Multiple Imputationsmodelle für *Knowledge Economy Indicators*

Theorie, Implementierung und Verbesserungsvorschläge

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Wirtschaftswissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

Vorgelegt von

Luis Augusto Huergo
geboren in La Plata, Argentinien

2010

Dekan: Prof. Dr. Kerstin Pull
Erstkorrektor: Prof. Dr. Dr. h. c. mult. Eberhard Schaich
Zweitkorrektor: Prof. Dr. Joachim Grammig
Tag der mündlichen Prüfung: 02. Juni 2010

Danksagung

Die vorliegende Arbeit wurde von der Wirtschaftswissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen im Juni 2010 als Dissertation angenommen. Sie fasst die wissenschaftlichen Erkenntnisse zusammen, die während meiner Zeit als Mitarbeiter des Projekts KEI (*Knowledge Economy Indicators*) der europäischen Kommission am Lehrstuhl für Statistik, Ökonometrie und Unternehmensforschung entstanden sind. An ihrem Gelingen besitzen viele Menschen Anteil, denen ich danken möchte.

Für die Ermöglichung einer Dissertation und die ständige Unterstützung im Arbeitsprozeß möchte ich mich in besonderem Maße bei meinem Doktorvater Prof. Dr. Dr. h. c. mult. Eberhard Schaich bedanken. Er hat mir geistige Freiräume gewährt und behielt stets ein offenes Ohr für Fragestellungen aller Art. Insgesamt hat er, sei es mit Rat oder Tat, mit unendlicher Geduld oder mit dem notwendigen Druck im entscheidenden Moment, maßgeblich zur Vollendung meiner Promotion beigetragen.

Mein Zweitgutachter Herr Prof. Dr. Grammig hat meine ersten Jahre an der Wirtschaftswissenschaftlichen Fakultät deutlich geprägt. Ihm verdanke ich meine ersten Schritte in der Statistik und insbesondere meine Ausbildung im Bereich der Ökonometrie, in dem ich seine methodischen Vorlieben und Ansichten als meine angenommen habe. Diese Kenntnisse haben sich als essenziell für die Methodenentwicklung im Rahmen des KEI-Projekts erwiesen.

Drei andere Professoren haben meine Arbeit entscheidend beeinflusst. In alphabetischer Reihenfolge: Prof. Dr. Michael Merz, Prof. Dr. Ralf Münnich und Prof. Dr. Martin Zerner.
Herr Prof. Dr. Merz in seiner doppelten Funktion als Kollege und Betreuer hat nahezu jeden Schritt in der Entwicklung meines mathematischen Verständnisses insbesondere in der letzten Phase meiner Dissertation mitgestaltet und zum Positiven beeinflusst. Mit wachem Geist und scharfem Blick hat er unermüdlich an meiner mathematischen Ausdrucksweise gefeilt und mich stets an seinem enormen Wissen teilhaben lassen.
Herr Prof. Dr. Münnich als Koordinator des KEI-Projekts hat mich bei meinen ersten ernsten wissenschaftlichen Vorhaben betreut und seither ständig begleitet. Er wurde zum Ratgeber und guten Freund.
Herr Prof. Dr. Zerner ist zweifelsohne dafür verantwortlich, dass aus einem ursprünglich geplanten Semester an der Mathematischen Fakultät sieben wurden. Seine offene und stets höfliche Art hat meine Anfangsängste als Nichtmathematiker gemildert und mir Lust auf mehr gemacht.

Meine Leidenschaft für die Mathematik hat sich in den Jahren meiner Promotion zum Leidwesen meiner Kollegen Jochen Heberle, Ramona Maier, Michael Merz und Dominik Ohly enorm gesteigert. Sie haben in unendlichen Diskussionen meine typische Mischung aus extremer Begeisterung

und mangelnder mathematischer Präzision über sich ergehen lassen. Dass sie dies geduldig ertragen und mich dazu ermuntert haben, weiter zu machen, hat meine Entwicklung der letzten Jahre ebenfalls ermöglicht. Herzlichen Dank dafür, Ihr Unsterblichen! Diese Rolle wird jetzt von meinen Kollegen an den Lehrstühlen von Prof. Dr. Grammig und Prof. Dr. Biewen, Stephan Jank, Franziska Peter, Andos Juhasz und Markus Niedergesäss übernommen. Hoffentlich können sie nicht erahnen, was auf sie zukommt.

Ohne Frau Eiting und Frau Bürger wäre ich wahrscheinlich arbeitsunfähig. Stellvertretend für jeden Arbeitstag möchte ich mich an dieser Stelle bei ihnen bedanken.

Zuletzt möchte ich meiner Frau Silja danken, die mir über den gesamten, von Höhen und Tiefen gezeichneten Promotionszeitraum hindurch immer zur Seite stand und oft mehr an mich glaubte als ich selbst, sowie meinen Eltern und meiner Tochter Pirjo-Sophie, die häufig, ohne es zu wissen, auf ihren Papa verzichten musste, damit ich meine Promotion abschließen konnte.

Marta, Silja und Pirjo, den drei wichtigsten Frauen in meinem Leben, ist diese Arbeit gewidmet.

Tübingen, im Juli 2010

Luis Huergo

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungs- und Symbolverzeichnis

Vorbemerkung: Bei der Verfassung dieser Arbeit wurde besonderer Wert darauf gelegt, die in der einschlägigen Fachliteratur übliche Notation zu verwenden. Da sich die Arbeit mit Ergebnissen aus verschiedenen Gebieten der Mathematik bzw. Statistik befasst und gleichzeitig deren typische Notation verwendet werden soll, lassen sich Abweichungen in der Notation der verschiedenen Kapitel nicht vermeiden. Zum Zweck der Übersichtlichkeit wurden daher die Symbole nach Kapitel sortiert und in der Erscheinungsreihenfolge im Text aufgelistet.

Abkürzungen

DA	*Data Augmentation*-Algorithmus
dim	Dimension eines Vektors
EM	*Expectation Maximization*-Algorithmus
i.i.d.	Identisch und unabhängig verteilt
max	Maximum
MCMC	Markov-Chain-Monte-Carlo
ML	Maximum Likelihood
MM	Mechanismus der fehlenden Daten (*Missingness mechanism*)
min	Minimum
MAR	*Missing at random*
MCAR	*Missing completely at random*
MNAR	*Missing not at random*
MI	Multiple Imputation
MP	Muster der fehlenden Daten (*Missingness pattern*)
O.B.d.A.	Ohne Beschränkung der Allgemeinheit
RSWP	*Reverse Sweep*-Operator
sup	Supremum
SWP	*Sweep*-Operator

Mathematische Symbole nach Kapitel

Kapitel 1

$Y, (y_{ij})$	Datenmatrix
Y_{obs}	Beobachtete Daten
Y_{mis}	Fehlende Daten
θ	Parametervektor der Verteilung der beobachteten Daten
$P(\cdot \mid \theta)$	Wahrscheinlichkeitsfunktion (allgemein) gegeben θ
$f(\cdot \mid \theta)$	Dichtefunktion gegeben θ
ξ	Parameter des Mechanismus der fehlenden Daten
R	Indikator Matrix
R^*	Sortierte Indikator Matrix

Kapitel 2

$x = (x_1, ..., x_n)$	Stichprobe vom Umfang n
Θ	Menge der zulässigen Parameterwerte θ
$\mathcal{L}_x(\theta); \mathcal{L}_x$	Likelihood-funktion für Parameter θ gegeben eine Stichprobe x
$L_x(\theta); L_x$	Log-Likelihood-Funktion für Parameter θ gegeben x
$\hat{\theta}$	Maximum-Likelihood Schätzer für den (möglicherweise vektorwertigen) Parameter θ
$T(x)$	Vektor suffizienter Statistiken der Stichprobe x
$E_{\theta_1}[\cdot]$	Erwartungswert gegeben Parameter θ_1
$H(\theta_1, \theta_2)$	H-Funktion gegeben zwei Parameter θ_1 und θ_2
$Q(\theta, \theta_i)$	Q-Funktion gegeben die i-te Schätzung für Parameter θ
θ_i	Schätzung für Parameter θ in der i-ten Iteration des EM-Algorithmus
$T_i(x)$	Vektor suffizienter Statistiken ausgewertet in x und θ_i

Kapitel 3

$\xrightarrow{\text{f.s.}}$	Fast sichere Konvergenz
$\xrightarrow{\text{d}}$	Konvergenz in Verteilung

$\xrightarrow{\text{p}}$	Stochastische Konvergenz
$\theta \in \mathbb{R}^k$	k-dimensionaler Parameter
$f(\theta)$	*a priori*-Verteilung von θ
$f(\theta\|\mathbf{y})$	*a posteriori*-Verteilung von θ
$\theta^{(1)}$	i-te Komponente von θ

Kapitel 4

μ, Σ	Vektor der Erwartungswerte und Varianz-Kovarianzmatrix einer normalverteilten Zufallsvariablen
$\mathcal{O}(s)$	Spalten mit beobachteten Daten im s-ten MP
$\mathcal{M}(s)$	Spalten mit unbeobachteten Daten im s-ten MP
$\mathcal{I}(s)$	Zeilen der Datenmatrix, welche das s-te MP aufweisen
A	*geswepte* θ-Matrix
$W^{-1}(m, \Lambda)$	normal-inverse-Wishart-Verteilung
τ, μ_0, m, Λ	Parameter der Wishart Verteilung

Kapitel 5

Ψ	Skalierungsmatrix (Varianz-Kovarianzmatrix)
μ	Vektor der Erwartungswerte
ν	Freiheitsgrade
q	Unobeobachtbare (latente) Variable
w	Vektor der Gewichte
d^2	Mahalanobis-Distanz
δ	Kontaminierungswahrscheinlichkeit
λ	Kontaminierungsparameter
y^*	Transformierte Variable
θ	Potenztransformationsparameter
z	Standardisierter Wert
$\overline{m}$	Vektor der Momentenbedingungen
I_k	Einheitsmatrix der Dimension k $(k \times k)$

Anhang A

■	Ende einer Definition
□	Ende eines Beweises

Einleitung

Die im März 2000 auf einem Sondergipfel der europäischen Staats- und Regierungschefs in Lissabon verabschiedete und im März 2005 durch den Europäischen Rat bestätigte **Lissabon Agenda**[1] hat sich zum Ziel gesetzt, einen Wendepunkt in der Unternehmens- und Innovationspolitik innerhalb der Europäischen Union zu signalisieren und „die E.U. zum wettbewerbsfähigsten und dynamischsten wissensgestützten Wirtschaftsraum der Welt zu machen“ (siehe Euractiv (2004)).

Schlüsselkonzepte der Lissabon Strategie sind der auf dem Werk von Joseph Schumpeter (1912) basierende Begriff der *Innovation* als Motor für Wirtschaftswachstum und die *Knowledge Economy* (siehe Rodrigues (2003)). Ein mit dem letztgenannten eng verwandter und im deutschsprachigen Raum verbreiteter Begriff ist die *Wissensgesellschaft* (siehe van Dülmen und Rauschenbach (2004)), welche eine Gesellschaftsform in hochentwickelten Ländern bezeichnet, in der sowohl das individuelle als auch das kollektive Wissen die Grundlage der sozialen und ökonomischen Entwicklung bilden (siehe Auer und Sturz (2007)).

Neben der Formulierung konkreter Ziele gehört die Machbarkeit der Erfolgsmessung zu den unabdingbaren strategischen Elementen eines jeden Vorhabens. Da die Wissensgesellschaft jedoch anstelle der Verwertung von Sachkapital von immateriellem Kapital ausgeht, welches nicht mit klassischen Methoden (Produkteinheiten pro Zeiteinheit) gemessen werden kann, müssen neuartige Messzahlen konzipiert und verwendet werden. Die Lissabon Strategie stellt somit die Europäische Kommission vor die Aufgabe, ihr Messinstrumentarium an die neue Zielsetzung anzupassen.

Die Umsetzung der Ziele der Lissabon Strategie in konkrete Maßnahmen erfolgt durch die so genannten *Framework Programmes for Research and Technological Development* der Europäischen Kommission.

Im Rahmen des *Sixth Framework Programme* wurden im September 2004 sechs europäische Forschungsinstitutionen: die Eberhard-Karls-Universität Tübingen, die Universität Trier, das Joint Research Center der Europäischen Kommission in Ispra/Italien, die Katholische Universität Leuven/Belgien, die Universität Maastricht/Niederlande und das Statistische Amt von Finnland mit der Aufgabe betraut, neue Indikatoren zur Messung der Zukunftsfähigkeit der Länder der Europäischen Union unter Berücksichtigung der Besonderheiten des der Lissabon Agenda zu Grunde gelegten Konzepts der Wissensgesellschaft (bzw. *knowledge economy*) zu entwickeln bzw. zu untersuchen. Das in diesem Rahmen entstandene und 2009 erfolgreich abgeschlossene KEI (*Knowledge Economy Indicators: Development of Innovative and Reliable Indicator Systems*)-Projekt hat sich mit der Entwicklung dieser Indikatoren sowie mit der Untersuchung ihrer mathematischen bzw. statistischen Eigenschaften befasst.

[1] Auch Lissabon Strategie genannt (siehe EUws (2005)).

Ein besonderer Verdienst innerhalb des KEI-Projekts gebührt der Universität Maastricht, die anhand extensiver Interviews mit Meinungsträgern verschiedenster Wirtschaftsbereiche eine Liste mit 116 Indikatoren zur Untersuchung bzw. Messung der Unterschiede in den wirtschaftlichen und technischen Entwicklungsstadien innerhalb Europas und der Nachhaltigkeit dieser Entwicklung im Hinblick auf die heutige Wissensgesellschaft erarbeitet hat. Dabei handelt es sich um bereits existierende, aber zum Teil erst seit wenigen Jahren von den statistischen Ämtern erhobene Indikatoren, welche bisher nur marginal zum Zustandekommen politischer Entscheidungen beigetragen haben. Diese Indikatoren werden jedoch in Zukunft als Maßstab für den Vergleich der Länder innerhalb der Europäischen Union verwendet und somit weittragende politische und wirtschaftliche Konsequenzen für die europäischen Länder haben (man denke z.B. an die finanzielle Förderung im Rahmen der Europäischen Union).

Um die Verfügbarkeit einer möglichst breiten Basis für die Untersuchung der Eigenschaften der Indikatoren zu sichern, wurde für das KEI-Projekt ein Datensatz zusammengestellt, in dem die 116 Indikatoren für 25 europäische (EU-25) und zwei in der Lissabon Agenda festgelegte, nicht europäische Länder (U.S. und Japan) im Zeitraum von 2001 bis 2004 Einzug finden.

Die hohe Anzahl an vorgeschlagenen Indikatoren spiegelt die Komplexität der modernen Wissensgesellschaft wieder und dient als Beleg dafür, dass diese nicht anhand weniger Maßzahlen zufriedenstellend erfasst werden kann.

Diese Fülle an Indikatoren birgt jedoch die Gefahr, sich als hemmend für die politische Entscheidungsfindung zu erweisen. Wenn auch die Komplexität der modernen Wirtschaft einerseits nicht anhand weniger Indikatoren erfasst werden kann, lassen sich andererseits keine politischen Maßnahmen anhand hunderter Kennzahlen ergreifen. Die Umsetzung der gewonnenen Erkenntnisse in politische Entscheidungen verlangt nach einer geeigneten Verdichtung der in den Indikatoren enthaltenen Informationen. Dieser Tatsache wird dadurch Rechnung getragen, dass aus den ursprünglichen Indikatoren mittels eines Aggregationsverfahrens zusammengesetzte Indikatoren (*Composite Indicators*) konstruiert werden[2]. Diese Aggregation einer Vielzahl an Kennzahlen in eine Einzige wirft wiederum sehr interessante Fragen auf, wie z.B. die Kompensierung. Man denke beispielsweise an die Anzahl abgeschlossener Ph.D's pro Jahr in einem Land. Diese Kennzahl hat zwar keine unmittelbare Aussagekraft bezüglich des gegenwärtigen Wohlstands, ist jedoch hoch relevant für die Nachhaltigkeit der Entwicklung eines jeden Landes. Im Gegensatz dazu ist die Anzahl an angemeldeten Patenten innerhalb eines Jahres eher eine Momentaufnahme, welche die momentane Innovationsfähigkeit eines Landes widerspiegelt. Würden diese Indikatoren aggregiert, um einen zusammengesetzten Indikator zu bilden, so wäre nicht von vornherein klar, ob ein Anstieg der angemeldeten Patente bei einem gleichzeitigen Rückgang der Anzahl an Ph.D. Abschlüssen tendenziell als Verbesserung oder als Verschlechterung der Lage angesehen werden sollte.

Aufgrund der Wichtigkeit dieser und anderer Fragen, welche mit der Problematik der Aggregation einhergehen, haben sich mehrere Forschungsinstitutionen im Rahmen des KEI-Projekts mit dieser Thematik befasst und eine Vielzahl an Aggregationsalgorithmen vorgeschlagen („Benefit of the doubt"-Methode, Analysis Mulitcriteria usw.). Diese Verfahren werden ausführlich im Bericht des KEI-Projekts behandelt (siehe KEI (2004)).

[2] Die Verwendung von zusammengesetzten Indikatoren ist keinesfalls eine Besonderheit des KEI-Projekts. In der Tat werden zusammengesetzte Indikatoren häufig zur Messung von komplexen Sachverhalten verwendet. Als Beispiel sei der *Innovationsindikator Deutschland* des Deutschen Instituts für Wirtschaftsforschung (DIW) genannt, welcher 150 Einzelindikatoren umfasst.

Trotz der Vielfältigkeit der vorgeschlagenen Verfahren ist allen Aggregationsmethoden die naheliegende Tatsache gemeinsam, dass sie einen kompletten Satz von Einzelindikatoren benötigen, um einen Wert des zusammengesetzten Indikators zu generieren.

Um die Konstruktion eines zusammengesetzten Indikators auf der Grundlage eines mit fehlenden Daten behafteten Datensatzes zu ermöglichen, müssen diese fehlenden Werte vervollständigt werden. Das Ergänzen fehlender Daten mit geeigneten Werten[3] wird in der Sprache der Statistik als *Imputation* bezeichnet.

Die Tatsache, dass der dem KEI-Projekt zur Verfügung stehende Datensatz, welcher die Grundlage für die Arbeit der meisten Partner-Institutionen darstellt, 42% fehlende Daten aufweist, verdeutlicht die Relevanz der Untersuchung von Imputationsmethoden im Rahmen des Projekts.

Dennoch geht die Problematik der fehlenden Daten weit über das Projekt hinaus, denn fehlende Werte sind, bis auf wenige Ausnahmen, in allen Bereichen anzutreffen, in denen mit Daten gearbeitet wird. Diese Problematik dürfte auch so alt sein, wie die Erhebung von Daten selbst. In der Tat ist die Idee der Imputation sehr alt. Belege für Imputationsverfahren lassen sich bereits in der frühen statistischen Literatur des zwanzigsten Jahrhunderts finden (siehe z.B. McKendrick (1926)). Diese Verfahren haben sich jedoch hauptsächlich auf Heuristiken gestützt, bis es Donald Rubin 1976 gelang, ein formales, mathematisches Modell für das Fehlen von Daten zu entwickeln. Dieses Modell liefert ein Kriterium anhand dessen die Qualität eines Imputationsverfahrens gemessen werden kann. Heutzutage basieren die anerkanntesten Imputationsverfahren auf dem Modell von Rubin.

Obwohl es durchaus Bereiche gibt, in denen Daten als direkte Konsequenz des wirtschaftlichen Geschehens entstehen (z.B. Handel an der Börse), ist die Informationsgewinnung meist mit einem erheblichen Geld- und Zeitaufwand verbunden. Man denke z.B. an die Vergütung von Probanden in einer klinisch-statistischen Studie (*Clinical Trial*), in der vier-, fünfstellige Geldbeträge für wenige Messungen ausgezahlt werden müssen.

Die Verwendung *aller* beobachteten Informationen durch die geeignete Imputation der fehlenden Werte erweist sich oft sowohl von einem statistischen als auch von einem finanziellen Standpunkt her als unerlässlich.

Die Relevanz der Problematik fehlender Daten für das KEI-Projekt im Besonderen sowie ihre Allgegenwärtigkeit im Allgemeinen haben die Abfassung dieser Arbeit motiviert.

Ziel der vorliegenden Dissertation ist die Untersuchung und Bereitstellung von Imputationsmodellen unter Berücksichtigung der Besonderheiten des KEI-Datensatzes. Zu diesem Zweck wird im ersten Kapitel die Aufgabe der Imputation definiert und die Problematik der fehlenden Daten unter verschiedenen Aspekten diskutiert. Dies geschieht zunächst von einem praktischen Standpunkt aus, indem bekannte und weit verbreitete Imputationsmethoden kurz erläutert werden. Anschließend wird das Modell von Rubin, ein formales Modell für fehlende Daten, eingeführt. Die gewählte Reihenfolge stimmt mit der chronologischen Entwicklung der Imputationsmethoden überein und macht die mangelnde theoretische Basis vieler in der Praxis eingesetzter Methoden deutlich.

[3] Diese etwas informale und unpräzise Definition lässt jedoch absichtlich die Frage offen, was unter *geeignet* verstanden werden soll. Die Antwort auf diese Frage ist Gegenstand der vorliegenden Arbeit.

Im zweiten Teil des ersten Kapitels wird die Beziehung zwischen dem Modell von Rubin für fehlende Daten und den bereits vorgestellten Imputationsverfahren näher erläutert. In diesem Zusammenhang kristallisiert sich bereits die Überlegenheit einer Methode, nämlich der multiplen Imputation, heraus. Die multiple Imputation im Allgemeinen und die Besonderheiten, welche aus der Beschaffenheit des KEI-Datensatzes hervorgehen, sind Gegenstand der letzten beiden Abschnitte des ersten Kapitels.

Um eine Basis für das Verständnis der ausgewählten Imputationsmodelle zu schaffen, wird im zweiten und dritten Kapitel die Theorie zweier Gruppen von Verfahren behandelt, welche den theoretischen Kern dieser Monographie darstellen und zu den mächtigsten Methoden der Modernen Statistik gehören: Der EM-Algorithmus und die Markov-Chain-Monte-Carlo-Methoden.

In Kapitel 2 wird der EM-Algorithmus eigeführt, eine Methode zur ML-Schätzung der Parameter eines statistischen Modells bei Vorhandensein unvollständiger Daten. Die relevantesten Merkmale dieser Methode werden ausführlich diskutiert.

Der Schätzung der Parameter des Imputationsmodells folgend, werden in Kapitel 3 Simulationsmethoden behandelt, welche zur Vervollständigung der mit fehlenden Daten behafteten Datensätze eingesetzt werden können. Diese i.d.R. bayesianischen Methoden sind unter dem Dachnamen Markov-Chain-Monte-Carlo-Verfahren bekannt und sind in der Lage, mit Hilfe stochastischer Prozesse, komplexe, multidimensionale Verteilungen zu simulieren, aus denen die Imputationswerte gezogen werden können. Da die Theorie der Markov-Chain-Monte-Carlo-Verfahren sehr eng mit der Theorie der Markov-Ketten gekoppelt ist, wird dieses Kapitel mit dem Anhang A ergänzt, in dem die wichtigsten Definitionen und Ergebnisse der Theorie der Markov-Ketten der Vollständigkeit halber aufgeführt sind.

Anschließend wird in Kapitel 4 ein Imputationsmodell vorgestellt, welches das Vorhandensein von Realisationen einer multivariaten Normalverteilung unterstellt und die Basis aller in dieser Arbeit behandelten Imputationsmodelle darstellt. Die Anwendung der in den vorangehenden Kapiteln eingeführten Methoden auf das Normalverteilungsmodell wird ausführlich diskutiert, wobei der Schwerpunkt von der statistischen Theorie auf die algorithmischen Fragestellungen verlagert wird. In diesem Kapitel werden die notwendigen Schritte zur vollständigen Implementierung dieses Basis-Imputationsmodells zur Behandlung von großen Datenmengen diskutiert und Lösungsansätze für die dabei resultierenden Probleme vorgestellt. Die besondere Berücksichtigung von Implementierungsproblemen wird den Rest dieser Dissertation prägen.

Verschiedene Möglichkeiten, das Basis-Modell auf die Besonderheiten des KEI-Datensatzes auszurichten, insbesondere bezüglich Verletzungen der Normalverteilungsannahme und der Problematik der kleinen Stichprobenumfänge, sind Gegenstand des fünften Kapitels. Aufgrund der Tatsache, dass die für den KEI-Datensatz typischen Probleme häufig in der Praxis anzutreffen sind, ist dieses Kapitel von einer hohen praktischen Relevanz, die weit über das KEI-Projekt hinausgeht.

Zum Schluss werden die Hauptresultate der Arbeit kritisch gewürdigt und die noch offenen Probleme diskutiert. Ein kleiner Ausblick rundet die Dissertation ab.

Für die Anfertigung dieser Arbeit und zur Imputation des oben genannten und im Anhang B beschriebenen Datensatzes wurden vom Verfasser alle vorzustellenden Verfahren in der Programmiersprache R implementiert.

Kapitel 1

Imputation

> „*When a data set contains missing values, multiple imputation for missing data appears to be an ideal technique. Most importantly, it allows valid statistical inferences. In contrast, any single imputation method, such as filling in the missing values with either their marginal means or their predicted values from linear regression, typically leads to biased estimates of parameters and thereby invalid inferences.*“ (Liu (1995, S. 139))

Unter Imputation wird die Technik der Vervollständigung eines mit fehlenden Werten behafteten Datensatzes vor seiner statistischen Analyse verstanden (vgl. Mendenhall et al. (2006, S. 373-74)). Es gibt eine Vielzahl an Imputationsverfahren, die sich deutlich voneinander unterscheiden sowohl in ihrer Komplexität als auch hinsichtlich der Qualität der ergänzten Werte. Aufgrund dieser Heterogenität ist es nicht möglich, ohne zusätzliche Kenntnis der verwendeten Imputationsmethode, eine Aussage bezüglich der Qualität der imputierten Daten zu treffen.

Im Folgenden werden ausgewählte Imputationsmethoden kurz vorgestellt, welche grundsätzlich auf Heuristiken basieren. Trotz der Verfügbarkeit moderner, auf einer soliden statistischen Theorie basierender Verfahren, erfreuen sich diese *ad hoc*-Methoden immer noch weiter Verbreitung. Anschließend wird ein formales Modell für die fehlenden Daten, das Modell von Rubin, eingeführt. Anhand des Modells von Rubin ist es möglich, Kriterien aufzulisten, welche die Qualität der imputierten Werte gewährleisten.

1.1 Bekannte *ad hoc*-Methoden

Eliminierungsverfahren:

Obwohl es sich dabei um keine Imputationsmethoden im engen Sinne handelt, dienen diese Methoden demselben Zweck wie die Imputation und werden daher hier kurz erläutert.

- Die ***listwise deletion***-Methode ist die wahrscheinlich bekannteste und am häufigsten verwendete Methode zur Korrektur von Datensätzen mit fehlenden Werten, um die statistische Analyse von unvollständigen Datensätzen mit herkömmlichen Verfahren zu ermöglichen.

Hierbei werden alle Beobachtungen[1] von der Analyse ausgeschlossen, welche fehlende Werte aufweisen. Die Entfernung der Daten erfolgt zeilenweise. Übrig bleiben also lediglich diejenigen Zeilen, welche vollständig beobachtet sind.

Diese Vorgehensweise ist sehr leicht anzuwenden und weit verbreitet. Sie ist jedoch nicht empfehlenswert bei Datensätzen mit mehr als 1% fehlende Werte, aufgrund folgender Tatsachen:

i) Ihre Anwendung geht auf Kosten der verwertbaren Information. Vor allem im Falle multivariater Datensätze kann dadurch das Problem entstehen, dass nur sehr wenige Zeilen ohne fehlende Daten resultieren.

ii) Wenn der Mechanismus, welcher das Fehlen von Daten verursacht, mit den zu bestimmenden Parametern verbunden ist, kann das Löschen der fehlenden Beobachtungen zu beachtlichen, systematischen Verzerrungen führen. Diese Problematik wird in Abschnitt 1.2.1 aufgegriffen und ausführlich diskutiert.

- Die ***omitted variable***-Methode unterscheidet sich von der *listwise deletion* lediglich darin, dass sie, anstatt Beobachtungen, Variable mit fehlenden Daten ausschließt.

Imputationsverfahren (im engen Sinne):

Wie bereits erwähnt, ist die Gruppe der Imputationsverfahren sehr heterogen. Einige herkömmliche Verfahren, welche häufig eingesetzt werden, sind:

1. ***Hot-Deck*-Imputation:** Bei dieser Methode wird versucht, eine Gruppe von beobachteten Daten zu finden, welche den fehlenden Daten bezüglich bestimmter Gesichtspunkten ähnlich sind. Aus dieser Gruppe wird dann eine Beobachtung mit Zurücklegen gezogen und diese als Ersatz für den fehlenden Wert verwendet.

2. **Unbedingte-Mittelwert-Imputation:** Hierbei werden die fehlenden Daten spaltenweise mit dem unbedingten arithmetischen Mittel der Spalte ersetzt.

3. **Bedingte-Mittelwert-Imputation:** Diese Methode stellt eine Verbesserung der unbedingten Mittelwert-Imputation dar. Für jede Spalte mit fehlenden Werten wird eine Hilfsvariable herangezogen, welche eine hohe Korrelation mit der zu ergänzenden Variablen aufweisen soll. Mittels einer linearen Regression auf diese Hilfsvariable werden dann Prognosewerte für die fehlenden Daten bestimmt. Die Verallgemeinerung auf mehrere prognostizierende Variable ist im Prinzip möglich. Die daraus resultierenden imputierten Werte haben bessere statistische Eigenschaften als diejenigen, welche von den restlichen hier vorgestellten Methoden generiert werden.

   Diese multidimensionale Methode bringt jedoch gewisse rechentechnische Schwierigkeiten mit sich, welche in Abschnitt 4.1.3.6 anhand eines Beispiels veranschaulicht werden.

Die bisher diskutierten Methoden sind grundsätzlich heuristisch motiviert. Es ist daher schwierig, Kriterien heranzuziehen, um die Qualität der Imputationen zu bewerten. Theoretisch fundierte Methoden dagegen basieren i.d.R. auf der Betrachtung der Likelihood einer Stichprobe und bedürfen somit eines theoretischen Rahmens, welcher im Folgenden eingeführt wird.

[1] Da für die Zwecke dieser Arbeit die Zeilen eines Datensatzes als Realisierungen einer vektorwertigen Zufallsvariablen aufgefasst werden, bezeichnet der Term *Beobachtung* eine Zeile des Datensatzes und nicht die Ausprägung einer einzelnen Variablen.

1.2 Formales Modell für die fehlenden Daten

Der springende Punkt bei der Behandlung von Datensätzen mit fehlenden Werten ist die Untersuchung des Mechanismus, welcher das Fehlen von Daten verursacht hat. Dies liegt darin begründet, dass das Fehlen auf unterschiedliche Ursachen zurückgeführt werden kann, welche die Anwendbarkeit bestimmter Imputationsmethoden deutlich beeinflussen können. Diese Tatsache wurde erst 1976 von Donald Rubin erkannt. Rubin gelang es zu zeigen, dass die Möglichkeit, Daten konsistent zu imputieren, eng mit diesem Mechanismus zusammenhängt.

Um die Behandlung dieser Thematik zu formalisieren wird eine $n \times p$ Datenmatrix $Y = (y_{ij})$ definiert, wobei $i = 1, ..., n$ die n Beobachtungen und $j = 1, ..., p$ die p Variablen bezeichnen.

Bei Vollständigkeit der Daten und unter der i.i.d. Annahme kann die Wahrscheinlichkeits- bzw. Dichtefunktionfunktion von Y folgendermaßen geschrieben werden:

$$P(Y|\theta) := \prod_{i=1}^{n} f(y_i|\theta), \tag{1.1}$$

wobei $f(y_i|\theta)$ die Wahrscheinlichkeits- bzw. Dichtefunktion einer Zeile y_i der Datenmatrix Y und θ einen unbekannten Parametervektor darstellt[2].

Wenn fehlende Daten vorhanden sind, kann jedoch die Wahrscheinlichkeits- bzw. Dichtefunktion von Y nicht mehr gemäß Gleichung (1.1) aufgestellt werden. Die Datenmatrix Y muss in zwei Gruppen zerlegt werden: die beobachteten Daten Y_{obs} und die fehlenden Daten Y_{mis}. Es gilt also $Y = (Y_{obs}, Y_{mis})$ (vgl. Schafer (1997, S. 10ff.)). Ferner wird eine $n \times p$ Indikatormatrix R eingeführt, welche folgendermaßen definiert wird

$$R := (r_{ij}), \text{ wobei } r_{ij} = \begin{cases} 1 & \text{wenn } y_{ij} \text{ nicht beobachtet wurde} \\ 0 & \text{wenn } y_{ij} \text{ beobachtet wurde} \end{cases}$$

für alle $i \in \{1, \ldots, n\}$ und $j \in \{1, \ldots, p\}$.

Die Imputation bezieht sich auf die Simulation der unbeobachtbaren Komponenten Y_{mis} von Y, um einen kompletten Datensatz zu erhalten.

1.2.1 Mechanismus der fehlenden Daten

1.2.1.1 Klassifizierung

Wie in Rubin (1976) beschrieben, ist der Mechanismus der fehlenden Daten[3] durch die bedingte Verteilung von R gegeben Y charakterisiert: $P(R \mid Y, \xi)$, wobei ξ einen unbekannten Parametervektor des MMs darstellt.

[2] Die etwas unübliche Verwendung von P zur Bezeichnung einer Wahrscheinlichkeits- bzw. Dichtefunktion ist typisch für die Fachliteratur auf diesem Gebiet.

[3] Fortan auch MM genannt, aus dem Englischen *missingness mechanism*.

Die Unterschiede in den Eigenschaften dieser bedingten Verteilung des MM-Indikators R ermöglicht die Einteilung der fehlenden Daten in drei Kategorien (nach Rubin (1976)):

(a) ***Missing Completely at Random*** **(MCAR):** Die fehlenden Werte sind eine einfache Zufallsstichprobe aller Daten. Das heißt, das Fehlen von Werten in Y ist unabhängig sowohl von Y_{obs} als auch von Y_{mis}. Die fehlenden Werte werden per Zufall nicht beobachtet und es können keine Informationen herangezogen werden, welche eine Tendenz zum Fehlen erklären könnten.

MCAR $\Longrightarrow P(R \mid Y, \xi) = P(R \mid \xi)$.

(b) ***Missing at Random*** **(MAR):** Diese ist eine weniger restriktive Kategorie der fehlenden Daten. Es wird hierbei nur verlangt, dass die fehlenden Werte eine Zufallsstichprobe innerhalb von Unterklassen beobachteter Werte sind. Das Fehlen von Werten in Y ist also lediglich von Y_{obs} abhängig, jedoch nicht von Y_{mis}.

MAR $\Longrightarrow P(R \mid Y, \xi) = P(R \mid Y_{obs}, \xi)$.

Obwohl ein Restzufall nach wie vor besteht, gibt es unter der MAR-Annahme einen Mechanismus, welcher eine Tendenz zum Fehlen erklären kann.

(c) ***Missing Not at Random*** **(MNAR)**[4]**:** Das Fehlen von Werten in Y hängt ausschließlich von Y_{mis} ab und kann nicht durch Y_{obs} erklärt werden.

MNAR $\Longrightarrow P(R \mid Y, \xi) = P(R \mid Y_{mis}, \xi)$.

Zu beachten ist, dass diese Definitionen keine Restriktionen bzgl. des Musters der fehlenden Werte in der Datenmatrix sind. Dieses Muster ist aus rechentechnischen Gründen ebenfalls relevant und wird in Abschnitt 1.2.2 behandelt.

In dieser Arbeit werden Modelle vorgestellt, welche unter einem MAR-Mechanismus gültig sind. Dies liegt darin begründet, dass ein Wahrscheinlichkeitsmodell für R angenommen wird, welches keine Abhängigkeit von Y_{mis} berücksichtigt.

Unter der MAR-Annahme kann die Abhängigkeitsstruktur im Datensatz verwendet werden, um prognostizierte Werte für die fehlenden Werte in Y_{mis} mit Hilfe der beobachteten Daten zu bestimmen.

Dass die Grenzen zwischen den verschiedenen Kategorien fließend sind, kann anhand folgenden Beispiels verdeutlicht werden:

Im Rahmen einer Stichprobenerhebung werden die Teilnehmer nach ihrem Einkommen gefragt. Es kann hierbei angenommen werden, dass Befragte mit einem sehr hohen Einkommen aus steuertechnischen Gründen und Befragte mit einem sehr geringen Einkommen aus Schamgefühl nicht bereit sind, auf diese Frage zu antworten[5].

Da das Fehlen von Daten ausschließlich eine Funktion der Höhe des Einkommens ist, handelt es sich dabei um einen klaren Fall von MNAR-Mechanismus. Das Heranziehen einer beobachteten Variablen, welche die Haushaltsausgaben erfasst, würde jedoch die Annahme eines MAR-Mechanismus plausibel machen, da das Fehlen der Einkommensdaten nun anhand der Haushaltsausgaben modelliert werden kann.

[4] Andere Autoren verwenden die leicht abweichende Bezeichnung ***Not Missing at Random*** **(NMAR).**

[5] Die Möglichkeit einer bewussten fehlerhaften Angabe wird in dieser Arbeit außer Acht gelassen.

Eine einleuchtende Aussage bezüglich der Beziehung zwischen den MAR- und MNAR-Annahmen, welche auf Schafer (1997, S. 27) zurückgeht, ist die folgende: In der Praxis wird unter der MAR-Annahme nicht verlangt, dass die fehlenden Daten komplett unabhängig von den Variablen in Y_{mis} sind, sondern lediglich, dass diese Abhängigkeit ebenfalls von Y_{obs} erklärt werden kann. Im Falle des Einkommmensbeispiels erklären die fehlenden Werte Y_{mis} der Variablen Einkommen und die beobachteten Werte Y_{obs} der Variablen Haushaltsausgaben den selben Teil an der Tendenz zum Fehlen. MAR-Modelle dürfen somit verwendet werden, denn es gilt

$$P(R \mid Y, \xi) = P(R \mid Y_{obs}, \xi).$$

1.2.1.2 *Distinctness* (Verschiedenheit) der Parameter

In dieser Arbeit wird davon ausgegangen, dass der Parameter des Datenmodells θ und derjenige des MM ξ *distinct* sind gemäß der Definition von Little und Rubin (2002, S. 119).

Distinctness der Parameter liegt vor, wenn der gemeinsame Parameterraum $\Omega_{\theta,\xi}$ von θ und ξ dem kartesischen Produkt der beiden Parameterräume von θ und ξ entspricht, d.h.

$$\Omega_{\theta,\xi} = \Omega_\theta \times \Omega_\xi.$$

Von einem bayesianischen Standpunkt her impliziert die *Distinctness*-Annahme, dass jede *a priori*-Verteilung für $(\theta,\ \xi)$ faktorisierbar ist in die Randverteilungen von θ und ξ (vgl. Schafer (1997, S. 11)).

Wenn beide Annahmen, MAR und *Distinctness* zwischen θ und ξ zutreffen, wird der Mechanismus der fehlenden Daten als ignorierbar bezeichnet.

1.2.1.3 Ignorabilität

Die Ignorabilitätsannahme ist von großem Nutzen für die Maximum-Likelihood-basierte Bestimmung der Parameter θ der Datenmatrix Y.

Zu diesem Zweck wird zuerst die Wahrscheinlichkeitsfunktion der beobachteten Daten explizit geschrieben

$$\begin{aligned} P(R, Y_{obs} \mid \theta, \xi) &= \int P(R, Y \mid \theta, \xi)\, \mathrm{d}Y_{mis} \\ &= \int P(R \mid Y, \theta, \xi)\, P(Y \mid \theta, \xi)\, \mathrm{d}Y_{mis} \\ &= \int P(R \mid Y, \xi)\, P(Y \mid \theta)\, \mathrm{d}Y_{mis}, \end{aligned} \tag{1.2}$$

wobei die Indikatormatrix R ebenfalls berücksichtigt werden muss, um die beobachteten Daten darzustellen.

Die rechte Seite von Gleichung (1.2) kann unter der MAR-Annahme wie folgt umgeformt werden:

$$\begin{aligned}\int P(R \mid Y, \xi) P(Y \mid \theta)\, \mathrm{d}Y_{mis} &= P(R \mid Y_{obs}, \xi) \int P(Y \mid \theta)\, \mathrm{d}Y_{mis} \\ &= P(R \mid Y_{obs}, \xi)\, P(Y_{obs} \mid \theta).\end{aligned}$$

Die zusätzliche Annahme der *Distinctness* stellt sicher, dass die Likelihood-basierte Inferenz bezüglich θ unabhängig von ξ, und somit auch vom Faktor $P(R \mid Y_{obs}, \xi)$ ist. Der Mechanismus der fehlenden Daten kann also ignoriert werden.

Für die Likelihood der beobachteten Daten[6] gilt dann

$$L(\theta \mid Y_{obs}) \propto P(Y_{obs} \mid \theta), \tag{1.3}$$

wohingegen die komplette Likelihood folgendermaßen definiert ist

$$L_{full}(\theta, \xi \mid R, Y_{obs}) \propto P(R, Y_{obs} \mid \theta, \xi).$$

Aufgrund der Ignorabilitätsannahme können also auf L_{ign} und auf L_{full} basierenden Inferenzen bezüglich θ als äquivalent betrachtet werden. Die Parameter für die gesamte Datenmatrix $Y = (Y_{obs}, Y_{mis})$ können mittels L_{ign} auf eine einfachere Art und Weise bestimmt werden.

Es sei an dieser Stelle darauf hingewiesen, dass eine Verletzung der Ignorabilitätsannahme nicht automatisch die Nicht-Anwendbarkeit der unter dieser Annahme entwickelten Methoden impliziert, zumindest in multivariaten Zusammenhängen (vgl. Schafer (1997, S. 20ff.)).

Aufgrund ihrer soliden statistischen Basis werden in dieser Arbeit ausschließlich Likelihood-basierte, ignorierbare Verfahren behandelt.

1.2.2 Muster der fehlenden Daten (*missing data pattern*)

Das Muster der fehlenden Daten[7] beschreibt die räumliche Verteilung der fehlenden Daten in einem Datensatz. Zusammen mit dem Mechanismus der fehlenden Daten bestimmt diese Verteilung die Anwendbarkeit der verschiedenen Imputationsmethoden. Für die Zwecke dieser Arbeit können die verschiedenen denkbaren Muster in die in Tabelle 1.1 dargestellten vier Kategorien eingeteilt werden:

(a) Monotones Muster: Aufgrund der speziellen Form dieses Musters ist es möglich, die Likelihood der beobachteten Daten derart zu faktorisieren, dass die Parameterschätzung auf ein Problem mit vollständig beobachteten Daten zurückgeführt werden kann.

[6] Diese Likelihood wird auch als *Likelihood, welche den MM ignoriert,* L_{ign}, bezeichnet (aus dem Englischen: *Likelihood ignoring the MM*).

[7] Fortan auch MP genannt.

Y_1	Y_2	Y_3
1	1	1
1	1	1
1	1	1
1	1	1
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0

$MP1$

Y_1	Y_2	Y_3
1	0	1
1	1	0
1	1	1
1	0	1
0	1	0
0	1	0
1	0	0
1	1	0
0	1	0
1	1	0
0	1	0
1	0	0
1	0	1
1	1	0
1	0	0
1	1	0

$MP2$

Y_1	Y_2	Y_3
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1

$MP3$

Y_1	Y_2	Y_3
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1

$MP4$

Tabelle 1.1: Verschiedene Muster an fehlenden Daten: $MP1$: Monotones Muster, $MP2$: Allgemeines Muster, $MP3$: Fehlende Variable, $MP4$: Vollständig unbeobachtete statistische Einheiten. Die Imputationsmethoden sind ausschließlich auf $MP1$ und $MP2$ anwendbar.

(b) Allgemeines Muster: Es ist keine direkte Faktorisierung der Likelihood möglich. Die Parameterschätzung muss anhand von Methoden für unvollständige Daten erfolgen und ist somit deutlich komplexer.

(c) Fehlende Variable: Da absolut keine Informationen bezüglich der Variablen Y_2 vorliegen, können keine Parameter für diese Variable geschätzt werden. Da die Parameterschätzung wiederum ein Bestandteil des Imputationsprozesses ist, kann diese Variable nicht imputiert werden.

(d) Fehlende statistische Einheiten: Von diesen statistischen Einheiten liegen bzgl. keiner Variablen Beobachtungen vor. Vollständig unbeobachtete Einheiten müssen entfernt werden, denn sie tragen nicht zur Likelihood bei.

Es muss an dieser Stelle darauf hingewiesen werden, dass die monotonen Muster als Untermenge der allgemeinen Muster aufgefasst werden können. Somit können Schätz- bzw Imputationsmethoden für ein allgemeines Muster ohne Probleme auf monotone Muster angewendet werden. Die vorliegende Arbeit berücksichtigt ausschließlich Methoden für ein allgemeines Muster an fehlenden Daten.

1.3 Das Modell von Rubin und die Imputationsmethoden

In diesem Abschnitt wird die Beziehung zwischen dem formalen Modell für die fehlenden Daten und den bereits vorgestellten Imputationsverfahren näher erläutert.

Es ist naheliegend, dass die Eliminierungsverfahren nur unter der MCAR-Annahme eine valide Inferenz ermöglichen. Unter MAR und MNAR stellen die vollständig beobachteten Daten keine unverzerrte Stichprobe aller Daten dar. Daher sind Parameterschätzungen und Tests, welche aus diesen verzerrten Daten hervorgehen, nicht repräsentativ für die zugrunde liegende Grundgesamtheit.

Die einfachen Imputationsmodelle sind unter der MAR-Annahme nicht grundsätzlich verzerrt, haben jedoch erhebliche Probleme, auf welche jetzt näher eingegangen wird.

Hot deck Imputationsverfahren können, je nach Art, unter der MCAR- oder MAR-Annahme gültig sein. Diese Methoden können jedoch nur bei sehr großen Datensätzen annähernd geeignete Ersatzwerte für die fehlenden Daten finden.

Die Imputation mittels des unbedingten arithmetischen Mittels ignoriert die Korrelationsstruktur des Datensatzes und verursacht somit eine Verringerung der resultierenden Korrelationen. Unbedingte und bedingte Mittelwert-Imputationsverfahren sind sowohl unter der MCAR-, als auch unter der MAR-Annahme anwendbar. Sie vernachlässigen jedoch die Streuung um das arithmetische Mittel. Die Varianz des Datenbestands wird dadurch künstlich verringert.

Die Imputation mittels einer Ziehung aus einem geschätzten Modell statt des (bedingten oder unbedingten) arithmetischen Mittels kann dieses Problem beheben. Dennoch sind Imputationen mittels Ziehungen aus einem geschätzten Modell insofern nicht optimal, als dadurch nicht berücksichtigt wird, dass die Modellparameter lediglich Schätzungen der wahren Parameter darstellen. Um dieser Parameterunsicherheit Rechnung zu tragen, sind folgende Schritte erforderlich:

a) Die gemeinsame Verteilung der Parameter und der fehlenden Daten gegeben die beobachteten Daten muss generiert werden.

b) Die prädiktive Verteilung der fehlenden Daten, d.h. die bedingte Verteilung der fehlenden Daten gegeben die beobachteten Daten für alle möglichen Parameter, muss bestimmt werden. Es handelt sich dabei um eine Randverteilung der Verteilung in a), in welcher die Parameter herausintegriert wurden.

Imputationsmethoden, welche Ziehungen aus einer solchen Verteilung erzeugen, werden als *proper* bezeichnet (vgl. Little und Rubin (2002, S. 89)).

Alle in Abschnitt 1.1 vorgestellten Methoden besitzen den Nachteil, dass sie jeden NA- mit einem einzigen Wert ersetzen. Sobald die fehlenden Werte ergänzt wurden, gilt in aller Regel der resultierende Datensatz als vollständig beobachtet und wird mit herkömmlichen Methoden für vollständige Daten analysiert (vgl. Schafer (1997, S. 2)). Ein imputierter Wert ist jedoch nicht einem beobachteten Wert gleichzusetzen. Die zusätzliche Unsicherheit aufgrund der fehlenden Daten darf nicht ignoriert werden. Einige Möglichkeiten, diese Problematik zu berücksichtigen, sind:

- Adjustierung der imputierten Werte eines einzelnen ergänzten Datensatzes: Diese Methoden werden oft in *Survey Sampling* eingesetzt und liefern richtige Standardfehler für einige Schätzmethoden. Da ihr Anwendungsspektrum sehr eingeschränkt ist (vgl. Little und Rubin (2002, S. 76)), werden diese Methoden nicht weiter berücksichtigt.
- Mehrfache Imputation mittels Computerintensiver Verfahren:

  i *Resampling*-Methoden (z.B. Bootstrap): Diese Methoden bestimmen die Standardfehler der geschätzten Parameter anhand der Variabilität der Schätzwerte, welche durch ein *resampling* der beobachteten Daten resultieren. Sowohl Parameterunsicherheit als auch Variabilität aufgrund der fehlenden Daten können mit diesen Methoden modelliert werden. Trotz der breiten Anwendbarkeit der Bootstrap-Verfahren sind diese zu Imputationszwecken suboptimal, denn:
  - Sie benötigen große Stichprobenumfänge.
  - Eine sehr hohe Anzahl an resultierenden (*resampled*) Datensätzen muss gespeichert werden, da die Imputation und die statistische Analyse der ergänzten Datensätze i.d.R. zu unterschiedlichen Zeitpunkten erfolgt.

  ii Multiple Imputation: Diese Methode wurde von Donald Rubin (1978) vorgeschlagen und vereinigt die Simplizität der einfachen Imputation mit der Berücksichtigung der Variabilität, welche typisch für *resampling*-Verfahren ist. Da die vorliegende Arbeit ausschließlich multiple Imputationsverfahren verwendet, wird diese Methode in Abschnitt 1.4 ausführlich diskutiert.

Tabelle 1.2 stellt verschiedene Imputationsmethoden gegenüber und hebt ihre Vor- und Nachteile hervor. Die Überlegenheit der multiplen Imputation wird dadurch ersichtlich.

Qualität der Imputation	Methode	Vorteil	Nachteil
	Eliminierungsverfahren		Verzerrt unter MAR und MNAR
	Unbedingte-Mittelwert Imputation	Unverzerrt unter ignorierbaren Mechanismen	Zerstörung der Korrelationsstruktur
	Bedingte-Mittelwert Imputation		
	Bedingte-Mittelwert Imputation (multivariate Modellierung)	Berücksichtigung der Korrelationsstruktur	Unterschätzung der Varianz der Daten
	Ziehung aus dem geschätzen multivariaten Modell	Korrektur der Varianz	Parameterunsicherheit nicht berücksichtigt
	Ziehung aus der prädiktiven Verteilung der fehlenden gegeben die beobachteten Daten (*proper*)	Berücksichtigung der Parameterunsicherheit	Variabilität aufgrund fehlender Daten nicht berücksichtigt
	Multiple Imputation		

Tabelle 1.2: Vergleich verschiedener Imputationsmethoden.

1.4 Multiple Imputation

Wie bereits erläutert wird die zusätzliche Variabilität aufgrund der fehlenden Daten von der einfachen Imputation ignoriert. Die multiple Imputation[8] stellt eine Möglichkeit dar, diese Variabilität zu berücksichtigen.

Im Prinzip kann jede Imputationsmethode, welche Ziehungen aus der prädiktiven Verteilung der fehlenden Daten gegeben die beobachteten Daten generiert, und somit *proper* ist, zur MI verwendet werden.

Schritte einer MI

Die Anwendung der Methode besteht aus drei Schritten: Imputation, Analyse und Zusammenführung.

