingewiesen, daß Empfehlungen in Bezug auf weniger bekannte Items für die Nutzer von größerem Interesse sein dürften. Schließlich werden den Nutzern die bekannteren Items ohnehin häufig von anderen Personen empfohlen. Teilweise werden Items, mit denen bereits zahlreiche Menschen Erfahrungen gesammelt haben, sogar von vielen Personen empfohlen. Sofern diese Empfehlungen von Personen stammen, die den Geschmack und die Bedürfnisse des Beratenen gut einschätzen können, ist anzunehmen, daß sie sogar zutreffender sein könnten als ihre maschinell bestimmten Gegenstücke. Daher ist es wahrscheinlich, daß ein bekanntes Item dem Nutzer bereits von mehreren Personen empfohlen wurde. Die Empfehlung dieses Items durch das Recommender-System kann dann bestenfalls als Bestätigung der vorherigen Empfehlungen angesehen

werden. Ein Hinweis auf ein Item, das dem Nutzer ansonsten nicht als Option bewußt geworden wäre, ist sie dagegen nicht.

Außerdem ist zu bedenken, daß aufgrund der dem Recommender-System vorliegenden Daten in Bezug auf einen bestimmten Nutzer meist nur ein kleiner Teil der Items identifizierbar ist, bezüglich derer der betreffende Nutzer bereits ausreichend informiert ist. Bei Items wie CDs, DVDs oder Büchern sind Empfehlungen für dem Nutzer bereits bekannte Items nicht hilfreich. Je bekannter ein Item ist, umso größer ist auch die Wahrscheinlichkeit, daß der Nutzer bereits von diesem Item gehört hat oder sogar eigene Erfahrungen mit dem betreffenden Item sammeln konnte. Da die wenigsten Nutzer jedes ihnen bekannte Item auch bewertet haben, steigt hierdurch die Wahrscheinlichkeit, daß den Nutzer Items empfohlen werden, die sie schon ausprobiert haben. Daher kann die Beschränkung auf vorwiegend bekannte Items den Nutzen der Empfehlungen mindern.

Selbstverständlich ist eine zutreffende Emfehlung für ein dem jeweiligen Nutzer bereits geläufiges Item im allgemeinen dazu geeignet, das Vertrauen in das Recommender-System erhöhen. Es ist aber nicht davon auszugehen, daß hierdurch ein erheblicher Beitrag zur Freizeitgestaltung oder zur Erhöhung der Lebensqualität des Nutzers geleistet werden kann. Zwar kann das den Empfehlungen entgegengebrachte Vertrauen insbesondere im Hinblick auf neue Nutzer ein wesentlicher Erfolgsfaktor hinsichtlich der Akzeptanz und weiteren Verwendung des Recommender-Systems sein. Zu diesem Zweck werden bereits seit geraumer Zeit eine Reihe anderer Methoden eingesetzt, die sich bestens in der Praxis bewährt haben. Eine erfolgreiche Maßnahme zur Erhöhung des Vertrauens scheinen Erklärungen der Empfehlungen zu sein (Herlocker et. al. (2000)). Amazon.de bietet seinen Kunden Begründungen für die ihnen unterbreiteten Empfehlungen an. Falls ein Kunde es als erklärungsbedürftig empfindet, warum ihm ein bestimmtes Item empfohlen wurde, kann er einen Button anklicken. Daraufhin wird er an Items erinnert, die dem empfohlenen Item ähneln und die er hoch bewertet hat. Dies hat den zusätzlichen Vorteil, daß der Nutzer die Empfehlungen besser verstehen und sie somit gezielter für seine Zwecke einsetzen kann. Beispielsweise kann es sein, daß ein Nutzer, der generell geistreiche Filme bevorzugt, sich nach einem besonders aufreibenden Tag weniger gern einen anspruchsvollen Film ansehen möchte. Die (automatisierte) Begründung seiner Empfehlungen durch das Recommender-System von Amazon.de kann es dem Nutzer ermöglichen, die Items besser einzuordnen und auf dieser Basis dasjenige Item auszuwählen, das

für den jeweiligen Moment am besten geeignet erscheint. (Gleichzeitig wird der Nutzer auf diese Weise mit seinen zu einem früheren Zeitpunkt abgegebenen Bewertungen erneut konfrontiert und kann diese bei Bedarf abändern.) Somit sind Erklärungen geeignet, nicht nur das Vertrauen in die Empfehlungen, sondern auch den Nutzen derselben zu erhöhen.

Vor diesem Hintergrund ist eine ideale Empfehlung für einen Nutzer nicht schon dadurch gegeben, daß der Nutzer den ihm empfohlenen Gegenstand auch tatsächlich für sehr gut oder hervorragend hält. Vielmehr ist eine Empfehlung erst dann ideal, wenn sie nicht nur zutreffend ist, sondern darüberhinaus auch auf ein Item hinweist, daß dem Nutzer ohne das Empfehlungssystem entweder gar nicht oder erst zu einem späteren Zeitpunkt empfohlen worden wäre. Je unbekannter ein Item ist, umso wertvoller kann eine Empfehlung in Bezug auf diesen Gegenstand sein.

Der Betreiber eines Online-Geschäfts kann durch hilfreiche Empfehlungen die Kundenzufriedenheit und die Kundenbindung erhöhen. Da der Wechsel zu einem anderen Online-Geschäft für die Nutzer mit allenfalls geringfügigem Zeitaufwand und vernachlässigbaren Kosten verbunden ist, sind die Kundenzufriedenheit und die Kundenbindung wichtige Erfolgsfaktoren eines Online-Geschäfts.

Zusätzlich ermöglicht eine zuverlässige Schätzung von Bewertungen hinsichtlich gänzlich unbekannter Items es auch abzuschätzen, welche und wieviele Nutzer Interesse an einem bestimmten neuen Item haben dürften. Dies könnte theoretisch zur Entscheidungsgrundlage hinsichtlich der Aufnahme eines neuen Items ins Sortiment gemacht werden. Da die Kaufhistorien der Kunden meist auch Rückschlüsse auf deren Zahlungsbereitschaft zulassen, wäre es denkbar, die Schätzer vorab zur Bestimmung des Preises (und zur Steuerung der damit verbundenen Nachfrage) einzusetzen. Außerdem wären auf dieser Basis auch gezielterere Werbemaßnahmen vorstellbar.

Im Rahmen der folgenden Argumentation werden als neue Items solche Items bezeichnet, bezüglich derer die Nutzer noch keine Bewertungen abgegeben haben. Entsprechend sind im Zusammenhang mit dem empirischen Vergleich der vorgestellten Verfahren neue Items diejenigen Items, hinsichtlich derer keine Bewertungen im Trainigsdatensatz zur Verfügung stehen.

Neben dem unbestreitbar höheren Nutzen, den Empfehlungen für neue Items für die Nutzer haben dürften, ist es in besonderen Maße für Betreiber und Manager eines Online-Geschäfts vorteilhaft, gerade auch Bewertungen für vollständig

unbekannte Items abschätzen zu können. Überdies könnten so selbst Schätzer für (noch) nicht existente Produkte bestimmt werden. Auf diese Weise könnten in Bezug auf neue Items geeignete Schätzer sogar auch zum Entwurf neuer Produkte verwendet werden.

Obwohl es Marketinggesichtspunkten überaus wichtig ist, Bewertungen für neue Items abschätzen zu können, gibt es wenig Literatur, die sich mit diesem Problem befaßt. Ein Ansatz, der wenigstens in diese Richtung geht, stammt von Ansari et. al. (2000). Die Autoren stellen ein hybrides Verfahren vor und evaluieren dessen Vorhersagegenauigkeit in Bezug auf im Trainigsdatensatz nicht vorhandene Items.

In der Tat ist es naheliegend, ein hybrides Verfahren für diese Aufgabe einzusetzen. Gerade in letzter Zeit sind aber sehr schnelle und effiziente kollaborative Verfahren wie die zweimodalen Clusterverfahren und das kollaborative hierarchische GP-Verfahren entwickelt worden. Mit diesen kollaborativen Methoden ist im Vergleich zu ähnlich genauen hybriden Verfahren (wie zum Beispiel dem linearen hierarchischen Regressionsansatz) erheblich weniger Rechenaufwand verbunden. Das motiviert den Versuch, besonders vielversprechende kollaborative Verfahren durch geringfügige kontentbasierte Erweiterungen oder Modifikationen zur Schätzung von Bewertungen hinsichtlich neuer Items geeignet zu machen.

9.1 Das Problem

Solange ein Item noch nicht bewertet wurde, kann man es mit Hilfe eines zweimodalen Clusterverfahrens keinem Cluster zuordnen. George, Merugu (2005) empfehlen die Verwendung der durchschnittlichen Bewertung $\bar{y}_{i.}$ als Schätzer für J^B, falls hinsichtlich des Items j keine Bewertungen im Trainingsdatensatz enthalten sind. Das ist genau die gleiche Vorhersage, die sich mittels eines Nutzer-basierten Ähnlichkeitsansatzes (Resnick et. al. (1994)) für neue Items ergibt. Dieser Ansatz hat zur Folge, daß jedes neue Item mit derselben Wahrscheinlichkeit empfohlen wird. In der überwiegenden Zahl der Fälle existiert eine hohe Anzahl von durch vergleichsweise viele Nutzer bewertete Items j^*, für die Bewertungen im Traingsdatensatz enthalten sind und hinsichtlich derer außerdem in Bezug auf einen bestimmten Nutzer i $\hat{Y}_{ij^*} > \bar{y}_{i.}$ gilt. Daher werden neue Items praktisch nie empfohlen, falls man sich mit $\bar{y}_{i.}$ als Schätzer begnügt. In den seltenen Fällen, in denen dennoch neue Items empfohlen werden, kann dies nur geschehen, wenn

der betreffende Nutzer generell extrem großzügige Bewertungen abgegeben hat, die sich zudem kaum voneinander unterscheiden. Es liegt auf der Hand, daß die Schätzer für neue Items auf andere Weise bestimmt werden müssen.

9.2 Lösungsmöglichkeiten

Sofern die relevanten Eigenschaften der Items bekannt sind, existieren verschiedene Strategien zur Lösung des beschriebenen Problems, die direkte und die indirekte Schätzung. Hinsichtlich beider Strategien werden verschiedene Varianten vorgestellt.

Eine bereits genannte Lösungsmöglichkeit ist, das Problem durch die Verwendung eines hybriden Verfahrens zu umgehen. Dieser Ansatz kann als direkte Schätzung bezeichnet werden. Die direkte Schätzung kann z.B. mittels eines hierarchischen Bayes'schen Regressionsansatzes erfolgen. Möglich ist auch, den Trainingsdatensatz vor der Verwendung des hierarchischen Bayes'schen Regressionsansatzes zweimodal zu clustern und die hierdurch gewonne Information dann (z.B. via die Prior) in das hierarchische Verfahren miteinfließen zu lassen.

Zudem kann Zusatzinformation über die Eigenschaften der Items dazu benutzt werden, die Clusterzugehörigkeit der neuen Items abzuschätzen. Dies wird hier als indirekte Schätzung bezeichnet.

9.2.1 Direkte Schätzung

Für die direkte Schätzung ohne Rekurs auf die Resultate eines zweimodalen Clusterverfahrens ist $d = 1$ und $Z = Z_1$. Da das Ziel die Schätzung von Bewertungen für Items ist, für die im Trainingsdatensatz noch keine Bewertungen vorhanden sind, sind für die betrachteten Items keine Clusterzugehörigkeiten bekannt. Daher kann auch ein auf zweimodalen Clustern basierendes Verfahren zur direkten Schätzung nur die Clusterzugehörigkeiten der Nutzer verwenden. Wie in Abschnitt 8.3.2 beschrieben, ist in diesem Fall im Rahmen des hierarchischen linearen Regressionsmodells entweder $d = K \wedge Z = Z_2 \in \mathbb{R}^{I,K}$ oder $d = K + 1 \wedge Z = Z_3 \in \mathbb{R}^{I,K+1}$. Alternativ kann man natürlich für jedes hybride Verfahren K getrennte Modelle für die verschiedenen Nutzer-Cluster bilden und diese Modelle getrennt berechnen.

9.2.2 Direkte Schätzung auf Basis clusterspezifischer Modelle

Eine weitere Möglichkeit ist die Bildung separater Modelle für jedes einzelne Nutzer-Cluster. Dies setzt das bereits das Ergebnis eines zweimodalen Clusterverfahrens voraus. Hierbei verwendet man für jedes einzelne Nutzer-Cluster k aus $\{1, ..., K\}$ die in der $k-$ten Zeile der Gewichtungsmatrix W, $(W_{k1}, ..., W_{kL})$, enthaltene verdichtete Information. Jedes Element $W_{kl}, l \in \{1, ..., L\}$, ist die durchschnittliche Bewertung der Nutzer als dem $k-$ten Cluster in Bezug auf die Items, die zum $l-$ten Item-Cluster gehören. Dementsprechend spiegelt ein hoher Wert für W_{kl} eine positive Einstellung der Nutzer des $k-$ten Clusters erster Modalität hinsichtlich der Items aus dem $l-$ten Item-Cluster wieder. Sofern es gelingt, die Eigenschaft(en) zu bestimmen, durch die sich die Items aller von den Nutzern aus dem $k-$ten Nutzer-Cluster besser bewerteten Item-Cluster von den übrigen Item-Clustern unterscheiden, so hat man hierdurch zumindest einen Teil der Eigenschaften identifiziert, die in Bezug auf die Nutzer des $k-$ten Nutzer-Clusters ausschlaggebend für die Bewertungen sein könnten. Die auf diese Weise bestimmten Merkmale können dann zur Kandidatenmenge einer separaten Modellbestimmung für das $k-$te Cluster werden.

Da alle Items Eigenschaften aufweisen, ist es möglich, in Bezug auf jedes einzelne Item-Cluster die durchschnittlich Ausprägung jeder Eigenschaft zu berechnen. Eine vergleichsweise hohe durchschnittliche Ausprägung der $\kappa-$ten Eigenschaft innerhalb der Items aus dem l-ten Item-Cluster weist die zugehörige Eigenschaft als ein mögliches Charakteristikum des l-ten Item-Clusters aus. Auf diese Weise kann man zumindest für die meisten Item-Cluster eine Reihe möglicherweise relevanter Eigenschaften bestimmen.

Diese möglichen Charakteristika können zur separaten Identifikation der vielleicht bewertungsrelevanten Eigenschaften in Bezug auf die unterschiedlichen Personen-Cluster verwendet werden. Eigenschaften, die gleichermaßen charakteristisch für vom betrachteten Nutzer-Cluster hoch wie niedrig bewertete Item-Cluster sind, brauchen nicht berücksichtigt werden.

Beispiel 9.1:

In diesem Beispiel werden bezüglich der zweimodalen Cluster aus Beispiel 5.11 und hinsichtlich der Eigenschaften aus Beispiel 4.1 die für die Personen aus

Item-Cluster	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$	$\kappa = 6$
$l = 1$	9,00	6,00	2,67	1,67	6,00	8,33
$l = 2$	6,00	8,50	5,00	5,50	5,50	3,50
$l = 3$	3,67	2,67	3,67	4,33	3,33	8,00

Tabelle 9.1: Durchschnittliche Merkmalsausprägungen für die einzelnen Item-Cluster (Beispiel 9.1)

dem Nutzer-Cluster $k = 1$ möglicherweise bewertungsrelevante Merkmale bestimmt. Aus den in Tabelle 4.1 angegebenen Eigenschaftsausprägungenen berechnet man die in Tabelle 9.1 dargestellten durchschittlichen Merkmalsausprägungen bezüglich der jeweiligen Nutzer-Cluster. Der Gewichtungsmatrix aus Beispiel 5.11

$$W = \begin{pmatrix} 3{,}78 & 1{,}00 & 4{,}50 \\ 2{,}38 & 3{,}60 & 4{,}00 \end{pmatrix}$$

entnimmt man, daß die Filme aus dem dritten Item-Cluster ($l = 3$) von den Nutzern des ersten Clusters erkennbar favorisiert werden. Die einzige verhältnismäßig stark ausgeprägte Eigenschaft dieser Filme ist die Eigenschaft $\kappa = 6$ (Charakterentwicklung). Deren durchschnittliche Merkmalsausprägung fällt auch im Hinblick auf die Filme aus dem ersten Item-Cluster hoch aus. Da auch die Filme des ersten Item-Clusters von den Personen aus dem ersten Nutzer-Cluster überwiegend als recht gut bewertet wurden, bestärkt das die Annahme, daß das Ausmaß der dargestellten Charakterentwicklung für die Personen aus dem dritten Nutzer-Cluster eine relevante Eigenschaft sein könnte. Geringe Ausprägungen dieser Eigenschaft finden sich dagegen bei den Items aus Cluster $l = 2$, die von den Mitgliedern des ersten Nutzer-Clusters deutlich abgelehnt werden. Daher ist das Ausmaß der dargesten Charakterentwicklung möglicherweise relevant.

9.2.3 Indirekte Schätzung

Eine Möglichkeit zur indirekten Schätzung basiert auf den zweimodalen Clusterverfahren. Zuerst verwendet man eines der in Abschnitt 5.6 (5.8) beschriebenen zweimodalen Clusterverfahren, um für alle betrachteten Nutzer und alle im Items $j \in J^B$, hinsichtlich derer Bewertungen im Trainingsdatensatz enthalten sind,

die Clusterzugehörigkeit zu bestimmen. Danach werden die Eigenschaften der bekannten Items $j \in J^B$ dazu benutzt, ein Modell zu bilden bzw. ein Verfahren zu trainieren, welches die Clusterzugehörigkeit $q_{jl}, j \in J^B$, eines Items zu einem der Cluster $l \in \{1, ..., L\}$ auf dessen Eigenschaften zurückführt. Das Modell bzw. der trainierte Algorithmus werden danach dazu verwendet, die Clusterzugehörigkeiten der neuen Items $j \in J^N$ zu schätzen. Hierdurch ergeben sich die Schätzer $\hat{q}_{jl}, j \in J^N$. Mittels dieser Schätzer $\hat{q}_{jl}, j \in J^N$ lassen sich im Rahmen eines zweimodalen $\hat{S}^1_Y$-Clusterverfahrens bereits die Schätzer

$$\hat{S}^{N1}_{Y,ij} = \sum_{k'=1}^{K} \sum_{l'=1}^{L} p_{ik'} w_{k'l'} \hat{q}_{jl'}, \;\; i = 1, ..., I, j \in J^N,$$

berechnen. Um die Schätzer $\hat{S}^2_{Y,ij}$ für $j \in J^N$ zu berechnen, muß außerdem noch eine Näherung für $\bar{y}_{.j}, j \in J^N$, gefunden werden. Es bietet sich an, hierzu die durchschnittliche Bewertung bekannter Kritiker bzw. professioneller Produkttester zu verwenden. Sei M^C_j die Menge der relevanten Kritiker bzw. Produkttester, die Bewertungen für das Item $j \in J^N$ abgegeben haben. Weiter sei $y^{MC}_{i_{MC}j}$ die Bewertung des Kritikers bzw. Produkttesters $i_{MC} \in M^C_j$ für das Item $j \in J^N$. Wichtig ist, daß allen Bewertungen der Kritiker $y^{MC}_{i_{MC}j}, i_{MC} \in M^C_j, j \in J^N$, dieselbe Skala zugrunde zu liegen hat wie den Bewertungsdaten der Nutzer $J^B, i = 1, ..., I, j \in J^B$. (Andernfalls müssen die Bewertungen der betreffenden Kritiker geeignet transformiert werden.) Im folgenden wird davon ausgegangen, daß den verwendeten Kritikerbewertungen dieselbe Skala wie den Nutzer-Bewertungen zugrundeliegt. Dann ist $\bar{y}^{MC}_{.j} = (1/|M^C_j|) \sum_{i'_{MC} \in M^C_j} y^{MC}_{i'_{MC}j}$ ein möglicher Ersatz für $\bar{y}_{.j}, j \in J^N$. Weiterhin ist zu bedenken, daß nicht nur $\bar{Y}_{.j}$ sondern vielmehr die gesamte Differenz

$$\bar{y}_{.j} - \sum_{l=1}^{L} q_{jl} \tilde{w}_{.l},$$

welche die Heterogenität des j-ten Items abbilden soll, zu approximieren ist. Da das Niveau der durchschnittlichen Nutzer-Bewertung von der durchschnittlichen Kritiker-Bewertung hinsichtlich eines Item-Clusters abweichen kann, sollte auch $\tilde{w}_{.l}$ durch die durchschnittliche Kritiker-Bewertung hinsichtlich des l-ten Item-Clusters ($l \in \{1, ..., L\}$)

$$\tilde{w}^{MC}_{.l} = \frac{\sum_{j' \in J^B} \sum_{i'_{MC} \in M^C_{j'}} q_{j'l} y^{MC}_{i'_{MC}j'}}{\sum_{j' \in J^B} \sum_{i'_{MC} \in C_{j'}} q_{j'l}}$$

ersetzt werden. Insgesamt ergibt sich daher

$$\hat{S}^{N2}_{Y,ij} = \sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{ik'} w_{k'l'} \hat{q}_{jl'} + \bar{y}_{i.} - \sum_{k'=1}^{K} p_{ik'} \tilde{w}_{k'.} + \bar{y}^{MC}_{.j} - \sum_{l'=1}^{L} \hat{q}_{jl'} \tilde{w}^{MC}_{.l'},$$

für $i = 1, ..., I, j \in J^N$. Sind genügend Kritiker-Bewertungen im Hinblick auf alle interessierenden neuen Items $j \in J^N$ vorhanden, so ist auch in Bezug auf den $\hat{S}^2_Y$-Schätzer das einzige noch zu lösende Problem die optimale Bestimmung von $\hat{q}_{jl}, l = 1, ..., L, j \in J^N$.

Sei $a_{j\kappa}$ die κ-te Eigenschaft des j-ten Items und sei κ_A die Anzahl der relevanten Eigenschaften der Items ($\kappa \in \{1, ..., \kappa_A\}$). Weiter sei $a_j = (a_{j1}, ..., a_{j\kappa_A})'$ der Eigenschaftsvektor, der das j-te Item beschreibt. Auf Basis dieser Eigenschaften, die sowohl für die bekannten Items aus dem Trainingsdatensatz als auch im Hinblick auf die neuen Items $j \in J^N$ bekannt sein müssen, lassen sich dann anhand der Trainingsdaten die Parameter eines Verfahrens bestimmen, das in der Lage ist, auf Basis der Eigenschaften eines beliebigen Items dessen Cluster-Zugehörigkeit zu schätzen. Da sich ein solches Verfahren auf die Item-Daten beschränkt, ist es im Vergleich zu Verfahren, die die gesamte Matrix y und die Item-Eigenschaften verwenden, mit vernachlässigbarem Rechenaufwand verbunden. Da die zweimodalen Clusterverfahren außerdem deutlich schneller als hierarchische Regressionsansätze und trotzdem vergleichsweise genau sind, lohnt sich die Beschäftigung mit Verfahren zur indirekten Schätzung. Daher wird in diesem Abschnitt eine Reihe verschiedener Verfahren vorgestellt, mittels derer die Cluster-Zugehörigkeit neuer Items geschätzt werden kann.

9.2.3.1 Distanz-Heuristiken

Heuristisch kann man die Distanz $d_\nu(j_1, j_2)$ zwischen zwei verschiedenen Items j_1 und j_2 als gewichteten euklidischen Abstand zwischen a_{j_1} und a_{j_2} mit Gewichtungsfaktoren $\nu_\kappa, \kappa \in \{1, ..., \kappa_A\}$ definieren:

$$d_\nu(j_1, j_2) = \sqrt{\sum_{\kappa'=1}^{\kappa_A} \nu_{\kappa'} (a_{j_1\kappa'} - a_{j_2\kappa'})^2}.$$

Die Abstände $d_\nu(j_1, j_2)$ der Items $j_1 \in J^N$ und $j_2 \in J^B$ sind die Grundlage dreier

verschiedener Heuristiken. Jede dieser Heuristiken benutzt die Grundidee eines zu anderen Zwecken einsetzbaren bekannten Verfahrens oder Ansatzes.

Das SL-Verfahren zur Approximation der Item-Cluster Zugehörigkeit ist nach dem einmodalen single-linkage Clusterverfahren (Sneath (1957)) benannt. Sei $j_1 \in J^N$ und sei $j_1' = \arg\min_{j_2 \in J^B} d_\nu(j_1, j_2)$ das bekannte Element zweiter Modalität, das dem Item $j_1 \in J^N$ bezüglich des gewichteten euklidischen Abstands am ähnlichsten ist. Dann ordnet das SL-Verfahren zur Approximation der Item-Cluster Zugehörigkeit (SL) j_1 dem Cluster zweiter Modalität zu, zu dem auch j_1' gehört:

$$\lambda_{SL}(j_1) = \arg \max_{l' \in \{1,...,L\}} q_{j_1' l'} \quad \rightarrow \quad \hat{q}^{SL}_{j_1 l} = \begin{cases} 1, & \text{falls } l = \lambda_{SL}(j_1) \\ 0, & \text{sonst} \end{cases} .$$

Hierbei werden Items als J^N, die bereits einem zweimodalen Cluster zugeordnet sind, nicht bei der Zuordnung anderer $j \in J^N$ zu einem Item-Cluster berücksichtigt.

Das KM-Verfahren benutzt wie das $k - means$ Clusterverfahren (McQueen (1967)) die Mittelwerte

$$\bar{a}^l_\kappa = \frac{\sum_{j' \in J^B} q_{j'l} a_{j'\kappa}}{\sum_{j' \in J^B} q_{j'l}}, \quad , \quad \kappa = 1, ..., \kappa_A \quad .$$

Die Vektoren $\bar{a}^l = (\bar{a}^l_1, ..., \bar{a}^l_{\kappa_A})'$ können als Zentren der bereits bestehenden Item-Cluster $l \in \{1, ..., L\}$ interpretiert werden. Mit Hilfe des gewichteten euklidischen Abstandes zu diesen Zentren

$$\tilde{d}(j_1, \bar{a}^l) = \sqrt{\sum_{\kappa'=1}^{\kappa_A} \nu_{\kappa'} (a_{j_1 \kappa'} - \bar{a}^l_{\kappa'})^2}$$

wird dann jedes neue Item $j_1 \in J^N$ dem Cluster $l \in \{1, ..., L\}$ zugeordnet, zu dessen Zentrum $\bar{a}^l, l \in \{1, ..., L\}$ seine Eigenschaften den kleinsten euklidischen Abstand aufweisen.
Somit gilt:

$$\lambda_{km}(j_1) = \arg \min_{l' \in \{1,...,L\}} \tilde{d}(j_1, \bar{a}^{l'}) \quad \rightarrow \quad \hat{q}^{km}_{j_1 l} = \begin{cases} 1, & \text{falls } l = \lambda_{km}(j_1) \\ 0, & \text{sonst} \end{cases} .$$

Es werden zur Berechnung der Zentren $\bar{a}^l, l \in \{1, ..., L\}$, nur Elemente der Menge J^B verwendet.

Ein weiterer heuristischer Ansatz hat entfernte Ähnlichkeit mit dem Bradley-Terry-Luce Modell (Bradley, Terry (1952), Luce (1959)) für die Wahrscheinlichkeit der Wahl einer von mehreren Alternativen. Dieser Ansatz wird hier als $\bar{\alpha}$-Methode ($\bar{\alpha}$) bezeichnet. Im Rahmen der $\bar{\alpha}$-Methode wird der Grad der Zugehörigkeit eines Items $j_1 \in J^N$ zu einem bestimmten Cluster $l \in \{1, ..., L\}$ vermöge der Formel

$$\tilde{q}_{j_1 l} = \frac{\sum\limits_{j_2 \in J^B} \left(\frac{1}{d_\nu(j_1, j_2)} \right)^{\bar{\alpha}} q_{j_2 l}}{\sum\limits_{l'=1}^{L} \sum\limits_{j_2 \in J^B} \left(\frac{1}{d_\nu(j_1, j_2)} \right)^{\bar{\alpha}} q_{j_2 l'}}$$

berechnet. Die kontinuierliche Version der $\bar{\alpha}$-Methode benutzt $\tilde{q}_{j_1 l}$ direkt als Schätzer $\hat{q}_{j_1 l}$. Es empfiehlt sich, $\bar{\alpha}$=1 oder $\bar{\alpha}$=2 zu wählen.

Die diskrete Version der $\bar{\alpha}$-Methode ordnet jedes Item $j_1 \in J^N$ demjenigen Cluster zu, für das sein Grad der Zugehörigkeit maximal ist:

$$\lambda_{\bar{\alpha}}(j_1) = \arg \max_{l' \in \{1,...,L\}} \tilde{q}_{j_1 l'} \quad \rightarrow \quad \hat{q}^{\bar{\alpha}}_{j_1 l} = \begin{cases} 1, & \text{falls } l = \lambda_{\bar{\alpha}}(j_1) \\ 0, & \text{sonst} \end{cases} .$$

Beispiel 9.2:

In den folgenden Beispielen wird wieder auf die Daten aus den Beispielen 4.1 und 5.1 zurückgegriffen. Zusätzlich hierzu sind die Eigenschaften eines neuen Items $j = 9$ gegeben durch $a_9 = (7, 8, 5, 4, 3, 5)'$. Bezüglich der Items $j = 1, ..., 9$ liegen die in Tabelle 9.2 aufgeführten Kritiker-Bewertungen vor. In diesem Beispiel werden Prognosen für Bernds Bewertung des neuen Items $j = 9$ auf Basis des $\hat{S}^1_Y$- und $\hat{S}^2_Y$-Schätzers mittels des SL-Verfahrens berechnet.

$y_{i_{MC}j}^{MC}$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$
$i_{MC}=1$	-	-	4	4	4	-	-	2	2
$i_{MC}=2$	2	4	-	4	-	4	5	3	2
$i_{MC}=3$	1	4	5	-	3	4	4	-	-

Tabelle 9.2: Kritikerbewertungen (Beispiel 9.2)

Genau wie die Bewertungen der Nutzer, sind die Bewertungen der Kritiker hier ganze Zahlen von 1 bis 5. Daher ist in diesem Beispiel keine Transformation der Kritikerbewertungen notwendig. Bezüglich des neuen Items ($j=9$) liegen noch keine Bewertungen der Nutzer vor. Es ergeben sich die folgenden euklidischen Distanzen zwischen dem neuen Item und allen übrigen Items:

$d_\nu(1,9)$	$d_\nu(2,9)$	$d_\nu(3,9)$	$d_\nu(4,9)$	$d_\nu(5,9)$	$d_\nu(6,9)$	$d_\nu(7,9)$	$d_\nu(8,9)$
5,39	8,12	4,36	8,19	8,77	7,94	6,16	2,65

Tabelle 9.3: Euklidische Distanzen (Beispiel 9.2)

Das Ziel dieses Beispiels ist die Berechnung eines Schätzers $\hat{Y}_{19}$ mittels des SL-Ansatzes zur Approximation der Item-Cluster Zugehörigkeit und zweimodaler Clusterverfahren.

Item $j=8$ weist die minimale Distanz zum neuen Item auf. Daher wird Item $j=9$ im Rahmen des hier als SL-Ansatz bezeichneten Verfahrens demselben Cluster zugeordnet wie Item $j=8$. Letzteres ist in Beispiel 5.11 ein Element des zweiten Item-Clusters und Bernd gehört im Rahmen von Beispiel 5.11 zum ersten Nutzer-Cluster. Daher folgt aus den Ergebnissen von Beispiel 5.11 der Schätzwert $\hat{Y}_{19}^{\hat{S}_Y^1} = w_{12} = 1$. Die Berücksichtigung von Kritiker-Bewertungen ist hierfür nicht erforderlich.
Bezüglich des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens werden die Cluster-Zugehörigkeiten aus Beispiel 5.12 verwendet. Bernd gehört im Hinblick auf letzteres Beispiel zum zweiten Nutzer-Cluster und das Item $j=8$ wurde in Beispiel 5.12 dem ersten Cluster zweiter Modalität zugeordnet.
Auf Basis dieser Clusterzugehörigkeiten erhält man mit Hilfe des SL-Verfahrens

die Beziehung $\hat{Y}_{19}^{\hat{S}_Y^2} = w_{21} + \bar{y}_{1.} - \tilde{w}_{2.} + \bar{y}_{.9}^{MC} - \tilde{w}_{.1}^{MC}$. Aus der Tabelle 9.2 erhält man $\bar{y}_{.9}^{MC} = 2$ und $\tilde{w}_{.1}^{MC} = 2$. Die restlichen Werte können Beispiel 5.12 entnommen werden. Es ergibt sich $\hat{Y}_{19}^{\hat{S}_Y^2} = 1 + 3,5 - 3,56 + 2 - 2 = 0,94$.

Beispiel 9.3:

In diesem Beispiel wird die Cluster-Zugehörigkeit des neuen Films mittels des KM-Verfahrens bestimmt. Hierzu wird eine zweimodale Klassifikation bekannter Items vorausgesetzt.
Für das zweimodale $\hat{S}_Y^1$-Clusterverfahren aus Beispiel 5.11 ergeben sich die folgenden Zentren $\bar{a}^l, l = 1, 2, 3$:

$$\bar{a}^1_{\hat{S}^1_Y} = \begin{pmatrix} 9{,}00 \\ 6{,}00 \\ 2{,}67 \\ 1{,}67 \\ 6{,}00 \\ 8{,}33 \end{pmatrix}, \bar{a}^2_{\hat{S}^1_Y} = \begin{pmatrix} 6{,}00 \\ 8{,}50 \\ 5{,}00 \\ 5{,}50 \\ 5{,}50 \\ 3{,}50 \end{pmatrix} \quad \text{und} \quad \bar{a}^3_{\hat{S}^1_Y} = \begin{pmatrix} 3{,}67 \\ 2{,}67 \\ 3{,}67 \\ 4{,}33 \\ 3{,}33 \\ 8{,}00 \end{pmatrix}.$$

Man erhält damit $\tilde{d}_{\hat{S}^1_Y}(9, \bar{a}^1_{\hat{S}^1_Y}) = 6,24$, $\tilde{d}_{\hat{S}^1_Y}(9, \bar{a}^2_{\hat{S}^1_Y}) = 3,46$ und $\tilde{d}_{\hat{S}^1_Y}(9, \bar{a}^3_{\hat{S}^1_Y}) = 7,11$. Daher ist das neue Item dem aus den Filmen 1 und 8 bestehenden zweitem Item-Cluster zuzuordnen. Somit ergibt sich derselbe Schätzer $\hat{Y}_{19}^{\hat{S}^1_Y}$ wie in Beispiel 9.1.

Auf der Grundlage der zweimodalen Klassifikation aus Beispiel 5.12 lassen sich die folgenden Zentren bestimmen:

$$\bar{a}^1_{\hat{S}^2_Y} = \begin{pmatrix} 6{,}00 \\ 8{,}50 \\ 5{,}00 \\ 5{,}50 \\ 5{,}50 \\ 3{,}50 \end{pmatrix}, \bar{a}^2_{\hat{S}^2_Y} = \begin{pmatrix} 7{,}00 \\ 3{,}67 \\ 3{,}33 \\ 3{,}33 \\ 4{,}67 \\ 8{,}00 \end{pmatrix} \quad \text{und} \quad \bar{a}^3_{\hat{S}^2_Y} = \begin{pmatrix} 5{,}67 \\ 5{,}00 \\ 3{,}00 \\ 2{,}67 \\ 4{,}67 \\ 8{,}33 \end{pmatrix}.$$

Die Zentren differieren von denen, die auf der Grundlage der im Rahmen des Beispiels 5.11 gewonnenen Klassifikation bestimmt wurden, da in Beispiel 5.12 die Elemente zweiter Modalität nicht alle denselben Clustern zugeordnet wurden wie in Beispiel 5.11. (Man erhält die Klassifikation aus Beispiel 5.11, indem man

in der im Beispiel 5.12 hergeleiteten Klassifikation die Items 2 und 5 vertauscht.) Es ergeben sich die Distanzen des neuen Elements zweiter Modalität von den Zentren $\tilde{d}_{\hat{S}^2_Y}(9, \bar{a}^1_{\hat{S}^2_Y}) = 3,46$, $\tilde{d}_{\hat{S}^2_Y}(9, \bar{a}^2_{\hat{S}^2_Y}) = 5,81$ und $\tilde{d}_{\hat{S}^2_Y}(9, \bar{a}^3_{\hat{S}^2_Y}) = 5,52$. Daher wird das neue Item hier dem Cluster 1 aus Beispiel 5.12 zugeordnet, welches aus dem ersten und dem achten Element zweiter Modalität besteht. Folglich ergibt sich derselbe Schätzer für Y_{19} wie in Beispiel 9.2.

Beispiel 9.4:

Es geht in diesem Beispiel darum, zuerst die Klassenzugehörigkeit des neuen Items mittels der $\bar{\alpha}$-Methode auf Basis einer bereits existierenden zweimodalen Klassifikation bekannter Items abzuschätzen und auf dieser Grundlage dann den Schätzer $\hat{Y}_{19}$ zu ermitteln. Hier wird $\bar{\alpha} = 1$ verwendet.

Legt man die Klassifikation aus Beispiel 5.11 zugrunde, so ergeben sich mittels der euklidischen Distanzen aus Beispiel 9.1 die Grade der Item-Cluster Zugehörigkeit $\tilde{q}^{\hat{S}^1_Y}_{91} = 0,32, \tilde{q}^{\hat{S}^1_Y}_{92} = 0,39$ und $\tilde{q}^{\hat{S}^1_Y}_{93} = 0,29$. Daher würde man das neue Item auch auf Basis der der $\bar{\alpha}$-Methode dem zweiten Cluster aus Beispiel 5.11 (Items 1 und 8) zuordnen. Es ergibt sich daher wieder derselbe Schätzer $\hat{Y}^{\hat{S}^1_Y}_{19} = 1$.

Für die im Rahmen von Beispiel 5.12 ermittelten Cluster-Zugehörigkeiten erhält man $\tilde{q}^{\hat{S}^2_Y}_{91} = 0,39, \tilde{q}^{\hat{S}^2_Y}_{92} = 0,33$ und $\tilde{q}^{\hat{S}^2_Y}_{93} = 0,28$. Deshalb ergibt sich wieder dieselbe approximative Cluster-Zuordnung des neuen Items wie in Beispiel 9.2 und 9.3 und somit auch derselbe Schätzer $\hat{Y}^{\hat{S}^2_Y}_{19} = 0,94$.