1. **Imputation:** Die unvollständigen Daten werden m Mal mittels Ziehungen aus der prädiktiven Verteilung der fehlenden gegeben die beobachteten Daten vervollständigt, wobei m i.d.R. zwischen 2 und 5 liegt.
2. **Analyse:** Jeder der m vervollständigten Datensätze wird der statistischen Analyse unterzogen.
3. **Zusammenführung:** Die Ergebnisse der m Analysen werden kombiniert, wobei einfache Regeln existieren, um diese Einzelergebnisse zu kombinieren.

Abbildung 1.1 veranschaulicht die drei Schritte einer MI (in Anlehnung an van Buuren (2008)).

Obwohl die MI eine sehr natürliche bayesianische Motivation besitzt, verfügt die daraus resultierende Inferenz ebenfalls über gute frequentistische Eigenschaften (vgl. Little und Rubin (2002, S. 87)).

Die vorliegende Arbeit beschäftigt sich ausschließlich mit dem ersten Schritt, d.h. der simulativen Erzeugung der prädiktiven Verteilung der fehlenden Daten gegeben die beobachteten Daten und der Generierung von Ziehungen aus dieser Verteilung. Die Regeln zur Zusammenführung der Ergebnisse werden in Rubin (1987) und Little und Rubin (2002) ausführlich diskutiert.

Eine sehr anschauliche Einführung in die MI stellt der Aufsatz von J. Schafer (1999) dar.

[8] Fortan MI.

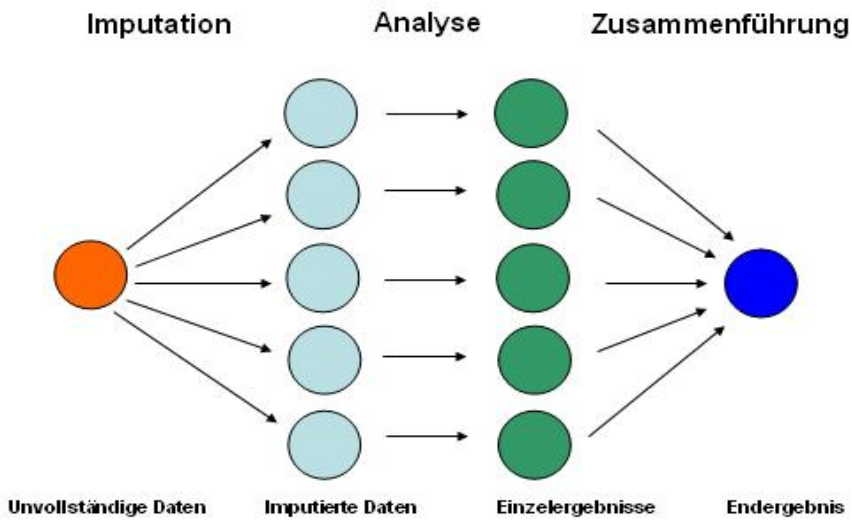

Abbildung 1.1: Schritte einer MI. Die m imputierten Datensätze können mit herkömmlichen Methoden analysiert werden. Die m Ergebnisse werden anschließend zusammengefügt.

1.5 Multiple Imputation und die *Knowledge Economy Indicators*

Die Grundannahme der in der vorliegenden Arbeit behandelten multiplen Imputationsverfahren besteht in der Betrachtung des Datenbestands als Realisationen einer multivariat-normalverteilten Zufallsvariablen.

Dieses multivariat-normale-Imputationsmodell wird in Kapitel 4 ausführlich behandelt und zeichnet sich durch eine sehr flexible und parameterarme Struktur aus. Aufgrund dessen ist es häufig in der Lage, Datensätze mit einem hohen Anteil an fehlenden Werten (NA's) zu imputieren.

Wie es oft in realistischen Situationen der Fall ist, zeigt jedoch die statistische Untersuchung des dem KEI-Projekt zugrunde liegenden Datensatzes empirische Befunde auf, welche die Plausibilität der Annahme einer multivariaten-Normalverteilung erschweren. Die vorliegende Arbeit konzentriert sich daher auf Probleme, welche bei diesen Daten zu beobachten sind und die Anwendbarkeit des Basismodells verhindern. Insbesondere schlägt sie Erweiterungen des Imputationsmodells vor, um folgende Probleme zu beheben:

1. Präsenz von Ausreißern: Unter Ausreißern werden Beobachtungen verstanden, deren Vorkommen unter den unterstellten Verteilungen sehr unplausibel ist (beispielsweise würde unter der Annahme einer Standardnormalverteilung eine Beobachtung mit einem Wert größer als 5 einen Ausreißer darstellen). Solche unplausiblen Beobachtungen kommen jedoch beim KEI-Datensatz häufig vor, z.B. aufgrund uneinheitlicher Definition der Indikatoren in den untersuchten Ländern. In Abschnitt 5.1 wird ein statistisches Kriterium zur Identifizierung von Ausreißern eingeführt und es werden Modelle behandelt, welche durch selektive Gewichtung den verzerrenden Effekt dieser Ausreißer neutralisieren können.

2. Abweichung von der Normalverteilungsannahme: Das Vorhandensein einer multivariaten Normalverteilung impliziert, dass alle Rand- und bedingten Verteilungen ebenfalls normal und die Regressionskurven[9] linear sind. Die Indikatoren des KEI-Datensatzes weisen jedoch häufig empirische Verteilungen und Abhängigkeitsstrukturen auf, deren Form deutlich von der einer Normalverteilung abweicht. Dies macht sich bemerkbar durch

   i) große Abweichungen der empirischen Randverteilungen von der Normalverteiltheit.

   ii) strikt positive Variable. Viele Indikatoren sind nur für Werte größer gleich Null sinnvoll. Man denke z.B. an die Anzahl an angemeldeten Patenten innerhalb eines Jahres.

   iii) Nichtlinearität der Beziehungen zwischen Variablen. Dieses Problem ergibt sich z.B. bei Vorhandensein von Randverteilungen mit unterschiedlich großen dritten Momenten. Im Falle einer multivariaten-Normalverteilung sind diese Momente gleich Null.

   In Abschnitt 5.2 werden Transformationsalgorithmen vorgestellt, welche die aufgelisteten Abweichungen verringern und die Daten auf eine Form bringen, die möglichst kompatibel mit der Annahme einer multivariaten Normalverteilung ist. Somit ergänzen sich die Gewichtungs- und die Transformationsmethoden gegenseitig, denn sie korrigieren unterschiedliche Abweichungen von der unterstellten Verteilung.

3. Kleine Stichproben: Dies resultiert aus der Tatsache, dass die statistischen Einheiten 25 europäische und zwei nichteuropäische Länder sind (siehe Beschreibung des KEI-Datensatzes in Anhang B). Um dieser Problematik Rechnung zu tragen, wird in Abschnitt 4.3 eine einfache Erweiterung des Basismodells vorgenommen, welche die Einbeziehung mehrerer Zeitperioden ermöglicht.

Die Charakteristika des KEI-Datensatzes, welche die Suche nach Weiterentwicklungen des Basismodells motiviert haben, sind häufig in praktischen Situationen zu beobachten, wodurch die breite Anwendbarkeit der in dieser Arbeit behandelten Methoden untermauert wird.

[9] In der mathematischen Statistik bezeichnet der Begriff „Regressionskurve“ die Menge der Punkte eines Zufallsvektors (X_1, X_2) mit der Form $(x_1, E[x_2|x_1])$ bzw. $(E[x_1|x_2], x_2])$. Man beachte die leicht abweichende Bedeutung des Begriffs im Vergleich zur jenem der Regressionsanalyse (vgl. Schaich und Münnich (2001, S. 78)).

Kapitel 2

EM-Algorithmus

Der *Expectation-Maximization*-Algorithmus[1] (kurz: EM-Algorithmus) ist ein iteratives Verfahren zur Bestimmung von Maximum-Likelihood-Schätzern in Situationen, die durch das Vorhandensein folgender Merkmale charakterisiert werden können:

- Die Daten werden nicht vollständig beobachtet bzw. können als unvollständig betrachtet werden.
- Die Komplexität der Likelihood-Funktion der beobachteten (unvollständigen) Daten erschwert die Bestimmung eines ML-Schätzers.
- Durch Ergänzung der unvollständigen Daten kann eine deutliche Vereinfachung der Likelihood-Funktion erzielt werden, welche die ML-Schätzung der zu Grunde liegenden Parameter ermöglicht.

Neben dem offensichtlichen Fall tatsächlich fehlender Werte, welcher den Kern der vorliegenden Abhandlung darstellt, gibt es eine Fülle an statistischen Fragestellungen, die ebenfalls diese Merkmale aufweisen können. Viele statistische Probleme scheinen zwar auf den ersten Blick nichts mit fehlenden Daten zu tun zu haben, können jedoch als ein Problem fehlender Daten aufgefasst werden: Man denke z.B an die Bestimmung von Parametern einer Mischung von Verteilungen anhand einer Stichprobe. Die ML-Schätzung ist in diesem Fall dadurch erschwert, dass es in aller Regel nicht bekannt ist, welche Verteilung der Mischung welche Beobachtungen hervorgebracht hat. Das Vorhandensein einer Indikator-Variablen, welche die Zugehörigkeit der Beobachtungen zu den verschiedenen Verteilungen der Mischung aufzeigt, könnte das Problem auf eine klassische ML-Schätzung zurückführen. Diese latenten Variablen können jedoch nicht beobachtet werden. Die Art und Weise, auf die die beobachteten Daten ergänzt werden und somit die ML-Schätzung ermöglicht wird, stellt das Hauptmerkmal des EM-Algorithmus dar.

[1] Der Begriff „Algorithmus“ ist insofern irreführend, als in der allgemeinen Definition des Verfahrens keine genauen Anweisungen für Rechenschritte gegeben werden, wie es der Begriff Algorithmus erfordern würde. In der Tat kann diese Methode je nach Fragestellung unterschiedliche Gestalten annehmen. Lediglich im Falle von Verteilungen der regulären Exponentialfamilie weist der EM-Algorithmus eine einheitliche Struktur auf. Diese wird in Abschnitt 2.1.6 erläutert.

Die Bedeutung des EM-Algorithmus für die moderne Statistik wird von Meng (2000) folgendermaßen eingeschätzt:

> *„The EM-Algorithm is the most popular and powerful method of the twentieth century for fitting models involving missing data and latent variables."*

Ziel dieses Kapitels ist es, die Struktur und Eigenschaften des Algorithmus zu diskutieren unter besonderer Berücksichtigung der Problematik fehlender Daten und der Möglichkeiten ihrer Imputation.

2.1 Allgemeine Theorie

2.1.1 Maximum-Likelihood-Methode

Aufgrund der bereits erwähnten Tatsache, dass der EM-Algorithmus ein iteratives Verfahren zur Gewinnung von ML-Schätzern ist, erweist es sich als zweckmäßig, einige mit der ML-Schätzmethode verbundene Begrifflichkeiten zu erläutern. Es sei jedoch explizit darauf hingewiesen, dass sich die vorliegende Arbeit nicht mit den Vor- und Nachteilen der ML-Schätzmethode als solche auseinandersetzt. Es wird also im weiteren Verlauf angenommen, dass Maximum-Likelihood eine geeignete Schätzmethode darstellt und das Hauptaugenmerk wird auf die Gewinnung des ML-Schätzers gelegt.

Definition: (Likelihood-Funktion)
Es sei X eine Zufallsvariable und $f(\cdot\,|\,\theta)$ ihre Wahrscheinlichkeits- bzw. Dichtefunktion, die bis auf einen d-dimensionalen Parametervektor $\theta \in \Theta$ eindeutig charakterisiert ist. Dabei ist $\Theta \subseteq \mathbb{R}^d$ die Menge der zulässigen Parameterwerte und x sei eine i.i.d. Stichprobe $x = (x_1, ..., x_n)$ aus X. Dann ist die Likelihood-Funktion der Stichprobe x durch folgende Abbildung definiert

$$\mathcal{L}_x : \Theta \to [\,0, \infty), \quad \theta \mapsto \mathcal{L}_x(\theta) := f(x|\,\theta) = \prod_{i=1}^{n} f(x_i|\,\theta). \tag{2.1}$$

□

Ziel der ML-Methode ist es, eine Schätzung $\hat{\theta}$ für θ zu bestimmen, welche diese Likelihood-Funktion maximiert. Die Berechnung einer Schätzung für θ wird somit auf die Bestimmung eines Maximums der Likelihood-Funktion zurückgeführt.

Definition: (Maximum-Likelihood Schätzer):
Der d-dimensionale Parametervektor $\hat{\theta} \in \Theta$ heißt Maximum-Likelihood-Schätzer (ML-Schätzer) für θ, falls

$$\mathcal{L}_x(\hat{\theta}) = \sup_{\theta \in \Theta} \mathcal{L}_x(\theta) \tag{2.2}$$

gilt.

□

Bei Verwendung der Definition ist jedoch zu beachten, dass Existenz und Eindeutigkeit des ML-Schätzers mit gewissen Regularitätsbedingungen verbunden sind (siehe Greene (2003), Kapitel 17).

Die Maximierung von Likelihood-Funktionen ist oft mit Schwierigkeiten verbunden[2]. Um diese Schwierigkeiten zu umgehen wird in der Regel die Log-Likelihood-Funktion L_x betrachtet

$$L_x : \Theta \to \mathbb{R} \cup \{-\infty\}, \quad \theta \mapsto L_x(\theta) := \begin{cases} \log \mathcal{L}_x(\theta) & \text{für } \mathcal{L}_x(\theta) > 0 \\ -\infty & \text{sonst.} \end{cases} \tag{2.3}$$

Aufgrund der Tatsache, dass der Logarithmus eine monoton wachsende und stetige Funktion ist, nehmen die Likelihood-Funktion und die Log-Likelihood-Funktion für dieselben Werte $\theta \in \Theta$ ihre Extrema an.

In vielen Fällen ist es jedoch unmöglich oder sehr aufwändig eine geschlossene Lösung des Maximierungsproblems zu bestimmen. Es ist dann notwendig, Verfahren zu benutzen, welche $\hat{\theta}$ iterativ berechnen. Ein solches Verfahren ist der EM-Algorithmus.

2.1.2 Fehlende Daten

Um das Verständnis des EM-Algorithmus zu erleichtern, müssen zunächst die fehlenden Daten formal definiert werden. Diese Definition kann in bestimmten Fällen mit der in Abschnitt 1.2 bereits vorgestellten übereinstimmen. Im Allgemeinen trifft dies jedoch nicht zu. Bei den fehlenden Daten im Modell von Rubin handelt es sich um unbeobachtete Werte eines Datensatzes (NA's), wohingegen es sich im Falle des EM-Algorithmus um eine viel allgemeinere Definition handelt, die z.B. latente Variable einschließt. Der wesentliche Unterschied dabei ist, dass latente Variable nicht zufallsbedingt, sondern per Definition nicht beobachtet werden. Aufgrund dieser allgemeinen Definition der fehlenden Daten geht die Anwendbarkeit des EM-Algorithmus deutlich über die Imputationsproblematik hinaus.

Für die folgende Betrachtung sei darauf hingewiesen, dass die gewählte Darstellung an Kapitel 3 der lesenswerten Abhandlung von Schürle (2004) angelehnt ist.

Gegeben sei ein Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathrm{P})$. Ferner definiere für $j \in \mathbb{N}$ die Abbildung

$$X : \begin{cases} \Omega \to \mathcal{X} \subseteq \mathbb{R}^j \\ \omega \mapsto X(\omega) \end{cases}$$

eine Zufallsvariable X, deren Bildraum mit $(\mathcal{X}, \mathfrak{F}_X, \mathrm{P}_X)$ gekennzeichnet sei. Diese Zufallsvariable bringt Realisationen hervor, welche i.A. nicht vollständig beobachtbar sind. Es wird angenommen, dass der beobachtbare Teil von X eine Zufallsvariable Y darstellt, welche für $k \in \mathbb{N}, k \leq j$ durch folgende Abbildung definiert ist

$$Y : \begin{cases} \Omega \to \mathcal{Y} \subseteq \mathbb{R}^k \\ \omega \mapsto Y(\omega). \end{cases}$$

[2] An dieser Stelle sei stellvertretend das *underflow*-Problem genannt, welches dadurch entsteht, dass die Werte der Dichtefunktion $f_X(x_i|\theta)$ für $i = 1, \ldots, n$ gemäß Gleichung (2.1) miteinander multipliziert werden. Mit wachsendem Stichprobenumfang wird dieses Produkt immer kleiner, bis es die Darstellungsmöglichkeiten jeglichen Computers unterschreitet.

Der Bildraum von Y ist dabei gegeben durch $(\mathcal{Y}, \mathfrak{F}_Y, \mathrm{P}_Y)$.

Ferner stellen die fehlenden Werte (unbeobachtete) Realisationen einer Zufallsvariablen Z dar, welche folgendermaßen definiert

$$Z : \begin{cases} \Omega \to \mathcal{Z} \subseteq \mathbb{R}^{j-k} \\ \omega \mapsto Z(\omega) \end{cases}$$

ist. Der Bildraum von Z sei durch $(\mathcal{Z}, \mathfrak{F}_Z, \mathrm{P}_Z)$ bezeichnet.

Der Definition der Bildräume von Y und Z ist zu entnehmen, dass $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ gilt. Somit lässt sich jedes Element $x \in \mathcal{X}$ als kartesisches Produkt von Elementen $y \in \mathcal{Y}$ und $z \in \mathcal{Z}$ darstellen.

Von den Elementen von X ist jedoch nur die Komponente Y beobachtbar. Wenn π die Projektion von $X = Y \times Z$ auf Y beschreibt, d.h. $\pi : \mathcal{X} \to \mathcal{Y}, (y, z) \mapsto \pi(y, z) := y$ gilt, so kann ein Element $x \in X$ durch die Zugehörigkeit zum Urbild $\pi^{-1}(\{y\})$ charakterisiert werden. Diese Abbildung ist jedoch im Allgemeinen nicht bijektiv, ein Element $x \in X$ kann also durch $\pi^{-1}(\{y\})$ nicht eindeutig charakterisiert werden. Durch die Kenntnis von y wird lediglich die Menge der Realisationen von x eingeschränkt.

Schließlich seien mit L_x und L_y die Log-Likelihood-Funktionen der vollständigen und beobachteten Daten gekennzeichnet. Die Beziehung der beiden Likelihood-Funktionen untereinander und deren Beziehung zum EM-Algorithmus sind Gegenstand des nächsten Abschnitts.

2.1.3 Darstellung des EM-Algorithmus

Ziel des EM-Algorithmus ist es, ein Maximum der Log-Likelihood der beobachteten Werte L_y zu bestimmen. Dies kann jedoch aufgrund ihrer Komplexität sehr schwierig bzw. unmöglich sein. Hingegen besitzt L_x, wie in Abschnitt 2 angenommen wurde, häufig eine viel einfachere funktionale Form. Diese Likelihood ist jedoch nicht beobachtbar. Der Zusammenhang zwischen beiden Likelihood-Funktionen wird in Abbildung (2.1) veranschaulicht.

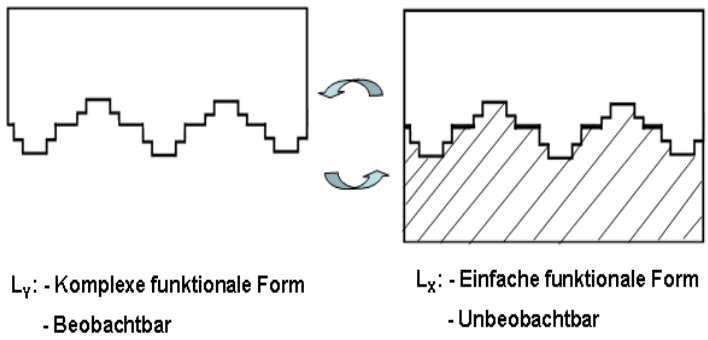

Abbildung 2.1: Zusammenhang zwischen L_y und L_x.

Die Log-Likelihood L_y kann also aufgrund ihrer Beschaffenheit nicht maximiert und Die Log-Likelihood L_x nicht beobachtet werden.

Die ML-Schätzung der vorliegenden Daten wird dadurch ermöglicht, dass der bedingte Erwartungswert der Likelihood der vollständigen Daten gegeben die beobachteten Daten, d.h. $E[L_x|y]$, anstelle von L_y maximiert wird. Um den bedingten Erwartungswert $E[L_x|y]$ berechnen zu können, wird jedoch der Parameter θ benötigt, welcher das eigentliche Ziel des Schätzverfahrens darstellt. Diese gegenseitige Abhängigkeit von Maximierung und Erwartungswertbildung deutet bereits auf eine iterative Lösung hin, welche den Kern des EM-Algorithmus bildet: Ausgehend von einem Startwert θ_0 für den Parameter θ wird der bedingte Erwartungswert der Likelihood der vollständigen Daten gegeben die beobachteten Daten berechnet und anschließend maximiert. Als Ergebnis der Maximierungsaufgabe erhält man einen neuen Schätzwert $\hat{\theta}_1$ für θ, welcher die Grundlage für eine neue Erwartungswertbildung bildet. Diese Vorgehensweise wird wiederholt bis ein Konvergenzkriterium erreicht wird.

Die aus Schätzung und Maximierung erhaltene Iterationsvorschrift wird als Estimation-Maximization-Algorithmus bezeichnet.

2.1.4 Formale Darstellung

Der EM-Algorithmus stellt eine Beziehung her zwischen der Likelihood-Funktion der beobachteten Daten $L_y(\theta)$ und dem Erwartungswert der Likelihood der vollständigen Daten gegeben die beobachteten Daten, d.h. $E[L_x(\theta)|Y = y, \theta_0]$. Er nutzt diese Beziehung aus, um iterativ ML-Schätzwerte für den unbekannten Parametervektor θ zu erhalten.

Zur Ableitung einer Beziehung zwischen $L_y(\theta)$ und $E[L_x(\theta)|Y = y, \theta_0]$ werden zunächst die betroffenen Dichtefunktionen aufgestellt:

Im Folgenden sei $f_X(\cdot|\theta)$, $\theta \in \Theta$ die Dichtefunktion von X. Es sei angenommen, dass die Bedingung $f_X(\cdot|\theta) > 0$ für alle $\theta \in \Theta$ und für alle Stichproben $x \subseteq \mathcal{X}$ gilt.

Die Dichtefunktion von Y ist eine Randdichte von $X = (Y, Z)$ und somit durch die Kenntnis der Verteilung von X vollständig charakterisiert.

$$f_Y(y|\theta) = \int_{-\infty}^{\infty} f_X\left((y, z)|\theta\right) \, \mathrm{d}z \; > 0. \tag{2.4}$$

Um den bedingten Erwartungswert $E[L_x(\theta)|Y = y, \theta]$ zu berechnen, benötigt man zuerst die bedingte Dichte von Z gegeben $Y = y$. Diese Dichte wird im Folgenden mit $f_Z(\,\cdot\,|Y = y, \theta)$ bezeichnet.

Es gilt:

$$f_Z(z|Y = y, \theta) = \frac{f_X\left((y, z)|\theta\right)}{f_Y(y|\theta)} \iff f_Y(y|\theta) = \frac{f_X\left((y, z)|\theta\right)}{f_Z(z|Y = y, \theta)}. \tag{2.5}$$

Das Logarithmieren der zweiten Gleichung in (2.5) ergibt die folgende Beziehung zwischen L_y und L_x:

$$\begin{aligned}\log f_Y(y|\theta) = L_y(\theta) &= \log f_X((y,z)|\theta) - \log f_Z(z|Y=y,\theta) \\ &= L_x(\theta) - \log f_Z(z|Y=y,\theta).\end{aligned} \tag{2.6}$$

Die Erwartungswertbildung der Funktionen in (2.6) ist in der Regel parameterabhängig. Hierfür seien zwei Parametervektoren θ_1 und $\theta_2 \in \Theta$ gegeben.

Mit Hilfe von θ_1 und θ_2 werden die folgenden, als existent vorausgesetzten Erwartungswerte von $\log f_X(x|\theta_1)$ und $\log f_Z(z|\theta_1)$ bezüglich der prädiktiven Verteilung der fehlenden Daten gegeben die beobachteten Daten, d.h. $f_Z(z|Y=y,\theta_2)$, gebildet:

$$\begin{aligned}E_{\theta_2}[\log f_X(x|\theta_1)|Y=y] &= \int_{-\infty}^{\infty} \log f_X(x|\theta_1) \cdot f_Z(z|Y=y,\theta_2)\,\mathrm{d}z \quad \text{und} \\ E_{\theta_2}[\log f_Z(z|Y=y,\theta_1)|Y=y] &= \int_{-\infty}^{\infty} \log f_Z(z|\theta_1) \cdot f_Z(z|Y=y,\theta_2)\,\mathrm{d}z.\end{aligned} \tag{2.7}$$

Aufgrund der Tatsache, dass $X = Y \times Z$ gilt, ist $\log f_X(\,\cdot\,|\theta)$ eine Funktion, welche für ein festes $y \in \mathcal{Y}$ nur abhängig von der Zufallsvariablen Z ist. Somit bildet man den Erwartungswert der Dichte f_X über die Zufallsvariable Z.

Bildung des bedingten Erwartungswertes in Gleichung (2.6) ergibt den folgenden Zusammenhang:

$$E_{\theta_2}[\log f_Y(y|\theta_1)|Y=y] = E_{\theta_2}[\log f_X((y,z)|\theta_1)\,|Y=y] - E_{\theta_2}[\log f_Z(z|Y=y,\theta_1)|Y=y]\,. \tag{2.8}$$

Man erhält aus (2.6):

$$\begin{aligned}L_y(\theta_1) &= E_{\theta_2}[\log f_X((y,z)|\theta_1)\,|Y=y] - E_{\theta_2}[\log f_Z(z|Y=y,\theta_1)|Y=y] \\ &= E_{\theta_2}[L_x(\theta_1)|Y=y] - E_{\theta_2}[\log f_Z(z|Y=y,\theta_1)|Y=y]\,.\end{aligned} \tag{2.9}$$

Somit wurde das Ziel erreicht, eine geeignete Beziehung zwischen der Log-Likelihood-Funktion von y und dem bedingten Erwartungswert der Log-Likelihood-Funktion von x herzustellen.

Mit Hilfe von (2.9) werden nun folgende Abbildungen definiert:

$$L : \Theta \longrightarrow \mathbb{R} \cup \{-\infty\},\ \theta \longmapsto L(\theta_1) := \log f_Y(y|\theta_1) \tag{2.10}$$

$$Q : \Theta \times \Theta \longrightarrow \mathbb{R} \cup \{-\infty\},\ (\theta_1,\theta_2) \longmapsto Q(\theta_1,\theta_2) := E_{\theta_2}[\log f_X(x|\theta_1)\,|Y=y] \tag{2.11}$$

$$H : \Theta \times \Theta \longrightarrow \mathbb{R} \cup \{-\infty\},\ (\theta_1,\theta_2) \longmapsto H(\theta_1,\theta_2) := E_{\theta_2}[\log f_Z(z|Y=y,\theta_1)|Y=y] \tag{2.12}$$

Mit dieser Notation folgt aus (2.9):

$$L(\theta_1) = Q(\theta_1,\theta_2) - H(\theta_1,\theta_2). \tag{2.13}$$

Mit Hilfe von Gleichung (2.13) kann nun der EM-Algorithmus definiert werden:

Gegeben seien die Zufallsvariablen X, Y und Z und die dadurch induzierten Abbildungen, wie in (2.10), (2.11) und (2.12) eingeführt. Ferner sei $\theta_0 \in \Theta$ ein Startvektor mit $L(\theta_0) > 0$. Ein EM-Algorithmus ist eine Iterationsvorschrift, welche eine Folge $(\theta_t)_{t\geq 0} \subseteq \Theta$ erzeugt, mit der Eigenschaft

$$\theta_t = \max_{\theta\in\Theta} Q(\theta, \theta_{t-1}) \quad \text{für alle } t \in \mathbb{N}. \tag{2.14}$$

D.h. θ_{t-1} ist ein fester Wert, während θ der zu maximierende Parameter ist.

Die Berechnung von $Q(\cdot, \theta_{t-1})$ wird als *Expectation*- und die Bestimmung eines Maximierers θ_t für $Q(\cdot, \theta_{t-1})$ als *Maximization*-Schritt bezeichnet.

Diese duale Funktion von θ deutet bereits auf die iterative Struktur des Algorithmus hin: Ausgehend von einem Startwert $\theta_0 \in \Theta$ mit $L(\theta_0) > 0$ wird der Erwartungswert der Log-Likelihood der vollständigen Daten gegeben die beobachteten Daten, $Q(\cdot, \theta_0)$, berechnet. Anschließend wird $Q(\theta, \theta_0)$ maximiert, d.h. $\theta_1 = \max_{\theta\in\Theta} Q(\theta, \theta_0)$ wird berechnet. Der resultierende Wert θ_1 wird dann für eine neue Erwartungswertbildung verwendet, wodurch eine neue zu maximierende Funktion $Q(\cdot, \theta_1)$ resultiert. Die Iteration wird fortgeführt, bis ein geeignetes Konvergenzkriterium erfüllt wird. Der nach Konvergenz erhaltene Parametervektor θ_τ, $\tau \in \mathbb{N}$ wird als ML-Schätzer für den unbekannten Parametervektor θ aufgefasst. Zwei Abbruchskriterien werden üblicherweise verwendet:

1. Aufeinanderfolgende Parameterwerte: Es wird die relative Veränderung zweier hintereinanderfolgender Werte θ_t, θ_{t-1} betrachtet, d.h. das Kriterium lautet:

   $$\frac{\|\theta_t - \theta_{t-1}\|}{\|\theta_{t-1}\|} \leq \varepsilon, \quad \text{für } \varepsilon > 0.$$

   Da θ in der Regel ein Parametervektor ist, konvergieren seine Komponenten $\theta_t^{(1)}, \ldots, \theta_t^{(p)}$ mit unterschiedlichen Geschwindigkeiten. Es ist daher sinnvoll, ein Kriterium heranzuziehen, anhand dessen diese Bedingung komponentenweise überprüft werden kann. Eine mögliche Wahl ist

   $$\max_j \left\{ \frac{\left|\theta_t^{(j)} - \theta_{t-1}^{(j)}\right|}{\left|\theta_{t-1}^{(j)}\right|} \right\} \leq \varepsilon, \quad \text{für } \varepsilon > 0.$$

2. Aufeinanderfolgende Werte der Likelihood der beobachteten Daten: Es wird die relative Veränderung der Likelihood hintereinanderfolgender Werte θ_t, θ_{t-1} betrachtet, d.h. das Kriterium lautet:

   $$\frac{|L(\theta_t) - L(\theta_{t-1})|}{|L(\theta_{t-1})|} \leq \varepsilon.$$

In Schafer (1997, S. 61) wird die Eignung beider Methoden hervorgehoben. Dennoch muss darauf hingewiesen werden, dass beide Kriterien Vor- und Nachteile aufweisen (vgl. Schürle (2004, S. 63). In Anwendungen, welche durch eine hohe Anzahl an fehlenden Daten gekennzeichnet sind,

kann die Implementierung beider Kriterien und ihre simultane Beobachtung wichtige Erkenntnisse bezüglich der Qualität der Parameterschätzung liefern (vgl. Schafer (1997, S. 61)).

Die wohl wichtigste Eigenschaft des EM-Algorithmus besteht darin, dass die Log-Likelihood-Funktion der beobachteten Daten $L(\theta)$ in jedem Schritt monoton ansteigt. Um dies zu zeigen, müssen die Abbildungen $Q(\cdot, \theta)$ und $H(\cdot, \theta)$ näher betrachtet werden:

H-Funktion:

Satz: *Es sei $\tilde{\theta} \in \Theta$ beliebig, dann gilt*

$$H(\theta, \tilde{\theta}) - H(\tilde{\theta}, \tilde{\theta}) \leq 0 \quad \textit{für alle } \theta \in \Theta.$$

Beweis: *Aus der Definition von H folgt, dass $H(\tilde{\theta}, \tilde{\theta})$* und *$H(\theta, \tilde{\theta}) > -\infty$. Es gilt also*:

$$\begin{aligned} H(\theta, \tilde{\theta}) - H(\tilde{\theta}, \tilde{\theta}) &= E_{\tilde{\theta}}[\log f_Z(z|Y = y, \theta)|Y = y] - E_{\tilde{\theta}}[\log f_Z(z|Y = y, \tilde{\theta})|Y = y] \\ &= E_{\tilde{\theta}}[\log f_Z(z|Y = y, \theta) - \log f_Z(z|Y = y, \tilde{\theta})|Y = y] \\ &= E_{\tilde{\theta}}\left[\log\left(\frac{f_Z(z|Y = y, \theta)}{f_Z(z|Y = y, \tilde{\theta})}\right)\middle| Y = y\right]. \end{aligned}$$

Aufgrund der Annahme über die Existenz der Erwartungswerte und die Konkavität der Logarithmus-Funktion lässt sich nun die Jensensche-Ungleichung anwenden

$$\begin{aligned} E_{\tilde{\theta}}\left[\log\left(\frac{f_Z(z|Y = y, \theta)}{f_Z(z|Y = y, \tilde{\theta})}\right)\middle| Y = y\right] &\leq \log E_{\tilde{\theta}}\left[\frac{f_Z(z|Y = y, \theta)}{f_Z(z|Y = y, \tilde{\theta})}\middle| Y = y\right] \\ &= \log\left(\int_{-\infty}^{\infty} \frac{f_Z(z|Y = y, \theta)}{f_Z(z|Y = y, \tilde{\theta})} \cdot f_Z(z|Y = y, \tilde{\theta})\,\mathrm{d}z\right) \\ &= \log\left(\int_{-\infty}^{\infty} f_Z(z|Y = y, \theta)\,\mathrm{d}z\right) \\ &= \log 1 = 0. \end{aligned}$$

□

Die H Funktion muss also nicht berücksichtigt werden. Das Hauptaugenmerk wird jetzt auf die Q-Funktion gelegt.

Q-Funktion:

Satz: *Für alle $t \in \mathbb{N}$ gilt*

$$Q(\theta_{t+1}, \theta_t) - Q(\theta_t, \theta_t) \geq 0.$$

Beweis: *Per Definition gilt $\theta_t = \max_{\theta \in \Theta} Q(\theta, \theta_{t-1})$. Dies impliziert $Q(\theta_{t+1}, \theta_t) \geq Q(\theta_t, \theta_t)$ und damit die Behauptung.*

□

Wird also eine Folge $(\theta_t)_{t\in\mathbb{N}}$ von dem EM-Algorithmus erzeugt, so ist die Folge $(L(\theta_t))_{t\in\mathbb{N}}$ monoton wachsend.

Diese Monotonieeigenschaft von $(\theta_t)_{t\in\mathbb{N}}$ erweist sich als sehr nützlich für ihre Konvergenz, impliziert sie jedoch nicht. Ist jedoch die Folge $(L(\theta_t))_{t\in\mathbb{N}}$ beschränkt, d.h. es existiert ein $M \in \mathbb{R}$, mit $(L(\theta_t))_{t\in\mathbb{N}} < M$, so ist die Konvergenz durch den bekannten Satz der monotonen Konvergenz gewährleistet.

Gestaltet sich der Beweis der Beschränktheit der Likelihood als nicht praktikabel, so bietet sich ein Umweg durch das Auffinden einer konvergenten Teilfolge an.

Die Intuition hinter diesem indirekten Beweis ist die Folgende: Ist es möglich, eine konvergente Teilfolge der monotonen Folge $(L(\theta_t))_{t\in\mathbb{N}}$ zu finden, so muss die Folge selbst beschränkt sein. Der Beweis dieser Aussage wird durch den Satz von Bolzano-Weierstrass (siehe Königsberger (2000, S. 46)) ermöglicht, welcher besagt, dass jede beschränkte Folge eine konvergente Teilfolge besitzt. Zwar gilt die Umkehrung nicht, d.h. es ist möglich, eine unbeschränkte Folge zu konstruieren, die eine konvergente Teilfolge besitzt. Ist die Folge jedoch monoton und besitzt sie eine konvergente Teilfolge, so muss sie beschränkt sein, und somit konvergent. In Schürle (2004, S. 65) wird ein Konvergenzbeweis für den EM-Algorithmus behandelt, der auf diesen Prinzipien basiert.

Es sei an dieser Stelle darauf hingewiesen, dass es im Allgemeinen nicht gewährleistet ist, dass der EM-Algorithmus gegen ein Maximum konvergiert. Bei Problemen, welche dadurch gekennzeichnet sind, dass die Log-Likelihood-Funktion unimodal und konkav auf dem ganzen Parameterraum Θ ist, konvergiert der EM-Algorithmus von jedem Startwert aus gegen ein eindeutiges Maximum. Entartete Fälle (Sattelpunkte, mehrfache Extrema usw.) werden ausführlich in Schafer (1997, S.51) diskutiert. Stellvertretend wird hier ein bekanntes Beispiel vorgestellt.

Gegeben sei die auf Tabelle 2.1 dargestellte Stichprobe aus einer bivariaten Normalverteilung mit unbekannten Parametern:

X	1	2	1	-	-	-
Y	-	-	-	3	5	1

Tabelle 2.1: Stichprobe mit fehlenden Daten.

Da kein Datenpaar beobachtet wurde, liegen keinerlei Informationen bezüglich des Korrelationskoeffizienten ρ vor. Dieser Parameter ist also anhand der Log-Likelihood-Funktion nicht schätzbar. In diesem Fall ist jeder Korrelationskoeffizient $\hat{\rho}$ im Intervall $[-1, 1]$ ein möglicher stationärer Punkt des EM-Algorithmus.

2.1.5 Vor- und Nachteile des EM-Algorithmus

In diesem Abschnitt werden abschließend die Vor- und Nachteile des EM-Algorithmus gegenübergestellt:

Die größten Vorteile des EM-Algorithmus liegen in seiner Einfachheit, seiner Stabilität und oftmals auch in seiner leichteren technischen Umsetzbarkeit im Vergleich zu konkurrierenden Verfahren (vgl. Schafer (1997, S.51)).

Ein möglicher Nachteil des EM-Algorithmus besteht darin, dass das Verfahren nach Konvergenz keinerlei Informationen über die Qualität des gefundenen Schätzwertes $\hat{\theta}$ im Sinne einer Varianz-Kovarianzmatrix liefert. Eine Möglichkeit, den Algorithmus zusätzlich mit einer solchen Schätzung auszustatten, ist der *Supplemented* EM-Algorithmus von Meng und Rubin (1991), welcher die numerische Bestimmung der Hesse-Matrix an der Stelle des Optimums und somit eine Schätzung der Varianz-Kovarianzmatrix des Schätzers ermöglicht.

Ein weiterer Nachteil des EM-Algorithmus ist seine vergleichsweise langsame Konvergenzgeschwindigkeit. Durch eine Taylor-Entwicklung der Likelihood-Funktion kann gezeigt werden, dass eine durch den EM-Algorithmus erzeugte und gegen einen stationären Punkt $\hat{\theta}$ konvergierende Folge $(\theta_t)_{t\in\mathbb{N}}$, in einer hinreichend kleinen Umgebung von $\hat{\theta}$ eine näherungsweise lineare Konvergenzgeschwindigkeit besitzt (vgl. Schafer (1997, S. 55)). Zum Beispiel ist die Konvergenzgeschwindigkeit des Newton-Raphson Algorithmus in der Nähe des Optimums im Vergleich dazu quadratisch. Seine lineare Konvergenzgeschwindigkeit wird i.d.R. als der größte Nachteil des EM-Algorithmus angesehen. Es wurden daher mehrere Möglichkeiten untersucht, die Konvergenzgeschwindigkeit des EM-Algorithmus zu erhöhen. Ein besonders nennenswertes Verfahren in diesem Zusammenhang wurde 2004 von Radi Varahdan entwickelt und ist in der Lage, durch externes Eingreifen die Konvergenzgeschwindigkeit des EM-Algorithmus zu erhöhen.

2.1.6 EM-Algorithmus für Exponentialfamilien

Wie zu Beginn des Kapitels bereits angemerkt, ist der Name „EM-Algorithmus“ insofern irreführend, als keine genauen Anweisungen für Rechenschritte gegeben werden. Dies hat zur Folge, dass die Implementierungen des EM-Algorithmus zur Lösung verschiedener statistischer Probleme auf den ersten Blick kaum Gemeinsamkeiten aufweisen. Im Falle der Exponentialfamilie (siehe Mittelhammer (1996)) basieren jedoch die verschiedenen Versionen des EM-Algorithmus in der Regel auf den gleichen Eigenschaften und besitzen somit einen gemeinsamen Kern, welcher, in Anlehnung an Tanner (1991), im Folgenden dargestellt wird.

Es sei angenommen, dass die Verteilung der vollständigen Daten $X = (Y, Z)$ zur Exponentialfamilie gehört, d.h. die Verteilung von X besitzt die Form

$$f_X(x|\theta) = b(x)\frac{e^{\theta' T(x)}}{a(\theta)},$$

wobei θ ein $d \times 1$ Parametervektor und $T(x)$ ein $1 \times d$ Vektor von suffizienten Statistiken der Stichprobe x ist. In diesem Fall ist die Q Funktion gegeben durch

$$\begin{aligned} Q(\theta, \theta_i) &= \int_{\mathcal{Z}} \log f_X(y, z|\theta) f_Z(z|Y = y, \theta_i)\, \mathrm{d}z \\ &= \int_{Z} \left[\log b(x) + \theta' T(x) - \log a(\theta)\right] f_Z(z|Y = y, \theta_i)\, \mathrm{d}z \\ &= \int_{\mathcal{Z}} \log b(x) f_Z(z|Y = y, \theta_i)\, \mathrm{d}z + \theta' \int_{\mathcal{Z}} T(x) f_Z(z|Y = y, \theta_i)\, \mathrm{d}z - \log a(\theta). \end{aligned} \tag{2.15}$$

Der erste Term auf der rechten Seite von (2.15) hängt nicht von θ ab und spielt somit bei der Maximierung von $Q(\theta, \theta_i)$ keine Rolle. Aufgrund dessen wird dieser Term im Folgenden nicht

mehr berücksichtigt. Der dritte Term auf der rechten Seite von (2.15) hingegen ist ausschließlich eine Funktion von θ und bleibt bei der Erwartungswertbildung unverändert.

Der Erwartungswert-Schritt beschränkt sich also auf die Berechnung von

$$E\left[T(x)|Y=y,\theta_i\right] = \int_{\mathcal{Z}} T(x) f_Z(z|Y=y,\theta_i)\,\mathrm{d}z = T_i$$

und der M-Schritt auf die Maximierung von $Q(\theta,\theta_i)$, welche in diesem Fall die einfache Form

$$\theta' T_i - \log a(\theta)$$

besitzt.

Nullsetzen der ersten Ableitungen von $Q(\theta,\theta_i)$ ergibt

$$\frac{\partial Q(\theta,\theta_i)}{\partial\theta} = -\frac{\partial \log a(\theta)}{\partial\theta} + \frac{\partial \theta' T_i}{\partial\theta} \stackrel{!}{=} 0.$$

Dies impliziert

$$\frac{\partial \log a(\theta)}{\partial\theta} = T_i. \tag{2.16}$$

Mit

$$a(\theta) = \int_{\mathcal{X}} b(x) e^{\theta' T(x)}\,\mathrm{d}x,$$

folgt für die linke Seite von (2.16)

$$\begin{aligned}\frac{\partial \log a(\theta)}{\partial\theta} &= \frac{\int_{\mathcal{X}} b(x)\frac{\partial e^{\theta' T(x)}}{\partial\theta}\,\mathrm{d}x}{a(\theta)} \\ &= \frac{\int_{\mathcal{X}} T(x) b(x) e^{\theta' T(x)}\,\mathrm{d}x}{a(\theta)} \\ &= E\left[T(x)|\theta\right].\end{aligned}$$

D.h. die Maximierung von $Q(\theta,\theta^i)$ bzgl. θ ist äquivalent zur Lösung der Gleichung

$$E\left[T(x)|\theta_i\right] = T_i. \tag{2.17}$$

Dieses Ergebnis wird in Kapitel 4 aufgegriffen.

2.2 Erweiterungen des EM-Algorithmus

2.2.1 ECM-Algorithmus

In Fällen, in denen sich der M-Schritt als zu schwierig gestaltet, ist es angebracht, in jedem Schritt den Wert der Q Funktion nur zu *erhöhen* anstatt ihn zu maximieren. Dies ist der Grundgedanke hinter dem *generalized* EM-Algorithmus (GEM-Algorithmus) von Dempster et al. (1977). Dieser Algorithmus ist in der Lage, die Likelihood in jeder Iteration zu erhöhen, verliert jedoch einige Konvergenzeigenschaften des EM-Algorithmus (vgl. Little und Rubin (2002, S. 179)). Ein Sonderfall des GEM mit noch relativ guten Konvergenzeigenschaften ist der *Expectation-Conditional Maximization*-Algorithmus (ECM-Algorithmus) von Meng und Rubin (1993). Der ECM-Algorithmus nutzt in gewissen Fällen die Einfachheit einer bedingten Maximierung. Ein M-Schritt des EM-Algorithmus wird durch $S > 1$ bedingte Maximierungsschritte ersetzt. D.h. in jedem CM-Schritt wird die Q-Funktion bezüglich eines Untervektors $\theta_s, s \in \{1 \dots S\}$ von θ maximiert, während die anderen Parameter konstant gehalten werden. Dies ist häufig aus numerischen Gründen einfacher als eine Maximierung über den ganzen Parameterraum von θ.

2.2.2 ECME-Algorithmus

Der *Expectation-Conditional Maximization Either* (ECME) Algorithmus von Liu und Rubin (1994) ist eine Verallgemeinerung des ECM-Algorithmus, welche die Tatsache ausnutzt, dass in gewissen bedingten Schritten direkt die Log-Likelihood anstelle ihres Erwartungswertes maximiert werden kann. In Fällen, in denen diese direkte Maximierung der (bedingten) Log-Likelihood möglich ist, kann auf diese Weise eine deutliche Erhöhung der Konvergenzgeschwindigkeit erzielt werden. Die Monotonieeigenschaften des EM-Algorithmus sind ebenfalls im ECME-Algorithmus vorhanden.

2.2.3 PX-EM-Algorithmus

Eine Weiterentwicklung des EM-Algorithmus, welche eine künstliche Erweiterung des Parameterraums mit dem Ziel vornimmt, die Konvergenzgeschwindigkeit zu erhöhen, ist der so genannte *Parameter-Expanded* EM (PX-EM)-Algorithmus. Dieses Verfahren bettet das zu schäztende Modell in ein größeres Modell mit zusätzlichen (bekannten bzw. frei wählbaren) Parametern ein, welche den Anteil an fehlenden Informationen verringern. Für eine ausführliche Beschreibung des PX-EM-Algorithmus siehe Liu et al. (1998).

Im Gegensatz zu den ersten zwei Methoden, welche im Rahmen der vorliegenden Arbeit implementiert und ausführlich getestet wurden, wurde der PX-EM-Algorithmus nicht verwendet. Dennoch wurde ein einfacher Ansatz, der von Kent et al. (1994) vorgeschlagen wurde, bei den robusten Modellen von Abschnitt 5.1 angewendet. Dieser Ansatz wird in Abschnitt 5.1.3.2 erläutert.

2.3 Historischer Rückblick

Die Idee hinter dem EM-Algorithmus ist intuitiv und natürlich. Aufgrund dessen wurden ähnliche Algorithmen bereits in der ersten Hälfte des 20. Jahrhunderts verwendet.

Meng (2000) berichtet über die Methode von McKendrick, der 1926 ein Verfahren zur statistischen Analyse einer Cholera-Epidemie in einem indischen Dorf entwickelte. Durch das Fehlschlagen eines in der Regel für solche Fälle geeignetes Poisson-Modells geleitet, versuchte McKendrick einen Umweg durch gezieltes Löschen abweichender Beobachtungen und ihre anschließende Ergänzung mittels prognostizierter Werte. Dieser heuristisch motivierter Ansatz legte den Grundstein für die spätere Entwicklung des EM-Algorithmus. In der heutigen statistischen Sprache ist dieses Verfahren als *Zero-inflated*-Poisson-Modell bekannt und gehört zur Klasse der *verallgemeinerten linearen Modelle (GLMs)*.

Trotz ihrer Originalität, welche dazu führte, dass sie im Sammelband *Breakthroughs in Statistics* Vol. III genannt wird, mangelte es dieser Arbeit an der notwendigen Allgemeinheit, um bereits als Version des EM-Algorithmus betrachtet zu werden. Der Ursprung des EM-Algorithmus wird deshalb mit der von H. O. Hartley (1958) in *Biometrics* veröffentlichten Arbeit in Verbindung gebracht, welche dem Verfahren die typische iterative Struktur verlieh und seine Nützlichkeit anhand einer Vielzahl von Beispielen hervorhob. Trotz der ursprünglichen Einschränkung auf diskrete Verteilungen hatte seine iterative Methode bereits die typischen Merkmale eines EM-Algorithmus. Eine Erweiterung der Methode auf stetige Verteilungen mit Hilfe numerischer Integrationsverfahren veröffentlichte Hartley 1971 in Zusammenarbeit mit R. R. Hocking.

Irwin (1963) griff die Methode von McKendrik auf und fügte dem Algorithmus ebenfalls zusätzliche und EM-ähnliche Iterationen hinzu, wodurch die ursprüngliche Schätzung der Parameter verbessert werden konnte.

Der Algorithmus gewann jedoch erst 1977 seine endgültige Form mit dem Beitrag von Dempster, Laird und Rubin, in dem die Methode in voller Allgemeinheit vorgestellt und ihre Eigenschaften untersucht wurden. Diese Autoren prägten den Begriff „EM-Algorithmus“ und zeigten anhand einer Fülle von unterschiedlichen Beispielen die breite Anwendbarkeit des Verfahrens. Obwohl Hartley in der Diskussion zu diesem Aufsatz zeigte, dass viele relevante Ergebnisse bereits in seiner Veröffentlichung von 1971 vorhanden waren, ist die Arbeit von Dempster, Laird und Rubin von bahnbrechender Bedeutung für die moderne Statistik.

Kapitel 3

Markov-Chain-Monte-Carlo-Methoden

In diesem Kapitel wird die Theorie der Markov-Chain-Monte-Carlo-Methoden[1] und ihre wichtigsten Vertreter erläutert. Hauptziele des Kapitels sind:

1. Die Darstellung der relevantesten Ergebnisse in einem möglichst allgemeinen Zusammenhang, der zuweilen weit über denjenigen einer multiplen Imputation hinausgeht.
2. Die Berücksichtigung der Didaktik. Zu diesem Zweck werden verwandte simulative Methoden vorgestellt, die das Verständnis der MCMC-Verfahren deutlich erleichtern. Darüber hinaus wird die Funktionsweise der behandelten Methoden anhand zahlreicher Beispiele veranschaulicht.

3.1 Einführung: Lösung eines multidimensionalen Integrationsproblems

In einem frequentistischen Kontext ist es häufig notwendig, mehrdimensionale Integrationen durchzuführen, welche sich nicht in geschlossener Form darstellen lassen. Man denke z.B. an die Angabe von Wahrscheinlichkeiten bei einer multivariaten Normalverteilung. In solchen Fällen ist man normalerweise auf numerische Verfahren angewiesen. Bei diesen Methoden macht sich jedoch ein Phänomen bemerkbar, welches von dem amerikanischen Mathematiker Richard Bellman untersucht und als *curse of dimensionality* bezeichnet wurde: Mit wachsender Anzahl an Dimensionen wächst die Komplexität bzw. der Rechenaufwand des zu bewältigenden Problems u.a. exponentiell. Dieses Phänomen trifft häufig in multidimensionalen Fragestellungen auf (vgl. Givens und Hoeting (2005, S. 144, 298)).

[1] Im Folgenden auch MCMC.

Die Monte-Carlo-Integrationsmethoden stellen eine Möglichkeit dar, dieses Problem zu vermeiden. Sie sind zwar bei niedrigdimensionalen Fragestellungen aufwändiger als die herkömmlichen numerischen Integrationsverfahren, jedoch steigt der Aufwand mit jeder zusätzlichen Dimension nur geringfügig. Diese Methoden sind somit bei hochdimensionalen Problemen nahezu unerlässlich.

Wenn auch die Monte-Carlo-Integrationsverfahren einen nicht zu unterschätzenden Nutzen für die frequentistische Statistik haben, bietet erst die Bayes-Statistik einen Rahmen, in dem diese Methoden ihr ganzes Potential entfalten. In der Tat besteht ein wesentlicher Bestandteil einer bayesianischen Analyse lediglich darin, Verteilungen zu charakterisieren. In diesem Zusammenhang merkt A. P. Dempster (1987) folgendes an:

> „*I believe that Bayesian statistics is fundamentally a computational theory whereby the implications of a set of statistical data are understood by constructing a probability model and computing a set of relevant probabilities and expectations.*“

Die von Dempster genannte Berechnung von Wahrscheinlichkeiten und Erwartungswerten führt unmittelbar zur Lösung eines Integrationsproblems. Somit zieht sich diese Integrationsproblematik durch die ganze Bayes-Statistik. Diese Aussage wird im Folgenden veranschaulicht (in Anlehnung an Brooks (1998)):

Gegeben seien eine Stichprobe $\mathbf{y}$ und deren Likelihood $\mathcal{L}_{\mathbf{y}}(\theta)$, wobei $\theta \in \mathbb{R}^k$ ein zu schätzender Parametervektor mit *a priori*-Verteilung[2] $f(\theta)$ ist. Mittels dem Bayes-Theorem erhält man die *a posteriori*-Verteilung $f(\theta|\mathbf{y})$ des Parametervektors

$$f(\theta|\mathbf{y}) = \frac{\mathcal{L}_{\mathbf{y}}(\theta)\, f(\theta)}{\int \mathcal{L}_{\mathbf{y}}(\theta)\, f(\theta)\, \mathrm{d}\theta}. \tag{3.1}$$

Das Integral auf der rechten Seite von (3.1) ist eine Proportionalitätskonstante, welche oft bestimmt werden muss.

Nun sei angenommen, dass nicht der ganze Parametervektor θ, sondern nur seine erste Komponente $\theta^{(1)}$ im Mittelpunkt der Betrachtung steht. Die Randverteilung von $\theta^{(1)}$ ist folgendermaßen definiert

$$f^{(1)}(\theta^{(1)}|\mathbf{y}) = \int \cdots \int f(\theta|\mathbf{y})\, \mathrm{d}\theta^{(2)} \ldots \mathrm{d}\theta^{(k)}.$$

Dieses ist ein typisches multidimensionales Integrationsproblem.

Schließlich könnte die Berechnung von Momenten dieser Verteilung erforderlich sein[3]. Z.B. gilt für das erste Moment

$$\mathrm{E}[\theta^{(1)}|\mathbf{y}] = \int \theta^{(1)} f^{(1)}(\theta^{(1)}|\mathbf{y}) \mathrm{d}\theta^{(1)}.$$

Die höheren Momente können auf gleiche Art und Weise berechnet werden. Dies ist ein dritter Fall, in dem die Monte-Carlo-Integration Einsatz findet.

[2] Es sei an dieser Stelle daran erinnert, dass für die Bayes-Statistik Parameter Zufallsvariable darstellen.
[3] Es wird angenommen, dass alle Momente dieser Verteilung endlich sind.

Der Monte-Carlo-Ansatz besteht darin, dass Erwartungswerte mittels arithmetischer Mittel approximiert werden, denn es gilt unter entsprechenden Annahmen

$$\frac{1}{m}\sum_{j=1}^{m}\theta_j \xrightarrow[m\to\infty]{\text{P bzw. f.s.}} \int \theta f(\theta)\ \mathrm{d}\theta.$$

Im Falle identisch und unabhängig-verteilter Zufallsvariabler wird diese Konvergenz durch das *schwache bzw. starke Gesetz der großen Zahlen* sichergestellt. Im ersten Fall konvergiert das arithmetische Mittel in Wahrscheinlichkeit gegen den Erwartungswert von θ und im zweiten Fall sogar fast sicher.

Mittels der Monte-Carlo-Methode können Erwartungswerte mit jedem beliebigen Genauigkeitsgrad approximiert werden, solange genug Rechenleistung bzw. Zeit zur Verfügung stehen (vgl. Casella und George (1992, S. 168)).

Diese einfache Methode setzt allerdings voraus, dass unabhängig und identisch-verteilte Beobachtungen vorliegen. Jedoch besteht die Mächtigkeit der Bayes-Statistik genau darin, dass hochentwickelte und realitätsnahe Modelle aufgestellt werden können. Die Verteilungen, die aus diesen Bayes-Modellen hervorgehen, sind meist multidimensional und von einer sehr komplexen strukturellen Form. Denn sie resultieren in aller Regel aus dem gemeinsamen Produkt von verschiedenen Dichtefunktionen. Das nicht Vorhandensein von Zufallszahlengeneratoren für diese Verteilungen erschwert die Anwendung der klassischen Monte-Carlo-Methode (vgl. Geyer (1992)).

Die MCMC-Verfahren konstruieren eine Markov-Kette, die eine vorgegebene Zielverteilung als invariantes Maß besitzt[4]. Lässt man diese Markov-Kette lange genug laufen, so erhält man *abhängige* Ziehungen aus der Zielverteilung, welche verwendet werden können, um Erwartungswerte durch arithmetische Mittel zu approximieren.

Somit approximieren diese Methoden Integrale mittels Monte-Carlo-Integration und die Markov-Ketten stellen den Zufallszahlengenerator dar, welcher die Ziehungen aus den Zielverteilungen realisiert (vgl. Rizzo (2008, S. 247)). Diese konzeptionelle Eigenschaft der MCMC-Methoden hat zu ihrer enormen Popularität maßgeblich beigetragen (vgl. Cappe und Robert (2000)), denn die Konstruktion von Zufallszahlengeneratoren ist für die meisten multivariaten Verteilungen sehr schwierig und bei komplexen stochastischen Modellen oftmals sogar unmöglich.

Die Denkweise hinter den MCMC-Methoden wird sehr anschaulich von Geyer (1992) im folgenden Zitat zusammengefasst:

> „*The basic idea is very simple. If one is unable to find a way to simulate independent realizations of some complicated stochastic process, it is almost as useful to be able to simulate* ***dependent*** *realizations* $x_1, x_2, \ldots$ *forming an irreducible Markov Chain having the distribution of interest* P *as its stationary distribution.*“

Ein mögliches Problem bei dieser Konstruktion besteht darin, dass die Markov-Kette zwar nach einiger Laufzeit Ziehungen aus der richtigen Verteilung hervorbringen kann, jedoch sind diese Ziehungen in aller Regel nicht voneinander unabhängig. Das arithmetische Mittel wird also anhand von Beobachtungen berechnet, die serielle Abhängigkeiten aufweisen.

[4] Der Begriff *invariantes Maß*, auch als *stationäres Maß* bekannt, wird später genauer präzisiert. Vorerst sei darunter die Grenzverteilung einer Markov-Kette verstanden.

Die Tatsache, dass die auf diese Art und Weise konstruierten Mittelwerte gegen die entsprechenden Erwartungswerte konvergieren wird durch die *Ergodensätze* gesichert. Diese sind eine Verallgemeinerung des starken Gesetzes der großen Zahlen auf Fälle, in denen die Zufallsvariablen nicht mehr unabhängig und identisch verteilt sind. Mit Hilfe dieser *ergodischen Mittelwerte* können die Momente der Zielverteilungen approximiert werden.

Da die Konstruktion von *ergodischen Mittelwerten* nicht für jede Markov-Kette sinnvoll ist, muss sicher gestellt werden, dass die Markov-Ketten mit gewissen Eigenschaften ausgestattet sind. Diese Eigenschaften sind *Irreduzibilität*, *positive Rekurrenz* und *Aperiodizität*. Eine zusätzliche und für die Konstruktion von Markov-Ketten sehr nützliche Eigenschaft, welche die oben genannten Bedingungen impliziert, ist die *Reversibilität*. Im Anhang A werden diese und weitere Eigenschaften von Markov-Ketten behandelt.

Geschichtliche Anmerkung

Obwohl die ersten Versionen von MCMC-Methoden bereits Anfang der fünfziger Jahre angewendet wurden und somit fast gleichzeitig zu den herkömmlichen Monte-Carlo-Methoden entwickelt wurden (vgl. Cappe und Robert (2000)), haben diese Techniken erst mit den Beiträgen von Gelfand und Smith (1990) und der Wiederentdeckung der Arbeiten von Metropolis et al. (1953) und Hastings (1970) an Verbreitung gewonnen. In den letzten 10 Jahren haben sich die Anwendungen und Weiterentwicklungen der MCMC-Methoden vor allem im Bereich der Bayes-Statistik explosionsartig entwickelt.

3.2 Einleitende Methoden

Auch wenn es sich bei *Acceptance-Rejection-* und *Importance-Sampling*-Methoden um keine iterativen Simulationsmethoden und somit um keine MCMC-Methoden handelt, ist ihre Behandlung insofern nützlich, als sie viele Elemente der allgemeinsten MCMC-Technik, des *Metropolis-Hastings*-Algorithmus, enthalten. Darüber hinaus ermöglicht die Einfachheit und Eleganz dieser Methoden einen leichteren Einstieg in die Theorie der MCMC-Verfahren.

3.2.1 *Acceptance-Rejection-Sampling*

Gegeben seien zwei Zufallsvariable X und Y mit Dichtefunktionen $f(X)$[5] bzw. $g(Y)$. Ferner sei c eine Konstante mit der Eigenschaft

$$\frac{f(t)}{g(t)} \leq c,$$

für alle t mit $f(t) > 0$. Die mit c multiplizierte Funktion g wird als majorisierende Funktion von f bezeichnet.

Es sei nun angenommen, dass aufgrund der Beschaffenheit von $f(X)$ die Ziehung von Zufallszahlen aus dieser Verteilung schwierig bzw. unmöglich ist. Die Ziehung von Zufallszahlen aus $g(Y)$ sei dagegen unproblematisch.

In solchen Fällen bietet die *Acceptance-Rejection*-Methode eine Möglichkeit, die einfachere Struktur von $g(Y)$ zu nutzen, um Ziehungen aus $f(X)$ zu generieren. Dies erfolgt unter Zuhilfenahme einer dritten Zufallsvariablen U mit Dichtefunktion $h(U)$, welche gleichverteilt ist im Intervall (0, 1).

Bei der Anwendung der *Acceptance-Rejection*-Methode wird ein Wert y aus $g(Y)$ gezogen und die Ziehung nur dann akzeptiert (d.h. Zielziehung x wird gleich y gesetzt), wenn eine unabhängige Ziehung u aus $h(U)$ kleiner als der Quotient $\frac{f(y)}{cg(y)}$ ist. Im anderen Fall wird sie abgelehnt.

Die Tatsache, dass die auf diese Art und Weise generierten Werte Ziehungen aus $f(X)$ darstellen, kann mit Hilfe des Theorems von Bayes gezeigt werden:

Zuerst sei das Ergebnis $\mathcal{A}$:= „Die Ziehung y aus $g(Y)$ wird akzeptiert“ und seine Wahrscheinlichkeit $\alpha := \mathrm{P}(\mathcal{A})$ definiert[6]. Die Wahrscheinlichkeit α wird auch als Akzeptanzwahrscheinlichkeit bezeichnet. Dann gilt

$$\begin{aligned} \mathrm{P}(\mathcal{A}|Y = y) &= \mathrm{P}\left(U < \frac{f(Y)}{cg(Y)}\Big|Y = y\right) \\ &= \frac{f(y)}{cg(y)}. \end{aligned} \tag{3.2}$$

[5] Beachte die vereinfachende Notation $f(X)$ für $f_X(x)$.
[6] Die Notation $\alpha := \mathrm{P}(\mathcal{A})$ wird im nächsten Abschnitt verwendet.

Im diskreten Fall ist die Wahrscheinlichkeit, dass die Ziehung u aus $h(U)$ akzeptiert wird

$$\begin{aligned} \mathrm{P}(\mathcal{A}) &= \sum_{y \in \{z:Y=z\}} \mathrm{P}(\mathcal{A}|Y=y)g(y) \\ &= \sum_{y \in \{z:Y=z\}} \frac{f(y)}{cg(y)} g(y) = \frac{1}{c}. \end{aligned} \tag{3.3}$$

Die Anwendung des Theorems von Bayes für eine beliebige Ziehung y^* ergibt

$$\mathrm{P}(y^*|\mathcal{A}) = \frac{\mathrm{P}(\mathcal{A}|y^*)\mathrm{P}(y^*)}{\mathrm{P}(\mathcal{A})}.$$

Mit (3.2) und (3.3) folgt somit

$$\begin{aligned} \mathrm{P}(y^*|\mathcal{A}) &= \frac{\frac{f(y^*)}{cg(y^*)} g(y^*)}{\frac{1}{c}} \\ &= f(y^*). \end{aligned}$$

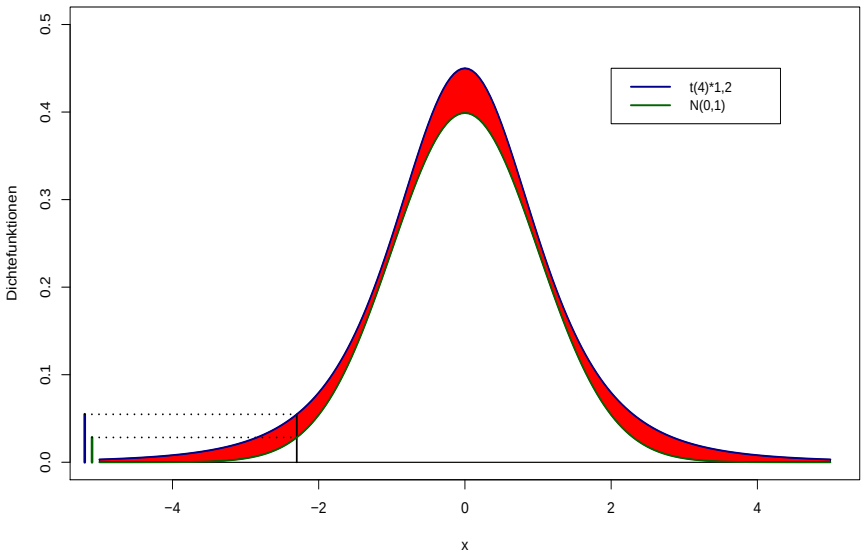

Abbildung 3.1: Dichtefunktion einer Standard-Normalverteilung und majorisierende (mit c= 1,2 skalierte) t-Verteilung mit vier Freiheitsgraden.

Abbildung 3.1 veranschaulicht die Vorgehensweise für den Fall einer Standardnormalverteilung f, die durch eine mit c=1,2 skalierte t-Verteilung mit vier Freiheitsgraden g majorisiert wird [7]. Die unabhängigen Ziehungen u aus $h(U)$ stellen sicher, dass im Durchschnitt $\frac{f(y)}{cg(y)} * 100\%$ aller Ziehungen y aus $g(Y)$ (blaue Linie) akzeptiert werden. Dies ist gleich $f(y)$ (grüne Linie). *Acceptance-Rejection-Sampling* ist somit eine exakte und keine approximative Methode.

[7] Das Beispiel ist insofern etwas künstlich, als es keineswegs einfacher ist, Werte aus einer t-Verteilung als aus einer Normalverteilung zu ziehen und dient lediglich der grafischen Darstellung der Methode.

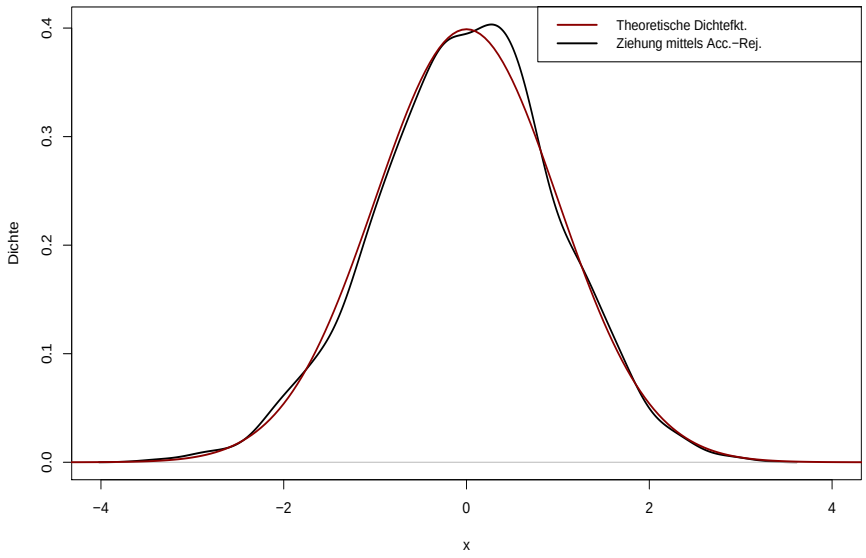

Abbildung 3.2: Kerndichteschätzung der mittels *Acceptance-Rejection* generierten Zufallszahlen und theoretische Zielverteilung.

Die Ergebnisse der Anwendung der *Acceptance-Rejection*-Methode zur Generierung von Zufallszahlen aus einer Standard-Normalverteilung werden in Abbildung 3.2 veranschaulicht. Obwohl die Werte aus einer t-Verteilung mit vier Freiheitsgraden gezogen wurden, weist die empirische Verteilung der Stichprobe (n = 800) in den Flanken die typische Gestalt einer Normalverteilung auf.

Zum Abschluß sei eine interessante Eigenschaft des Algorithmus erwähnt: Aufgrund der Einführung einer Akzeptanzwahrscheinlichkeit α, ist die vom *Acceptance-Rejection-Sampler* benötigte Zeit zur Generierung einer Stichprobe der Größe n selbst zufällig. Algorithmen, die eine solche Eigenschaft besitzen, werden oft als „probabilistische Algorithmen“ bezeichnet (vgl. Fishman (2006)).

3.2.2 *Importance-Sampling*

3.2.2.1 Faktorisierung einer Funktion

Im Mittelpunkt der folgenden Betrachtung steht eine integrierbare Funktion h mit Definitionsbereich D. Mit H sei das Integral von h über ihrem Definitionsbereich definiert. Das Ziel ist nun die Berechnung von $H := \int_D h(x)\mathrm{d}x$. Die Idee ist die Funktion h derart zu faktorisieren, dass einer der Faktoren die Eigenschaften einer Dichtefunktion besitzt. D.h. es gelte

$$h(x) = g(x)f(x), \tag{3.4}$$

mit

$$f(x) \geq 0 \quad \text{für alle } x \in D \quad \text{und} \quad \int_D f(x)\mathrm{d}x = 1.$$

In diesem Fall kann dann das Integral H als der Erwartungswert der transformierten Zufallsvariablen $g(X)$[8] aufgefasst werden. D.h. es gilt

$$H = \mathrm{E}[g(X)] = \int_D g(x)f(x)\mathrm{d}x. \tag{3.5}$$

Das Integral in (3.5) kann nun anhand einer Stichprobe $x_1 \dots x_n$ aus $f(X)$ und des Schätzers

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

approximiert werden. Diese Vorgehensweise, welche die Grundlage der einfachen Monte-Carlo-Integration darstellt, setzt allerdings voraus, dass aus $f(X)$ auf effiziente Weise Stichproben gezogen werden können.

3.2.2.2 *Importance-Sampling*-Faktorisierung

Es ist offensichtlich, dass die Faktorisierung in (3.4) nicht eindeutig ist. Die Wahl der Dichtefunktion $f(X)$ von X hat somit einen erheblichen Einfluss auf die Varianz des Schätzers $\hat{H}$. Mit der Wahl einer möglichst geeigneten Dichtefunktion $f(X)$ für X beschäftigt sich die *Importance-Sampling*-Methode, welche die Faktorisierung in (3.4) folgendermaßen bewerkstelligt

$$\begin{aligned} H &= \int_D h(x)\mathrm{d}x \\ &= \int_D \frac{h(x)}{f(x)} f(x)\mathrm{d}x. \end{aligned} \tag{3.6}$$

Die Funktion f ist dabei wieder eine Dichtefunktion über D und wird als *importance*-Verteilung bezeichnet[9]. Der Schätzer $\hat{H}$ bei einer i.i.d. Stichprobe $x_1 \dots x_n$ aus $f(X)$ ist dann

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n} \frac{h(x_i)}{f(x_i)}.$$

Da es sich um ein arithmetisches Mittel handelt, hat dieser Schätzer die Varianz

$$\mathrm{Var}(\hat{H}) = \frac{1}{n}\,\mathrm{Var}\left(\frac{h(X)}{f(X)}\right).$$

Die Varianz des Schätzers $\hat{H}$ ist somit eine Funktion der Varianz des Quotienten $h(X)/f(X)$. Das Ziel ist also, durch geeignete Wahl von f diese Varianz zu minimieren. Mit Hilfe des Verschiebungssatzes kann dieser Ausdruck wie folgt dargestellt werden

$$\mathrm{Var}\left(\frac{h(X)}{f(X)}\right) = \mathrm{E}\left[\frac{h^2(X)}{f^2(X)}\right] - \left(\mathrm{E}\left[\frac{h(X)}{f(X)}\right]\right)^2. \tag{3.7}$$

[8] Im Gegensatz zum Abschnitt 3.2.1 bezeichnet nun $g(X)$ eine Zufallsvariable und nicht die Dichtefunktion einer Zufallsvariablen. D.h. die Funktion g besitzt i.A. nicht die Eigenschaften einer Dichtefunktion.

[9] Die Bezeichnung *importance*-Funktion ist ebenfalls gebräuchlich.

Zum Zweck der Minimierung lässt sich der Ausdruck auf der rechten Seite von (3.7) vereinfachen, indem man die ursprüngliche Form des zu berechnenden Integrals heranzieht. Dadurch wird der zweite Term auf der rechten Seite von (3.7) zu einer Funktion von h, die bei der Minimierung nicht berücksichtigt werden muss:

$$\left(\mathrm{E}\left[\frac{h(X)}{f(X)}\right]\right)^2 = \left(\int_D h(x)\mathrm{d}x\right)^2.$$

Mit Hilfe der Jensen*schen* Ungleichung erhält man für den ersten Term auf der rechten Seite von (3.7) eine untere Schranke:

$$\begin{aligned}\mathrm{E}\left[\frac{h^2(X)}{f^2(X)}\right] &\geq \left(\mathrm{E}\left[\frac{|h(X)|}{f(X)}\right]\right)^2 \\ &= \left(\int_D |h(x)|\mathrm{d}x\right)^2.\end{aligned} \tag{3.8}$$

Die Dichte $f(X)$ wird nun so gewählt, dass die Ungleichung (3.8) in eine Gleichung umgewandelt wird. Dies wird durch die folgende Wahl von $f(X)$ erreicht

$$f(x) := \frac{|h(x)|}{\int_D |h(x)|\mathrm{d}x}. \tag{3.9}$$

Denn dann gilt

$$\begin{aligned}\mathrm{E}\left[\frac{h^2(X)}{f^2(X)}\right] &= \int_D \frac{h^2(x)}{\left(\frac{|h(x)|}{\int_D |h(x)|\mathrm{d}x}\right)^2} f(x)\mathrm{d}x \\ &= \int_D \frac{h^2(x)}{\left(\frac{|h(x)|}{\int_D |h(x)|\mathrm{d}x}\right)^2} \frac{|h(x)|}{\int_D |h(x)|\mathrm{d}x}\mathrm{d}x \\ &= \int_D \left(\int_D |h(x)|\mathrm{d}x\right)^2 \frac{|h(x)|}{\int_D |h(x)|\mathrm{d}x}\mathrm{d}x \\ &= \int_D |h(x)|\mathrm{d}x \cdot \int_D |h(x)|\mathrm{d}x \\ &= \left(\int_D |h(x)|\mathrm{d}x\right)^2.\end{aligned}$$

Dieses Ergebnis ist insofern nicht konstruktiv, als $\int_D |h(x)|\mathrm{d}x$ in (3.9) bis auf die Betragsfunktion mit dem zu approximierenden Integral $\int_D h(x)\mathrm{d}x$ übereinstimmt. Dennoch ermöglicht dieses Ergebnis folgende Erkenntnis: Um eine möglichst gute Approximation des Integrals zu erzielen, sollte die Dichtefunktion f so gewählt werden, dass sie proportional zur ursprünglichen Funktion h ist.

Auf nahezu gleiche Art und Weise kann die Methode zur Schätzung von Erwartungswerten eingesetzt werden.

Sei nun H wie folgt definiert

$$H = E[g(X)] = \int_D g(x)h(x)\mathrm{d}x. \tag{3.10}$$

D.h. $h(X)$ ist nun die Dichtefunktion der Zufallsvariablen $g(X)$. Diese Gleichung kann analog zu (3.6) auf Seite 38 als

$$H = \int_D g(x) \frac{h(x)}{f(x)} f(x)\,\mathrm{d}x \tag{3.11}$$

dargestellt werden, wobei $f(X)$ die Dichtefunktion einer Zufallsvariablen X sei. Das Integral in (3.11) ermöglicht eine neue Interpretation des Verfahrens: Die Zufallsvariable $g(X)$ mit ursprünglicher Dichtefunktion h erhält eine neue Dichtefunktion f und einen Gewichtungsfaktor $h(t)/f(t)$ für $t \in D$, der die neue Dichtefunktion „korrigiert“. Der Ausdruck $h(t)/f(t)$ für $t \in D$ wird als *importance weight* von f bezeichnet.

Der Schätzer $\hat{H}$ für (3.11) hat somit die Form

$$\hat{H} = \frac{1}{n} \sum_{i=1}^{n} g(x_i) \frac{h(x_i)}{f(x_i)}$$

wobei die $x_1 \ldots x_n$ i.i.d. Ziehungen aus $f(X)$ sind.

Beispiel: Schätzung des Erwartungswertes einer $\Gamma(4, 1)$ Verteilung mittels *Importance-Sampling*. Zwei *importance*-Verteilungen[10] $f(X)$ werden in Betracht gezogen:

(a) Lognormal(1,34; 0,55)

(b) Lognormal(2,2; 0,35)

Die Form der *importance*-Verteilungen und der Zielverteilung sind in Abbildung 3.3 zu sehen. Offensichtlich sind sich (a) und die Zielverteilung relativ ähnlich, während (b) die Wahrscheinlichkeitsmasse anders verteilt.

Die Simulationen für Stichprobenumfänge zwischen eins und 500 sind in Abbildung 3.4 zu sehen. Der wahre Erwartungswert 4,0 der $\Gamma(4, 1)$-Verteilung ist mit der waagrechten roten Linie hervorgehoben. Der Vergleich beider Simulationen zeigt deutlich, welchen Einfluss die richtige Wahl der *Importance*-Verteilung auf die Qualität der Schätzung ausübt. Während sich die Konvergenz des arithmetischen Mittels gegen den wahren Erwartungswert im Fall (a) bereits bei einem Stichprobenumfang von 100 einzustellen beginnt, sind im Fall (b) große Ausschläge der arithmetischen Mittel zu verzeichnen. Dieser sprunghafte Verlauf ist eine Konsequenz des Quotienten $h(x)/f(x)$, der bei dieser *importance*-Verteilung sehr große Werte annehmen kann.

Zum Schluß sei angemerkt, dass beide Methoden in der Tat als Bestandteil der noch zu behandelnden MCMC-Verfahren eingesetzt werden, um deren Eigenschaften zu verbessern bzw. um ihre Benutzung überhaupt erst zu ermöglichen (vgl. Gilks et al. (1996, S. 101 f) bzw. Gilks und Wild (1992)).

[10] Um den Vergleich mit den in Abschnitt 3.3.3.2 auf Seite 50 zu behandelnden Methoden zu ermöglichen, werden im Folgenden die *importance*-Verteilungen auch als „Kandidaten generierende Verteilungen“ bezeichnet. Dabei sind (a) und (b) jeweils die Kandidaten generierenden Verteilungen Nr. 1 und 2.

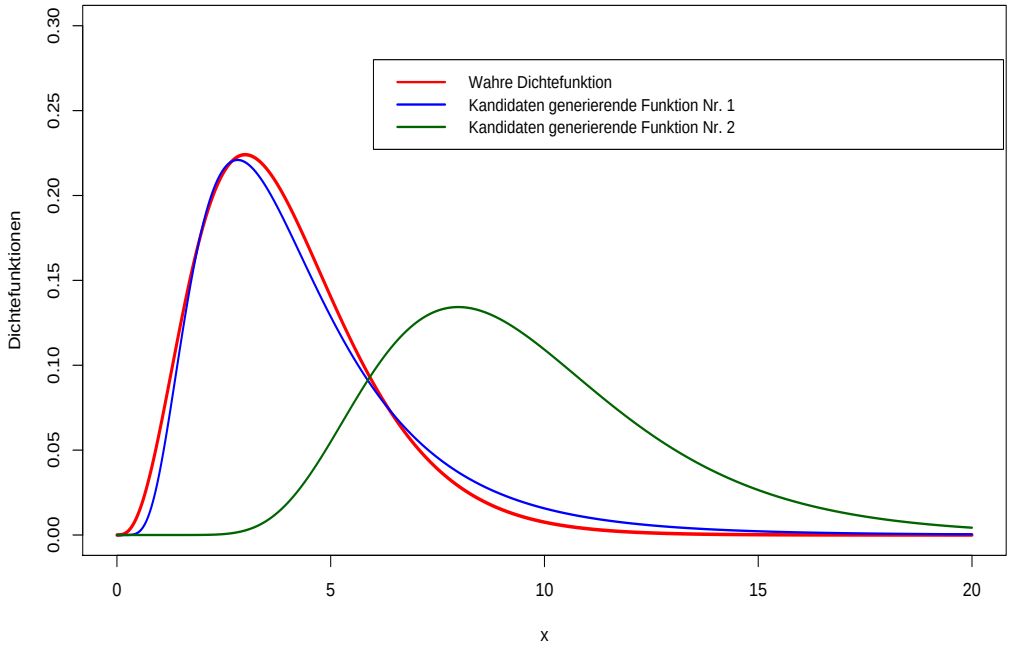

Abbildung 3.3: $\Gamma(4,1)$ Verteilung und ihre Kandidaten generierenden Verteilungen.

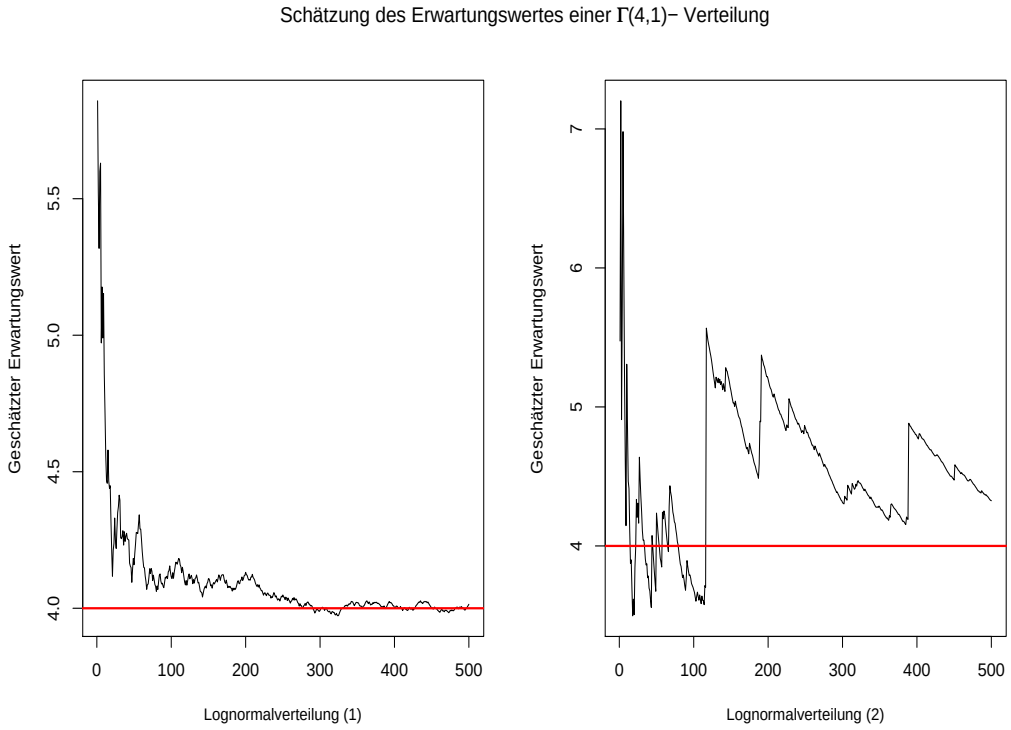

Abbildung 3.4: Simulative Schätzung des Erwartungswertes einer $\Gamma(4,1)$ Verteilung mit dem *Importance-Sampling*-Algorithmus und zwei unterschiedlichen *importance*-Verteilungen und $n = 1, \ldots, 500$.

3.3 Metropolis-Hastings-Algorithmus

3.3.1 Einführung

Der Metropolis-Hastings-Algorithmus stellt eine Art Verallgemeinerung der *Acceptance-Rejection*-Methode aus Abschnitt 3.2.1 dar[11], die Werte aus geeigneten, jedoch nahezu beliebigen Verteilungen generiert und diese Ziehungen derart „korrigiert", dass sie sich für große Stichprobenumfänge wie Ziehungen aus der Zielverteilung verhalten (vgl. Brooks (1998, S. 72)). Beiden Methoden ist die Tatsache gemeinsam, dass sie mit einer durchschnittlichen Wahrscheinlichkeit $1 - \alpha$[12] einen vorgeschlagenen Wert ablehnen. Jedoch stellt beim Metropolis-Hastings-Algorithmus, im Gegensatz zur *Acceptance-Rejection*-Methode, eine hohe Akzeptanzwahrscheinlichkeit α im Allgemeinen kein geeignetes Kriterium dar, um die Qualität der generierten Markov-Ketten zu bewerten (vgl. Cappe und Robert (2000)).

Der Name des Metropolis-Hastings-Algorithmus geht auf die Arbeiten von Metropolis et al. (1953) und Hastings (1970) zurück. Eine Reihe anderer Autoren haben jedoch in den sechziger und siebziger Jahren des letzten Jahrhunderts zur Entwicklung der Methode beigetragen (vgl. z.B. Gamerman (1997, S. 161)). Der erste Beitrag zum Metropolis-Hastings-Algorithmus wurde 1953 von N. Metropolis et. al. in einer Zeitschrift der chemischen Physik veröffentlicht und hat interessanterweise hauptsächlich heuristische Beweise geliefert.

Der Metropolis-Hastings-Algorithmus hat in den letzten Jahren enorm an Bedeutung gewonnen und wird in verschiedenen Bereichen angewendet (siehe z.B. Merz und Wüthrich (2008) für eine Anwendung in der Versicherungsmathematik).

3.3.2 Mathematische Grundlagen

Den Kern der Theorie der Markov Prozesse für die MCMC-Methoden stellen die folgenden zwei Tatsachen dar (vgl. Brooks (1998, S. 71)):

- Jede Markov-Kette, welche irreduzibel, positiv rekurrent und aperiodisch ist (siehe Anhang A.3 auf Seite 153 für eine Erläuterung dieser Begriffe), besitzt eine invariante Verteilung.
- Der t-Schritt-Übergangskern (Siehe Anhang A.1) konvergiert für $t \to \infty$ gegen diese Verteilung [13].

Es ist also ausreichend, um eine Markov-Kette mit invarianter Verteilung π zu generieren, einen Übergangskern K zu finden, für den $\pi K = \pi$ gilt. Die Suche nach einem geeigneten Übergangskern kann allerdings erheblich vereinfacht werden, wenn man die Bedingung der zeitlichen Reversibilität heranzieht, welche wie folgt definiert ist:

$$K(x, y)\pi(x) = K(y, x)\pi(y)$$

für alle x, y aus dem Definitionsbereich von π.

[11] Gleichzeitig kann dieser Algorithmus als Verallgemeinerung der *Importance-Sampling*-Methode aus Abschnitt 3.2.2.2 angesehen werden, wie in 3.3.3.2 auf Seite 50 erörtert wird.

[12] Siehe Definition der Akzeptanzwahrscheinlichkeit α in Abschnitt 3.2.1 auf Seite 35.

[13] Auf die Form der Konvergenz (Konvergenz in totaler Variation) wird an dieser Stelle nicht eingegangen. Der Leser sei auf die Literatur zu Markov Prozessen, z.B. Bremaud (1999) hingewiesen.

Im Folgenden wird gezeigt, dass die Reversibilität von π bezüglich eines Übergangskerns K seine Invarianz bezüglich K impliziert

Diskreter Zustandsraum:

$$\begin{aligned}(\pi K)(j) =& \sum_{i\in S} \pi(i)K(i,j)\\ =& \sum_{i\in S} \pi(j)K(j,i)\\ =& \pi(j) \underbrace{\sum_{i\in S} K(j,i)}_{=1}\\ =& \pi(j)\end{aligned}$$

Stetiger Zustandsraum: Für das invariante Maß muss gelten

$$\pi(\mathrm{d}y) = \pi(y)\mathrm{d}y = \int K(x,\mathrm{d}y)\pi(x)\mathrm{d}x, \tag{3.12}$$

und somit

$$\int K(x,A)\pi(x)\mathrm{d}x = \int \left[\int_A K(x,y)\mathrm{d}y\right] \pi(x)\mathrm{d}x.$$

Der Satz von Tonelli ermöglicht die Vertauschung der Integrationsreihenfolge

$$\int K(x,A)\pi(x)\mathrm{d}x = \int_A \int K(x,y)\pi(x)\,\mathrm{d}x\,\mathrm{d}y. \tag{3.13}$$

Bei Reversibilität von π bzgl. K ist es möglich $K(x,y)\pi(x)$ mit $K(y,x)\pi(y)$ zu ersetzen. Es gilt also

$$\begin{aligned}\int K(x,A)\pi(x)\mathrm{d}x &= \int_A \pi(y) \underbrace{\int K(y,x)\,\mathrm{d}x}_{=1}\,\mathrm{d}y\\ &= \int_A \pi(y)\mathrm{d}y.\end{aligned}$$

Für ein vorgegebenes Maß π ist es in praktischen Situationen i.d.R. nahezu unmöglich, einen Übergangskern K aufzufinden, bzgl. dem π reversibel ist. Die Aufgabe kann jedoch durch folgende Überlegungen beträchtlich erleichtert werden.

Es sei angenommen, dass eine Funktion k existiert, welche die Reversibilitätsbedingung $k(x,y)\pi(x) = k(y,x)\pi(y)$ erfüllt. Ferner sei der Übergangskern K folgendermaßen definiert

$$K(x,dy) := k(x,y)\mathrm{d}y + r(x)\delta_x(\mathrm{d}y), \tag{3.14}$$

wobei $k(x,x) = 0$, $\delta_x(\mathrm{d}y) = 1$ für $x \in \mathrm{d}y$ und 0 sonst gelte. Außerdem ist $r(x) = 1 - \int k(x,y)\mathrm{d}y$ die Wahrscheinlichkeit dafür, dass die Kette in x bleibt. Aufgrund der Tatsache, dass $r(x) \neq 0$ gilt, ist das Integral $\int k(x,y)\mathrm{d}y \neq 1$.

Es wird nun gezeigt, dass die schwächere Bedingung der Reversibilität für die Funktion k anstelle von K ausreichend für die Existenz eines invarianten Maßes π ist. Es gilt

$$\int K(x,A)\pi(x)\mathrm{d}x = \int \left[\int_A k(x,y)\mathrm{d}y\right] \pi(x)\mathrm{d}x + \int r(x)\delta_x(A)\pi(x)\mathrm{d}x. \tag{3.15}$$

Die Vertauschung der Integrationsreihenfolge im ersten Term auf der rechten Seite von (3.15) erfolgt analog zu (3.13). Ferner kann das zweite Integral auf der rechten Seite von (3.15) auf den Bereich A eingeschränkt werden, wodurch sich folgender Ausdruck ergibt.

$$\int K(x,A)\pi(x)\mathrm{d}x = \int_A \left[\int k(x,y)\pi(x)\mathrm{d}x\right] \mathrm{d}y + \int_A r(x)\pi(x)\mathrm{d}x.$$

Die Reversibilität impliziert weiter

$$\begin{aligned}\int K(x,A)\pi(x)\mathrm{d}x &= \int_A \left[\int k(y,x)\pi(y)\mathrm{d}x\right] \mathrm{d}y + \int_A r(x)\pi(x)\mathrm{d}x \\ &= \int_A \pi(y) \left[\int k(y,x)\mathrm{d}x\right] \mathrm{d}y + \int_A r(x)\pi(x)\mathrm{d}x.\end{aligned}$$

Das Integral in eckigen Klammern ist jedoch gemäß der Definition von $r(\cdot)$ gleich $1 - r(y)$

$$\begin{aligned}\int K(x,A)\pi(x)\mathrm{d}x &= \int_A (1 - r(y))\pi(y)\mathrm{d}y + \int_A r(x)\pi(x)\mathrm{d}x \\ &= \int_A \pi(y)\mathrm{d}y.\end{aligned}$$

Im Folgenden wird für den Fall eines diskreten Zustandsraums gezeigt, wie der Metropolis-Hastings-Algorithmus eine solche Funktion $k(x,y)$ konstruiert. Der stetige Fall erfolgt analog.

Sei S ein diskreter und endlicher Zustandsraum, π ein Wahrscheinlichkeitsmaß auf S und K eine stochastische Matrix auf S. Ferner sei folgende Abbildung definiert:

$$g : [0,\infty] \to [0,1],\ x \mapsto g(x) = \begin{cases} 0 & \text{für } x = 0 \\ 1 & \text{für } x = \infty \\ xg(\frac{1}{x}) & \text{für } 0 < x < \infty. \end{cases} \tag{3.16}$$

Schließlich sei für alle $i, j \in S$ die Akzeptanzwahrscheinlichkeit

$$\alpha(i,j) := \begin{cases} 0 & \text{falls } \pi(i)K(i,j) = 0 = \pi(j)K(j,i) \\ g\left(\dfrac{\pi(j)K(j,i)}{\pi(i)K(i,j)}\right) & \text{sonst} \end{cases}$$

und die „korrigierte“ stochastische Matrix $\tilde{K}$

$$\tilde{K}(i,j) := \begin{cases} K(i,j)\,\alpha(i,j) & \text{falls } i \neq j \\ K(i,i) + \sum\limits_{k \neq i} K(i,k)(1-\alpha(i,k)) & \text{falls } i = j \end{cases} \tag{3.17}$$

definiert. Diese „Kandidaten-generierende-Matrix“[14] schlägt zu gegebenem $i \in S$ einen Übergang $i \to j \in S$ vor. Dieser Vorschlag wird mit der Wahrscheinlichkeit $\alpha(i,j)$ akzeptiert, wodurch der neu angenommene Zustand j wird. Mit Wahrscheinlichkeit $1-\alpha(i,j)$ wird der Vorschlag abgelehnt und Markov-Kette bleibt im Zustand i stehen. Es gibt verschiedene Alternativen, eine Funktion g zu definieren, welche die Bedingungen in (3.16) erfüllt. Die wohl bekannteste und für die Zwecke dieser Arbeit relevante Version von g wurde von Hastings (1970) vorgeschlagen und lautet wie folgt:

$$g(x) = \min\{x, 1\}.$$

Diese Funktion erfüllt die Bedingungen in (3.16), denn für $x \leq 1$ gilt $xg(\frac{1}{x}) = x \cdot 1 = x = g(x)$ und für $x \geq 1$, $xg(\frac{1}{x}) = x \cdot \frac{1}{x} = 1 = g(x)$.

Wie bereits erwähnt ist die Matrix $\tilde{K}$ stochastisch. Dies kann leicht anhand einer beliebigen Zeile $i \in S$ gezeigt werden:

$$\sum_{j \in S} \tilde{K}(i,j) \stackrel{!}{=} 1,$$

denn gemäß (3.17) gilt

$$\begin{aligned} \sum_{j \in S} \tilde{K}(i,j) &= \sum_{j \neq i} K(i,j)\alpha(i,j) + \underbrace{K(i,i) + \sum_{k \neq i} K(i,k)(1-\alpha(i,k))}_{i=j} \\ &= \sum_{j \in S} K(i,j) = 1. \end{aligned}$$

Weniger offensichtlich ist die Tatsache, dass $\tilde{K}$ bzgl. π reversibel ist. D.h. es gilt für alle $i, j \in S$:

$$\pi(i)\tilde{K}(i,j) = \pi(j)\tilde{K}(j,i). \tag{3.18}$$

Diese Gleichung ist für die Fälle $i = j$ und $\pi(i)K(i,j) = \pi(j)K(j,i) = 0$ trivialerweise erfüllt. Weiter gilt:

Falls $i \neq j$ und $\pi(i)K(i,j) \neq 0$ gilt:

$$\begin{aligned} \pi(i)\tilde{K}(i,j) &= \pi(i)K(i,j)\alpha(i,j) \\ &= \pi(i)K(i,j)\, g\left(\frac{\pi(j)K(j,i)}{\pi(i)K(i,j)}\right). \end{aligned} \tag{3.19}$$

[14] Auch als „Vorschlagsmatrix“ bezeichnet.

Mit $g(x) = xg(\frac{1}{x})$ folgt daraus weiter

$$\begin{aligned}\pi(i)\tilde{K}(i,j) &= \pi(i)K(i,j)\frac{\pi(j)K(j,i)}{\pi(i)K(i,j)}\, g\left(\frac{\pi(i)K(i,j)}{\pi(j)K(j,i)}\right)\\ &= \pi(j)K(j,i)\,\alpha(j,i)\\ &= \pi(j)\tilde{K}(j,i).\end{aligned}$$

Falls $i \neq j$ und $\pi(i)K(i,j) = 0$ gilt:

$$\pi(i)\tilde{K}(i,j) = \pi(j)K(j,i)\, g\left(\frac{\overbrace{\pi(i)K(i,j)}^{=0}}{\pi(j)K(j,i)}\right) = 0. \tag{3.20}$$

Der Quotient im dritten Term von (3.20) ist wohldefiniert, denn der Ausdruck im Nenner ist ungleich Null. Falls K schon reversibel zu π, so gilt $K = \tilde{K}$.

3.3.3 Metropolis-Hastings-Aktualisierungsschema

Das Metropolis-Hastings Aktualisierungsschema[15] wurde von Hastings (1970) als Verallgemeinerung des Algorithmus von Metropolis et al. (1953) vorgeschlagen. In Anlehnung an Chib und Greenberg (1995) wird die Methode im Folgenden beschrieben:

Analog zu den Erläuterungen in Abschnitt 3.2.2.2 sei eine Kandidaten generierende Dichtefunktion[16]

$$f(x,y) \quad \text{mit} \quad \int f(x,y)\,\mathrm{d}y = 1$$

vorhanden. Diese Kandidaten generierende Dichtefunktion ist das stetige Pendant zu einer Übergangsmatrix und wird folgendermaßen interpretiert: Wenn sich der Prozeß in x befindet, generiert diese Dichtefunktion einen Kandidaten y. Diese Dichtefunktion ist im Allgemeinen abhängig vom Vorgängerwert x und hat somit die Eigenschaften eines Übergangskerns[17].

Gegeben sei nun ein Parametervektor $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, $k \in \mathbb{N}$. Um die $i-te$ Komponente x_i, $i \in k$ zu aktualisieren wird ein Wert aus

$$f(x_i^{t-1}, y) \quad t \in \mathbb{N}$$

gezogen und die Ziehung mit einer bestimmten Wahrscheinlichkeit

$$\alpha(x_i^{t-1}, y)$$

[15] Die Charakterisierung von MCMC-Algorithmen anhand ihrer Aktualisierungsschemata ist üblich für die Fachliteratur auf dem Gebiet.

[16] Die bereits verwendete Bezeichnung Kandidaten generierende Verteilung bzw. Funktion wird häufig in der Literatur zu den MCMC-Methoden benutzt (vgl. Chib und Greenberg (1995)) und macht deutlich, dass die Ziehungen nur mit einer Wahrscheinlichkeit α angenommen werden.

[17] Im Folgenden werden als Übergangskerne bedingte Dichtefunktionen verwendet. Dies macht sich bereits in der Notation $f(\cdot,\cdot)$ anstelle von $K(\cdot,\cdot)$ bemerkbar.

akzeptiert (vgl. Abschnitt 3.2.1 auf Seite 35). Wird die Ziehung akzeptiert, so wird $x_i^t = y$ gesetzt, ansonsten $x_i^t = x_i^{t-1}$. Auch wenn sowohl die *Acceptance-Rejection-* als auch die Metropolis-Hastings-Methode Akzeptanzwahrscheinlichkeiten verwenden, weisen sie Unterschiede auf, die in der folgenden Tabelle zusammengefasst werden:

Acceptance-Rejection

1. Alle Ziehungen sind unabhängig voneinander.
2. Die Ablehnung einer Ziehung führt dazu, dass sie verworfen wird. Die Größe der Stichprobe erhöht sich nicht.

Metropolis-Hastings

1. Die Ziehungen weisen serielle Abhängigkeiten auf.
2. Bei Ablehnung einer Ziehung bleibt der Prozeß im vorherigen Zustand. Die Ziehung ist jedoch gültig. Durch das Verbleiben der Markov-Kette auf einem Punkt erhöht sich die Wahrscheinlichkeitsmasse (bzw. Wahrscheinlichkeitsdichte) des Punktes.

3.3.3.1 Intuitive Motivation der Akzeptanzwahrscheinlichkeiten

Die Anwendung von Akzeptanzwahrscheinlichkeiten stellt den Kern des Metropolis-Hastings-Algorithmus dar und liefert eine Methode, Markov-Ketten mit den in Abschnitt 3.3.2 aufgelisteten Eigenschaften zu konstruieren. Aufgrund dieser konstruktiven Natur ist es sinnvoll, die Akzeptanzwahrscheinlichkeiten näher zu betrachten.

Gegeben sei eine Zielverteilung π und eine Kandidaten generierende Verteilung (Übergangskern) $f(\cdot,\cdot)$[18]. In Abschnitt 3.3.2 auf Seite 42 wurde gezeigt, dass die Reversibilität einer Verteilung π bezüglich eines Übergangskerns $f(\cdot,\cdot)$ eine hinreichende Bedingung dafür ist, dass eine mit $f(\cdot,\cdot)$ als Übergangskern konstruierte Markov-Kette π als invariante Verteilung besitzt. Es sei nun angenommen, dass π nicht reversibel für $f(\cdot,\cdot)$ ist. Dann gilt für ein Paar (x_0, y_0)

$$\pi(x_0)f(x_0,y_0) > \pi(y_0)f(y_0,x_0) \text{ oder}$$
$$\pi(x_0)f(x_0,y_0) < \pi(y_0)f(y_0,x_0).$$

Ohne Beschränkung der Allgemeinheit sei $\pi(x_0)f(x_0,y_0) > \pi(y_0)f(y_0,x_0)$ angenommen. Durch Einführung einer Wahrscheinlichkeit $\alpha(x_0,y_0) < 1$ wird nun die Anzahl der Bewegungen von x_0 nach y_0 verringert. Die Bewegungen von y_0 nach x_0 werden hingegen immer akzeptiert. Dazu wird die Akzeptanzwahrscheinlichkeit wie folgt definiert:

$$\alpha(x_0,y_0) := \frac{\pi(y_0)f(y_0,x_0)}{\pi(x_0)f(x_0,y_0)}.$$

[18] Die Notation $f(\cdot,\cdot)$ macht deutlich, dass es sich dabei um eine Funktion von zwei Argumenten handelt. Im Falle von Funktionen eines einzigen Argumentes wird im Folgenden vereinfachend auf die Klammern verzichtet.