9.2.3.2 Logistische Regression (LR)

Ein weiteres möglicherweise geeignetes Verfahren ist das Multinomiale Logit-Modell (McFadden (1974)) unter Verwendung der Item-Cluster Zugehörigkeit als endogene Variable (deren Realisationen $q^{class}_j = \max_{l' \in \{1,...,L\}} q_{jl'}, j \in J^B$, sind) und der Item-Eigenschaften als exogene Variablen (die hinsichtlich jedes Items $j \in J^B$ durch die Komponenten des Vektors $a_j = (a_{j1}, ..., a_{j\kappa_A})'$ realisiert sind). Da jedes Item $j \in J^B$ Element genau eines Clusters $C_2(1), ..., C_2(L)$ ist, ist q^{class}_j gleich dem jeweilige Wert l und daher eine wohldefinierte abhängige Größe. Die Parameter dieses Modells $\beta^l_{A,0}, \in \mathbb{R}^1, \beta^l_A \in \mathbb{R}^{\kappa_A}, l \in \{1, ..., L\}$, können mittels eines Maximum-Likelihood Ansatzes auf Basis der hinsichtlich der bezüglich der bekannten Items $j \in J^B$ vorhandenen Bewertungen geschätzt werden. Mittels der resultierenden Schätzer $\hat{\beta}^l_{A,0}, \in \mathbb{R}^1, \hat{\beta}^l_A \in \mathbb{R}^{\kappa_A}, l \in \{1, ..., L\}$, lassen sich bezüglich

aller neuen Items $j \in J^N$ die geschätzten Wahrscheinlichkeit der Zugehörigkeit zu einem bestimmten Item-Cluster $l \in \{1, ..., L\}$ auf Basis der entsprechenden Eigenschaften a_j mit dem Ansatz

$$\widehat{P}(j \in C_2(l)|a_j) = \frac{\exp(\hat{\beta}^l_{A,0} + a'_j\hat{\beta}^l_A)}{\sum_{l'=1}^{L} \exp(\hat{\beta}^{l'}_{A,0} + a'_j\hat{\beta}^{l'}_A)}, \quad l = 1, ..., L,$$

berechnen. (Da die Logistische Regression bereits hinlänglich bekannt ist, wird hier nicht näher auf diese Methode eingegangen.) Auf diese Weise lassen sich Wahrscheinlichkeiten für die Zugehörigkeit zu einem Cluster berechnen. Diese Wahrscheinlichkeiten lassen sich auf dieselbe Weise, auf die man im Rahmen der $\bar{\alpha}$-Methode den Grad der Zugehörigkeit eines neuen Items dazu benutzt hat, es einem Cluster zuzuordnen, dazu verwenden, die Schätzer $\hat{q}_{jl}, l \in \{1, ..., L\}$ für die neuen Items $j \in J^N$ zu bestimmen. Zur Berechnung wurde die TDA-Software (Rohwer, Pötter (2005)) benutzt.

9.3.2.3 Entscheidungsbäume ($C4.5$)

Außerdem eignen sich Entscheidungsbaum-Verfahren wie der C4.5 Algorithmus (Quinlan (1993)) zum Lösen von Klassifikationsproblemen.

Jedem Entscheidungsbaum liegt ein gerichteter azyklischer Graph zugrunde. Zu Beginn werden die gesamten zur Bestimmung des Baumes verfügbaren Trainingsdaten $(q_j^{class}, a_j), j \in J^B$, einem einzigen Knoten (der Wurzel des zu generierenden Entscheidungsbaumes) zugeordnet. Danach wird in jedem Schritt die dem betrachteten Knoten zugeordnete Daten(teil)menge so in Untermengen aufgeteilt, daß das zugehörige Auswahlmaß (beim $C4.5$ der relative Informationsgewinn) maximal ist. Ist das Auswahlmaß oberhalb eines a priori festgelegten Schwellenwerts, so wird jeder der Untermengen ein neuer Knoten zugeordnet. Andernfalls wird die Datenmenge nicht aufgeteilt und bleibt dem betrachteten Knoten zugeteilt, der dann nicht weiter bearbeitet wird. Nachdem aufgrund des Schwellenwerts keine weiteren Aufteilungen der noch nicht aufgeteilten Knoten mehr möglich sind, wird ein sogenannter Pruning-Schritt durchgeführt, der weniger überzeugende Aufteilungen rückgängig macht, um eine Überanpassung an die verwendeten Trainingsdaten zu verhindern. Die Item-Cluster Zugehörigkeit der neuen Items läßt sich dann bestimmen, indem man den auf Basis des Trainingsdatensatzes erzeugten Entscheidungsbaum-Algorithmus auf die Eigenschaften der

neuen Items $a_j, j \in J^N$, anwendet. Da auch der $C4.5$-Algorithmus mittlerweile zu einem Standardinstrumentarium der Datenanalyse geworden ist, wird auch auf eine nähere Darstellung des $C4.5$-Algorithmus verzichtet. Die Berechnung erfolgte mittels der Weka-Software (Witten, Frank (2005)).

9.3.2.4 Bayes'sche Multinetzwerke (BMN)

Bayes'sche Multinetzwerke (BMN) sind ein weiterer möglicherweise hilfreicher (wenngleich weniger bekannter) Machine Learning Algorithmus, der auf Basis der vorhandenen Daten bezüglich der bekannten Items J^B trainiert werden kann, um die Cluster-Zugehörigkeit der neuen Items abzuschätzen.

Da Bayes'sche Multinetzwerke aus (lokalen) Bayes'schen Netzwerken bestehen, ist es notwendig, zuerst die Bayes'schen Netzwerke zu erklären. Jedes Bayes'sche Netzwerk $(\mathfrak{B}, \Theta) = (\mathfrak{X}, \mathfrak{E}, \Theta)$ besteht aus einem gerichteten azyklischen Graphen $\mathfrak{B} = (\mathfrak{X}, \mathfrak{E})$, dessen Knoten $\mathfrak{X}_1, ..., \mathfrak{X}_{M_\mathfrak{B}}$ $(\mathfrak{X} = \{\mathfrak{X}_1, ..., \mathfrak{X}_{M_\mathfrak{B}}\})$ Zufallsvariablen repräsentieren können und dessen Kanten $\mathfrak{E}$ die bedingten Abhängigkeiten zwischen den Knoten darstellen.

Seien alle $\mathfrak{X}_1, ..., \mathfrak{X}_{M_\mathfrak{B}}$ diskrete Zufallsvariablen und sei $\pi_\mathfrak{B}(\mathfrak{X}_{\bar{\mu}})$ die Menge aller Knoten, von denen in $\mathfrak{B}$ eine Kante zu $\mathfrak{X}_{\bar{\mu}}$ führt. Die Menge der Knoten $\pi_\mathfrak{B}(\mathfrak{X}_{\bar{\mu}})$ werden auch als „Eltern" des Knoten $\mathfrak{X}_{\bar{\mu}}$ bezeichnet. $\Theta_{\bar{\mu}} = P(\mathfrak{X}_{\bar{\mu}}|\pi_\mathfrak{B}(\mathfrak{X}_{\bar{\mu}}))$ ist dann die bedingte Wahrscheinlichkeitsfunktion des Knoten $\mathfrak{X}_{\bar{\mu}}, \bar{\mu} = 1, ..., M_\mathfrak{B}$. Es gilt $\Theta = \{\Theta_1, ..., \Theta_{M_\mathfrak{B}}\}$. Daher kodiert jedes Bayes'sche Netzwerk die gemeinsame Wahrscheinlichkeitsverteilung der (diskreten) Zufallsvariablen $\mathfrak{X}_1, ..., \mathfrak{X}_{M_\mathfrak{B}}$. $M_\mathfrak{B}$ ist die Anzahl der Knoten von $\mathfrak{B}$.

Weil Bayes'sche Netzwerke azyklisch sind, kann man ihre Knoten ordnen, so daß alle Vorfahren eines Knoten kleinere Indizes als der betreffende Knoten haben. Daraus folgt für die gemeinsame Wahrscheinlichkeitsfunktion

$$P(\mathfrak{X}_1, ..., \mathfrak{X}_{M_\mathfrak{B}}) = \prod_{\bar{\mu}'=1}^{M_\mathfrak{B}} P(\mathfrak{X}_{\bar{\mu}'}|\mathfrak{X}_{\bar{\mu}'-1}, ..., \mathfrak{X}_1).$$

Außerdem wird angenommen, daß nur die „Eltern" eines Knoten seine bedingte Wahrscheinlichkeitsfunktion beeinflussen:

$$P(\mathfrak{X}_{\bar{\mu}}|\mathfrak{X}_{\bar{\mu}-1}, ..., \mathfrak{X}_1) = P(\mathfrak{X}_{\bar{\mu}}|\pi_\mathfrak{B}(\mathfrak{X}_{\bar{\mu}})), \bar{\mu} = 1, ..., M_\mathfrak{B}.$$

Damit ergibt sich

$$\log P(\mathfrak{X}_1, ..., \mathfrak{X}_{M_{\mathfrak{B}}}) = \sum_{\bar{\mu}'=1}^{M_{\mathfrak{B}}} \log P(\mathfrak{X}_{\bar{\mu}'} | \pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}'})). \tag{9.1}$$

Deshalb kann $P(\mathfrak{X}_{\bar{\mu}}|\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}}))$ für jeden einzelnen Knoten $\mathfrak{X}_{\bar{\mu}}$ des Netzwerks getrennt optimiert werden ($\bar{\mu} = 1, ..., M_{\mathfrak{B}}$). Dies ist der Grundgedanke der bekanntesten Algorithmen zum Erlernen Bayes'scher Netzwerke. Falls die diskrete Zufallsvariable $\mathfrak{X}_1$ die Item-Cluster Zugehörigkeit und die übrigen Zufallsvariablen $\mathfrak{X}_2, ..., \mathfrak{X}_{M_{\mathfrak{B}}}$ die Item-Eigenschaften sind ($M_{\mathfrak{B}} = 1+\kappa_A$), kann ein Bayes'sches Netzwerk zur Klassifikation von Items anhand ihrer Eigenschaften eingesetzt werden. Dann sind $x_{j1} = q_j^{class}, x_{j2} = a_{j1}, ..., x_{jM_{\mathfrak{B}}} = a_{j\kappa_A}, j \in J^B$, die Realisationen von $\mathfrak{X}_1, ..., \mathfrak{X}_{M_{\mathfrak{B}}}$. Hierzu muß zuerst $\mathfrak{B}$ anhand dieses Trainingsdatensatzes bestimmt werden. Dies bedeutet, daß ermittelt werden muß, welche Elternknoten $\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}})$ jedem Knoten $\mathfrak{X}_{\bar{\mu}}, \bar{\mu} = 1, ..., M_{\mathfrak{B}}$ zugeordnet sind. Dies kann durch den sogenannten K2-Algorithmus nach Cooper, Herskovits (1992) erreicht werden. Cooper, Herskovits (1992) haben bewiesen, daß unter der Voraussetzung, daß die einzelnen Realisationen $x_{j1}, ..., x_{jM_{\mathfrak{B}}}$ von $\mathfrak{X}_1, ..., \mathfrak{X}_{M_{\mathfrak{B}}}$ diskret, vollständig und voneinander unabhängig sind, gilt

$$P(\mathfrak{B}, (x_{j1}, ..., x_{jM_{\mathfrak{B}}}), j \in J^B) = P(\mathfrak{B}) \prod_{\bar{\mu}'=1}^{M_{\mathfrak{B}}} g_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}'}, \pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}'}))$$

mit

$$g_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}'}, \pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}'})) = \begin{cases} \prod\limits_{\bar{x}' \in \mathfrak{M}(\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}}))} \dfrac{(|\mathfrak{M}(\mathfrak{X}_{\bar{\mu}'})| - 1)!}{(\mathfrak{V}_{\bar{\mu}'\bar{x}'} + |\mathfrak{M}(\mathfrak{X}_{\bar{\mu}'})| - 1)!} \\ \qquad \cdot \prod\limits_{\bar{m}' \in \mathfrak{M}(\mathfrak{X}_{\bar{\mu}})} \alpha^{\mathfrak{B}}_{\bar{\mu}'\bar{x}'\bar{m}'}!, & \text{für } \pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}'}) \neq \emptyset \\ \dfrac{(|\mathfrak{M}(\mathfrak{X}_{\bar{\mu}'})| - 1)!}{(\mathfrak{V}_{\bar{\mu}'} + |\mathfrak{M}(\mathfrak{X}_{\bar{\mu}'})| - 1)!} \prod\limits_{\bar{m}' \in \mathfrak{M}(\mathfrak{X}_{\bar{\mu}})} \breve{\alpha}^{\mathfrak{B}}_{\bar{\mu}'\bar{m}'}!, & \text{für } \pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}'}) = \emptyset \end{cases}.$$

Hier ist $\mathfrak{M}(\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}}))$ die Menge aller in den Daten vorhandenenen unterschiedlichen Kombinationen von Realisationen der Elemente der Menge $\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}})$. (Sei beispielsweise $\{\mathfrak{X}_{\bar{\mu}_1}, \mathfrak{X}_{\bar{\mu}_2}\} = \pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}})$, dann sind alle unterschiedlichen Kombinationen $(x_{j\bar{\mu}_1}, x_{j\bar{\mu}_2})$, $j \in J^B$, Elemente der Menge $\mathfrak{M}(\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}}))$.) $\mathfrak{M}(\mathfrak{X}_{\bar{\mu}})$ bezeich-

net die Menge der Werte, die $\mathfrak{X}_{\bar{\mu}}$ bezüglich der gegebenen Daten annimmt. Es gilt $\bar{x} \in \mathfrak{M}(\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}}))$ und $\bar{m} \in \mathfrak{M}(\mathfrak{X}_{\bar{\mu}})$. Weiter ist $\alpha^{\mathfrak{B}}_{\bar{\mu}\bar{x}\bar{m}}$ die Anzahl der Fälle bezüglich der Items aus J^B bei denen $\mathfrak{X}_{\bar{\mu}}$ den Wert $\bar{m}$ annimmt und darüberhinaus die Knoten aus $\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}})$ die Wertekombination $\bar{x}$ realisieren. Analog gilt $\breve{\alpha}^{\mathfrak{B}}_{\bar{\mu}\bar{m}} = |\{j \in J^B | x_{j\bar{\mu}} = \bar{m}\}|$. $\mathfrak{V}_{\bar{\mu}\bar{x}}$ sei die Anzahl der Fälle, in denen ein Elternknoten von $\mathfrak{X}_{\bar{\mu}}$ den Wert $\bar{x}$ annimmt. Außerdem ist wegen der vorausgesetzten Vollständigkeit der Daten $\mathfrak{V}_{\bar{\mu}}$ gleich der Anzahl der im Hinblick auf das betreffende Bayes'sche Netzwerk betrachteten Daten. Der zu bestimmende gerichtete azyklische Graph $\mathfrak{B}$ wird im jeweils n-ten Schritt des Algorithmus durch $\mathfrak{B}(n) = (\mathfrak{X}, \mathfrak{E}(n))$ ersetzt. Der K2-Algorithmus verwendet die Unabhängigkeit der Knoten $\mathfrak{X}_{\bar{\mu}}$ von allen übrigen Knoten $\mathfrak{X}_{\bar{\mu}} \in \mathfrak{X} \backslash \pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}})$ (vgl. Formel (9.1)). Es wird in diesem Zusammenhang davon ausgegangen, daß es - auch bevor $\mathfrak{B}$ bestimmt ist - möglich ist, die Variablen $\mathfrak{X}_1, ..., \mathfrak{X}_{1+\kappa_A}$ so zu ordnen, daß alle möglichen Eltern $\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}})$ einer Variable $\mathfrak{X}_{\bar{\mu}}$ kleinere Indizes als $\bar{\mu}$ haben. Im folgenden wird vorausgesetzt, daß $\mathfrak{X}_{\bar{\mu}}$ bereits eine auf diese Weise geordnete Menge ist. Die Menge $Pred(\mathfrak{X}_{\bar{\mu}}|\mathfrak{X}) = \{\mathfrak{X}_{\bar{\mu}-1}, ..., \mathfrak{X}_1\}$ enthält dann alle Knoten aus der Menge $\mathfrak{X}$, die Eltern des Knoten $\mathfrak{X}_{\bar{\mu}}$ sein könnten. Im Rahmen des K2-Algorithmus wird für jeden einzelnen Knoten $\mathfrak{X}_{\bar{\mu}}$ die zugehörige Elternmenge bestimmt. Man beginnt für jeden Knoten $\mathfrak{X}_{\bar{\mu}}$ mit einer leeren Menge von Elternknoten. Man bestimmt zunächst in jedem n-ten Schritt die Menge $M_n^{cand}(\mathfrak{X}_{\bar{\mu}}) = Pred(\mathfrak{X}_{\bar{\mu}}|\mathfrak{X}) \setminus \pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}})$. Sodann bestimmt man den Knoten

$$\mathfrak{X}_{\underline{\mu}} = \arg \max_{\mathfrak{X}_{\mu^*} \in M_n^{cand}(\mathfrak{X}_{\bar{\mu}})} g_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}} | \pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}}) \cup \{\mathfrak{X}_{\mu^*}\}),$$

wobei $g_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}}, \pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}}))$ völlig analog zu $g_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}}, \pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}}))$ zu bilden ist. (Bei der Bildung beider Funktionen sind nur die Kanten aus $\mathfrak{B}$ ($\mathfrak{B}(n)$) zu berücksichtigen, die $\mathfrak{X}_{\bar{\mu}}$ mit Elementen aus der Menge $\pi_{\mathfrak{B}}(\mathfrak{X}_{\bar{\mu}})$ verbinden.)

Falls $g_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}} | \pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}}) \cup \{\mathfrak{X}_{\underline{\mu}}\}) > g_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}} | \pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}}))$ gilt, wird $\mathfrak{X}_{\underline{\mu}}$ der Elternmenge im n-ten Schritt $\pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}})$ hinzugefügt und man erhält dadurch die neue Elternmenge $\pi_{\mathfrak{B}(n+1)}(\mathfrak{X}_{\bar{\mu}}) = \pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}}) \cup \{\mathfrak{X}_{\underline{\mu}}\}$. Sobald sich im Hinblick auf den betrachteten Knoten $\mathfrak{X}_{\bar{\mu}}$ die Elternmenge bezüglich $\mathfrak{B}(n)$ nicht mehr vergrößern läßt oder die Anzahl der Elternknoten eine a priori festgesetzte Anzahl E^{max} erreicht, betrachtet man die Elternmenge des jeweiligen Knoten als vollständig und wendet sich einem anderen Knoten zu. Dessen Elternmenge wird analog bestimmt. Die Heuristik ist zu Ende, sobald für alle Knoten Elternmen-

Startwerte: $n = 0$, $\mathfrak{B}(0) = \{\mathfrak{X}, \mathfrak{E}(0)\}$, $\mathfrak{E}(0) = \emptyset$, $\bar{\mu}' = \bar{\mu}_{min}$
Solange $(\bar{\mu}' \leq \kappa_A)$ {
$v^{ctrl} = 0$
$\pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'}) = \emptyset$
Solange $(v^{ctrl} = 0 \wedge |\pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'})| < E^{max})$ {
$M_n^{cand}(\mathfrak{X}_{\bar{\mu}'}) = Pred(\mathfrak{X}_{\bar{\mu}'}|\mathfrak{X}) \setminus \pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'})$
$\mathfrak{X}_{\underline{\mu}} = \arg \max_{\mathfrak{X}_{\mu^*} \in M_n^{cand}(\mathfrak{X}_{\bar{\mu}'})} g_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'}|\pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'}) \cup \{\mathfrak{X}_{\mu^*}\})$
Falls $(g_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'}|\pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'}) \cup \{\mathfrak{X}_{\underline{\mu}}\}) > g_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'}|\pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'})))$ {
$\pi_{\mathfrak{B}(n+1)}(\mathfrak{X}_{\bar{\mu}'}) = \pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'}) \cup \{\mathfrak{X}_{\underline{\mu}}\}$
}
Sonst {
$\pi_{\mathfrak{B}(n+1)}(\mathfrak{X}_{\bar{\mu}'}) = \pi_{\mathfrak{B}(n)}(\mathfrak{X}_{\bar{\mu}'})$
$v^{ctrl} = 1$
}
$n \leftarrow n + 1$
}
$\bar{\mu}' \leftarrow \bar{\mu}' + 1$
}

Abbildung 9.1: K2-Algorithmus nach Cooper, Herskovits (1992)

gen bestimmt sind. Der K2-Algorithmus ist in Abbildung 9.1 dargestellt. $\bar{\mu}_{min}$ bezeichnet den kleinsten Index $\bar{\mu}$ bezüglich der betrachteten Knotenmenge.

Unabhängig vom Wert, den $\mathfrak{X}_1$ annimmt, wird ein einmal anhand eines spezifischen Trainingsdatensatzes erstelltes Bayes'sches Netzwerk dieselben bedingten Abhängigkeiten (bzw. Unabhängigkeiten) zwischen den Zufallsvariablen wiederspiegeln. Das wird auch dann der Fall sein, wenn für bestimmte Werte von $\mathfrak{X}_1$ ein Teil dieser bedingten Abhängigkeitsbeziehungen im Trainingsdatensatz nicht existiert (Geiger, Heckerman (1996)). Sofern es sich bei den Items beispielsweise um Filme handelt, könnte es durchaus sein, daß für bestimmte Film-Cluster, die hauptsächlich aus Horror-, Action- und Abenteuerfilmen bestehen, andere Abhängigkeitsbeziehungen existieren (und in den Trainingsdaten zu Tage treten) als in Film-Clustern, deren Elemente hauptsächlich Komödien, Romanzen, Dramen und Dokumentarfilme sind. Bayes'sche Multinetzwerke (Geiger, Heckerman (1996)) erlauben die Modellierung verschiedener Abhängigkeitsstrukturen hin-

sichtlich unterschiedlicher Realisationen von $\mathfrak{X}_1$. Ein Bayes'sches Multinetzwerk besteht aus L Bayes'schen Netzwerken $\mathfrak{B}_l$, die jeweils die gemeinsame Wahrscheinlichkeitsverteilung von $\{\mathfrak{X}_2, ..., \mathfrak{X}_{M_{\mathfrak{B}}}\} \equiv \mathfrak{X}(\mathfrak{B}_l)$ unter der Voraussetzung $\mathfrak{X}_1 = l$ kodieren, und den Wahrscheinlichkeiten $P(\mathfrak{X}_1 = l), l = 1, ..., L$. Für jede Realisation $l \in \{1, ..., L\}$ von $\mathfrak{X}_1$ wird separat das Bayes'sches Netzwerk $\mathfrak{B}_l$ hinsichtlich der bedingten Wahrscheinlichkeitsfunktion $P(\mathfrak{X}_2, ..., \mathfrak{X}_{M_{\mathfrak{B}}}|\mathfrak{X}_1 = l)$ auf Basis aller Daten $(q_j^{class}, a_j), j \in J^B$, für die $q_j^{class} = l$ ist, bestimmt. Auf dieser Grundlage ergibt sich

$$\hat{P}(\mathfrak{X}_1 = l|\mathfrak{X}_2, ..., \mathfrak{X}_{1+\kappa_A}) = \frac{\hat{P}(\mathfrak{X}_2, ..., \mathfrak{X}_{1+\kappa_A}|\mathfrak{X}_1 = l)\hat{P}(\mathfrak{X}_1 = l)}{\sum_{l'=1}^{L} \hat{P}(\mathfrak{X}_2, ..., \mathfrak{X}_{1+\kappa_A}|\mathfrak{X}_1 = l')\hat{P}(\mathfrak{X}_1 = l')},$$

für $l = 1, ..., L$. Hier bedeutet die Schreibweise $\mathfrak{X}_1 = l$, daß $\mathfrak{X}_1$ den Wert l annimmt. Es gilt

$$\hat{P}(\mathfrak{X}_1 = l|\mathfrak{X}_2, ..., \mathfrak{X}_{1+\kappa_A}) = \frac{\prod_{\bar{\mu}'=2}^{1+\kappa_A} \hat{P}(\mathfrak{X}_{\bar{\mu}'}|\pi_{\mathfrak{B}l}(\mathfrak{X}_{\bar{\mu}'}), \mathfrak{X}_1 = l)\hat{P}(\mathfrak{X}_1 = l)}{\sum_{l'=1}^{L} \prod_{\bar{\mu}'=2}^{1+\kappa_A} \hat{P}(\mathfrak{X}_{\bar{\mu}'}|\pi_{\mathfrak{B}l'}(\mathfrak{X}_{\bar{\mu}'}), \mathfrak{X}_1 = l')\hat{P}(\mathfrak{X}_1 = l')},$$

wobei wieder $l = 1, ..., L$ gilt. Friedman et. al. (1997) konnten empirisch belegen, daß Bayes'sche Multinetzwerke zu besseren Klassifikationen als Naive-Bayes Verfahren und Bayes'sche Netzwerke führen und überdies auch im Vergleich mit dem C4.5 Algorithmus gute Ergebnisse liefern.

Beispiel 9.5:

Es werden wieder dieselben Daten wie in den Beispielen 9.2, 9.3 und 9.4 betrachtet. Im Unterschied zu den bisherigen Beispielen werden aber nur die Eigenschaften $\kappa = 1, 2, 3, 4$ benutzt. Es werden die transformierten Daten

$$a_{j\kappa}^0 = \begin{cases} 1, & \text{falls } a_{j\kappa} > \frac{1}{J}\sum_{j'=1}^{J} a_{j'\kappa} \\ 0, & \text{sonst} \end{cases}$$

verwendet. Diese werden auf Basis der in Beispiel 5.11 mittels des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens bestimmten Klassifikation unterteilt:

$a_{j\kappa}^0$	$l=1$			$l=2$		$l=3$		
	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$\kappa=1$	1	1	1	0	0	0	0	0
$\kappa=2$	0	0	1	1	1	0	0	0
$\kappa=3$	0	0	0	1	1	1	1	0
$\kappa=4$	1	0	0	1	1	1	1	1

Tabelle 9.4: Transformierte Eigenschaften (Beispiel 9.5)

In diesem Beispiel soll die Bewertung des ersten Nutzers hinsichtlich des neuen Items ($j = 9$) mittels eines Bayes'schen Multinetzwerks geschätzt werden. Die Eigenschaft κ wird repräsentiert durch den Knoten $\mathfrak{X}_{\kappa+1}$. $\mathfrak{X}_1$ repräsentiert die Cluster-Zugehörigkeit. Hier ist $\mathfrak{X}(\mathfrak{B}_l) = \{\mathfrak{X}_2, ..., \mathfrak{X}_5\}$. Auf Basis von Tabelle 9.4 ergeben sich bezüglich des ersten Item-Clusters ($\mathfrak{X}_1 = 1$) im Hinblick auf $\mathfrak{X}_2$ die Werte $\mathfrak{V}_2 = 3, \breve{\alpha}_{20}^{\mathfrak{B}1} = 3, \breve{\alpha}_{21}^{\mathfrak{B}1} = 0, |\mathfrak{M}(\mathfrak{X}_2)| = 1$. Wegen $Pred(\mathfrak{X}_2|\mathfrak{X}(\mathfrak{B}_1)) = \emptyset$ hat $\mathfrak{X}_2$ bezüglich $\mathfrak{X}(\mathfrak{B}_1)$ keine Eltern. Hinsichtlich $\mathfrak{X}_3$ erhält man

$$g_{\mathfrak{B}1}(\mathfrak{X}_3|\emptyset) = \frac{(|\mathfrak{M}(\mathfrak{X}_3)|-1)!}{(\mathfrak{V}_3+|\mathfrak{M}(\mathfrak{X}_3)|-1)!}\breve{\alpha}_{30}^{\mathfrak{B}1}!\breve{\alpha}_{31}^{\mathfrak{B}1}! = \frac{(2-1)!}{(3+2-1)!}2!1! = \frac{1}{12}.$$

Weil $Pred(\mathfrak{X}_3|\mathfrak{X}(\mathfrak{B}_1)) = \{\mathfrak{X}_2\}$ ist, ergibt sich

$$g_{\mathfrak{B}1}(\mathfrak{X}_3|\mathfrak{X}_2) = \frac{(|\mathfrak{M}(\mathfrak{X}_3)|-1)!}{(\mathfrak{V}_{31}+|\mathfrak{M}(\mathfrak{X}_3)|-1)!}\alpha_{310}^{\mathfrak{B}1}!\alpha_{311}^{\mathfrak{B}1}! = \frac{(2-1)!}{(3+2-1)!}2!1! = \frac{1}{12}.$$

Da $g_{\mathfrak{B}1}(\mathfrak{X}_3|\mathfrak{X}_2)$ nicht größer als $g_{\mathfrak{B}1}(\mathfrak{X}_3|\emptyset)$ ist, hat auch $\mathfrak{X}_3$ keine Eltern in $\mathfrak{B}_1$. Auf diese Weise untersucht man jeden Knoten von $\mathfrak{B}_1, \mathfrak{B}_2$ und $\mathfrak{B}_2$ bezüglich der Daten aus dem jeweils relevanten Cluster. Weil durch den K2-Algorithmus für keinen dieser Knoten Eltern gefunden werden, sind für die Berechnung der geschätzten Wahrscheinlichkeit der Zugehörigkeit eines Items zu einem bestimmten Item-Cluster nur die Werte $\hat{\Theta}_{\bar{\mu}\bar{m}}^l = \hat{P}(\mathfrak{X}_{\bar{\mu}} = \bar{m}|\mathfrak{X}_1 = l), \bar{\mu} = 2, ..., 5, \bar{m} = 0, 1, l = 1, 2, 3,$

und $\hat{\Theta}^l = \hat{P}(\mathfrak{X}_1 = l), l = 1, 2, 3$, erforderlich. Auf Basis der Werte in Tabelle 9.4 ergeben sich die in Tabelle 9.5 aufgeführten Werte:

$\hat{\Theta}^l_{\bar{\mu}1}$ $(\hat{\Theta}^l_{\bar{\mu}0})$	$\bar{\mu} = 2$	$\bar{\mu} = 3$	$\bar{\mu} = 4$	$\bar{\mu} = 5$
$l = 1$	1 (0)	$\frac{1}{3}$ $(\frac{2}{3})$	0 (1)	$\frac{1}{3}$ $(\frac{2}{3})$
$l = 2$	0 (1)	1 (0)	1 (0)	1 (0)
$l = 3$	0 (1)	0 (1)	$\frac{2}{3}$ $(\frac{1}{3})$	1 (0)

Tabelle 9.5: Geschätzte Wahrscheinlichkeiten (Beispiel 9.5)

Außerdem gilt $\hat{\Theta}^1 = \hat{\Theta}^3 = \frac{3}{8}$ und $\hat{\Theta}^2 = \frac{1}{4}$. Bezüglich der ersten 4 Merkmale des 9-ten Items gilt $a^0_{91} = 0$ und $a^0_{92} = a^0_{93} = a^0_{94} = 1$. Deshalb gilt:

$$\hat{P}(\mathfrak{X}_1 = l|\mathfrak{X}_2 = 0, \mathfrak{X}_3 = 1, \mathfrak{X}_4 = 1, \mathfrak{X}_5 = 1) = \frac{\hat{\Theta}^l_{20}\hat{\Theta}^l_{31}\hat{\Theta}^l_{41}\hat{\Theta}^l_{51}\hat{\Theta}^l}{\sum_{l'=1}^{L}\hat{\Theta}^{l'}_{20}\hat{\Theta}^{l'}_{31}\hat{\Theta}^{l'}_{41}\hat{\Theta}^{l'}_{51}\hat{\Theta}^{l'}}, \quad l = 1, ..., L.$$

Wegen

$$\begin{aligned}
\hat{\Theta}^1_{20}\hat{\Theta}^1_{31}\hat{\Theta}^1_{41}\hat{\Theta}^1_{51}\hat{\Theta}^1 &= 0 \cdot \tfrac{1}{3} \cdot 0 \cdot \tfrac{1}{3} \cdot \tfrac{3}{8} = 0 \\
\hat{\Theta}^2_{20}\hat{\Theta}^2_{31}\hat{\Theta}^2_{41}\hat{\Theta}^2_{51}\hat{\Theta}^2 &= 1 \cdot 1 \cdot 1 \cdot 1 \cdot \tfrac{1}{4} = \tfrac{1}{4} \\
\hat{\Theta}^3_{20}\hat{\Theta}^3_{31}\hat{\Theta}^3_{41}\hat{\Theta}^3_{51}\hat{\Theta}^3 &= 1 \cdot 0 \cdot \tfrac{2}{3} \cdot 1 \cdot \tfrac{3}{8} = 0
\end{aligned}$$

ist das 9-te Item dem zweiten Item-Cluster (aus Beispiel 5.11) zuzuordnen. Daher erhält man auch mittels des Bayes'schen Multinetzes $\hat{Y}^{\hat{S}^1_Y}_{19} = w_{12} = 1$.

Die Klassifikation anhand des Bayes'schen Multinetzwerkes wurde mit Hilfe des BN PowerConstructors (Cheng (2006)) durchgeführt.

Logistischer Modellbaum (LMT)

Die Logistischen Modellbäume (LMT) wurden von Landwehr et. al. (2005) entwickelt. Ziel dieses Verfahrens ist die Verbindung von Logistischer Regression mit Entscheidungsbaumverfahren. Die Idee des Verfahrens ist, die Datenmenge zunächst durch einen möglichst groben Entscheidungsbaum aufzuteilen und

dann auf den zu den Knoten gehörenden Teilmengen logistische Regressionen auszuführen. Vorhersagen für unbekannte Items können gemacht werden, indem zunächst aufgrund ihrer Eigenschaften mittels des Entscheidungsbaums der zugehörige logistische Regressionsansatz identifiziert wird. Der betreffende logistische Regressionsansatz wird dann in Verbindung mit den Eigenschaften dazu benutzt, die Wahrscheinlichkeiten zu berechnen, mit denen das betrachtete neue Item zu den einzelnen Clustern gehört. Das unbekannte Item wird dann dem Cluster zugeordnet, zu dem es am wahrscheinlichsten gehört.

Zu Beginn werden die zur Verfügung stehenden Daten (die bekannten Items J^B) in eine Strukturerzeugungsmenge J^S und eine Kalibrierungsmenge J^{cal} unterteilt ($J^B = J^S \cup J^{cal} \wedge J^S \cap J^{cal} = \emptyset$). Basierend auf J^S erzeugt man eine Entscheidungsbaumstruktur durch Maximieren des Informationsgewinnverhältnisses. Das Informationsgewinnverhältnis ist das Auswahlmaß, welches vom C4.5-Algorithmus (Quinlan (1993)) verwendet wird. Zusätzlich werden nur Knoten weiter aufgeteilt, zu denen noch mindestens 15 Items gehören und zu den resultierenden Knoten jeweils mindestens 2 Items gehören. Durch diesen noch unbeschnittenen Entscheidungsbaum werden die zur Verfügung stehenden Trainingsdaten an jedem Knoten des Baums je nach Ausprägung der dem Knoten entsprechenden Eigenschaft weiter in disjunkte Teilmengen unterteilt. Der Wurzel $\mathfrak{X}_{\mu_1}$ eines Entscheidungsbaumes ist die gesamte zur Erzeugung dieser Entscheidungsbaumstruktur verwendete Datenmenge $J[\mathfrak{X}(\mu_1)] = J^S$ zugeordnet. Sei $J[\mathfrak{X}(\mu_1)] = J^S$ und seinen weiter $\mu_2, ..., \mu_{\underline{n}}$ Eigenschaften aus $\{1, ..., \kappa_{MAX}\}$. (κ_{MAX} ist die Anzahl aller verfügbaren Variablen.) $A_{s_{\underline{n}}}(\mu_{\underline{n}})$ bezeichne das $s_{\underline{n}}$-te Intervall bezüglich der Eigenschaft $\mu_{\underline{n}}, \underline{n} = 2, ..., \underline{N}, \underline{N} \leq \kappa_{MAX}$. Bezüglich aller Eigenschaften $\mu_{\underline{n}} \in \{1, ..., \kappa_{MAX}\}$, ist die Vereinigung $A_1(\mu_{\underline{n}}) \cup A_2(\mu_{\underline{n}}) \cup ... \cup A_{\bar{\mathfrak{s}}}(\mu_{\underline{n}})$ der gesamte Bereich der Werte, der hinsichtlich die Eigenschaft $\mu_{\underline{n}}$ angenommen werden kann. Es gilt $s_{\underline{n}} = 1, ..., \bar{\mathfrak{s}}$. (Meistens wird $\bar{\mathfrak{s}} = 2$ verwendet.) Außerdem benutzt man $\underline{a}_j = (\underline{a}_{j1}, ..., \underline{a}_{j\kappa_{MAX}})'$. Dieser Vektor kann zusätzlich zu den Komponenten von a_j noch weitere Eigenschaften des Items j enthalten. Es ergeben sich die Mengen $J[\mathfrak{X}(\mu_1, \mu_2^{s_2})] = \{j \in J[\mathfrak{X}(\mu_1)] | \underline{a}_{j\mu_2} \in A_{s_2}(\mu_2)\}$, $s_2 = 1, ..., \bar{\mathfrak{s}}$, auf Basis von $J[\mathfrak{X}(\mu_1)]$.

Auf diese Weise wird durch den Entscheidungsbaum für jeden aufteilbaren Knoten $\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}-1}^{s_{\underline{n}-1}})$ die zugehörige Datenmenge $J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}-1}^{s_{\underline{n}-1}})]$ auf Grundlage einer bestimmten Eigenschaft $\mu_{\underline{n}} \in \{2, ..., \kappa_A\} \setminus \{\mu_1, ..., \mu_{\underline{n}-1}\}$ in weitere $\bar{\mathfrak{s}}$ Teilmengen $J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}-1}^{s_{\underline{n}-1}}, \mu_{\underline{n}}^{s_{\underline{n}}})] = \{j \in J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}-1}^{s_{\underline{n}-1}})] | \underline{a}_{j\mu_{\underline{n}}} \in A_{s_{\underline{n}}}(\mu_{\underline{n}})\}$,

$s_{\underline{n}} \in \{1, ..., \bar{\mathfrak{s}}\}$, unterteilt, wobei jede dieser Teilmengen jeweils dem entsprechenden Knoten $\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_{\underline{n}}})$, $s_{\underline{n}} = 1, ..., \bar{\mathfrak{s}}$, zugeordnet wird. Die vereinfachende Notation $J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}-1}^{s_{\underline{n}-1}}, \mu_{\underline{n}}^{s_{\underline{n}}})] = J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_{\underline{n}}})]$ bezeichnet eine anhand von $\underline{n}$ verschiedenen Schritten (Variablen) bestimmte Datenteilmenge. Alle Elemente dieser Menge haben bezüglich aller jeweiligen Eigenschaften $\mu_2, ..., \mu_{\underline{n}}$ in den entsprechenden genau definierten Intervallen zu liegen. Auf die vollständige Angabe dieser Intervalle wurde zugunsten einer allgemeingültigen Darstellung verzichtet, da die genaue Kenntnis dieser Intervalle zur Erklärung der rekursiven Aufteilung der Daten durch den Entscheidungsbaum nicht erforderlich ist.