Die Korrigierte Gleichung lautet somit

$$\alpha(x_0, y_0)\pi(x_0)f(x_0, y_0) = \pi(y_0)f(y_0, x_0)$$

und die Reversibilität ist für das Paar (x_0, y_0) erfüllt.

Die Korrektur durch Einführung einer Akzeptanzwahrscheinlichkeit $\alpha(x, y)$ für alle Paare (x, y) führt zu der folgenden Konstruktion

$$\alpha(x, y) = \begin{cases} \min\left\{\dfrac{\pi(y)f(y, x)}{\pi(x)f(x, y)}, 1\right\}, & \text{für } \pi(x)f(x, y) > 0 \\ 1, & \text{sonst.} \end{cases} \tag{3.21}$$

Ein wichtiger Sonderfall entsteht durch die Wahl einer symmetrischen Kandidaten generierenden Verteilung. In diesem Fall gilt

$$f(x, y) = f(y, x)$$

und die Akzeptanzwahrscheinlichkeit in Gleichung (3.21) vereinfacht sich zu

$$\alpha(x, y) = \begin{cases} \min\left\{\dfrac{\pi(y)}{\pi(x)}, 1\right\}, & \text{für } \pi(x)f(x, y) > 0 \\ 1, & \text{sonst.} \end{cases} \tag{3.22}$$

Dies ist die von Metropolis et. al. 1953 vorgeschlagene Akzeptanzwahrscheinlichkeit.

Diese vereinfachte Akzeptanzwahrscheinlichkeit lässt sich wie folgt motivieren: Wenn sich die Kette aufwärts bewegt, d.h. $\pi(y) > \pi(x)$, werden die Bewegungen immer akzeptiert. Hingegen bewegt sich die Kette abwärts, d.h. $\pi(y) < \pi(x)$, mit einer Wahrscheinlichkeit gleich $\frac{\pi(y)}{\pi(x)}$ (vgl. Chib und Greenberg (1995)). Abbildung (3.5) veranschaulicht diese Aussage.

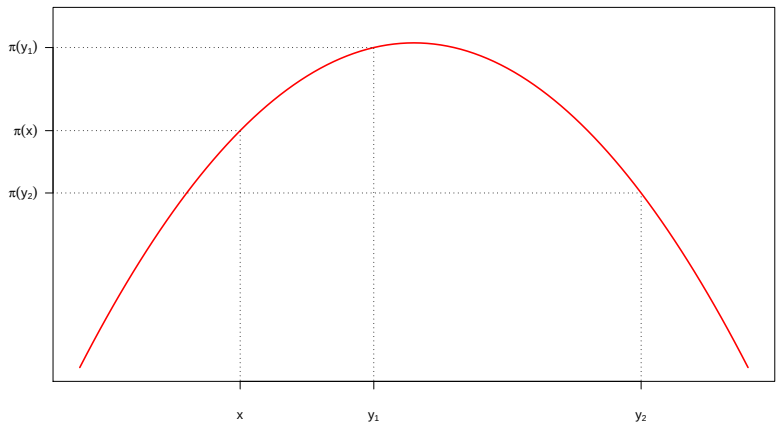

Abbildung 3.5: Übergangswahrscheinlichkeit mit symmetrischer Kandidaten generierender Funktion. Diese Wahrscheinlichkeiten sind ausschließlich eine Funktion der Position der vorgeschlagenen Kandidaten zum Vorgänger.

3.3.3.2 Wahl der Kandidaten generierenden Funktion

Obwohl es im Prinzip unendlich viele Möglichkeiten für die Wahl der Funktion $f(\cdot,\cdot)$ gibt (vgl. Brooks (1998)), sind zwei Klassen von Funktionen von besonderer Relevanz und werden im Folgenden behandelt.

***Random-Walk*-Funktion:** Eine Klasse von Kandidaten generierenden Funktionen $f(\cdot,\cdot)$, welche bereits in der Arbeit von Metropolis et. al. benutzt wurde, hat die Form $f(x,y) = h(y-x)$ für eine beliebige Dichtefunktion h, die uni- bzw. multivariat sein kann (vgl. Brooks (1998)). Unter Verwendung dieser Kandidaten generierenden Funktion hat der Kandidat für $t+1$ die Struktur

$$y^{t+1} = x^t + z, \quad \text{mit } z \sim h. \tag{3.23}$$

Durch (3.23) ist ein *Random-Walk*-Prozeß definiert (vgl. dazu Anhang A.1).

Da die Kandidaten generierenden Funktionen h in der Regel symmetrisch sind,[19] hat die Akzeptanzwahrscheinlichkeit die Form (3.22). Häufig wird die Sprungweite, d.h. die erlaubte Länge der Bewegung des *Random-Walk*s von einem Skalierungsparameter der Kandidaten generierenden Funktion gesteuert, der sorgfältig ausgewählt werden muss. Denn die Effizienz des Simulationsalgorithmus wird maßgeblich von dieser Größe beeinflusst. Bevor dieser Zusammenhang anhand eines Beispiels verdeutlicht wird, wird die Rolle der Akzeptanzwahrscheinlichkeit bei der Simulation von Verteilungen näher betrachtet.

Akzeptanzwahrscheinlichkeit und Güte der Approximation: Wie bereits in Abschnitt 3.3 auf Seite 42 erwähnt, stellt die Höhe der Akzeptanzwahrscheinlichkeit im Allgemeinen kein geeignetes Kriterium dar, um die Güte der Anpassung zu beurteilen: Bei einer geringen Streuung der Kandidaten generierenden Funktion bleiben die neuen Kandidaten in der Nähe des Vorgängers und die Werte von α sind tendenziell nahe bei 1, wodurch fast alle Kandidaten akzeptiert werden. Jedoch braucht die Markov-Kette eine große Anzahl an Iterationen, um den ganzen Definitionsbereich von h zu durchlaufen. Dies verlangsamt die Konvergenz gegen die invariante Verteilung. Andererseits wird der Definitionsbereich bei einer großen Streuung der Kandidaten generierenden Funktion zwar schnell abgedeckt, jedoch werden die meisten Bewegungen abgelehnt und die Kette bleibt für lange Zeit an einem Punkt stehen. Der Simulationsalgorithmus ist dann ineffizient (vgl. Roberts (1996, S. 55)).

Beispiel: Sei $f(x,Y)$ eine gleichverteilte Kandidaten generierende Funktion. Dann ist $Y \sim Gl(x-a, x+a)$ verteilt mit Erwartungswert $E[Y|x] = x$ und Varianz $\mathrm{Var}[Y|x] = \mathrm{Var}[Y] = \frac{a^2}{3}$. Die Zielverteilung π sei $\Gamma(4,1)$. Ferner werden drei Werte für den Parameter a betrachtet: $a_1 = 0.25$, $a_2 = 7.5$ und $a_3 = 25$. Die Varianzen von Y betragen dann 0,021, 18,75 bzw. 208,33. Alle drei Markov-Ketten führen 5000 Iterationen durch. Schließlich wird eine Zufallstichprobe von Umfang 5000 aus einer $\Gamma(4,1)$ zum Vergleich gezogen.

Wie auf der linken Grafik in Abbildung 3.6 zu sehen ist, hat die Markov-Kette nach 5000 Iterationen ungefähr das Intervall (0, 8) durchlaufen. Die Approximation ist trotz einer Akzeptanzrate

[19] In der Regel ist h die Dichte einer Gleichverteilung, einer Normalverteilung oder einer t-Verteilung. Dies gilt sowohl für den univariaten als auch für den multivariaten Fall.

von 97,80% offensichtlich mangelhaft. Die Betrachtung der dazugehörigen Markov-Kette (obere Grafik in Abbildung 3.7) gibt Aufschluss über die Ursache dieser schlechten Approximation: Die Markov-Kette weist eine sehr hohe zeitliche Abhängigkeit zwischen den Realisationen auf. Dies hat zur Folge, dass sich der Prozeß sehr langsam entlang des Definitionsbereiches der Zielverteilung bewegt. Die Fähigkeit, den Definitionsbereich zu durchlaufen, wird in der Literatur häufig als *mixing* bezeichnet (vgl. Gilks et al. (1996, S. 67)). Diese Markov-Kette hat somit ein schlechtes *mixing*.

Die rechte Grafik in Abbildung 3.6 zeigt das Ergebnis der Markov-Kette mit der höchsten Streuung. Obwohl der Algorithmus den Definitionsbereich problemlos durchlaufen kann, verursachen die weiten Sprünge des *Random Walks* eine niedrige Akzeptanzrate (11,22%) und die Kette bleibt dadurch für lange Zeit am selben Punkt stehen (siehe Abbildung 3.7, untere Grafik). Dies an den waagrechten Strecken erkennbar und führt zu der zackigen Gestalt der approximierenden Dichte.

Schließlich zeigt die mittlere Grafik in Abbildung 3.6 die Ergebnisse einer adäquat gewählten Streuung der Kandidaten generierenden Funktion. Die Akzeptanzrate von 37,88% steht im Einklang mit den in der Literatur vorgeschlagenen Werten von $25\% - 50\%$ für univariate Ketten (vgl. Roberts (1996)). Die Ergebnisse des Metropolis-Hastings-Algorithmus und der Zufallsziehung sind in etwa vergleichbar. Die dazugehörige Markov-Kette (siehe Abbildung 3.7, mittlere Grafik) weist ein schnelles *mixing* auf. Dies ist dadurch erkennbar, dass hohe Sprünge, jedoch keine waagrechten Strecken zu beobachten sind.

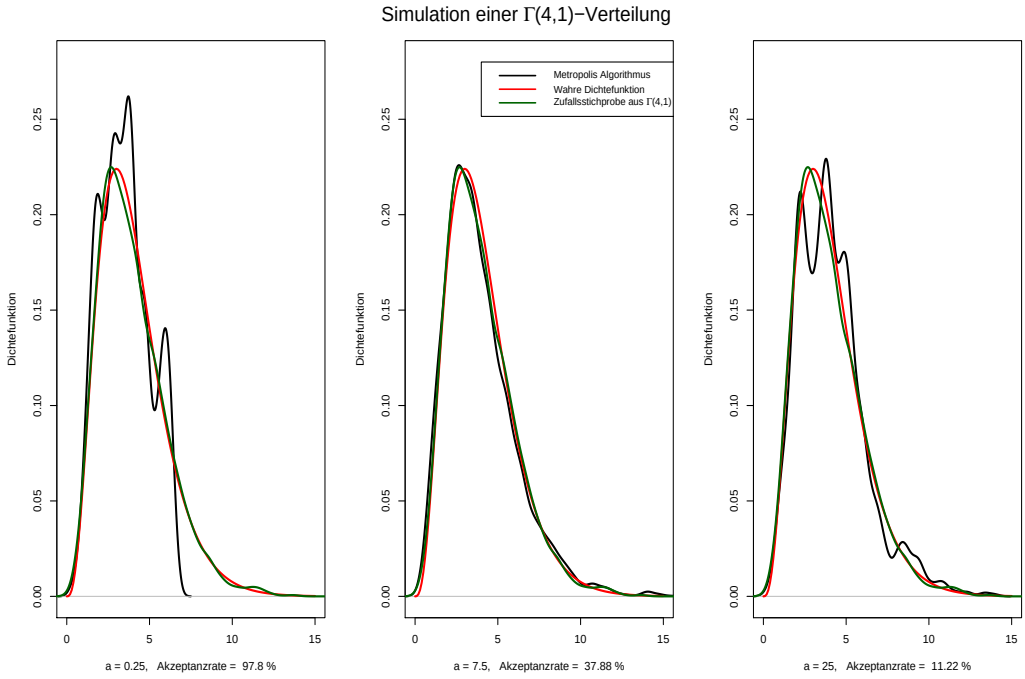

Abbildung 3.6: Simulation einer $\Gamma(4,1)$-Verteilung mit dem *Random-Walk*-Metropolis-Algorithmus mit gleichverteilter Kandidaten generierender Funktion h für drei verschiedene Werte des Parameters a.

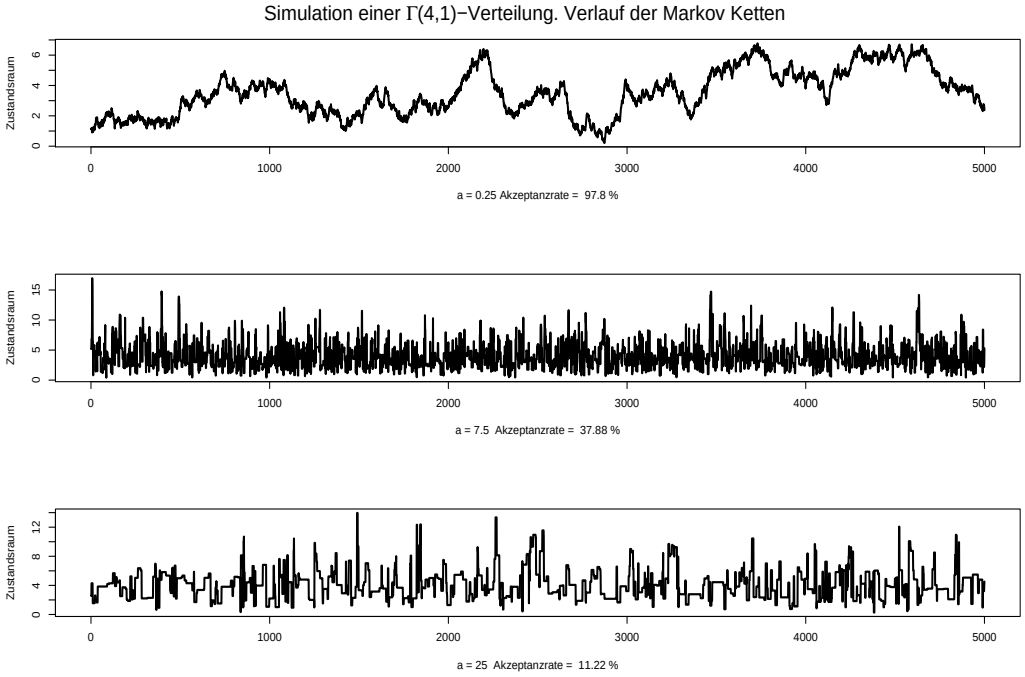

Abbildung 3.7: Verlauf der Markov-Ketten bei der Simulation einer $\Gamma(4,1)$-Verteilung mit dem *Random-Walk*-Metropolis-Algorithmus mit gleichverteilter Kandidaten generierender Funktion h für drei verschiedene Werte des Parameters a.

***Independence-Sampler*:** Diese Klasse von Kandidaten generierenden Funktionen wurde in Hastings (1970) eingeführt. Der Name *Independence-Sampler* geht auf Tierney (1994) zurück. Im Gegensatz zum *Random-Walk* ist die Generierung eines Kandidaten zum Zeitpunkt $t+1$ unabhängig vom Zustand des Prozesses im Zeitpunkt t. D.h. die Kandidaten generierende Funktion hat die Form

$$f(x,y) = h(y),$$

wobei h eine Dichtefunktion ist. Die Akzeptanzwahrscheinlichkeit hat somit folgende Struktur

$$\alpha(x,y) = \begin{cases} \min\left\{\dfrac{q(y)}{q(x)}, 1\right\}, & \text{für } q(x) > 0 \\ 1, & \text{sonst} \end{cases} \tag{3.24}$$

mit $q := \frac{\pi(x)}{h(x)}$.

Brooks (1998) weist darauf hin, dass q als das *Importance*-Gewicht interpretiert werden kann, welches von einem *Importance-Sampler* (siehe Abschnitt 3.2.2.2 auf Seite 38) benutzt wird, um Ziehungen aus π mit Hilfe der Funktion h zu generieren.

In der Tat benutzen beide Methoden Gewichte, um die Wahrscheinlichkeitsmasse auf gewissen Punkten bzw. Intervallen aufzubauen. Sie unterscheiden sich jedoch in der Art, wie sie das bewerkstelligen:

Importance-Sampling

- Punkte mit einem großen Gewicht werden relativ häufig gewählt.

Independence-Sampler von Metropolis-Hastings

- Die Kette verbleibt für lange Zeitintervalle auf Zustandspunkten mit hohen Gewichten.

Es ist interessant, dass im Falle eines *Independence-Samplers* die Kandidaten generierende Funktion keine Markov*sche* Struktur aufweist. Der Übergangskern setzt sich allerdings aus der Kandidaten generierenden Funktion h und der Akzeptanzwahrscheinlichkeit α zusammen, die vom Wert des vorigen Zeitpunkts abhängig ist. Dadurch wird doch noch die Markov-Eigenschaft induziert (vgl. Gamerman (1997, S. 168)).

Schließlich sei darauf hingewiesen, dass der *Independence Sampler* dazu neigt, ein polarisiertes Verhalten aufzuweisen. D.h. er funktioniert entweder sehr gut oder sehr schlecht. Dieses Verhalten hängt von der Ähnlichkeit zwischen der Kandidaten generierenden Funktion und der Zielfunktion ab, welche durch die Größe $\inf_x \frac{h(x)}{\pi(x)}$ ausgedrückt wird (vgl. Roberts (1996, S. 54)). Deshalb sollte die Kandidaten generierende Funktion h so gewählt werden, dass sie die Dichtefunktion π gut approximieren kann. Wenn dies jedoch nicht möglich ist, sollten andere Algorithmen herangezogen werden. In diesem Zusammenhang kommentiert Roberts (1996, S. 55):

> $[\cdots]$ „*it is rare for the Independence-Sampler to be useful as a stand-alone algorithm. However, within a hybrid strategy which combines and mixes different Markov Chain Monte Carlo methods, this method is extremely easy to implement and often very effective.*“

Beispiel (Fortsetzung): Nun wird die $\Gamma(4,1)$-Verteilung aus dem letzten Beispiel mittels des *Independence-Sampler*-Metropolis-Hastings-Algorithmus simuliert. Zwei unterschiedliche Kandidaten generierende Verteilungen werden in Betracht gezogen:

(1) Lognormal(1,34, 0,55)

(2) Lognormal(2,2, 0,35)

Die Form der Kandidaten generierenden Verteilungen und der Zielverteilung $\Gamma(4,1)$ sind in Abbildung 3.3 auf Seite 41 zu sehen. Offensichtlich sind sich die Verteilung (1) und die Zielverteilung relativ ähnlich während die Verteilung (2) die Wahrscheinlichkeitsmasse anders verteilt. Wie oben aufgeführt ist die Ähnlichkeit zwischen der Kandidaten generierenden Funktion und der Zielfunktion maßgeblich für die Leistung des *Independence-Sampler*-Algorithmus. Die linke Grafik in Abbildung 3.8 zeigt die Approximation mittels der Verteilung (1). Die Ähnlichkeit beider Verteilungen ermöglicht eine gute Approximation. Die hohe Akzeptanzrate von 91,08% ist eine direkte Konsequenz dieser Tatsache und anders als beim *Random-Walk*-Algorithmus als geeignetes Anpassungskriterium zu interpretieren. Auch die dazugehörige Markov-Kette (siehe Abbildung 3.9, obere Grafik) zeigt keine erkennbaren waagrechten Strecken, die typisch für die Ablehnung der Kandidaten für mehrere aufeinanderfolgende Iterationen sind.

Die rechte Grafik in Abbildung 3.8 zeigt die Approximation der Zielverteilung mittels der Verteilung (2). Die zackige Gestalt der approximierenden Dichte ist typisch für die konsekutive Ablehnung der Kandidaten. Denn wenn die Markov-Kette an einem Punkt verbleibt, erhöht sich die Wahrscheinlichkeitsdichte für diesen Punkt. Schließlich bestätigt die Betrachtung der unteren Grafik in Abbildung 3.9 diese Behauptung. Der Prozeß weist lange waagrechte Strecken auf. Besonders auffällig ist die erste Strecke, die sich über mehr als 1000 Iterationen zieht.

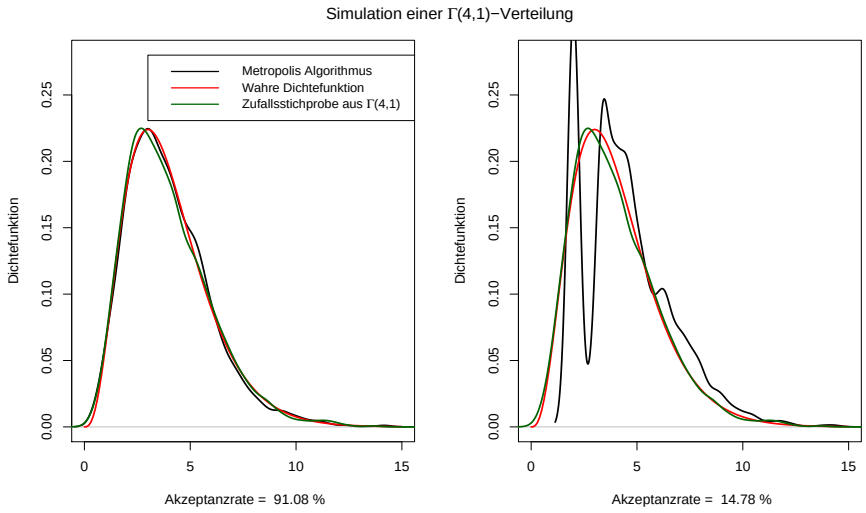

Abbildung 3.8: Vergleich der approximierenden Dichten bei der Simulation einer $\Gamma(4,1)$-Verteilung mit dem *Independence-Sampler*-Metropolis-Algorithmus.

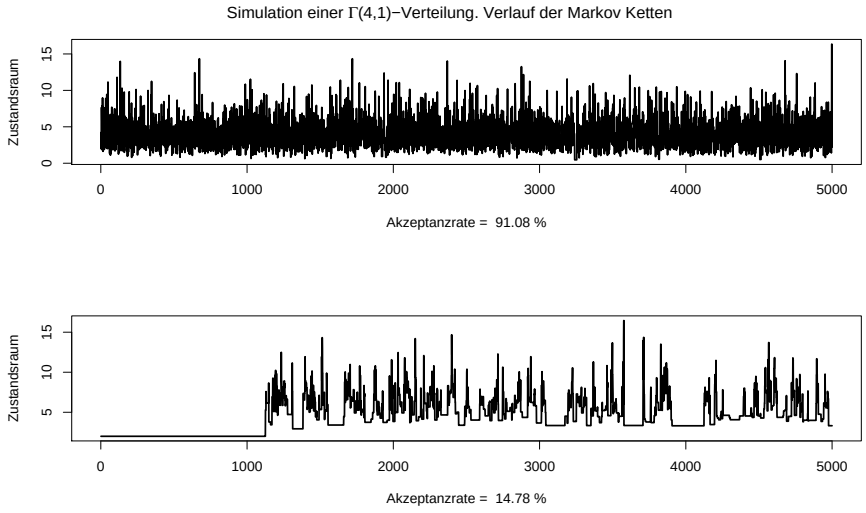

Abbildung 3.9: Verlauf der Markov-Ketten bei der Simulation einer $\Gamma(4,1)$-Verteilung mit dem *Independence-Sampler*-Metropolis-Algorithmus für zwei verschiedene Lognormal-Verteilungen als Kandidaten generierende Funktionen.

Aufgrund des erörterten Nachteils wird im Folgenden der *Independence-Sampler*-Metropolis-Hastings-Algorithmus nicht für die Simulation von Verteilungen eingesetzt.

3.4 *Gibbs Sampler* und *Data Augmentation*-Algorithmus

Die in diesem Abschnitt vorzustellenden MCMC-Algorithmen weisen Ähnlichkeiten auf, die ihre gemeinsame Behandlung rechtfertigen[20].

Um das Verständnis der Grundgedanken hinter den Algorithmen zu erleichtern, ist es jedoch sinnvoll, zunächst eine mathematische Methode zur numerischen Lösung von Fixpunkt-Integralgleichungen einzuführen, welche unter dem Namen Substitutionsmethode bekannt ist (vgl. Rall (1969, S. 64-74)).

3.4.1 Von bedingten Verteilungen zu Randverteilungen: Die Substitutionsmethode

Die Substitutionsmethode bildet die Grundlage der beiden MCMC-Algorithmen *Gibbs Sampler* von Geman und Geman (1984) und *Data Augmentation*-Algorithmus von Tanner und Wong (1987). Bei der Substitutionsmethode handelt es sich um die iterative Lösung einer Fixpunkt-Integralgleichung, d.h. einer Gleichung der Form

$$f(x) = \int k(x,t)f(t)\,\mathrm{d}t, \tag{3.25}$$

wobei $k(\cdot,\cdot)$ ein geeigneter Kern[21] ist.

Falls die Gleichung (3.25) eine eindeutige Lösung besitzt, muss diese ein Fixpunkt sein, denn die Funktion f kommt an beiden Seiten der Gleichung vor.

Es sei nun angenommen, dass der folgendermaßen definierte Integraloperator T

$$Tf(x) := \int k(x,t)f(t)\,\mathrm{d}t$$

für integrierbare Funktionen f existiert und selbst integrierbar ist. Eine Lösung der Integralgleichung kann dann wie folgt erhalten werden (vgl. Tanner und Wong (1987, S. 530)):

1. Eine Approximation $f_0(x)$ für $f(x)$ wird gewählt.
2. Für $i \in \mathbb{N}$ wird sukzessive

   $$f_{i+1}(x) = Tf_i(x) \tag{3.26}$$

   berechnet, bis ein Konvergenzkriterium erreicht wird.

Die Iterationsvorschrift in (3.26) definiert eine Fixpunktgleichung, deren eindeutige Lösung f ist (vgl. Gelfand und Smith (1990)). Der Name Substitutionsmethode deutet auf die substitutive Natur des Algorithmus hin.

[20] In der Tat haben sie auch gewisse Ähnlichkeiten mit dem in Kapitel 2 vorgestellten EM-Algorithmus.
[21] Zur allgemeinen Theorie der Integralgleichungen siehe Pipkin (1991).

Es war die Leistung von Tanner und Wong (1987) zu erkennen, dass dieser Algorithmus zur Simulation von (Rand-) Verteilungen eingesetzt werden kann, wenn bestimmte bedingte Verteilungen vorliegen. Diese Beobachtung hat zur Entwicklung des *Data Augmentation*-Algorithmus geführt, der sehr erfolgreich bei der Betrachtung und Analyse von fehlenden Daten eingesetzt wird (vgl. Schafer (1997) bzw. Little und Rubin (2002)). Ähnliche Erkenntnisse haben den *Gibbs Sampler* motiviert, welcher auf der Arbeit von Besag (1974) basierend, 1984 von Geman und Geman für die bayesianische Rekonstruktion von Bildern eingeführt wurde (siehe Geman und Geman (1984)). Beide Algorithmen sind Gegenstand der nächsten beiden Abschnitte.

3.4.2 Die Substitutionsmethode für Dichtefunktionen im bivariaten Fall

Gegeben seien zwei Zufallsvariable X und Y mit unbekannter gemeinsamer Verteilung $f_{X,Y}(x,y)$ und bekannten bedingten Verteilungen $f_{X|Y}(x|y)$ und $f_{Y|X}(y|x)$. Für die Randverteilung $f_X(x)$ von X gilt also

$$\begin{aligned} f_X(x) &= \int f_{X,Y}(x,y)\,\mathrm{d}y \\ &= \int f_{X|Y}(x|y) f_Y(y)\,\mathrm{d}y. \end{aligned} \tag{3.27}$$

Analog erhält man für

$$f_Y(y) = \int f_{Y|X}(y|x) f_X(x)\,\mathrm{d}x. \tag{3.28}$$

Substitution von (3.28) in (3.27) ergibt

$$f_X(x) = \int f_{X|Y}(x|y) \int f_{Y|X}(y|t) f_X(t)\,\mathrm{d}t\,\mathrm{d}y.$$

Der Satz von Tonelli ermöglicht die Vertauschung der Integrationsgrenzen

$$f_X(x) = \int \left[\int f_{X|Y}(x|y) f_{Y|X}(y|t)\,\mathrm{d}y \right] f_X(t)\,\mathrm{d}t.$$

Der Term in eckigen Klammern kann als Übergangskern (siehe dazu Abschnitt 3.3.2) von x nach x (vertreten durch die *dummy* Variable t) angesehen werden. Es gilt also

$$k(x,t) = \int f_{X|Y}(x|y) f_{Y|X}(y|t)\,\mathrm{d}y \tag{3.29}$$

und somit

$$f_X(x) = \int k(x,t) f_X(t)\,\mathrm{d}t. \tag{3.30}$$

D.h. man erhält eine Fixpunkt-Integralgleichung des Typs (3.25).

Neben der von Tanner und Wong verwendeten funktionalanalytischen Argumentation ist es möglich, mit Hilfe bekannter Eigenschaften der Markov-Ketten die Methode zu motivieren, indem man Gleichung (3.29) mit der in Abschnitt A.1 auf Seite 148 dargestellten Chapman-Kolmogorov-Gleichung vergleicht. In der Tat handelt es sich beim Kern $k(\cdot,\cdot)$ um die zustandskontinuierliche Version der Gleichung (A.1), welche die k-Schritt-Übergangsmatrix des Markov-Prozesses darstellt.

3.4.3 *Substitution-Sampling*

Nun wird die Substitutionsmethode[22] zur Simulation von Dichtefunktionen in algorithmischer Form dargestellt. Dabei blieben die Bedingungen von Abschnitt 3.4.2 erhalten, d.h. gesucht seien die Randverteilungen f_X und f_Y und die bedingten Verteilungen $f_{X|Y}$ und $f_{Y|X}$ seien als bekannt vorausgesetzt:

1. Ein Startwert x^0 wird aus einer Startdichtefunktion f_X^0 gewählt. Die Wahl eines nicht-zufälligen Startwertes stellt keine Einschränkung der Methode dar.

2. Bedingt auf den Startwert wird ein Wert y^1 aus $f_{Y|X}(y|x^0)$ gezogen. Dieser Wert hat eine Randdichte, welche durch

$$f_Y^1(y) = \int f_{Y|X}(y|x^0) f_X^0(x) \mathrm{d}x$$

   gegeben ist.

3. Anschließend wird ein Wert aus $f_{X|Y}(x|y^1)$ gezogen, dessen Randdichte

$$f_X^1(x) = \int f_{X|Y}(x|y^1) f_Y^1(y) \mathrm{d}x = \int k(x,t) f_X^0(t) \mathrm{d}t$$

   ist (vgl. dazu (3.30)). Analog erhält man die Randdichte von Y.

4. Die Wiederholung dieses Schemas führt zu einer Folge $(x^i, y^i)_{i\in\mathbb{N}}$ mit der Eigenschaft

$$F_X^i \xrightarrow[i\to\infty]{d} F_X \quad \text{und} \quad F_Y^i \xrightarrow[i\to\infty]{d} F_Y.$$

Im bivariaten Fall stellt diese Vorgehensweise die Basis sowohl des *Data Augmentation*-Algorithmus als auch des *Gibbs Sampler*s dar. D.h. im bivariaten Fall sind beide Algorithmen identisch.

Mit Hilfe funktionalanalytischer Hilfsmittel haben Tanner und Wong (1987) zahlreiche Eigenschaften des Substitutionsalgorithmus zur Simulation von Dichtefunktionen bewiesen. Diese werden in Abschnitt 3.4.6.2 aufgelistet.

[22] Die Verwendung der Substitutionsmethode zur Simulation von Dichtefunktionen wird zuweilen auch als *Substitution-Sampling* bezeichnet.

3.4.4 Mehr als zwei Variable

Auf analoge Weise kann der Substitutionsalgorithmus auf k Variable verallgemeinert werden. Z.B. erhält man für den Fall dreier Zufallsvariablen X, Y und Z:

$$f_X(x) = \iint f_{X,Y,Z}(x,y,z)\,\mathrm{d}y\,\mathrm{d}z \tag{3.31}$$

$$= \iint f_{X|YZ}(x|y,z) f_{YZ}(y,z)\,\mathrm{d}y\,\mathrm{d}z, \tag{3.32}$$

wobei die Zufallsvariablen Y und Z wie eine Einheit behandelt werden. Aus

$$f_{YZ}(y,z) = \int f_{YZ|X}(y,z|x) f_X(x)\,\mathrm{d}x$$

erhält man durch Einsetzen in (3.32)

$$f_X(x) = \iint f_{X|YZ}(x|y,z) \int f_{YZ|X}(y,z|x) f_X(x)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$$

und die Vertauschung der Integrationsgrenzen führt schließlich zum Ausdruck

$$f_X(x) = \int \left[\iint f_{X|YZ}(x|y,z) f_{YZ|X}(y,z|t)\,\mathrm{d}y\,\mathrm{d}z \right] f_X(t)\,\mathrm{d}t. \tag{3.33}$$

Obwohl die Substitution nun zu einem anderen Integraloperator als in (3.29) führt, hat die Integralgleichung als eindeutige Lösung wieder $f_X(x)$. Dieses Beispiel zeigt, dass es mehrere Möglichkeiten gibt, eine Fixpunkt-Integralgleichung aufzustellen, welche $f_X(x)$ als eindeutige Lösung besitzt. Der *Data Augmentation*-Algorithmus und der *Gibbs Sampler* unterscheiden sich hauptsächlich in der Wahl der bedingten Verteilungen, die zur Konstruktion der Integralgleichung verwendet werden, und sind unter gewissen Bedingungen äquivalent zueinander. Dies wird in Abschnitt 3.4.6.3 gezeigt.

3.4.5 *Gibbs Sampler*

Die wohl bekannteste MCMC-Methode ist der *Gibbs Sampler* von Geman und Geman (1984), der im Kontext der Bildrekonstruktion entwickelt wurde und schnell eine weite Verbreitung in verwandten Bereichen wie z.B. Neuronale Netze und Expertensysteme gefunden hat. Diese Methoden sind dadurch gekennzeichnet, dass sie aus einer sehr hohen Anzahl an Variablen bestehen, deren gemeinsame Verteilung aufgrund ihrer Komplexität nicht angegeben werden kann. Ein Umweg bietet sich durch die Verwendung von bedingten Verteilungen an, welche häufig in solchen Modellen nur von einer kleinen Anzahl an Variablen abhängen und somit entweder in geschlossener Form dargestellt oder effizient simuliert werden können. In den neunziger Jahren haben die Arbeiten von Gelfand und Smith (1990) und Casella und George (1992) die Anwendbarkeit des *Gibbs Sampler*s für eine breite Palette an statistischen Fragestellungen hervorgehoben und somit einen wichtigen Beitrag zur Verbreitung der MCMC-Methoden geleistet.

In den letzten Jahren hat der *Gibbs Sampler* zunehmend an Verbreitung gewonnen und in erheblichem Maße zum Aufblühen der Bayes-Statistik beigetragen[23]. Gilks et al. (1996) zufolge,

[23] Die frei erhältliche Software BUGS (WinBUGS), deren Name für „Bayes Inference Using Gibbs Sampling“ steht, wird erfolgreich in verschiedenen Bereichen eingesetzt. Siehe dazu http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

verwenden heutzutage die meisten auf MCMC-Verfahren basierenden statistischen Anwendungen den *Gibbs Sampler*.

Da der Grundstein zum Verständnis des *Gibbs Sampler*s bereits durch die Behandlung des *Substitution-Sampling*-Algorithmus gelegt wurde, ist es ausreichend, das zugehörige Aktualisierungsschema zu betrachten.

3.4.5.1 Aktualisierungsschema des *Gibbs Sampler*s

Der *Gibbs Sampler* zeichnet sich dadurch aus, dass der Übergangskern immer mit Hilfe aller voll-bedingten Verteilungen[24] konstruiert wird. Dass die Menge aller voll-bedingten Verteilungen unter gewissen schwachen Regularitätsbedingungen ausreichend ist, um die gemeinsame Dichte zu charakterisieren, ist ein wichtiges Ergebnis der Stochastik, welches bereits 1974 von Besag untersucht wurde. Neben der wegweisenden Arbeit von Geman und Geman (1984) haben Roberts und Smith (1994) und Tierney (1994) formale Beiträge zum Verständnis der Konvergenzeigenschaften dieses Aktualisierungsschemas geleistet.

Gegeben sei ein k-dimensionaler Zufallsvektor $\mathbf{X} := (X_1, \ldots, X_k)$ dessen o.B.d.A. stetigen Randverteilungen unbekannt sind. Die voll-bedingten Verteilungen $f_{X_i}(x_i|\mathbf{X}_{-i})$ seien jedoch bekannt, wobei $\mathbf{X}_{-i}$ der Zufallsvektor ist, der aus $\mathbf{X}$ entsteht, wenn die i-te Komponente entfernt wird.

Das Gibbs*sche* Aktualisierungsschema besitzt folgende Struktur:

1. Startwerte $x_1^0, \ldots, x_k^0$ werden gewählt.
2. Ein Wert x_1^1 wird aus der Dichte $f_{X_1|\mathbf{X_{-1}}}(\cdot|x_2^0, \ldots, x_k^0)$ gezogen.
3. Unter Verwendung von x_1^1 wird ein Wert x_2^1 aus der Dichte $f_{X_2|\mathbf{X_{-2}}}(\cdot|x_1^1, x_3^0, \ldots, x_k^0)$ gezogen.
4. Alle Komponenten werden auf gleiche Art und Weise aktualisiert. Im k-ten Aktualisierungsschritt entsteht dadurch ein vollständiger *Gibbs*-Zyklus.
5. Das Schema wird fortgesetzt bis ein Stoppkriterium erreicht wird.

[24] Aus dem Englischen *full conditionals*. Eine voll-bedingte Verteilung ist die bedingte Verteilung der i-ten Komponente eines Zufallsvektors gegeben alle anderen Komponenten.

3.4.5.2 Eigenschaften des *Gibbs Sampler*s

In ihrem einflussreichen Diskussionsbeitrag zu MCMC-Verfahren listen Gelfand und Smith (1990) folgende Eigenschaften des *Gibbs Sampler*s auf:

1. (Konvergenz) Sei $X_i,\ i \in \{1, \ldots, k\}$ die i-te Zufallsvariable eines k-dimensionalen Zufallsvektors, deren Dichtefunktion mit Hilfe des *Gibbs Sampler*s geschätzt werden soll. Ferner bezeichne X_i^j die j-te Iteration des *Gibbs Sampler*s. Dann gilt

   $$X_i^j \xrightarrow[j\to\infty]{\mathrm{d}} X_i \quad \text{für} i = 1, \ldots, k.$$

   D.h. für großes j konvergiert X_i^j in Verteilung gegen X_i und zwar unabhängig von der Reihenfolge, in der die Variablen besucht werden, solange sie unendlich oft besucht werden können[25]. Diese Bedingung ist relativ schwach und in aller Regel gewährleistet.

2. (Rate) Die Verteilung von $(X_1^j, \ldots, X_k^j)$ konvergiert für $j \to \infty$ gegen die wahre Verteilung $(X_1, \ldots, X_k)$ geometrisch[26] in der Supremum Norm für die natürliche Reihenfolge, d.h. $1 \to 2, \ldots, k \to 1$. Die Konvergenzrate bedarf einer kleinen Korrektur für andere *io*-Besuchschemata.

3. (Ergodensatz) Es sei $g(X_1, \ldots, X_k)$ eine messbare Funktion und der Erwartungswert von $g(X_1, \ldots, X_k)$ existiere. Dann gilt

   $$\lim_{m\to\infty} \frac{1}{m} \sum_{j=1}^{m} g(X_1^j, \ldots, X_k^j) \xrightarrow{\text{f.s.}} \mathrm{E}[g(X_1, \ldots, X_k)].$$

   D.h. das arithmetische Mittel von $g(X_1, \ldots, X_k)$ konvergiert mit Wahrscheinlichkeit eins gegen ihren Erwartungswert.

3.4.5.3 *Gibbs Sampler* als Sonderfall des Metropolis-Hastings-Algorithmus

Obwohl *Gibbs Sampler* und der Metropolis-Hastings Algorithmus Strukturen besitzen, die auf den ersten Blick sehr unterschiedlich erscheinen, kann gezeigt werden, dass der *Gibbs Sampler* als Instanz des Metropolis-Hastings-Algorithmus aufgefasst werden kann. Der Vollständigkeit halber wird dieses Resultat in Anlehnung an Brooks (1998, S. 73) im Folgenden vorgestellt.

Sei $\mathbf{X} = (X_1, \ldots, X_k),\ k \in \mathbb{N}$ ein k-dimensionaler Zufallsvektor und X_i die i-te Komponente von X, die mittels eines Metropolis-Hastings-Algorithmus aktualisiert werden soll. Ferner sei mit $\mathbf{X}_{-i}$ wieder der Vektor bezeichnet, der aus $\mathbf{X}$ entsteht, wenn die i-te Komponente entfernt wird. $\mathbf{X}$ sei also darstellbar als $\mathbf{X} = (X_i, X_{-i})$. Außerdem sei folgende Kandidaten generierende Funktion definiert:

[25] In diesem Zusammenhang bedeutet die Aussage, dass die Komponente eines Zufallsvektors *besucht* wird, eine Ziehung aus ihrer Dichte- bzw. Wahrscheinlichkeitsfunktion. Schemata, bei denen Zufallsvariable unendlich oft besucht werden, werden fortan als *io*-Schemata bezeichnet (*io*: *infinitely often*).

[26] In der Fachliteratur zu MCMC werden i.d.R. zwei Konvergenzarten betrachtet: gleichmässige und geometrische Konvergenz. Die geometrische Konvergenz ist schwächer als die gleichmässige Konvergenz und ist vom Startwert abhängig.

$$k(x,y) = \begin{cases} \pi(y_i|x_{-i}) & y_{-i} = x_{-i}, \text{ für } i = 1,\ldots,k \\ 0 & \text{sonst.} \end{cases}$$

Mit dieser Kandidaten generierenden Funktion ist die resultierende Akzeptanzwahrscheinlichkeit gleich 0, falls $y_{-i} \neq x_{-i}$ gilt, und

$$\begin{aligned} \alpha(x,y) &= \frac{\pi(y)k(y,x)}{\pi(x)k(x,y)} \\ &= \frac{\pi(y)/\pi(y_i|x_{-i})}{\pi(x)/\pi(x_i|y_{-i})} \quad \text{für } y_{-i} = x_{-i}. \end{aligned}$$

Die Tatsache, dass $y_{-i} = x_{-i}$ gilt, impliziert weiter

$$\alpha(x,y) = \frac{\pi(y)/\pi(y_i|y_{-i})}{\pi(x)/\pi(x_i|x_{-i})}.$$

Mit $y = (y_i, y_{-i})$ folgt $\pi(y) = \pi(y_{-i})\pi(y_i|y_{-i})$. D.h. man erhält

$$\alpha(x,y) = \frac{\pi(y_{-i})}{\pi(x_{-i})}$$

und wegen $y_{-i} = x_{-i}$ folgt schließlich $\alpha(x,y) = 1$.

Es gilt somit insgesamt

$$\alpha(x,y) = \begin{cases} 0, & \text{für } y_{-i} \neq x_{-i} \\ 1, & \text{für } y_{-i} = x_{-i}. \end{cases}$$

Der Prozeß kann also in jeder Iteration nur Zustände y besuchen, die gleich x in allen Komponenten bis auf die i-te sind, und diese Bewegungen werden mit Wahrscheinlichkeit $\alpha = 1$ akzeptiert. Dies zeigt, dass es sich bei diesem besonderen Metropolis-Hastings-Algorithmus um den *Gibbs Sampler* handelt, was zu zeigen war.

3.4.5.4 Beispiel: *Gibbs Sampler* zur Ziehung aus einer bivariaten Normalverteilung mit bekannten Parametern

Das folgende Beispiel dient lediglich dem Zweck, die Funktionsweise des *Gibbs Sampler*s grafisch darzustellen. Denn die unabhängige Ziehung aus einer bivariaten Normalverteilung ist relativ unproblematisch und es gibt zahlreiche Algorithmen, die dies effizient bewerkstelligen können. Die Generierung einer bestimmten Korrelationsstruktur erfolg in aller Regel durch Multiplikation unabhängiger Ziehungen mit der Choleski-Zerlegung der Ziel-Varianz-Kovarianz-Matrix. Dennoch ist es aus didaktischen Gründen interessant, sich mit einer multivariaten Verteilung zu beschäftigen, deren bedingte Verteilungen und Randverteilungen zur gleichen Verteilungsfamilie gehören.

Die Ausgangssituation sei gegeben durch eine bivariat normalverteilte Zufallsvariable Y mit bekanntem Parametervektor θ bestehend aus $\mu_1 = 1, \mu_2 = 0{,}5, \sigma_1^2 = 1, \sigma_2^2 = 1$ und $\rho = 0.5$. Gesucht seien Ziehungen aus dieser bivariaten Verteilung.

Bekannt seien die beiden bedingten Verteilungen:

$$f_{Y_1|Y_2=y_2;\theta} \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2);\ \sigma_1^2(1-\rho^2)\right)$$

und

$$f_{Y_2|Y_1=y_1;\theta} \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1 - \mu_1);\ \sigma_2^2(1-\rho^2)\right).$$

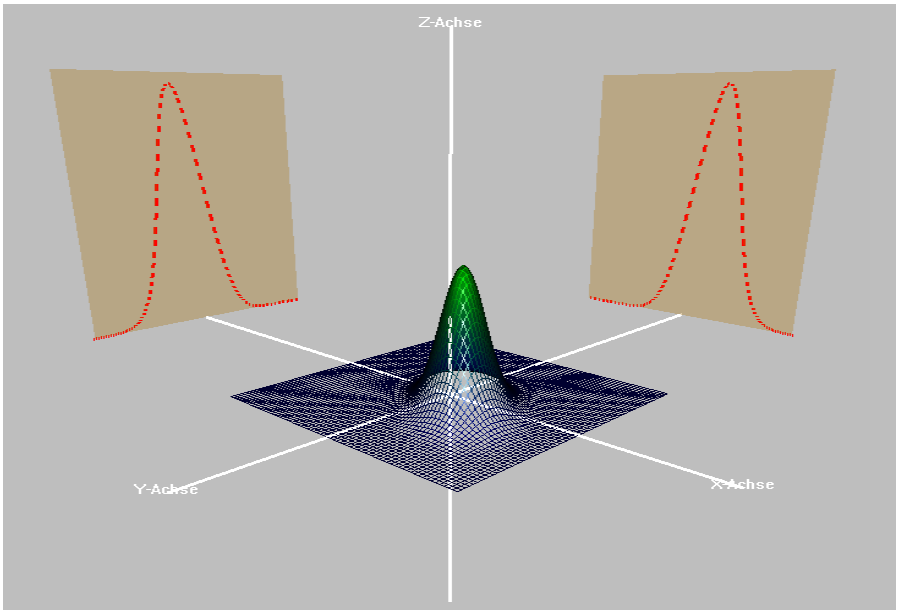

Abbildung 3.10: Beginn der Simulation: Gesucht sind Ziehungen aus der bivariaten Verteilung in der Mitte der Grafik. Zur zusätzlichen Kontrolle sind beide Randverteilungen eingezeichnet (rote, gestrichelte Linien). Es sei angemerkt, dass weder direkt aus der bivariaten Verteilung noch aus den Randverteilungen gezogen wird und sie nur dem Vergleich mit den Simulationsergebnissen dienen.

Aufgrund der Tatsache, dass der Parametervektor θ als bekannt vorausgesetzt wird, stellen die beiden Verteilungen $f_{Y_1|Y_2=y_2;\theta}$ und $f_{Y_2|Y_1=y_1;\theta}$ den ganzen Satz an voll-bedingten Verteilungen dar. Somit sind die Voraussetzungen für die Anwendung des *Gibbs Sampler*s erfüllt.

Zunächst wird ein Startwert benötigt, der im Prinzip beliebig sein kann, vorausgesetzt, dass seine Wahrscheinlichkeitsdichte in der Zielverteilung größer als Null ist[27]. Es sei o.B.d.A. angenommen, dass y_1^0 der Startwert aus der Verteilung von Y_1 ist. Für die erste Ziehung wird dieser Wert als fest angenommen, womit die Verteilung $f_{Y_2|Y_1=y_1^0;\theta}$ vollständig charakterisiert ist. Aus dieser bedingten Verteilung wird wiederum ein Wert y_2^1 gezogen, welcher die Basis für eine Ziehung aus $f_{Y_1|Y_2=y_2^1;\theta}$ darstellt. Dieser Vorgang wird wiederholt bis genügend Beobachtungen vorliegen, um die bivariate Verteilung zu charakterisieren.

[27] Genau genommen kommen noch andere Anforderungen hinzu, die jedoch an dieser Stelle nicht näher betrachtet werden.

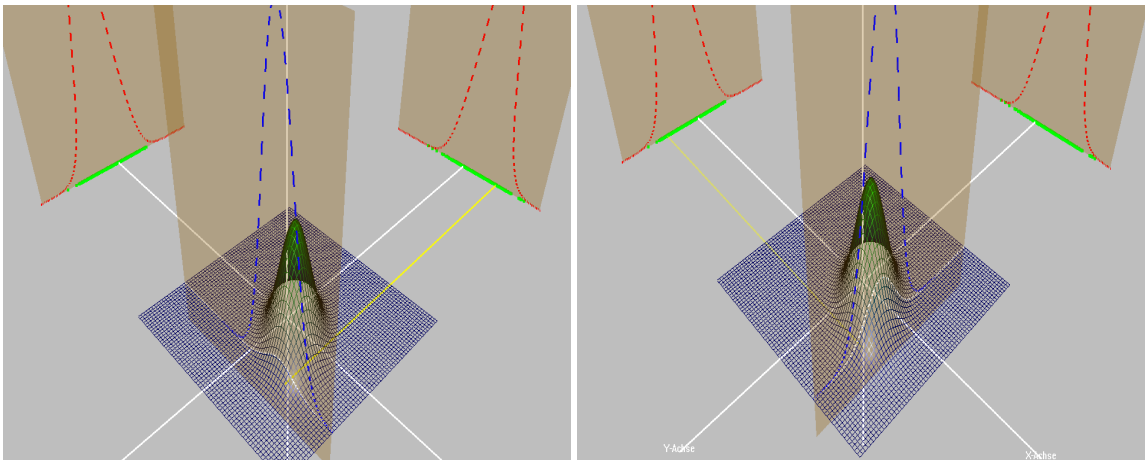

Abbildung 3.11: Ziehung aus den bedingten Verteilungen. In der rechten Grafik ist erkennbar, dass sich die bedingte Verteilung (in einen orangenen Rahmen eingebettet) genau auf dem in der linken Grafik gezogenen Punkt befindet, der durch die gelbe Linie farblich hervorgehoben wird.

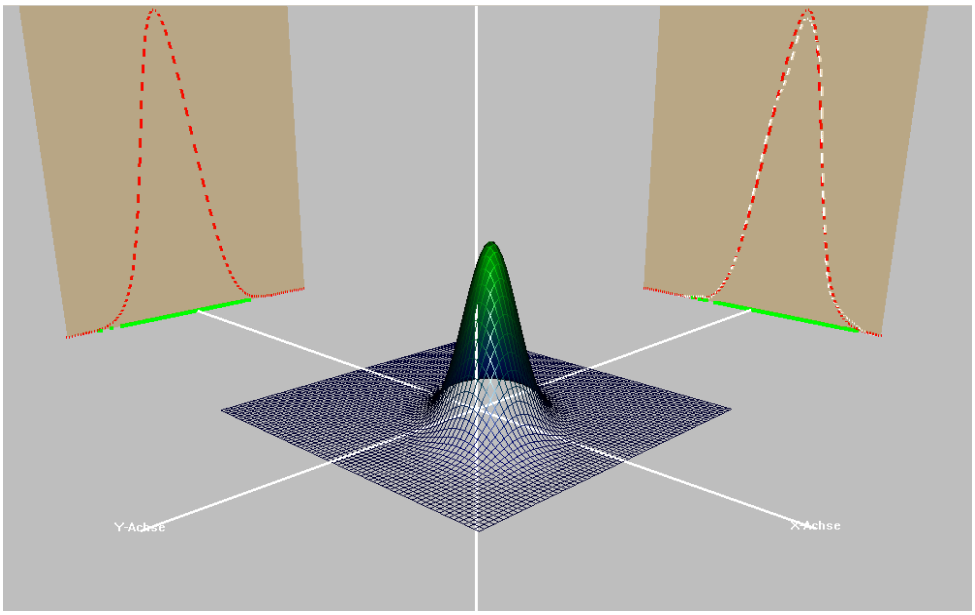

Abbildung 3.12: Erste Randverteilung: Die gestrichelte weiße Linie zeigt die Kerndichteschätzung der ersten 500 Ziehungen.