Nachdem auf diese Weise ein Entscheidungsbaum erzeugt worden ist, wird bezüglich jedes Knoten $\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_{\underline{n}}})$ die in Anhang F dargestellte Version des LogitBoost-Algorithmus von Friedman et. al. (2000) benutzt, um die Logit- Regression hinsichtlich der Daten $J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_{\underline{n}}})]$ durchzuführen. Dieser Algorithmus hat den Vorteil, daß zusätzlich zur Schätzung der Parameter die hinsichtlich der Datenmenge $J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_{\underline{n}}})]$ optimalen exogenen Variablen bestimmt werden. Um der unterschiedlichen Bestimmung der Schätzer Rechnung zu tragen, werden die Schätzer, die mittels des LogitBoost-Algorithmus bestimmt werden, anders bezeichnet, als die üblicherweise im Rahmen der logistischen Regression bestimmten Maximum-Likelihood Schätzer $\hat{\beta}_{A,0}^{l}, \beta_A^l, l = 1, ..., L$. Der LogitBoost-Algorithmus ermittelt bezüglich der Daten $J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_{\underline{n}}})]$ die Schätzer

$$\hat{\mathfrak{G}}_{l,0}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]}, \hat{\mathfrak{G}}_{l,\kappa}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]}, \kappa = 1, ..., \kappa_{MAX}, l = 1, ..., L.$$

Alle Schätzer $\hat{\mathfrak{G}}_{l,\kappa}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]}, \kappa = 1, ..., \kappa_{MAX}$, deren zugehörige Eigenschaft κ nicht zur Menge der hinsichtlich der Daten $J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_{\underline{n}}})]$ ermittelten exogenen Variablen gehört, werden im Rahmen des LogitBoost-Algorithmus gleich Null gesetzt. Mittels

$$\hat{\mathfrak{G}}_{jl}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]} = \hat{\mathfrak{G}}_{l,0}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]} + \underline{a}_j' \hat{\mathfrak{G}}_{l}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]}, l = 1, ..., L,$$

und $\hat{\mathfrak{G}}_{l}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s})]} = (\hat{\mathfrak{G}}_{l,1}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]}, ..., \hat{\mathfrak{G}}_{l,\kappa_{MAX}}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]})'$ ergibt sich

$$\hat{P}(j \in C_2(l) | \underline{a}_j, J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_{\underline{n}}})]) = \frac{\exp\left(\hat{\mathfrak{G}}_{jl}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]}\right)}{\sum_{l'=1}^{L} \exp\left(\hat{\mathfrak{G}}_{jl'}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_{\underline{n}}})]}\right)}.$$

Für jeden Knoten $\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_n})$ wird ein logistisches Regressionsmodell, mit maximal $\underline{n}\xi_{MAX}$ exogenen Variaben geschätzt. ξ_{MAX} ist ein Parameter. Daher ist das zum Knoten $\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_n})$ gehörende Logistische Regressionsmodell im allgemeinen umso komplizierter, je höher $\underline{n}$ ist.

Nachdem der Entscheidungsbaum auf Basis von J^S erzeugt ist und für alle Knoten logistische Regressionen durchgeführt wurden, wird der Entscheidungsbaum mittels des sogenannten Error-Complexity Prunings, das von Breiman et. al. (1984) im Rahmen des CART Algorithmus verwendet wird, beschnitten. Die Beschneidung erfolgt auf der Grundlage der zu den jeweils betrachteten Knoten gehörenden Fehlklassifikationsrate bezüglich J^{cal}, die auf der Klassifikation beruht, die sich auf der Basis des lokal an dem betreffenden Knoten durch den LogitBoost-Algorithmus geschätzten logistischen Regressionsansatzes ergibt.

Weil das zu einem Knoten $\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_n})$ gehörende Logistische Regressionsmodell umso komplizierter sein kann, je höher $\underline{n}$ ist, werden durch den Pruning-Schritt alle zu komplizierten Modelle, die zu Überanpassungseffekten führen, entfernt. Es wird versucht, verschiedene Teilmengen der Items so zu bestimmen, daß das zugehörige auf Basis des LogitBoost-Algorithmus berechnete logistische Regressionsmodell optimal zur Vorhersage von Daten geeignet ist.

Eine Vorhersage für ein neues Item erfolgt, indem man dieses aufgrund der Ausprägungen seiner Eigenschaften einem bestimmten Endknoten $\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_n})$ des Logistischen Modellbaums zuordnet und dann die zum entsprechenden Endknoten gehörenden (geschätzten) Parameter $\hat{\mathfrak{G}}_{l,0}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})]}$, $\hat{\mathfrak{G}}_{l,\kappa}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})]}$, für alle $\kappa = 1, ..., \kappa_{MAX}$, $l = 1, ..., L$, zur Berechnung der zugehörigen geschätzten Wahrscheinlichkeit $\hat{P}(j \in C_2(l)|\underline{a}_j, J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_n})])$ $l = 1, ..., L$, benutzt. Auf dieser Basis ist das jeweilige neue Item dem Cluster zuzuordnen, zu dem es am wahrscheinlichsten gehört.

Es konnte empirisch belegt werden, daß die Logistischen Modellbäume häufig zu deutlich besseren Resultaten als die Entscheidungsbaum-Algorithmen C4.5 und CART sowie die logistische Regression führen (Landwehr et. al. (2005)).

Zur Klassifikation der Items mittels des Logistischen Modellbaums wurde die Weka-Software (Witten, Frank (2005)) eingesetzt.

Beispiel 9.6:

Wieder werden dieselben Daten wie in den Beispielen 9.2, 9.3 und 9.4 betrachtet.

Da die Anzahl der im Trainingsdatensatz zur Verfügung stehenden Items unter 15 ist, kann die Wurzel $\mathfrak{X}(\mu_1)$ nicht aufgeteilt werden. Daher entfällt die Erzeugung einer Entscheidungsbaumstruktur und es wird lediglich für die gesamte Datenmenge J^B ein Logit-Modell mittels des LogitBoost Algorithmus nach Friedman et. al. (2000) erstellt. Für die Cluster-Zugehörigkeiten aus Beispiel 5.11 ergeben sich an der Wurzel $\mathfrak{X}(\mu_1)$ die folgenden Endergebnisse hinsichtlich der Schätzer $\hat{\mathfrak{G}}_{jl}, l = 1, 2, 3$:

$$\begin{aligned}
\hat{\mathfrak{G}}_{j1}^{J[\mathfrak{X}(\mu_1)]} &= -1,76 + 0,01\underline{a}_{j6} \\
\hat{\mathfrak{G}}_{j2}^{J[\mathfrak{X}(\mu_1)]} &= -3,91 + 0,51\underline{a}_{j1} + 0,31\underline{a}_{j2} \\
\hat{\mathfrak{G}}_{j3}^{J[\mathfrak{X}(\mu_1)]} &= 5,06 - 0,51\underline{a}_{j1} - 0,31\underline{a}_{j2} \quad .
\end{aligned}$$

Unter Verwendung von $\underline{a}_j = a_j$ ergibt sich:

$$\begin{aligned}
\hat{\mathfrak{G}}_{91}^{J[\mathfrak{X}(\mu_1)]} &= -1,76 + 0,01 \cdot 5 = -1,71 \\
\hat{\mathfrak{G}}_{92}^{J[\mathfrak{X}(\mu_1)]} &= -3,91 + 0,51 \cdot 7 + 0,31 \cdot 8 = 2,14 \\
\hat{\mathfrak{G}}_{93}^{J[\mathfrak{X}(\mu_1)]} &= 5,06 - 0,51 \cdot 7 - 0,31 \cdot 8 = -0,99 \quad .
\end{aligned}$$

Daraus folgt:

$$\begin{aligned}
\hat{P}(j \in C_2(1)|\underline{a}_9, J[\mathfrak{X}(\mu_1)]) &= \frac{\exp\left(\hat{\mathfrak{G}}_{91}^{J[\mathfrak{X}(\mu_1)]}\right)}{\exp\left(\hat{\mathfrak{G}}_{91}^{J[\mathfrak{X}(\mu_1)]}\right)+\exp\left(\hat{\mathfrak{G}}_{92}^{J[\mathfrak{X}(\mu_1)]}\right)+\exp\left(\hat{\mathfrak{G}}_{93}^{J[\mathfrak{X}(\mu_1)]}\right)} \\
&= 0,02.
\end{aligned}$$

Analog erhält man die Wahrscheinlichkeiten $\hat{P}(j \in C_2(2)|\underline{a}_9, J[\mathfrak{X}(\mu_1)]) = 0,94$ und $\hat{P}(j \in C_2(3)|\underline{a}_9, J[\mathfrak{X}(\mu_1)]) = 0,04$. (Alle Ergebnisse gelten in Bezug auf die Klassifikation aus Beispiel 5.11.)

Auf dieser Basis folgen die diskreten Schätzer $\hat{q}_{91}^{\hat{S}_Y^1} = 0$, $\hat{q}_{92}^{\hat{S}_Y^1} = 1$ und $\hat{q}_{93}^{\hat{S}^1} = 0$. Somit gilt im Hinblick auf den Schätzer hinsichtlich Bernds Bewertung in Bezug auf das neue Item $\hat{Y}_{19}^{\hat{S}_Y^1} = w_{12} = 1$.

9.2.3.6 Neuronale Netze

Ein Mehrschichtiges Neuronales Netzwerk mit Rückpropagation wurde bereits in Abschnitt 4.2 kurz vorgestellt. Das im Rahmen von Kapitel 4 benutzte Neuronale

Netzwerk diente nur zur Zuordnung zu zwei disjunkten Klassen $\bar{z} = 0$ oder $\bar{z} = 1$. Durch die Definition

$$g_{\bar{z}}^{mult}(\mathbf{T}_{\bar{z}j}) = \frac{\exp(\mathbf{T}_{\bar{z}j})}{\sum_{l'=1}^{L} \exp(\mathbf{T}_{l'j})}, \quad \bar{z} = 1, ..., L,$$

läßt sich der in Abschnitt 4.2 dargestellte Ansatz unter Verwendung der Fehlerfunktion

$$\mathcal{F}_{item}^{NN} = \sum_{\bar{z}=1}^{L} \sum_{j \in J^B} (q_{j\bar{z}} - g_{\bar{z}}^{mult}(\mathbf{T}_{\bar{z}j}))^2$$

und der Beziehungen

$$\mathbf{T}_{\bar{z}j} = \underline{\beta}_{0\bar{z}} + \underline{\beta}'_{0\bar{z}} Z_j, \bar{z} = 1, ..., L, j \in J^B,$$
$$Z_j = (Z_{1j}, ..., Z_{Mj}), \text{ mit } Z_{mj} = A_{akt}(\underline{\alpha}_{0m} + \underline{\alpha}'_m a_j), m = 1, ..., M, j \in J^B,$$

auf den L-Klassen Fall verallgemeinern und zur Klassifikation der Items auf Basis ihrer Eigenschaften einsetzen. Hierzu wurde in dieser Arbeit die Statistik-Software R (Hornik (2004)) verwendet.

Neuronale Netzwerke erfordern die Schätzung von Parametern, die häufig als Gewichte bezeichnet werden. Im Hinblick auf das vorgestellte Neuronale Netzwerk sind $\underline{\alpha}_m \in \mathbb{R}^{\kappa_A}, \underline{\alpha}_{0m}, m = 1, ..., M$, sowie die Gewichte $\underline{\beta}_{\bar{z}} \in \mathbb{R}^M, \underline{\beta}_{0\bar{z}}, \bar{z} = 1, ..., L$, zu berechnen. Daher ist die Bestimmung von insgesamt $M(\kappa_A + 1) + L(M + 1)$ Parametern erforderlich. Je höher M gewählt wird, umso besser können Neuronale Netzwerke Nichtlinearitäten berücksichtigen. Andererseits ist zu beachten, daß bei einer hohen Anzahl von Eigenschaften und zu schätzender Klassen eine hohe Anzahl von „hidden units“ M schnell zu einer sehr hohen Anzahl an Parametern führt, wodurch es zur Überanpassung kommen kann.

9.3 Empirische Ergebnisse

Als Datengrundlage dienen wieder die beiden 418 Items umfassenden Teildatensätze D1 und D3 des MovieLens-Datensatzes. Aus der Menge von 418 Items

$\hat{S}^2_Y$	MW	SL	$\bar{\alpha}$	KM	$C4.5$	LMT	BMN	LR	NN
R^2	0,137	0,172	0,061 (0,175)	0,156	0,198 (0,206)	0,156 (0,196)	0,193 (0,203)	0,049 (0,116)	0,158 (0,179)
AAD	0,840	0,791	0,865 (0,811)	0,819	0,798 (0,792)	0,815 (0,797)	0,801 (0,795)	0,867 (0,834)	0,812 (0,799)
$Prec.$	-	0,427	0,391 (0,416)	0,415	0,422 (0,429)	0,404 (0,424)	0,435 (0,424)	0,371 (0,384)	0,408 (0,395)
$Rec.$	-	0,368	0,396 (0,360)	0,450	0,422 (0,435)	0,416 (0,417)	0,416 (0,398)	0,414 (0,401)	0,407 (0,396)
R_B	55,99	70,83	68,54 (70,12)	70,75	71,67 (72,43)	71,04 (71,33)	71,66 (71,44)	68,44 (69,51)	69,66 (70,15)

Tabelle 9.6: Ergebnisse der indirekten Schätzung auf Basis zweimodalen $\hat{S}^2_Y$-Clusterverfahrens ($K = L = 10$) unter Verwendung der Variablen-Kombination M_2

wurden 118 per Zufallsgenerator ausgewählt. Diese Items bilden die Menge der neuen Items J^N. Entsprechend werden die verbleibenden 300 Items als Menge der bekannten Items J^B verwendet. Zusätzlich zu den Bewertungen werden die Eigenschaften dieser Filme sowie bezüglich dieser Filme verfügbare Kritikerbewertungen benutzt.

9.3.1 Ergebnisse der indirekten Schätzung

Zu Beginn dieser empirischen Untersuchung werden einfach alle Bewertungen in Bezug auf die Items J^B als Trainingsdatensatz verwendet. Im späteren Verlauf dieses Abschnitts werden auch verschieden große Teile der Bewertungen der Items J^B als Trainingsdatensatz eingesetzt. In beiden Fällen besteht der Testdatensatz aus allen Bewertungen mit Bezug auf Items aus J^N.

In Tabelle 9.6 sind die Ergebnisse der indirekten Schätzung auf Basis des zweimodalen $\hat{S}^2_Y$-Clusterverfahrens mit den Clustergrößen $K = L = 10$ bezüglich der verschiedener Verfahren zur Approximation der Item-Cluster Zugehörigkeit dargestellt. Zur Approximation der Item-Cluster Zugehörigkeit der neuen Items wurden das SL-Verfahren zur Approximation der Item-Cluster Zugehörigkeiten (SL), die KM-Methode zur Approximation der Item-Cluster Zugehörigkeit (KM), der $\bar{\alpha}$-Ansatz zur näherungsweisen Bestimmung der Item-Cluster Zugehörigkeit

($\bar{\alpha}$), der C4.5 Algorithmus ($C4.5$), ein Logistischer Modellbaum (LMT) und ein Bayes'sches Multinetzwerk (BMN) sowie ein logistischer Regressionsansatz (LR) verwendet. In Bezug auf das Mehrschichtige Neuronale Netzwerk mit Rückpropagation (NN) wurde $M = 6$ benutzt. Als Basis-Modell dient die von George, Merugu (2005) vorgeschlagene Methode der Vorhersage des Nutzer-Mittelwerts (MW). Hinsichtlich aller Verfahren wurden alle 14 Eigenschaften (Regressorenkombination M_2) zur Basis der Approximation der Item-Cluster Zugehörigkeit gewählt. Die Werte in Klammern beziehen sich auf die kontinuierliche Version des entsprechenden Verfahrens, bei der nicht die geschätzte Item-Cluster Zugehörigkeit sondern der (normierte) Grad der Zugehörigkeit zu den verschiedenen Item-Clustern zur Berechnung der Schätzer verwendet wird.

Das SL- und das KM-Verfahren sowie der C4.5 Algorithmus, der logistische Modellbaum und das Bayes'sche Multinetzwerk führen in Bezug auf den Testdatensatz immerhin zu deutlich besseren Ergebnissen als das von George, Merugu vorgeschlagene MW-Verfahren. Insgesamt lassen alle auf Basis des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens erzielten Ergebnisse in Bezug auf die Genauigkeit (AAD und R^2) zu wünschen übrig. Trotzdem fallen die Breese-Werte teilweise erfreulich hoch aus. Da der zum Zwecke der Schätzung neuer Items leicht abgewandelte $\hat{S}_{Y,ij}^{N2}$-Schätzer neben dem Nutzer-Mittelwert $\bar{Y}_{i.}$ auch die durchschnittliche Kritikerbewertung $\bar{Y}_{.j}^{MC}$ enthält, hängen die Schätzer nicht allein von der Zuordung des betrachteten Items j zu einem Item-Cluster ab. Zudem dürften hierdurch die Schätzer für allgemein populäre Items in der Regel etwas höher ausfallen. Hierdurch können die erfreulichen Breese-Werte erklärt werden.

Die Verwendung anderer Clustegrößen (L) für die Item-Cluster und die Verwendung anderer Regressorenkombinationen ($M3$, $M8$, $M1$, usw.) führen auf Basis des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens nicht zu signifikanten Ergebnisverbesserungen.

Auch durch die Verwendung des ordinalen zweimodalen Clusterverfahrens lassen sich kaum Verbesserungen erzielen.

Außer dem SL-Verfahren zur Approximation der Item-Cluster Zugehörigkeit führen alle Methoden der indirekten Schätzung auf Basis des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens zu Ergebnissen, die nicht einmal die Resultate des Basis-Verfahrens (MW) nach George, Merugu (2005) übertreffen. Grund hierfür ist, daß das zweimodale $\hat{S}_Y^1$-Clusterverfahren empfindlicher auf weniger glückliche Cluster-Zuordnungen reagiert. Außerdem ist zu bemerken, daß die involvierteren

R^2 (Test)	R^2 (Train)	AAD (Test)	AAD (Train)	$Prec.$	$Rec.$	R_B
0,270	0,392	0,765	0,684	0,597	0,149	70,77

Tabelle 9.7: Ergebnisse der indirekten Methode zur Berechnung der Schätzer für neue Items auf Basis des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens ($K = L = 10$) und des SL-Verfahrens zur Approximation der Item-Cluster Zugehörigkeit unter Verwendung der Variablenkombination M_2 hinsichtlich D1 (mit den Zusätzen „Test“ und „Train“ in Klammern sind die entsprechenden Gütemaße in Bezug auf die Test- bzw. Trainingsdaten gemeint)

Verfahren wie der C4.5 Algorithmus, das Bayes'sche Multinetzwerk und die logistischen Modellbäume normalerweise eher zu Klassifikationsaufgaben mit deutlich weniger als 10 Klassen eingesetzt werden. Dennoch ergeben sich auch bei niedrigeren Item-Clustergrößen ($L = 5, 7$) bezüglich der involvierteren Verfahren keine nennenswerten Verbesserungen. Überraschenderweise sind die Resultate der indirekten Schätzung mittels des SL-Verfahrens zur Approximation der Item-Cluster Zugehörigkeit auf Basis des $\hat{S}_Y^1$-Clusterverfahrens ziemlich gut (siehe Tabellen 9.7). Es wurden wieder die Clusteranzahlen $K = L = 10$ verwendet. Die Approximation der Item-Cluster Zugehörigigkeit erfolgte auf Basis der Variablenkombination M_2. (Alle Variablen wurden gleich gewichtet.)

R^2, AAD, und Präzision sind deutlich besser als die entsprechenden Resultate der indirekten Schätzung auf Basis des zweimodalen $\hat{S}^2$-Clusterverfahrens. (Das selbe Muster wurde bei anderer Aufteilung in Test- und Trainingsdatensatz reproduziert.) Trotzdem ist der Breese-Wert R_B etwas niedriger als bei den indirekten Schätzverfahren, die auf dem zweimodalen $\hat{S}_Y^2$-Clusterverfahren und dem C4.5 Algorithmus, dem logistischen Modellbaum und dem Bayes'schen Multinetzwerk basieren. Die Tabellen 9.8 und 9.9 enthalten die Ergebnisse der indirekten Schätzverfahren auf Basis des SL-Ansatzes zur Approximation der Item-Cluster Zugehörigkeit mittels des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens (das im folgenden als $SL/\hat{S}_Y^1$-Heuristik bezeichnet wird) unter Verwendung verschiedener Variablenmengen und unterschiedlicher Größen der Item-Cluster. Man erkennt, daß die Güte der Schätzer eher von der zur Berechnung der Distanzen verwendeten Variablenkombinationen als von der Anzahl der verwendeten Item-Cluster abhängt. Die Variablenkombination M_3, die bereits a priori im Rahmen der Variablenselektion auf Basis der MCMC Modellbestimmung in Verbindung mit dem

	Variablenmenge M_2			Variablenmenge M_3			Variablenmenge M_8		
L	5	7	10	5	7	10	5	7	10
R^2	0,262	0,255	0,270	0,302	0,299	0,294	0,272	0,272	0,270
AAD	0,773	0,778	0,765	0,753	0,754	0,756	0,763	0,767	0,759
$Prec.$	0,684	0,706	0,597	0,674	0,637	0,647	0,649	0,635	0,684
$Rec.$	0,127	0,136	0,149	0,249	0,256	0,345	0,173	0,194	0,145
R_B	70,57	70,71	70,77	72,91	72,94	73,75	72,17	72,44	72,26

Tabelle 9.8: R^2,AAD, Präzision, Recall und Breese-Wert R_B des indirekten Schätzverfahrens auf Basis des SL-Verfahrens zur Approximation der Item-Cluster Zugehörigkeit und des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens unterschiedlichen Item-Clustergrößen L und auf Basis der Variablenkombinationen M_2, M_3 und M_8 bezüglich D1 (1067 Nutzer, 418 Items)

	Variablenmenge M_2			Variablenmenge M_3			Variablenmenge M_8		
L	5	7	10	5	7	10	5	7	10
R^2	0,249	0,263	0,264	0,304	0,303	0,298	0,262	0,285	0,285
AAD	0,784	0,777	0,775	0,750	0,751	0,754	0,776	0,782	0,763
$Prec.$	0,717	0,742	0,707	0,649	0,655	0,641	0,570	0,603	0,614
$Rec.$	0,065	0,171	0,164	0,213	0,257	0,324	0,075	0,218	0,157
R_B	73,35	75,78	75,77	77,43	77,37	77,45	75,51	76,09	76,27

Tabelle 9.9: R^2,AAD, Präzision, Recall und Breese-Wert R_B des indirekten Schätzverfahrens auf Basis des SL-Verfahrens zur Approximation der Item-Cluster Zugehörigkeit und des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens unterschiedlichen Item-Clustergrößen L und auf Basis der Variablenkombinationen M_2, M_3 und M_8 bezüglich D3 (2020 Nutzer, 418 Items)

unvollständigen balancierten Blockplan identifiziert wurde, erweist sich als am besten geeignet. Die Ergebnisse der $SL/\hat{S}_Y^1$-Heuristik auf Basis der Variablenkombinationen M_0 und M_3 für unterschiedlich große Trainingsdatensätze auf Basis von D1 und D3 sind in den Tabellen 9.10 und 9.11 abgebildet. Zur Berechnung dieser Werte wurden als Trainingsdatensätze unterschiedlich große Teilmenge der insgesamt in Bezug auf die 300 bekannten Items abgegebenen Bewertungen verwendet. Als Testdatensatz wurden in allen Fällen die Bewertungen für neue Items einge-

verwendeter	Variablenkombination M_0				Variablenkombination M_3			
J^B-Datenanteil	AAD	$Prec.$	$Rec.$	R_B	AAD	$Prec.$	$Rec.$	R_B
100%	0,769	0,569	0,139	71,53	0,754	0,647	0,345	73,75
90%	0,777	0,624	0,135	70,21	0,749	0,588	0,291	73,51
80%	0,790	0,692	0,107	70,37	0,754	0,545	0,320	74,14
70%	0,789	0,595	0,113	69,58	0,763	0,526	0,291	72,62
60%	0,803	0,680	0,077	67,04	0,764	0,543	0,307	73,09
50%	0,796	0,531	0,137	68,90	0,770	0,525	0,320	73,36
40%	0,799	0,505	0,174	68,22	0,805	0,469	0,356	70,85
30%	0,811	0,517	0,112	67,50	0,813	0,466	0,257	70,25
20%	0,857	0,551	0,051	65,93	0,844	0,391	0,218	68,03
10%	0,937	0,421	0,152	63,17	0,898	0,355	0,319	68,80

Tabelle 9.10: AAD, Präzision, Recall und Breese-Werte bei unterschiedlich großen Trainingsdatensätzen (dh. verschieden großen Anteilen des Trainingsdatensatzes an der bezüglich der bekannten Items J^B im Datensatz D1 vorhanden Datenmenge) für das $SL/\hat{S}_Y^1$-Verfahren im Hinblick auf die Variablenkombinationen M_0 (links) und M_3 (rechts)

setzt. Daher sind bei jedem prozentualen Anteil des Trainingsdatensatzes an den insgesamt in Bezug auf die bekannten Items von den betrachteten Nutzern abgegebenen Bewertungen für jeden Nutzer dieselben Bewertungen im Testdatensatz. Im Unterschied zur in den vorangegangenen Abschnitten 5.10 und 8.4 benutzten Unterteilung in Test- und Trainingsdatensatz, sind daher auch, wenn hinsichtlich eines Nutzers fast alle Bewertungen aus dem Trainingsdatensatz entfernt worden sind, nicht mehr Bewertungen dieses Nutzers im Testdatensatz enthalten. Daher verringern sich die Breese-Werte mit abnehmender Trainingsdatenmenge nicht so stark.

Auffällig ist, daß die Ergebnisse auf Basis der Variablenkombination M_3 deutlich besser ausfallen. Zudem verschlechtern sich die auf Basis der Variablenkombination M_3 berechneten Ergebnisse vor allem in Bezug auf die resultierenden Breese-Werte mit abnehmender Trainingsdatenmenge weniger stark. Dies unterstreicht ein weiteres Mal die Bedeutung der verwendeten Variablenkombination.

Die Ergebnisse hinsichtlich des ordinalen zweimodalen Clusterverfahrens in Verbindung mit den Methoden zur Approximation der Item-Cluster Zugehörig-

verwendeter	Variablenkombination M_0				Variablenkombination M_3			
J^B-Datenanteil	AAD	$Prec.$	$Rec.$	R_B	AAD	$Prec.$	$Rec.$	R_B
100%	0,773	0,554	0,161	75,79	0,754	0,641	0,324	77,45
90%	0,778	0,735	0,073	75,52	0,754	0,587	0,331	77,22
80%	0,799	0,535	0,092	74,54	0,756	0,554	0,343	77,55
70%	0,781	0,567	0,196	75,77	0,769	0,594	0,248	77,58
60%	0,787	0,517	0,162	75,73	0,779	0,487	0,406	76,41
50%	0,809	0,532	0,111	71,55	0,779	0,521	0,377	77,70
40%	0,811	0,612	0,072	73,31	0,793	0,526	0,374	76,63
30%	0,829	0,661	0,078	72,32	0,809	0,491	0,377	76,08
20%	0,877	0,531	0,092	71,39	0,822	0,456	0,363	74,78
10%	0,934	0,441	0,223	72,75	0,941	0,356	0,357	73,15

Tabelle 9.11: AAD, Präzision, Recall und Breese-Werte bei unterschiedlich großen Trainingsdatensätzen (dh. verschieden großen Anteilen des Trainingsdatensatzes an der bezüglich der bekannten Items J^B im Hinbick auf die betrachteten Nutzer verfügbaren Datenmenge) für das $SL/\hat{S}_Y^1$-Verfahren zu den Variablenkombinationen M_0 (links) und M_3 (rechts) auf Basis des größeren Ausschnitts aus dem MovieLens-Datensatz D2 (2020 Nutzer, 418 Items)

keit fallen überwiegend schwach aus.

In Tabelle 9.12 sind die Resultate des ordinalen zweimodalen Clusterverfahrens (OZC) und des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens in Verbindung mit der SL-Methode zur Approximation der Item-Cluster Zugehörigkeit neuer Items hinsichtlich der Variablenkombination M_3 abgebildet. Verwendet wurden wieder die bewährten Clustergrößen $K = L = 10$.

Die Berücksichtigung des ordinalen Skalenniveaus der Bewertungsdaten im Rahmen des ordinalen zweimodalen Clusterverfahrens führt in Verbindung mit der SL-Methode zur Approximation der Item-Cluster Zugehörigkeit teilweise zu Ergebnisverbesserungen im Vergleich zur $SL/\hat{S}_Y^2$-Heuristik. Dennoch sind die durch die $SL/\hat{S}_Y^1$-Heuristik erzielten Resultate den in Tabelle 9.12 aufgelisteten Ergebnissen klar überlegen. Folglich scheint die Berücksichtigung des ordinalen Skalenniveaus in diesem Zusammenhang nicht hilfreich zu sein.

verwendeter	zweimodale $\hat{S}_Y^2$-Clustermethode				OZC			
J^B-Datenanteil	AAD	$Prec.$	$Rec.$	R_B	AAD	$Prec.$	$Rec.$	R_B
100%	0,791	0,427	0,368	70,83	0,766	0,431	0,414	71,19
90%	0,808	0,401	0,440	70,56	0,764	0,416	0,410	71,04
80%	0,806	0,424	0,374	69,99	0,769	0,423	0,417	70,79
70%	0,820	0,403	0,428	70,39	0,756	0,438	0,422	71,91
60%	0,839	0,391	0,490	70,66	0,787	0,397	0,425	70,50
50%	0,830	0,392	0,449	70,69	0,778	0,406	0,439	70,87
40%	0,840	0,392	0,447	70,72	0,807	0,387	0,443	69,38
30%	0,861	0,385	0,439	69,88	0,812	0,366	0,452	69,83
20%	0,892	0,375	0,486	70,81	0,845	0,364	0,445	69,19
10%	0,936	0,341	0,447	67,26	0,897	0,330	0,413	67,05

Tabelle 9.12: AAD, Präzision, Recall und Breese-Werte bei unterschiedlich großen Trainingsdatensätzen (dh. verschieden großen Anteilen des Trainingsdatensatzes an der bezüglich der bekannten Items J^B im Hinblick auf die betrachteten Nutzer verfügbaren Datenmenge) für das SL-Verfahren in Verbindung mit den zweimodalen $\hat{S}_Y^2$-Clusterverfahren (links) und dem ordinalen zweimodalen Clusterverfahren (rechts) unter Benutzung der Variablenkombination M_3 auf Basis des Ausschnitts D1 aus dem MovieLens-Datensatz (1067 Nutzer, 418 Items)

9.3.2 Ergebnisse der direkten Schätzung

Geeignete Methoden zur direkten Schätzung sind das hybride GP-Verfahren und der HBLR-Ansatz.

Eine in Bezug auf das HBLR-Verfahren möglicherweise geeignete Methode zur Variablenselektion ist die in Abschnitt 8.1.4 vorgestellte Kombination aus einer MCMC Modellbestimmung und einem balancierten unvollständigen Blockplan. In Abschnitt 8.4 erfolgte die mit diesem Ansatz durchgeführte Modellbestimmung auf Basis aller Items. Um festzustellen, ob es möglich ist, durch diesen Ansatz unter Verwendung der hinsichtlich der bekannten Items abgegebenen Bewertungen die optimale Variablenkombination zu bestimmen, muß dieses Verfahren erneut ausschließlich auf Basis der sich auf bekannte Items beziehenden Bewertungen durchgeführt werden. Um den numerischen Aufwand zu begrenzen,

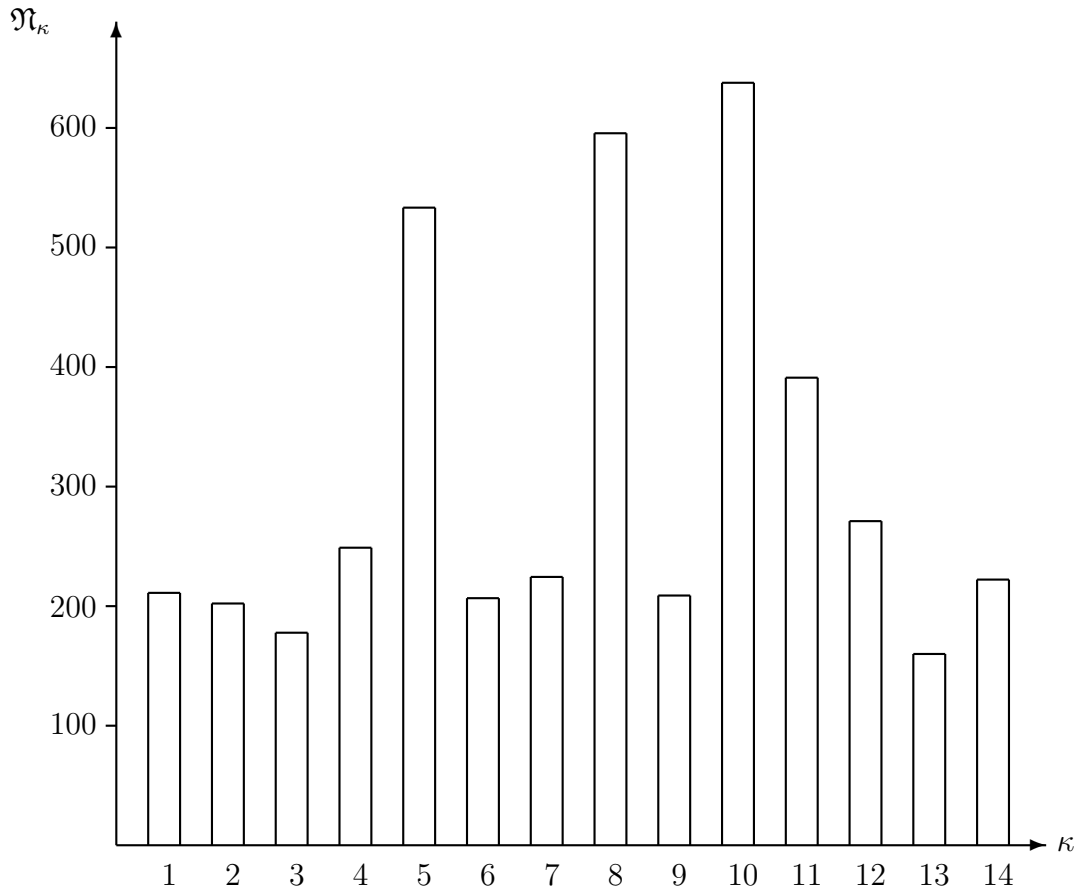

Abbildung 9.2: auf Basis von 10% der hinsichtlich der bekannten Items abgegebenen Bewertungen aus D1 bestimmte Häufigkeit $\mathfrak{N}_\kappa$, mit der die jeweilige Eigenschaft κ im Modell mit der höchsten Posterior-Wahrscheinlichkeit für einen einzelnen Nutzer enthalten ist, $\kappa = 1, ..., 14$

werden hierzu nur 10% der in Bezug auf die Items aus J^B gegebenen Bewertungen verwendet. Die Ergebnisse dieses Modellbestimmungsansatzes sind in Abbildung 9.2 graphisch dargestellt. In wesentlichen Punkten entsprechen diese Ergebnisse den in Bezug auf alle 418 Items berechneten Ergebnissen (vgl. Abbildung 8.13). Der Actionfilmcharakter, der Grad der erzeugten Spannung und der Grad der dargestellten Charakterentwicklung sind auch bei Beschränkung auf die hinsichtlich der Items aus der Menge J^B abgegebenen Bewertungen mit großem Abstand die wichtigsten Variablen. Der Grad der Exzentrizität ($\kappa = 11$) scheint in Bezug auf die bekannten Items etwas wichtiger zu sein als allgemein. Aufgrund dieser Ergebnisse ist die erfolgversprechendsten Variablenkombination M_3.

	verwendeter J^B-Datenanteil:								
	100 %	90 %	80 %	70 %	60 %	50 %	40 %	30 %	20 %
M_0	194879	175353	156336	136540	118006	97636	78986	59245	39349
M_1	189265	170385	152016	132701	114706	94965	77003	57627	38365
M_2	192484	173291	154540	135027	116512	96663	78298	59146	39034
M_3	192327	173107	154416	134832	116556	96449	78023	58563	38962
M_4	191016	171910	153401	134002	115777	95893	77621	58191	38772
M_5	190752	171741	153143	133671	115586	95701	77480	58132	38731
M_6	192266	173024	154361	134783	116529	96923	77978	58530	38947
M_7	190357	171342	152856	133474	115352	95562	77370	58029	38683
M_8	190360	171294	152914	133621	115029	95556	77299	58012	38701

Tabelle 9.13: *DIC*-Werte des konservativen HBLR-Verfahrens bezüglich der Regressorenkombinationen M_0, M_1, M_2, M_3, M_4, M_5, M_6, M_7 und M_8 (D1)

verwendeter	Variablenkombination M_0				Variablenkombination M_2			
J^B-Datenanteil	*AAD*	*Prec.*	*Rec.*	R_B	*AAD*	*Prec.*	*Rec.*	R_B
100%	0,758	0,602	0,032	68,15	0,780	0,611	0,052	67,36
90%	0,759	0,582	0,032	68,23	0,782	0,642	0,053	67,29
80%	0,760	0,620	0,036	68,19	0,783	0,638	0,048	67,33
70%	0,761	0,587	0,032	68,20	0,784	0,626	0,047	67,45
60%	0,763	0,588	0,035	68,27	0,788	0,624	0,047	66,94
50%	0,762	0,579	0,035	68,31	0,790	0,615	0,045	66,49
40%	0,768	0,540	0,024	68,61	0,793	0,538	0,032	67,78
30%	0,772	0,591	0,036	68,61	0,784	0,525	0,034	67,33
20%	0,782	0,528	0,031	68,85	0,804	0,497	0,029	66,89
10%	0,803	0,466	0,030	68,21	0,819	0,465	0,031	66,65

Tabelle 9.14: *AAD*, Präzision, Recall und Breese-Werte bei unterschiedlich großen Trainingsdatensätzen auf Basis von D1 in Bezug auf das HBLR-Verfahren auf Grundlage der Variablenkombinationen M_0 (links) und M_2 (rechts)

In Tabelle 9.13 sind die *DIC*-Werte des konservativen HBLR-Verfahrens im

verwendeter	Variablenkombination M_1				Variablenkombination M_3			
J^B-Datenanteil	AAD	$Prec.$	$Rec.$	R_B	AAD	$Prec.$	$Rec.$	R_B
100%	0,755	0,636	0,067	69,87	0,762	0,613	0,043	70,02
90%	0,757	0,639	0,065	69,85	0,763	0,629	0,048	70,06
80%	0,758	0,635	0,063	69,80	0,764	0,642	0,047	70,01
70%	0,760	0,609	0,060	69,99	0,764	0,639	0,047	70,14
60%	0,764	0,628	0,064	69,78	0,768	0,624	0,046	70,03
50%	0,766	0,609	0,063	69,24	0,768	0,606	0,045	69,81
40%	0,768	0,550	0,040	70,12	0,771	0,630	0,041	69,78
30%	0,774	0,613	0,051	69,59	0,776	0,616	0,041	69,63
20%	0,783	0,557	0,041	69,84	0,785	0,563	0,045	69,64
10%	0,792	0,566	0,045	69,66	0,793	0,459	0,040	69,53

Tabelle 9.15: AAD, Präzision, Recall und Breese-Werte bei unterschiedlich großen Trainingsdatensätzen (D1) in Bezug auf das HBLR-Verfahren auf Basis der Variablenkombinationen M_1 (links) und M_3 (rechts)

Hinblick auf unterschiedliche Regressorenkombinationen aufgelistet. Es werden wieder unterschiedlich große Trainingsdatensätze betrachtet. Alle Trainingsdatensätze enthalten einen unterschiedlichen prozentualen Anteil der Bewertungen, die die betrachteten 1067 Nutzer in Bezug auf die 300 bekannten Items abgegeben haben (D1). Das komplexeste Modell (M_1) weist bei allen Anteilen des Trainingsdatensatz an der insgesamt im Hinblick auf die verwendeten Nutzer bezüglich der bekannten Items vorhandenen Menge der Bewertungen die kleinsten DIC-Werte auf.

Bezüglich des HBLR-Ansatzes wurde die bereits in Abschnitt 8.4 beschrieben konservative Parametrisierung verwendet. Die Tabellen 9.14 und 9.15 enthalten die Ergebnisse in Bezug auf den Testdatensatz hinsichtlich der konservativen HBLR-Verfahren auf Basis der Regressorenmengen M_0, M_1, M_2 und M_3. Das komplexeste Modell M_1 erweist sich hinsichtlich der Anpassung an die Daten (gemessen durch den AAD-Wert) tatsächlich als das beste Modell. Trotzdem fallen die Breese-Werte im Hinblick auf das Modell M_3 überwiegend geringfügig besser aus als auf Basis der Regressorenkombination M_1.

Dies legt die Vermutung nahe, daß im Hinblick auf ein konservatives HBLR-Verfahren die unter Anpassungs- bzw. Vorhersagegesichtspunkten optimale Variablenenkombination tatsächlich auf der Grundlage des DIC-Werts gewählt werden kann. Diese Wahl muß jedoch nicht zu den nützlichsten Empfehlungen führen.

Während aufgrund der DIC-Werte das komplexeste Modell (M_1) zu wählen wäre, legen die Ergebnisse der in Abschnitt 8.1.4 vorgestellten Kombination aus einer MCMC Modellbestimmung und einem balancierten unvollständigen Blockplan eher nahe, das weniger komplexe Modell M_3 zu verwenden. Vor dem Hintergrund der resultierenden Breese-Werte wäre auch diese Wahl zu rechtfertigen.

Dennoch sind die Unterschiede zwischen dem alle Variablen enthaltenden Modell M_1 und der aus nur 5 Variablen bestehenden Regressorenkombination M_3 so gering, daß im Hinblick auf das konservative HBLR-Verfahren eine genaue Bestimmung der im Hinblick auf den zu erwartenden Nutzen optimalen Regressorenkombination verzichtet werden kann.

Dies ist ein wichtiger Vorteil des konservativen HBLR-Verfahrens. Während die Vorhersagekraft klassische Regressionsansätze sensibel von der sorgfältigen und faktenbasierten Auswahl geeigneter exogener Variablen abhängt, scheint das konservative HBLR-Verfahren weniger stark von der Auswahl der unabhängigen Variablen abzuhängen. Das konservative HBLR-Verfahren eignet sich daher speziell für Fälle, in denen die Schätzung genereller Tendenzen von untergeordneter Bedeutung ist und die Schätzung oder Vorhersage individuellen Verhaltens im Vordergrund des Interesses steht. Obwohl es ein Regressionsmodell ist, ist es selbst unter Voraussetzung der gewählten konservativen Parametrisierung, selbst dann noch zu zuverlässigen Vorhersagen in der Lage, wenn vorher keine genaue Modellbestimmung durchgeführt wird oder werden kann. Hinsichtlich der Vorhersage von Bewertungen erwies sich die gewählte konservative Parametrisierung auf Basis der Datensätze D1 und D3 als optimal.

Tabelle 9.16 enthält die AAD-Werte des hybriden GP-Verfahrens auf Basis des 1067 Nutzer und 418 Items umfassenden Ausschnitts D1 aus dem MovieLens-Datensatz unter Verwendung der Variablenkombinationen M_0 bis M_8. Es werden wieder verschieden große Anteile der von den 1067 Nutzern hinsichtlich der 300 bekannten Filme abgegebenen Bewertungen als Trainingsdatensatz verwendet.

AAD	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
100%	0,733	0,733	0,733	0,734	0,730	0,733	0,735	0,732	0,731
90%	0,730	0,731	0,726	0,728	0,727	0,726	0,735	0,730	0,726
80%	0,731	0,730	0,726	0,731	0,724	0,728	0,732	0,727	0,719
70%	0,731	0,726	0,727	0,727	0,720	0,727	0,723	0,722	0,722
60%	0,732	0,731	0,729	0,727	0,724	0,724	0,727	0,720	0,718
50%	0,733	0,736	0,735	0,726	0,724	0,722	0,733	0,729	0,727
40%	0,741	0,743	0,744	0,731	0,730	0,728	0,734	0,731	0,728
30%	0,745	0,756	0,750	0,737	0,740	0,736	0,744	0,740	0,738
20%	0,763	0,769	0,774	0,754	0,750	0,751	0,758	0,756	0,757
10%	0,784	0,808	0,821	0,765	0,772	0,775	0,789	0,787	0,784

Tabelle 9.16: *AAD*-Werte bei unterschiedlich großen Trainingsdatensätzen in Bezug auf das GP-Verfahren auf Basis der Variablenkombinationen M_0 bis und M_8 bezüglich der D1-Daten (1067 Nutzer, 418 Items)

R_B	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
100%	68,21	68,26	68,26	68,34	68,17	68,38	68,51	68,23	68,14
90%	68,21	68,05	68,17	68,40	68,10	68,53	68,43	68,22	68,07
80%	68,16	67,73	67,66	68,23	68,15	68,15	68,01	67,65	67,85
70%	68,21	67,54	67,57	68,14	67,98	68,23	68,46	67,76	67,92
60%	68,21	66,82	67,02	68,11	67,47	67,96	67,90	67,94	67,88
50%	68,21	66,46	67,16	68,10	67,55	67,95	67,71	67,40	67,47
40%	68,20	66,28	66,21	67,87	67,22	67,80	67,25	66,53	67,07
30%	68,21	65,22	65,79	67,80	66,47	67,90	67,31	66,92	67,07
20%	68,08	65,43	65,21	67,14	66,96	66,75	66,63	66,30	66,53
10%	68,12	65,12	64,88	67,02	66,81	66,74	66,17	66,25	65,59

Tabelle 9.17: Breese-Werte bei unterschiedlich großen Trainingsdatensätzen in Bezug auf das GP-Verfahren auf Basis der Variablenkombinationen M_0 bis und M_8 hinsichtlich D1 (1067 Nutzer, 418 Items)

Prec./	verwendete Variablenkombination:								
Rec.	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
100%	0,582	0,535	0,535	0,570	0,575	0,573	0,562	0,566	0,565
	0,091	0,145	0,145	0,101	0,114	0,116	0,104	0,124	0,120
90%	0,560	0,537	0,538	0,564	0,582	0,570	0,536	0,547	0,585
	0,095	0,153	0,147	0,108	0,126	0,126	0,111	0,131	0,127
80%	0,571	0,534	0,541	0,567	0,570	0,567	0,554	0,559	0,564
	0,101	0,161	0,158	0,116	0,128	0,130	0,113	0,133	0,135
70%	0,581	0,539	0,530	0,581	0,567	0,567	0,673	0,560	0,562
	0,100	0,168	0,165	0,116	0,133	0,134	0,131	0,148	0,142
60%	0,555	0,525	0,531	0,560	0,560	0,559	0,562	0,560	0,551
	0,098	0,165	0,166	0,122	0,133	0,136	0,128	0,147	0,146
50%	0,548	0,508	0,520	0,558	0,546	0,549	0,553	0,540	0,530
	0,102	0,180	0,182	0,125	0,142	0,139	0,132	0,147	0,142
40%	0,551	0,510	0,505	0,554	0,548	0,545	0,533	0,527	0,526
	0,111	0,187	0,179	0,129	0,151	0,139	0,138	0,155	0,149
30%	0,524	0,474	0,480	0,520	0,518	0,516	0,511	0,507	0,521
	0,121	0,182	0,186	0,132	0,149	0,144	0,154	0,171	0,156
20%	0,477	0,465	0,450	0,490	0,492	0,491	0,483	0,488	0,488
	0,126	0,182	0,185	0,148	0,154	0,155	0,137	0,156	0,152
10%	0,410	0,398	0,417	0,424	0,411	0,417	0,423	0,425	0,407
	0,142	0,190	0,191	0,159	0,175	0,164	0,171	0,181	0,154

Tabelle 9.18: Präzision und Recall (untereinander) des GP-Verfahrens auf Basis unterschiedlicher Regressorenkombinationen und verschieden großer D1-Trainingsdatenmengen (1067 Nutzer, 418 Items)

Die *AAD*-Werte fallen allgemein erheblich kleiner aus als die auf Basis der *SL*-Heuristiken und des HBLR-Verfahrens berechneten *AAD*-Werte. Insbesondere wenn nur wenige Bewertungen im Trainingsdatensatz zur Verfügung stehen, erzielt man auf Basis der Regressorenkombination M_3 die besten Ergebnisse. In diesem Bereich führen komplexere Modelle zu deutlich größeren *AAD*-Werten. Die bei eher geringen Trainingsdatenmengen bei den komplexesten Modellen zu beobachtenden Überanpassungseffekte fallen erheblich schwächer aus als bei den Experimenten, bei denen alle Bewertungen, die nicht mehr im Testdatensatz ver-

AAD	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
100%	0,741	0,745	0,757	0,742	0,736	0,741	0,742	0,739	0,738
90%	0,736	0,742	0,751	0,735	0,733	0,735	0,736	0,735	0,730
80%	0,738	0,736	0,750	0,733	0,728	0,726	0,733	0,729	0,725
70%	0,736	0,737	0,750	0,732	0,728	0,732	0,731	0,729	0,728
60%	0,740	0,741	0,750	0,734	0,731	0,732	0,734	0,728	0,727
50%	0,744	0,747	0,757	0,736	0,734	0,731	0,740	0,735	0,736
40%	0,746	0,751	0,767	0,738	0,740	0,736	0,741	0,740	0,737
30%	0,755	0,763	0,775	0,745	0,750	0,745	0,751	0,750	0,754
20%	0,774	0,792	0,798	0,760	0,766	0,762	0,773	0,775	0,781
10%	0,793	0,817	0,833	0,778	0,780	0,781	0,788	0,789	0,789

Tabelle 9.19: *AAD*-Werte bei unterschiedlich großen Trainingsdatensätzen in Bezug auf das GP-Verfahren auf Basis der Variablenkombinationen M_0 bis und M_8 bezüglich D3 (2020 Nutzer, 418 Items)

R_B	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
100%	72,17	71,66	69,87	71,93	71,67	71,88	72,05	71,77	71,65
90%	72,13	71,54	69,88	71,94	71,68	71,85	72,02	71,50	71,71
80%	72,16	71,28	69,32	71,81	71,56	71,78	71,89	71,69	71,39
70%	72,21	70,85	69,34	71,90	71,47	71,78	71,72	71,47	71,52
60%	72,20	70,52	69,22	71,81	71,29	71,35	71,48	71,21	71,31
50%	72,19	70,17	68,53	71,83	70,95	71,45	71,35	71,05	71,06
40%	72,16	69,70	68,27	71,43	70,97	71,44	71,25	70,68	70,88
30%	72,20	69,75	67,71	70,93	70,76	71,11	71,12	70,18	70,21
20%	72,16	69,67	67,38	70,22	70,57	70,47	70,89	70,07	70,08
10%	72,13	68,68	66,68	70,26	69,45	69,98	69,72	69,95	69,78

Tabelle 9.20: Breese-Werte bei unterschiedlich großen Trainingsdatensätzen in Bezug auf das GP-Verfahren auf Basis der Variablenkombinationen M_0 bis und M_8 bezüglich D3 (2020 Nutzer, 418 Items)

wendet wurden, Bestandteil des Testdatensatzes waren (vgl. Abschnitt 5.10).
Der mittels der Breese-Werte geschätzte Nutzen der resultierenden Empfehlungslisten ist meist klein im Vergleich zum Nutzen der auf Basis des HBLR-Ansatzes

Prec./	verwendete Variablenkombination:								
Rec.	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
100%	0,573	0,524	0,541	0,568	0,566	0,564	0,558	0,549	0,564
	0,110	0,154	0,125	0,111	0,120	0,121	0,113	0,128	0,128
90%	0,560	0,530	0,540	0,566	0,558	0,556	0,554	0,541	0,580
	0,112	0,172	0,135	0,120	0,130	0,129	0,125	0,141	0,140
80%	0,563	0,521	0,531	0,570	0,559	0,569	0,555	0,553	0,563
	0,114	0,174	0,140	0,133	0,142	0,145	0,133	0,148	0,150
70%	0,562	0,522	0,524	0,580	0,543	0,553	0,555	0,544	0,552
	0,117	0,178	0,149	0,133	0,144	0,144	0,141	0,154	0,150
60%	0,548	0,513	0,516	0,542	0,548	0,533	0,548	0,541	0,546
	0,123	0,186	0,162	0,139	0,152	0,150	0,149	0,161	0,158
50%	0,544	0,502	0,499	0,545	0,534	0,544	0,534	0,524	0,527
	0,129	0,193	0,155	0,143	0,154	0,150	0,148	0,164	0,156
40%	0,527	0,483	0,471	0,515	0,531	0,527	0,524	0,516	0,527
	0,138	0,191	0,166	0,147	0,159	0,168	0,152	0,175	0,172
30%	0,504	0,473	0,466	0,493	0,498	0,508	0,511	0,494	0,483
	0,142	0,202	0,176	0,167	0,171	0,167	0,158	0,169	0,175
20%	0,465	0,453	0,432	0,461	0,472	0,469	0,464	0,464	0,460
	0,142	0,196	0,177	0,173	0,177	0,187	0,174	0,185	0,165
10%	0,384	0,397	0,390	0,392	0,392	0,392	0,394	0,401	0,395
	0,160	0,192	0,169	0,183	0,168	0,176	0,162	0,154	0,179

Tabelle 9.21: Präzision und Recall (untereinander) des GP-Verfahrens auf Basis unterschiedlicher Regressorenkombinationen und verschieden großer Trainingsdatenmengen hinsichtlich D3 (2020 Nutzer, 418 Items)

oder der $SL/\hat{S}^1_Y$-Heuristik bestimmten Empfehlungslisten. Genauere Schätzer führen nicht in jedem Fall auch zu den nützlichsten Empfehlungen. Die Tabellen 9.19 bis 9.21 enthalten die Ergebnisse des hybriden GP-Ansatzes auf Basis des etwas größeren Datensatzes D3 (2020 Nutzer, 418 Items). Tabelle 9.22 enthält die Ergebnisse des HBLR-Verfahrens bezüglich der Variablenkombinationen M_0 und M_8 auf der gleichen Datengrundlage. Im Rahmen des konservativen HBLR-Ansatzes erzielt man durch die zusätzliche Berücksichtigung weiterer Faktoren außer der durchschnittlichen Kritikerbewertung Ergebnisverbesserungen in Be-

verwendeter	Variablenkombination M_0				Variablenkombination M_8			
J^B-Datenanteil	AAD	$Prec.$	$Rec.$	R_B	AAD	$Prec.$	$Rec.$	R_B
100%	0,762	0,638	0,034	72,67	0,762	0,614	0,049	73,28
90%	0,763	0,621	0,032	72,67	0,763	0,597	0,050	73,18
80%	0,763	0,628	0,034	72,69	0,764	0,618	0,049	73,16
70%	0,765	0,608	0,036	72,72	0,767	0,612	0,041	72,92
60%	0,768	0,577	0,029	72,85	0,770	0,602	0,040	73,16
50%	0,770	0,641	0,032	72,92	0,772	0,600	0,039	73,10
40%	0,773	0,610	0,033	73,05	0,774	0,605	0,036	73,42
30%	0,775	0,624	0,037	72,64	0,780	0,612	0,040	73,14
20%	0,781	0,612	0,036	72,51	0,783	0,618	0,041	73,06
10%	0,788	0,536	0,031	72,40	0,788	0,542	0,039	72,95

Tabelle 9.22: AAD, Präzision, Recall und Breese-Werte bei unterschiedlich großen Trainingsdatensätzen (dh. verschieden großen Anteilen des Trainingsdatensatzes an der bezüglich der bekannten Items J^B im Hinbick auf die betrachteten Nutzer verfügbaren Datenmenge D3) für das HBLR-Verfahren zu den Variablenkombinationen M_0 und M_8

zug auf den geschätzten Nutzen der resultierenden Empfehlungslisten. Dies war im Rahmen des GP-Ansatzes nicht möglich. Dafür ließen sich auf Basis des GP-Verfahrens unter Verwendung der Variablenkombination M_3 kleinere AAD-Werte als mittels der Variablenmenge M_0 erreichen. Im Hinblick auf große bis mittlere Trainingsdatensätze lassen sich mit Hilfe der $SL/\hat{S}_Y^1$-Heuristik die höchsten Breese-Werte erzielen. Sofern nur wenige Trainingsdaten zur Verfügung stehen, führt die Verwendung des HBLR-Ansatzes zu höheren Breese-Werten.

Im Rahmen der linearen Bayes'schen hierarchischen Regression (HBLR) werden unter Verwendung von $Z = Z_1 = e_I$ die allgemeinen Tendenzen der Schätzer $\beta^i, i = 1, ..., I$, aller Nutzer einfach in einem Vektor Δ' gebündelt. Die durch das zweimodale $\hat{S}_Y^2$-Clusterverfahren auf Basis aller im D1-Datensatz enthaltenen Bewertungen, die sich auf die 300 bekannten Items beziehen, bestimmten Cluster-Zugehörigkeiten der Nutzer können durch die Wahl $Z = Z_2 = P$ bzw. $Z = Z_3 = (e_I P)$ anstelle von $Z = Z_1 = e_I$ zum Bestandteil des HBLR-Verfahrens gemacht werden (vgl. Abschnitt 8.3.2).

	R^2	AAD	$Prec.$	$Rec.$	R_B
Z_1	0,281	0,762	0,613	0,043	70,02
Z_2	0,284	0,760	0,644	0,051	70,52
Z_3	0,284	0,760	0,641	0,050	70,53

Tabelle 9.23: Ergebnisse des HBLR-Verfahrens bezüglich unterschiedlicher Wahlen für Z (Z_1, Z_2 und Z_3) unter Verwendung der Regressorenmenge M_3 auf Basis von D1

	Konstante	Action	Spannung	Charakter *	$\bar{Y}_{\cdot j}^{MC}$
allg.	**1,34** (0,96)	**-0,014** [0,68]	**0,028** (0,99)	**0,057** (1,00)	**0,371** (1,00)
$k=1$	**1,29** (0,99)	**0,045** (1,00)	**-0,017** [0,99]	**-0,030** [1,00]	**-0,181** [1,00]
$k=2$	**-0,71** [0,87]	**-0,062** [1,00]	**0,018** (0,98)	**0,043** (1,00)	**0,206** (1,00)
$k=3$	**-0,58** (0,68)	**-0,020** [0,95]	**0,006** (0,77)	**0,019** (0,97)	**0,234** (1,00)
$k=4$	**0,31** (0,69)	**-0,008** [0,77]	**0,004** (0,72)	-0,001 [0,55]	**0,045** (0,73)
$k=5$	**0,24** (0,87)	**0,004** (0,68)	**-0,005** [0,76]	**-0,012** [0,87]	**0,066** (0,79)
$k=6$	**0,62** (0,97)	**-0,045** [1,00]	**0,011** (0,92)	0,001 (0,58)	-0,005 [0,55]
$k=7$	**-0,76** [0,96]	**-0,011** [0,79]	0,001 (0,53)	**0,037** (0,99)	**0,230** [1,00]
$k=8$	**0,30** (0,68)	**-0,005** [0,69]	**0,010** (0,90)	**0,030** (0,98)	**-0,033** [0,65]
$k=9$	0,04 (0,48)	**0,039** (0,99)	**0,010** (0,94)	**-0,023** [0,99]	**0,041** (0,75)
$k=10$	**0,38** (0,74)	**0,039** (0,97)	0,002 (0,65)	**-0,005** [0,67]	0,005 (0,54)

Tabelle 9.24: Δ-Koeffizienten des HBLR-Verfahrens (Z_3-Prior) unter Verwendung der auf Basis des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens bezüglich der Bewertungen im Hinblick auf bekannte Items berechneten Nutzer-Cluster Zugehörigkeiten (D1)

Durch die zusätzliche Berücksichtigung der Cluster-Zugehörigkeiten der Nutzer via die Prior lassen sich immerhin geringe Ergebnisverbesserungen gegenüber der Standardwahl Z_1 erzielen (Tabelle 9.23).

Tabelle 9.24 zeigt die Δ-Koeffizienten des HBLR-Verfahrens (Z_3-Prior) auf Basis der mittels des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens hinsichtlich aller Bewertungen aus D1, die sich auf bekannte Items beziehen, berechneten Nutzer-Cluster Zugehörigkeiten. Die Werte in runden Klammern sind die Wahrscheinlichkeiten, mit der der jeweilige Koeffizient ein positives Vorzeichen hat. In eckigen Klammern stehen die Wahrscheinlichkeiten dafür, daß der betrachtete Koeffizient ein negatives Vorzeichen aufweist. Alle fettgedruckten Δ-Koeffizienten sind signifikant hinsichtlich eines Signifikanzniveaus von 5%. Die erste Zeile der Tabelle enthält die erste Zeile der Matrix Δ. Diese Zeile enthält die allgemeinen Tendenzen aller Nutzer. Alle übrigen Zeilen enthalten die allgemeinen Tendenzen innerhalb eines bestimmten Nutzer-Clusters.

Die Δ-Koeffizienten legen nahe, daß Filme dann eher höher bewertet werden, wenn sie von den Kritikern für gut gehalten werden, sich eingehend mit der Charakterentwicklung der handelnden Personen befassen und von den Zuschauern als spannend empfunden werden. Dabei hat ein stark ausgeprägter Actionfilmcharakter in vielen Fällen eher einen leicht negativen Effekt auf die Bewertung. Dies kann der ersten Zeile der Tabelle entnommen werden.

Die übrigen Zeilen enthalten Informationen hinsichtlich der einzelnen Nutzer-Cluster. Das erste Nutzer-Cluster ($k = 1$) hat von allen Nutzer-Clustern den höchsten Achsenabschnittsterm. Alle übrigen Koeffizienten hinsichtlich des ersten Nutzer-Clusters $\Delta_{2\kappa}, \kappa = 2, ..., 5$, haben exakt das entgegengesetzte Vorzeichen der entsprechenden Koeffizienten $\Delta_{1\kappa}, \kappa = 1, ..., 5$, im generellen Modell. Hierdurch kommt den Eigenschaften der Filme geringere Bedeutung zu. Das erste Nutzer-Cluster weist von allen Nutzer-Clustern die positivste Einstellung in Bezug auf den Grad des Actionfilmcharakters auf. Zudem scheinen die Nutzer aus dem ersten Cluster Filme besser zu bewerten, die sich eingehend mit der Charakterentwicklung der dargestellten Figuren befassen. Trotzdem scheinen die diesem Cluster zugeordneten Personen weniger Wert auf die Charakterentwicklung zu legen, als alle übrigen Cluster. Spannung scheint für die Personen aus diesem Nutzer-Cluster nicht besonders wichtig zu sein.

Tabelle 9.25 enthält die Gewichtematrix W für das zweimodale $\hat{S}_Y^2$-Clusterver-

k	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$	$l=7$	$l=8$	$l=9$	$l=10$
1	3,72	3,65	3,58	3,59	3,61	3,30	3,65	3,40	3,98	3,65
2	2,60	2,14	3,84	3,94	3,33	4,21	3,13	2,70	3,26	3,83
3	3,31	2,43	4,18	3,67	3,73	4,10	3,13	2,78	3,87	3,70
4	3,42	3,19	3,81	3,71	3,94	4,11	3,64	2,63	3,78	3,58
5	3,26	2,97	3,98	3,97	3,32	3,62	3,57	2,95	3,95	3,11
6	3,33	3,09	3,98	4,02	3,85	3,98	3,64	3,51	3,42	4,20
7	2,92	2,58	4,02	3,90	3,27	4,21	3,38	2,29	3,89	3,30
8	3,31	2,97	3,83	3,80	3,80	4,34	3,37	3,34	3,75	2,90
9	3,33	3,25	3,69	2,97	3,23	3,71	3,36	2,85	4,13	2,58
10	3,48	3,37	3,97	3,25	3,68	4,05	3,30	3,29	4,13	3,94

Tabelle 9.25: Gewichtematrix W für das zweimodale $\hat{S}_Y^2$-Clusterverfahren auf Basis der Bewertungen von 1067 Nutzern in Bezug auf 300 bekannte Filme (D1)

fahren auf Basis der Bewertungen von 1067 Nutzern in Bezug auf 300 bekannte Filme (D1). Der ersten Zeile der Matrix W kann entnommen werden, daß die zum ersten Nutzer-Cluster gehörenden Personen generell eher hohe Bewertungen und zudem recht ähnliche Bewertungen hinsichtlich aller Film-Cluster abgegeben haben. Das einzige Cluster, daß diese Nutzer geringfügig zu bevorzugen scheinen, ist Item-Cluster $l = 9$. Auf den ersten Blick hat es den Anschein, als ob diese Nutzer entweder keine besonders stark ausgeprägten Präferenzen haben oder diese Präferenzen durch ihre Tendenz, generell hohe Bewertungen abzugeben, gut verbergen. Ein Vergleich der Elemente der sechsten Spalte der Gewichtematrix W führt zu der Beobachtung, daß die Filme des sechsten Item-Cluster von fast allen Nutzer-Clustern durchschnittlich auffallend hoch bewertet wurden. Die niedrigste Durchschnittsbewertung stammt von dem ansonsten so großzügigen ersten Nutzer-Cluster. Gleichzeitig ist dies die niedrigste Durchschnittsbewertung in der gesamten ersten Zeile der Matrix W. Dies begründet die Vermutung, daß die Personen aus dem ersten Nutzer-Cluster sich von allen anderen Nutzern vor allem durch ihre weniger positive Einstellung hinsichtlich des sechsten Film-Clusters unterscheiden.

Im Hinblick auf das zweite Nutzer-Cluster sind sowohl die Δ-Koeffizienten als auch die Einträge der entsprechenden Zeile der Gewichtematrix erheblich aussagekräftiger. Die Δ-Koeffizienten belegen, daß die zum zweiten Nutzer-Cluster gehörenden Personen eine ausgeprägte Abneigung gegenüber Actionfilmelemeten

κ	Item-Cluster l										Max.	Min.
	1	2	3	4	5	6	7	8	9	10		
1	3,35	3,63	3,82	3,38	3,95	2,65	3,43	3,44	3,50	3,60	3,95	2,65
2	5,72	4,63	4,71	5,13	5,86	5,19	5,70	5,50	3,29	4,50	5,86	3,29
3	2,45	2,02	1,79	1,69	2,18	2,46	1,85	1,56	2,21	3,40	3,40	1,56
4	6,18	6,72	6,25	5,84	6,72	6,84	5,85	6,54	5,43	7,45	7,45	5,43
5	4,91	4,24	3,83	3,78	6,13	5,31	5,15	4,50	3,64	4,05	6,13	3,64
6	4,21	4,46	5,00	4,59	4,05	4,61	4,50	3,69	4,14	5,00	5,00	3,69
7	4,15	3,68	4,21	3,78	3,95	3,46	3,45	3,52	3,92	4,90	4,90	3,46
8	5,15	5,00	4,63	5,03	5,59	4,73	5,55	5,10	3,93	4,75	5,59	3,93
9	5,42	5,90	5,46	6,13	5,59	5,27	5,60	5,42	5,21	6,35	6,35	5,21
10	5,79	6,02	6,54	6,84	6,09	5,99	6,08	5,69	5,50	6,50	6,84	5,50
11	3,79	4,21	4,17	3,77	4,18	2,84	4,05	3,38	3,36	2,70	4,21	2,70
12	6,72	7,46	6,83	7,09	7,32	6,46	7,03	6,92	6,29	7,30	7,46	6,29
13	5,63	6,20	5,83	5,88	6,27	6,07	5,93	5,98	5,07	6,90	6,90	5,07
14	2,94	3,93	2,21	2,97	4,18	3,46	3,65	3,21	1,79	3,20	4,18	1,79

Tabelle 9.26: Durchschnittswerte der einzelnen Eigenschaften innerhalb verschiedener Item-Cluster in Bezug auf die 300 bekannten Filme

haben und sehr deutlich Filme bevorzugen, die von den professionellen Kritikern durchschnittlich hoch bewertet wurden und sich eingehend mit der Charakterentwicklung der handelnden Personen befassen. Der Grad der durch den Film evozierten Spannung scheint auch für die Mitglieder der zweiten Nutzer-Clusters von untergeordneter Bedeutung zu sein. Aus der Betrachtung der zweiten Zeile der Gewichtematrix folgt, daß das zweite Nutzer-Cluster Filme aus dem vierten und sechsten Item-Cluster bevorzugt und eine Abneigung gegenüber Filmen aus dem zweiten Item-Cluster zu haben scheint.

9.3.3 Ergebnisse der clusterspezifischen Modelle

Die direkte Schätzung auf Basis clusterspezifischer Modelle wurde bereits in Abschnitt 9.2.2 erläutert. Zuerst werden für jedes einzelne Nutzer-Cluster auf Basis der Gewichtematrix W diejenigen Item-Cluster identifiziert, deren Filme durchschnittlich von den Personen aus dem betrachtenen Nutzer-Cluster erkennbar besser oder schlechter bewertet wurden. Die Durchschnittswerte der einzelnen

Eigenschaften innerhalb verschiedener Item-Cluster in Bezug auf die 300 bekannten Filme werden zur Bestimmung von Eigenschaften, die in einem bestimmten Item-Cluster stärker oder schwächer ausgeprägt sind, benutzt. Diese Information wird kombiniert, um für jedes Nutzer-Cluster Eigenschaften zu finden, die möglicherweise im Hinblick auf das betrachtete Nutzer-Cluster ausschlaggebend für das Ge- oder Mißfallen der Items sind. Die auf diese Weise ermittelten Eigenschaften werden dann in Bezug auf das betrachtete Nutzer-Cluster auf ihre Signifikanz (gegebenenfalls neben anderen aus anderen Quellen bekannten möglicherweise relevanten Eigenschaften) überprüft. Im Hinblick auf jedes Nutzer-Cluster werden nur die in Bezug auf das betrachtete Nutzer-Cluster relevanten Variablen ins zugehörige clusterspezifische Modell aufgenommen.

Grundsätzlich gilt, daß es möglich ist, daß Nutzer dieselbe Eigenschaften im Hinblick auf verschiedene Gruppen von Items unterschiedlich bewerten können. Beispielsweise ist es möglich, daß einem Nutzer, der Thriller und Komödien mag, im Hinblick auf einen Thriller Spannung sehr wichtig und Komik eher unwichtig ist, während es ihm in Bezug auf Komödien nicht auf den Grad der erzeugten Spannung aber sehr wohl auf den Grad an Komik ankommt. Wenn ein Nutzer-Cluster zwei verschiedene Item-Cluster besonders gut bewertet und eine bestimmte Eigenschaft in Bezug auf das erste Item-Cluster stark ausgeprägt ist, spricht es daher nicht gegen die Relevanz dieser Eigenschaft, wenn sie in dem anderen hochbewerteten Item-Cluster nur schwach ausgeprägt ist. Dagegen würde gegen die Relevanz der Eigenschaft sprechen, wenn ein drittes Item-Cluster von den Personen aus dem betrachteten Nutzer-Cluster überwiegend schlecht bewertet worden wäre und dieses die betreffende Eigenschaft auch in starkem Maße aufweist.

Tabelle 9.26 enthält die Durchschnittswerte der einzelnen Eigenschaften innerhalb verschiedener Item-Cluster in Bezug auf die 300 bekannten Filme. Auf Basis der in Tabellen 9.25 und 9.26 enthaltenen Information lassen sich möglicherweise relevante Eigenschaften hinsichtlich aller Nutzer-Cluster bestimmen. Das genaue Vorgehen wird im folgenden zunächst am Beispiel des zweiten Nutzer-Clusters und dann am Beispiel des schwierigeren ersten Nutzer-Clusters erläutert.

Die Analyse der Gewichtematrix ergibt, daß hinsichtlich des zweiten Nutzer-Clusters die Item-Cluster $l = 4$ und $l = 6$ die höchsten Durchschnittsbewertungen aufweisen und das Item-Cluster $l = 2$ besonders schlecht bewertet wurde.

Tabelle 9.26 ist zu entnehmen, daß das Item-Cluster $l = 4$ in Bezug auf den Grad der dargestellten Charakterentwicklung ($\kappa = 10$) den höchsten Durch-

schnittswert aufweist. Hinsichtlich der beiden Cluster $l = 2$ und $l = 6$ scheint der Grad der dargestellten Charakterentwicklung im Vergleich zu den übrigen Clustern nicht besonders wichtig zu sein. Daher könnte der Grad der dargestellten Charakterentwicklung bezüglich des zweiten Nutzer-Clusters eine relevante Eigenschaft sein. Der Actionfilmcharkter ($\kappa = 5$) scheint in Bezug auf die Filme aus Cluster $l = 4$ schwach ausgeprägt zu sein. Weder Cluster $l = 2$ noch Cluster $l = 6$ weisen im Vergleich zu den anderen Item-Clustern besonders stark ausgeprägte Werte auf. Daher könnte sich der Actionfilmcharakter ($\kappa = 5$) als relevante Eigenschaft erweisen.

Das sechste Item-Cluster weist einen relativ zu den übrigen Werten hohen (6,84) wenngleich auch nicht den höchsten Wert hinsichtlich der Mainstream-Tendenzen ($\kappa = 4$) auf. Cluster $l = 4$ hat einen vergleichsweise niedrigen Wert in Bezug auf diese Eigenschaft. Der Wert des zweiten Item-Clusters fällt ähnlich aus wie der des sechsten Clusters. Da das zweite Item-Cluster überwiegend aus Filmen besteht, die den Personen aus dem zweiten Nutzer-Cluster zu mißfallen scheinen, spricht dies gegen die Relevanz der Mainstream-Tendenzen.

Cluster $l = 2$ hat von allen Item-Clustern die höchsten Werte hinsichtlich des Grads der cinematographischen Perfektion ($\kappa = 12$) und eine vergleichsweise hohen Wert in Bezug auf den Grad an Exzentrität ($\kappa = 11$). Im Hinblick auf den Grad der cinematographischen Perfektion weist auch das vierte Item-Cluster einen hohen Wert auf. Da zum einen der Grad der cinematographischen Perfektion zwischen den Item-Clustern nur schwach variiert und zum anderen dieser Wert in Bezug auf drei Cluster merklich höher als in Bezug auf das vierte Cluster ausfällt, ist ein hoher Grad an cinematographischen Perfektion nicht charakteristisch für Item-Cluster $l = 4$. Cluster $l = 6$ weist einen vergleichsweise niedrigen Grad an cinematographischer Perfektion auf. (Dagegen ergibt sich hinsichtlich der Filme aus dem zweiten Item-Cluster der höchste Durchschnittswert in Bezug auf diese Eigenschaft.) Da somit ein hohes Maß an cinematographischer Perfektion charakteristisch für ein vom zweiten Nutzer-Cluster schlecht bewertetes Item-Cluster ist und zudem ein niedriges Maß an cinematographischer Perfektion ($\kappa = 12$) eine Besonderheit eines von diesen Personen mit hohen Bewertungen versehenen Item-Clusters ist, könnte sich der Grad der cinematographischen Perfektion in Bezug auf das erste Nutzer-Cluster als relevante Eigenschaft erweisen. Der Grad an Exzentrität ($\kappa = 11$) hat in Bezug auf das sechste Item-Cluster einen vergleichsweise geringen Wert und hinsichtlich des Clusters $l = 4$ zumindest kei-

nen vergleichsweise hohen Wert. Daher könnte auch der Grad an Exzentrität für die Personen aus dem zweiten Nutzer-Cluster relevant sein.

Diese Betrachtungen führen auf den Actionfilmcharakter, den Grad der dargestellten Charakterentwicklung, den Grad an Exzentrität und die cinematographische Perfektion als Menge der möglicherweise relevanten Variablen. Bis auf den Grad der cinematographischen Perfektion erweisen sich alle diese Variablen als (hinsichtlich der von Personen aus dem zweiten Nutzer-Cluster in Bezug auf die bekannten Items abgegebenen Bewertungen) signifikant.

Im Vergleich zu den übrigen Nutzer-Clustern fallen die Bewertungen des ansonsten eher großzügig bewertenden ersten Nutzer-Clusters in Bezug auf Item-Cluster $l = 6$ auffallend niedrig aus. Außerdem haben die Nutzer aus dem ersten Cluster das Item-Cluster $l = 6$ von allen Item-Clustern am schlechtesten bewertet. Obwohl der absolute Wert des Koeffizienten w_{16} nicht besonders klein ist, kann dies ein Hinweis auf eine Abneigung des ersten Nutzer-Clusters hinsichtlich der Filme des sechsten Item-Clusters sein. In Bezug auf Item-Cluster $l = 6$ ist ein hoher Grad an Mainstream-Tendenzen ($\kappa = 4$) charakteristisch. Das einzige andere Cluster was in diesem Zusammenhang relevant sein könnte ist das neunte Item-Cluster. Dieses wurde von den Personen aus dem ersten Nutzer-Cluster geringfügig besser beurteilt als die übrigen Item-Cluster. Es weist hinsichtlich vieler Eigenschaften minimale Durchschnittswerte auf. Hieraus kann wenig gefolgert werden.

Da eine möglicherweise relevante Variable neben der durchschnittlichen Kritikerbewertung etwas wenig ist, sollte an dieser Stelle Information aus anderen Quellen miteinbezogen werden. Da die Ergebnisse der vorherigen Bayes'schen Regression (Tabelle 6.23) belegen, daß der Actionfilmcharakter und der Grad an dargestellter Charakterentwicklung relevante Eigenschaften sind, sollten diese beiden Variablen nicht vernachlässigt werden.