Abbildung 3.10 veranschaulicht die Zielverteilung zusammen mit ihren beiden Randverteilungen. Vor Beginn der Simulation sind keine bedingten Verteilungen eingezeichnet. Der Vorgang ist in Abbildung 3.11 bei einem fortgeschrittenen Simulationsstand abgebildet, welcher durch die grünen Punkte unterhalb der Randdichten erkennbar ist. Beide Bilder stellen zwei aufeinander folgende Schritte der Simulation grafisch dar. Die zweite bedingte Verteilung liegt genau auf dem zuletzt gezogenen Wert, wodurch die Nachbildung der Korrelationsstruktur mit wachsendem Stichprobenumfang bewerkstelligt wird.

Nach 500 Iterationen wird die Zielverteilung mit der Kerndichteschätzung der gezogenen Stichprobe verglichen. Außerdem werden die Randverteilungen eingezeichnet, wie in Abbildungen 3.12, 3.13 und 3.14 zu sehen ist. Trotz der kleinen Abweichungen zwischen der bivariaten theoretischen

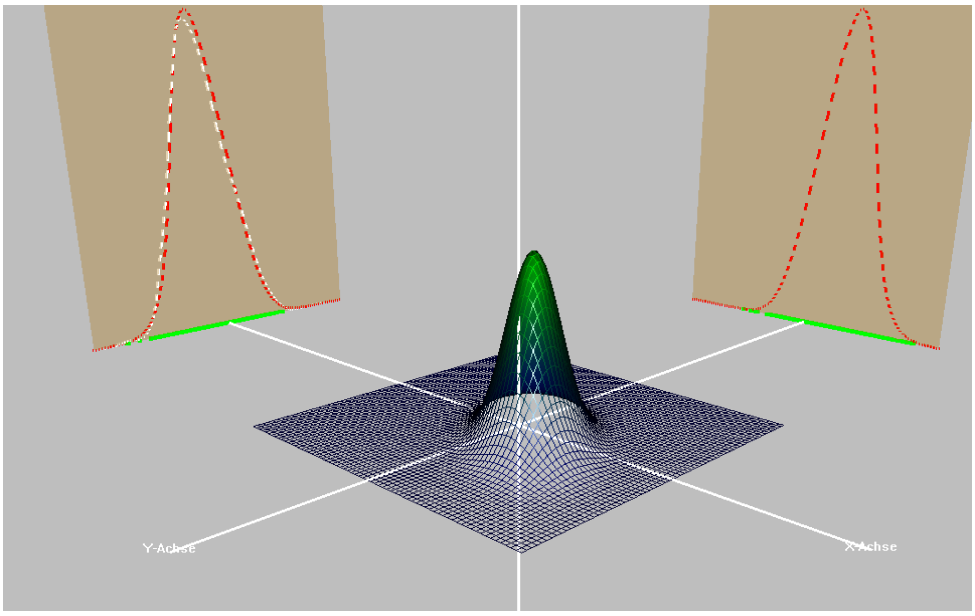

Abbildung 3.13: Zweite Randverteilung: Nach ca. 500 Ziehungen ist die Approximation der Randdichte recht gut.

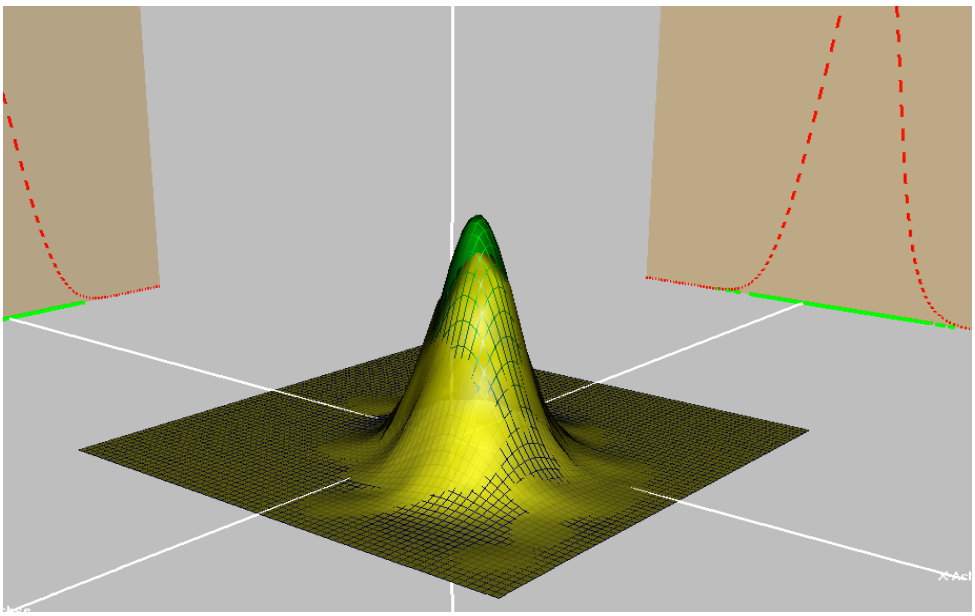

Abbildung 3.14: Approximation der bivariaten Verteilung: Das Gelbe Flächendiagramm ist die Kerndichteschätzung der ersten 500 Paare von Ziehungen aus den bedingten Verteilungen. Der Vergleich mit der geplotteten theoretischen bivariaten Verteilung zeigt schon eine recht gute Approximation. Es ist dabei auch erkennbar, dass die Abhängigkeitsstruktur der Verteilung erhalten geblieben ist.

und der simulierten Verteilung, ist die Approximation bereits recht gut. Aufgrund der Existenz der bivariaten Verteilung und der Tatsache, dass die Ziehungen aus ihren voll-bedingten Verteilungen erfolgt, kann mit wachsendem Stichprobenumfang jeder beliebige Genauigkeitsgrad in der Approximation erreicht werden.

3.4.6 *Data Augmentation*-Algorithmus

Der *Data Augmentation*-Algorithmus ist ein iteratives Verfahren, welches 1987 von Tanner und Wong eingeführt wurde. Der Name beruht darauf, dass durch eine künstliche Erweiterung des zu analysierenden Modells, sei es mit neuen Parametern oder Daten, eine Vereinfachung der ursprünglichen Verteilungen erzielt werden kann.

Dieser Algorithmus kann aus zwei verschiedenen Perspektiven betrachtet werden: Einerseits stellt die Substitutionsmethode, wie im Fall des *Gibbs Sampler*s, die mathematische Basis des Algorithmus dar. Unter dieser Perspektive handelt es sich dabei um eine Abwandlung des *Gibbs Sampler*s und es ist relativ einfach zu beweisen, dass der *Data Augmentation*-Algorithmus bei Vorhandensein der voll-bedingten Verteilungen ein *Gibbs Sampler* ist.

Andererseits lässt die Betrachtung der algorithmischen Struktur des Verfahrens eine hohe Ähnlichkeit zum EM-Algorithmus erkennen. In der Tat nehmen beide Verfahren eine künstliche Erweiterung des zu Grunde liegenden Modells vor, die eine Erleichterung seiner Schätzung bzw. Simulation bewirkt (vgl. Dempster (1987)). Diese künstliche Modellerweiterung induziert eine charakteristische zweischrittige Struktur, welche beide Algorithmen prägt. Infolgedessen wird der *Data Augmentation*-Algorithmus oft als stochastische Version des EM-Algorithmus betrachtet (vgl. Tanner (1991)).

Aufgrund der Wichtigkeit des *Data Augmentation*-Algorithmus für die vorliegende Arbeit werden nach seiner Einführung beide Perspektiven untersucht.

3.4.6.1 Formale Struktur des *Data Augmentation*-Algorithmus

Gegeben seien die in Kapitel 2 betrachteten Zufallsvariablen X, Y und Z, wobei Y den beobachtbaren und Z den unbeobachtbaren bzw. latenten Teil von X bezeichnen. Es gelte also $X = (Y, Z)$. Die Verteilung von X sei abhängig von einem Parameter θ. Ferner ist

$$f_{\theta|Y}(\vartheta|y) = \int f_{\theta|X}(\vartheta|z, y)\, f_{Z|Y}(z|y)\, \mathrm{d}z \tag{3.34}$$

die *a posteriori*-Verteilung des Parameters θ gegeben die Realisationen y aus Y. Dabei ist $f_{\theta|X}(\vartheta|z, y)$ die bedingte Dichte des Parameters θ gegeben die vollständige Datenmenge $x = (y, z)$ und $f_{Z|Y}(z|y)$ ist die prädiktive Verteilung der latenten Zufallsvariablen Z gegeben $Y = y$ darstellen. Diese bedingte Dichte $f_{Z|Y}(z|y)$ kann wiederum dargestellt werden als

$$f_{Z|Y}(z|y) = \int f_{Z|\theta,Y}(z|\vartheta, y)\, f_{\theta|Y}(\vartheta|y)\, \mathrm{d}\vartheta. \tag{3.35}$$

Die Ähnlichkeit zu (3.27) auf Seite 55 ist offensichtlich.

Für den Ausdruck in (3.34) erhält man somit

$$\begin{aligned} f_{\theta|Y}(\vartheta|y) &= \int f_{\theta|Z,Y}(\vartheta|z, y) \cdot \int f_{Z|\theta,Y}(z|\vartheta, y)\, f_{\theta|Y}(\vartheta|y)\, \mathrm{d}\vartheta\, \mathrm{d}z \\ &= \int \left[\int f_{\theta|Z,Y}(\vartheta|z, y)\, f_{Z|\theta,Y}(z|\vartheta, y)\, \mathrm{d}z \right] f_{\theta|Y}(\vartheta|y)\, \mathrm{d}\vartheta, \end{aligned}$$

wobei das Integral in eckigen Klammern als

$$k(\vartheta, \phi) := \int f_{\theta|Z,Y}(\vartheta|z, y) f_{Z|\theta,Y}(z|\phi, y)\,\mathrm{d}z$$

analog zu Abschnitt 3.3.3 als Übergangskern von ϑ nach ϑ (vertreten durch das *dummy* Argument ϕ) dargestellt werden kann. Die Struktur der Fixpunkt-Integralgleichung ist somit erkennbar.

Der Algorithmus hat die folgende allgemeine Struktur:

1. Den Startpunkt stellt eine Approximation $f^0_{\theta|Y}(\cdot|y)$ für die Dichte $f_{\theta|Y}(\cdot|y)$ dar.
2. Eine Stichprobe von Umfang m, $(z^1, \ldots, z^m)$, aus der aktuellen Approximation der Dichte $f_{Z|Y}(\cdot|y)$ wird folgendermaßen generiert.
   (a) Ein Wert ϑ^0 wird aus $f^0_{\theta|Y}(\cdot|y)$ gezogen.
   (b) Bedingt auf den erhaltenen Wert ϑ^0 wird die Stichprobe z aus $f_{Z|\theta,Y}(\cdot|\vartheta^0, y)$ generiert.
3. Die neue Approximation für $f_{\theta|Y}(\cdot|y)$ ist

   $$f^1_{\theta|Y}(\vartheta|y) = \frac{1}{m}\sum_{j=1}^{m} f_{\theta|Z,Y}(\vartheta|z^j, y).$$

   D.h. $f^1_{\theta|Y}(\vartheta|y)$ ist ein arithmetisches Mittel aus den m bedingten Dichten $f_{\theta|Z,Y}(\cdot|z^j, y)$. Für großes m stellt das arithmetische Mittel von $f_{\theta|Y,Z}(\vartheta|y, z)$ über die m erweiterten Daten eine gute Approximation für $f_{\theta|Y}(\vartheta|y)$ dar, denn es gilt

   $$\frac{1}{m}\sum_{j=1}^{m} f_{\theta|Z,Y}(\vartheta|z^j, y) \xrightarrow[m\to\infty]{} \int f_{\theta|Z,Y}(\vartheta|z, y)\, f_{Z|Y}(z|y)\,\mathrm{d}z.$$

4. Die Schritte eins bis drei werden wiederholt, bis ein Konvergenzkriterium erreicht wird.

3.4.6.2 Eigenschaften des *Data Augmentation*-Algorithmus

Mit Hilfe funktionalanalytischer Methoden und unter Annahmen, welche der Irreduzibilität und positiven Rekurrenz einer Übergangsmatrix im Falle eines diskreten Zustandraums ähneln, haben Tanner und Wong (1987) folgende Eigenschaften des Algorithmus bewiesen:

1. Die durch den *Data Augmentation* Algorithmus induzierten Markov-Ketten konvergieren eindeutig gegen die Zielverteilung. Der Einfachheit halber sei hier die Zielverteilung mit f bezeichnet.
2. Die Folge $\int |f_i(t) - f(t)|\,\mathrm{d}t$ konvergiert für $i \to \infty$ geometrisch gegen Null.

Eine zusätzliche und zugleich veblüffende Eigenschaft dieses Algorithmus ist die Tatsache, dass die Konvergenz für alle Stichprobenumfänge m gewährleistet ist. Das heißt, sogar im Fall $m = 1$ ist die Lösung eindeutig (vgl. Schafer (1997, S. 71)). Dieses Resultat wird im nächsten Abschnitt aufgegriffen.

3.4.6.3 *Data Augmentation* als Sonderfall des *Gibbs Sampler*s

In Anlehnung an Gelfand und Smith (1990) werden für die folgende Ausführung die *Substitution-Sampling*- und die *Data Augmentation*-Methode als äquivalente Verfahren betrachtet.

Der Startpunkt des Vergleichs sind drei beliebige Zufallsvariable X, Y, Z mit gemeinsamer Dichtefunktion $f_{XYZ}(\cdot)$[28]. Dieser trivariate Fall ist insofern wichtig, als er den einfachsten Fall darstellt, der die Unterschiede zwischen beiden Verfahren aufzeigen kann. Die drei Randdichten von f_{XYZ} sind darstellbar als

$$\begin{aligned} f_X(x) &= \iint f_{X|Z,Y}(x|z,y)\, f_{Z|Y}(z|y)\, f_Y(y)\, \mathrm{d}z\, \mathrm{d}y \\ f_Y(y) &= \iint f_{Y|X,Z}(y|x,z)\, f_{X|Z}(x|z)\, f_Z(z)\, \mathrm{d}z\, \mathrm{d}x \\ f_Z(z) &= \iint f_{Z|Y,X}(z|y,x)\, f_{Y|X}(y|x)\, f_X(x)\, \mathrm{d}x\, \mathrm{d}y. \end{aligned} \tag{3.36}$$

Die Substitutionsmethode benötigt alle sechs bedingten Verteilungen auf der rechten Seite von (3.36). D.h. insbesondere, dass die drei voll-bedingten Verteilungen $f_{X|Z,Y}, f_{Y|Y,Z}$ und $f_{Z|Y,X}$ benötigt werden. Im Gegensatz dazu braucht der *Gibbs Sampler* ausschließlich diese voll-bedingten Verteilungen[29].

Gelfand and Smith zeigen, dass die von der Substitutionsmethode zusätzlich benötigten Dichtefunktionen $f_{Z|Y}, f_{X|Z}$ und $f_{Y|X}$ leicht mit Hilfe der voll-bedingten Verteilungen und der Randverteilungen f_X, f_Y und f_Z durch Hinzufügen von sogenannten *Gibbs*-Unterzyklen (siehe 3.4.5.1) simuliert werden können. Diese neue Reihenfolge weicht zwar von der ursprünglichen Reihenfolge des *Gibbs Sampler*s ab. Nichtsdestotrotz kann leicht gezeigt werden, dass diese auch ein *io*-Schema darstellt. D. h., aufgrund der in Abschnitt 3.4.5.2 ersten aufgelisteten Eigenschaft des *Gibbs Sampler*s, ist dies bereits ausreichend, um die Konvergenz des Algortihmus zu gewährleisten. Die Erweiterung auf k Zufallsvariable kann auf gleiche Art und Weise bewerkstelligt werden.

Es wurde somit gezeigt, dass bei Vorhandensein der voll bedingten Verteilungen der *Data Augmentation*-Algorithmus immer durch einen *Gibbs Sampler* ersetzt werden kann. Ein möglicher Nachteil hiervon könnte eine eventuelle Verringerung der Konvergenzgeschwindigkeit sein, bei Vorhandensein zusätzlicher bedingter Verteilungen, welche der *Gibbs Sampler* nicht berücksichtigt. Gelfand und Smith (1990, S. 401) haben diesen Fall simulativ untersucht und sind zu dem Schluß gekommen, dass diese Verlangsamung der Konvergenzgeschwindigkeit meist von geringer praktischer Bedeutung ist.

Zum Schluß sei angemerkt, dass es ebenso möglich ist, den *Gibbs Sampler* als Sonderfall des *Data Augmentation*-Algorithmus darzustellen. Wie in Abschnitt 3.4.6.2 angeführt, ist die Konvergenz des *Data Augmentation*-Algorithmus für alle Stichprobenumfänge m gewährleistet. Das iterative Schema, welches sich bei $m = 1$ ergibt, ist nichts anderes als eine Instanz des *Gibbs Sampler*s,

[28] Um den Vergleich des *Gibbs Sampler*s mit dem *Data Augmentation*-Algorithmus zu ermöglichen, gilt in diesem Abschnitt X=(Y,Z) i.A. nicht.

[29] Das ist auch die kleinste Auswahl an bedingten Verteilungen, welche die gemeinsame Dichte eindeutig charakterisieren kann (vgl. Gelfand und Smith (1990, S. 401).

welche häufig zur statistischen Behandlung fehlender Daten eingesetzt wird. Diese Methode stellt die Basis der noch ausführlich zu behandelnden Imputationsmodellen dar, die dem praktischen Teil dieser Arbeit zu Grunde liegen[30].

3.4.6.4 Algorithmische Struktur des *Data Augmentation*-Algorithmus

Wenn seine algorithmische Struktur in den Vordergrund gestellt wird, weist der *Data Augmentation*-Algorithmus eine hohe Ähnlichkeit zum bereits in Kapitel 2 vorgestellten EM-Algorithmus auf. In der Tat kombiniert der *Data Augmentation*-Algorithmus die Eigenschaften des EM-Algorithmus und der multiplen Imputation, um die *a posteriori*-Verteilung von θ zu simulieren. Ebenfalls besitzt der *Data Augmentation*-Algorithmus eine zweischrittige Struktur, welche der des EM-Algorithmus sehr ähnlich ist. Der Hauptunterschied zum EM-Algorithmus besteht darin, dass die deterministischen E- und M-Schritte durch zwei stochastische Schritte ersetzt werden, welche mit I- bzw. P-Schritt bezeichnet werden.

(a) Imputation-Schritt (I-Schritt)

Um die fehlenden Werte zu imputieren, wird eine Schätzung des Parameters θ benötigt. Im Gegensatz zum EM-Algorithmus, welcher bedingte Erwartungen verwendet, benutzt der *Data Augmentation*-Algorithmus Ziehungen aus der bedingten Verteilung der fehlenden Daten gegeben die beobachteten Daten und die aktuellen Parameter. Der Imputation-Schritt kann wie folgt geschrieben werden:

$$Y_{mis}^{(t+1)} \sim P(Y_{mis} \mid Y_{obs}, \theta^{(t)}). \tag{3.37}$$

Beziehung zum EM-Algorithmus: Vereinfachend kann der Wert für $Y^{(t+1)}$, welcher mittels des EM-Algorithmus bestimmt wird, als der Erwartungswert der Verteilung in (3.37) betrachtet werden.

(b) Posterior-Schritt (P-Schritt)

Mit den vervollständigten Daten vom I-Schritt, $Y^{(t+1)} = (Y_{obs}, Y_{mis}^{(t+1)})$, kann eine neue Schätzung des Parameters vorgenommen werden. Diese Schätzung stellt die Informationen über die Parameter dar, die in den Daten enthalten sind. Diese Daten-basierten Informationen werden mit den *a priori*-Informationen kombiniert, welche in Form einer *a priori*-Verteilung vorliegen[31]. Aus der Kombination der *a priori* und der in den Daten enthaltenen Informationen resultiert eine neue Verteilung der Parameter, aus der neue Werte gezogen werden können. Der P-Schritt kann also wie folgt geschrieben werden

$$\theta^{(t+1)} \sim P\left(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}\right). \tag{3.38}$$

Beziehung zum EM-Algorithmus: Aufgrund der Tatsache, dass es sich um einen ML-Schätzer handelt, stellt der Wert $\theta^{(t+1)}$ des EM-Algorithmus[32] meist den Modalwert der Verteilung in Gleichung (3.38) dar.

[30] In der Literatur wird jedoch der Fall $m = 1$ meist als *Data Augmentation* Algorithmus und nicht als *Gibbs Sampler* bezeichnet, um seine Beziehung zu Problemen mit unvollständigen Daten hervorzuheben (vgl. Schafer (1997, S. 72)).

[31] Diese Betrachtung des Algorithmus hat einen deutlichen bayesianischen Charakter. Auf die Wahl der *a priori*-Verteilung wird in Abschnitt 4.2 eingegangen.

[32] Dies gilt unter zusätzlichen Regularitätsbedingungen, die im engen Zusammenhang mit der Wahl der *a priori*-Verteilung stehen.

3.4.6.5 Beispiel: Prädiktive Verteilung eines fehlenden Wertes (NA) im Fall einer multivariaten Normalverteilung.

Gegeben sei eine Stichprobe vom Umfang 400 aus einer 5-dimensionalen Normalverteilung mit den als unbekannt vorausgesetzten Parametern

$$\mu = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 2 \\ 1 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 4{,}000 & 1{,}340 & 1{,}200 & 0{,}304 & 0{,}144 \\ 1{,}340 & 1{,}000 & 0{,}750 & 0{,}208 & 0{,}120 \\ 1{,}200 & 0{,}750 & 2{,}250 & 0{,}744 & 0{,}531 \\ 0{,}304 & 0{,}208 & 0{,}744 & 0{,}640 & 0{,}264 \\ 0{,}144 & 0{,}120 & 0{,}531 & 0{,}264 & 0{,}360 \end{pmatrix} \tag{3.39}$$

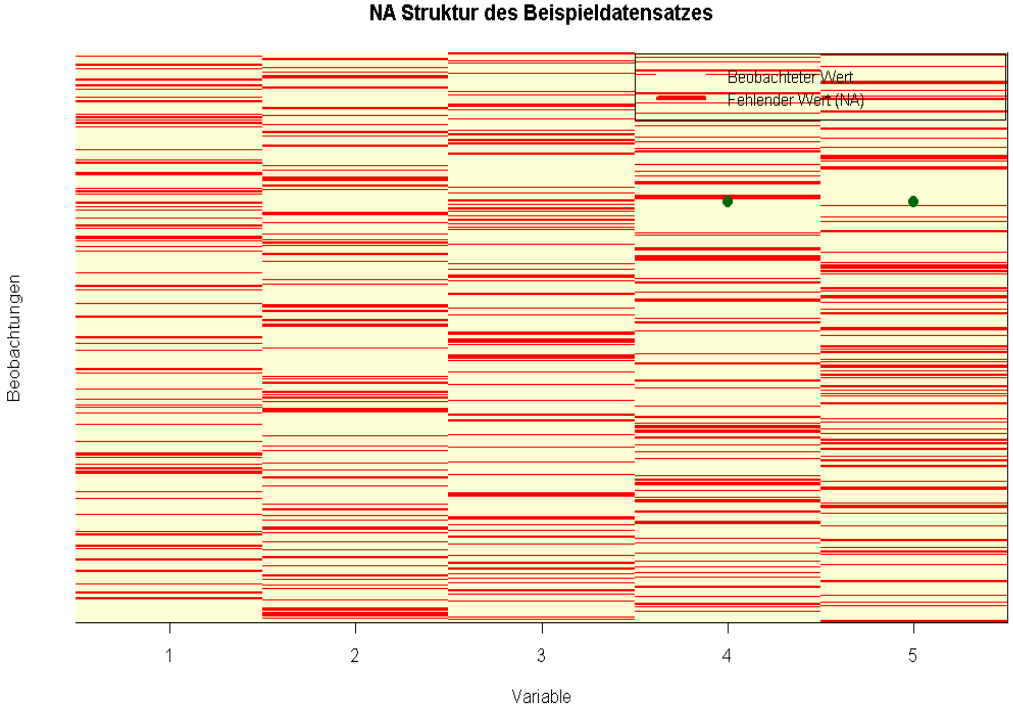

Abbildung 3.15: NA-Struktur des Datensatzes. Die grünen Punkte kennzeichnen die Werte, deren Verteilung geschätzt wird.

Für die Simulation werden 400 Werte aus dem Datensatz als NA gesetzt. Diese Werte werden so gewählt, dass keine einzige Zeile der Datenmatrix vollständig beobachtet ist. Die resultierende Struktur des Datensatzes ist in Abbildung 3.15 zu sehen. Gesucht sei nun die prädiktive Verteilung eines bestimmten Paares an fehlenden Daten, gegeben die Informationen in der Stichprobe, zwecks Imputation dieser unbeobachteten Werte.

Um die gemeinsame prädiktive Verteilung beider Werte zu simulieren, wird ein *Data Augmentation*-Algorithmus eingesetzt, wobei der Einfachheit halber der Wert $m = 1$ gewählt wird. Es handelt sich somit genau genommen um einen *Gibbs Sampler*, wie in Abschnitt 3.4.6.3 erörtert wurde.

Um die Struktur des Algorithmus möglichst verständlich darzustellen, wird zunächst ein Vergleich zwischen der in Abschnitt 3.4.6.1 eingeführten theoretischen Struktur und der vorliegenden Anwendung vorgenommen:

Mit θ seien die unbekannten Parameter μ und Σ bezeichnet, deren gemeinsame Verteilung sich aus $f_\theta = f_\Sigma \cdot f_{\mu|\Sigma}$ ergibt. Ferner wird mit $\mathbf{x}$[33] die vorliegende Stichprobe bezeichnet. Aufgrund des Vorhandenseins von fehlenden Werten setze sich $\mathbf{x}$ in diesem Fall aus den beobachteten Daten $\mathbf{y}$ und den unbeobachteten Daten $\mathbf{z}$ zusammen. Es gelte also $\mathbf{x} = (\mathbf{y}, \mathbf{z})$.

Die Dichtefunktion $f_{\theta|X}(\vartheta|\mathbf{x}) = f_{\theta|Y,Z}(\mu, \Sigma|\mathbf{y}, \mathbf{z})$ charakterisiert somit die *a posteriori*-Verteilung der Parameter μ und Σ gegeben die mit fehlenden Werten behafteten Stichprobe $\mathbf{x}$.

Wenn alle Werte beobachtbar wären, würde man durch Anwendung des Theorems von Bayes

$$f_{\theta|X}(\vartheta|\mathbf{x}) \propto f_\theta(\vartheta) f_{X|\theta}(\mathbf{x}|\vartheta), \tag{3.40}$$

erhalten. Dabei bezeichnet $f_\theta(\vartheta)$ die *a priori*-Verteilung von μ und Σ und das Symbol $\propto$ verdeutlicht, dass diese Verteilung bis auf eine Proportionalitätskonstante $\int f_\theta(\vartheta) f_{X|\theta}(\mathbf{x}|\vartheta)\,\mathrm{d}\vartheta$ eindeutig bestimmt ist. Im betrachteten Fall einer multivariaten Normalverteilung ist es unproblematisch, $f_{\theta|X}(\vartheta|\mathbf{x})$ in geschlossener Form anzugeben. Die Berechnung der *a posteriori*-Verteilung der Parameter unter Verwendung sowohl von konjugierten als auch von uninformativen *a priori*-Verteilungen wird ausführlich in Gelman et al. (2004, S. 87-88) behandelt.

Bei Vorhandensein fehlender Daten ist die Situation jedoch deutlich komplexer. Der Ausdruck $f_{\theta|X}(\vartheta|\mathbf{x})$ kann nicht mehr analytisch berechnet werden und man ist auf simulative Methoden angewiesen.

Mit $f_{Z|Y}(z|y)$ wird die prädiktive Verteilung der fehlenden Daten $\mathbf{z}$ gegeben die beobachteten Daten $\mathbf{y}$ bezeichnet. Die Abhängigkeit vom Parameter θ ist dabei durch Integration bereits eliminiert worden. D.h. es gilt

$$f_{Z|Y}(z|y) = \int f_{Z|\theta,Y}(\mathbf{z}|\vartheta, \mathbf{y}) f_{\theta|Y}(\vartheta|\mathbf{y})\,\mathrm{d}\vartheta. \tag{3.41}$$

Der Integrator in (3.41) liegt nicht in geschlossener Form vor und die prädiktive Verteilung $f_{Z|Y}(z|y)$ muss deshalb ebenfalls simuliert werden. Die simulative Bestimmung dieser prädiktiven Verteilung ist das ultimative Ziel eines jeden multiplen Imputationsverfahrens.

Der *Data Augmentation*-Algorithmus für dieses Beispiel geht Folgendermaßen vor:

1. Startwerte $\vartheta_0 = \{\Sigma_0, \mu_0\}$ für $\theta = \{\Sigma, \mu\}$ werden gewählt.
2. Bedingt auf ϑ_0 und die beobachteten Daten $\mathbf{y}$ werden Werte $\mathbf{z}_0$ aus $f_{Z|\theta,Y}(\mathbf{z}|\vartheta_0, \mathbf{y})$ für die fehlenden Werte gezogen. Dabei handelt es sich um eine Normalverteilung deren Erwartungswert ϕ und Varianz-Kovarianzmatrix Ψ durch die multivariate lineare Regression von $\mathbf{z}$ auf $\mathbf{y}$ bzw. durch die zugehörige Varianz-Kovarianzmatrix der Residuen gegeben sind (im Folgenden werden beide Parameter als ϕ zusammengefasst).

[33] Um die Unterscheidung zwischen Datenmatrizen und Datenvektoren zu ermöglichen wird für dieses Beispiel die Notation $\mathbf{x}$ für die Stichprobe x verwendet. Gleiches gilt für $\mathbf{y}$ und $\mathbf{z}$.

3. Mit den vervollständigten Daten $\mathbf{x}_0$ =können neue Werte für die Parameter berechnet werden:

   (a) Gegeben die vollständigen Daten $\mathbf{x}_0 = (\mathbf{y}, \mathbf{z}_0)$ ist die neue Varianz-Kovarianz Matrix Wishart[34] verteilt mit Skalierungsparameter $\hat{\Sigma}_0$, wobei

   $$\hat{\Sigma}_0 = \frac{1}{n}\left(\mathbf{x}_0'\mathbf{x}_0 - \bar{x}_0\bar{x}_0'\right) \quad \text{und} \tag{3.42}$$

   $$\bar{x}_0^j = \frac{1}{n}\sum_{i=1}^{n} x_0^{i,j} \quad \text{für } j \in \{1:5\}. \tag{3.43}$$

   Dabei ist $\bar{x}_0 = (\bar{x}_0^1, \ldots, \bar{x}_0^5)$ der Vektor der spaltenweise berechneten arithmetischen Mittel der vervollständigten Datenmatrix.

   (b) Bedingt auf Σ_0 und $\mathbf{x}_0$ wird μ_1 aus einer Normalverteilung mit den Parametern $\bar{x}_0$ und Σ_0/n) gezogen, wobei n die Anzahl der Zeilen der Matrix $\mathbf{x}_0$ bezeichnet. Der *Posterior*-Schritt ist somit vollständig.

4. Es wird zwischen Imputation- und *Posterior*-Schritt iteriert bis ein Konvergenzkriterium erreicht wird.

Die Werte der vierten und fünften Spalte der 106. Zeile des Beispieldatensatzes wurden für die Simulation als NA gesetzt. Diese Werte werden fortan als $z_{106,4}$ und $z_{106,5}$ bezeichnet. Die Wahl der 106. Zeile ist relativ willkürlich, wohingegen die vierte und fünfte Spalte aufgrund der Tatsache ausgewählt worden sind, dass sie hohe Korrelationen (0,62 bzw. 0,59) zur dritten Spalte aufweisen. Somit ist eine gewisse Qualität in der Imputation gewährleistet, welche die didaktischen Zwecke dieses Beispiels unterstützt.

Ziel der Untersuchung ist die Simulation der gemeinsamen Verteilung der beiden Werte gegeben die beobachteten Daten, d.h. die Simulation von $f(z_{106,4}, z_{106,5}|\mathbf{y})$.

Nach einer *burn in*-Phase[35] von 2000 Iterationen werden 1000 Iterationen des *Data Augmentation*-Algorithmus zur Schätzung der Verteilung der fehlenden Werte verwendet. Als Vergleich dienen:

1. Die Verteilung dieser Daten bei bekannten Parametern und gegeben die anderen Beobachtungen der ausgewählten Zeile. Es handelt sich dabei um eine Normalverteilung mit

   - Vektor der Erwartungswerte: ergibt sich aus der linearen Regression der Werte in der vierten und fünften Spalte auf die beobachteten Werte in der Zeile 106.
   - Varianz-Kovarianzmatrix: ist gleich der Varianz-Kovarianzmatrix der Residuen dieser Regression.

[34] Die Wishart-Verteilung stellt eine Art multivariate Version der Chi-Quadrat-Verteilung dar. Diese Verteilung wird im normalverteilten Modell (siehe Kapitel 4) als Stichprobenverteilung der Varianz Kovarianz Martix der Daten angenommen.

[35] Beim *burn in* handelt es sich um die ersten Zyklen eines MCMC-Algorithmus, in denen die Markov-Kette noch nicht gegen ihre stätionäre Verteilung konvergiert hat. Diese Zyklen werden nicht berücksichtigt.

Die nötigen Regressionsparameter resultieren aus der Faktorisierung der Parameter der multivariaten Normalverteilung und können mit Hilfe des *Sweep*-Operators für alle möglichen Kombinationen von beobachteten und fehlenden Daten berechnet werden (siehe Abschnitt 4.1.2). Die bedingte Verteilung der fehlenden Werte, gegeben die Parameter ϕ und restlichen Werte $x_{106,\{1,2,3\}}$ der Zeile 106, d.h. $f(z_{106,4}, z_{106,5}|\phi, x_{106,\{1,2,3\}})$[36], stellt die maximal zu erreichende Kenntnis über die fehlenden Werte $z_{106,4}$ und $z_{106,5}$ dar und dient dem Zweck, einen grafischen Maßstab zu liefern, um die Güte der anderen simulierten Verteilungen zu beurteilen.

2. Die prädiktive Verteilung der zu imputierenden Daten im Fall eines vollständig beobachteten Datensatzes. Hierbei ist die Imputation eher als Prognose zu verstehen: Die vollständig beobachteten Daten ermöglichen die Schätzung der Parameter der gemeinsamen Verteilung. Die geschätzten Parameter $\hat{\phi}$ und die beobachteten Daten der 106ten Zeile ergeben dann die *a posteriori*-Verteilung der zu simulierenden Werte, d.h. $f(z_{106,4}, z_{106,5}|\phi = \hat{\phi}, \mathbf{x})$. Um die Abhängigkeit von geschätzten Parametern zu eliminieren, ist bei bayesianischen Imputationstechniken üblich, die prädiktive Verteilung der fehlenden gegeben die beobachteten Werte zu konstruieren:

$$f(z_{106,4}, z_{106,5}|\mathbf{x}) = \int_{\Phi} f(z_{106,4}, z_{106,5}|\phi, \mathbf{x}) f(\phi|\mathbf{x}) \mathrm{d}\phi.$$

Diese Verteilung wurde ebenfalls simuliert. Aufgrund der Tatsache, dass die Daten vollständig beobachtet sind, kann die Simulation der nötigen Verteilungen mittels herkömmlicher Monte-Carlo-Verfahren bewerkstelligt werden.

Abbildung 3.16 veranschaulicht die Ergebnisse der Simulation. Trotz der hohen Anzahl an fehlenden Daten sind die mittels des *Data Augmentation*-Algorithmus simulierten Randdichten (grüne Linie) in der Lage, die theoretische bedingte Verteilung der fehlenden Daten (schwarze Linie) zu approximieren. Die Tatsache, dass die prädiktiven Verteilungen etwas breiter und niedriger als die theoretischen Verteilungen sind, ist eine natürliche Konsequenz des Herausintegrierens der Parameter.

Als zusätzlicher Vergleich wurde ein EM-Algorithmus auf die Daten angewendet. Die Ergebnisse sind durch eine vertikale gepunktete schwarze Linie in Abbildung 3.16 gekennzeichnet. Der EM-Algorithmus liefert die Werte: $\hat{z}_{106,4} = 1{,}686$ und $\hat{z}_{106,5} = 0{,}785$. Beide Werte befinden sich sehr nahe dem Modalwert der jeweiligen theoretischen Verteilung, wie auf der Abbildung zu sehen ist.

Die theoretische Korrelation zwischen $z_{106,4}$ und $z_{106,5}$ beträgt 0,281[37]. Aufgrund der vorhandenen Korrelationsstruktur der beobachteten Variablen kann die ursprüngliche Korrelation von 0,55 zwischen der fünften und der sechsten Spalte nicht vollständig rekonstruiert werden. Die Simulation bei unbekannten Parametern und voll beobachteten Daten liefert eine empirische Korrelation von 0,246. Die simulierten Werte bei Vorhandensein von fehlenden Daten weisen eine Korrelation von 0,309 auf. Abbildung 3.17 zeigt die simulierten Daten aus den drei bivariaten Verteilungen. Aus allen geht eine ähnliche Korrelationsstruktur hervor.

[36] Im Falle bekannter Parameter spielen die anderen Werte des Datensatzes, für die Schätzung dieser bedingten Verteilung, keine nennenswerte Rolle.

[37] Diese Korrelation darf nicht mit der theoretischen Korrelation zwischen der fünften und der sechsten *Spalte* verwechselt werden.

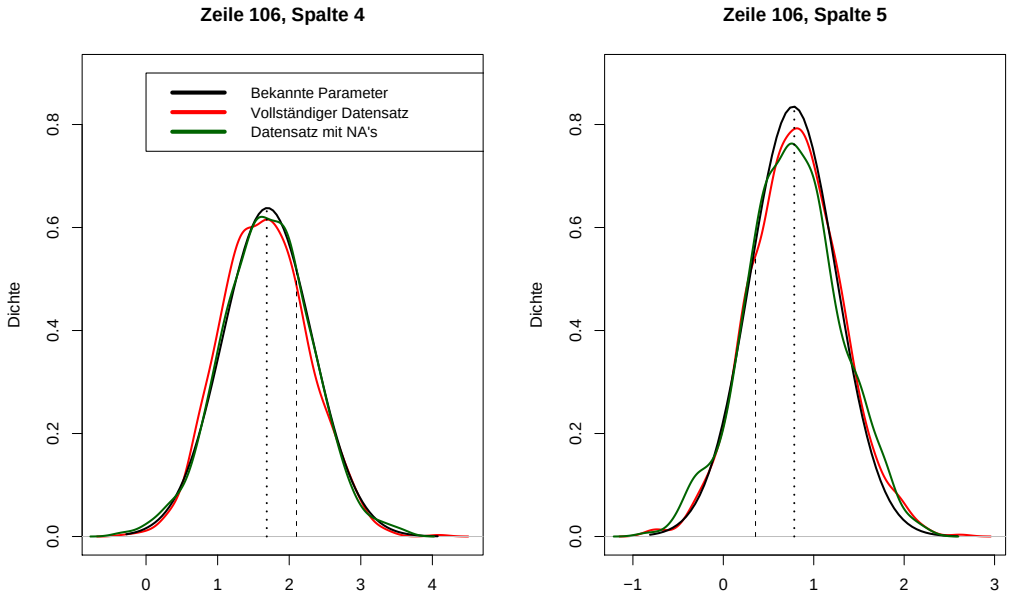

Abbildung 3.16: Vergleich der Randverteilungen. Die bedingte Verteilung der fehlenden Werte bei bekannten Parametern (schwarze Linie) wird mit den Kerndichteschätzungen der prädiktiven Verteilungen bei voll beobachteten Daten (rote Linie) und bei Vorhandensein von fehlenden Werten (grüne Linie) verglichen. Die gestrichelten vertikalen Linien kennzeichnen die ursprünglichen und für die Simulation entfernten Beobachtungen. Die gepunkteten vertikalen Linien kennzeichnen den Schätzwert des EM-Algorithmus.

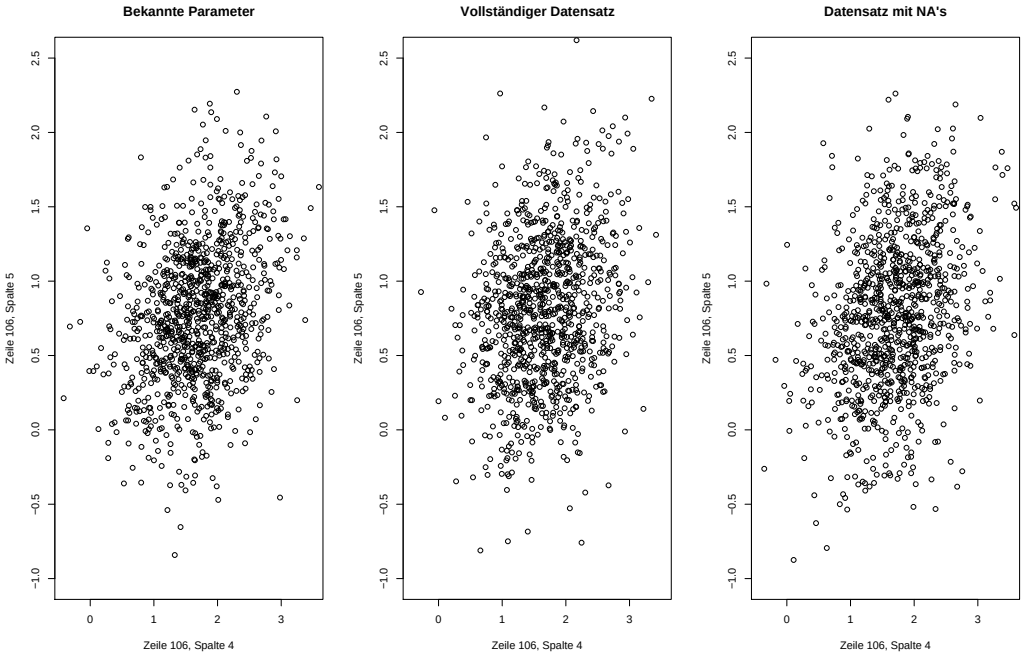

Abbildung 3.17: Die drei Streudiagramme veranschaulichen die bivariaten Verteilungen der simulierten Daten in den drei betrachteten Fällen, wobei das erste die theoretisch zu erwartende Korrelationsstruktur wiedergibt. Es ist ersichtlich, dass der *Data Augmentation*-Algorithmus in der Lage ist, diese Korrelationsstruktur nachzubilden.

3.5 *Grid Sampler*

Die Anwendung des *Gibbs Sampler*s und des *Data Augmentation*-Algorithmus setzt voraus, dass aus allen bedingten Verteilungen Stichproben auf effiziente Art und Weise gezogen werden können.

Das Vorhandensein aller bedingten Verteilungen kann in praktischen Situationen jedoch oft nur durch sehr restriktive Annahmen an die zugrunde liegenden Modelle erkauft werden. Die in Abschnitt 3.2 beschriebenen Methoden *Acceptance-Rejection-* und *Importance-Sampling* stellen eine Möglichkeit dar, bedingte Verteilungen ohne restriktive Annahmen zu approximieren. Diese Methoden setzen jedoch Kenntnisse der zu approximierenden Funktionen voraus, welche nicht immer vorhanden sind.

Eine Alternative zu diesen Simulationsmethoden stellt der *Grid Sampler* (siehe Tanner und Wong (1987) oder Ritter und Tanner (1992)) dar. Dieser Algorithmus basiert auf der Evaluation der zu approximierenden (unnormierten) Dichtefunktion in einem Punktgitter und der anschließenden Annäherung ihrer Quantilfunktion mittels einfacher, stückweise definierter Funktionen (Linear- bzw Splinefunktionen).

Dieses Verfahren erweist sich als besonders nützlich, wenn die zu approximierende Verteilung

- univariat ist,
- eine unbekannte Form besitzt und
- nicht normiert ist.

Im Falle multivariater Verteilungen ist jedoch der *Grid Sampler* aufgrund des in der Einleitung bereits erwähnten schnellen Anstiegs der Anzahl an Stützstellen (*Curse of dimensionality*) nicht zu empfehlen.

3.5.1 Vorbereitende Schritte: Anzahl und Lage der Stützstellen

Bevor der *Grid Sampler* verwendet werden kann, müssen sowohl die Anzahl als auch die Lage der Stützstellen festgelegt werden. Da beide Entscheidungsgrößen einen erheblichen Einfluss auf die Qualität der Annäherung durch den *Grid Sampler* haben, erweist es sich als zweckmäßig, diese Problematik näher zu betrachten.

Es ist naheliegend, dass eine gleichmäßige Verteilung der Stützstellen nur dann eine gute Annäherung der Zielfunktion liefern kann, wenn das Punktgitter sehr engmaschig gewählt wird. Der resultierende Aufwand wird jedoch selten durch die erzielte Qualität der Annäherung kompensiert. Andererseits werden durch ein grobes Gitter, insbesondere im Falle von Dichtefunktionen mit einer hohen Kurtosis, Regionen mit geringer Wahrscheinlichkeitsmasse erheblich überschätzt, während Regionen mit hoher Wahrscheinlichkeitsmasse unterschätzt werden. Abbildung 3.18 veranschaulicht diese Aussagen.

Der von Ritter und Tanner (1992) vorgeschlagene Ansatz besteht darin, die Stützstellen so zu wählen, dass sie mit der Funktion Bereiche gleicher Wahrscheinlichkeitsmasse einschließen. Dieser

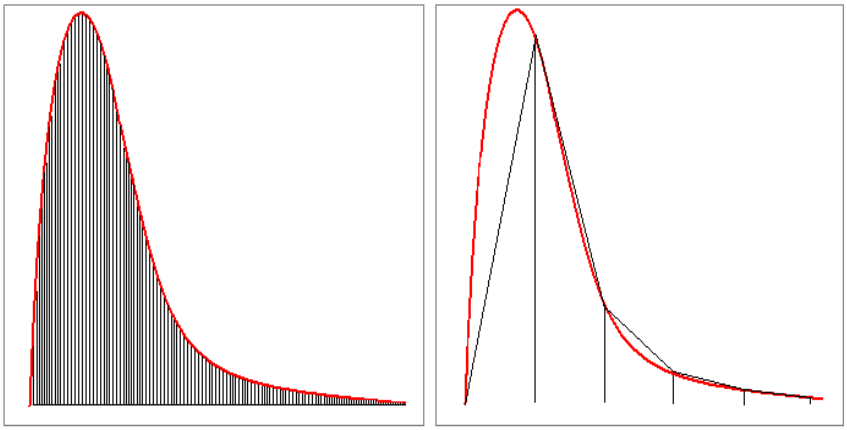

Abbildung 3.18: Approximation einer Zielfunktion (rote Linie) durch stückweise definierte lineare Funktionen in einem gleichmässig verteilten Gitter. **Linke Grafik:** Aufgrund ihrer geringen Wahrscheinlichkeitsmasse könnte die rechte Flanke mit einer viel geringeren Anzahl an Stützstellen zu einem vergleichbaren Genauigkeitsgrad approximiert werden. **Rechte Grafik:** Durch ein grobes Gitter werden Regionen mit einer hohen Wahrscheinlichkeitsmasse unter- und mit einer geringen Wahrscheinlichkeitsmasse überschätzt.

Vorschlag muss jedoch im hier relevanten Fall unbekannter Funktionen auf iterative Weise bewerkstelligt werden, denn die eingeschlossene Wahrscheinlichkeitsmasse ist wiederum eine Funktion der Lage der Stützstellen.

Der für die Zwecke dieser Arbeit entwickelte Ansatz weicht vom oben genannten Verfahren ab und basiert auf der Untersuchung der Krümmung der Zielfunktion in einem Punktgitter und ihrer Approximation mit einfachen, linearen Funktionen. Die Begründung für diesen Ansatz ist wie folgt: Um eine möglichst gute Annäherung der Zielfunktion durch stückweise definierte lineare Funktionen zu ermöglichen, müssen diejenigen Bereiche, welche durch eine ausgeprägte Krümmung der Funktion gekennzeichnet sind, an mehr Stützstellen evaluiert werden, als diejenigen mit einer geringen Krümmung. Ein geeignetes Maß für die Krümmung einer Funktion an einer bestimmten Stelle ist der Betrag ihrer zweiten Ableitung an dieser Stelle. Das vorgeschlagene Verfahren basiert somit auf der numerischen Bestimmung der zweiten Ableitung an verschiedenen Auswertungsstellen und verteilt mit Hilfe dieser Informationen die Stützstellen für den *Grid Sampler*. Dieses Verteilungskriterium wird in Abbildung 3.19 veranschaulicht.

Da dieser Algorithmus speziell für die vorliegende Arbeit entwickelt wurde, wird er im Folgenden ausführlich beschrieben.

1. Eine Ober- und Untergrenze für den Definitionsbereich der Funktion sowie die Anzahl an Evaluationspunkten für die zweite Ableitung werden festgelegt. Da am Anfang keine Informationen bezüglich der Krümmung der zu evaluierenden Funktion vorliegen, werden die Evaluationspunkte gleichmäßig verteilt.

   Zu diesem Zweck wird der festgelegte Definitionsbereich der Funktion in Intervalle zerlegt, welche die Breite $\delta = (\gamma_2 - \gamma_1)/p$ besitzen. Dabei bezeichnet γ_2 die Obergrenze, γ_1 die Untergrenze und p die gewählte Anzahl an Evaluationspunkten bezeichnen.

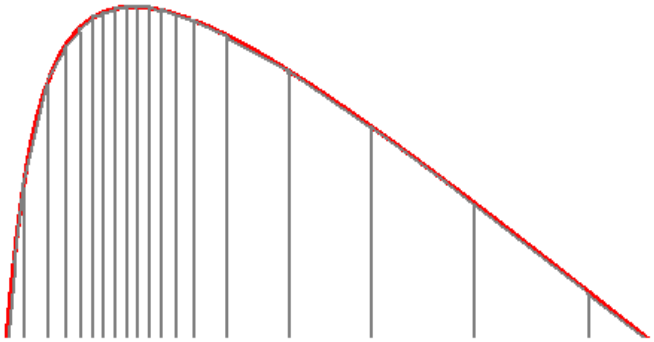

Abbildung 3.19: Approximation einer Zielfunktion (rote Linie) durch stückweise definierte lineare Funktionen. In Regionen mit einer ausgeprägten Krümmung der Zielfunktion liegen die Stützstellen dichter beieinander.

2. In der Mitte von jedem Intervall wird ein Stützpunkt platziert und die zweite Ableitung an diesem Punkt numerisch bestimmt. In Abbildung 3.20 werden die ersten zwei Schritte im Falle einer Standardnormalverteilung veranschaulicht.

3. Der Wert der zweiten Ableitung wird einer Potenztransformation unterzogen. Diese Transformation dient dem Zweck, eine exzessive Konzentration der Stützstellen im Falle von Funktionen zu verhindern, welche große Unterschiede in ihrem Krümmungsverhalten aufweisen. Niedrige Werte des Potenzparameters λ ($\lambda \ll 1$) ermöglichen eine eher gleichmäßige Verteilung der Stützstellen. Werte nahe 1 dagegen bewirken eine hohe Konzentration der Stützstellen in Bereichen mit einer ausgeprägten Krümmung.

4. Anschließend werden die in Schritt 3. resultierenden Werte durch ihre Summe geteilt. Diese relativen Werte entscheiden dann über die Verteilung der Stützstellen im Definitionsbereich der Funktion.

   Abbildung 3.21 zeigt die Verteilung der Stützstellen für drei Werte 0,3, 0,6 und 1 des Parameters λ.

5. Der *Grid Sampler* wird für die resultierenden Stützstellen konstruiert.

3.5.2 Implementierung des *Grid Sampler*s

Im Folgenden wird der *Grid Sampler* in der Version von Tanner und Wong (1987) beschrieben. Dieser Algorithmus wird häufig auch *Griddy Gibbs Sampler* genannt (vgl. Ritter und Tanner (1992), Little und Rubin (2002)). Diese Bezeichnung ist jedoch insofern irreführend, als der Algorithmus nicht iterativer Natur ist. Zwar kann dieses Verfahren die Anwendung des *Gibbs Sampler*s erleichtern, seine Anwendungsgebiete beschränken sich jedoch nicht auf MCMC-Methoden.

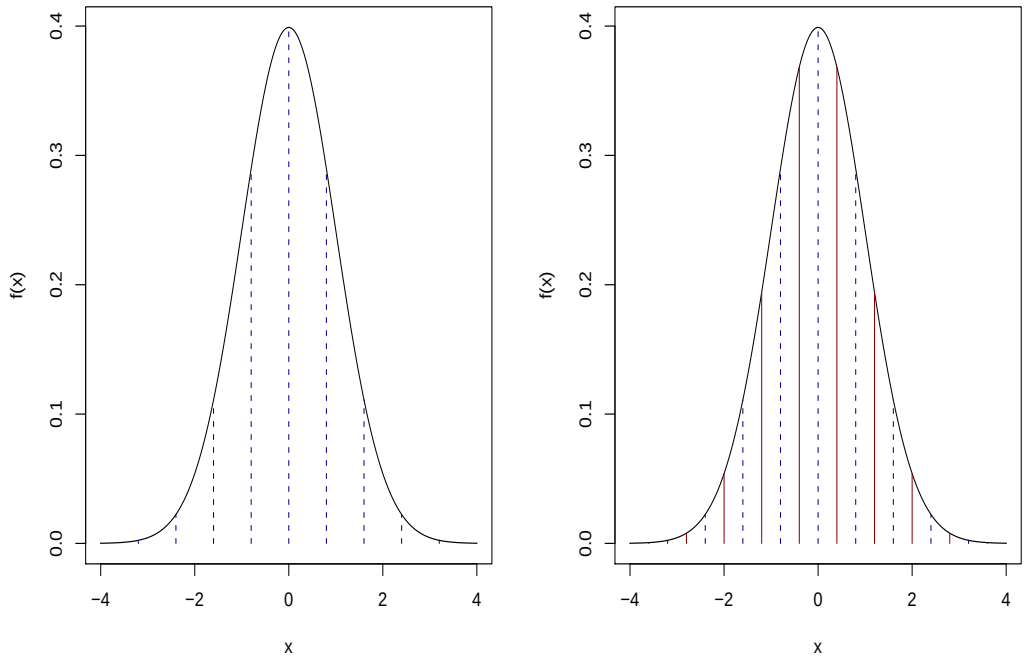

Abbildung 3.20: **Approximation einer Standardnormalverteilung:** Die linke Grafik zeigt die gleichmäßige Einteilung in zu evaluierende Bereiche (gestrichelte Linien). Die rechte Grafik zeigt die Stützstellen zur Berechnung der zweiten Ableitung (durchgezogene Linien). In Abhängigkeit vom Betrag der zweiten Ableitung wird pro Bereich die Anzahl an Stützstellen zur Evaluation der Funktion bestimmt.

Das Verfahren besteht aus den folgenden Schritten:

1. Es werden die Werte der Zielfunktion an den gewählten Stützstellen bestimmt. Zwischen den Stütztellen werden alle Werte durch lineare Interpolation berechnet. Abbildung 3.22 zeigt die Annäherung einer Standardnormalverteilung mittels stückweise linearer Funktionen für 30 Stützstellen. Trotz der geringen Anzahl an Stützstellen sind die theoretische Funktion (ganze Linie) und die Annäherung (gestrichelte Linie) kaum zu unterscheiden.

2. Die relative kumulierte Summe der Funktionswerte wird gebildet. Zu diesem Zweck werden alle Werte bis zur n-ten Stützstelle aufsummiert und durch ihre Gesamtsumme geteilt[38]. Aufgrund dessen ist der *Grid Sampler* auch für Dichtefunktionen geeignet, welche bis auf eine Proportionalitätskonstante bekannt sind (unnormierte Dichtefunktionen).

   Es entsteht also eine Folge von Werten zwischen 0 und 1, welche in ihrer Interpretation einer empirischen Verteilungsfunktion entspricht. Aufgrund ihrer Linearität ist diese stückweise konstruierte Funktion leicht invertierbar. Die Inverse dieser Funktion liefert eine Approximation für die Quantilfunktion, welche die Ziehung von Stichproben aus der approximierenden Verteilung ermöglicht.

[38] Die ersten zwei Schritte sind analog zur numerischen Integration einer Dichtefunktion mittels der Trapezregel.

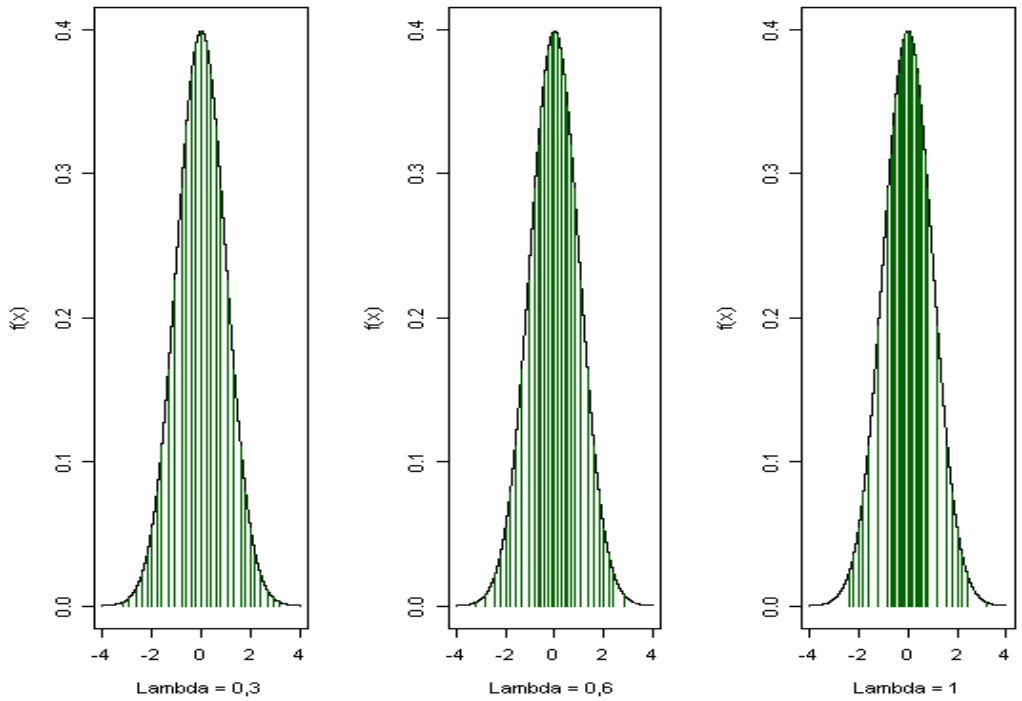

Abbildung 3.21: **Approximation einer Standardnormalverteilung:** Niedrige Werte des Parameters Lambda bewirken eine eher gleichmäßige Verteilung der Stützstellen, wohingegen große Werte eine starke Konzentration der Stützstellen in Regionen mit großer Krümmung verursachen. Simulationsergebnisse deuten darauf hin, dass Werte zwischen $\lambda = 0{,}5$ und $\lambda = 0{,}7$ für die meisten Anwendungen gut geeignet sind, um eine gute Verteilung der Stützstellen zu erzielen.

3. Um aus der approximierenden Funktion zu ziehen wird eine (0,1)-rechtecksverteilte Zufallszahl generiert. Mittels der Quantilfunktion wird der entsprechende Wert aus der linearen Approximation der Zielfunktion bestimmt. Abbildung (3.23) zeigt die Kerndichteschätzung einer mittels des *Grid Sampler*s generierten Stichprobe von Umfang n= 100.

Schließlich sei erwähnt, dass eine Vielzahl an Weiterentwicklungen existiert, welche durch adaptive Ansätze die Leistungsfähigkeit des *Grid Sampler*s, insbesondere in Verbindung mit dem *Gibbs Sampler*, erhöhen können. Einige dieser Ansätze wurden in der vorliegenden Arbeit verwendet und werden in Abschnitt 5.1.3.3 erläutert.

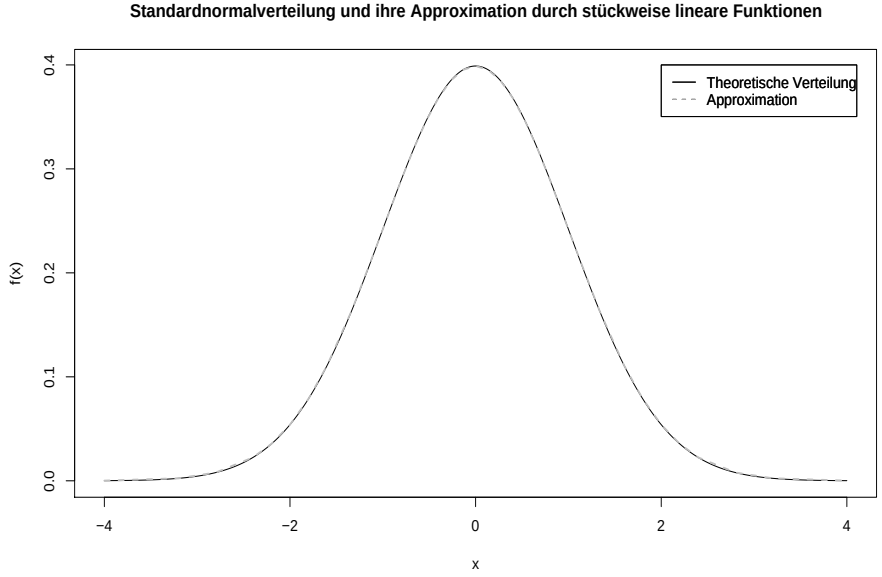

Abbildung 3.22: **Approximation einer Standardnormalverteilung:** Die gestrichelte Linie zeigt die Annäherung der theoretischen Verteilung (ganze Linie) mittels stückweise linearer Funktionen für 30 Stützstellen.

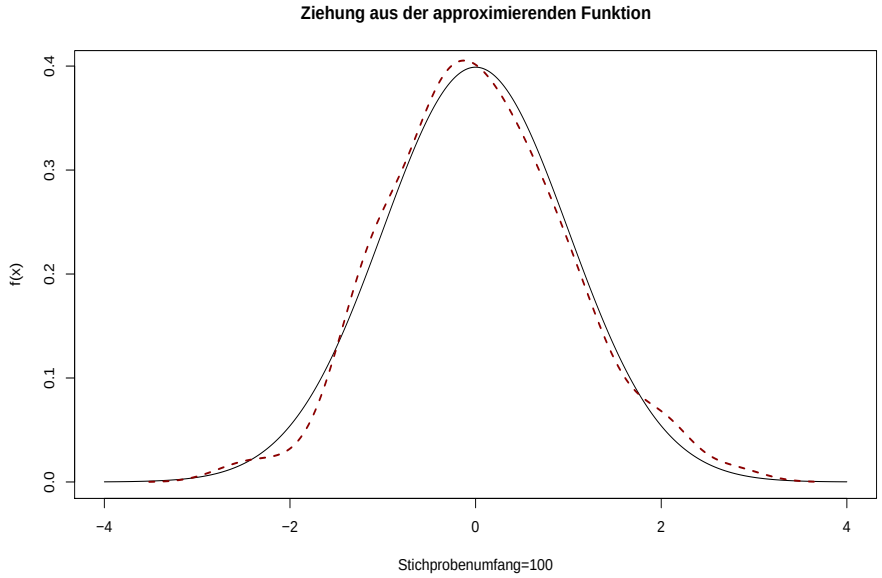

Abbildung 3.23: **Approximation einer Standardnormalverteilung:** Die gestrichelte Linie zeigt die Kerndichteschätzung einer Stichprobe vom Umfang $n = 100$ aus der Standardnormalverteilung (ganze Linie) generiert mittels *Grid Sampling*. Trotz des geringen Stichprobenumfangs ist die gezogene Stichprobe bereits eine gute Annäherung der wahren Verteilung.

Kapitel 4

Imputationsmodell für multivariat-normalverteilte Daten

Dieses und die folgenden Kapitel sind den algorithmischen Aspekten der Imputation gewidmet. Um bei den folgenden Ausführungen eine möglichst hohe Übersichtlichkeit zu gewährleisten, wird teilweise eine etwas abweichende, vereinfachte Notation verwendet.

Ziel dieses Kapitels ist es, die in Kapitel 2 und 3 behandelten Methoden, den EM-Algorithmus und die MCMC-Verfahren, auf die konkrete Problematik der Imputation anzuwenden.

4.1 EM-Algorithmus für multivariat-normalverteilte Daten

In diesem Abschnitt wird eine Version des EM-Algorithmus zur Schätzung der Parameter einer multivariaten Normalverteilung bei Vorhandensein nicht beobachteter Daten diskutiert. In Abschnitt 4.1.2 erfolgt dies von einem theoretischen Standpunkt aus. Anschließend werden in Abschnitt 4.1.3 die Details diskutiert, welche die Implementierung des EM-Algorithmus ermöglichen. Da sich der Algorithmus jedoch gewisse Eigenschaften der Normalverteilung zunutze macht, werden diese zunächst kurz vorgestellt.

4.1.1 Alternative Parameterisierung der Normalverteilung

Gegeben sei ein p-dimensionaler Zufallsvektor $\mathbf{Z} \sim MVN(\mu, \Sigma)$. Ferner sei $p_1 \in \mathbb{N}$ mit $p_1 < p$. Der Vektor $\mathbf{Z}$ kann dann in zwei Vektoren zerlegt werden, $\mathbf{Z}' = (\mathbf{Z}_1', \mathbf{Z}_2')$, wobei $\dim \mathbf{Z}_1 = p_1$ und $\dim \mathbf{Z}_2 = p - p_1$ gilt. D.h. $\mathbf{Z}_1$ besteht o.B.d.A aus den ersten p_1 und $\mathbf{Z}_2$ aus den letzten $p - p_1$ Elementen von $\mathbf{Z}$. Diese Partition des Vektors $\mathbf{Z}$ induziert für die Parameter der multivariaten Verteilung folgende Zerlegung:

$$\begin{aligned} \mu' &= (\mu_1', \mu_2'), \\ \Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \end{aligned}$$

wobei $E[\mathbf{Z}_i] = \mu_i$, $\text{Var}[\mathbf{Z}_i] = \Sigma_{ii}$ für i aus $\{1, 2\}$ und $\text{Cov}[\mathbf{Z}_1, \mathbf{Z}_2] = \Sigma_{12} = \Sigma_{21}'$.

Da $\mathbf{Z}$ multivariat-normalverteilt ist, ist die bedingte Verteilung von $\mathbf{Z}_2$ gegeben $\mathbf{Z}_1$ ebenfalls normal mit Parametern

$$E[\mathbf{Z}_2|\mathbf{Z}_1 = x] = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x - \mu_1) = \underbrace{(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1)}_{:=\alpha_{2\cdot1}} + \underbrace{(\Sigma_{21}\Sigma_{11}^{-1})}_{:=\beta_{2\cdot1}} x \tag{4.1}$$

und

$$\text{Var}[\mathbf{Z}_2|\mathbf{Z}_1 = x] = \underbrace{\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}_{:=\Sigma_{2\cdot1}}, \tag{4.2}$$

wobei $\alpha_{2\cdot1}$ der Vektor von Achsenabschnitten, $\beta_{2\cdot1}$ die Matrix der Steigungsparameter und $\Sigma_{2\cdot1}$ die Varianz-Kovarianzmatrix der Residuen der multivariaten Regression von $\mathbf{Z}_2$ auf $\mathbf{Z}_1$ darstellen.

Die Parameter in Gleichungen (4.1) und (4.2) können mittels des *Sweep*-Operators für alle möglichen Faktorisierungen von $\mathbf{Z}$ aus den (unbedingten) Parametern einer multivariaten Normalverteilung ermittelt werden. Die Eigenschaften dieses Operators werden im nächsten Abschnitt diskutiert.

4.1.2 *Sweep*-Operator

4.1.2.1 Allgemeines zum *Sweep*-Operator

Der *Sweep*-Operator ist ein vielseitig einsetzbarer Matrix-Operator, welcher in der Lage ist, Parameter von uni- und multivariaten linearen Regressionen mit Hilfe einer symmetrischen Parametermatrix auf sehr effiziente Art und Weise zu berechnen. Aufgrund dieser Effizienz ist er ein wertvolles Hilfsmittel für die Anwendung des EM-Algorithmus bei multivariat-normalverteilte Daten. Denn in diesem Fall können die bedingten Erwartungswerte mit Hilfe von linearen Regressionen exakt bestimmt werden (vgl. (4.1)).

Die außergewöhnliche Vielseitigkeit des *Sweep*-Operators zeigt sich z.B. dadurch, dass er nach kleinen Modifikationen zur

- Invertierung einer quadratischen Matrix,
- Generierung einer verallgemeinerten Inversen,
- Anwendung der Gauss-Jordan Eiliminationsmethode,
- Berechnung der Choleski Zerlegung und
- Berechnung von Determinanten

eingesetzt werden kann.

Als zusätzlichen Vorteil des *Sweep*-Operators hebt Goodnight (1979, S. 149) die Tatsache hervor, dass jedes Element einer *geswepten*[1] Matrix identifiziert werden kann und über statistische Bedeutung verfügt. Im Falle einer multivariaten Normalverteilung liefert dieser Operator die Parameter der bedingten Verteilung einer oder mehrerer Variabler gegeben die anderen Variablen.

Die Struktur und die wichtigsten Eigenschaften des *Sweep*-Operators werden im Folgenden kurz erläutert. Für eine anschauliche und ausführliche Einführung in den *Sweep*-Operator siehe Goodnight (1979). Der *Sweep*-Operator speziell im Zusammenhang mit dem EM-Algorithmus wird in Little und Rubin (2002) und Schafer (1997) behandelt.

Gegeben sei eine symmetrische $p \times p$ Matrix $G := (g_{ij})_{ij}$. Die *Sweeping*-Operation der Matrix G auf Position k ($k \in 1, \ldots, p$), wird als $SWP[k]$ bezeichnet und erzeugt eine $p \times p$ Matrix H:

$$SWP[k]\, G = H,$$

deren Elemente durch:

$$\begin{aligned} h_{kk} &:= -1/g_{kk} \\ h_{jk} = h_{kj} &:= g_{jk}/g_{kk} \qquad \text{für} \quad j \neq k \\ h_{jl} = h_{lj} &:= g_{jl} - g_{jk}g_{kl}/g_{kk} \qquad \text{für} \quad j \neq k \quad \text{und} \quad l \neq k \end{aligned}$$

definiert sind.

Wird der *Sweep*-Operator sukzessive auf alle p Positionen der Matrix G angewendet, so erhält man die negative Inverse von G[2]. D.h. es gilt

$$SWP[1, \ldots, p]\, G := SWP[1] \ldots SWP[p]\, G = -G^{-1}. \tag{4.3}$$

Eine für die Zwecke dieser Arbeit sehr nützliche Eigenschaft des *Sweep*-Operators ist seine Kommutativität

$$SWP[k_1]\, SWP[k_2]\, G = SWP[k_2]\, SWP[k_1]\, G$$

für alle $k_1 \neq k_2$ mit $k_1, k_2 \in \{1, \ldots, p\}$. Die Reihenfolge der *Sweep*s in Gleichung (4.3) beeinflusst also das Endergebnis nicht.

Die *Sweep*-Operation besitzt eine Inverse, welche als *Reverse Sweep*-Operation bezeichnet wird. Die Anwendung des *Reverse Sweep*-Operators auf eine $p \times p$ Matrix H und auf Position k liefert

$$RSWP[k]\, H = G,$$

[1] Engl.: *swept*. Die deutsch-englische Mischung *geswept* ist zwar grammatikalisch nicht korrekt, besitzt jedoch eine leichtere Aussprache als *gesweept* und ermöglicht, im Gegensatz zu *swept*, ihre Deklination.

[2] Diese Aussage setzt voraus, dass das *Sweeping* keine Division durch 0 verursacht.

mit den Elementen

$$\begin{aligned} g_{kk} &:= -1/h_{kk} \\ g_{jk} = g_{kj} &:= -h_{jk}/h_{kk} \qquad \text{für} \quad j \neq k \\ g_{jl} = g_{lj} &:= h_{jl} - h_{jk}h_{kl}/h_{kk} \qquad \text{für} \quad j \neq k \quad \text{und} \quad l \neq k. \end{aligned}$$

Es lässt sich leicht zeigen, dass

$$RSWP[k]\, SWP[k]\, G = G,$$

für alle $k \in 1, \ldots, p$ gilt. Ferner kann man zeigen, dass der *Reverse Sweep*-Operator ebenfalls kommutativ ist.

4.1.2.2 *Sweep*-Operator und EM-Algorithmus für normalverteilte Daten

Es seien mit $\theta = (\vartheta_{ij})_{ij} := \{\mu, \Sigma\}$ die Parameter einer multivariaten Normalverteilung bezeichnet. Die Elemente $\vartheta_{i,j}$ von θ müssen zunächst in die vom *Sweep*-Operator benötigte Struktur eingeordnet werden. Dazu betrachtet man die *erweiterte* Varianz-Kovarianzmatrix:

$$\theta = \begin{bmatrix} -1 & \mu' \\ \mu & \Sigma \end{bmatrix} = \begin{bmatrix} -1 & \mu_1' & \mu_2' \\ \mu_1 & \Sigma_{11} & \Sigma_{12} \\ \mu_2 & \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

wobei das Element $\vartheta_{1,1}$ in der Regel als Position 0 bezeichnet wird[3]. Die *Sweep*-Operationen von θ auf die Positionen $1, \ldots, p_1$ liefern dann das Ergebnis

$$SWP[1, \ldots, p_1]\, \theta = \begin{bmatrix} -1 - \mu_1'\Sigma_{11}^{-1}\mu_1 & \mu_1'\Sigma_{11}^{-1} & \mu_2' - \mu_1'\Sigma_{11}^{-1}\Sigma_{12} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} & \Sigma_{11}^{-1}\Sigma_{12} \\ \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 & \Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{bmatrix}.$$

Diese *geswepte* Matrix enthält die Parameter der bedingten Verteilung von $\mathbf{Z}_2$ gegeben $\mathbf{Z}_1$ und eine $(p_1 + 1) \times (p_1 + 1)$ Submatrix mit den Parametern der Randverteilung von $\mathbf{Z}_1$

$$\begin{bmatrix} -1 & \mu_1' \\ \mu_1 & \Sigma_{11} \end{bmatrix} \Rightarrow \begin{bmatrix} -1 - \mu_1'\Sigma_{11}^{-1}\mu_1 & \mu_1'\Sigma_{11}^{-1} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} \end{bmatrix}$$

in *geswept*er Form. Das Element $\vartheta_{1,1}$ ist gleich -1. Dies liegt darin begründet, dass θ in seiner ursprünglichen Form als eine *bereits auf Position 0 geswepte Matrix angenommen wird*. Die *Reverse Sweep*-Operation von θ auf Position 0 liefert

$$RSWP[0]\, \theta = \begin{bmatrix} 1 & \mu' \\ \mu & \Sigma + \mu\mu' \end{bmatrix} \tag{4.4}$$

[3] Man beachte die leicht abweichende Notation gegenüber dem vorhergehenden Abschnitt, in dem die erste Position mit 1 gekennzeichnet wurde.

liefert die Parameter der multivariaten Normalverteilung parametrisiert nach den ersten zwei unzentrierten Momenten. Diese *ungeswepte* Form der Matrix[4] wird zur Bestimmung der Maximum-Likelihood-Schätzer benutzt.

Bisher wurde die Zerlegung von $\mathbf{Z}$ derart vorgenommen, dass $\mathbf{Z}_1$ die *ersten* p_1 Elemente enthielt. Diese Situation wird nun verallgemeinert.

Das Ergebnis eines einfachen oder mehrfachen *Sweep*s is eine Matrix H mit vier Untermatrizen $-A, B, B'$ und C, deren Elemente i.d.R. keine leicht erkennbare Ordnung besitzen. Um diese Untermatrizen identifizieren zu können, wird folgende Notation eingeführt, welche bereits auf die Problematik der fehlenden Daten hindeutet:

- Mit $\mathcal{O}$ (observed) seien die Positionen in der erweiterten Varianz-Kovarianzmatrix gekennzeichnet, welche zu Spalten der ursprünglichen Datenmatrix gehören, deren Elemente beobachtet werden.
- Mit $\mathcal{M}$ (missing) seien die Positionen in der erweiterten Varianz-Kovarianzmatrix gekennzeichnet, welche zu Spalten der ursprünglichen Datenmatrix gehören, die fehlende Daten aufweisen.

Diese Notation wird sich im nächsten Abschnitt als nützlich erweisen. Für die folgende Behandlung der Algorithmen ist es ausreichend, $\mathcal{O}$ und $\mathcal{M}$ als Mengen zu betrachten, mit den beiden Eigenschaften $\mathcal{O} \cup \mathcal{M} = \{1, \ldots, p\}$ und $\mathcal{O} \cap \mathcal{M} = \emptyset$.

Beschreibung der Untermatrizen:

- B: Matrix der Regressionskoeffizienten der multiplen Regression der nicht-*geswepten* Variablen auf die *geswepten*. Die Elemente dieser Matrix werden definiert als:

  $$b_{i,j} := \{h_{ij} \,|\, i \in \mathcal{O} \text{ und } j \in \mathcal{M}\}\,.$$

  Durch B' ist die transponierte Matrix von B gegeben.

- C: Varianz-Kovarianzmatrix der Residuen. Da es sich i.d.R. um eine multivariate Regression handelt, ist C eine Matrix und kein Skalar.

  $$c_{i,j} := \{h_{ij} \,|\, i, j \in \mathcal{M}\}\,.$$

- $-A$: Matrix der invertierten zweiten Momente der Regressoren. Diese Matrix liefert durch Multiplikation mit den jeweiligen Varianzen der Residuen die Varianzen und Kovarianzen der geschätzten Regressionsparameter.

  $$-a_{i,j} := \{h_{ij} \,|\, i, j \in \mathcal{O}\}\,.$$

  Da für die verschiedenen abhängigen Variablen die Regressoren immer dieselben sind, wird die Matrix A mit den jeweiligen Hauptdiagonalelementen der Matrix C multipliziert, um die Standardfehler der Regressionskoeffizienten zu bestimmen.

Anhand der behandelten Verfahren wird im nächsten Abschnitt die Implementierung des EM-Algorithmus für normalverteilte Daten mit fehlenden Werten erläutert.

[4] Man beachte, dass die Elemente dieser Matrix unbedingte Momente sind. Aufgrund dessen wird sie als *ungeswept* bezeichnet.

4.1.3 Implementierung des EM-Algorithmus

4.1.3.1 Überblick

Gegeben sei eine Datenmatrix Y mit n unabhängigen Realisationen aus einer multivariaten Normalverteilung $MVN(\mu, \Sigma)$[5]. Die multivariate Normalverteilung gehört zur Exponentialfamilie. Wie bereits in Abschnitt 2.1.6 erläutert, ist die Loglikelihood Funktion für Verteilungen aus der Exponentialfamilie nicht linear in den Daten, sondern in den suffizienten Statistiken für die Parameter. Die Bestimmung der Erwartungswerte dieser suffizienten Statistiken ist das Ziel des E-Schritts des EM-Algorithmus.

Die suffizienten Statistiken beider Parameter der multivariaten Normalverteilung können wie folgt dargestellt werden: $T^{(1)} := Y'\mathbf{1}$ mit $\mathbf{1} := (1, 1, \ldots, 1)'$, und $T^{(2)} := Y'Y$. Diese beiden Statistiken werden oftmals in eine $(p+1) \times (p+1)$ Matrix T angeordnet werden

$$T = \begin{bmatrix} n & T^{(1)'} \\ T^{(1)} & T^{(2)} \end{bmatrix}. \tag{4.5}$$

Damit der *Sweep*-Operator verwendet werden kann, müssen die Parameter μ und Σ in Form einer erweiterten Varianz-Kovarianzmatrix vorliegen (vgl. Abschnitt 4.1.2.2)

$$\theta = \begin{bmatrix} -1 & \mu' \\ \mu & \Sigma \end{bmatrix}.$$

Die ML-Schätzung in diesem Modell kann dann mittels eines *Reverse Sweep*-Operators (vgl. Gleichung (4.4)) dargestellt werden. Die ML-Schätzer für die beiden Parameter μ und Σ ergeben sich als Lösung der Gleichung:

$$RSWP[0]\,\theta = n^{-1}T.$$

Aus den Eigenschaften des *Sweep*-Operators folgt

$$\hat{\theta} = SWP[0]\,n^{-1}T. \tag{4.6}$$

Dieses Resultat kann verwendet werden, um eine kompakte Darstellung des EM-Algorithmus für normalverteilte Daten zu erhalten. Zu diesem Zweck werden die beiden Schritte des EM-Algorithmus in folgender Gleichung zusammengefasst

$$\hat{\theta}_{t+1} = SWP[0]\,n^{-1}\,E\left[T \mid Y_{obs}, \hat{\theta}_t\right], \tag{4.7}$$

wobei $\hat{\theta}_t$ und $\hat{\theta}_{t+1}$ zwei aufeinanderfolgende Parameterschätzungen für θ darstellen[6]. Der Erwartungswert auf der rechten Seite von Gleichung (4.7) wird durch die lineare Regression von Y_{mis} auf Y_{obs} ermittelt, deren Bestimmung von Parameter θ abhängig ist. Die Bestimmung von θ erfolgt mittels des *Sweep*-Operators und stellt das Ziel des M-Schritts dar.

In den nächsten Abschnitten wird im Detail die Implementierung des EM-Algorithmus erläutert.

[5] Um die Notation möglichst einfach zu halten bezeichnet Y sowohl die Zufallsvariable als auch die Wertematrix ihrer Realisationen.

[6] Im Folgenden wird der Schätzwert $\hat{\theta}$ für θ ebenfalls mit θ bezeichnet.

4.1.3.2 Preliminäre Datenvorbereitung

Der hier betrachtete EM-Algortihmus ist in der Lage, die Parameter einer multivariaten Normalverteilung aus Daten zu schätzen, welche ein allgemein gehaltenes Muster an fehlenden Daten (MP)[7] aufweisen. Dies impliziert, dass der Algorithmus bei komplexen Mustern eine sehr hohe Anzahl an Rechenschritten durchführen muss. Es ist also aus Effizienzgründen ratsam, die Daten derart zu sortieren, dass die Anzahl an benötigten *Sweep*-Operationen minimiert wird. Dies geschieht dadurch, dass die Zeilen der Datenmatrix nach ihrem Muster an fehlenden Daten gruppiert werden. Zu diesem Zweck werden die folgenden Schritte vorgenommen:

(a) Eine Indikator Matrix R mit Dimension $n \times p$ wird eingeführt, deren Elemente folgendermaßen definiert sind

$$r_{ij} := \begin{cases} 1, & \text{wenn } y_{ij} \text{ beobachtet wird} \\ 0, & \text{wenn } y_{ij} \text{ nicht beobachtet wird,} \end{cases}$$

für alle $i \in \{1, \ldots, n\}$ und $j \in \{1, \ldots, p\}$. Dies wird in Abbildung 4.1 veranschaulicht. Diejenigen Zeilen, welche lediglich aus fehlenden Daten bestehen, müssen eliminiert werden, denn sie tragen nicht zur beobachteten Likelihood bei und verlangsamen die Konvergenz des EM-Algorithmus, indem sie den Anteil an fehlenden Informationen erhöhen.

Die so definierte Matrix R hat die nützliche Eigenschaft, die eindeutige Identifikation der verschiedenen MPs zu ermöglichen. Dies liegt darin begründet, dass ihre Zeilen aus Nullen und Einsen bestehen, welche, gemeinsam betrachtet, als binäre Darstellung einer Zahl aufgefasst werden können. Diese Zahl charakterisiert ein MP eindeutig und ermöglicht somit die Gruppierung der Zeilen nach ihren Mustern. Die *Sweep*-Operation muss lediglich einmal pro MP und nicht für alle Zeilen durchgeführt werden. Bei Datensätzen, in denen die Anzahl an Zeilen wesentlich größer als die Anzahl an Spalten ist, ist die Effizienzgewinnung von einem beträchtlichen Ausmaß.

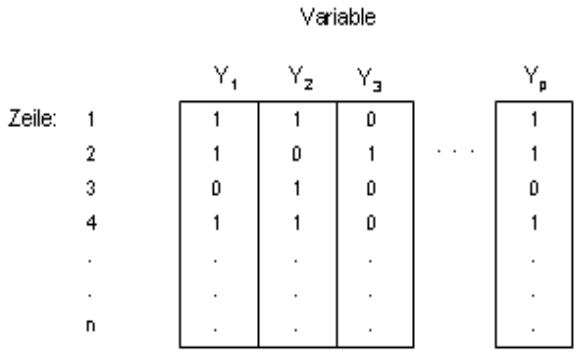

Abbildung 4.1: Die R Matrix. Die Zeilen eins und vier weisen das gleiche MP auf.

[7] Vgl. 1.2.2. Diese Bezeichnung wird aus Übersichtlichkeitsgründen sowohl für das Muster eines ganzen Datensatzes als auch für das Muster einer Zeile verwendet, wenn aus dem Kontext ersichtlich ist, ob ein ganzer Datensatz oder eine Zeile gemeint ist.

Es sei nun mit S die Anzahl an MPs gekennzeichnet. Nach der Gruppierung besteht die Datenmatrix aus S Untermatrizen. Die Untermatrizen werden mit $s = 1, \ldots, S$ indiziert. Dies wird in Abbildung 4.2 veranschaulicht.

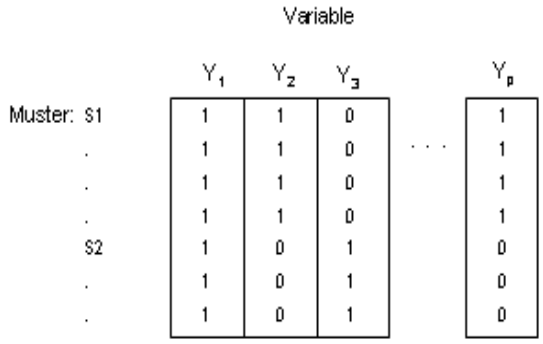

Abbildung 4.2: Die sortierte R Matrix. Alle Zeilen mit dem gleichen MP werden gruppiert.

(b) Eine zusätzliche Verdichtungsstufe wird vorgenommen, indem eine Indikator-Matrix R^* mit Dimension $S \times p$ konstruiert wird, welche lediglich die verschiedenen MPs beinhaltet (siehe Abbildung 4.3). Die Elemente dieser Matrix werden folgendermaßen definiert

$$r^*_{sj} := \begin{cases} 1, & \text{wenn } y_j \text{ in der Gruppe } s \text{ beobachtet wird} \\ 0, & \text{wenn } y_j \text{ in der Gruppe } s \text{ nicht beobachtet wird.} \end{cases}$$

Für alle $s \in \{1, \ldots, S\}$ und $j \in \{1, \ldots, p\}$.

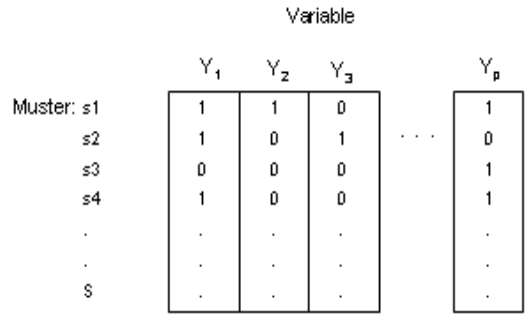

Abbildung 4.3: Die R^* Matrix. Lediglich die verschiedenen MPs werden berücksichtigt.

Die beiden Matrizen R und R^* enthalten alle Kennzahlen, welche zur effizienten Implementierung des EM-Algorithmus benötigt werden.

Analog zum vorhergehenden Abschnitt wird die Notation $\mathcal{O}(s)$ und $\mathcal{M}(s)$ verwendet, um die beobachteten und fehlenden Variablen[8] für alle S Gruppen zu kennzeichnen. Dabei wird die Zugehörigkeit der Kennzahl zu einer Gruppe durch (s) hervorgehoben. Die betrachteten Mengen sind also

$$\mathcal{O}(s) = \{j : r_{sj} = 1\},$$
$$\mathcal{M}(s) = \{j : r_{sj} = 0\}.$$

Schließlich seien in der Menge $\mathcal{I}(s) \subseteq \{1, \ldots, n\}$ die Zeilen der ursprünglichen Datenmatrix enthalten, welche das s-te MP aufweisen.

4.1.3.3 Erwartungswertschritt (E-Schritt)

Die vorgenommene Behandlung der Daten ermöglicht die Berechnung des Erwartungswertes der suffizienten Statistiken der vollständigen Daten bzgl. $P(Y_{mis}|Y_{obs}, \theta)$[9] für einen gegebenen Wert von θ. Diese Statistiken haben für eine Normalverteilung bekanntlich die Form $\sum_i y_{ij}$, $\sum_i y_{ij}^2$ und $\sum_i y_{ij}y_{ik}$. Der E-Schritt besteht also darin, dass die Erwartungswerte von y_{ij}, y_{ij}^2 und $y_{ij}y_{ik}$ bzgl. $P(Y_{mis} \mid Y_{obs}, \theta)$ bestimmt werden.

Aufgrund der Annahme der Unabhängigkeit der Zeilen gilt

$$P(Y_{mis} \mid Y_{obs}, \theta) = \prod_{i=1}^{n} P(y_{i(mis)} \mid y_{i(obs)}, \theta) \quad \text{für } i = 1, \ldots, n. \tag{4.8}$$

Dabei bezeichnen $y_{i(mis)}$ und $y_{i(obs)}$ die Untervektoren von y_i mit jeweils beobachteten und fehlenden Werten.

Die Verteilung $P(y_{i(mis)} \mid y_{i(obs)}, \theta)$ ist multivariat-normal, wobei der Erwartungswert durch die multivariate lineare Regression von $y_{i(mis)}$ auf $y_{i(obs)}$ gegeben ist (vgl. Gleichung (4.1)). Die Regressionsparameter können mittels einer *Sweep*-Operation der Matrix θ auf die Positionen der Variablen in $y_{i(obs)}$ bestimmt werden und leicht anhand der in Abschnitt 4.1.2.2 eingeführten Notation in der *geswepten* Matrix identifiziert werden.

Dieses Vorgehen kann kompakt wie folgt dargestellt werden

$$A = SWP[\mathcal{O}(s)]\theta, \tag{4.9}$$

wobei A die *geswepte* Matrix bezeichnet.

Die Erwartungswertbildung der suffizienten Statistiken für eine allgemeine Zeile i mit Muster (MP) s setzt sich wie folgt zusammen:

[8] Aufgrund der Gruppierung sind die Spalten innerhalb einer Gruppe entweder völlig beobachtet oder fehlend. Deswegen werden sie jetzt als beobachtete bzw. fehlende „Variable“ bezeichnet.

[9] Der vereinfachte Ausdruck $P(Y_{mis} \mid Y_{obs}, \theta)$ steht für $f_{Z|Y,\theta}(z|y, \vartheta)$. Da in diesem Abschnitt das Hauptaugenmerk auf der Implementierung des Algorithmus liegt, wird die Notation möglichst einfach und intuitiv gehalten.

- Beobachtete Werte (Spalten in $\mathcal{O}(s)$) besitzen folgende (erste und zweite) Momente

$$\begin{aligned} E[y_{i(obs)} \mid Y_{obs}, \theta] &= y_{i(obs)} \\ \mathrm{Cov}[y_{i(obs)}, y_{ik} \mid Y_{obs}, \theta] &= 0 \quad \text{für alle } k \in \{1, \ldots, p\}. \end{aligned}$$

  Dies folgt unmittelbar aus der Tatsache, dass $y_{i(obs)}$ als Konstante betrachtet wird.

- Unbeobachtete Werte (Spalten in $\mathcal{M}(s)$) besitzen folgende erste und zweite Momente

$$\begin{aligned} E[y_{i(mis)} \mid Y_{obs}, \theta] &= a_{0(mis)} + \sum_k a_{(obs)(mis)} y_{i(obs)} \\ \mathrm{Cov}[y_{i(mis_l)}, y_{i(mis_m)} \mid Y_{obs}, \theta] &= a_{(mis_l)(mis_m)} \quad \text{für } l, m \in \mathcal{M}(s). \end{aligned}$$

  Die Koeffizienten $a_{0(mis)}$ und a_{kj} können dabei leicht aus der *geswepten* Matrix (siehe Abschnitt 4.1.2.2) abgelesen werden.

Unter Berücksichtigung folgender Beziehung

$$E[y_{ij}, y_{ik} \mid Y_{obs}, \theta] = \mathrm{Cov}[y_{ij}, y_{ik} \mid Y_{obs}, \theta] + E[y_{ij} \mid Y_{obs}, \theta] E[y_{ik} \mid Y_{obs}, \theta],$$

wobei

$$\mathrm{Cov}[y_{ij}, y_{ik} \mid Y_{obs}, \theta] = \begin{cases} 0, & \text{für } j \in \mathcal{O}(s) \\ a_{jk}, & \text{für } j, k \in \mathcal{M}(s) \end{cases}$$

erhält man für y_{ij} und $y_{ij} y_{ik}$

$$E[y_{ij} \mid Y_{obs}, \theta] = \begin{cases} y_{ij}, & \text{für } j \in \mathcal{O}(s) \\ y^*_{ij}, & \text{für } j \in \mathcal{M}(s) \end{cases}$$

bzw.

$$E[y_{ij} y_{ik} \mid Y_{obs}, \theta] = \begin{cases} y_{ij} y_{ik}, & \text{für } j, k \in \mathcal{O}(s) \\ y^*_{ij} y_{ik}, & \text{für } j \in \mathcal{M}(s) \text{ und } k \in \mathcal{O}(s) \\ a_{jk} + y^*_{ij} y^*_{ik}, & \text{für } j, k \in \mathcal{M}(s) \end{cases}$$

mit

$$y^*_{ij} = a_{0j} + \sum_k a_{kj} y_{ik}.$$

Aufgrund seiner Linearität ist der Erwartungswert der suffizienten Statistiken gleich der Summe der einzelnen, wie oben aufgeführt berechneten Erwartungswerte. Die Summation erfolgt zuerst innerhalb der MPs und in einem zweiten Schritt werden diese Summen addiert. Das Ergebnis des E-Schrittes kann dann kompakt dargestellt werden als

$$E\left[T \mid Y_{obs}, \theta\right],$$

wobei die Matrix T die suffizienten Statistiken enthält (vgl. (4.5)). Anhand dieser Erwartungswerte können nun die ML-Schätzwerte bestimmt werden.

4.1.3.4 Maximierungsschritt (M-Schritt)

Die Bestimmung des ML-Schätzers für θ erfordert lediglich die Anwendung von Gleichung (4.6), d.h. eine *Sweep*-Operation auf Position 0, auf die Matrix $E\,[\,T \mid Y_{obs}, \theta\,]$.

Der neue E-Schritt erfolgt wie oben dargestellt, wobei die *Sweep*-Operationen auf die im Maximierungschritt berechnete Matrix θ_1 anstelle der initialisierenden Matrix θ_0 angewendet werden. Die E- und M-Schritte werden fortgeführt bis ein geeignetes Abbruchkriterium erreicht wird. Für geeignete Abbruchkriterien siehe Abschnitt (2.1.4).

4.1.3.5 Wahl der Startwerte

In Abschnitt 4.1.3.3 wurde die Berechnung der Erwartungswerte der suffizienten Statistiken für ein gegebenes θ beschrieben. Da die Erwartungswertbildung bereits vom Parameter θ abhängt, muss ein geeigneter Startwert θ_0 gewählt werden. Auf die Wahl dieses Startwertes wird hier kurz eingegangen.

Aufgrund der Tatsache, dass die vom *Sweep*-Operator verwendete Matrix sich aus der Varianz-Kovarianzmatrix und dem Vektor der Mittelwerte der Daten zusammensetzt, ist es naheliegend, die empirischen Momente der beobachteten Daten

$$\theta^{(0)} = \begin{bmatrix} -1 & \bar{x}_{obs}^{(0)\prime} \\ \bar{x}_{obs}^{(0)} & \Sigma_{obs}^{(0)} \end{bmatrix}$$

als Startwerte zu verwenden. Dabei bezeichnet $\bar{x}_{obs}^{(0)}$ den Vektor der arithmetischen Mittel der beobachteten Daten und $\Sigma_{obs}^{(0)}$ ihre Varianz-Kovarianzmatrix.

Ein einfacherer, alternativer Ansatz besteht darin, dass für den ersten Schritt die Varianzen als Eins und die Kovarianzen als Null angenommen werden. Beide Methoden wurden ausführlich in Simulationen getestet. Obwohl die Verwendung der empirischen Momente der beobachteten Daten bei einem geringen Anteil an fehlenden Werten die Anzahl an benötigten Iterationen verringert, ergaben die Simulationen in realistischen Situationen keine nennenswerten Unterschiede zwischen beiden Ansätzen. Little und Rubin (2002) behandeln zusätzliche Möglichkeiten zur Wahl der Startwerte.

4.1.3.6 Ein einfaches Beispiel

Das folgende Beispiel veranschaulicht die Überlegenheit des EM-Algorithmus gegenüber einer KQ-Regression im Falle einer bivariaten Normalverteilung bei einem hohen Anteil an fehlenden Werten. Die ersten zwei Spalten der Tabelle 4.1 beinhalten Beobachtungen aus einer bivariaten Normalverteilung. Die letzten zwei Spalten der Tabelle enthalten die selben Ausgangsdaten, nachdem 40% der Werte eliminiert wurden. Die vollständig beobachteten Wertepaare werden in Abbildung 4.4 als grüne Punkte gekennzeichnet, während die gelöschten Werte als graue Kreise

eingezeichnet sind. Da eine KQ-Regression vollständig beobachtete Daten zur Schätzung der Parameter benötigt, ist sie nicht in der Lage, die unvollständig beobachteten Daten zu verarbeiten. Aufgrund dessen verfügt diese Schätzmethode nicht über 60% sondern lediglich über 10% der ursprünglichen Daten. Dies spiegelt sich in einer ungenauen Schätzung der Parameter wider, wie in Abbildung 4.4 (blaue Linie) veranschaulicht wird.

Der EM-Algorithmus ist dagegen in der Lage, unvollständige Daten mit einzubeziehen. Die Genauigkeit der Schätzung der Regressionsparameter ist daher höher. Dies ist in Abbildung 4.4 (gelbe Linie) erkennbar.

Vollständig beobachtete Daten		Datensatz mit fehlenden Werten	
-0.2317790	1.5645425	-0.2317790	NA
1.3977843	3.4177385	NA	3.4177385
1.0965123	1.5624080	NA	1.5624080
4.2057202	5.0685650	4.2057202	NA
5.1601624	6.3115680	NA	6.3115680
-0.6176621	0.3965697	-0.6176621	0.3965697
2.1914055	2.5698609	NA	2.5698609
0.6822623	0.8115055	NA	0.8115055
0.8397043	1.0817316	0.8397043	1.0817316
2.4754820	3.5060047	NA	3.5060047
0.2276409	0.2446433	0.2276409	NA
-1.1512896	1.8198668	-1.1512896	NA
0.8433541	3.0730297	NA	3.0730297
1.8972929	2.8986880	1.8972929	2.8986880
3.5244921	4.0894709	NA	4.0894709
2.4650780	5.0531604	2.4650780	NA
1.9660875	3.2031609	1.9660875	NA
1.0907652	3.0170962	1.0907652	NA
-0.4410336	3.7961236	-0.4410336	3.7961236
-1.1765625	-1.7791738	-1.1765625	NA

Tabelle 4.1: Datensatz vor und nach der Entfernung von Beobachtungen. Lediglich vier Datenpaare werden vollständig beobachtet.

Ein realistischeres Beispiel wurde bereits im Zusammenhang mit dem Data-Augmentation-Algorithmus in Abschnitt 3.4.6.5 ausführlich behandelt.

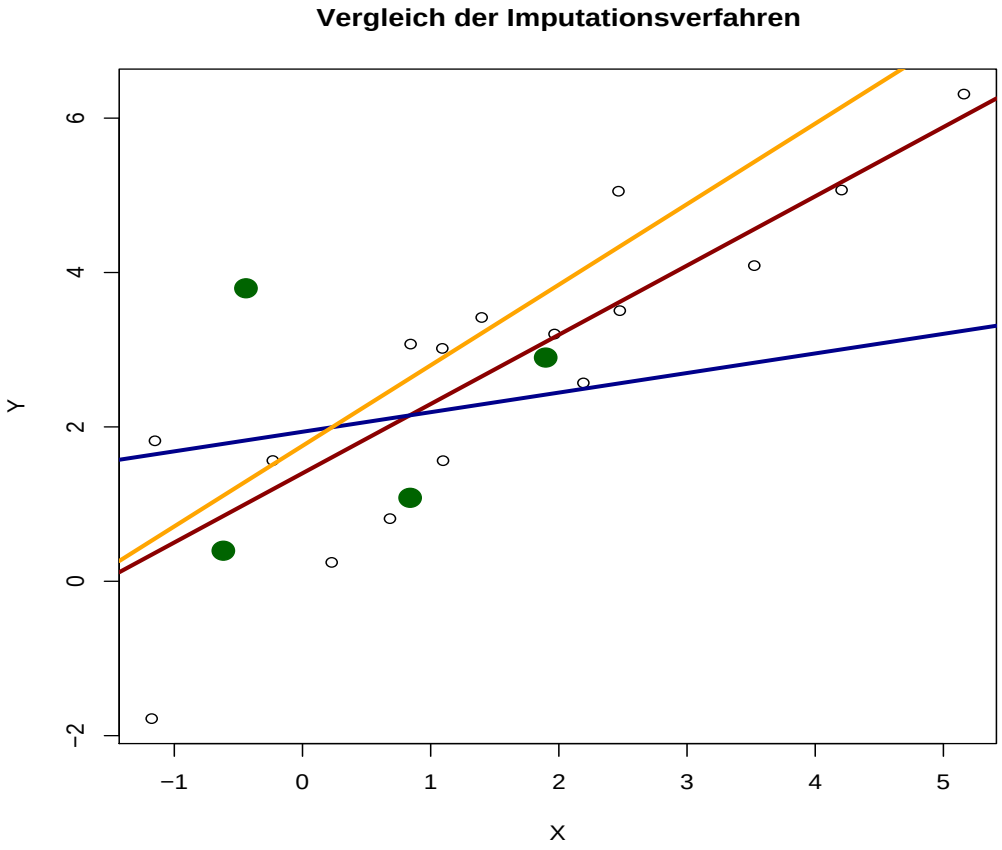

Abbildung 4.4: Vergleich einer KQ-Regression der beobachteten Daten mit dem EM-Algorithmus. Die braune Linie bezeichnet die Regressionsgerade bei vollständig beobachteten Daten. Diese wurde eingezeichnet, um den Vergleich der Methoden zu ermöglichen. Die blaue Linie wurde mittels einer KQ-Regression basierend auf den vier vollständig beobachteten Paaren bestimmt. Die Parameter der gelben Linie wurden mittels des EM-Algorithmus geschätzt. Die Überlegenheit des EM-Algorithumus unter den o.g. Bedingungen ist offensichtlich.