Generell sollte es nicht dabei belassen werden, die Signifikanz einer Variable hinsichtlich eines Nutzer-Clusters zu belegen. Insbesondere sollte das Vorzeichen des entsprechenden Δ-Koeffizienten auch mit dem Argument für die mögliche Relevanz dieser Eigenschaft in Bezug auf das betrachtete Nutzer-Cluster konsistent sein. Falls eine Variable zwar signifikant ist aber die zugehörige Komponente des Schätzers ein vor dem Hintergrund der vorherigen Überlegungen unerwartetes Vorzeichen aufweist, ist dies ein Hinweis darauf, daß noch nicht alle relevanten Einflußgrößen entdeckt sind.

k	signifikante Variablen
1	Grad der Mainstream-Tendenzen, Actionfilmcharacter, Grad der dargestellten Charakterentwicklung
2	Actionfimcharacter, Grad der Exzentrität, Grad der dargestellten Charakterentwicklung
3	Actionfilmcharakter, Grad der erzeugten Spannung, Grad der Exzentrität, Grad der dargestellten Charakterentwicklung
4	Grad an Komik, Grad an erzeugter Spannung, Grad an Exzentrizität, Grad der dargestellten Characterentwicklung
5	Grad der dargestellten Gewalt, Grad der dargestellten Charakterentwicklung, Grad der Exzentrizität
6	Actionfilmcharakter, Grad der erzeugten Spannung, Grad des Exzentrizität, Grad der dargestellten Charakterentwicklung
7	Grad der dargestellten Charakterentwicklung, Grad der Exzentrizität
8	Grad der dargestellten Charakterentwicklung, Grad der Exzentrizität
9	Grad der Familienfreundlichkeit, Grad der Mainstream-Tendenzen, Grad der erzeugten Spannung, Grad der dargestellten Charakterentwicklung, Grad der cinematographischen Perfektion
10	Grad der Mainstream-Tendenzen, Grad der erzeugten Spannung, Grad der dargestellten Charakterentwicklung, Grad der cinematographischen Perfektion

Tabelle 9.27: Regressoren der clusterspezifischen Modelle (zusätzlich zu den in dieser Abbildung aufgelisteten Variablen enthält jedes Modell die mittlere durchschnittliche Kritikerbewertung)

Den Nutzern aus dem ersten Cluster scheinen die Items aus dem sechsten Item-Cluster nicht zu gefallen. Da der Grad der Mainstream-Tendenzen hinsichtlich dieses Cluster die einzige besonders stark ausgeprägte Eigenschaft war, sollte das Vorzeichen des entsprechenden Δ-Koeffizienten negativ sein. Falls der Grad der Mainstream-Tendenzen sich als signifikant erweist aber der zugehörige Δ-Koeffizient positiv ist, erklärt der Grad der Mainstream-Tendenzen nicht die Unbeliebtheit des sechsten Item-Clusters bei den Personen aus dem ersten Nutzer-Cluster. In diesem Fall sollte man versuchen, den wahren Grund dieser Unbeliebtheit zu finden, da hierdurch das Regressionsmodell verbessert werden kann. Da in diesem Beispiel keine andere Eigenschaft einen verhältnismäßig

R^2	AAD	$Prec.$	$Rec.$	R_B
0,305	0,748	0,638	0,089	71,44

Tabelle 9.28: Ergebnisse des kombinierten Modells

hohen Item-Cluster Durchschnittswert aufweist, kann man in diesem Fall ohne zusätzliche Information nur versuchen, Eigenschaften mit besonders niedrigen Ausprägungen hinsichtlich des sechsten Item-Clusters zu identifizieren. Hierbei ist allerdings zu bedenken, daß die Nutzer dieselbe Eigenschaft in Bezug auf eine Gruppe von Filmen schätzen können, während sie ihnen hinsichtlich einer anderen Filmgruppe eher gleichgültig sind. Daher ist die besonders niedrige Ausprägung einer Eigenschaft immer ein deutlich schwächeres Argument für die Relevanz einer Eigenschaft.

Auf diese Weise lassen sich die Ergebnisse einer vorherigen Regression und die in der Gewichtematrix W enthaltene zusätzliche Information dazu benutzen, für jedes der zehn verschiedenen Nutzer-Cluster ein separates (grobes) Modell zu bestimmen. Eine Aufstellung der in den einzelnen Modellen enthaltenen Regressoren findet sich in Tabelle 9.27.

Da die Nutzer-Cluster disjunkte Mengen sind, ist es möglich, die in Tabelle 9.27 skizzierten clusterspezifischen Modelle zu einem Modell für die gesamten Daten zu vereinen. Die Ergebnisse des resultierenden kombinierten Modells sind Tabelle 9.28 zu entnehmen.

Die beschriebene Erzeugung clusterspezifischer Modelle setzt neben den Ergebnissen eines Clusterverfahrens auch darüberhinausgehende Information wie die in Tabelle 9.24 dargestellten Ergebnisse einer vorherigen Regression voraus. Durch die dargestellte Methode lassen sich wichtige Variablen identifizieren. Es kann aber nicht davon ausgegangen werden, daß dieses Verfahren in der Lage ist, alle wichtigen Variablen zu identifizieren. Außerdem ist die vorgestellte Methode nur zur Bestimmung grober (möglicherweise unvollständiger) Modelle geeignet, die den Ausgangspunkt weiterer Modellbestimmungsüberlegungen bilden können.

Durch die Verwendung eines aus clusterspezifischen Modellen kombinierten Modells lassen sich immerhin bessere Ergebnisse erzielen als auf Basis aller bisher betrachteten HBLR-Modelle $M_0, M_1, M_2, M_3, M_4, M_5, M_6, M_7$ und M_8. Nichtsdestotrotz ist die wesentlich einfachere $SL/\hat{S}_Y^1$-Heuristik unter Verwendung der Variablenbombinationen M_3 und M_8 auch diesem Ansatz insgesamt überlegen. Es

läßt sich zwar durch Verwendung des kombinierten Modells ein etwas geringeres AAD erzielen, dafür fallen aber sowohl der Breese-Wert als auch der Recall im Hinblick auf die $SL/\hat{S}_Y^1$-Heuristik auf Basis der Variablenbombinationen M_3 und M_8 deutlich größer aus als unter Verwendung des kombinierten Modells. Hinsichtlich der Variablenkombinationen M_0 und M_2 erzielt man jedoch nicht nur wesentlich höhere AAD-Werte als auf Basis des kombinierten Bayes'schen Modells sondern auch deutlich geringere Breese-Werte. Daher ist die $SL/\hat{S}_Y^1$-Heuristik dem kombinierten Bayes'schen Modell nur dann überlegen, wenn die richtige Variablenkombination verwendet wird. Dafür ist die $SL/\hat{S}_Y^1$-Heuristik einfacher, numerisch weniger aufwendig, vollständig automatisiert (somit weniger arbeitsintensiv und weniger fehleranfällig) und erfordert keine Vorinformation bis auf die zu verwendende Variablenkombination.

9.3.4 Herleitung des optimalen Verfahrens

Da sich die Ergebnisse der $SL/\hat{S}_Y^1$-Heuristik durch die Wahl der optimalen Variablenkombination erheblich verbessern lassen, ist es sinnvoll, verschiedene Möglichkeiten zur Bestimmung der optimalen Variablenkombination auf Basis der Trainingsdaten zu vergleichen.

In vielen Fällen kann eine gelungene Variablenselektion zu besseren und zuverlässigeren Ergebnissen führen. Bereits in Abschnitt 9.3.2 wurden Methoden zur Variablenselektion im Zusammenhang mit den Methoden der direkten Schätzung erörtert.

Bezüglich der $SL/\hat{S}_Y^1$-Heuristik sollte die zu verwendende Variablenkombination direkt auf der Basis der $SL/\hat{S}_Y^1$-Heuristik bestimmt werden. Hierzu können die Daten, auf deren Grundlage die Schätzer berechnet werden sollen, in zwei disjunkte Mengen unterteilt werden. Auf Basis der größeren dieser Datenmengen lassen sich zunächst unter Verwendung einer möglichst großen Anzahl von in Betracht kommenden Variablen Schätzer mittels der $SL/\hat{S}_Y^1$-Heuristik berechnen. Der andere Teil der Daten, auf deren Grundlage die Schätzer zu berechnen sind, kann als Kalibrierungsmenge dazu verwendet werden, einen Eindruck von der Wirksamkeit der unterschiedlichen Variablenkombinationen zu gewinnen.

AAD	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
100%	0,769	0,773	0,765	0,756	0,751	0,755	0,751	0,770	0,759
90%	0,777	0,791	0,772	0,749	0,766	0,752	0,765	0,771	0,765
80%	0,790	0,775	0,773	0,754	0,759	0,769	0,743	0,765	0,790
70%	0,789	0,782	0,778	0,763	0,766	0,759	0,759	0,773	0,775
60%	0,803	0,789	0,779	0,764	0,769	0,758	0,760	0,769	0,782
50%	0,796	0,789	0,788	0,770	0,784	0,775	0,772	0,793	0,781
40%	0,799	0,793	0,816	0,805	0,781	0,783	0,773	0,780	0,799
30%	0,811	0,812	0,811	0,813	0,800	0,814	0,792	0,803	0,809
20%	0,857	0,865	0,833	0,844	0,849	0,825	0,834	0,841	0,844
10%	0,937	0,961	0,928	0,898	0,946	0,930	0,921	0,924	0,920

Tabelle 9.29: AAD-Werte bei unterschiedlich großen Trainingsdatensätzen in Bezug auf die $SL/\hat{S}_Y^1$-Heuristik auf Basis der Variablenkombinationen M_0 bis und M_8 auf der Datengrundlage D1 (1067 Nutzer, 418 Items)

R_B	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
100%	71,53	71,00	70,77	73,75	73,80	72,39	73,72	72,38	72,26
90%	70,21	70,88	71,92	73,51	72,71	72,18	73,63	71,32	72,16
80%	70,37	70,78	71,28	74,14	72,94	72,44	73,99	73,29	69,53
70%	69,58	71,43	71,11	72,62	72,76	73,06	73,35	72,54	71,68
60%	67,04	69,89	71,68	73,09	72,88	72,79	73,77	72,81	71,98
50%	68,90	70,84	72,07	73,36	72,78	73,07	73,30	72,59	71,80
40%	68,22	71,94	70,28	70,85	72,39	72,10	72,92	73,17	71,36
30%	67,50	69,90	69,29	70,25	71,30	69,87	72,49	70,79	71,38
20%	65,93	68,04	69,47	68,03	68,19	69,42	70,24	69,84	69,39
10%	63,17	65,09	65,86	68,80	65,22	66,40	66,77	63,72	65,64

Tabelle 9.30: Breese-Werte (R_B) bei unterschiedlich großen Trainingsdatensätzen in Bezug auf die $SL/\hat{S}_Y^1$-Heuristik auf Basis der Variablenkombinationen M_0 bis und M_8 bezüglich D1 (1067 Nutzer, 418 Items)

Analog zur Steinmetz-Methode können sukzessive Variablen weggelassen werden, wenn sich dadurch die auf Basis der Kalibrierungsdaten bestimmten Breese-Werte verbessern. Sobald keine weiteren Verbesserungen mehr durch Weglassen von Va-

Prec./	verwendete Variablenkombination:								
Rec.	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
100%	0,569	0,705	0,597	0,647	0,648	0,591	0,634	0,609	0,684
	0,139	0,125	0,149	0,345	0,180	0,144	0,133	0,163	0,145
90%	0,624	0,679	0,698	0,588	0,616	0,517	0,608	0,605	0,571
	0,135	0,137	0,130	0,291	0,140	0,055	0,141	0,143	0,146
80%	0,692	0,665	0,664	0,545	0,619	0,608	0,634	0,602	0,553
	0,107	0,132	0,121	0,320	0,203	0,135	0,185	0,187	0,114
70%	0,598	0,668	0,647	0,526	0,600	0,667	0,653	0,624	0,556
	0,113	0,152	0,168	0,291	0,189	0,157	0,137	0,181	0,078
60%	0,680	0,627	0,706	0,543	0,643	0,611	0,625	0,607	0,529
	0,077	0,069	0,109	0,307	0,179	0,211	0,160	0,224	0,068
50%	0,531	0,735	0,611	0,525	0,557	0,568	0,675	0,582	0,634
	0,137	0,124	0,132	0,320	0,070	0,134	0,071	0,060	0,116
40%	0,505	0,675	0,646	0,469	0,618	0,635	0,622	0,561	0,576
	0,174	0,147	0,152	0,356	0,193	0,082	0,183	0,203	0,208
30%	0,517	0,637	0,703	0,466	0,605	0,487	0,559	0,576	0,490
	0,112	0,108	0,079	0,257	0,125	0,084	0,154	0,176	0,091
20%	0,551	0,604	0,551	0,391	0,532	0,511	0,561	0,658	0,510
	0,051	0,070	0,142	0,218	0,123	0,143	0,118	0,104	0,151
10%	0,421	0,432	0,468	0,355	0,352	0,364	0,303	0,346	0,369
	0,152	0,115	0,136	0,319	0,133	0,156	0,047	0,138	0,191

Tabelle 9.31: Präzision und Recall der $SL/\hat{S}_Y^1$-Heuristik auf Basis unterschiedlicher Regressorenkombinationen und verschieden großer Trainingsdatenmengen auf Basis von D1 (1067 Nutzer, 418 Items)

riablen zu erreichen sind, sollten analog zur Mauerer-Methodologie sukzessive weitere Variablen hinzugefügt werden, sofern sich dadurch die auf der Grundlage der Kalibrierungsdaten berechneten Breese-Werte vergrößern lassen, bis sich auch durch Hinzufügen weiterer Variablen keine Ergebnisverbesserungen mehr erreichen lassen. Hierzu kann iterativ vorgegangen werden. Zunächst versucht man, die Variablenmenge solange durch Weglassen der Variable, ohne die der höchste Breese-Wert erreicht werden kann, zu verkleinern, bis sich der Breese-Wert nicht mehr durch Entfernen von Variablen erhöhen läßt (Schritt 1). Danach

wird solange sich der Breese-Wert durch die Hinzunahme einer weiteren Variablen erhöhen läßt, die resultierende Variablenmenge durch Hinzufügen derjenigen Variable, deren Aufnahme in die verwendete Variablenmenge mit der stärksten Erhöhung des Breese-Werts verbunden ist, vergrößert (Schritt 2). Man iteriert beide Schritte solange, bis sich entweder keine oder nur noch vernachlässigbare Ergebnisverbesserungen erzielen lassen.

Die sich auf die bekannten 300 Items beziehenden Bewertungen aus dem Datensatz D1 wurden hierzu weiter unterteilt. Von den 300 bekannten Items wurden zunächst 50 Items per Zufallszahl ausgewählt. 30 % der in Bezug auf diese 50 Items abgegebenen Bewertungen wurden als Kalibrierungsdaten eingesetzt. Weiterhin dienten 30 % der übrigen Daten zur Berechnung der Schätzer. Das oben beschriebene Vorgehen führt auf dieser Datenbasis zur Variablenkombination M_3 als optimale Variablenmenge.

Auf Basis der zufällig selektierten 30 % der in Bezug auf die 118 neuen Items abgegebenen Bewertungen läßt sich die Eignung unterschiedlicher Variablenkombinationen zur Generierung von Empfehlungslisten (oder zur Vorhersage) überprüfen. In den Tabellen 9.29, 9.30 und 9.31 sind die resultierenden Ergebnisse für verschiedene Variablenkombinationen aufgelistet. Die Variablenkombinationen M_3 und M_6 scheinen am besten geeignet zur Vorhersage von Bewertungen hinsichtlich neuer Items und auch zur Generierung von Empfehlungslisten in Bezug auf neue Items zu sein. (Modell M_3 und M_6 unterscheiden sich nur um eine einzige Variable.) Vor dem Hintergrund dieser Ergebnisse scheint die hier dargestellte Methode zur Bestimmung einer im Hinblick auf das heuristische $SL/\hat{S}_Y^1$-Verfahren geeigneten Variablenkombination ihren Zweck im Hinblick auf diese Heuristik nicht schlechter zu erfüllen als eine MCMC Modellbestimmung in Verbindung mit Methoden der Versuchsplanung hinsichtlich des HBLR-Ansatzes. Da die dargestellte Kombination aus Steinmetz- und Maurer-Methode im Vergleich zu einer MCMC Modellbestimmung nur geringen numerischen Aufwand erfordert, ist dies ein weiterer Vorteil der $SL/\hat{S}_Y^1$-Heuristik.

Allerdings ist in diesem Zusammenhang erwähnenswert, daß die Ergebnisse auf Basis des konservativen HBLR-Verfahrens kaum von der verwendeten Variablenmenge abzuhängen scheinen, solange in dieser wenigstens die wichtigsten Variablen enthalten sind. Daher kann bei der Verwendung des konservativen HBLR-Ansatzes in vielen Fällen auf eine Variablenselektion verzichtet werden.

In Bezug auf das konservative HBLR-Verfahren kann eine Menge geeigne-

ter Regressoren mittels der MCMC Modellbestimmung in Verbindung mit einem balancierten unvollständigen Blockplan bestimmt werden. Die Verwendung der DIC-Werte zur Auswahl der Regressoren kann im Hinblick auf den durch die Breese-Werte geschätzten Nutzen zu suboptimalen Ergebnissen führen. Hinsichtlich der $SL/\hat{S}_Y^1$-Heuristik ist die beschriebene iterative Verbindung der Steinmetz-Methode mit dem Maurer-Vorgehen zu empfehlen.

Insgesamt lassen sich in Bezug auf die neuen Items mittels des hybriden GP-Verfahrens die kleinsten AAD-Werte erreichen. Trotzdem führt das hybride GP-Verfahren in Bezug auf neue Items zu vergleichsweise schwachen Breese-Werten. Falls genügend Bewertungen hinsichtlich bekannter Items zur Berechnung der Schätzer zur Verfügung stehen, erweist sich vor allen Dingen in Bezug auf den Breese-Wert die $SL/\hat{S}_Y^1$-Heuristik als vorteilhaft.

Da es insbesondere wichtig ist, Nutzern bereits aufgrund weniger Bewertungen hilfreiche Empfehlungen gerade auch in Bezug auf neue Items machen zu können, kommt dem konservativen HBLR-Ansatz trotz des mit ihm verbundenen hohen numerischen Aufwands eine gewisse praktische Bedeutung zu.

Falls die meisten Nutzer eines Empfehlungssystems nur wenige Bewertungen abgegeben haben, ist vor dem Hintergrund der Empfehlungsqualität die Verwendung des konservativen HBLR-Ansatzes zu empfehlen. Dies wird vor allem in der Einführungsphase des Systems vorteilhaft sein.

Zu bedenken ist aber auch, daß die Zeit, die ein Recommender-System zur Generierung seiner Empfehlungen benötigt, ebenso wie die Empfehlungsqualität ein wichtiger Erfolgsfaktor ist. Gelingt es dem System nicht, einem Nutzer zeitnah zur Abgabe seiner Bewertungen sinnvolle Empfehlungen zu unterbreiten, könnte sich dieser von dem System abwenden.

Bezüglich der Teilmenge aus dem D1-Datensatz, welche die Bewertungen hinsichtlich der 300 bekannten Items enthält, ergeben sich mittels des HBLR-Ansatzes erst, wenn nur noch 20 % dieser Daten zum Training der Verfahren verwendet werden, bessere Ergebnisse als auf Basis der $SL/\hat{S}_Y^1$-Heuristik (bezüglich der Variablenkombination M_3). Die in diesem Bereich bestimmten Breese-Werte sind etwas höher. Der in diesem Bereich mit der Verwendung des HBLR-Verfahrens verbundene mäßige Gewinn bezüglich der Vorhersagequalität rechtfertigt nicht die hiermit einhergehende Erhöhung der Reaktionszeit, die das Recommender-System benötigt, um auf Basis der Bewertungen Empfehlungen zu ermitteln. Im Hinblick auf die Variablenkombination M_0 ergeben sich bereits

wenn noch 60 % dieses Teils der D1-Daten zum Training der Verfahren verwendet wird auf Basis des $SL/\hat{S}_Y^1$-Verfahrens Ergebnisse, die erheblich schlechter als die des HBLR-Verfahrens sind. Daher sollte die $SL/\hat{S}_Y^1$-Methode dem HBLR-Verfahren nur vorgezogen werden, wenn genügend Daten vorhanden sind, um vorab eine Variablenselektion durchzuführen. Dies ist zumindest dann erfüllbar, wenn bezüglich der betrachteten Item-Kategorie gut vergleichbare Bewertungen verfügbar sind. Hinsichtlich des HBLR-Verfahrens sind die resultierenden Empfehlungen im Hinblick auf neue Items weniger stark von der verwendeten Variablenmenge und der Anzahl der zur Schätzung nutzbaren Bewertungen abhängig. Somit erscheint nur in den Fällen, in denen keine oder zu wenig Bewertungen verfügbar sind, die Verwendung des konservativen HBLR-Verfahrens im Zusammenhang mit der Empfehlung neuer Items gerechtfertigt. Selbst in diesen Fällen ist die (temporäre) Verwendung des HBLR-Ansatzes nicht die einzige Option. In diesem Zusammenhang ist die folgende einfache Überlegung hilfreich. In dem Teil des D1-Datensatzes, der sich auf die 300 bekannten Items bezieht, sind ca. 70000 Bewertungen enthalten. Folglich sind pro Nutzer durchschnittlich 66 Bewertungen verfügbar. Zumindest, wenn 30 % dieses Datensatzes vorhanden sind, können mittels der $SL/\hat{S}_Y^1$-Heuristik (auf Basis der Variablenmenge M_3) vergleichsweise nützliche Empfehlungen abgegeben werden (vgl. Tabelle 9.10). Hierzu sind im betrachteten Datensatz durchschnittlich ca. 20 Bewertungen pro Nutzer erforderlich. Daraus folgt in Bezug auf den verwendeten Datensatz, daß bereits auf Basis von 20 Bewertungen hinsichtlich einer vorgegebenen Menge von 300 bekannten Items, Vorhersagen mittels der $SL/\hat{S}_Y^1$-Heuristik (für beliebig viele unbekannte Items) gemacht werden können. Diese Vorhersagen übertreffen im Hinblick auf die resultierenden Breese-Werte die entsprechenden Prognosen auf Basis aller übrigen zur Vorhersage neuer Items eingesetzten Verfahren.

Im allgemeinen sollte es einem Online-Shop möglich sein, an der Benutzung eines Recommender-Systems interessierte Kunden durch wenig kostenintensive Anreize wie die Teilnahme an einer Verlosung oder die einmalige Gewährung eines geringen Rabatts zur Abgabe von 20 Bewertungen zu bewegen. Hierdurch könnte von Anfang an ein Mindestmaß an Zuverlässigkeit und Nützlichkeit der personalisierten Empfehlungen erreicht werden.

Macht man einen Nutzer oder Kunden auf ein bestimmtes Produkt im Rahmen einer Direkt-Marketing Aktion aufmerksam, so nimmt man zunächst die Zeit dieser Personen in Anspruch. Es muß davon ausgegangen werden, daß letz-

tere dies nur solange als akzeptabel empfinden, wie der Nutzen der Angebote bzw. Empfehlungen sie für den Zeitverlust entschädigt. Es ist fraglich, ob die Nutzer eines Online-Geschäfts zwischen persönlichen Empfehlungen des zum Geschäft gehörenden Empfehlungssystems und Direkt-Marketing Aktionen differenzieren. Daher könnte eine unglückliche Direkt-Marketing Aktion durchaus das Empfehlungssystem in Mißkredit bringen. Vor allen Dingen können auch unglückliche Offerten oder häufige gezielte Werbeabtionen zu Unzufriedenheit mit dem betreffenden Online-Geschäft führen. Deshalb gelten für Direkt-Marketing Aktionen dieselben Kriterien wie für die Empfehlungen eines Recommenders: Vor allen Dingen sollten unzutreffende Empfehlungen (bzw. unglückliche Angebote) vermieden werden. Auch hier kommt es darauf an, den Nutzen der Empfehlungen zu maximieren. Präzision ist wichtiger als Recall. Daher eignen sich dieselben Verfahren als Ausgangspunkt von Direkt-Marketing Aktionen, die auch hinsichtlich eines Empfehlungssystems optimal sind.

Die $SL/\hat{S}^1$-Heuristik scheint für Marktforschungszwecke vergleichsweise gut geeignet zu sein. Hierfür sprechen die Präzisions- und Recall-Werte. Da für Marktforschungszwecke vor allem die möglichst korrekte Vorhersage von Höchstbewertungen von Interesse ist, stehen hier Präzision und Recall im Vordergrund. Bei Verwendung der Variablenmenge M_3 erhält man mittels der $SL/\hat{S}^1_Y$-Heuristik zwar eine niedrigere Präzision aber dafür auch einen sehr viel größeren Recall-Wert als man bei Verwendung des konservativen HBLR-Ansatzes oder des hybriden GP-Verfahrens erhalten würde. Da es im Hinblick auf einige Marktforschungsfragestellungen durchaus auch auf den Recall-Wert ankommen kann, legt dies die Vermutung nahe, daß die $SL/\hat{S}^1$-Heuristik auch in dieser Hinsicht geeigneter als die genannten Bayes'schen Verfahren ist.

Es sind allerdings auch Marktforschungsfragestellungen vorstellbar, hinsichtlich derer der Recall sogar ebenso wichtig wie die Präzision ist.

Ein Beispiel für eine Problemstellung aus dem Bereich der Markforschung, bei der es gleichermaßen auf Präzision wie auf den Recall ankommt, ist eine Abschätzung des prozentualen Anteils der Nutzer, die an einem eventuell neu ins Sortiment eines Online-Geschäfts einzuführenden Item interessiert sein könnten. Hierbei wird eine prognostizierte Höchstbewertung eines Nutzers in Bezug auf ein Item als vorhergesagtes Interesse des jeweiligen Nutzers an dem betreffenden Item interpretiert.

In diesen Fällen kann zur empirischen Beurteilung der Eignung der betreffen-

J^B-Datenanteil	$SL/\hat{S}_Y^1$	$SL/\hat{S}_Y^2$	SL/OZC	GP	$HBLR$
100%	0,450	0,395	0,422	0,172	0,080
90%	0,389	0,420	0,413	0,181	0,089
80%	0,403	0,397	0,420	0,189	0,088
70%	0,375	0,415	0,430	0,193	0,088
60%	0,392	0,435	0,411	0,193	0,086
50%	0,398	0,419	0,422	0,209	0,084
40%	0,405	0,418	0,413	0,209	0,077
30%	0,331	0,410	0,404	0,211	0,076
20%	0,280	0,424	0,400	0,227	0,083
10%	0,336	0,387	0,367	0,231	0,074

Tabelle 9.32: F_1-Wert bei unterschiedlich großen Trainingsdatensätzen (dh. verschieden großen Anteilen des Trainingsdatensatzes an der bezüglich der bekannten Items J^B im Hinbick auf die betrachteten Nutzer verfügbaren Datenmenge D1) unterschiedlicher Verfahren unter Verwendung der Variablenkombination M_3

den Verfahren der F_1-Wert herangezogen werden (vgl. Abschnitt 5.9).

Mittels des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens und des ordinalen Clusterverfahrens lassen sich unter Benutzung des SL-Verfahrens zur Approximation der Item-Cluster Zugehörigkeit im allgemeinen sogar noch etwas bessere F_1-Werte erzielen als mit der $SL/\hat{S}_Y^1$-Heuristik. Deshalb eignen sich diese Verfahren sogar noch etwas besser zur Identifikation von Höchstbewertungen. Dagegen erscheinen das hybride GP-Verfahren und der HBLR-Ansatz für solche Zwecke ungeeignet zu sein.

Sofern bezüglich aller hinsichtlich eines einzuführenden Neuprodukts relevanten Segmente eine ausreichende Anzahl von Bewertungsdaten vorhanden ist, ließe sich auf diese Weise sogar auch der prozentuale Anteil der Nutzer, die an unterschiedlichen Variationen dieses Neuprodukts interessiert sein könnten, abschätzen. Auf diese Weise ließen sich die online generierten Bewertungsdaten möglicherweise sogar zur Entscheidungsunterstützung im Rahmen der Neuprodukteinführung einsetzen.

Aufgrund der Tatsache, daß selbst die besten F_1-Werte nicht allzu groß ausfallen, ist fraglich, ob allein auf der Basis der in diesem Kapitel vorgestellten Verfahren Entscheidungen über die Aufnahme eines Produkts ins Sortiment eines

Online-Shops getroffen werden sollten. Mit dem gleichen Argument kann auch bezweifelt werden, ob die dargestellten Methoden einen substanziellen Beitrag zur Entscheidungsunterstützung im Rahmen der Neuprodukteinführung leisten können. Bestechend ist, daß diese Verfahren in der Lage wären, eine extrem große und zudem auch bereits vorhandene und darum kostenlose Datenbasis zu nutzen. Fraglich ist dennoch, ob und inwieweit durch die Breite der genutzten Datenbasis die Ungenauigkeit der Vorhersagen kompensiert werden kann.

Abschließend soll noch einmal bemerkt werden, daß die $SL/\hat{S}_Y^1$-Heuristik und das konservative HBLR-Verfahren zum gegenwärtigen Zeitpunkt die zur Generierung von Empfehlungen (auch im Rahmen des Direkt-Marketing) geeignetsten Methoden sind. Hierbei ist zu beachten, daß die $SL/\hat{S}_Y^1$-Heuristik in Bezug auf den D1-Datensatz für alle Nutzer verwendet werden sollte, die mindestens 20 Bewertungen abgegeben haben. Für alle übrigen Nutzer ist das konservative HBLR-Verfahren im Hinblick auf den zu erwartenden Nutzen der resultierenden Empfehlungen geeigneter.

Außerdem erweist sich das $SL/\hat{S}_Y^1$-Verfahren auch hinsichtlich der verzerrten Datensätze als dem HBLR-Ansatz und dem hybriden GP-Verfahren überlegen. Dies ergibt ein empirischer Vergleich dieser Verfahren, der in Anhang B.5 wiedergegeben ist. Es konnte belegt werden, daß die Breese-Werte hinsichtlich der $SL/\hat{S}_Y^1$-Heuristik bei allen Verzerrungsgraden größer ausfallen, als die auf Basis des HBLR-Verfahens oder des hybriden GP-Verfahrens bestimmten Breese-Werte (vgl. Anhang B.5).

Vor diesem Hintergrund scheint insgesamt die $SL/\hat{S}_Y^1$-Heuristik das beste Verfahren zur Vorhersage neuer Items zu sein. Fraglich ist, wie in Bezug auf Nutzer vorzugehen ist, die erst wenige Items bewertet haben. Möglich wäre, im Hinblick auf diese Personen das HBLR-Verfahren einzusetzen, bis genügend Bewertungen für diese Nutzer vorliegen. Alternativ könnten diese Nutzer durch geringfügige Anreize zur Abgabe einer moderaten Anzahl von Bewertungen bewegt werden.

Kapitel 10

Zusammenfassung und Ausblick

10.1 Zusammenfassende Beurteilung

Online-Geschäfte betreiben Recommender-Systeme, um die Kundenzufriedenheit und die Kundenbindung zu erhöhen. Da der Wechsel zu einem anderen Anbieter im Internet für die Kunden i.d.R. mit geringfügigem Zeitaufwand und vernachlässigbaren Kosten verbunden ist, sind sowohl die Kundenzufriedenheit aus auch die Kundenbindung entscheidende Faktoren im Hinblick auf den Erfolg einer via Internet Handel treibenden Unternehmung. Dementsprechend ist das wirtschaftliche Interesse, den Nutzern dieser Recommender-Systeme Empfehlungen unterbreiten zu können, die diese als hilfreich wahrnehmen, hoch. Vor diesem Hintergrund ist es nicht verwunderlich, daß sich eine Vielzahl wissenschaftlicher Arbeiten mit der Vorhersage online-generierter ranggeordneter Bewertungsdaten befaßt, um auf Basis solcher Vorhersagen den Nutzern der Recommender-Systeme wertvollere Empfehlungen zur Verfügung stellen zu können.

In den letzten Jahren sind in der Literatur zahlreiche kollaborative Verfahren vorgeschlagen worden. Problematisch ist, daß in der Literatur empirische Verfahrensvergleiche hauptsächlich auf Basis des *AAD*-Wertes angestellt werden. Der Nutzen der aus den Prognosen resultierenden Empfehlungslisten, der besser durch den Breese-Wert quantitativ erfaßt werden kann, bleibt hierdurch weitgehend unberücksichtigt. Es konnte in der vorliegenden Arbeit gezeigt werden, daß Verfahrensvergleiche auf Basis des *AAD*-Werts zu unterschiedlichen Schlußfolgerungen verleiten können, als sich mittels einer (sachgerechteren) Gegenüberstellung dieser Methoden anhand der resultierenden Breese-Werte ergeben würden.

Eine weitere Schwierigkeit ist, daß der Vergleich der unterschiedlichen Metho-

den in der Literatur unter idealisierten Annahmen vorgenommen wird, die in der Praxis nicht erfüllt sein dürften. Insbesondere ist davon auszugehen, daß sich die Menge der abgegebenen Bewertungen erheblich von den fehlenden Bewertungen unterscheidet, da die Nutzer im allgemeinen nur Items bewerten können, von denen sie sich zumindest zu einem Zeitpunkt in der Vergangenheit etwas versprochen haben. Daher ist davon auszugehen, daß der Durchschnitt der von einem Nutzer bewerteten Items i.d.R. höher ausfällt, als der Durchschnitt seiner Bewertungen im Hinblick auf die Items ausfallen würde, mit denen er sich bislang nicht befaßt hat. Dieser Umstand bleibt bei der üblichen Aufteilung der Datengrundlage in Trainings- und Testdatensatz per Zufallszahlgenerator unberücksichtigt. Hierdurch wird die grundsätzliche Verschiedenheit beurteilter und unbeurteilter Items vernachlässigt. Dieser wesentliche Unterschied wird im Rahmen der in dieser Arbeit angestellten empirischen Untersuchung erstmals berücksichtigt. Es zeigt sich, daß auch die Vernachlässigung dieser Problematik zu einer fehlerhaften Vorstellung in Bezug auf die praktische Eignung der Verfahren führen kann (vgl. Abschnitte 5.10, 8.4, 9.3 und Anhang B.5).

Die meisten vorgestellten Verfahren ignorieren fehlende Werte. Das SVD-basierte Verfahren nach Sarwar et. al. (2000b) imputiert die fehlenden Einträge in Y durch die Mittelwerte vorhandener Daten. Beide Vorgehensweisen setzen die MCAR-Eigenschaft voraus. Da ein Zusammenhang besteht zwischen der Tatsache, ob ein Nutzer ein Item bewerten kann, und seiner positiven Einstellung in Bezug auf dieses Item in der Vergangenheit, ist das Fehlen einer Bewertung nicht unabhängig davon, wie gut diese Bewertung ausfallen würde, sofern sie existent wäre. Deshalb ist nicht anzunehmen, daß die MAR-Annahme (und folglich die MCAR-Annahme) erfüllt ist (vgl. Abschnitt 3.3). Die festgestellten Ergebnisverschlechterungen mit zunehmendem Verzerrungsgrad illustrieren die möglichen Folgen der gängigen Praxis, nicht vorhandene Daten bei der Berechnung der Schätzer zu ignorieren.

Da die Verfahren dazu dienen, hinsichtlich jedes Nutzers von diesem bisher unbewertete Items zu entdecken, die seinem Geschmack entsprechen, hilft die Information, daß die meisten unbewerteten Items schlechter bewertet werden würden als die bewerteten Gegenstände, nicht weiter. Auch die Idee von Little (1986), Teilmittelwerte geeignet zu gewichten, um hierdurch den zu erwartenden Verzerrungen der Schätzer entgegenzuwirken, führt auf Basis der bekannten zweimodalen Clusterverfahren nicht zu nützlicheren Vorhersagen. Beleg hierfür sind

die nur geringfügigen Ergebnisverbesserungen, die sich durch Verwendung des zweimodalen $\hat{S}_Y^{2*}$-Clusterverfahrens gegenüber dem herkömmlichen zweimodalen $\hat{S}_Y^2$-Clusterverfahren erzielen lassen.

Das hybride Verfahren nach Pazzani (1999) führt bei hohen Testdatenanteilen bezüglich des D1-Datensatzes zu den besten Breese-Werten. Deshalb scheint es in besonderem Maße dazu geeignet zu sein, für neu hinzukommende Nutzer, die erst wenige Bewertungen abgegeben haben, Prognosen zu berechnen, auf deren Basis hilfreiche Empfehlungen generiert werden können. Außerdem führt es bei hohen und mittleren Verzerrungsgraden zu den größten Breese-Werten. Deshalb dürfte die Qualität der resultierenden Empfehlungen im Hinblick auf dieses Verfahren weniger stark durch den Unterschied zwischen den (zur Berechnung verwendeten) gegebenen Bewertungen und den zu schätzenden Bewertungen beeinträchtigt werden. Bei kleinen und mittleren Anteilen des Testdatensatzes an der Datenmenge D1 führt das Verfahren nach Pazzani (1999) zu vergleichsweise guten, wenngleich nicht den besten Breese-Werten. Deshalb erweist sich dieses Verfahren in Bezug auf das ökonomisch sinnvollste Gütemaß als beste Option. Wie alle hybriden Verfahren erfordert das hybride Verfahren nach Pazzani (1999) quantitativ erfaßte Item-Eigenschaften und kann daher nicht auf die D2-Daten angewandt werden.

Das SVD-basierte Verfahren nach Sarwar et. al. (2000b) erreicht in Bezug auf die D1-Daten bei hohen Testdatensätzen beinahe ebenso überzeugende Breese-Werte wie das hybride Verfahren nach Pazzani (1999). Bei hohen Verzerrungsgraden führt es allerdings zu sehr niedrigen Breese-Werten. Bezüglich der D2-Daten erzielt man auf Basis dieser Methode nicht die besten Ergebnisse. Wegen der starken Überlegenheit des zweimodalen $\hat{S}_Y^2$-Clusterverfahrens bezüglich des großen Datenteils D2 und hinsichtlich hoher Verzerrungsgrade scheint das zweimodale $\hat{S}_Y^2$-Clusterverfahren besser zur Generierung hilfreicher Prognosen in der Praxis geeignet zu sein.

Das hybride GP-Vefahren nach Yu et. al. (2006) ist bezüglich großer Anteile des Testdatensatzes an des gesamten Datenteilmenge D1 im Hinblick auf die Breese-Werte etwas besser als das zweimodale $\hat{S}_Y^2$-Clusterverfahren (aber schlechter als das SVD-basierte Verfahren). Trotzdem erweist es sich im Hinblick auf den großen Datensatz D2 hinsichtlich hoher und mittlerer Testdatenanteile in Bezug auf die resultierenden Breese-Werte als weniger geeignet als das zweimodale $\hat{S}_Y^2$-Clusterverfahren. Dafür schneidet das hierarchische GP-Verfahren nach Yu et. al. (2006) bei hohen Verzerrungsgraden fast so gut ab wie die deutlich langsamere

hybride Methode nach Pazzani (1999).

Insgesamt ist daher aufgrund der empirischen Ergebnisse nicht eindeutig festzustellen, ob in den Fällen, in denen keine geeignete Information hinsichtlich der Items verfügbar ist, eher das hierarchische GP-Verfahren oder das zweimodale $\hat{S}_Y^2$-Clusterverfahren zu empfehlen ist. Beide Verfahren haben sich jedoch unter der genannten Voraussetzung aufgrund der empirischen Ergebnisse als den übrigen Ansätzen überlegen erwiesen.

Alle kollaborativen Verfahren und auch das hybride Verfahren nach Pazzani (1999) sind nicht zu Vorhersagen in Bezug auf neue Items geeignet und führen im Hinblick auf Items, die deutlich seltener als die übrigen im Trainingsdatensatz enthaltenen Items bewertet wurden, zu unzuverlässigen Prognosen. Prognosen bezüglich neuer Items können ohnehin nicht mittels dieser Methoden berechnet werden. Da unzutreffende Empfehlungen zu vermeiden sind, ist es ratsam, auch zur Vorhersage weniger bekannter Items andere Verfahren einzusetzen.

Im Hinblick auf den Nutzen der Empfehlungen ist zu bedenken, daß auch der Breese-Wert den Neuigkeitsgrad der empfohlenen Items nicht berücksichtigt. Gerade vor dem Hintergrund, daß der Breese-Wert ein heuristisches Maß für den Nutzen der resultierenden Empfehlungslisten bilden soll, ist problematisch, daß hierbei vernachlässigt wird, wie hoch der Neuigkeitsgrad der mit Hilfe des jeweils zu evaluierenden Prognoseverfahrens generierten Empfehlungen ist. Je neuer und unbekannter die empfohlenen Items sind, umso unwahrscheinlicher ist es, daß die Nutzer bereits aus anderen Quellen Hinweise auf diese Items erhalten haben. Daher ist mit zutreffenden Empfehlungen hinsichtlich neuer Items ein höherer Beitrag zur Freizeitgestaltung und zur Erhöhung der Lebensqualität der Nutzer verbunden. Demnach können Empfehlungen bezüglich neuer Items einen höheren Beitrag zur Kundenzufriedenheit und Kundenbindung leisten. Hieraus folgt, daß unter marketingrelevanten Gesichtspunkten der Neuigkeitsgrad der resultierenden Empfehlungen entweder a priori gegeben sein muß oder eine im Rahmen eines empirischen Verfahrensvergleichs quantitativ zu erfassende Erfolgsdimension ist.

Es ist auch unter dieser Voraussetzung noch berechtigt, den Breese-Wert als ein erheblich besseres Maß für den Nutzen der resultierenden Empfehlungen als den *AAD*-Wert zu bezeichnen. Zum gegenwärtigen Zeitpunkt existiert weder ein anderes allgemein anerkanntes Maß für den geschätzten Nutzen der auf der Basis der Prognosen generierbaren Empfehlungslisten, noch wäre die Verwendung einer Modifikation des Breese-Wertes, die zusätzlich die Berücksichtigung des

Neuigkeitsgrades gestattet, unproblematisch. Falls die Eignung zur Generierung zutreffender Empfehlungen, die möglichst weit oben in der Empfehlungsliste angeordnet sind, nicht von dem Neuigkeitsgrad derselben getrennt wird, ist nicht mehr klar erkennbar, ob der Vorteil eines Verfahrens in resultierenden zutreffenden Empfehlungen oder dem Neuigkeitsgrad dieser Empfehlungen besteht. Daher ist eher die Verwendung des bekannten Breeese-Werts und eine davon unabhängige Betrachtung des Neuigkeitsgrades der resultierenden Empfehlungen zu empfehlen. (Eine weitere Schwierigkeit bei der Definition eines idealen Maßes für den Nutzen wäre, wie man den Schaden durch besonders unglückliche Empfehlungen quantifiziert, die bei Prognosen auf Basis kollaborativer Verfahren hinsichtlich weniger bekannter Items in verstärktem Maß zu erwarten sind.)

Das hybride Verfahren nach Pazzani (1999) verwendet ebenso wie das Nutzerbasierte Ähnlichkeitsverfahren vor allem die (im Trainingsdatensatz vorhandenen) Bewertungen anderer Nutzer, die mit dem Nutzer, dessen Bewertungen vorherzusagen sind, stark (positiv oder negativ) korreliert sind, zur Bestimmung der Vorhersagen. Bezüglich eines selten bewerteten Items sind definitionsgemäß wenige Nutzer vorhanden, die dieses Item beurteilt haben. Somit sind unter diesen Voraussetzungen im Hinblick auf das betreffende Item nicht unbedingt Bewertungen von Nutzern verfügbar, die mit dem jeweiligen Nutzer, dessen Bewertung hinsichtlich des betrachteten Items zu prognostizieren ist, stark (positiv oder negativ) korreliert sind. Daher ist davon auszugehen, daß die Bewertungen hinsichtlich dieser Items im allgemeinen nicht besonders stark vom Mittelwert der Bewertungen des betrachteten Nutzers abweichen dürften. Somit ist es vergleichsweise unwahrscheinlich, daß auf Basis des hybriden Verfahrens ein weniger bekanntes Item empfohlen wird. Wegen der Vernachlässigung des Neuigkeitsgrads durch den Breese-Wert sind daher die auf Basis der Methode nach Pazzani (1999) resultierenden Breese-Werte eher ein Beleg dafür, daß im Rahmen dieses Verfahrens unzutreffende Bewertungen vermieden werden, indem überwiegend Empfehlungen für bekannte Items abgegeben werden. Vor diesem Hintergrund erscheinen die auf Basis dieses Verfahrens erreichten hohen Breese-Werte weniger beeindruckend.

Problematisch ist, daß bis heute erstaunlich wenig Literatur existiert, die sich mit der Vorhersage von Bewertungen in Bezug auf Items befaßt, hinsichtlich derer noch keine Bewertungen verfügbar sind.

Dem Nutzer werden ständig seitens seiner Familie, seiner Freunde, Kollegen und Nachbarn bekannte Items empfohlen. Daher besteht kein großer Bedarf im

Hinblick auf Empfehlungen, die sich auf Items beziehen, die bereits im allgemeinen Bewußtsein präsent sind. Erheblich hilfreicher wären dagegen Empfehlungen für neue Items.

Daher war es das Hauptanliegen dieser Arbeit, Verfahren, die zur Vorhersage und Empfehlung neuer Items eingesetzt werden können, zu entwickeln und unter möglichst praxisnahen Bedingungen empirisch miteinander und mit bereits vorhandenen Alternativen zu vergleichen.

Die in dieser Arbeit vorgestellte Methode zur indirekten Schätzung (die als SL-Heuristik bezeichnet wird) ermöglicht Prognosen bezüglich neuer Items auf Basis zweimodaler Clusterverfahren. (Hierzu ist allerdings erforderlich, daß die Eigenschaften der betrachteten Items bekannt und quantitativ bestimmbar sind.)

Die auf der Grundlage von Bewertungen hinsichtlich bekannter Items mittels zweimodaler Clusterverfahren erzeugten Item-Cluster werden in Verbindung mit den (quantitativ erfaßten) Eigenschaften dieser bekannten Items dazu benutzt, auf Basis der Eigenschaften der neuen Items deren Clusterzugehörigkeit zu schätzen. Dieses Klassifikationsproblem wurde mittels einer Reihe von Verfahren aus der Literatur bestmöglich zu lösen versucht.

Der empirische Vergleich der resultierenden Kombinationen dieser Klassifikationsverfahren mit drei verschiedenen zweimodalen Clusterverfahren (Abschnitt 9.3) hat gezeigt, daß eine einfache (durch das single-linkage Verfahren motivierte) Heuristik in Verbindung mit dem zweimodalen $\hat{S}_Y^1$-Clusterverfahren in den meisten Fällen zu den besten Ergebnissen führt. Das resultierende Verfahren wird als $SL/\hat{S}_Y^1$-Heuristik bezeichnet.

Neben diesem Ansatz wurde auch ein linearer Hierarchischer Bayes'scher Regressionsansatz zur Vorhersage der Bewertungen bezüglich neuer Items verwendet, der dem Verfahren von Ansari et. al. (2000) ähnelt.

Zusätzlich ist es gelungen, ein kollaboratives Hierarchisches Bayes'sches Modell (das sogenannte Hierarchische GP-Verfahren nach Yu et. al. (2006)) so zu modifizieren, daß es zu Regressionszwecken einsetzbar ist und Prognosen hinsichtlich der Bewertungen, die sich auf neue Items beziehen, auf Basis der Eigenschaften dieser Items ermöglicht.

Außerdem wurden vielfältige Ansätze vorgestellt und empirisch evaluiert, die es ermöglichen, die Ergebnisse eines zweimodalen Clusterverfahrens (via die Prior) zum Bestandteil des Hierarchischen Bayes'schen linearen Regressionsmodells nach Rossi et. al. (1996) zu machen. Obwohl auf der Basis clusterspezifischer Mo-

delle eine deutliche Verbesserung im Vergleich zum Hierarchischen Bayes'schen linearen Regressionsverfahren erzielt werden konnte, werden die mit dieser Methode erreichten Ergebnisse von dem besten Resultat der deutlich schnelleren $SL/\hat{S}_Y^1$-Heuristik übertroffen.

Die Verwendung einer geeigneten Menge von Eigenschaften ist im Rahmen der $SL/\hat{S}_Y^1$-Heuristik insbesondere, wenn bezüglich der bekannten Items vergleichsweise wenig Beobachtungen vorliegen, ausschlaggebend für die Nützlichkeit der aus den zugehörigen Vorhersagen resultierenden Empfehlungen. Sofern es gelingt, eine geeignete Menge von Variablen zu bestimmen, führt die $SL/\hat{S}_Y^1$-Heuristik erst, wenn nur noch wenige Bewertungen hinsichtlich bekannter Items zur Berechnung der Schätzer eingesetzt werden, zu schlechteren Ergebnissen in Bezug auf die resultierenden Breese-Werte als das Hierarchische Bayes'sche lineare Regressionsmodell. Das vorgestellte Hybride Hierarschische GP-Verfahren führt zwar zu den kleinsten *AAD*-Werten, aber auch zu den schlechtesten Breese-Werten. (Da alle Verfahren in Abschnitt 9.3 nur auf ihre Eignung zur Vorherage bzw. Empfehlung neuer Items hin untersucht wurden, ist es irrelevant, daß der Breese-Wert den Neuigkeitsgrad der empfohlenen Items nicht berücksichtigt.) Insgesamt kann die $SL/\hat{S}_Y^1$-Heuristik empfohlen werden, sofern nicht zu wenig Bewertungen zur Berechnung der Schätzer verfügbar sind und vorab eine sachgerechte Menge von Variablen identifiziert werden kann. (Zum letztgenannten Zweck konnte eine geeignete Methode vorgestellt werden.) Falls zu wenig Beobachtungen gegeben sind, bietet es sich an, entweder den (in diesem Fall überlegenen) Hierarchischen Bayes'schen Regressionsansatz nach Rossi et. al. (1996) zu verwenden oder die Nutzer, bezüglich derer zuwenige Bewertungen vorhanden sind, durch das Setzen geeigneter Anreize zur Abgabe weiterer Bewertungen zu bewegen.

Insgesamt führt die Berücksichtigung des ordinalen Skalenniveaus erstaunlich selten zu Ergebnisverbesserungen gegenüber guten linearen Approximationen.

Auch die subtilere Erfassung individueller Besonderheiten durch die Hierarchischen Bayes'schen Verfahren führt nicht immer zu besseren Resultaten.

Wenn nicht möglichst zutreffende Empfehlungen im Hinblick auf neue Items im Vordergrund des Interesses stehen, sondern die zutreffende Identifikation neuer Items beabsichtigt wird, von denen anzunehmen ist, daß bestimmte Nutzer an ihnen interessiert sind, kommt es gleichermaßen auf Precision und Recall an. Es konnte empirisch erwiesen werden, daß die zweimodalen Clusterverfahren zu diesem Zweck besser geeignet sind als die beiden hybriden Bayes'schen Verfahren.

Beispielhaft für eine Anwendung, bei der es auf die zuverlässige Bestimmung von Items ankommt, die von bestimmten Nutzern bevorzugt werden, ist die (approximative) Ermittlung der (Anzahl der) Nutzer, die an einem gegebenenfalls neu in das Sortiment eines Online-Shops einzuführenden Item interessiert sind.

10.2 Ausblick

Gerade vor dem Hintergrund des mit den in dieser Arbeit betrachteten Problemen und den hergeleiteten Lösungsmöglichkeiten verbundenen wirtschaftlichen Interesses laden die erzielten Ergebnisse zu weiterführender Forschung ein.

Es ist im Rahmen dieser Arbeit gelungen, Verfahren zu entwickeln, mittels derer (bessere) Vorhersagen in Bezug auf neue Items abgeben werden können. Auch Items, die zu einem bestimmten Zeitpunkt neu waren, werden i.d.R. zu späteren Zeitpunkten bewertet. Daher wäre die Entwicklung von Methoden wichtig, die es ermöglichen, die ersten Bewertungen, die in Bezug auf diese Items erfolgen, zur Verbesserung des Prognoseverfahrens im Hinblick auf den zu erwartenden Nutzen der resultierenden Empfehlungslisten zu verwenden.

Eine bisher nicht genutzte Chance ist die bessere Berücksichtigung der Zeitabhängigkeit der Bewertungsdaten. Bewertungen reflektieren den Geschmack einer bestimmten Person zu einer bestimmten Zeit. Da sich der Geschmack eines Menschen mit der Zeit ändern kann, kann es sein, daß insbesondere ältere Bewertungen eines Nutzers nicht mehr dessen aktuellen Geschmack wiederspiegeln. Als lohnend könnte sich auch die Beschäftigung mit Mustern erweisen, die sich auf die zeitliche Entwicklung des Geschmacks der Nutzer beziehen.

Ließen sich mehrmodale Clusterverfahren entwickeln, bei denen verzerrte Daten nicht zu verzerrten Partitionen führen würden, könnten durch die geeignete Gewichtung der Partitionsmittelwerte in Anlehnung an Little (1986) im Rahmen des zweimodalen $\hat{S}_Y^{2*}$-Schätzers mehr als nur geringfügigen Verbesserungen im Verleich zum zweimodalen $\hat{S}_Y^2$-Schätzer erreicht werden. Daher besteht weiterer Forschungsbedarf in Bezug auf die Entwicklung von zweimodalen Cluster-Algorithmen.

Eine zusätzliche Möglichkeit zur Verbesserung der Qualität von Empfehlungen ist die bessere Berücksichtigung der Situation des Nutzers. Periodisch wiederkehrende Interessen wie die Suche nach einem unterhaltsamen neuen Kinofilm am Samstagabend, der gleichermaßen Frauen wie Männern zusagt, oder die Suche

nach familienfreundlichen Filmen in der Weihnachtszeit, sollten besser erkannt und berücksichtigt werden.

In einer Reihe von Arbeiten (Gaul, Schmidt-Thieme (2000), (2001), (2002)) wurden Methoden entwickelt, die (u.a.) dazu geeignet sind, das Navigationsverhalten der Nutzer einer Website und dessen Implikationen eingehend zu analysieren. Es stellt sich die Frage, ob sich dem individuellen Navigationsverhalten der Nutzer auch Informationen entnehmen lassen, die Rückschlüsse auf deren Beurteilung von Produkten und Dienstleistungen erlauben. So ist beispielsweise anzunehmen, daß Technik-affine Nutzer ein effizienteres Navigationsverhalten an den Tag legen, da sie i.d.R. mit den neuen interaktiven Medien vertrauter sind und oft über beträchtliches Internet-spezifisches Hintergrundwissen verfügen. Es wäre zu überprüfen, ob und inwiefern ein effizienteres (weniger effizientes) Navigationsverhalten Schlußfolgerungen hinsichtlich der Beurteilung (bestimmter Gruppen) von Items zuläßt. Zudem lassen sich unter Umständen durch die Analyse der Logfiles für jeden Nutzer Item-Genres identifizieren, nach denen der betrachtete Nutzer häufiger sucht. Hieraus können möglicherweise Anhaltspunkte für individuelle Genre-Vorlieben der Nutzer abgeleitet werden. Daher wäre durch zukünftige Forschung zu klären, ob eine praktikabele Möglichkeit besteht, durch zusätzliche Verwendung von Navigationsdaten (Logfiles) die Qualität der resultierenden Empfehlungen zu erhöhen.

Anders als die im MovieLens-Datensatz enthaltenen persönlichen Angaben beruhen die Logfiles der Nutzer nicht auf Selbstauskünften. Sie sind daher zuverlässiger. Im Unterschied zu den abgefragten Daten ist ihre Erhebung für den Nutzer nicht mit Mühe verbunden. Bei der Erhebung von Navigationsdaten besteht auch nicht die Gefahr, Nutzer durch persönliche Fragen zu verärgern oder ihnen gar das Gefühl zu geben, ausgeforscht zu werden. Daher ist die Erhebung von Navigationsdaten im Unterschied zur expliziten Befragung der Nutzer aus ökonomischer Sicht unproblematisch.

Empfehlungen dienen einem bestimmten Zweck. Daher ist die Eignung der Verfahren zur Erfüllung dieses Zwecks zu diskutieren. Ist dieser Zweck die Versorgung der Nutzer mit hilfreichen Empfehlungen, so ist davon auszugehen, daß insbesondere Empfehlungen für neue und weniger bekannte Items sinnvoller sind als Empfehlungen für allgemein bekannte Items. Dennoch lassen sich in Bezug auf allgemein bekannte Items leichter niedrigere *AAD*-Werte und auch hohe Precision-, Recall- und Breese-Werte erzielen, obwohl nicht unbedingt davon ausgegangen

werden kann, daß die auf Grundlage dieser Schätzer bestimmten Empfehlungen dem Nutzer oder dem Online-Geschäft weiterhelfen, da der betreffende Nutzer in vielen Fällen das Buch, den Film oder das Musikstück schon (zu) gut kennt oder sich bereits aus anderen Gründen entschieden hat, sich mit dem betreffenden Gegenstand nicht näher vertraut zu machen. Daher dürfen die Verfahren nicht ausschließlich anhand der Werte geeigneter Gütemaße beurteilt werden, sondern es sollte vielmehr entscheidend sein, wie gut diese Werte in Bezug auf die Schätzung der Bewertung neuer Items ausfallen.

Es wurde in dieser Arbeit empirisch illustriert, daß viele betrachtete Gütemaße die Interessen der Betreiber eines Empfehlungssystems (Online-Geschäfts) zu ignorieren scheinen. Kein Betreiber ist an genauen Prognosen per se (AAD-Wert) interessiert.

Wegen der hohen Bedeutung, die der Kundenzufriedenheit und Kundenbindung im Hinblick auf die Erfolgsaussichten eines Online-Geschäfts zukommen, wurde im Rahmen dieser Arbeit im Einklang mit der Literatur vorausgesetzt, daß der Zweck der Empfehlungen ist, daß diese von den Nutzern als wertvolle Beiträge zu ihrer Freizeitgestaltung und geeignete Mittel zur Erhöhung ihrer Lebensqualität wahrgenommen werden. Es wäre aber durchaus nicht uninteressant, zu erforschen, ob die Empfehlungen darüberhinaus noch weiteren ökonomischen Zwecken dienen können, ohne daß sich hierdurch das Niveau der Kundenzufriedenheit und Kundenbindung signifikant verringert. In der Regel darf davon ausgegangen werden, daß zumindest die Betreiber eines Online-Geschäfts das Empfehlungssystem letzlich dazu nutzen möchten, die Gewinne ihres Unternehmens zu maximieren. Ob sich hierzu die aus der Sicht der Nutzer hilfreichsten Empfehlungen in allen Fällen eignen, darf bezweifelt werden. Neben dem mit den generierten Empfehlungslisten verbundenen Nutzen der Kunden (Breese-Wert) sollte auch der aufgrund der generierten Empfehlungslisten zu erwartende Beitrag zum Gewinn des Online-Geschäfts geschätzt werden und Beachtung finden.

Die vorgestellten Verfahren zur Prognose von Bewertungen im Hinblick auf neue Items ermöglichen auch Vorhersagen im Hinblick auf (noch) nicht existierende, rein hypothetische Items, wie z.B. sie im Rahmen von Heuristiken zur Neuprodukteinführung betrachtet werden. Auf Basis der im Rahmen dieser Arbeit betrachteten Methoden ließen sich bezüglich jedes fiktiven Items die Bewertungen aller Nutzer schätzen. Die SL-Heuristik in Verbindung mit zweimodalen Clusterverfahren hat sich hinsichtlich der Identifikation neuer Items, an

denen bestimmte Nutzer Interesse haben könnten, als geeigneter als die beiden Bayes'schen Verfahren erwiesen. Es wäre empirisch zu überprüfen, ob die mittels der $SL/\hat{S}_Y^1$-, $SL/\hat{S}_Y^2$- oder SL/OZC-Heuristik bestimmbaren Vorhersagen der Bewertungen bestimmter Nutzer bezüglich neu einzuführender Produkte zur Entscheidungshilfe hinsichtlich der Aufnahme eines neuen Produkts ins Sortiment eines Online-Geschäfts dienen sollten. Ebenso scheint es eine interessante Fragestellung zu sein, ob die Prognosen bezüglich neuer Items im Zusammenhang mit der Neuprodukteinführung oder Produktlinienoptimierung gewinnbringend eingesetzt werden könnten.

Literaturverzeichnis

Adomavicius, G., Tuzhilin, A. (2005): Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, *IEEE Transactions on Knowledge and Data Engineering,* 17, 734-749

Agresti, A. (1984): *Analysis of Ordinal Categorical Data*, Wiley

Agresti, A. (2002): *Categorical Data Analysis*, Wiley Series in Probability and Statistics

Albrecht, P. (1980): On the Correct Use of the Chi-Square Goodness-of-Fit-Test, *Scandinavian Actuarial Journal*, 149-160

Allison, P.D. (1999): *Logistic Regression Using the SAS System: Theory and Application*, Wiley

Anderson, J.A. (1984): Regression and Ordered Categorical Variables, *Journal of the Royal Statistical Society B,* 46, 1-30

Anderson, A.B., Basilevsky, A., Hum, D.J. (1983): Missing Data: A Review of the Literature, In: Rossi, P.H., Wright, J.D., Anderson, A.B. (Eds.): *Handbook of Survey Research,* Academic Press, 415-493

Ansari, A., Essegaier, S., Kohli, R. (2000): Internet Recommendation Systems, *Journal of Marketing Research,* 37, 363-375

Arbuckle, J.L. (1996): Full Information Estimation in the Presence of Incomplete Data, In: Marcoulides, G.A., Schumacker, R.E. (Eds.): *Advanced Structural Equation Modeling: Issues and Techniques,* Lawrence Erlbaum, 243-277

Azen, S., Van Guilder, M. (1981): Conclusions Regarding Algorithms for Handling Incomplete Data, *Proceedings of the Statistical Computing Section, American Statistical Association 1981,* 53-56

Baier, D., Gaul, W., Schader, M. (1997): Two-Mode Overlapping Clustering with Applications in Simultaneous Benefit Segmentation and Market Structuring, In: Klar, R., Opitz, O. (Eds.): *Classification and Knowledge Organization,* Springer, 557-566

Balabanovic, M., Shoham, Y. (1997): Content-Based, Collaborative Recommendation, *Communications of the ACM,* 40, 66-72

Baltagi, B.H. (1998): *Econometrics*, Springer

Banerjee, A., Dhillon, I.S., Ghosh, J., Merugu, S., Modha, D.S. (2004): A Generalized Maximum Entropy Approach to Bregman Co-Clustering and Matrix Approximation, *Proceedings of the 10th ACM SIGKKD International Conference on Knowledge Discovery and Data Mining,* 509-514

Bankhofer, U. (1995): *Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse*, Josef Eul

Basu, C., Hirsh, H., Cohen, W. (1998): Recommendation as Classification: Using Social and Content-Based Information in Recommendation, *Proceedings of the 15th National Conference on Artificial Intelligence,* 714-720

Ben-Akiva, M., Lerman, S.R. (1985): *Discrete Choice Analysis: Theory and Application in Travel Demand*, MIT Press

Berger, J.O., Bernardo, J.M. (1992): On the Development of Reverence Priors, In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.): *Bayesian Statistics 4*, 35-60

Bernardo, J.M., Giron, F.J. (1988): A Bayesian Analysis of Simple Mixture Problems, *Bayesian Statistics,* 3, 67-78

Berry, M.W. (1992): Large Scale Singular Value Computations, *International Journal of Supercomputer Applications,* 6, 13-69

Billsus, D., Pazzani, M.J. (1998): Learning Collaborative Information Filters, *Proceedings of the 15th International Conference on Machine Learning,* 46-54

Blum, A., Hellerstein, L., Littlestone, N. (1995): Learning in the Presence of Finitely or Infinitely Many Irrelevant Attributes, *Journal of Computer and System Sciences,* 50, 32-40

Both, M., Gaul, W. (1987): Ein Vergleich zweimodaler Clusteranalyseverfahren, *Methods of Operations Research,* 57, 593-605

Boyen, X., Koller, D. (1998): Tractable Inference from Complex Stochastic Systems, *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence,* 33-42

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

Breese, J.S., Heckerman, D., Kadie, C. (1998): Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence,* 43-52

Breiman, L. (1959): The Strong Law of Large Numbers for a Class of Markov Chains, *The Annals of Mathematical Statistics*, 31, 801-803

Breiman, L., Friedman, H., Olshen, J.A., Stone, C.J. (1984): *Classification and Regression Trees*, Chapman & Hall

Brick, J.M., Kalton, G. (1996): Handling Missing Data in Survey Response, *Statistical Methods in Medical Research,* 5, 215-238

Brooks, S.P., Roberts, G.O. (1998): General Methods for Monitoring Convergence of Iterative Simulations, *Journal of Computational and Graphical Statistics,* 7, 434-455

Brown, C.H. (1983): Asymptotic Comparison of Missing Data Procedures for Estimating Factor Loadings, *Psychometrika,* 48, 269-291

Brown, M.B., Benedetti, J.K. (1977): Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables, *Journal of the American Statistical Association,* 72, 309-315

Buck, S.F. (1960): A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer, *Journal of the Royal Statistical Society B,* 22, 302-306

Buckley, C., Salton, G., Allan, J. (1993): Automatic Retrieval with Locality Information Using SMART, In: Harman, D.K. (Ed.): *The 1st Text Retrieval Conference,* National Institute of Standards and Technology, Special Publication 500-207, 59-72

Buckley, C., Salton, G., Allan, J. (1994): The Effect of Adding Relevance Information in a Relevance Feedback Environment, *Proceedings of the 17th International ACM/SIGIR Conference on Research and Development in Information Retrieval,* 292-300

Calderon-Benavides, M.L., Gonzales-Caro, C.N., Perez-Alcazar, J., Garcia-Diaz, J.C., Delgado, J. (2004): A Comparison of Several Predictive Algorithms for Collaborative Filtering on Multi-Valued Ratings, *Proceedings of the 2004 ACM Symposium on Applied Computing,* 1033-1039

Cameron, T.A. (1987): The Impact of Grouping Coarseness in Alternative Grouped-Data Regression Models, *Journal of Econometrics,* 35, 37-57

Cameron, T.A. (1992): Errata: The Impact of Grouping Coarseness in Alternative Grouped-Data Regression Models, *Journal of Econometrics,* 52, 419-421

Caudill, S.B. (1992): More on Grouping Coarseness in Linear Normal Models, *Journal of Econometrics,* 52, 407-417

Chan, L.S., Dunn, O.J. (1972): The Treatment of Missing Values in Discriminant Analysis, *Journal of the American Statistical Association,* 67, 473-477

Chen, P.Y., Popovich, P.M. (2002): *Correlation: Parametric and Nonparametric Measures*, Sage

Cheng, J. (2006): BN PowerConstructor Software, *http://www.cs.ualberta.ca/jcheng/bnsoft.htm*

Cheng, Y., Church, G.M. (2000): Biclustering of Gene Expression Data, *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology,* 93-103

Chien, Y.-H., George, E.I. (1999): A Bayesian Model for Collaborative Filtering, *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*, http://www-stat.wharton.upenn.edu/∼edgeorge/Research_papers/Bcollab.pdf

Chu, W., Ghahramani, Z. (2005): Gaussian Processes for Ordinal Regression, *Journal of Machine Learning Research,* 6, 1019-1041

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M. (1999): Combining Content-Based and Collaborative Filters in an Online Newspaper, *Proceedings of the ACM SIGIR'99 Workshop on Recommender Systems*, http://www.cs.wpi.edu/∼claypool/papers/content-collab/

Clogg, C.C., Shihadeh, E.S. (1994): *Statistical Models for Ordinal Variables*, Sage

Cochran, W.G. (1954): Some Methods of Strengthening the Common χ^2-Tests, *Biometrics,* 10, 417-451

Cohen, J., Cohen, P. (1975): *Applied Multiple Regression, Correlation Analysis for the Behavioral Sciences*, Erlbaum

Colledge, M.J., Johnson, J.H., Paré, R., Sande, I.G. (1978): Large Scale Imputation of Survey Data, *Proceedings of the American Statistical Association 1978,* 431-436

Constanzo, C.M., Halperin, W.C., Gale, N., Richardson, G.D. (1982): An Alternative Method for Assessing Goodness-of-Fit for Logit Models, *Environment and Planning A,* 14, 963-971

Cooper, G.F., Herskovits, E. (1982): A Bayesian Method for the Induction of Probabilistic Networks from Data, *Machine Learning,* 9, 309-347

Cowles, M.K., Carlin, B.P. (1996): Markov Chain Monte Carlo Convergence Diagnostic: A Comparative Review, *Journal of the American Statistical Association,* 91, 883-904

Cox, D.R. (1972): Regression Models and Life Tables, *Journal of the Royal Statistical Society B*, 34, 187-220

Cox, D.R., Snell, E.J. (1989): *The Analysis of Binary Data*, Chapman & Hall

Cox, D.R., Wermuth, N. (1992): A Comment on the Coefficient of Determination for Binary Responses, *The American Statistician,* 46, 1-4

Cragg, J.G., Uhler, R. (1970): The Demand for Automobiles, *Canadian Journal of Economics,* 3, 386-406

David, M., Little, R.J.A., Samuhel, M.E., Triest, R.K. (1986): Alternative Methods for CPS Income Imputation, *Journal of the American Statistical Association,* 81, 29-41

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (1990): Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science,* 41, 391-407

DeMaris, A. (1992): *Logit Modeling: Practical Applications*, Sage

Dempster, A., Laird, N., Rubin, D. (1977): Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society B,* 39, 1-38

DeSarbo, W.S. (1982): GENNCLUS: New Models for General Nonhierarchical Clustering Analysis, *Psychometrika,* 47, 449-475

Deshpande, M., Karypis, G. (2004): Item-Based Top-N Recommendation Algorithms, *ACM Transactions on Information Systems,* 22, 143-177

Dillon, W.R., Madden, T.J., Kirmani, A., Mukheerjee, S. (2001): Understanding What's in a Brand Rating: A Model for Assessing Brand and Attribute Effects and Their Relationship to Brand Equity, *Journal of Marketing Research,* 38, 415-429

Eckart, C., Young, G. (1936): The Approximation of one Matrix by Another of Lower Rank, *Psychometrika,* 1, 211-218

Edgett, G.L. (1956): Multiple Regression With Missing Observations Among the Independent Variables, *Journal of the American Statistical Association,* 51, 122-132

EDWARDS, D., THURSTONE, L. (1952): An Internal Consistency Check for Scale Values Determined by the Method of Successive Intervals, *Psychometrika*, 17, 169-180

EFRON, B. (1978): Regression and ANOVA With Zero-One Data: Measures of Residual Variation, *Journal of the American Statistical Association,* 73, 113-121

ELIASHBERG, J., SHUGAN, S.M. (1997): Film Critics: Influencers or Predictors?, *Journal of Marketing,* 61, 68-78

ESPEJO, E., GAUL, W. (1986): Two-Mode Hierarchical Clustering as an Instrument for Marketing Research, In: Gaul, W. Schader, M. (Eds.): *Classification as a Tool of Research*, Springer, 121-128

FAHRMEIR, L., TUTZ, G. (2001): *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer

FEDERSPIEL, C.F., MONROE, R.J., GREENBERG, B.G. (1959): *An Investigation of Some Multiple Regression Methods for Incomplete Samples*, University of North Carolina, Institute of Statistics, Memoe Series, No. 236

FIENBERG, S.E. (1980): *The Analysis of Cross-Classified Categorical Data*, MIT Press

FISHER, D., HILDRUM, K., HONG, J., NEWMAN, M., THOMAS, M., VUDUC, R. (2000): A Framework for Collaborative Filtering Algorithm Development and Evaluation, *Proceedings of the 23nd Annual International ACM SIGIR,* 366-368

FLETCHER, R., REEVES, C.M. (1964): Function Minimization by Conjugate Gradients, *Computing Journal*, 7, 149-154

FRANE, J.W. (1978): Missing Data and BMDP: Some Pragmatic Approaches, *ASA Proceedings of the Statistical Computing Section*, 27-33

FRIEDMAN, N., GEIGER, D., GOLDSZMIDT, M. (1997): Bayesian Network Classifiers, *Machine Learning,* 29, 131-163

FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. (2000): A Statistical View of Boosting, *The Annals of Statistics,* 28, 337-374

Gaul, W. (2004): *Moderne Marktforschung*, Vorlesungsskript WS 2004/2005, Universität Karlsruhe

Gaul, W. (2004): *Ausgewählte Statistische Verfahren im Marketing*, Vorlesungsskript SS 2005, Universität Karlsruhe

Gaul, W., Schader, M. (1994): Pyramidal Classification Based on Incomplete Dissimilarity Data, *Journal of Classification*, 11, 171-193

Gaul, W., Schader, M. (1996): A New Algorithm for Two-Mode Clustering, In: Bock, H.H., Polasek, W. (Eds.): *Data Analysis and Information Systems,* Springer, 15-23

Gaul, W., Schmidt-Thieme, L. (2000): Mining Web Navigation Path Fragments, *Proceedings of the Workshop Web Mining for E-Commerce, The 6th ACM SIGKDD International Conference on Data Mining*, 105-110

Gaul, W., Schmidt-Thieme, L. (2001): Mining Generalized Association Rules for Sequential and Path Data, *Proceedings of the 2001 International Conference on Data Mining*, 593-596

Gaul, W., Schmidt-Thieme, L. (2002): Recommender Systems Based on User Navigational Behavior in the Internet, *Behaviormetrika*, 29, 1-22

Gaul, W., Schader, M., Bomhardt, C. (2006): *Two-Mode Clustering With Missing Values*, Working Paper

Geiger, D., Heckerman, D. (1996): Knowledge Representation and Inference in Similarity Networks and Bayesian Multinets, *Artificial Intelligence,* 82, 45-74

Gelfand, A.E., Smith, A.F.M. (1990): Sampling-Based Approaches for Calculating Marginal Densities, *Journal of the American Statistical Association,* 85, 398-409

Gelman, A., Carlin, J., Stern, H., Rubin, D. (1995): *Bayesian Data Analysis*, Chapman & Hall

Gelman, A. (1996): Inference and Monitoring Convergence, In: Gilks, W.R., Richardson, S., Spiegelhalter, J. (Eds.): *Markov Chain Monte Carlo in Practice*, Chapman & Hall, 131-143

Gelman, A., Rubin, D. (1992a): Inference From Iterative Simulation Using Multiple Sequences, *Statistical Science,* 7, 457-472

Gelman, A., Rubin, D. (1992b): A Single Series From the Gibbs Sampler Provides a False Sense of Security , In: Bernado, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.): *Bayesian Statistics 4,* Claredon Press, 625-631

Geman, S., Geman, D. (1984): Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721-741

George, T., Merugu, S. (2005): A Scalable Collaborative Filtering Framework Based on Co-Clustering, *Proceedings of the 5th International IEEE Conference on Data Mining (ICDM)*, 625-628

Geweke, J. (1992): Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments, In: Bernado, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.): *Bayesian Statistics,* Volume 4, Oxford University Press, 169-194

Gibbons, J.D. (1993): *Nonparametric Measures of Association*, Sage

Gilks, W.R., Wang, C.C., Coursaget, P., Yvonnet, B. (1993): Random-Effects Models for Longitudinal Data Using Gibbs Sampling, *Biometrics,* 49, 441-453

Gleason, T.C., Staelin, R.A. (1975): A Proposal for Handling Missing Data, *Psychometrika,* 40, 229-252

Goldberg, K., Roeder, T., Gupta, D., Perkins, C. (2001): Eigentaste: A Constant Time Collaborative Filtering Algorithm, *Information Retrieval,* 4, 133-151

Good, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B., Herlocker, J., Riedl, J. (1999): Combining Collaborative Filtering with Personal Agents for Better Recommendations, *Proceedings of the Conference of the American Association for Artificial Intelligence,* 463-449

Goodman, L.A. (1970): The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications, *Journal of the American Statistical Association,* 65, 226-256

Goodman, L.A., Kruskal, W.H. (1954): Measures of Association for Cross Classification, *Journal of the American Statistical Association,* 49, 732-764

Greene, W.G. (2003): *Econometric Analysis,* 5th Edition, Prentice Hall

Greenless, W.S., Reece, J.S., Zieschang, K.D. (1982): Imputation of Missing Values When the Probability of Response Depends on the Variables Being Imputed, *Journal of the American Statistical Association,* 77, 251-261

Haberman, S.J. (1978): *Analysis of Qualitative Data,* Volume I, Academic Press

Hagle, T.M., Mitchell, G.E. (1992): Goodness-of-Fit Measures for Probit and Logit, *American Journal of Political Science,* 36, 762-784

Haitovsky, Y. (1968): Missing Data in Regression Analysis, *Journal of the Royal Statistical Society B* 30, 67-81

Hanson, R.H. (1978): *The Current Population Survey: Design and Methodology*, Technical Paper No. 40, U.S. Bureau of the Census

Harrell, F.E. Jr. (2001): *Regression Modelling Strategies*, Springer

Hartung, J. (1989): *Multivariate Statistik,* 4. Auflage, Oldenbourg Verlag

Hartung, J., Elpelt, B. (1995): *Multivariate Statistik,* 5. Auflage, Oldenbourg Verlag

Hasselblad, V., Stead, A.G., Galke, W. (1980): Analysis of Coarsely Grouped Data from the Lognormal Distribution, *Journal of the American Statistical Association,* 75, 771-778

Hastings, W.K. (1970): Monte Carlo Sampling Methods using Markov Chains and Their Applications, *Biometrika,* 57, 97-109

Heckerman, D., Geiger, D., Chickering, D. (1995): Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, *Machine Learning,* 20, 197-243

Herlocker, J.A., Konstan, J.A., Borchers, A., Riedl, J. (1999): An Algorithmic Framework for Collaborative Filtering, In: *Proceedings of the 22nd Annual International ACM SIGIR*, 230-237

Herlocker, J.A., Konstan, J.A., Riedl, J. (2000): Explaining Collaborative Filtering Recommendations, In: *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, 241-250

Herzog, T.N., Rubin, D.B. (1983): Using Multiple Imputations to Handle Nonresponse in Sample Surveys, In: Madow, W.G., Nisselson, H., Olkin, I. (Eds.): *Incomplete Data in Sample Surveys,* 2, 209-245

Hill, W., Stead, L., Rosenstein, M., Furnas, G. (1995): Recommending and Evaluating Choices in a Virtual Community of Use, *Proceedings of the Conference on Human Factors in Computing Systems,* 194-201

Hoeting, J., Raftery, A.E., Madigan, D. (1996): A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression, *Computational Statistics and Data Analysis,* 22, 251-270

Hofstede, T.F., Wedel, M., Steenkamp, J.-B. E.M. (2002): Identifying Spatial Segments in International Markets, *Marketing Science,* 21, 160-177

Hornik, K. (2004): *The R FAQ - R Foundation for Statistical Computing*, http://www.ci.tuwien.ac.at/hornik/R/

Hosmer, D.W., Lemeshow, S. (2000): *Applied Logistic Regression*, Wiley

Ittner, D.J., Lewis, D.D., Ahn, D.D. (1995): Text Categorization of Low Quality Images, *Symposium on Document Analysis and Information Retrieval,* 301-315

Jackson, E.C. (1968): Missing Values in Linear Multiple Discriminant Analysis, *Biometrics,* 24, 835-844