4.2 DA-Algorithmus für multivariat-normalverteilte Daten

4.2.1 Implementierung des DA-Algorithmus

4.2.1.1 I-Schritt

Aufgrund der Tatsache, dass die Zeilen der Datenmatrix Y als bedingt unabhängig gegeben θ angenommen werden, kann der Vektor $y_{i(mis)}$ mit den fehlenden Werten der i-ten Zeile durch eine unabhängige Ziehung aus der folgenden Verteilung erzeugt werden:

$$y_{i(mis)}^{(t+1)} \sim P(y_{i(mis)} \mid y_{i(obs)}, \theta^{(t)}).$$

Der I-Schritt besteht also in der Generierung, für jede Zeile der Datenmatrix Y mit fehlenden Werten, von unabhängigen Vektoren aus der bedingten Verteilung der fehlenden Daten gegeben die beobachteten Daten und die aktuellen Parameterwerte.

Im multivariaten normalen Modell ist die Verteilung für die i-te Zeile im s-ten MP multivariat-normal mit Erwartungswert

$$E[y_{ij} \mid Y_{obs}, \theta] = a_{0j} + \sum_{k \in \mathcal{O}(s)} a_{kj} y_{ik} \tag{4.10}$$

und Varianz-Kovarianzmatrix[10]

$$\mathrm{Cov}[y_{ij}, y_{im} \mid Y_{obs}, \theta] = a_{jm} \quad \text{für } j, m \in \mathcal{M}(s). \tag{4.11}$$

Dabei können die Koeffizienten a_{0j}, a_{kj} und a_{jm} aus der *geswepten* Matrix abgelesen werden (siehe Abschnitt 4.1.2.2).

4.2.1.2 P-Schritt

***A priori*-Verteilung:** Wie in der Einführung von Kapitel 3 erläutert, haben die MCMC-Methoden entscheidend zur explosionsartigen Entwicklung der Bayes Statistik in den letzten Jahren beigetragen. Dennoch besitzt die mathematische Herleitung dieser Algorithmen keine bayesianischen Elemente[11]. Die MCMC-Verfahren sind in der Lage, iterativ hochdimensionale Zufallsvariable zu simulieren, deren gemeinsame Dichte- bzw. Verteilungsfunktion in den meisten Fällen nicht analytisch darstellbar ist. Die Simulation einer Markov-Kette, welche die gewünschte Zielverteilung als invariantes Maß besitzt, lässt auf den ersten Blick keine Verbindung zur Bayes Statistik erkennen.

Der Beitrag der Bayes Statistik besteht in der Betrachtung unbekannter Parameter als Zufallsvariable, deren Verteilung die verfügbaren Informationen über diese Parameter widerspiegeln.

Anhand der Daten wird versucht, genauere Erkenntnisse über diese Parameter zu gewinnen. Diese Informationsgewinnung kann sukzessiv erfolgen, wodurch sich die Bayes Statistik von der frequentistischen Statistik unterscheidet.

[10] Diese Momente sind analog zum EM-Algorithmus für normalverteilte Daten definiert (vgl. Abschnitt 4.1.3.3).
[11] Lediglich die Bezeichnung Posterior-Schritt des DA-Algorithmus deutet auf die Bayes Statistik hin.

Vor der Analyse der vorhandenen Daten liegt eine *a priori*-Verteilung vor. Die Daten liefern neue Informationen über die Parameter. Diese werden üblicherweise durch die Likelihood Funktion ausgedrückt. D.h. mit Hilfe der Likelihood wird die *a priori*-Verteilung modifiziert. Das Ergebnis der Kombination der *a priori*-Informationen mit der Likelihood wird als *a posteriori*-Verteilung bezeichnet.

Wenn die MCMC-Methoden zur Simulation einer Parameterverteilung eingesetzt werden, müssen jedoch diese Algorithmen gemäß der Regeln der Bayes Statistik behandelt werden. Da die in dieser Arbeit im Detail behandelte Version des DA-Algorithmus im t-ten Schritt einen Kandidaten $\theta^{(t)}$ aus der bedingten Verteilung des Parameters gegeben die beobachteten Daten Y_{obs} und die aktuelle Schätzung der fehlenden Daten $Y_{mis}^{(t)}$ generiert, d.h.

$$\theta^{(t)} \sim P(\theta \mid Y_{obs}, Y_{mis}^{(t)}),$$

liegt die Simulation einer Parameterverteilung und somit eine Schätzung im Rahmen der Bayes Statistik vor. Die aus den Daten gewonnenen Informationen müssen also mit einer *a priori*-Verteilung kombiniert werden, deren Gestalt je nach Kenntnisgrad über die Parameter sehr stark variieren kann.

Im normalen Modell und und bei Nichtvorhandensein von *a priori*-Informationen wird häufig die bekannte Jeffreys-Verteilung[12]

$$P(\mu, \Sigma) \propto |\Sigma|^{\frac{-(p+1)}{2}} \tag{4.12}$$

verwendet, wobei p die Anzahl der Variablen ist. Diese Verteilung hat folgende nützliche Eigenschaften:

(a) Sie ist nicht informativ. D.h. die Gestalt der *a posteriori*-Verteilung wird von der Likelihood und somit ausschließlich von den Daten geprägt.
(b) Die resultierende *a posteriori*-Verteilung kann auf eine einfache Art und Weise simuliert werden.

***A posteriori*-Verteilung:** Wie in Schafer (1997, Abschnitte 5.2.2 und 5.2.3) und Little und Rubin (2002, Kap. 11) beschrieben, ist die *a posteriori*-Verteilung der vervollständigten Daten $P(\theta \mid Y_{obs}, Y_{mis})$, unter Verwendung einer *Jeffreys a priori*-Verteilung, eine normal-inverse-Wishart-Verteilung. Der P-Schritt besteht also im Wesentlichen darin, eine geeignet skalierte normal-inverse-Wishart-Verteilung zu simulieren und aus dieser Verteilung die Parameter zu ziehen:

$$\Sigma | Y_{obs}, Y_{mis} \sim W^{-1}(n-1, S),$$
$$\mu \mid \Sigma, Y_{obs}, Y_{mis} \sim N(\bar{y}, n^{-1}\Sigma),$$

wobei $W^{-1}(n-1, S)$ die inverse-Wishart-Verteilung mit $n-1$ Freiheitgraden und Skalierungsmatrix S darstellt (siehe Gelman et al. (2004)). Der Vektor $\bar{y}$ und die Skalierungsmatrix S können mit Hilfe der *a priori*-Verteilung, der beobachteten Daten und den imputierten Werten $Y_{mis}^{(t)}$ des vorigen I-Schritts bestimmt werden.

Um eine Wishart-Verteilung zu simulieren, wird eine obere Dreiecksmatrix B folgendermaßen zusammengestellt:

[12] Die Jeffreys-Verteilung stellt die bekannteste *a priori*-Verteilung für normalverteilte Daten dar (vgl. Gelman et al. (2004, S. 66)).

(a) Die Elemente auf der Hauptdiagonalen sind $\chi^2_{(n-j)}$-verteilt. Dabei ist $j = 1, \ldots, p$ und n der Stichprobenumfang.

(b) Die Elemente oberhalb der Hauptdiagonalen sind standard-normalverteilt.

Für die Matrix $B'B$ gilt $B'B \sim W(n, I)$, wobei I die Einheitsmatrix bezeichnet. Anschließend wird die Matrix

$$M = (B')^{-1}C,$$

berechnet, wobei C die Cholesky Zerlegung von S^{-1} ist. D.h. es gilt $S^{-1} = C'C$. Dann ist $\Sigma = M'M$ *inverted* Wishart, denn es gilt

$$(M'M)^{-1} = C^{-1}B'B(C')^{-1} \sim W(n, S). \tag{4.13}$$

Diese Methode zur Simulation einer Wishart-Verteilung ist bekannt als **Bartlett Zerlegung**.

Schließlich erhält man den Erwartungswertvektor μ

$$\mu = \bar{y} + n^{-1/2}M'z \mid \Sigma \sim N(\bar{y}, n^{-1}\Sigma), \tag{4.14}$$

wobei $z \sim N(0, I)$ ein $p \times 1$ Vektor mit unabhängigen standard-normalverteilten Elementen ist.

Der P-Schritt in der t-ten Iteration kann somit wie folgt zusammengefasst werden:

(a) Die Matrix $\Sigma^{(t)}$ wird aus einer inversen-Wishart-Verteilung $W^{-1}(n-1, S^{(t)})$ gemäß Gleichung (4.13) gezogen, wobei $S^{(t)}$ die Varianz-Kovarianzmatrix der ergänzten Daten der t-ten Iteration darstellt[13].

(b) Bedingt auf $\Sigma^{(t)}$ wird $\mu^{(t)}$ gemäß Gleichung (4.14) aus einer Normalverteilung $N(\bar{y}^{(t)}, n^{-1}\Sigma^{(t)})$ generiert. Mit $\bar{y}^{(t)}$ wird der Vektor der arithmetischen Mittel der vervollständigten Daten bezeichnet.

4.3 Panel Struktur

Eine interessante Möglichkeit des klassischen Imputationsmodells, welche kaum in der Fachliteratur Erwähnung findet, besteht in der Einbeziehung vollständig beobachteter, binärer Variablen (sog. *Dummy*-Variablen) zur Modellierung komplexerer Datenstrukturen. Mittels *Dummy*-Variablen ist es möglich, vorhandener struktureller Heterogenität der Daten, z.B. unterschiedliche Zeitpunkte, Rechnung zu tragen. Dieser Sachverhalt war für das KEI-Projekt aufgrund folgender Tatsachen von besonderer Relevanz:

- Die vorhandenen Stichproben sind per Definition von einem kleinen Umfang, da die statistischen Einheiten zur Messung von Indikatoren die europäischen Länder sind.
- Es liegen Werte für die KEI-Indikatoren im Zeitraum 2001 bis 2004 vor.

Die folgende Erweiterung des Basismodells ist durch eine Bemerkung von Schafer (1997, S.35) motiviert.

[13] Man beachte die abweichende Anzahl der Freiheitsgrade.

4.3.1 Motivation der Modellierung mittels *Dummy*-Variablen

Der folgenden Erweiterung des Basis-Modells liegt die Vorstellung zugrunde, dass sich die Verteilung der Grundgesamtheit von einem Zeitpunkt zum anderen lediglich in ihrem Erwartugswert ändert, während die Abhängigkeitsstrukur konstant bleibt. Die zu den verschiedenen Zeitpunkten gehörenden Daten stellen also unabhängige Ziehungen aus multivariaten Normalverteilungen dar, welche sich ausschließlich durch den Vektor der Erwartungswerte μ unterscheiden. Diese zeitliche Stabilität der Varianz-Kovarianzmatrix impliziert die Konstanz der durch *Sweep*-Operationen bestimmten Steigungsparameter und ermöglicht somit ihre Modellierung mittels *Dummy*-Variablen.

Die Einbeziehung von Daten aus verschiedenen Zeitpunkten ohne explizite Modellierung ihrer Heterogenität würde unterstellen, dass die Erwartungswerte entweder konstant bleiben oder um den gleichen Prozentsatz steigen bzw. abnehmen, und dass die Abhängigkeitsstruktur konstant bleibt. Ändert sich jedoch der Erwartungswert der Verteilung in der Grundgesamtheit, so können die vorgeschlagenen Methoden fehlerhafte Ergebnisse liefern. Dies wird in Abbildung 4.5 veranschaulicht.

Auch wenn beide Annahmen, je nach Datenlage, für kurze Zeitperioden plausibel sein können, eignen sie sich nicht zur Modellierung von Daten mit einer ausgeprägten zeitlichen Dimension.

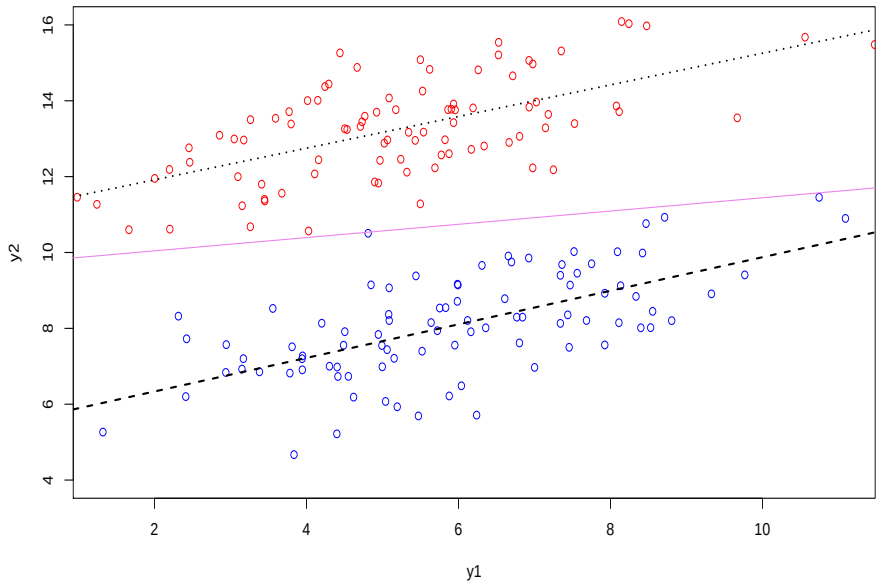

Abbildung 4.5: Die untere Punktwolke stellt eine i.i.d. Ziehung aus einer Normalverteilung mit Vektor der Erwartungswerte $\mu_1 = (6, 8)$ und Varianz-Kovarianzmatrix Σ dar. Bei der Generierung der oberen Wolke hat sich lediglich der Vektor der Erwartungswerte verändert. Die erste Komponente hat um 20% abgenommen, während die zweite Komponente um 60% gestiegen ist. D.h. der neue Vektor ist $\mu_2 = (4{,}8, 12{,}8)$. Da die Abhängigkeitsstruktur unverändert geblieben ist, sind beide Steigungsparameter (siehe gestrichelte und gepunktete Linien) bis auf Stichprobeneffekte identisch. Die Nicht-Berücksichtigung dieser Änderung verursacht jedoch Fehler in der Parameterschätzung bei Verwendung des EM-Algorithmus (durchgezogene Linie).

Obwohl die Annahme einer Veränderung in der Erwartungswertstruktur durchaus kompatibel mit dem gewählten Modellierungsansatz ist, erscheint es auf den ersten Blick merkwürdig, ein Imputationsmodell, welches das Vorhandensein einer multivariaten Normalverteilung unterstellt, mit Variablen zu ergänzen, deren Randverteilungen sehr von einer Normalverteilung abweichen. Diese Abweichungen verursachen jedoch keine Schätzprobleme, denn:

1. Diese Variablen werden vollständig beobachtet (müssen also nicht imputiert werden).
2. Unter Vorhandensein einer gemeinsamen Normalverteilung der restlichen Variablen sind alle bedingten Verteilungen gegeben diese *Dummy*-Variablen ebenfalls normal. Dies wird durch die Linearität der Regressionskurven gewährleistet.

Die zweite Aussage wird anhand des *Sweep*-Operators veranschaulicht. In Anlehnung an Abschnitt 4.1.2.2 wird die erweiterte Varianz-Kovarianzmatrix eines bivariat-normalverteilten Zufallsvektors Y erstellt. Diese Matrix wird im Folgenden mit θ bezeichnet

$$\theta = \begin{bmatrix} -1 & \mu_1 & \mu_2 \\ \mu_1 & \sigma_{11} & \sigma_{12} \\ \mu_2 & \sigma_{12} & \sigma_{22} \end{bmatrix}. \tag{4.15}$$

Mittels einer *Sweep*-Operation werden jetzt die Parameter der bedingten Verteilung von Y_2 gegeben $Y_1 = y_1$ bestimmt. Die auf die Position 1 *geswepte* Matrix wird in Anlehnung an Abschnitt 4.1.2.2 mit A und deren Elemente mit a_{ij} gekennzeichnet. Man erhält dann:

$$A = SWP[1]\theta = \begin{bmatrix} -(1 + \frac{\mu_1^2}{\sigma_{11}}) & \frac{\mu_1}{\sigma_{11}} & \mu_2 - (\frac{\sigma_{12}}{\sigma_{11}})\mu_1 \\ \frac{\mu_1}{\sigma_{11}} & \frac{-1}{\sigma_{11}} & \frac{\sigma_{12}}{\sigma_{11}} \\ \mu_2 - (\frac{\sigma_{12}}{\sigma_{11}})\mu_1 & \frac{\sigma_{12}}{\sigma_{11}} & \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}} \end{bmatrix}. \tag{4.16}$$

Die Maximum-Likelihood Schätzer der Matrizen (4.15) und (4.16) sind gegeben durch

$$\hat{\theta} = \begin{bmatrix} -1 & \bar{y}_1 & \bar{y}_2 \\ \bar{y}_1 & s_{11} & s_{12} \\ \bar{y}_2 & s_{12} & s_{22} \end{bmatrix} \tag{4.17}$$

bzw.

$$\hat{A} = SWP[1]\hat{\theta} = \begin{bmatrix} -(1 + \frac{\bar{y}_1^2}{s_{11}}) & \frac{\bar{y}_1}{s_{11}} & \bar{y}_2 - (\frac{s_{12}}{s_{11}})\bar{y}_1 \\ \frac{\bar{y}_1}{s_{11}} & \frac{-1}{s_{11}} & \frac{s_{12}}{s_{11}} \\ \bar{y}_2 - (\frac{s_{12}}{s_{11}})\bar{y}_1 & \frac{s_{12}}{s_{11}} & s_{22} - \frac{s_{12}^2}{s_{11}} \end{bmatrix}, \tag{4.18}$$

wobei $\bar{y}_1$ und $\bar{y}_2$ die arithmetischen Mittel und $s_{i,j}$ die empirischen Varianzen ($i = j$) bzw. Kovarianzen ($i \neq j$) darstellen.

Im Folgenden wird das Hauptaugenmerk auf die dritte Spalte der Matrix $\hat{A}$ gelegt. Wie bereits in Abschnitt 4.1.2 erörtert, liefern die Einträge $\hat{a}_{13}$ und $\hat{a}_{23}$ Schätzwerte der Parameter für den

bedingten Erwartungswert der Zufallsvariablen Y_2 gegeben $Y_1 = y_1$, während $\hat{a}_{33}$ ihre bedingte Varianz angibt. Es ist unmittelbar ersichtlich, dass die Ausdrücke $\bar{y}_2 - (s_{12}/s_{11})\bar{y}_1$ und s_{12}/s_{11} den Achsenabschnitt und den Steigungsparameter einer KQ-Regression von y_2 auf y_1 darstellen. Lediglich die Varianz der Residuen unterscheidet sich vom KQ-Schätzer für die Varianz der Störterme durch die Anzahl an Freiheitsgraden (bei der Größe in $\hat{a}_{33}$ handelt es sich um einen ML-Schätzer). Diese Regressionsparameter behalten bekannterweise ihre Struktur, wenn y_1 mit einer binären Variablen ersetzt wird (Regression mit *Dummy*-Variablen).

Solange die *Dummy*-Variablen vollständig beobachtet werden, ist die bedingte Verteilung von Y_2 gegeben $Y_1 = y_1$ normal mit Parametern $\mu_{2\cdot1} = (\mu_2 - (\sigma_{12}/\sigma_{11})\mu_1, \sigma_{12}/\sigma_{11}y_1)'$ und $\sigma_{22\cdot1} = \sigma_{22} - (\sigma_{12}^2/\sigma_{11})$ (vgl. Abschnitt 4.1.1) auch im Fall einer *Dummy*-Variablen y_1. Dies trifft jedoch nicht mehr zu, wenn die *Dummy*-Variablen NAs aufweisen. In diesem Fall muss die bedingte Verteilung von Y_1 gegeben $Y_2 = y_2$ bestimmt werden, welche nicht mittels linearer Regressionen modelliert werden kann.

Durch diese Modellerweiterung kann die Zeitdimension auf eine einfache und sehr parametersparsame Art berücksichtigt werden. Dies ist angesichts der kleinen Stichprobenumfänge der im Rahmen des KEI-Projektes verwendeten Variablen besonders wichtig.

Ein zusätzlicher Vorteil dieser metrisch-binären-Struktur besteht darin, dass sie unmittelbar auf die Weiterentwicklungen des klassischen Modells übertragen werden kann. Diese werden in Kapitel 5 betrachtet.

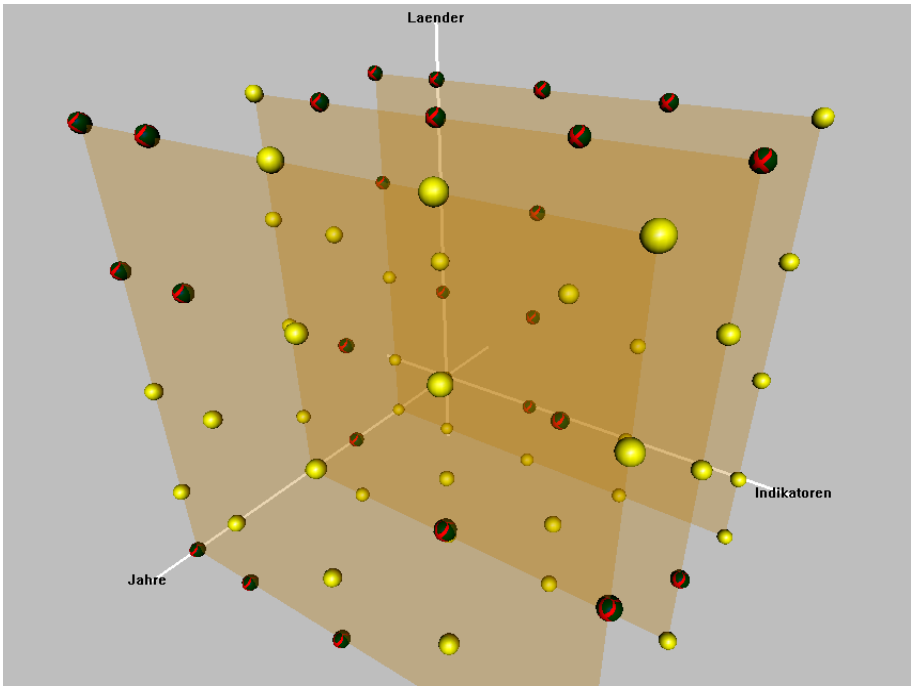

Abbildung 4.6: NA-Struktur des KEI-Datensatzes. Die gekreuzten Punkte kennzeichnen Werte, die nicht beobachtet wurden. Aufgrund der NA-Struktur müssen alle Zeitpunkte symmetrisch behandelt werden, d.h. spätere Zeitpunkte können durchaus verwendet werden, um fehlende Werte des vorhergehenden Jahres zu imputieren.

Hinweis: Die Modellierung von Zeiteffekten mittels *Dummy*-Variablen ist nicht unumstritten. Allerdings ist, vor allem unter Berücksichtigung der kleinen Stichprobenumfänge, die Frage gerechtfertigt, ob detailliertere Methoden eine bessere Modellierung ermöglichen können.

In Abbildung 4.7 wird das Imputationsmodell veranschaulicht[14]. Die Verwendung sowohl des EM- als auch des DA-Algorithmus zur Imputation weist gewisse Vorteile auf, insbesondere im Hinblick auf die Weiterentwicklungen des Basis-Imputationsmodells. Diese Weiterentwicklungen sind Gegenstand des nächsten Kapitels.

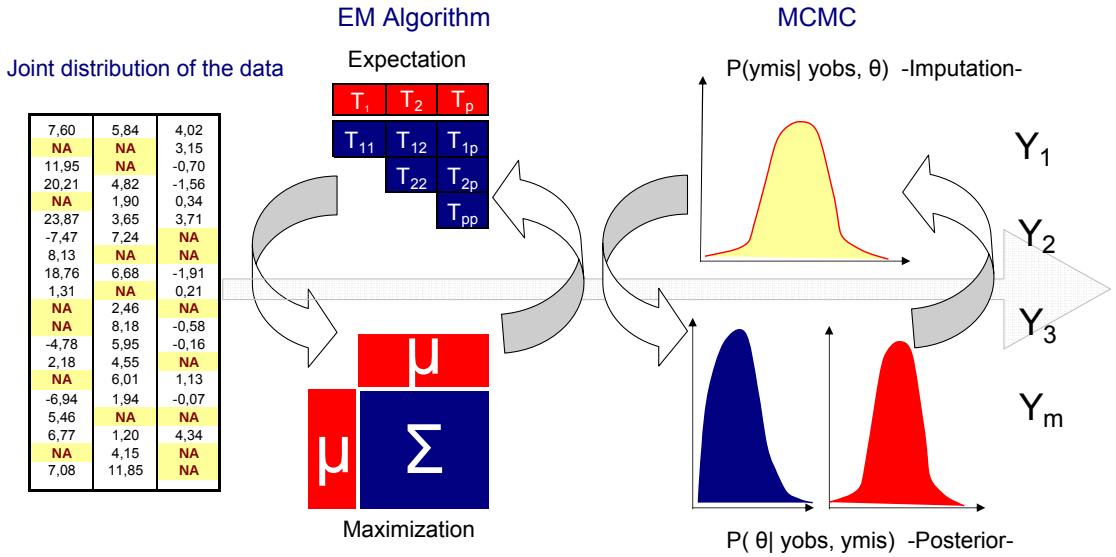

Abbildung 4.7: Allgemeines Imputationsschema. Während der EM-Algorithmus die Parameter eines multivariaten Datensatzes mit fehlenden Werten schätzt und somit optimale Startwerte für die Markov-Ketten bestimmt, liefert ein MCMC-Algorithmus m imputierte Daten pro fehlenden Wert.

4.3.2 Zur Überprüfung der Qualität der Imputationen

Die Überprüfung der Qualität der verschiedenen Imputationen erfolgt i.d.R. durch künstliches Löschen einer Beobachtung und Simulation ihrer prädiktiven Verteilung gegeben die restlichen beobachteten Daten. Dies wurde bereits in Abschnitt 3.4.6.5 anhand eines Beispiels veranschaulicht. In diesem Abschnitt wurde gezeigt, dass, unter Gültigkeit der unterstellten Verteilungsannahmen und bei Vorhandensein einer geeigneten Abhängigkeitsstruktur[15], die MCMC-Algorithmen durchaus in der Lage sind, die Daten zu rekonstruieren.

Die Qualitätsuntersuchung der imputierten Daten wurde im Rahmen des KEI-Projekts (siehe KEI (2004)) ausführlich dokumentiert und wird an dieser Stelle nicht weiter erörtert.

[14] Aufgrund der engen Beziehung zwischen dieser Arbeit und dem KEI-Projekt und der Notwendigkeit, in dessen Rahmen Zwischenergebnisse vorzustellen, sind einige Abbildungen auf Englisch beschriftet.

[15] Wenn die Daten schwach miteinander korrelieren, sind nicht genug Informationen vorhanden, um prädiktive Verteilungen zu simulieren, auch unter Gültigkeit der Verteilungsannahmen.

Kapitel 5

Weiterentwicklungen des Basis-Imputationsmodells

> „*David et al.(1986) found little evidence of bias in ignorable procedures that imputed missing values of income on the basis of other demographic and questionnaire items that were observed. This evidence came from knowledge of the missing values obtained from an external source, the actual wages and salary reported to the Internal Revenue Service on the individuals' tax returns. David et al.(1986) also concluded that* ***further improvements in the missing-data procedures would probably come from better modeling of the multivariate structure of the data, not from nonignorable modeling.***"
>
> (In Schafer (1997, S. 29))

Das vorliegende Kapitel hat die Untersuchung und Bereitstellung von Verfahren zum Ziel, welche den Schlussfolgerungen von David *et al.* Rechnung tragen. Somit wird die Anwendbarkeit verbreiteter Imputationsmethoden für normalverteilte Daten auf realistischere Datensituationen ausgeweitet.

Einleitung

Die Basis der in der vorliegenden Arbeit behandelten Imputationsverfahren stellen der EM- und der *Data Augmentation*-Algorithmus für normalverteilte Daten dar. Beide Verfahren betrachten den Datenbestand als Realisationen einer multivariat-normalverteilten Zufallsvariablen.

Daten, welche in realistischen Situationen beobachtet werden, weichen häufig von der Normalverteilung ab. Diese Abweichungen können in zwei Gruppen eingegliedert werden:

1. Die von der Annahme der Normalverteilung abweichenden Daten gehören zur Klasse der *elliptischen* Verteilungen[1]. In diesem Fall sind Weiterentwicklungen der Basis-Modelle anwendbar, welche die vorliegenden Daten als Realisierungen einer endlichen bzw. unendlichen

[1] Unter elliptischen Verteilungen wird eine Klasse symmetrischer, unimodaler Verteilungen verstanden, zu der die Normal-, t- und Cauchy Verteilungen gehören. Elliptische Verteilungen zeichnen sich dadurch aus, dass sie eine lineare Abhängigkeitsstruktur besitzen.

Mischung normalverteilter Zufallsvariabler betrachten. Diese Modelle versehen die Daten mit Gewichten und ermöglichen somit eine unverzerrte Schätzung der zu Grunde liegenden Parameter. Diese können wiederum zur Imputation herangezogen werden. Durch diese Modelle können Daten berücksichtigt werden, welche folgende Abweichungen von der Normalverteilung aufweisen:

- Präsenz von Ausreißern.
- Flanken mit mehr Wahrscheinlichkeitsmasse, als es bei Vorhandensein einer Normalverteilung zu erwarten wäre (*heavy tailed* Verteilungen).

Die auf Mischverteilungen basierenden Methoden werden in Kapitel 5.1 behandelt.

2. Die Daten gehören nicht zur Klasse der elliptischen Verteilungen. In diesem Fall sind keine Weiterentwicklungen sondern grundlegende Änderungen der Algorithmen erforderlich. Die vorgeschlagenen Verfahren, welche auf Datentransformationen basieren, ermöglichen die Berücksichtigung von asymmetrischen und strikt positiven Verteilungen und werden in Kapitel 5.2 erörtert.

5.1 Robuste Modelle mittels selektiver Gewichtung

Dieser Abschnitt befasst sich mit Weiterentwicklungen des in Kapitel 4 vorgestellten multivariaten normalen Modells, sowohl für den EM- als auch für den DA-Algorithmus. Ziel dieser Modelle ist die Identifikation und Neutralisierung von Ausreißern, deren Vorhandensein sowohl die Schätzung der Parameter des zu Grunde liegenden Modells als auch die Simulation ihrer *a posteriori*-Verteilungen beeinträchtigen. Modelle, welche imstande sind, den Effekt von Ausreißern zu neutralisieren, werden in der Literatur als *robust* bezeichnet (siehe Little und Rubin (2002)). Die Neutralisierung der Ausreißer erfolgt mittels selektiver Gewichtung und beruht auf der Vorstellung, dass die vorhandenen Daten aus einer Verteilung stammen, welche durch eine endliche bzw. unendliche Mischung verschiedener Normalverteilungen darstellbar ist. Die Anwendbarkeit dieser Modelle setzt folgende Bedingungen voraus:

- Die Daten generierende Verteilung ist elliptisch.
- Innerhalb eines Modells sind alle Verteilungen normal mit gleichem Erwartungswert μ.
- Die Varianzen der verschiedenen Verteilungen haben eine gemeinsame Komponente σ^2 und unterscheiden sich lediglich durch einen multiplikativen Term $1/q_i$, d.h. die Varianz der i-ten Verteilung ist folgendermaßen definiert:

$$\mathrm{Var}_i = \frac{\sigma^2}{q_i}.$$

Somit wäre im univariaten Fall die i-te Verteilung wie folgt charakterisiert

$$P_i(Y|\theta) \sim N\left(\mu, \frac{\sigma^2}{q_i}\right).$$

5.1.1 Allgemeines Mischungsmodell

5.1.1.1 Parametrisierung

Zusätzlich zu den Parametern μ und Σ des klassischen multivariaten normalen Modells werden positive, unbeobachtete Skalare q_i $(i = 1, 2, \ldots, n)$ berücksichtigt, welche i.i.d. Realisationen einer Zufallsvariablen Q mit Dichtefunktion $h_Q(q)$ darstellen. Unter dem neuen Modell kann die Verteilung von Y wie folgt dargestellt werden

$$P(y_i \mid \theta, q_i) \sim N_p(\mu, \Sigma/q_i), \tag{5.1}$$

wobei p die Dimension von Y darstellt. Wie bereits in Abschnitt 2.1.6 erläutert, gehört die Normalverteilung zur Exponentialfamilie. Aufgrund dessen ist die Q-Funktion in Gleichung (2.15) nicht linear in den Daten sondern in einem Satz an suffizienten Statistiken. Im erweiterten Modell ist die Loglikelihood also eine Funktion der suffizienten Statistiken[2] $T_0 = \sum_{i=1}^n q_i$, $T_1 = \sum_{i=1}^n q_i y_i$ und $T_2 = \sum_{i=1}^n q_i\, y_i y_i'$, welche, in Analogie zum normalen Modell, in eine $(p+1) \times (p+1)$ Matrix eingeordnet werden können:

$$T = \begin{bmatrix} T_0 & T_1' \\ T_1 & T_2 \end{bmatrix}.$$

Wenn q und Y vollständig beobachtet würden, wären die ML-Schätzer für $\theta = (\mu, \Sigma)$ äquivalent zu den gewichteten KQ-Schätzern

$$\hat{\mu} = \frac{T_1}{T_0} \tag{5.2}$$

$$\hat{\Sigma} = \frac{1}{n}\left(T_2 - \frac{T_1 T_1'}{T_0}\right). \tag{5.3}$$

Dies ist jedoch nicht der Fall. Insbesondere sind die Skalare q_i als unbeobachtbar definiert.

5.1.1.2 Implementierung

Die folgende Ausführung lehnt sich an Little (1988, S. 24ff) an.

E-Schritt: Aufgrund des Vorhandenseins fehlender Daten werden im E-Schritt die suffizienten Statistiken der vollständigen Daten mittels ihrer bedingten Erwartungswerte geschätzt:

$$E\left[T_0 \mid Y_{obs}, \theta\right] = E\left[\sum_{i=1}^n q_i \,\middle|\, Y_{obs}, \theta\right] = \sum_{i=1}^n w_i^{(t)}$$

mit den geschätzten Gewichten $w_i^{(t)} = E[q_i \mid Y_{obs}, \theta^{(t)}]$.

[2] Der Ausdruck T_2 ist offensichtlich nicht linear in y.

Die j-te Komponente von $E\left[T_1 \mid Y_{obs}, \theta\right]$ ist:

$$\begin{aligned} E\left[\sum_{i=1}^{n} q_i y_{ij} \,\middle|\, Y_{obs}, \theta\right] &= \sum_{i=1}^{n} E\big[q_i\, E(y_{ij} \mid Y_{obs}, q_i, \theta) \mid Y_{obs}, \theta\big] \\ &= \sum_{i=1}^{n} w_i^{(t)} E[y_{ij} \mid Y_{obs}, \theta], \\ \text{d.h. } E[q_i y_{ij} \mid Y_{obs}, \theta] &= \begin{cases} w_i^{(t)} y_{ij}, & \text{für } j \in \mathcal{O}(s) \\ w_i^{(t)} y_{ij}^*, & \text{für } j \in \mathcal{M}(s), \end{cases} \end{aligned}$$

wobei

$$y_{ij}^* = a_{0j} + \sum_k a_{kj} y_{ik} \quad \text{(vgl. Abschnitt 4.1.3.3 auf Seite 87).}$$

Das (j,k)-te Element von $E\left[T_2 \mid Y_{obs}, \theta\right]$ ist:

$$\begin{aligned} E\left[\sum_{i=1}^{n} q_i y_{ij} y_{ik} \mid Y_{obs}, \theta\right] &= \sum_{i=1}^{n} E[q_i E(y_{ij} y_{ik} \mid Y_{obs}, q_i, \theta) \mid Y_{obs}, \theta] \\ &= \sum_{i=1}^{n} E\Big\{ q_i \left[E(y_{ij} \mid Y_{obs}, \theta) E(y_{ik} \mid Y_{obs}, \theta)\right. \\ &\qquad \left. + \operatorname{Cov}(y_{ij} y_{ik} \mid Y_{obs}, q_i, \theta)\right] \Big| Y_{obs}, \theta \Big\} \\ &= \sum_{i=1}^{n} w_i\, E[y_{ij} \mid Y_{obs}, \theta]\, E[y_{ik} \mid Y_{obs}, \theta] + \Sigma_{obs,i}, \\ \text{d.h. } E[q_i y_{ij} y_{ik} \mid Y_{obs}, \theta] &= \begin{cases} w_i\, y_{ij} y_{ik}, & \text{für } j,k \in \mathcal{O}(s) \\ w_i\, y_{ij}^* y_{ik}, & \text{für } j \in \mathcal{M}(s), k \in \mathcal{O}(s) \\ w_i\, y_{ij}^* y_{ik}^* + a_{jk}, & \text{für } j,k \in \mathcal{M}(s) \end{cases} . \end{aligned}$$

M-Schritt: Im M-Schritt werden die neuen Schätzungen $\mu^{(t+1)}$ und $\Sigma^{(t+1)}$ anhand von Gleichungen (5.2) und (5.3) analog zum normalen Modell bestimmt, wobei T_0, T_1 und T_2 die im vorigen E-Schritt geschätzten Statistiken darstellen. Eine leichte Modifizierung, welche typisch für den PX-EM-Algorithmus (siehe Abschnitt 2.2) ist, erhöht die Konvergenzgeschwindigkeit dadurch, dass der Nenner n in Gleichung (5.3) mit der Summe der im E-Schritt bestimmten Gewichten, $\sum_{i=1}^{n} w_i^{(t)}$, ersetzt wird.

Bevor die unterschiedlichen Modelle behandelt werden, erweist es sich als zweckmäßig, ein Maß zur Identifizierung der Ausreißer einzuführen.

5.1.1.3 Mahalanobis-Distanz

Ein weit verbreitetes Maß zur Messung der (statistischen) Distanz zwischen zwei Punkten einer Verteilung, welches die Kovarianz zwischen Zufallsvariablen berücksichtigt, ist die nach dem

Indischen Mathematiker Prasanta Chandra Mahalanobis (1936) genannte Mahalanobis Distanz. Diese Distanz misst den Abstand zum Zentroid (Vektor der Erwartungswerte) aller Punkte einer (multivariaten)[3] Verteilung, unter Berücksichtigung ihrer Kovarianz-Struktur. Aufgrund dieser Tatsache wird die Mahalanobis-Distanz häufig zur Identifikation von Ausreißern eingesetzt (siehe Schafer und Ghosh-Dastidar (2006)).

Um Ausreißer zu identifizieren, wird den beobachteten Daten ein Gewichtsterm w_i zugeordnet, welcher sich als Funktion der Mahalanobis Distanz dieser Daten zum Zentroid ergibt.

Die Mahalanobis-Distanz für die i-te Zeile lautet

$$d_i^{(t)} = \sqrt{(y_{obs,i} - \mu_{obs,i}^{(t)})' \Sigma_{obs,i}^{(t)-1} (y_{obs,i} - \mu_{obs,i}^{(t)})},$$

wobei $\mu_{obs,i}$ den Vektor der Erwartungswerte *der beobachteten Werte der i-ten Zeile* und $\Sigma_{obs,i}$ die Varianz-Kovarianz Matrix bezeichnen. Große Werte von d_i^2 deuten auf Ausreißer hin. Zeilen mit großen Werten für d_i^2 werden also heruntergewichtet. Die genaue Struktur der Gewichte ist jedoch abhängig vom gewählten Modell. Zwei Modelle werden in dieser Arbeit diskutiert: das kontaminierte-normale-Modell und das t-Modell.

5.1.2 Kontaminiertes-normales-Modell

Dieses Mischungsmodell erweist sich als sehr nützlich zur Modellierung von Grundgesamtheiten, die hauptsächlich Realisationen aus einer multivariaten Normalverteilung hervorbringen, mit Ausnahme von wenigen Fällen, die durch eine große Abweichung von der unterstellten Verteilung gekennzeichnet sind (vgl. McLachlan und Peel (2000, 17)).

Um das kontaminierte-(multivariate)-normale-Modell herzuleiten, wird folgende Verteilung für q_i angenommen:

$$h_Q(q_i) = \begin{cases} 1-\delta & \text{für } q_i = 1 \\ \delta & \text{für } q_i = \lambda \\ 0 & \text{sonst,} \end{cases} \tag{5.4}$$

wobei $0 < \delta < 1$, $\lambda > 0$ mit bekannter Kontaminierungswahrscheinlichkeit δ und bekanntem Parameter λ. Dann ist die bedingte Verteilung von y_i eine Mischung aus zwei Normalverteilungen

$$N(\mu, \Sigma) \text{ und } N(\mu, \Sigma/\lambda).$$

Für das kontaminierte-normale-Modell wird der Wert für den Parameter λ üblicherweise so gewählt, dass gilt $\lambda \ll 1$. Erfahrungswerte liegen zwischen 0,05 und 0,3 (siehe Schafer und Ghosh-Dastidar (2006). Little und Rubin zeigen, dass das i-te Gewicht w_i durch die direkte Anwendung des Bayes-Theorems abgeleitet werden kann.

Für die i-te Zeile liefert Gleichung (5.4) folgendes Gewicht:

$$w_i^{(t)} = \frac{1 - \delta + \delta\lambda^{1+\frac{k_i}{2}} \exp\left\{(1-\lambda)\frac{d_i^{(t)2}}{2}\right\}}{1 - \delta + \delta\lambda^{\frac{k_i}{2}} \exp\left\{(1-\lambda)\frac{d_i^{(t)2}}{2}\right\}}.$$

[3] Im Falle einer univariaten Verteilung ist die Berechnung der Mahalanobis-Distanz äquivalent zur Standardisierung $z = (x - E[X])/\sqrt{\text{Var}(X)}$.

Aufgrund seiner Konstruktion weist die empirische Verteilung der durch das kontaminierte-normale-Modell hervorgebrachten Gewichte eine starke Konzentration auf zwei Werte auf:

- Große Gewichte und hohe Häufigkeiten für diejenigen Werte, welche niedrige Distanzwerte d_i^2 haben.
- Niedrige Gewichte und niedrige Häufigkeiten für die als Ausreißer identifizierten Werte.

Dies wird am Ende des Beispiels 5.1.3.3 veranschaulicht.

Beispiel: Parameterschätzung im kontaminierten-normalen-Modell

Gebeben sei eine Stichprobe vom Umfang $n = 680$ aus einer multivariaten Normalverteilung mit folgenden Parametern:

$$\mu = (2, 3, 4, 2, 1, 3, 4, 6)$$

und

$$\Sigma = \begin{pmatrix} 4{,}0 & 1{,}3 & 1{,}2 & 0{,}3 & 0{,}1 & 0{,}5 & 1{,}0 & 0{,}9 \\ 1{,}3 & 1{,}0 & 0{,}7 & 0{,}2 & 0{,}1 & 0{,}2 & 0{,}5 & 0{,}4 \\ 1{,}2 & 0{,}7 & 2{,}2 & 0{,}6 & 0{,}3 & 0{,}4 & 0{,}5 & 0{,}4 \\ 0{,}3 & 0{,}2 & 0{,}6 & 0{,}6 & 0{,}2 & 0{,}3 & 0{,}4 & 0{,}2 \\ 0{,}1 & 0{,}1 & 0{,}3 & 0{,}2 & 0{,}3 & 0{,}2 & 0{,}3 & 0{,}1 \\ 0{,}5 & 0{,}2 & 0{,}4 & 0{,}3 & 0{,}2 & 1{,}0 & 0{,}8 & 0{,}4 \\ 1{,}0 & 0{,}5 & 0{,}5 & 0{,}4 & 0{,}3 & 0{,}8 & 4{,}0 & 0{,}9 \\ 0{,}9 & 0{,}4 & 0{,}4 & 0{,}2 & 0{,}1 & 0{,}4 & 0{,}9 & 1{,}4 \end{pmatrix}.$$

Die Beobachtungen werden mit 10 % Ziehungen aus einer anderen Normalverteilung kontaminiert, welche den gleichen Erwartungswert und die Varianz-Kovarianzmatrix $\Sigma^* = \frac{1}{\lambda}\Sigma$ für $\lambda = 0{,}05$ besitzt.

Um die Aufmerksamkeit auf die Neutralisierung der Ausreißer zu richten, wird o.B.d.A. im Folgenden von einem vollständig beobachteten Datensatz ausgegangen. Die Parameter werden mit den üblichen ML-Schätzer im Falle einer Normalverteilung (siehe Anhang 5.2.2.2) und mit dem in Abschnitt 5.1.2 eingeführten EM-Algorithmus für eine kontaminierte Normalverteilung geschätzt. Schließlich wird die Genauigkeit der Schätzung mit den folgenden Formeln bestimmt

$$D^{mod}_{\mu_{(i)}} = \frac{\hat{\mu}^{mod}_{(i)} - \mu_{(i)}}{\mu_{(i)}} \cdot 100 \quad \text{und} \quad D^{mod}_{\sigma_{(ij)}} = \frac{\hat{\sigma}^{mod}_{(ij)} - \sigma_{(ij)}}{\sigma_{(ij)}} \cdot 100, \tag{5.5}$$

wobei $i, j \in \{1, \ldots, 8\}$ und *mod* das betrachtete Modell bezeichnet.

Ergebnisse:

Die prozentualen Abweichungen zwischen den theoretischen Parametern μ und Σ und ihren ML-Schätzern $\hat{\mu}_{norm}$ und $\hat{\Sigma}_{norm}$ betragen

$$D_{\mu_{(i)}}^{norm} = (-2{,}7,\ -2{,}0,\ -3{,}5,\ -4{,}6,\ -4{,}9,\ -2{,}4,\ -3{,}9,\ 0{,}2)$$

und

$$D_{\Sigma}^{norm} = \begin{pmatrix} 156{,}8 & 165{,}0 & 167{,}4 & 206{,}3 & 254{,}5 & 246{,}1 & 260{,}2 & 193{,}3 \\ 165{,}0 & 195{,}1 & 185{,}3 & 225{,}8 & 204{,}4 & 302{,}4 & 306{,}8 & 251{,}6 \\ 167{,}4 & 185{,}3 & 195{,}3 & 191{,}2 & 165{,}3 & 296{,}4 & 495{,}9 & 309{,}2 \\ 206{,}3 & 225{,}8 & 191{,}2 & 177{,}0 & 165{,}7 & 251{,}7 & 454{,}3 & 227{,}8 \\ 254{,}5 & 204{,}4 & 165{,}3 & 165{,}7 & 138{,}5 & 196{,}9 & 254{,}1 & 331{,}5 \\ 246{,}1 & 302{,}4 & 296{,}4 & 251{,}7 & 196{,}9 & 244{,}4 & 329{,}8 & 248{,}9 \\ 260{,}2 & 306{,}8 & 495{,}9 & 454{,}3 & 254{,}1 & 329{,}8 & 268{,}9 & 299{,}2 \\ 193{,}3 & 251{,}6 & 309{,}2 & 227{,}8 & 331{,}5 & 248{,}9 & 299{,}2 & 181{,}1 \end{pmatrix}.$$

Es ist offensichtlich, dass die ML-Schätzung der Varianz-Kovarianzmatrix von den Ausreißern beeinflusst wird, wodurch Verzerrungen von bis zu ca. 500 % entstehen. Da sowohl die kontaminierende als auch die zu schätzende Verteilung den gleichen Erwartungswertvektor besitzen, ist der verzerrende Einfluss im Falle des Erwartungswertes von einem geringeren Umfang. Das kontaminierte-normale-Modell liefert die Parameterschätzungen $\hat{\mu}_{kont-norm}$ und $\hat{\Sigma}_{kont-norm}$ mit den folgenden prozentualen Abweichungen

$$D_{\mu}^{kont-norm} = (-1{,}5,\ -0{,}7,\ -1{,}3,\ -1{,}0,\ -2{,}6,\ 0{,}0,\ 0{,}1,\ 0{,}2)$$

und

$$D_{\Sigma}^{kont-norm} = \begin{pmatrix} 4{,}2 & 15{,}7 & 4{,}5 & -3{,}4 & 11{,}4 & 16{,}3 & 19{,}1 & 15{,}2 \\ 15{,}7 & 18{,}8 & 14{,}1 & 21{,}8 & 7{,}3 & 30{,}8 & 26{,}1 & 28{,}3 \\ 4{,}5 & 14{,}1 & 3{,}0 & 4{,}3 & -3{,}0 & 11{,}1 & 4{,}8 & 22{,}6 \\ -3{,}4 & 21{,}8 & 4{,}3 & 19{,}3 & 11{,}5 & 20{,}4 & 22{,}8 & 17{,}6 \\ 11{,}4 & 7{,}3 & -3{,}0 & 11{,}5 & 2{,}5 & 5{,}3 & 3{,}9 & 29{,}4 \\ 16{,}3 & 30{,}8 & 11{,}1 & 20{,}4 & 5{,}3 & 14{,}0 & 12{,}2 & 22{,}0 \\ 19{,}1 & 26{,}1 & 4{,}8 & 22{,}8 & 3{,}9 & 12{,}2 & 18{,}7 & 22{,}7 \\ 15{,}2 & 28{,}3 & 22{,}6 & 17{,}6 & 29{,}4 & 22{,}0 & 22{,}7 & 11{,}9 \end{pmatrix}.$$

Trotz der noch vorhandenen Abweichungen ist die höhere Qualität der Schätzung mittels des kontaminierten-normalen-Modells leicht ersichtlich. Durch die selektive Gewichtung der Beobachtung ist dieses Modell in der Lage, den verzerrenden Effekt der Ausreißer zu verringern.

Es gibt jedoch Situationen, in denen sich die Abweichungen von der Normalverteilung nicht durch wenige auffällige Werte, sondern durch die gesamte Verteilungsgestalt bemerkbar machen. Einen typischen Fall stellen Verteilungen mit mehr Wahrscheinlichkeitsmasse in den Flanken als die Normalverteilung dar. Aufgrund dieser unterschiedlichen Gestaltung der Flanken kommen sehr große und sehr kleine Werte mit einer höheren Wahrscheinlichkeit vor, als die Normalverteilung erwarten ließe. Um solche Abweichungen von der Normalverteiltheit zu korrigieren, ist eine gleichmäßigere Gewichtung der Beobachtungen erwünscht. Das multivariate t-Modell ist dazu konzipiert, solche Situationen zu behandeln.

5.1.3 Multivariates t-Modell

5.1.3.1 t-Modell (mit bekannten Freiheitsgraden ν)

Eine andere Möglichkeit für die Herleitung einer funktionalen Form der Gewichte ist die Annahme, dass die q_i so verteilt sind, dass gilt: $q_i\,\nu$ ist Chi-quadrat-verteilt mit Freiheitsgraden gleich ν, d.h.

$$q_i|\nu \sim \Gamma(\nu/2, \nu/2)$$

(vgl. Liu (1995, 140)). Durch die Anwendung von Gleichung (5.1) ergibt sich folgende Randverteilung für y_i

$$y_i \sim t_k(\mu, \Sigma, \nu),$$

wobei t_k eine k-variate (skalierte) Student's t-Verteilung mit folgender Dichtefunktion bezeichnet

$$P(Y \mid \mu, \Sigma, \nu) = \frac{\Gamma\left(\frac{\nu+k}{2}\right) \mid \Sigma \mid^{-\frac{1}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\left\{\Gamma\left(\frac{1}{2}\right)\right\}^k \nu^{\frac{k}{2}}} \times \left(1 + \frac{(Y-\mu)'\Sigma^{-1}(Y-\mu)}{\nu}\right)^{-\left(\frac{\nu+k}{2}\right)}.$$

Little und Rubin (2002) zeigen, dass die Verteilung der Gewichte anhand des Bayes-Theorems hergeleitet werden kann.