Jeffreys, A. (1961): *The Theory of Probability,* Cambridge University Press

Johnson, V.E. (1996): Studying Convergence of MCMC Algorithms Using Coupled Sample Paths, *Journal of the American Statistical Association,* 91, 154-166

Kamakura, W.A., Wedel, M. (1997): Statistical Data Fusion for Cross-Tabulation, *Journal of Marketing Research,* 34, 485-498

KARYPIS, G. (2000): Evaluation of Item-Based Top-N Recommendation Algorithms, *Proceedings of the 10th International Conference on Information and Knowledge Management,* 247-254

KENDALL (1938): A New Measure of Rank Correlation, *Biometrika,* 30, 81-93

KIM, J.O., CURRY, J. (1977): The Treatment of Missing Data in Multivariate Analysis, *Sociological Methods an Research,* 6, 215-239

KONSTAN, J.A., MILLER, B.N., MALTZ, D., HERLOCKER, J.L., GORDON, L.R., RIEDL, J. (1997): GroupLens: Applying Collaborative Filtering to Usenet News, *Communications of the ACM,* 40, 77-87

KRANTZ, D.H., LUCE, R.D., SUPPES, P., TVERSY, A. (1971): *Foundations of Measurement, Volume I, Additive and Polynomial Representations*, Academic Press

KVALSETH, T.O. (1985): Cautionary Note about R^2, *The American Statistician,* 39, 279-285

LAM, W., BACCHUS, F. (1994): Learning Bayesian Belief Networks - An Approach based on the MDL Principle, *Computational Intelligence,* 10, 269-293

LANG, K. (1995): NewsWeeder: Learning to Filter Netnews, *Proceedings of the 12th International Conference on Machine Learning,* 331-336

LAURITZEN, S.L. (1992): Propagation of Probabilities, Means and Variances in Mixed Graphical Association Models, *Journal of the American Statistical Association,* 87, 1098-1108

LAURITZEN, S.L., SPIEGELHALTER, D.J. (1988): Local Computations with Probabilities on Graphical Structures and their Applications to Expert Systems, *Journal of the Royal Statistical Society B,* 50, 157-224

LI, P., YAMADA, S. (2004): A Movie Recommender System Based on Inductive Learning, *Proceedings of the 2004 IEEE,* 318-323

LINDEN, G., SMITH, B., YORK, Y. (2003): Amazon.com Recommendations, *IEEE Internet Computings,* 7, 76-80

Little, R.J.A. (1986): Survey Nonresponse Adjustments for Estimates of Means, *International Statistical Review,* 54, 139-157

Little, R.J.A. (1988): A Test of Missing Completely at Random for Multivariate Data With Missing Values, *Journal of the American Statistical Association,* 83, 1198-1202

Little, R.J.A. (1993): Post-Stratification: A Modeler's Perspective, *Journal of the American Statistical Association,* 88, 1001-1012

Little, R.J.A., Rubin, D.B. (2002): *Statistical Analysis With Missing Data*, Wiley Series in Probability and Statistics

Little, R.J.A., Smith, P.J. (1987): Editing and Imputation for Quantitative Survey Data, *Journal of the American Statistical Association,* 82, 58-68

Littlestone, N., Warmuth, M. (1994): The Weighted Majority Algorithm, *Information and Computation,* 108, 212-261

Liu, C., Liu, J., Rubin, D.B. (1992): A Variational Control Variable for Assessing the Convergence of the Gibbs Sampler, *Proceedings of the American Statistical Association, Statistical Computing Section*, 74-78

Liu, C.H., Rubin, D.B. (1996): Markov-Normal Analysis of Iterative Simulations before their Convergence, *Journal of Econometrics,* 75, 69-78

Liu, C.H., Rubin, D.B. (2002): Markov-Normal Analysis of Iterative Simulations before their Convergence: Redesign for Better Convergence , *Statistica Sinica,* 12, 751-767

Lösel F., Wüstendörfer, W. (1974): Zum Problem unvollständiger Datenmatrizen in der empirischen Sozialforschung, *Kölner Zeitschrift für Soziologie und Sozialpsychologie,* 26, 342-357

Lord, F.M. (1955): Estimation of Parameters from Incomplete Data, *Journal of the American Statistical Association,* 50, 870-876

Luce, R.D. (1959): *Individual Choice Behavior: A Theoretical Analysis*, Wiley

Maddala, G.S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press

MADIGAN, D., YORK, J. (1995): Bayesian Graphical Models for Discrete Data, *International Statistical Review*, 63, 215-232

MAGEE, L. (1990): R^2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests, *The American Statistician*, 44, 250-253

MAGNUS, J.R., NEUDECKER, H. (1988): *Matrix Differential Calculus With Applications in Statistics and Econometrics*, Wiley

MATTHAI, A. (1951): Estimation of Parameters from Incomplete Data With Application to Design of Sample Surveys, *Sankhya: The Indian Journal of Statistics,* 2, 145-152

MCCULLAGH, P. (1980): Regression Models for Ordinal Data, *Journal of the Royal Statistical Society B,* 42, 109-142

MCCULLAGH, P., NELDER, J.A. (1989): *Generalized Linear Models,* Chapman & Hall

MCCULLOCH, R.E., ROSSI, P.E. (1994): An Exact Likelihood Analysis of the Multinomial Probit Model With Fully Identified Parameters, *Journal of Econometrics,* 64, 207-240

MCFADDEN (1974): The Measurement of Urban Travel Demand, *Journal of Public Economics*, 3, 303-328

MCQUEEN (1967): Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,* 281-297

MELVILLE, P., MOONEY, R.J., NAGARAJAN, R. (2002): Content-Boosted Collaborative Filtering for Improved Recommendations, *Proceedings of the 18th National Conference on Artificial Intelligence*, 187-192

MENARD, S. (1995): *Applied Logistic Regression Analysis*, Sage

METROPOLIS, N., ROSENBLUTH, A.W., ROSENBLUTH, M.N., TELLER, A.H., TELLER, E. (1953): Equation of State Calculations by Fast Computing Machine, *Journal of Chemical Physics,* 21, 1087-1091

Mild, A., Natter, M. (2002): Collaborative Filtering or Regression Models for Internet Recommendation Systems?, *Journal of Targeting, Measurement and Analysis for Marketing*, 10, 4, 304-313

Minka,T.M. (2001a): *A Family of Algorithms for Approximate Bayesian Inference*, Dissertation, MIT

Minka,T.M. (2001b): Expectation Propagation for Approximate Bayesian Inference, *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*, 362-369

Möntmann, V., Bollinger, G., Herrmann, A. (1983): Tests auf Zuverlässigkeit auf „Missing Data“, In: Wilke, H. (Ed.): *Statistiksoftware in der Sozialforschung*, Quorum

Montgomery, A.L., Rossi, P.E. (1999): Estimating Price Elasticities with Theory-Based Priors, *Journal of Marketing Research,* 36, 413-423

Mooney, R.J., Roy, L. (2000): Content-based Book Recommending Using Learning for Text Categorization, *Proceedings of the 5th ACM Conference on Digital Libraries*, 195-204

Myrtveit, I., Stensrud, E., Olsson, U.H. (2001): Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods, *IEEE Transactions on Software Engineering*, 27, 999-1013

Nagelkerke, N.J.D. (1991): A Note on a General Definition of the Coefficient of Determination, *Biometrika*, 78, 691-692

Nocedal, J., Wright, S.J. (1999): *Numerical Optimization*, Springer

Norusius, M.J. (1997): *SPSS Professional Statistics 7.5*, SPSS Publications

Oh, H.L., Scheuren, F.S. (1983): Weighting Adjustment for Unit Nonresponse, In: Madow, W.G., Olkin, I., Rubin, D.B. (Eds.): *Incomplete Data in Sample Surveys*, 2, 143-184

O'Hagan, A. (1978): Curve Fitting and Optimal Design for Prediction, *Journal of the Royal Statistical Society B*, 40, 1-42

Opper, M., Winther, O. (1999): A Bayesian Approach to Online-Learning, In: Saad, D. (Ed.): *Online Learning in Neural Networks*, 363-378

Pazzani, M., Billsus, D. (1997): Learning and Revising User Profiles: The Identification of Interesting Web Sites, *Machine Learning*, 27, 313-331

Pazzani, M. (1999): A Framework for Collaborative, Content-Based and Demographic Filtering, *Artificial Intelligence Review*, 13, 393-408

Pearl, J. (1982): Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach, *Proceedings of the Second National Conference on Artificial Intelligence*, 133-136

Pearl, J. (1986): Fusion, Propagation, and the Structuring in Belief Networks, *Artificial Intelligence*, 29, 241-288

Pearson, K. (1900): On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables in such that it can be Reasonably Supposed to Have Arisen from Random Sampling, *Philosophical Magazine*, 5, 157-175

Pearson, K. (1901): On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine*, 6, 559-572

Phillips, M.J. (1993): Contingency Tables With Missing Data, *The Statistician*, 42, 9-18

Platek, R., Gray, G.B. (1983): Imputation Methodology: Total Survey Error, In: Madow, W.G., Olkin, I., Rubin, D.B. (Eds.): *Incomplete Data in Sample Surveys*, Volume 2, Academic Press, 259-272

Polak, E., Ribiere, G. (1969): *Note sur la convergence de méthodes de directions conjugées*, Revue Francaise d'Informatique et de Recherche Opérationelle, 16, 35-43

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (2002): *Numerical Recipes in C++*, Cambridge University Press

Quinlan, J.R. (1993): *C4.5: Programs for Machine Learning*, Morgan Kaufman

Raftery, A.E. (1986): A Note on Bayes-Factors for Log-linear Contingency Tables With Vague Prior Information, *Journal of the Royal Statistical Society B*, 48, 249-250

Raftery, A.E. (1995): Bayesian Model Selection in Social Research, *Sociological Methodology*, 25, 111-163

Raftery, A.E., Madigan, D., Hoeting, J. (1997): Bayesian Model Averaging for Linear Regression Models, *Journal of the American Statistical Association*, 92, 179-191

Rao, C.R., Toutenburg, H. (1995): *Linear Models - Least Squares and Alternatives*, Springer

Rasch, D., Herrendörfer, G. (1982): *Statistische Versuchsplanung*, VEB Deutscher Verlag der Wissenschaften

Rasch, D., Kubinger, K.D. (2006): *Statistik*, Elsevier

Rasmussen, C.E. (1996): *Evaluation of Gaussian Processes and other Methods for Non-Linear Regression*, Dissertation, University of Toronto

Rasmussen, C.E., Williams, C.K.I. (2006): *Gaussian Processes for Machine Learning*, MIT Press

Rennie, J.D.M., Srebro, N. (2005): Fast Maximum Margin Matrix Factorization for Collaborative Prediction, *Proceedings of the 22nd International Conference on Machine Learning*, 713-719

Resnick, P., Neophytos, I., Mitesh, S., Bergstrom, P., Riedl, J. (1994): GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *Proceedings of the Computer Supported Cooperative Work Conference*, 175-186

Richardson, S., Green, P.J. (1997): On Bayesian Analysis of Mixtures with an unknown Number of Components (with Discussion), *Journal of the Royal Statistical Society B*, 59, 731-792

Ritter, C., Tanner, M.A. (1992): Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler, *Journal of the American Statistical Association*, 87, 861-868

ROBERTS (1979): *Measurement Theory*, Springer

ROBERTS (1992): Convergence Diagnostics of the Gibbs Sampler, In: Berger, J., Bernardo, J.M., Dawid, A.P., Smith, A.F.M. (Eds.): *Bayesian Statistics 4*, 775-782

ROBERTSON, S.E., WALKER, S., JONES, S., HANCOCK-BEAULIEU, M.M., GATFORD, M. (1979): Okapi at TREC-3, *Proceedings of the 3rd Text Retrieval Conference*, 109-126

ROHWER, G., PÖTTER, U. (2005): *Transitional Data Analysis*, http://www.stat.ruhr-uni-bochum.de/tman.html

ROJAS, R. (1993): *Theorie der neuronalen Netze*, Springer

ROSSI, P.E., MCCULLOCH, R.E., ALLEBY, G.M. (1996): The Value of Purchase History Data in Target Marketing, *Marketing Science*, 15, 321-340

RUBIN, D.B. (1976): Inference and Missing Data, *Biometrika*, 63, 581-592

RUBIN, D.B. (1978): Multiple Imputation in Sample Surveys, *Proceedings of the American Statistical Association 1978*, 20-34

RUBIN, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys*, Wiley

RUBIN, D.B. (1996): Multiple Imputation after 18+ years, *Journal of the American Statistical Association*, 91, 473-489

RUMELHART, D.E., HINTON, G., WILLIAMS, R. (1986): Learning Internal Representations by Error Propagation, In: Rumelhart, D.E., McCelland, J.L. (Eds.): *Parallel Distributed Processing*, MIT Press, 318-362

RUMMEL, R.J. (1970): *Applied Factor Analysis*, Evanston

SANDOR, Z., WEDEL, M. (2001): Designing Conjoint Choice Experiments Using Managers' Prior Beliefs, *Journal of Marketing Research*, 28, 430-444

SALTON, G., BUCKLEY, C. (1988): Term-Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, 24, 513-523

SARWAR, B., KARYPIS, G., KONSTAN, J., RIEDL, J. (2000a): Analysis of Recommendation algorithms for E-Commerce, *Proceedings of the ACM E-Commerce 2000 Conference*, 158-167

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. (2000b): Application of Dimensionality Reduction in Recommender System - A Case Study, *Proceedings of the WEBKDD*, 82-90

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. (2000): Item-based Collaborative Filtering Recommendation Algorithms, *Proceedings of the 10th International Conference on the World Wide Web*, 285-295

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. (2002): Recommender Systems for Large-Scale E-Commerce: Scalable Neighborhood Formation Using Clustering, *Proceedings of the 5th International Conference on Computer and Information Technology (ICCIT 2002)*, 407-412

Schafer, J.L. (1997): *Analysis of Incomplete Multivariate Data*, Chapman & Hall

Schlecht, V., Gaul, W. (2004): Fuzzy Two-Mode Clustering vs. Collaborative Filtering, In: Weihs, C., Gaul, W. (Eds.): *Classification the Ubiquitous Challenge*, Springer, 410-417

Schnell, R. (1986): *Missing-Data Probleme in der empirischen Sozialforschung*, Dissertation, Ruhr-Universität Bochum

Schwab, G. (1991): *Fehlende Werte in der angewandten Statistik*, Deutscher Universitätsverlag

Schwarz, G. (1978): Estimating the Dimension of a Model, *The Annals of Statistics*, 6, 461-464

Shachter, R. (1986): Evaluating Influence Diagrams, *Operations Research*, 34, 321-340

Shardanand, U., Maes, P. (1995): Social Information Filtering: Algorithms for Automating „Word of Mouth“, *Proceedings of CHI'95 Conference on Human Factors in Computing Systems*, 210-217

Shepard, R.N., Arabie, P. (1979): Additive Clustering Representation of Similarities as Combinations of Discrete Overlapping Properties, *Psychological Review*, 86, 87-123

Simonoff, J.S. (1998): Logistic Regression, Categorical Predictors, and Goodness-of-Fit: It Depends on Who You Ask, *The American Statistician*, 52, 10-14

Simonoff, J.S. (2003): *Analyzing Categorical Data*, Springer

Singhal, A., Buckley, C., Mitra, M. (1996): Pivoted Document Length Normalization, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 21-29

Sneath, P.H.A. (1957): The Application of Computers to Taxonomy, *Journal of General Microbiology*, 17, 201-226

Snedecor, G. R., Cochran, W.G. (1989): *Statistical Models*, Iowa State University Press

Soboroff, I.M., Nicholas, C.K. (1999): Combining Content und Collaboration in Text Filtering, *Proceedings of the IJCAI'99 Workshop on Machine Learning for Information Filtering*, 86-91

Somers, R.H. (1962a): A New Asymmetric Measure of Association for Ordinal Variables, *American Sociological Review*, 27, 799-811

Somers, R.H. (1962b): A Similarity Between Goodman and Kruskal's tau and Kendall's tau, With a Partial Interpretation of the Latter, *Journal of the American Statistical Association*, 57, 804-812

Spearman, C. (1904): The Proof and Measurement of Association Between two Things, *American Journal of Psychology*, 15, 72-101

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Linde, v.d.A. (2002): Bayesian Measures of Model Complexity and Fit, *Journal of the Royal Statistical Society B*, 64, 583-639

Srebro, N., Jaakkola, T. (2003): Weighted Low-Rank Approximations, *Proceedings of the 20th International Conference on Machine Learning*, 720-727

Srebro, N., Rennie, J.D.M., Jaakkola, T.S. (2005): Maximum-Margin Matrix Factorization, *Advances in Neural Information Processing Systems*, 17, 1329-1336

STEWART, G.W. (1993): On the Early History of the Singular Value Decomposition, *SIAM Review*, 35, 511-566

STEWART, M.B. (1983): On Least Squares Estimation When the Dependent Variable is Grouped, *Review of Economic Studies*, 50, 737-753

STRANG, G. (1998): *Lineare Algebra*, Springer

STUART, A. (1904): The Estimation and Comparison of Strength of Association in Contingency Tables, *Biometrika*, 40, 105-110

TANNER, M.A., WONG, W.H. (1987): The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, 82, 528-550

TIERNEY, L. (1994): Markov Chains for Exploring Posterior Distributions, *Annals of Statistics*, 22, 1701-1762

TRAIN, K.E. (2003): *Discrete Choice Models With Simulation*, Cambridge University Press

TRAN, T., COHEN, R. (2000): Hybrid Recommender Systems for Electronic Commerce, *Proceedings of the 17th National Conference on Artificial Intelligence*, 78-84

UPTON, G. (1991): The Explanatory Analysis of Survey Data Using Log-linear Models, *The Statistician*, 40, 169-182

VAN RIJSBERGEN, C.J. (1979): *Information Retrieval*, Butterworths

VEALL, M.R., ZIMMERMANN, K.F. (1992): *Pseudo-R^2's in the Ordinal Probit Model*, Butterworths

VEHTARI, A. (2001): *Bayesian Model Assessment and Selection using Expected Utilities*, Dissertation, Technical University of Helsinki

WALKER, S.H., DUNCAN, D.B. (1967): Estimation of the Probability of an Event as a Function of Serveral Independent Variables, *Biometrika*, 54, 167-178

WEDEL, M., PIETERS, R. (2000): Eye Fixation on Advertisements and Memory for Brands: A Model and Findings, *Marketing Science*, 19, 297-312

Wickens (1989): *Multidimensional Contingency Tables Analysis in the Social Sciences*, Lawrence Erlbaum

Wilks, S.S. (1932): Moments and Distribution of Estimates of Population Parameters from Fragmentary Samples, *The Annals of Mathematical Statistics*, 3, 163-195

Williams, C.K.I. (1996): Regression with Gaussian Processes, In: Ellacott, S.W., Mason, J.C., Anderson, I.J. (Eds.): *Mathematics of Neural Networks: Models, Algorithms, Applications*, Kluver, 378-382

Williams, C.K.I., Rasmussen, C.E. (1996): Gaussian Processes for Regression, In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (Eds.): *Advances in Neural Information Processing Systems*, 8, 514-520

Witten, I. H., Frank, E. (2005): *Data Mining*, Elsevier

Wothke, W. (1993): Nonpositive Definite Matrices in Structural Modeling, In: Bollen, K.A., Long, J.S. (Eds.): *Testing Structural Equation Models*, Sage Publications

Yu, S., Yu, K., Tresp, V., Kriegel, H.-P. (2006): Collaborative Ordinal Regression, to appear in: *Proceedings of the 23rd International Conference on Machine Learning*

Zeger, S.L., Karim, M.R. (1991): Generalised Linear Models With Random Effects, *Journal of the American Statistical Association*, 86, 79-86

Zellner, A. (1971): *An Introduction to Bayesian Inference in Econometrics*, Wiley

Zellner, A., Min, C.-K. (1995): Gibbs Sampler Convergence Criteria, *Journal of the American Statistical Association*, 90, 921-927

Ziegler, C.N., McNee, S.M., Konstan, J.A., Lausen, G. (2005): Improving Recommendation Lists through Topic Diversification, *Proceedings of the 14th International Conference on the World Wide Web*, 22-32

Anhang A

Schätzung des Nutzens

Als Maß für den Nutzen einer Liste von Empfehlungen, die auf Basis von Schätzern für einen Nutzer generiert wird, wird der Breese-Wert (Abschnitt 5.8) verwendet. Da der Breese-Wert stark mit Präzision und Recall korreliert ist, ergeben sich durch die zusätzliche Angabe des Breese-Werts nicht immer neuen Erkenntnisse.

A.1 Breese-Werte Nicht-Bayes'scher Verfahren

Die Breese-Werte R_B einiger Nicht-Bayes'scher Methoden auf Grundlage des 1067 Nutzer und 418 Items umfassenden Teils des MovieLens-Datensatzes sind der folgenden Tabelle zu entnehmen:

	Anteil des Testdatensatzes am gesamten Datensatz:								
	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
$\hat{S}_Y^1$	85,87	74,37	71,76	67,37	61,01	64,22	57,98	54,58	48,56
$\hat{S}_Y^2$	86,84	77,62	72,27	68,25	66,86	63,51	63,17	61,67	65,34
OZC	87,05	77,91	71,45	68,51	65,55	63,31	61,98	60,65	58,30
OMF	82,89	76,50	72,11	69,55	67,30	65,84	63,90	61,87	56,14
HP	85,28	77,19	73,06	70,10	68,59	68,41	68,03	67,17	66,88

Tabelle A.1: Breese-Werte R_B verschiedener Nicht-Bayes'scher kollaborativer Verfahren und des hybriden Verfahrens nach Pazzani (1999) hinsichtlich des Testdatensatzes bei unterschiedlichen Anteiles des Testdatensatzes am gesamten Datensatz

Die wichtigsten Nicht-Bayes'schen kollaborativen Verfahren sind das zweimodale $\hat{S}_Y^1-$ und $\hat{S}_Y^2-$Clusterverfahren ($\hat{S}_Y^1$ und $\hat{S}_Y^2$), das ordinale zweimodale Clusterverfahren (OZC) und das Verfahren zur ordinalen Matrixfaktorisation (OMF) nach Rennie, Srebro (2005). Ein weiteres Nicht-Bayes'sches Verfahren ist das hybride Verfahren nach Pazzani (HP). Bei hohen Anteilen des Testdatensatzes an der gesamten Datenmenge sind die auf Grundlage des hybride Verfahren nach Pazzani berechneten Schätzer erwartungsgemäß besser zur Generierung von Empfehlungslisten geeignet, da sie zusätzliche Information miteinbeziehen.

A.2 Breese-Werte der Bayes'schen Verfahren

Die Breese-Werte der Bayes'schen Verfahren und deren Diskussion sind Abschnitt 8.4 zu entnehmen.

Anhang B

Auswirkungen verzerrter Datenstrukturen

In diesem Teil des Anhangs werden die Auswirkungen verzerrter Datenstrukturen auf die Ergebnisse verschiedener Verfahren untersucht.

B.1 Auswirkungen auf den geschätzten Nutzen

In Tabelle sind die Breese-Werte bei unterschiedlichen Verzerrungsgraden für das zweimodale $\hat{S}_Y^1-$Clusterverfahren ($\hat{S}_Y^1$), das zweimodale $\hat{S}_Y^2-$Clusterverfahren ($\hat{S}_Y^2$), das ordinale zweimodale Clusterverfahren (OZC) und das Verfahren der ordinalen Matrixfaktorisation (OMF) abgebildet.

Die Breese-Werte nehmen im allgemeinen mit steigendem Verzerrungsgrad ab. Der unerwartete Anstieg der Breese-Werte beim maximalen Verzerrungsgrad (90%) ist dadurch zu erklären, daß bei diesem Verzerrungsgrad nur noch sehr wenige hohe Bewertungen im Testdatensatz vorhanden sind und gleichzeitig die (auf Basis der Trainingsdaten berechneten) Nutzer-Durchschnittsbewertungen $\bar{y}_{i.}, i = 1, ..., I$, sehr groß ausfallen. Hierdurch werden auch die idealen Nutzen-Werte $R_{B,i}^{MAX}$ kleiner. Der Anstieg der Breese-Werte bei extrem hohen Verzerrungsgraden kann als Anzeichen dafür interpretiert werden, daß die Eignung der Schätzer zur Generierung guter Empfehlungslisten durch verzerrte Trainingsdaten nur in begrenztem Umfang verschlechtert werden kann.

Gütemaße	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
$\hat{S}_Y^1$	68,90	67,09	64,93	62,28	65,74
$\hat{S}_Y^2$	69,69	68,03	65,34	63,93	65,49
OZC	70,05	67,36	65,42	64,34	65,86
OMF	68,56	67,00	65,22	63,71	64,96

Tabelle B.1: Breese-Werte verschiedener Nicht-Bayes'scher kollaborativer Verfahren (zweimodales $\hat{S}_Y^1-$ und $\hat{S}_Y^2-$Clusterverfahren, OZC und OMF) bei unterschiedlichen Graden der Verzerrung

Gütemaße	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
AAD	0,729	0,760	0,805	0,884	0,936
Präzision (Schwellenwert: 4,5)	0,615	0,541	0,465	0,382	0,255
Recall (Schwellenwert: 4,5)	0,210	0,219	0,218	0,263	0,278
Präzision (Schwellenwert: 4,75)	0,693	0,657	0,577	0,502	0,317
Recall (Schwellenwert: 4,75)	0,084	0,093	0,093	0,103	0,097
Präzision (Schwellenwert: 5,0)	0,772	0,700	0,667	0,609	0,400
Recall (Schwellenwert: 5,0)	0,025	0,023	0,025	0,028	0,025
Breese-Wert (R_B)	71,77	69,46	69,07	68,74	68,77

Tabelle B.2: AAD, Präzision und Recall (untereinander) bei unterschiedlichen Graden der Verzerrung in Bezug auf das hybride Verfahren nach Pazzani (1999)

B.2 Auswirkungen auf das hybride Verfahren nach Pazzani (1999)

In Tabelle B2 sind das *AAD*, die Präzision und der Recall in Bezug auf das hybride Verfahren nach Pazzani (1999) aufgelistet. Präzision und Recall werden zu drei verschiedenen Präzision-und-Recall Schwellenwerten (4,5, 4,75 und 5,0) angegeben.

Es mag auf den ersten Blick verwunderlich erscheinen, daß das hybride Verfahren, das zusätzlich zu den Bewertungen auch die Eigenschaften der Items benutzen kann, eine Verzerrung der Daten hinsichtlich des *AAD* nur unerheblich

besser als rein kollaborative Verfahren kompensieren kann. Hierbei ist aber zu bedenken, daß das Verfahren nach Pazzani kein selbstständiges hybrides Verfahren ist, sondern lediglich eine kontentbasierte Ergänzung des Nutzer-basierten Ähnlichkeitsverfahrens. Auch wenn die Eigenschaften der Items zur Berechnung der Korrelationen zwischen den Nutzern verwendet werden, ändert das nichts daran, daß die Schätzer eines Ähnlichkeitsverfahrens nur gewichtete Summen der Bewertungsdaten sind. Sind diese verzerrt, so überträgt sich deren Verzerrung auf die Schätzer. Dies führt zur Vergrößerung der *AAD*-Werte mit steigendem Verzerrungsgrad. Dennoch erhält man im Hinblick auf das Verfahren nach Pazzani (1999) Dank der Verwendung zusätzlicher Information immerhin etwas höhere Breese-Werte als für die übrigen Verfahren.

B.3 Auswirkungen auf das kollaborative GP- Verfahren

Die Ergebnisse der hierarchischen GP-Verfahrens nach Yu et. al. (2006) bei unterschiedlichen Verzerrungsgraden sind in Tabelle A2 wiedergegeben:

Gütemaß	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
AAD	0,689	0,734	0,802	0,920	1,008
Prec. (Schwellenwert:4,5)	0,691	0,651	0,567	0,492	0,354
Rec. (Schwellenwert:4,5)	0,206	0,238	0,226	0,268	0,282
Prec. (Schwellenwert:4,0)	0,446	0,392	0,309	0,198	0,114
Rec. (Schwellenwert:4,0)	0,557	0,603	0,603	0,671	0,704
Prec. (Schwellenwert:5,0)	0,813	0,736	0,702	0,663	0,416
Rec. (Schwellenwert:5,0)	0,022	0,020	0,021	0,026	0,027
Breese-Wert (R_B)	72,79	71,31	69,41	67,28	68,20

Tabelle B.3: AAD, Präzision und Recall (untereinander) bei unterschiedlichen Graden der Verzerrung (hierarchisches GP-Verfahren nach Yu et. al. (2006)

Bemerkenswert ist, daß das kollaborative GP-Verfahren nach Yu et. al. (2006) mit

deutlich weniger Information und weitaus geringerem numerischen Aufwand bei geringen Verzerrungsgraden bessere Ergebnisse erzielt als das hybride Verfahren nach Pazzani (1999). Bei höheren Verzerrungsgraden ist das hybride Verfahren nach Pazzani dem kollaborativen GP-Verfahren zunehmend überlegen. Dies kann als Hinweis auf die bessere Eignung hybrider Verfahren zur Vorhersage von Bewertungsdaten in der Praxis interpretiert werden.

B.4 Auswirkungen auf die hybriden Bayes'schen Verfahren

In diesem Abschnitt werden die Auswirkungen verzerrter Datenstrukturen auf das hierarchische lineare Regressionsmodell (HBLR) und das hybride GP-Verfahren untersucht. Für beide Verfahren wurde die Regressorenkombination M_8 verwendet. Da der Trainingsdatensatz verhältnismäßig groß ist, wären bei unverzerrten Daten trotz des mittleren Komplexitätsgrades des M_8-Modells auch im Rahmen des hybriden GP-Verfahrens keine Überanpassungseffekte zu erwarten.

Gütemaß	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
AAD	0,748	0,779	0,833	0,925	0,980
Prec. (Schwellenwert:4,5)	0,627	0,587	0,528	0,408	0,274
Rec. (Schwellenwert:4,5)	0,093	0,112	0,123	0,163	0,177
Breese-Wert (R_B)	66,76	64,13	60,79	57,89	57,15

Tabelle B.4: AAD, Präzision, Recall und Breese-Wert bei unterschiedlichen Graden der Verzerrung im Hinblick auf das hierarchische lineare Regressionsmodell (HBLR) unter Verwendung der Regressorenkombination M_8

Der konservative HBLR-Ansatz führt zu etwas größeren AAD-Werten als das hybride GP-Verfahren und weist bei höheren Verzerrunggraden kleinere Breese-Werte auf. Hieraus kann gefolgert werden, daß der weitaus rechenzeitintensivere HBLR-Ansatz im Hinblick auf verzerrte Datensätze starke Anpassungsprobleme hat und weniger geeignet zur Generierung nützlicher Empfehlungslisten erscheint.

Gütemaß	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
AAD	0,734	0,752	0,874	0,892	0,894
Prec. (Schwellenwert:4,5)	0,546	0,475	0,390	0,293	0,325
Rec. (Schwellenwert:4,5)	0,157	0,198	0,218	0,253	0,234
Breese-Wert (R_B)	66,27	63,72	61,33	61,30	61,77

Tabelle B.5: AAD, Präzision, Recall und Breese-Wert bei unterschiedlichen Graden der Verzerrung in Bezug auf das hybride hierarchische GP-Modell unter Verwendung der Regressorenkombination M_8

Der HBLR-Ansatz bleibt hinsichtlich der Breese-Werte insbesondere bei hohen Verzerrungsgraden hinter dem hybriden Verfahren von Pazzani (1999) und dem kollaborativen GP-Verfahren zurück. Beide hybriden Bayes'schen Verfahren führen hinsichtlich verzerrter Datensätze zu kleineren Breese-Werten als die Nicht-Bayes'schen kollaborativen Verfahren.

B.5 Auswirkungen auf Schätzer für neue Items

In diesem Abschnitt wird die Auswirkung verzerrter Trainingsdatensätze auf die Schätzer im Hinblick auf neue Items untersucht. Hierzu müssen die verzerrten Test- und Trainingsdatensätze (Kapitel 4) geeignet modifiziert werden. Aus den verzerrten Trainingsdatensätzen werden nur die Bewertungen verwendet, die sich aus bekannte Items aus der Menge J^K beziehen. Entsprechend sind nur noch die Bewertungen Bestandteil der zugehörigen Testdatensätze, die neue Items aus J^N zum Gegenstand haben. Hierdurch werden die Trainingsdatensätze etwas kleiner und die Testdatensätze erheblich kleiner. Da die Testdatensätze auch vorher eher klein gewählt wurden (um möglichst hohe Verzerrungsgrade zu erreichen), stehen in den modifizierten Testdatensätzen pro Nutzer oft nur wenige Items zur Verfügung. Dadurch erhöhen sich die Breese-Werte im Vergleich zu den ursprünglichen verzerrten Test- und Trainingsdatensätzen. Somit sind die folgenden Breese-Werte nicht geeignet, um den Nutzen der Empfehlungslisten für neue Items bei verzerrter Datenstruktur mit dem Nutzen der Empfehlungslisten bezüglich bekannter Items bei verzerrten Test- und Trainingsdaten zu vergleichen. Vielmehr ist der Zweck dieser Untersuchung, die verschiedenen Verfahren

zur Vorhersage neuer Items untereinander im Hinblick auf ihre Eignung für verzerrte Datensätze zu vergleichen.

Die Mittelwerte der modifizierten verzerrten Test- und Trainingsdatensätze ändern sich erwartungsgemäß nur geringfügig:

	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
Trainingsdaten	3,54	3,62	3,70	3,80	3,85
Testdaten	3,24	3,07	2,87	2,69	2,56

Tabelle B.6: Durchschnittsbewertung der modifizierten verzerrten Test- und Trainingsdaten

Das Basis-Modell (MW) bildet Schätzer für neue Items $j \in J^N$, indem es bezüglich jedes Nutzers i dessen mittlere Bewertung $\bar{y}_{i.}$ verwendet: $\hat{S}_{ij}^{MW} = \bar{y}_{i.}$. Erstellt man auf der Grundlage solcher Schätzer Empfehlungslisten für jeden einzelnen Nutzer, so ist die Reihenfolge der Items dem Zufall überlassen. Deshalb sind im allgemeinen keine hohen Breese-Werte im Hinblick auf den MW-Ansatz zu erwarten. Daher illustrieren die folgenden Werte, wie viel größer die auf Basis des beschriebenen Testdatensatzes berechneten Breese-Werte ausfallen:

Gütemaß	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
AAD	0,860	0,899	0,946	1,031	1,079
$Prec.$ (Schwellenwert:4,5)	0,429	0,417	0,333	0,235	0,250
$Rec.$ (Schwellenwert:4,5)	0,003	0,007	0,009	0,014	0,015
R_B	80,65	79,47	77,14	75,63	76,29

Tabelle B.7: AAD, Präzision, Recall und Breese-Wert bei unterschiedlichen Graden der Verzerrung (MW-Methode)

Tabelle B.8 enthält AAD, Präzision, Recall und Breese-Wert im Hinblick auf das SL-Verfahren zur Approximation der Item-Cluster Zugehörigkeit auf der Basis

Gütemaß	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
AAD	0,743	0,740	0,737	0,720	0,719
Prec. (Schwellenwert:4,5)	0,675	0,461	0,351	0,279	0,082
Rec. (Schwellenwert:4,5)	0,203	0,159	0,038	0,041	0,029
R_B	89,29	88,40	87,85	83,67	82,02

Tabelle B.8: AAD, Präzision, Recall und Breese-Wert bei unterschiedlichen Graden der Verzerrung im Hinblick auf den SL-Ansatz zur Approximation der Item-Cluster Zugehörigkeit und das zweimodalen $\hat{S}_Y^1$-Clusterverfahren auf Basis der Variablenkombination M_3 ($K = L = 10$)

Gütemaß	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
AAD	0,767	0,760	0,762	0,729	0,736
Prec. (Schwellenwert:4,5)	0,601	0,490	0,398	0,281	0,018
Rec. (Schwellenwert:4,5)	0,145	0,098	0,092	0,055	0,015
R_B	88,28	88,22	87,07	83,05	82,34

Tabelle B.9: AAD, Präzision, Recall und Breese-Wert bei unterschiedlichen Graden der Verzerrung im Hinblick auf den SL-Ansatz zur Approximation der Item-Cluster Zugehörigkeit und das zweimodalen $\hat{S}_Y^1$-Clusterverfahren auf Basis der Variablenkombination M_8 ($K = L = 10$)

des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens mit den Clustergrößen $K = L = 10$ und der in Bezug auf diesen Ansatz erfolgreichsten Variablenkombination M_3. Ein Vergleich mit den Breese-Werten des Basis-Modells gibt ein Gefühl dafür, wie stark der ungewöhnlich kleine Trainingsdatensatz die Breese-Werte erhöht. Im Gegensatz zu allen anderen beschriebenen Verfahren zur Approximation der Item-Cluster Zugehörigkeit erweist sich der SL-Ansatz auf Basis des zweimodalen $\hat{S}_Y^1$-Clusterverfahrens als extrem robust im Hinblick auf verzerrte Datensätze. Mit zunehmendem Verzerrungsgrad werden die Einträge in der Gewichtematrix W immer ähnlicher und die Prognosen fallen daher zunehmend konservativ aus.

In Bezug auf den Testdatensatz führt der steigende Verzerrungsgrad vor allen Dingen zu einer deutlichen Erhöhung der neutralen Bewertungen (3) auf Kosten

Gütemaß	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
AAD	0,771	0,797	0,857	0,934	0,984
Prec. (Schwellenwert:4,5)	0,606	0,473	0,523	0,337	0,217
Rec. (Schwellenwert:4,5)	0,065	0,059	0,084	0,102	0,110
Breese-Wert (R_B)	84,23	82,78	81,98	79,06	80,07

Tabelle B.10: AAD, Präzision, Recall und Breese-Wert hinsichtlich der Berechnung von Schätzern für unbekannte Items bei unterschiedlichen Graden der Verzerrung im Hinblick auf das hierarchische lineare Regressionsmodell (HBLR) unter Verwendung der Regressorenkombination M_8

Gütemaß	Grad der Verzerrung:				
	20 %	40 %	60 %	80 %	90 %
AAD	0,761	0,820	0,911	1,025	1,109
Prec. (Schwellenwert:4,5)	0,518	0,446	0,427	0,258	0,165
Rec. (Schwellenwert:4,5)	0,155	0,161	0,191	0,201	0,272
Breese-Wert (R_B)	82,70	81,86	81,46	80,14	80,52

Tabelle B.11: AAD, Präzision, Recall und Breese-Wert hinsichtlich der Berechnung von Schätzern für unbekannte Items bei unterschiedlichen Graden der Verzerrung im Hinblick auf das hybride hierarchische GP-Verfahren unter Verwendung der Regressorenkombination M_8

eines geringeren Anteils extremer Bewertungen (4 und 5). Daher können sich die *AAD*-Werte durch die mit steigendem Verzerrungsgrad immer konservativeren Prognosen leicht verbessern, obwohl der (geschätzte) Nutzen der Prognosen abnimmt. Es ist allerdings zu bedenken, daß Präzision und Recall trotzdem sehr niedrig ausfallen. In den extrem wenigen Fällen, in denen der $\hat{S}_Y^1$-Schätzer den Schwellenwert überschreitet, handelt es sich meist nicht um eine Höchstbewertung. Daran erkennt man, daß diesem Verfahren mit steigendem Verzerrungsgrad die Identifikation der (immer weniger werdenden) Höchstbewertungen zunehmend schwerer fällt. Hinsichtlich der Variablenkombination M_8 erhält man ähnliche Ergebnisse (Tabelle B.9).

Die Ergebnisse des HBLR-Verfahrens (Tabelle B.10) und die Resultate der

GP-Methode (Tabelle B.11) bezüglich der Variablenkombination M_8 sind im Hinblick auf *AAD*- und Breese-Wert denen des *SL*-Ansatzes zur Approximation der Item-Cluster Zugehörigkeit in Verbindung mit dem zweimodalen $\hat{S}_Y^1$-Clusterverfahren unterlegen. Dafür fallen bei hohen Verzerrungsgraden die Werte für Präzision und Recall immerhin besser aus als im Rahmen des $SL/\hat{S}_Y^1$-Ansatzes. Das HBLR-Verfahren, die hybride GP-Methode und die $SL/\hat{S}_Y^1$-Heuristik führen bei allen Verzerrungsgraden zu besseren Ergebnissen als das Basis-Modell (MW).

Anhang C

Versuchsanlage

Der balancierte unvollständige Blockplan ($a = 14, n_a = 4, b = 91, r = 26, \lambda_V = 6$) ist in der folgenden Abbildung dargestellt:

Block	Behandlungen				Block	Behandlungen			
1	1	2	8	9	47	5	9	13	14
2	1	3	8	10	48	6	8	10	14
3	1	4	8	11	49	7	8	9	11
4	1	5	8	12	50	1	2	4	14
5	1	6	8	13	51	1	3	7	13
6	1	7	8	14	52	1	5	6	11
7	2	3	9	10	53	1	9	10	12
8	2	4	9	11	54	2	3	5	8
9	2	5	9	12	55	2	6	7	12
10	2	6	9	13	56	2	10	11	13
11	2	7	9	14	57	3	4	6	9
12	3	4	10	11	58	3	11	12	14
13	3	5	10	12	59	4	5	7	10
14	3	6	10	13	60	4	8	12	13
15	3	7	10	14	61	5	9	13	14
16	4	5	11	12	62	6	8	10	14
17	4	6	11	13	63	7	8	9	11
18	4	7	11	14	64	1	2	4	14
19	5	6	12	13	65	1	3	7	13
20	5	7	12	14	66	1	5	6	11

Block	Behandlungen				Block	Behandlungen			
21	6	7	14	14	67	1	9	10	12
22	1	2	4	14	68	2	3	5	8
23	1	3	7	13	69	2	6	7	12
24	1	5	6	11	70	2	10	11	13
25	1	9	10	12	71	3	4	6	8
26	2	3	5	8	72	3	11	12	14
27	2	6	7	12	73	4	5	7	10
28	2	10	11	13	74	4	8	12	13
29	3	4	6	9	75	5	9	13	14
30	3	11	12	14	76	6	8	10	14
31	4	5	7	10	77	7	8	9	11
32	4	8	12	13	78	1	2	4	14
33	5	9	13	14	79	1	3	7	13
34	6	8	10	14	80	1	5	6	11
35	7	8	9	11	81	1	9	10	12
36	1	2	4	14	82	2	3	5	8
37	1	3	7	13	83	2	6	7	12
38	1	5	6	11	84	2	10	11	13
39	1	9	10	12	85	3	4	6	9
40	2	3	5	8	86	3	11	12	14
41	2	6	7	12	87	4	5	7	10
42	2	10	11	13	88	4	8	12	13
43	3	4	6	9	89	5	9	13	14
44	3	11	12	14	90	6	8	10	14
45	4	5	7	10	91	7	8	9	11
46	4	8	12	13					

Anhang D

Konjugiertes Gradientenverfahren

Ziel dieses Teils des Anhangs ist, die Wirkungsweise des in den Abschnitten 5.4 und 5.7 benutzten Konjugierten Gradientenverfahrens kurz darzustellen.
Sei $\gamma = (\gamma_{11}, ..., \gamma_{1C-1}, ..., \gamma_{I1}, ..., \gamma_{IC-1})'$ und

$$\frac{\partial \mathcal{Z}_O^x}{\partial \gamma'}(\gamma) = \begin{pmatrix} \sum_{j \in J_1} \chi_{1j}^1 h'(\chi_{1j}^1(\gamma_{11} - \hat{S}_{Y,1j}^x)) \\ \vdots \\ \sum_{j \in J_1} \chi_{1j}^{C-1} h'(\chi_{1j}^{C-1}(\gamma_{1C-1} - \hat{S}_{Y,1j}^x)) \\ \vdots \\ \sum_{j \in J_I} \chi_{Ij}^1 h'(\chi_{Ij}^1(\gamma_{I1} - \hat{S}_{Y,Ij}^x)) \\ \vdots \\ \sum_{j \in J_I} \chi_{Ij}^{C-1} h'(\chi_{Ij}^{C-1}(\gamma_{IC-1} - \hat{S}_{Y,Ij}^x)) \end{pmatrix}.$$

Für den zweckmäßig zu wählenden Startwert für γ, γ^0, ergeben sich die weiteren Startwerte

$$\mathbf{d}_0 = \mathbf{r}_0 = -\frac{\partial Z_O^x}{\partial \gamma'}(\gamma^0).$$

Aufbauend auf diesen Startwerten ergibt sich der umseitig dargestellte Algorithmus des (nichtlinearen) Konjugierten Gradientenverfahrens.

1. Wähle Startwerte $n = 0$, γ^0, $\mathbf{r}_0$ und $\mathbf{d}_0$.
2. Solange $n < n_{MAX}$ ist:
   a) Bestimme $\alpha_n = \arg\min_{\tilde{\alpha}_n} \{Z_O^x(\gamma^n + \tilde{\alpha}_n \mathbf{d}_n)\}$ mittels Sekantenverfahren.
   b) $\gamma^{n+1} = \gamma^n + \alpha_n \mathbf{d}_n$
   c) $\mathbf{r}_{n+1} = -\frac{\partial Z_O^x}{\partial \gamma'}(\gamma^{n+1})$
   d) $\tilde{\mathfrak{c}}_{n+1} = \max\left\{\frac{\mathbf{r}'_{n+1}(\mathbf{r}_{n+1}-\mathbf{r}_n)}{\mathbf{r}'_n\mathbf{r}_n}, 0\right\}$ (Polak-Ribière'sche Modifikation)
   e) Falls $\frac{\mathbf{r}'_{n+1}\mathbf{r}_n}{\mathbf{r}'_n\mathbf{r}_n} \geq \epsilon_{res}$ gilt, setzte $\tilde{\mathfrak{c}}_{n+1} = 0$. (Neustart)
   f) $\mathbf{d}_{n+1} = \mathbf{r}_{n+1} + \tilde{\mathfrak{c}}_{n+1}\mathbf{d}_n$
   g) $n \longleftarrow n + 1$

Abbildung D.1: Algorithmus des Konjugierten Gradientenverfahrens

Dem Sekantenverfahren in Schritt 2a) liegt in seiner m-ten Iteration die Taylor-Reihe zweiter Ordnung der Funktion

$$\frac{d}{d\tilde{\alpha}_n}\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha}_n^{m+1}\mathbf{d}_n) = \frac{d}{d\tilde{\alpha}_n}\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha_n}\mathbf{d}_n)\Big|_{\tilde{\alpha}_n=\tilde{\alpha}_n^{m+1}}$$

um den Entwicklungspunkt $\gamma^n + \tilde{\alpha}_n^m \mathbf{d}_n$ zu Grunde. Für die zweite Ableitung wird die Approximation

$$\begin{aligned}\frac{d^2}{d\tilde{\alpha}_n^2}\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha}_n^m\mathbf{d}_n) &\approx \frac{\frac{d}{d\tilde{\alpha}_n}\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha}_n^m\mathbf{d}_n) - \frac{d}{d\tilde{\alpha}_n}\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha}_n^{m-1}\mathbf{d}_n)}{\tilde{\alpha}_n^m - \tilde{\alpha}_n^{m-1}} \\ &= \frac{(\nabla\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha}_n^m\mathbf{d}_n))'\,\mathbf{d}_n - \left(\nabla\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha}_n^{m-1}\mathbf{d}_n)\right)'\mathbf{d}_n}{\tilde{\alpha}_n^m - \tilde{\alpha}_n^{m-1}}\end{aligned}$$

verwendet. Setzt man diese Näherung in die Taylor-Reihe zweiter Ordnung ein, erhält man in der m-ten Iteration des Sekantenverfahrens

$$\tilde{\alpha}_n^{m+1} = \tilde{\alpha}_n^m - (\tilde{\alpha}_n^m - \tilde{\alpha}_n^{m-1})\frac{(\nabla\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha}_n^{m-1}\mathbf{d}_n))'\,\mathbf{d}_n}{(\nabla\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha}_n^m\mathbf{d}_n))'\,\mathbf{d}_n - (\nabla\mathcal{Z}_O^x(\gamma^n + \tilde{\alpha}_n^{m-1}\mathbf{d}_n))'\,\mathbf{d}_n}.$$

Die Startwert $\tilde{\alpha}_n^0$ und $\tilde{\alpha}_n^1$ sollten so gewählt werden, daß das Intervall $[\tilde{\alpha}_n^1, \tilde{\alpha}_n^0]$ den hinsichtlich der Optimierung relevanten Wertebereich von $\tilde{\alpha}_n$ überdeckt. Das Sekantenverfahren wird solange ausgeführt, bis die Bedingung $|\tilde{\alpha}_n^{m+1} - \tilde{\alpha}_n^m| > \epsilon$ für ein beliebig aber fest zu wählendes $\epsilon > 0$ nicht mehr erfüllt ist.
In Schritt 2d) hätte anstelle der angegebenen Modifikation durch Polak und Ribière (Polak, Ribière (1969)) auch die Fletcher-Reeves Formel (Fletcher, Reeves (1964))

$$\tilde{\mathfrak{c}}_{n+1} = \frac{\mathbf{r}'_{n+1}\mathbf{r}_{n+1}}{\mathbf{r}'_n\mathbf{r}_n}$$

verwendet werden können. Sofern es sich bei der Funktion h um eine konvexe quadratische Form handelt und α_n exakt ermittelt wird, sind beide Formeln wegen der Orthogonalität der Gradienten $\mathbf{r}_n$ und $\mathbf{r}_{n+1}$ äquivalent. Letztere Voraussetzung ist hier nicht gegeben. In diesem Fall ist die Variante nach Polak und Ribière vorzuziehen, da diese sich in der Praxis als robuster und effizienter erwiesen hat.

Handelt es sich bei der Funktion h um eine konvexe quadratische Form, so liegt ein (lineares) Konjugiertes Gradientenverfahren vor. Letzteres konvergiert in maximal I(C-1) Schritten gegen ein globales Minimum.

Sofern es sich bei der Funktion h nicht um eine konvexe quadratische Form handelt, muß das Konjugierte Gradientenverfahren nicht in maximal I(C-1) Schritten konvergieren. In diesem Fall spricht man von Nichtlinearen Konjugierten Gradientenverfahren. Durch das Neustarten erhofft man sich, in einen Bereich zu gelangen, in dem die Funktion h ihr Minimum aufweist. Hinter der verwendeten Heuristik steht die Vermutung, daß die Funktion h umso besser durch eine konvexe quadratische Form angenähert werden kann, je näher man dem Minimum kommt. In diesem Fall erhofft man sich eine schnellere Konvergenz des Nichtlinearen Konjugierten Gradientenverfahrens. Man kann zeigen, daß $\mathbf{r}'_{n+1}\mathbf{r}_n = 0$ gilt, falls h eine konvexe quadratische Form ist (siehe z.B. Nocedal, Wright (2006)). Daher kann man vermuten, daß

$$\frac{\mathbf{r}'_{n+1}\mathbf{r}_n}{\mathbf{r}'_n\mathbf{r}_n}$$

umso größer ausfallen wird, je weiter der aktuelle Wert von h vom Minimum entfernt ist. Falls die obige Größe einen vorher festgelegten Schwellenwert ϵ_{res}

überschreitet, versucht man daher durch den Neustart in einen anderen Bereich der Funktion h zu gelangen.

Jedes nichtlineare Konjugierte Gradientenverfahren mit Neustart-Stufe (Schritt 2e) des dargestellten Algorithmus) konvergiert gegen einen stationären Punkt.

Beim (linearen) Konjugierten Gradientenverfahren kann Stufe 2e) entfallen, da für konvexe quadratische Formen h gilt $\mathbf{r}_{n+1} \cdot \mathbf{r}_{\tilde{u}} = 0, \tilde{u} = 1, ..., n,$.

Anhang E

Konvergenzdiagnostik für Markovketten

Ein Konvergenzdiagnostikverfahren ist eine Methode, die dazu dient, zu überprüfen, ob nach einem bestimmten Glied einer homogenen Markovkette alle weiteren Glieder derselben näherungsweise als Glieder der stationären Verteilung der Markovkette aufgefaßt werden können. Hierdurch gelingt eine Abschätzung der Länge n_{BURN} der sogenannten Einbrennphase.

Das bekannteste und gebräuchlichste konvergenzdiagnostische Verfahren ist ein Ansatz, der auf Gelman, Rubin (1992) zurückgeht. Dieses Verfahren ist sehr intuitiv, hat aber dafür den Nachteil, daß es entweder nur auf skalare Parameter θ oder nur auf die einzelnen Komponenten eines vektoriellen Parameters θ angewandt werden kann. Es werden $NC \geq 2$ Markovketten $\{\theta_n(n_C)\}_{n=1}^{2n_{V,B}}, n_C \in \{1, ..., NC\}$ mit möglichst unterschiedlichen Startwerten nach der Verfahrensvorschrift (Metropolis-Hastings Algorithmus bzw. Gibbs Sampling angewandt auf eine bestimmte Posterior auf Basis der gegebenen Daten) erzeugt, hinsichtlich der n_{BURN} zu bestimmen ist.

Sofern $\theta_n(n_C)$ ein Vektor mit $dim(\theta_n(n_C))$ Komponenten ist, bezeichnet $\tilde{\zeta}_{nn_C}^{o_\zeta}$ die o_ζ-te Komponente von θ, $o_\zeta \in \{1, ..., dim(\theta_n(n_C))\}$. Für eindimensionale (skalare) θ gilt $\tilde{\zeta}_{nn_C}^{o_\zeta} = \tilde{\zeta}_{nn_C}^{1} = \theta_n$.
Mittels der Definitionen

$$\overline{\tilde{\zeta}_{.n_C}^{o_\zeta}} = \frac{1}{n_{V,B}} \sum_{n=n_{V,B}+1}^{2n_{V,B}} \tilde{\zeta}_{nn_C}^{o_\zeta} \quad \text{und} \quad \overline{\tilde{\zeta}_{..}^{o_\zeta}} = \frac{1}{NC} \sum_{n_C=1}^{NC} \overline{\tilde{\zeta}_{.n_C}^{o_\zeta}}$$

läßt sich die Summe der quadratischen Abweichungen in die Summe der quadratischen Abweichungen innerhalb der Ketten und die Summe der quadratischen Abweichungen zwischen den Ketten aufspalten:

$$Q_G = \sum_{n_C=1}^{NC} \sum_{n=n_{V,B}+1}^{2n_{V,B}} (\tilde{\zeta}_{nn_C}^{o_\zeta} - \overline{\tilde{\zeta}_{..}^{o_\zeta}})^2 = \underbrace{\sum_{n_C=1}^{NC} \sum_{n=n_{V,B}+1}^{2n_{V,B}} (\tilde{\zeta}_{nn_C}^{o_\zeta} - \overline{\tilde{\zeta}_{.n_C}^{o_\zeta}})^2}_{= Q_{innerhalb}}$$
$$\underbrace{+ n_{V,B} \sum_{n_C=1}^{NC} (\overline{\tilde{\zeta}_{.n_C}^{o_\zeta}} - \overline{\tilde{\zeta}_{..}^{o_\zeta}})^2}_{= Q_{zwischen}} = Q_{innerhalb} + Q_{zwischen}.$$

Auf $Q_{innerhalb}$ entfallen $NC(n_{V,B}-1)$ Freiheitsgrade, die restlichen $NC-1$ Freiheitsgrade gehören zu $Q_{zwischen}$. Durch die Division der beiden Summen der quadratischen Abweichungen durch ihre jeweiligen Freiheitsgrade erhält man - sofern die Kette nach $n_{V,B}$ Gliedern stationär ist - unverzerrte Schätzer der Varianz der stationären Verteilung (Posterior). Es ergibt sich

$$B(o_\zeta, n_{V,B}) = \frac{Q_{zwischen}}{NC-1} = \frac{n_{V,B}}{NC-1} \sum_{n_C=1}^{NC} (\overline{\tilde{\zeta}_{.n_C}^{o_\zeta}} - \overline{\tilde{\zeta}_{..}^{o_\zeta}})^2.$$

Mit

$$\hat{s}_{n_C}^2(o_\zeta, n_{V,B}) = \frac{1}{n_{V,B}-1} \sum_{n=n_{V,B}+1}^{2n_{V,B}} (\tilde{\zeta}_{nn_C}^{o_\zeta} - \overline{\tilde{\zeta}_{.n_C}^{o_\zeta}})^2$$

erhält man

$$W(o_\zeta, n_{V,B}) = \frac{Q_{innerhalb}}{NC(n_{V,B}-1)} = \frac{1}{NC} \sum_{n_C=1}^{NC} \hat{s}_{n_C}^2(o_\zeta, n_{V,B}).$$

Daher ist neben $W(o_\zeta, n_{V,B})$ und $B(o_\zeta, n_{V,B})$ das gewichtete Mittel

$$\hat{V}(o_\zeta, n_{V,B}) = \left(1 - \frac{1}{n_{V,B}}\right) W(o_\zeta, n_{V,B}) + \frac{1}{n_{V,B}} B(o_\zeta, n_{V,B}).$$

ein unverzerrter Schätzer für die Varianz der stationären Verteilung. Sofern nach den ersten $n_{V,B}$ Gliedern der Markovkette keine Abhängigkeit von der ursprünglichen Startwerten beobachtet werden kann, dürfen alle weiteren Glieder der Markovkette(n) als Ziehungen aus der stationären Verteilung (also der Posterior) betrachtet werden. In der Literatur wird in diesem Zusammenhang häufig von Konvergenz gesprochen. (Hierin findet der Terminus Konvergenzdiagnostik seinen Ursprung.)

Falls die einzelnen Markovketten noch unter dem Einfluß ihrer Trajektorien stehen, nähern die Markovketten nach $n_{V,B}$ Gliedern noch nicht ihre stationäre Verteilung (also die Posterior) an. In diesem Fall wird vor allem $B(o_\zeta, n_{V,B})$ erheblich größer ausfallen als die Varianz der Posterior. Dafür kann die Stichprobenvarianz zwischen den Ketten $W(o_\zeta, n_{V,B})$ bei einer kleinen Anzahl von Gliedern $n_{V,B}$ kleiner als die Varianz der Posterior ausfallen, weil alle oder zumindest ein erheblicher Teil der Ketten bislang erhebliche Teile ihres Wertebereichs noch nicht abdecken konnten. Folglich ist $\sqrt{\hat{R}(o_\zeta, n_{V,B})} = \sqrt{\frac{\hat{V}(o_\zeta, n_{V,B})}{W(o_\zeta, n_{V,B})}}$ größer als 1, falls Ketten nach $n_{V,B}$ Gliedern noch nicht unabhängig von ihren Startwerten geworden sind und hierdurch ihre stationären Verteilungen erreicht haben. Andernfalls sind sowohl $\hat{V}(o_\zeta, n_{V,B})$ als auch $W(o_\zeta, n_{V,B})$ konsistente Schätzer für die Varianz. Es sollte für den Faktor $\sqrt{\hat{R}(o_\zeta, n_{V,B})} = \sqrt{\frac{\hat{V}(o_\zeta, n_{V,B})}{W(o_\zeta, n_{V,B})}}$ gelten

$$\lim_{n_{V,B} \longrightarrow \infty} \sqrt{\hat{R}(o_\zeta, n_{V,B})} = 1.$$

Gelman (1996) schlägt vor, $n_{V,B}$ als Näherung für n_{BURN} zu verwenden, wenn $\sqrt{\hat{R}(o_\zeta, n_{V,B})} < 1,2$ gilt.

Ein Nachteil dieses in der Praxis sehr beliebten Verfahrens ist, daß es bei mehrdimensionalen Parametern θ nur für die einzelnen Komponenten ζ_ζ^o getrennt durchgeführt werden kann. Außerdem ist die ursprünglich von Gelman, Rubin vorgeschlagene Größe für $\sqrt{\hat{R}(o_\zeta, n_{V,B})}$ erheblich komplizierter als die angegebene Formel. Brooks, Gelman (1998) ist es gelungen, für alle Komponenten eines vektoriellen θ eine gemeinsame Obergrenze aller ursprünglich von Gelman und Rubin vorgeschlagenen komplizierteren Versionen von $\sqrt{\hat{R}(o_\zeta, n_{V,B})}$ zu berechnen.

Cowles, Carlin (1996) stellen fest, daß dieses Verfahren die Verwendung unterschiedlicher Startwerte erfordert, von denen einige von der stationären Verteilung

hinreichend weit aber dennoch keiner von ihr zu weit entfernt sein sollte. Dies setzt aber bereits Kenntnisse über die stationäre Verteilung voraus. Dem ist entgegenzuhalten, daß in vielen Anwendung zumindest die ungefähre Größenordnung und ein grober Wertebereich der Größe θ, deren Verteilung simuliert werden soll, bekannt ist.

Als problematischer ist anzusehen, daß das vorgestellte Verfahren implizit voraussetzt, daß die gesuchte Posterior durch die Normalverteilung approximierbar ist. Daher scheint es in einer Vielzahl von Fällen, in denen dringend die Konvergenz überprüft werden sollte, nicht anwendbar zu sein.

Das Verfahren nach Geweke (1992) verwendet nur die Glieder $\tilde{\zeta}_{n1}$ einer einzigen Markovkette. (Hier ist daher $n_C = 1$.) Dieses Verfahren basiert auf den Definitionen

$$\overline{\tilde{\zeta}_{\cdot}^{o_\zeta}(n_1, n_2, n_{V,B})} = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_{V,B}+n_1}^{n_{V,B}+n_2} \tilde{\zeta}_{n1}^{o_\zeta}$$

und

$$\hat{s}^2(o_\zeta, n_1, n_2, n_{V,B}) = \frac{1}{n_2 - n_1} \sum_{n=n_{V,B}+n_1}^{n_{V,B}+n_2} \left(\tilde{\zeta}_{n1}^{o_\zeta} - \overline{\tilde{\zeta}_{\cdot}^{o_\zeta}(n_1, n_2, n_{V,B})} \right)^2 .$$

Unter der Voraussetzung, daß die Markovkette nach $n_{V,B}$ Gliedern ihre stationäre Verteilung erreicht hat, $n_a + n_b < n_{MAX}$ gilt und die Verhältnisse n_a/n_{MAX} und n_b/n_{MAX} konstant sind konvergiert

$$\mathcal{Z}_G^{o_\zeta}(n_{V,B}, n_a, n_b) = \frac{\overline{\tilde{\zeta}_{\cdot}^{o_\zeta}(1, n_b, n_{V,B})} - \overline{\tilde{\zeta}_{\cdot}^{o_\zeta}(n_{MAX} - n_a + 1, n_{MAX}, n_{V,B})}}{\sqrt{\frac{\hat{s}^2(o_\zeta, 1, n_b, n_{V,B})}{n_b} + \frac{\hat{s}^2(o_\zeta, n_{MAX} - n_a + 1, n_{MAX}, n_{V,B})}{n_a}}}$$

im Limes $n_{MAX} \longrightarrow \infty$ gegen die Standardnormalverteilung. Geweke (1992) empfiehlt $n_a = 0,5 \cdot n_{MAX}$ und $n_b = 0,1 \cdot n_{MAX}$. Notwendig aber keinesfalls hinreichend für die Stationarität der betrachteten Markovkette ab dem $n_{V,B}$-ten Glied ist, daß die Größe $\mathcal{Z}_G^{o_\zeta}(n_{V,B}, n_a, n_b)$ kleine Werte annimmt. Beispielsweise ist es möglich, daß $\mathcal{Z}_G^{o_\zeta}(n_{V,B}, n_a, n_b)$ für eine stark autokorrelierte Markovkette

klein ausfällt, obwohl diese Kette nach dem $n_{V,B}$-ten Glied noch nicht stationär ist. Somit läßt sich durch die Geweke-Konvergenzdiagnostik nur in einigen Fällen nachweisen, daß die Konvergenz der betrachteten Kette nach dem $n_{V,B}$-ten Glied noch nicht eingetreten ist. Gelingt ein solcher Nachweis nicht, kann daraus nicht auf die Stationarität der Markovkette $\{\tilde{\zeta}_{n1}^{o_\zeta}\}_{n=n_{V,B}}^{\infty}$ geschlossen werden. Vielmehr ist ein Mißlingen dieses Nachweises lediglich als Indikator für als oder Hinweis auf dieselbe zu interpretieren.

In der dargestellten Form ist auch die Geweke-Konvergenzdiagnostik ein univariates Verfahren. In der Literatur wird vorgeschlagen, anstelle der Mittelwerte zweier Teile der Kette die entsprechenden Mittelwerte des (-2)-fachen der Posterior $\mathcal{P}(\theta)$ miteinander zu vergleichen (Cowles, Carlin (1996), Gamerman (1997)). Mit

$$\overline{LP(\theta, n_1, n_2, n_{V,B})} = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_{V,B}+n_1}^{n_{V,B}+n_2} -2 \log \mathcal{P}(\theta_n)$$

und

$$\hat{s}_{\mathcal{P}}^2(\theta, n_1, n_2, n_{V,B}) = \frac{1}{n_2 - n_1} \sum_{n=n_{V,B}+n_1}^{n_{V,B}+n_2} \left(-2 \log \mathcal{P}(\theta_n) - \overline{LP(\theta, n_1, n_2, n_{V,B})} \right)^2.$$

erhält man die multivariate Größe

$$\mathcal{Z}_G^{\mathcal{P}}(n_{V,B}, n_a, n_b) = \frac{\overline{LP(\theta, 1, n_b, n_{V,B})} - \overline{LP(\theta, n_{MAX} - n_a + 1, n_{MAX}, n_{V,B})}}{\sqrt{\frac{\hat{s}_{\mathcal{P}}^2(\theta, 1, n_b, n_{V,B})}{n_b} + \frac{\hat{s}_{\mathcal{P}}^2(\theta, n_{MAX} - n_a + 1, n_{MAX}, n_{V,B})}{n_a}}}.$$

In dieser Version eignet sich das Verfahren nach Geweke (1992) auch für vektorielle θ. Auch in dieser Variante kann die Geweke-Konvergenzdiagnostik immer nur eine conditio sine qua non in Bezug auf die Stationarität der Markovkette $\theta_{n_{V,B}}, \theta_{n_{V,B}+1}, ...$ sein.

Eine Reihe weiterer konvergenzdiagnostischer Verfahren eignen sich nur für Markovketten, die mittels Gibbs Sampling erzeugt wurden (z.B. Roberts (1992), Ritter, Tanner (1992), Zellner, Min (1995), Liu et. al. (1992)).

Auf Basis ihres umfassenden Vergleichs der Konvergenzdiagnoseverfahren kommen Cowles, Carlin (1996) zu dem Schluß, daß keines der bekannten Verfahren zum direkten Nachweis der Stationarität (und damit zur Bestimmung von n_{BURN}) geeignet ist. Die Autoren empfehlen die zusätzliche Beschaffung von Information hinsichtlich der Posterior vor Beginn der Konvergenzdiagnose und die komponentenweise graphische Darstellung der betrachtete(n) Kette(n). Zudem kommen Cowles, Carlin (1996) zu dem Schluß, daß zu jeder Konvergenzdiagnose sowohl der Vergleich mehrerer Ketten mit unterschiedlichen Startwerten (Verfahren nach Gelman, Rubin (1992)) als auch der Vergleich unterschiedlicher Teile einer sehr langen Kette (Geweke-Konvergenzdiagnostik (1992)) herangezogen werden sollten. Auch Gamerman (1997) legt nahe, immer mehrere Verfahren zu verwenden. Offenbar besteht in Bezug auf die näherungsweise Bestimmung von n_{BURN} weiterer Forschungsbedarf.

Anhang F

Der LogitBoost-Algorithmus

Der LogitBoost-Algorithmus nach Friedman et. al. (2000) ist ein Algorithmus zur Bestimmung der Parameter eines Multinomialen Logit-Modells. In der vorliegenden Arbeit wird dieser Algorithmus im Rahmen der Logistischen Modellbäume (Landwehr et. al. (2005)) benutzt, um an jedem Knoten $\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_n})$ des mittels des C4.5 Algorithmus generierten Entscheidungsbaums Multinomiale Logit Modell für die Daten $J[\mathfrak{X}(\mu_1, ..., \mu_{\underline{n}}^{s_n})]$ zu berechnen. Ein Item j mit dem Eigenschaftsvektor a_j gehört im Rahmen des Multinomialen Logit-Modells mit der Wahrscheinlichkeit

$$\bar{p}(\mathfrak{G}_{jl}) = \frac{\exp(\mathfrak{G}_{jl})}{\sum_{l'=1}^{L} \exp(\mathfrak{G}_{jl'})}$$

ins Cluster $l \in \{1, ..., L\}$. Zur Vereinfachung der Notation wird anstelle von $\mathfrak{G}_{jl}^{J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})]}$ der Ausdruck $\mathfrak{G}_{jl}$ verwendet. Die Werte $\mathfrak{G}_{jl}$ sind latente Größen, die den Realisationen $q_{jl}, j = 1, ..., L, l = 1, ..., L$, zugrundeliegen. Die iterativ auf Basis der jeweiligen Eigenschaften a_j zu bestimmenden Schätzer von $\mathfrak{G}_{jl}, l \in \{1, ..., L\}$, seien $\hat{\mathfrak{G}}_{jl}$, wobei $l \in \{1, ..., L\}$ gilt.
Friedman et. al. (2000) verwenden die Größe

$$z_{jl} = \frac{q_{jl} - \bar{p}(\hat{\mathfrak{G}}_{jl})}{\bar{p}(\hat{\mathfrak{G}}_{jl})(1 - \bar{p}(\hat{\mathfrak{G}}_{jl}))}$$

als Approximation für den Fehler. Für den Schätzer der latenten Größe wird $\hat{\mathfrak{G}}_{jl} = \hat{\mathfrak{G}}_{l,0} + \hat{\mathfrak{G}}_l' a_j$, mit $\hat{\mathfrak{G}}_l = (\hat{\mathfrak{G}}_{l,1}, ..., \hat{\mathfrak{G}}_{l,\kappa_A})'$, verwendet.

1. Startwerte: $\bar{p}(\hat{\mathfrak{G}}_{jl}^0) = \frac{1}{L}, \forall j \in J^{cal}[\mathfrak{X}(\mu_1)], \hat{\mathfrak{G}}_{l,\kappa}^0 = 0, \forall l \in \{1,...,L\}, \forall \kappa \in \{0,...,\kappa_A\}$. Falls $\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n}) \neq \mathfrak{X}(\mu_1)$ gilt, verwendet man als Startwerte die Resultate des Eltern-Knoten: $\bar{p}(\hat{\mathfrak{G}}_{jl}^0) = \bar{p}(\hat{\mathfrak{G}}_{jl}), \forall j \in J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})], \hat{\mathfrak{G}}_{l,\kappa}^0 = \hat{\mathfrak{G}}_{l,\kappa}, \forall l \in \{1,...,L\}$, $\forall \kappa \in \{0,...,\kappa_A\}$.
2. Für $\xi = 1,...,\xi_{MAX}$:
   a) (i) Berechne
   $$z_{jl}^{\xi} = \frac{q_{jl} - \bar{p}(\hat{\mathfrak{G}}_{jl}^{\xi-1})}{\bar{p}(\hat{\mathfrak{G}}_{jl}^{\xi-1})(1-\bar{p}(\hat{\mathfrak{G}}_{jl}^{\xi-1}))}, \quad \omega_L^{\xi}(j,l) = \bar{p}(\hat{\mathfrak{G}}_{jl}^{\xi-1})(1-\bar{p}(\hat{\mathfrak{G}}_{jl}^{\xi-1})),$$
   für alle $j \in J[\mathfrak{X}_{\mu_1 \cdots \mu_{\underline{n}}^s}]$ und alle $l \in \{1,...,L\}$.
   (ii) Bestimme WLS-Schätzer $\hat{\gamma}_0^{\xi l\kappa}, \hat{\gamma}_1^{\xi l\kappa}$ durch WLS-Regression mit $z_{1l}^{\xi},...,z_{|J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})]|l}^{\xi}$ als Realisationen der endogenen Variable und $a_{1\kappa},...,a_{|J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})]|\kappa}$ als Realisationen der einzigen exogenen Variable und Gewichten $w_L^{\xi}(j,l), \forall\, \kappa \in \{1,...,\kappa_{MAX}\}, \forall\, l \in \{1,...,L\}$.
   (iii) Wähle für jedes $l \in \{1,...,L\}$ die Eigenschaft $\kappa_l^* \in \{1,...,\kappa_A\}$ die zur besten Anpassung der Daten führt.
   b) Setze $\hat{\mathfrak{G}}_{l,0}^{\xi} = \hat{\mathfrak{G}}_{l,0}^{\xi-1} + \frac{L-1}{L}(\hat{\gamma}_0^{\xi l\kappa_l^*} - \frac{1}{L}\sum_{l'=1}^{L}\hat{\gamma}_0^{\xi l'\kappa_l^*}), l = 1,...,L,$
   $\hat{\mathfrak{G}}_{l,\kappa_l^*}^{\xi} = \hat{\mathfrak{G}}_{l,\kappa_l^*}^{\xi-1} + \frac{L-1}{L}(\hat{\gamma}_{\kappa_l^*}^{\xi l\kappa_l^*} - \frac{1}{L}\sum_{l'=1}^{L}\hat{\gamma}_{\kappa_l^*}^{\xi l'\kappa_l^*}), l = 1,...,L.$
   c) Berechne $\bar{p}(\hat{\mathfrak{G}}_{jl}^{\xi}) = \dfrac{\exp(\hat{\mathfrak{G}}_{jl}^{\xi})}{\sum_{l'=1}^{L}\exp(\hat{\mathfrak{G}}_{jl'}^{\xi})}$ mit $\hat{\mathfrak{G}}_{jl}^{\xi} = \hat{\mathfrak{G}}_{l,0}^{\xi} + \hat{\mathfrak{G}}_{l,\kappa_l^*}^{\xi} a_{j\kappa_l^*}$.

Abbildung F.1: LogitBoost Variante für Logistische Modellbäume

Hinsichtlich jedes Knotens $\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})$ werden in der jeweils ξ-ten Iteration die Vektoren $(z_{1l}^{\xi},...,z_{|J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})]|l}^{\xi})'$ und $(a_{1\kappa},...,a_{|J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})]|\kappa})'$ bezüglich der am Knoten $\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})$ vorhandenen Datenmenge $J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s})]$ gebildet. Jeder Vektor $(z_{1l}^{\xi},...,z_{|J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})]|l}^{\xi})', l = 1,...,L$, wird dann durch WLS-Regression an jeweils einen Vektor $(\underline{a}_{1\kappa},...,\underline{a}_{|J[\mathfrak{X}(\mu_1,...,\mu_{\underline{n}}^{s_n})]|\kappa})', \kappa = 1,...,\kappa_{MAX}$, angepaßt. Somit werden an jedem Knoten $L\kappa_{MAX}$ WLS-Regressionen ausgeführt. Hierbei werden die Gewichte $\omega_L(j,l) = \bar{p}(\hat{\mathfrak{G}}_{jl})(1-\bar{p}(\hat{\mathfrak{G}}_{jl}))$ verwendet.

Auf diese Weise werden alle Eigenschaften einzeln als exogene Variablen ausprobiert. Es ergeben sich die Schätzer $\hat{\gamma}_0^{\xi l\kappa}$ und $\hat{\gamma}_1^{\xi l\kappa}, \kappa = 1,...,\kappa_{MAX}, l = 1,...,L$. Für jedes Cluster $l \in \{1,...,L\}$ wählt man die Eigenschaft $\kappa_l^* \in \{1,...,\kappa_{MAX}\}$ aus, mit der sich die beste Anpassung an die betrachtete Datenmenge erhalten ließ und berechnet auf dieser Grundlage

$$\begin{array}{l}\hat{\mathfrak{G}}_{l,0}^{\xi} = \hat{\mathfrak{G}}_{l,0}^{\xi-1} + \frac{L-1}{L}(\hat{\gamma}_0^{\xi l \kappa_l^*} - \frac{1}{L}\sum_{l'=1}^{L}\hat{\gamma}_0^{\xi l' \kappa_l^*}), \\ \hat{\mathfrak{G}}_{l,\kappa_l^*}^{\xi} = \hat{\mathfrak{G}}_{l,\kappa_l^*}^{\xi-1} + \frac{L-1}{L}(\hat{\gamma}_{\kappa_l^*}^{\xi l \kappa_l^*} - \frac{1}{L}\sum_{l'=1}^{L}\hat{\gamma}_{\kappa_l^*}^{\xi l' \kappa_l^*})\end{array} \quad , l = 1, ..., L.$$

Damit berechnet man $\hat{\mathfrak{G}}_{jl}^{\xi} = \hat{\mathfrak{G}}_{l,0}^{\xi} + \hat{\mathfrak{G}}_{l,\kappa_l^*}^{\xi} a_{j\kappa_l^*}$ und $\bar{p}(\hat{\mathfrak{G}}_{jl}^{\xi}), l = 1, ..., L, j \in J^K$.

Für jeden Knoten wird dies ξ_{MAX}-mal ausgeführt. Hierdurch erreicht man, daß in Bezug auf die Knoten auf der ersten Stufe des Baumes maximal ξ_{MAX} Eigenschaften zur Bestimmung der Schätzer verwendet werden und hinsichtlich der Knoten auf der $\underline{n}$-ten Hierarchie-Ebene höchstens $\underline{n}\xi_{MAX}$ Eigenschaften zur Berechnung der Schätzer genutzt werden. Die zu einem Sohnknoten gehörenden Schätzer können aber müssen nicht auf Basis von mehr Eigenschaften ermittelt worden sein als die des Elternknoten.

Die optimale Anzahl für ξ_{MAX} wird durch fünffache Kreuzvalidierung bestimmt. Es reicht aus, den optimalen Wert für ξ_{MAX} an der Wurzel zu bestimmen und diesen Wert für alle Knoten zu verwenden (Landwehr et. al. (2005)).

Die im Rahmen der Logistischen Modellbäume verwendete Variante des LogitBoost Algorithmus ist in Abbildung F.1 dargestellt.

Index