Für die i-te Zeile gilt

$$w_i^{(t)} = E[q_i \mid Y_{obs}, \theta^{(t)}] = \frac{(\nu + k_i)}{\left(\nu + d_i^{(t)2}\right)}. \tag{5.6}$$

Um die Gewichte zu bestimmen, unterstellen sowohl das kontaminierte-normale-Modell als auch das bisher vorgestellte t-Modell das Vorhandensein fester Parameterwerte λ und ν. Eine höhere Flexibilität in der Modellierung wird dadurch erreicht, dass die Parameter, welche die Verteilungform der Gewichte bestimmen, mitgeschätzt werden.

5.1.3.2 Adaptives-t-Modell (unbekannte Freiheitsgrade ν)

Diese Weiterentwicklung des t-Modells lockert die Annahme fester (bekannter) Freiheitsgrade ν auf und erreicht somit eine höhere Anpassungsfähigkeit an die Datenlage. Die Freiheitsgrade ν in Gleichung (5.6) werden durch einen Schätzwert $\nu^{(t)}$ ersetzt. Der M-Schritt, welcher die neuen Schätzungen $\mu^{(t+1)}, \Sigma^{(t+1)}$ für μ und Σ bestimmt, wird erweitert, um die Bestimmung eines Schätzwertes $\nu^{(t+1)}$ zu ermöglichen.

Die Parameter dieses erweiterten t-Modells werden mittels des ECME-Algorithmus geschätzt (siehe Kapitel 2.2). Diese Variante des EM-Algorithmus ersetzt den Maximierungsschritt durch zwei bedingte Schritte:

CM1: Die Schätzwerte $\mu^{(t+1)}$ und $\Sigma^{(t+1)}$ werden durch Maximierung der Q-Funktion bezüglich $\theta = (\mu, \Sigma)$ bestimmt.

CM2: Die tatsächliche Loglikelihood wird bezüglich ν für fest gehaltene Parameter $(\mu^{(t+1)}, \Sigma^{(t+1)})$ maximiert

$$\begin{aligned} L(\nu, \mu, \Sigma \mid Y_{obs}, \mu^{(t+1)}, \Sigma^{(t+1)}) &= -\frac{n}{2}\log|\Sigma| + n\log\left(\Gamma(\frac{\nu+k}{2})\right) \\ &\quad -\frac{nk}{2}\log(\nu) - n\log\left(\Gamma(\frac{\nu}{2})\right) \\ &\quad -\frac{\nu+k}{2}\sum_{i=1}^{n}\left(\log(1+\frac{(y_i-\mu)\Sigma^{-1}(y_i-\mu)'}{\nu})\right). \end{aligned}$$

5.1.3.3 Ziehungen aus der *a posteriori*-Verteilung

Während die Implementierung des DA-Algorithmus für die Modelle in den Abschnitten 5.1.2 und 5.1.3.1 keinerlei Schwierigkeiten bereitet, erfordert das *adaptive*-multivariate-t-Modell die Ziehung aus der *a posteriori*-Verteilung von ν. Diese Verteilung hat jedoch keine typische funktionale Form. Unter der Verwendung einer nichtinformativen, konstanten *a priori*-Verteilung wird die funktionale Form der *a posteriori*-Verteilung vollständig durch die Likelihood bestimmt. Es ist im t-ten Schritt also erforderlich, aus der skalierten bedingten Likelihood gegeben die Schätzwerte $\mu^{(t)}$ und $\Sigma^{(t)}$ einen Schätzwert für $\nu^{(t)}$ zu ziehen.

Diese Ziehung aus der Likelihood ist mit numerischen Problemen verbunden, da ihre Werte einen numerischen *underflow* verursachen[4].

Der P-Schritt wird in zwei Unterschritte P1 und P2 geteilt:

P1: Analog zum P-Schritt im normalen Modell, neue Schätzwerte $\Sigma^{(t+1)}$ und $\mu^{(t+1)}$ werden gemäß den Verteilungen in Gleichungen (4.13) und (4.14) generiert.

P2: Bedingt auf die generierten Werte $\Sigma^{(t+1)}$ und $\mu^{(t+1)}$ und unter der Annahme einer konstanten *a priori*-Verteilung für ν wird ein Wert aus der skalierten bedingten Likelihood gezogen.

Es gibt verschiedene Methoden, welche die Ziehung aus dieser Verteilung ermöglichen. Eine Möglichkeit besteht in der Verwendung des in Kapitel 3 behandelten *Grid Samplers*, welcher die Likelihood in einem Punktgitter evaluiert und anhand dieser Werte eine empirische Verteilungsfunktion konstruiert, aus der Werte gezogen werden können.

Die Programmierung eines *Grid Sampler*s für das t-Modell erfordert die folgenden Schritte:

1. Werte der bedingten Likelihood werden in einem Gitter $v_1, v_2, \ldots, v_k$ bestimmt. Um eine gute Approximation der Likelihood zu erzielen, müssen die Stützstellen so gewählt werden, dass viele von ihnen in Regionen platziert werden, in denen die skalierte Likelihood eine hohe Wahrscheinlichkeitsmasse aufweist. Die Regionen der Likelihood mit niedriger Wahrscheinlichkeitsmasse können anhand weniger Punkte approximiert werden. Um eine

[4] Die Likelihoodfunktion ergibt sich aus dem Produkt der einzelnen Dichtewerten. Diese gemeinsamen Produkte liegen i.d.R. unterhalb der Darstellungsmöglichkeiten selbst modernster Computer.

möglichst gute Verteilung der Stützstellen gemäß den oben genannten Kriterien zu erzielen, wurde in dieser Arbeit eine Methode entwickelt, welche die zweiten Ableitungen der Likelihood-Funktion verwendet. Dies liegt darin begründet, dass die Regionen der Likelihood, in denen sich das Krümmungsverhalten schnell ändert, mit größerer Sorgfalt (und somit mit mehr Stützstellen) angenähert werden müssen als diejenigen, in denen die Kurve einen konstanten Verlauf aufweist.

2. Eine Approximation der Verteilungsfunktion $P(\nu^{(t+1)} \mid Y, \Sigma^{(t+1)}, \mu^{(t+1)})$ wird anhand der Funktionswerte an den Stützstellen konstruiert.

3. Die Quantilfunktion zu dieser empirischen Verteilungsfunktion wird erzeugt.

4. Ein Wert p aus einer Gleichverteilung im Intervall $[0, 1]$ wird gezogen.

5. Der Wert $\nu^{(t+1)}$ wird als derjenige Wert gewählt, der das p-Quantil dieser empirischen Verteilung darstellt.

Beispiel (Fortsetzung): Parameterschätzung im adaptiven-t-Modell

Um die Funktionsweise des adaptiven-t-Modells zu veranschaulichen wird eine Stichprobe von Umfang $n = 750$ aus einer skalierten t-Verteilung mit den Parametern μ, Σ (siehe Beispiel 5.1.2) und $\nu = 5$ gezogen. Zur Parameterschätzung werden die üblichen ML-Schätzer im Falle einer Normalverteilung (siehe Anhang 5.2.2.2) und der in Abschnitt 5.1.3.2 vorgestellte ECME-Algorithmus für die t-Verteilung verwendet. Schließlich wird die Genauigkeit der Schätzungen mit den Formeln in (5.5) bestimmt.

Ergebnisse:

Die prozentualen Abweichungen zwischen den theoretischen Parametern μ und Σ und ihre ML-Schätzer im Falle einer Normalverteilung, $\hat{\mu}_{norm}$ und $\hat{\Sigma}_{norm}$, betragen

$$D_{\mu}^{norm} = (1{,}9, -0{,}9, 3{,}4, 2{,}2, 3{,}3, 1{,}5, -2{,}7, 1{,}7)$$

und

$$D_{\Sigma}^{norm} = \begin{pmatrix} 53{,}0 & 55{,}5 & 50{,}7 & 75{,}8 & 69{,}7 & 80{,}9 & 68{,}8 & 52{,}1 \\ 55{,}5 & 58{,}3 & 45{,}9 & 56{,}9 & 75{,}5 & 77{,}3 & 52{,}2 & 66{,}7 \\ 50{,}7 & 45{,}9 & 54{,}0 & 61{,}8 & 47{,}8 & 59{,}7 & 81{,}5 & 79{,}9 \\ 75{,}8 & 56{,}9 & 61{,}8 & 57{,}2 & 60{,}4 & 69{,}5 & 119{,}6 & 75{,}1 \\ 69{,}7 & 75{,}5 & 47{,}8 & 60{,}4 & 57{,}9 & 60{,}3 & 74{,}7 & 103{,}5 \\ 80{,}9 & 77{,}3 & 59{,}7 & 69{,}5 & 60{,}3 & 77{,}2 & 66{,}8 & 75{,}4 \\ 68{,}8 & 52{,}2 & 81{,}5 & 119{,}6 & 74{,}7 & 66{,}8 & 61{,}9 & 54{,}7 \\ 52{,}1 & 66{,}7 & 79{,}9 & 75{,}1 & 103{,}5 & 75{,}4 & 54{,}7 & 47{,}6 \end{pmatrix}.$$

Analog zum Beispiel (5.1.2) werden die Schätzungen durch Werte verzerrt, welche unter der Annahme einer Normalverteilung Ausreißer darstellen. Im Gegensatz zur kontaminierten Normalverteilung handelt es sich beim t-Modell nicht um wenige extreme Beobachtungen sondern um eine unterschiedliche Gestaltung der Flanken der Verteilung, welche viel mehr Wahrscheinlichkeitsmasse als diejenigen einer Normalverteilung aufweisen. Der ECME-Algorithmus liefert die folgenden Ergebnisse:

$$D_\mu^t = (-2{,}0,\ -1{,}0,\ 3{,}0,\ 2{,}2,\ 2{,}6,\ 1{,}1,\ -1{,}5,\ 0{,}4)$$

und

$$D_\Sigma^t = \begin{pmatrix} -3{,}7 & -3{,}7 & -8{,}7 & -6{,}7 & -9{,}2 & 10{,}2 & 16{,}4 & 2{,}4 \\ -3{,}7 & 0{,}0 & -7{,}6 & -2{,}3 & -0{,}6 & 9{,}0 & 1{,}9 & 10{,}1 \\ -8{,}7 & -7{,}6 & 2{,}4 & 3{,}5 & -6{,}4 & 2{,}8 & -6{,}0 & 8{,}0 \\ -6{,}7 & -2{,}3 & 3{,}5 & 2{,}0 & -0{,}5 & 5{,}2 & 13{,}3 & -1{,}3 \\ -9{,}2 & -0{,}6 & -6{,}4 & -0{,}5 & -0{,}3 & -9{,}7 & -7{,}0 & 6{,}3 \\ 10{,}2 & 9{,}0 & 2{,}8 & 5{,}2 & -9{,}7 & 4{,}0 & 8{,}2 & 6{,}7 \\ 16{,}4 & 1{,}9 & -6{,}0 & 13{,}3 & -7{,}0 & 8{,}2 & 4{,}1 & 6{,}3 \\ 2{,}4 & 10{,}1 & 8{,}0 & -1{,}3 & 6{,}3 & 6{,}7 & 6{,}3 & -0{,}2 \end{pmatrix}.$$

Auch in diesem Fall ist die Überlegenheit des robusten Modells offensichtlich. Der Schätzwert für den Freiheitgradsparameter ν beträgt 5,47 und stellt somit eine gute Approximation des wahren Parameters dar.

Schließlich seien in Abbildung 5.1 die Histogramme der Gewichte von Beispiel 5.1.2, (rechte Grafik) und Beispiel 5.1.3.3, (linke Grafik), vergleichend dargestellt. Während das t-Modell eine glockenförmige Verteilung der Gewichte hervorbringt, zeichnen sich die Gewichte des kontaminierten-normalen-Modells durch eine Konzentration auf wenige Werte aus.

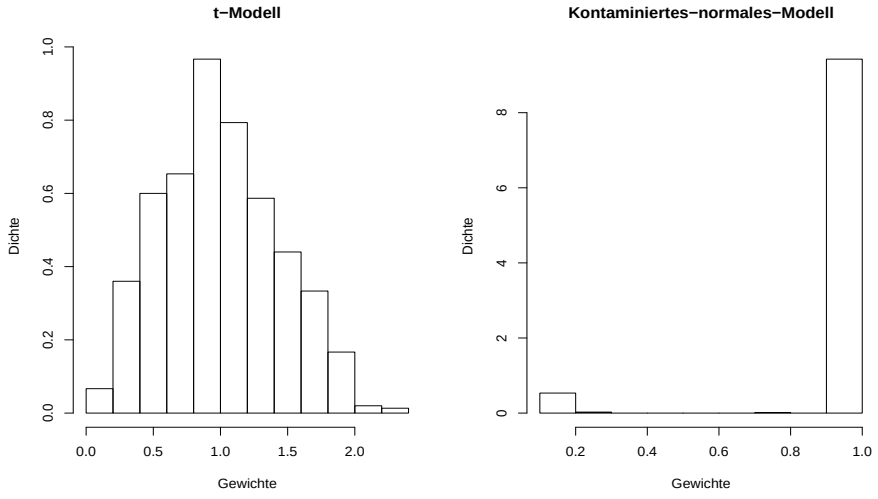

Abbildung 5.1: Vergleich der Verteilungen der Gewichte eines adaptiven-t- und eines kontaminierten-normalen-Modells. Während sich die Gewichte des t-Modells glockenförmig verteilen, konzentrieren sich diejenigen des kontaminierten-Modells auf wenige Ausprägungen.

5.2 Robuste Modelle mittels Datentransformation

5.2.1 Begründung der Notwendigkeit einer Transformation der Daten

Insbesondere bei dem in der Praxis häufig beobachteten Fall rechtsschiefer, strikt positiver Daten, können Algorithmen für normalverteilte Daten zu Fehlimputationen führen, welche nicht durch selektive Gewichtung korrigiert werden können. Abbildung 5.2 veranschaulicht diese Situation. Die diagonal-schraffierte Dichtefunktion im Vordergrund stellt die wahre Verteilung der Daten dar. Der EM-Algorithmus für normalverteilte Daten verwendet die aus diesen Daten gewonnenen empirischen Parameter, arithmetisches Mittel und Varianz, um die Verteilung der Daten zu charakterisieren. Denn im Falle einer Normalverteilung bestimmen diese Parameter die Verteilung eindeutig. Die aus der Normalverteilungsannahme und den empirischen Parametern resultierende Verteilung wird im Hintergrund in Abbildung 5.2 veranschaulicht. Obwohl ihre ersten beiden empirischen Momente identisch sind, unterscheiden sich beide Verteilungen deutlich voneinander. Insbesondere stellen die Werte der Abszisse unter der rot markierten Fläche Werte dar, welche unter der ursprünglichen Verteilung eine Wahrscheinlichkeitsdichte von Null besitzen und somit unzulässig als Imputationswerte sind.

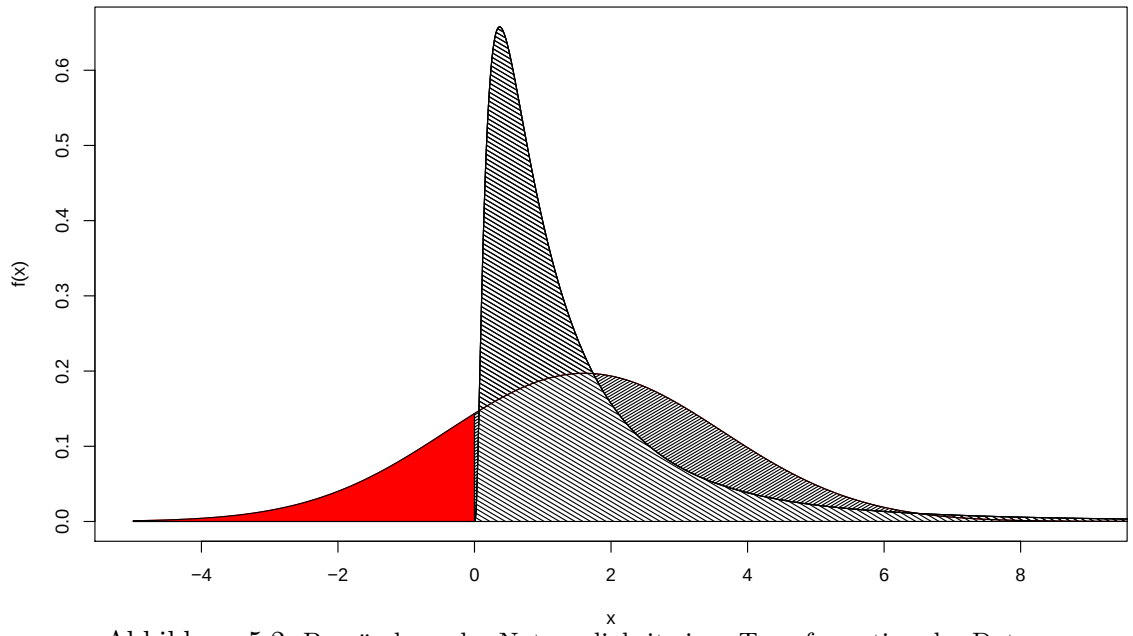

Abbildung 5.2: Begründung der Notwendigkeit einer Transformation der Daten.

Die Konsequenzen der Nichtbeachtung dieser Tatsache werden anhand des folgenden Beispiels verdeutlicht.

Beispiel:

Gegeben seien unvollständig beobachtete Realisationen einer positiven zweidimensionalen Zufallsvariablen. Ferner seien ihre Randverteilungen rechtsschief, jedoch mit unterschiedlichen Schiefewerten. Dieser Sachverhalt macht sich in Abbildung 5.3 durch einen nichtlinearen Verlauf der Punktwolke bemerkbar. Zum Zwecke der Übersichtlichkeit der grafischen Darstellung wurden folgende Vereinfachungen vorgenommen:

- Variable 1 weist fehlende Werte auf, während Variable 2 vollständig beobachtet wird.
- Die Imputation wird nicht mit dem DA-, sondern mit dem EM-Algorithmus durchgeführt. Dadurch wird erreicht, dass die imputierten Daten auf einer Geraden liegen. Dies erleichtert die grafische Darstellung der Problematik.

Die klassischen EM- und DA- Algorithmen imputieren mittels linearer Regressionen. Dies hat zur Folge, dass unzulässige Werte imputiert werden können, wenn die Annahme der multivariaten Normalverteilung verletzt wird. Abbildung 5.3 veranschaulicht diese Aussage.

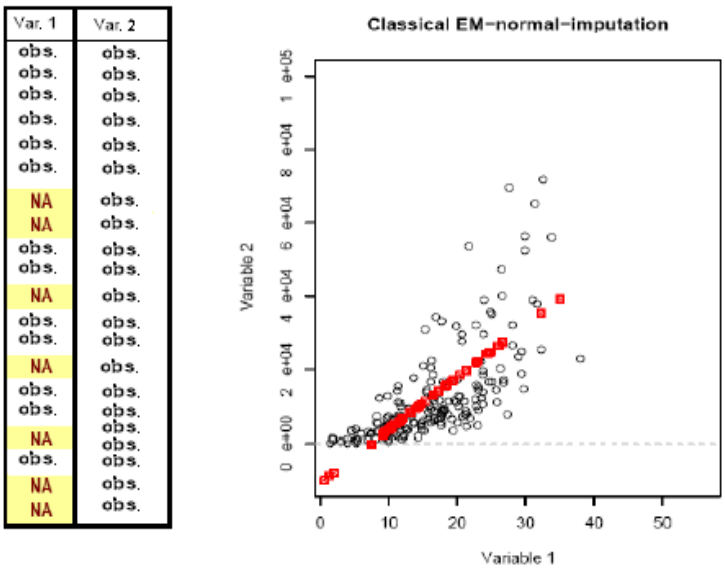

Abbildung 5.3: Unzulässig imputierte Daten aufgrund der Verletzung der Normalverteilungsannahme.

Da die explizite Modellierung der tatsächlichen Verteilung der Daten eine kaum zu bewältigende Komplexität mit sich bringen würde, erweist es sich als zweckmäßiger, die Daten so zu transformieren, dass die Normalverteilungsannnahme möglichst plausibel gemacht werden kann.

Schafer bestätigt die Ratsamkeit dieser Vorgehensweise:

> „*Datasets encountered in the real world often deviate from multivariate normality, but in many cases the normal model will be useful even when the actual data are nonnormal. [...] Sometimes the normality assumption may be made more plausible by applying suitable transformations to one or more variables.*“ (Schafer (1997, S. 29 und 147)).

Somit können Imputationsmodelle für normal- auf heterogen-verteilte Daten angewendet werden. Insbesondere bei multivariaten Datensätzen mit unterschiedlich verteilten Variablen stellt diese Möglichkeit einen notwendigen Kompromiss zwischen statistischer Genauigkeit und Plausibilität dar.

5.2.1.1 Alternativen für die Transformation

Eine *Conditio sine qua non* für eine Transformationsmethode ist ihre Umkehrbarkeit. Nur wenn diese Bedingung erfüllt ist, können transformierte Daten auf ihre ursprüngliche Form ohne Informationsverlust gebracht werden. Typische Möglichkeiten, die Verteilungsform von Daten zu transformieren, sind die folgenden:

(a) Addition einer Konstanten (Verschiebung).

(b) Multiplikation mit einer Konstanten (Skalierung).

Die Operationen (a) und (b) können unter dem Begriff der *affin linearen Transformation* zusammengefasst werden.

(c) Logarithmierung.

(d) Exponierung.

(e) Potenzierung.

(a,b) Affin lineare Transformationen: Eine affin lineare Transformation einer Zufallsvariablen X bzw. ihrer Realisationen $x_i,\ i \in \{1,\ldots,n\}$, wird als $Y := f(x) = a + b\,x$ für $b > 0$ definiert. Diese Transformation ist bijektiv und somit umkehrbar.

Aufgrund der Tatsache, dass die Normalverteilung invariant bezüglich affin linearer Transformationen ist, kann diese jedoch nicht zur Korrektur der Form der Daten herangezogen werden. Es ist also notwendig, die Suche nach einer geeigneten Transformation auf nichtlineare Funktionen einzuschränken.

(c,d) Logarithmierung und Exponierung: Die Logarithmierung ist eine weit verbreitete nichtlineare Transformation, welche häufig in der Lage ist, rechtsschiefe Verteilungen auf eine approximativ normalverteilt aussehende Form zu bringen. Zu ihren Vorteilen zählt ihre Einfachheit und insbesondere ihre Invertierbarkeit, d.h. $e^{\log(Y)} = Y$. Ein möglicher Nachteil dieser Transformation besteht darin, dass die tatsächliche Form der Daten nicht berücksichtigt wird. In der Tat werden durch diesen Ansatz Daten mit unterschiedlichen Gestalten der gleichen Transformation unterzogen. Gleiches gilt für ihre Umkehrtransformation, die Exponierung.

Die Möglichkeit, die tatsächliche Form der Daten mit einzubeziehen, erweist sich somit als wünschenswert. Eine Transformation, die diese Eigenschaft besitzt, ist die Potenzierung.

(e) Potenzierung: Die Potenzierung einer strikt positiven Zufallsvariablen X bzw. ihrer Realisationen $x_i,\ i \in \{1,\ldots,n\}$, wird als $Y := f(x) = x^r$, $r \in \mathbb{R}$ definiert. Diese ist streng monoton wachsend für $r > 0$ und streng monoton fallend für $r < 0$, und somit umkehrbar (vgl. Königsberger (2000, S. 31)). Ferner ist $g(x) = x^{\frac{1}{r}}$ die inverse Funktion von $f(x) = x^r$.

Eine Art der Potenztransformation, welche die Logarithmustransformation als Sonderfall beinhaltet, ist die Box-Cox Transformation.

5.2.1.2 Box-Cox Transformation

Diese Potenztransformation wurde bereits 1964 von Box und Cox vorgeschlagen und besitzt die Form

$$y = \begin{cases} (x^\lambda - 1)/\lambda & \text{für } \lambda \neq 0 \\ \ln(x) & \text{für } \lambda = 0. \end{cases} \tag{5.7}$$

Der Fall $x^\lambda = \ln(x)$ ergibt sich durch Grenzwertbildung und Anwendung der Regel von L'Hospital

$$\begin{aligned} \lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} &= \lim_{\lambda \to 0} \frac{\frac{\mathrm{d}(x^\lambda - 1)}{\mathrm{d}\lambda}}{1} \\ &= \lim_{\lambda \to 0} x^\lambda \ln(x) = \ln(x). \end{aligned}$$

Für $\lambda \neq 0$ kann der Ausdruck in (5.7) umgeformt werden

$$(x^\lambda - 1)/\lambda = \frac{1}{\lambda} x^\lambda - \frac{1}{\lambda}. \tag{5.8}$$

Es ist nun leicht ersichtlich, dass die rechte Seite von (5.8) die Form $a + bx^c$ besitzt. Aufgrund der Invarianzeigenschaft der Normalverteilung bezüglich affin linearer Transformationen können die Faktoren $a = \frac{1}{\lambda}$ und $b = \frac{1}{\lambda}$ keinen Einfluss auf die Form der Verteilung ausüben. Diese Terme dienen lediglich dem Zweck, die Stetigkeit von y in (5.7) für $\lambda = 0$ zu gewährleisten.

Aufgrund ihrer guten Eigenschaften basiert die in dieser Arbeit vorgeschlagene Transformation auf der Potenzierungsmethode und kann als Instanz der Box-Cox Transformation betrachtet werden. Das Hauptaugenmerk wird nun auf die Bestimmung des Potenzparameters gelegt.

Bestimmung eines optimalen Potenzparameters:

> „*In the analysis of data it is often assumed that observations $y_1, y_2, \ldots, y_n$ are independently normally distributed with constant variance and with expectations specified by a model linear in the set of parameters θ. In this paper we make the less restrictive assumption that such a normal, homoscedastic linear model is appropriate after some suitable transformation has been applied to the ys.*“ (Box und Cox (1964))

Bereits im Abstract ihres wegweisenden Beitrags erarbeiten Box und Cox drei Bedingungen, die sie als hinreichend für das Vorhandensein einer Normalverteilung betrachten. Diese Bedingungen sind:

1. Normalität der Randverteilungen.
2. Homoskedastizität.
3. Additivität, d.h. keine Interaktionen zwischen Variablen.

Die verbreitesten Möglichkeiten, den optimalen Potenzparameter zu bestimmen, konzentrieren sich auf zwei Methoden: Das „Ausprobieren“[5] (vgl. Chatfield (1997)) und die von Box und Cox verwendete ML-Schätzung, welche ebenfalls durch eine manuelle Suche des optimalen Parameters gekennzeichnet ist. Hamilton (1994, S. 126) bietet eine anschauliche Erklärung der ML-Methode zur Bestimmung des Potenzparameters:

> „*One approach is to pick a particular value of λ and maximize the likelihood function for $Y_t^{(\lambda)}$ under the assumption that $Y_t^{(\lambda)}$ is gaussian [...]. The value of λ that is associated with the highest value of the maximized likelihood is taken as the best transformation. However Nelson and Granger (1979) reported discouraging results from this method in practice.*“

Auch die Möglichkeit, über eine eindeutige Vorgehensweise zur Wahl des optimalen Parameters zu verfügen, ist eine erstrebenswerte Eigenschaft einer Transformationsmethode.

Wie bereits erwähnt, kann die in der vorliegenden Arbeit vorgeschlagene Methode als Instanz der Box-Cox Transformation aufgefasst werden. Sie basiert jedoch nicht auf der Likelihood-Funktion, sondern macht sich gewisse Regelmäßigkeiten der Normalverteilung zunutze. Diese Regelmäßigkeiten, welche als Momenten- bzw. Orthogonalitätsbedingungen aufgefasst werden können, ermöglichen die Verwendung der Struktur einer Klasse von Schätzmethoden, bekannt als GMM-Schätzer (vgl. Hayashi (2000, S. 446)), zur Bestimmung des optimalen Potenzparameters.

Somit liefert die vorgeschlagene Methode ein explizites Verfahren zur Lösung der Optimierungsaufgabe.

Diese noch näher zu erläuternde Transformation versucht die ersten zwei der von Box und Cox genannten Bedingungen zu erfüllen, indem sie die vorliegenden Daten auf eine Form bringt, welche mit einer multivariaten Normalverteilung vereinbar ist. Die Annahmen der Normalität der Randverteilung, Homoskedastizität und Linearität der Erwartungen sind bei Vorhandensein einer gemeinsamen Normalverteilung automatisch erfüllt.

In seinem Abschnitt über die Box-Cox Transformation hebt Greene (2003, S. 173,174) zwei Punkte hervor, in welchen sich diese und die vorgeschlagene Methode unterscheiden:

> „*In principle, each regressor could be transformed by a different value of λ, but, in most applications, this level of generality becomes excessively cumbersome, and λ is assumed to be the same for all the variables in the model.*
> *At the same time, it is also possible to transform y, say, by y^θ. Transformation of the dependent variable, however, amounts to a specification of the whole model, not just the functional form.*“

In dieser Arbeit wurde explizit die Möglichkeit eingeräumt, dass die Zufallsvariablen, welche die verschiedenen Spalten eines Datensatzes hervorgebracht haben, durchaus unterschiedliche Verteilungen haben können. Diese Überlegung motiviert die Notwendigkeit der Bestimmung eines optimalen Transformationsparameters *für jede Variable*, um eine approximative multivariate

[5] Eng.: *guesstimate.*

Normalverteilung zu erreichen. Um Greenes Vorbehalten bezüglich des Aufwands Rechnung zu tragen, muss dies jedoch auf eine einfache Art und Weise bewerkstelligt werden.

Darüber hinaus macht die vorgeschlagene Transformation keinen Unterschied zwischen abhängigen und unabhängigen Variablen. Diese symmetrische Behandlung der Variablen ist eine direkte Konsequenz der Existenz von fehlenden Werten. Bei einem allgemeinen Muster an fehlenden Daten sind alle Variablen abwechselnd abhängige Variable und Regressoren. Abbildung 5.4 veranschaulicht diese Aussage.

Während im Falle der Maximum-Likelihood Methode es nicht offensichtlich ist, warum der optimale Transformationsparameter λ die Likelihood-Funktion maximieren soll, besitzen die Momentenbedingungen eine hohe Transparenz, d.h. es ist von vornherein klar, welches Ziel erreicht werden soll und warum der optimale Parameter sich als Minimierer herauskristallisiert. Es sei jedoch an dieser Stelle die Tatsache erwähnt, dass im Falle der Normalverteilung ML- und Momentenschätzer identisch sind. Dieser Sachverhalt wird für die Auswahl der Bestimmungsgleichungen in Abschnitt 5.2.2.2 ausgenutzt.

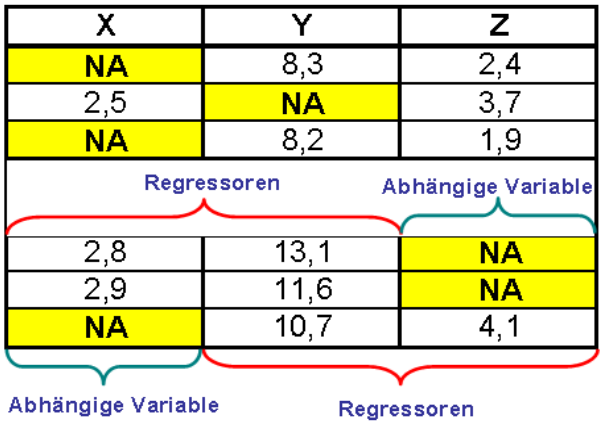

Abbildung 5.4: Begründung einer symmetrischen Behandlung der Variablen.

5.2.2 Univariate Transformation

5.2.2.1 Einleitung

Wie bereits in Abschnitt 5.2.1.2 erwähnt, macht sich die Transformationsmethode gewisse Regelmäßigkeiten der Normalverteilung zunutze, um die empirische Verteilung der Daten zu korrigieren. Diese Regelmäßigkeiten beziehen sich auf die Schiefe und Wölbung der Verteilung, welche im Falle der Normalverteilung feste Werte annehmen.

Die Hauptaufgabe der Transformationsmethode besteht in der Bestimmung eines Transformationsparameters derart, dass ein noch zu definierendes *Abstandskriterium*[6] zwischen der empirischen Schiefe und Wölbung und diesen festen Werten minimal wird.

Da diese Minimierung nicht analytisch gelöst werden kann, müssen numerische Verfahren eingesetzt werden. Der vorgeschlagene Algorithmus hat eine der GMM-Schätzmethode sehr ähnliche Struktur.

5.2.2.2 Bestimmungsgleichungen

Sei X eine normalverteilte Zufallsvariable mit Parametern μ und σ^2, es gelte also $X \sim N(\mu, \sigma^2)$. Dann besitzt X die folgenden Momente

$$
\begin{aligned}
&E[X] = \mu \in \mathbb{R} : \text{ Erwartungswert, erstes Moment.} \\
&E[(X-\mu)^2] = \sigma^2 \in \mathbb{R}_+ : \text{ Varianz, zweites zentriertes Moment.} \\
&E[(X-\mu)^3] = 0 : \text{ Schiefe, drittes zentriertes Moment.} \\
&\frac{E[(X-\mu)^4]}{\sigma^4} = 3 : \text{ Kurtosis, viertes skaliertes, zentriertes Moment.}
\end{aligned}
$$

Darüber hinaus sind alle weiteren ungeraden Momente gleich Null[7].

Die Grundidee des Transformationsalgorithmus besteht darin, dass die bekannte Struktur der Momente einer Normalverteilung ausgenutzt werden kann, um einen optimalen Transformationsparameter zu bestimmen.

Die ersten zwei Momente sind jeweils auf $\mathbb{R}$ und $\mathbb{R}+$ definiert. Aufgrund ihrer Variabilität tragen sie nicht zur Bestimmung des Potenzparameters bei und werden im Prinzip mitgeschätzt. Die Bestimmung eines optimales Transformationsparameters und der ersten zwei Momente ist somit ein dreidimensionales Problem. Dieses dreiparametrische Optimierungsproblem kann jedoch auf ein einparametrisches zurückgeführt werden, indem man μ und σ^2 durch ihre Maximum-Likelihood Schätzer

[6] Die Methode weist in diesem Sinne ebenfalls Ähnlichkeiten zu den sogenannten *Minimum Distance Estimators* (siehe Wolfowitz (1957) bzw. Drossos (1980)) auf.

[7] Man beachte, dass die oben dargestellte nicht die allgemeine Definition der Schiefe ist. Im Falle der Normalverteilung stimmt jedoch diese mit der Schiefe überein.

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{und}$$
$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{ML})^2$$

ersetzt. Diese Maximum-Likelihood Schätzer gehen in die Formeln der höheren Momente ein und verringern somit die Anzahl an Dimensionen des Optimierungsproblems. Dass diese *plug-in* Strategie keine Inkonsistenzen verursacht, wird im Folgenden erläutert.

Begründung der *plug-in*-Strategie: Obwohl es sich dabei um Standardergebnisse der Statistik bzw. Ökonometrie handelt, wird der Vollständigkeit halber im Folgenden gezeigt, dass das Einsetzen (*plug-in*) der ML-Schätzer für μ und σ^2 in die GMM-ähnliche Struktur des Algorithmus zur Bestimmung des optimalen Potenzparameters keine Inkonsistenzen verursacht. Zu diesem Zweck wird zuerst gezeigt, dass ML- und Momentenschätzer für die Parameter μ und σ^2 einer Normalverteilung als äquivalent angesehen werden können. Anschließend wird gezeigt, dass der (iterierte) verallgemeinerte Momentenschätzer (GMM) die Cramér-Rao Schranke erreicht und somit ebenfalls ein ML-Schätzer ist[8]. Da der MM-Schätzer als ML-Schätzer bereits die Cramér-Rao Schranke erreicht, können die von GMM berücksichtigten Momentenbedingungen keinerlei Effizienzvorteile mit sich bringen. Alle drei Schätzer sind in diesem Falle äquivalent und somit vertauschbar.

Aufgrund dieser Tatsache ist es möglich, die Bedingungen der ersten zwei Momente direkt mit den ML-Schätzern $\hat{\mu}_{ML}$ und $\hat{\sigma}^2_{ML}$ für μ und σ^2 zu ersetzen und diese in die Bedingungen der dritten und vierten zentrierten Momente einzusetzen. Dadurch wird ein erheblicher Zugewinn in der numerischen Stabilität des Algorithmus erzielt.

ML- und MM-Schätzer

(i) Maximum-Likelihood-Schätzer: Gegeben sei eine Stichprobe vom Umfang n mit Realisationen einer univariat normalverteilten Zufallsvariablen Y, es gelte also $Y \sim N(\mu, \sigma^2)$. Die Loglikelihood-Funktion dieser Stichprobe ist dann:

$$L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mu)^2}{\sigma^2}. \tag{5.9}$$

Um ihre ML-Schätzer zu bestimmen, wird zuerst die Loglikelihood-Funktion (5.9) nach den Parametern μ und σ^2 abgeleitet. Anschließend wird das resultierende Gleichungssystem gleich Null gesetzt.

Die partielle Ableitung der Loglikelihood-Funktion nach μ ergibt

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu). \tag{5.10}$$

[8] Die Betrachtung des ML-Schätzers, unter gewissen Bedingungen, als Sonderfall eines GMM-Schätzers ist ebenfalls geläufig (vgl. Hamilton (1994, S. 409)).

Nullsetzen von Gleichung (5.10) ergibt ferner

$$\sum_{i=1}^{n} y_i = n\mu. \tag{5.11}$$

Gleichung (5.11) führt schließlich zum Schätzer

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}. \tag{5.12}$$

Die partielle Ableitung der Loglikelihood-Funktion nach σ^2 ist

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2. \tag{5.13}$$

Das Nullsetzen von Gleichung (5.13) führt zu folgendem Ausdruck

$$-n\sigma^2 + \sum_{i=1}^{n}(y_i - \mu)^2 = 0. \tag{5.14}$$

Einsetzen von (5.12) in (5.14) ergibt schließlich den Schätzer für σ^2

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{5.15}$$

Anhand der Schätzer in (5.12) und (5.15) können die Bedingungen für das dritte und vierte zentrierte Moment ausschließlich als Funktion des Transformationsparameters θ aufgestellt werden.

(ii) Momentenschätzer: Es sei mit

$$\bar{m}_k := \frac{1}{n}\sum_{i=1}^{n} y_i^k$$

das k-te unzentrierte Moment einer einfachen Zufallsstichprobe aus $\mathrm{N}(\mu, \sigma^2)$ definiert. Dann gilt

$$\operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} y_i = \operatorname{plim} \bar{m}_1 = E(y_i) = \mu$$

und

$$\operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \operatorname{plim} \bar{m}_2 = \operatorname{Var}(y_i) + \mu^2 = \sigma^2 + \mu^2.$$

Daraus folgt

$$\hat{\mu}_{MM} = \bar{m}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

und

$$\hat{\sigma}^2_{MM} = \bar{m}_2 - \bar{m}_1^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Es gilt also[9],

$$\hat{\mu}_{ML} = \hat{\mu}_{MM} = \bar{y}$$

und

$$\hat{\sigma}^2_{ML} = \hat{\sigma}^2_{MM} = \frac{1}{n}\sum(y_i - \bar{y})^2,$$

was zu zeigen war.

□

Cramér-Rao Schranke für ML und (G)MM

(i) Maximum Likelihood: Um die Varianz-Kovarianzmatrix der Schätzer $\hat{\mu}_{ML}$ und $\hat{\sigma}^2_{ML}$ zu bestimmen, wird die Hesse-Matrix H der Loglikelihood-Funktion bestimmt

$$H(\mu, \sigma^2) = \begin{bmatrix} \frac{\partial^2 L}{\partial \mu \partial \mu} & \frac{\partial^2 L}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 L}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 L}{\partial \sigma^2 \partial \sigma^2} \end{bmatrix} = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum(y_i - \mu) \\ -\frac{1}{\sigma^4}\sum(y_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum(y_i - \mu)^2 \end{bmatrix}.$$

Der mit minus Eins multiplizierte Erwartungswert der Hesse Matrix I

$$I := -E[H(\mu, \sigma^2)] = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

stellt die sogenannte Informationsmatrix dar.

Unter der Annahme gewisser Regularitätsbedingungen für die zu Grunde liegende Dichtefunktion ist die Varianz eines Schätzers für einen Parameter θ immer mindestens so groß wie die Inverse der Informationsmatrix (vgl. Greene (2003, S. 889)). Diese Inverse wird als Cramér-Rao Schranke bezeichnet:

$$I\left[(\mu, \sigma^2)\right]^{-1} = \frac{1}{n}\begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}.$$

[9] Diese Ergebisse gelten nicht nur für die Normalverteilung, sondern allgemein für die Exponentialfamilie, zu welcher die Normalverteilung gehört (siehe Fahrmeir und Tutz (2001)).

(ii) Verallgemeinerte Momentenmethode (GMM): Wenn mehr Momentenbedingungen als Parameter θ vorliegen, können die MM-Schätzer die Momentenbedingungen nicht exakt erfüllen. Es müssen also andere Kriterien herangezogen werden, um diese Schätzer zu bestimmen. Der übliche GMM-Ansatz besteht in der Minimierung bezüglich θ einer gewichteten quadratischen Form dieser Momentenbedingungen:

$$\min_{\theta} Q(\theta) = \bar{m}' \, \mathrm{W} \, \bar{m}.$$

In seinem einflussreichen Beitrag hat Hansen (1982) gezeigt, dass die Inverse der Varianz-Kovarianzmatrix der Momentenbedingungen $\mathbf{S}^{-1}$ insofern eine optimale Gewichtungsmatrix W darstellt, als dadurch GMM-Schätzer bestimmt werden können, welche die kleinste Varianz-Kovarianzmatrix $\mathbf{V}$ aufweisen. Zu diesem Zweck muss zunächst die Varianz-Kovarianzmatrix der Momentenbedingungen bestimmt werden[10], wobei in diesem Fall $\theta = (\mu, \sigma^2)$ gilt:

$$\begin{aligned}
\mathbf{S} &= \mathrm{E}[\bar{m}(\hat{\theta})\,\bar{m}(\hat{\theta})'] \\
&= \mathrm{E}\left[\begin{pmatrix} (y_i-\mu) \\ (y_i-\mu)^2-\sigma^2 \\ (y_i-\mu)^3 \\ (y_i-\mu)^4-3\sigma^4 \end{pmatrix} \begin{pmatrix} (y_i-\mu) \\ (y_i-\mu)^2-\sigma^2 \\ (y_i-\mu)^3 \\ (y_i-\mu)^4-3\sigma^4 \end{pmatrix}' \right] \\
&= \mathrm{E}\begin{bmatrix} (y_i-\mu)^2 & \underbrace{(y_i-\mu)(y_i-\mu)^2-(y_i-\mu)\sigma^2}_{=0} & & \\ & (y_i-\mu)^4 & \underbrace{(y_i-\mu)^5-(y_i-\mu)3\sigma^4}_{=0} & \\ \vdots & \dots & & \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & 0 & 3\,\sigma^4 & 0 \\ 0 & 2\,\sigma^4 & 0 & 12\,\sigma^6 \\ 3\,\sigma^4 & 0 & 15\,\sigma^6 & 0 \\ 0 & 12\,\sigma^6 & 0 & 96\,\sigma^8 \end{bmatrix},
\end{aligned}$$

wobei

$$\begin{aligned}
&\mathrm{E}[y_i-\mu] = 0 \\
&\mathrm{E}[(y_i-\mu)^2] = \sigma^2 \\
&\mathrm{E}[(y_i-\mu)^3] = 0 \qquad \Big(\mathrm{E}[(y_i-\mu)^{3+2k}] = 0 \quad \text{für } k = 0, 1 \ldots\Big) \\
&\mathrm{E}[(y_i-\mu)^4] = 3\sigma^4 \\
&\mathrm{E}[(y_i-\mu)^6] = 15\sigma^6 \\
&\mathrm{E}[(y_i-\mu)^8] = 105\sigma^8.
\end{aligned}$$

Die partiellen ersten Ableitungen der Momentenbedingungen lauten

$$\mathfrak{D} = \mathrm{E}\begin{bmatrix} \frac{\partial \bar{m}_1}{\partial \mu} & \frac{\partial \bar{m}_1}{\partial \sigma^2} \\ \frac{\partial \bar{m}_2}{\partial \mu} & \frac{\partial \bar{m}_2}{\partial \sigma^2} \\ \frac{\partial \bar{m}_3}{\partial \mu} & \frac{\partial \bar{m}_3}{\partial \sigma^2} \\ \frac{\partial \bar{m}_4}{\partial \mu} & \frac{\partial \bar{m}_4}{\partial \sigma^2} \end{bmatrix} = \mathrm{E}\begin{bmatrix} -1 & 0 \\ -2(y_i-\mu) & -1 \\ -3(y_i-\mu)^2 & 0 \\ -4(y_i-\mu)^3 & -6\sigma^2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}.$$

[10] Der Verfasser bedankt sich bei Herrn Professor Dr. Jung für die Unterstützung bei der Zusammenstellung der Ergebnisse.

Mit $\mathbf{S}^{-1}$ als Gewichtungsmatrix ist die Varianz-Kovarianzmatrix des GMM-Schätzers

$$
\begin{aligned}
&\mathrm{Cov}(\hat{\theta}) = \mathbf{V} = \frac{1}{n}(\mathfrak{D}'\,\mathbf{S}^{-1}\,\mathfrak{D})^{-1} \\
&= \frac{1}{n}\left(\begin{bmatrix} -1 & 0 & -3\sigma^2 & 0 \\ 0 & -1 & 0 & -6\sigma^2 \end{bmatrix} \begin{bmatrix} \frac{5}{2\sigma^2} & 0 & \frac{-1}{2\sigma^4} & 0 \\ 0 & \frac{2}{\sigma^4} & 0 & \frac{-1}{4\sigma^6} \\ \frac{-1}{2\sigma^4} & 0 & \frac{-1}{6\sigma^6} & 0 \\ 0 & \frac{-1}{4\sigma^6} & 0 & \frac{1}{24\sigma^8} \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}\right)^{-1} \\
&= \frac{1}{n\frac{1}{\sigma^2}\frac{1}{2\sigma^4}} \begin{bmatrix} \frac{1}{2\sigma^4} & 0 \\ 0 & \frac{1}{\sigma^2} \end{bmatrix} \\
&= \frac{1}{n} \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}.
\end{aligned}
$$

Dieser Ausdruck stellt die Cramér-Rao-Schranke für $\hat{\mu}_{GMM}$ und $\hat{\sigma}^2_{GMM}$ dar, was zu zeigen war. □

Höhere Momente: Im Gegensatz zu den ersten zwei Momenten nehmen Schiefe und Kurtosis bei der Normalverteilung feste Werte an und zwar für alle möglichen Ausprägungen der Parameter μ und σ^2. Diese festen Werte stellen die Anhaltspunkte für die Transformation dar.

Ziel der Transformation ist also, einen Potenzparameter so zu bestimmen, dass die empirischen zweiten und dritten Momente der Daten möglichst nahe an jeweils null und drei liegen.

Dass diese höheren Momente in der Tat Auskunft über die Gestalt der Normalverteilung geben, kann anhand des bekannten Jarque-Bera Tests erkannt werden. Dieser Test auf Normalverteiltheit, welcher von Carlos M. Jarque und Anil K. Bera 1980 vorgeschlagen wurde, ist ein statistischer Test, der anhand der empirischen Schiefe und Kurtosis einer Stichprobe die Vereinbarkeit des Datenbefunds mit der Annahme einer Normalverteilung überprüft.

Die Teststatistik JB ist definiert als

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right),$$

wobei S die Schiefe und K die Kurtosis, wie in 5.2.2.2 aufgeführt, bezeichnen, und n die Anzahl der Beobachtungen angibt.

5.2.2.3 Der Transformationsalgorithmus

In seiner univariaten Version hat der Transformationsalgorithmus folgende Struktur:

1. **Empirische Momente:** Die ersten zwei Momente μ and σ^2 werden mit ihren ML- Schätzern für die transformierten Daten $y^\star = y^\theta_{obs}$ ersetzt[11]:

$$\begin{aligned} \bar{y}^\star &= \frac{1}{n}\sum y_i^\star \\ s^\star &= \sqrt{\frac{1}{n}\sum (y_i^\star - \bar{y}^\star)^2}. \end{aligned}$$

2. **Z-Transformation:** Aufgrund der Tatsache, dass eine verschwindende Varianz die perfekte Erfüllung aller Momentenbedingungen bewirken würde, ist es notwendig, die Daten zu standardisieren. Es wird also eine z-Transformation vorgenommen

$$z_i = \frac{y_i^\star - \bar{y^\star}}{s^\star},$$

   dann besitzt z ein arithmetisches Mittel von Null und eine Varianz von Eins.

3. **Momentenbedingungen:** Die Momentenbedingungen der dritten und vierten zentrierten Momente für die standardisierten Variablen werden aufgestellt:

$$\overline{m}(\hat{\theta}) = \begin{bmatrix} \overline{m}_1 \\ \overline{m}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

   wobei

$$\begin{aligned} \overline{m}_1 &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \mu_z)^3 = \frac{1}{n}\sum_{i=1}^{n} z_i^3 \\ \overline{m}_2 &= \frac{1}{n}\sum_{i=1}^{n}\left((z_i - \mu_z)^4 - 3\sigma^4\right) = \frac{1}{n}\sum_{i=1}^{n} z_i^4 - 3. \end{aligned}$$

   Diese Momentenbedingungen geben den Abstand zwischen empirischen und theoretischen Momenten in Abhängigkeit von θ an.

   Aufgrund der Tatsache, dass es mehr Bedingungen ($\overline{m}_1$ und $\overline{m}_2$) als zu schätzende Parameter (θ) gibt, kann kein Schätzer $\hat{\theta}$ die empirischen Momentenbedingungen eindeutig lösen. Die an GMM erinnernde Idee besteht darin, dass ein $\hat{\theta} \in \Theta$ gesucht wird, das diese empirischen Momentenbedingungen *möglichst gut* erfüllt.

[11] Um eine mögliche Verwechslung mit dem ebenfalls in diesem Kapitel betrachteten Kontaminierungsparameter λ zu vermeiden, wird der Transformationsparameter mit θ bezeichnet.

4. **Minimierung:** Die Suche nach einem optimalen θ wird anhand einer quadratischen Form der empirischen Momentenbedingungen $\overline{m}$ vorgenommen, welche es zu minimieren gilt. Analog zu GMM können die einzelnen Gleichungen mit einer Gewichtung versehen werden. Für die Zwecke der Vorstellung des Algorithmus und ohne Beschränkung der Allgemeinheit, wird an dieser Stelle die Einheitsmatrix I_2 benutzt, so dass beide Bedingungen gleiche Gewichte erhalten.

   Der Transformationsparameter $\hat{\theta}$ ist dann als derjenige Parameter $\theta \in \Theta$ definiert, welcher folgendes Minimierungsproblem löst

   $$\min_{\theta} Q(\theta) = \overline{m}' \, \mathrm{I}_2 \, \overline{m}.$$

   Da der Algorithmus auf der Berechnung der höheren Momente der Normalverteilung basiert, ist zu erwarten, dass die Methode eher für große Stichproben geeignet ist (vgl. Hayashi (2000, S. 215)). Analog zur GMM-Methode ist ebenfalls für die Schätzbarkeit des optimalen Potenzparameters die Stationarität der Daten im Sinne von Cochrane (2005, S. 198) eine notwendige Bedingung.

5.2.2.4 Simulative Untersuchung der Eigenschaften der Transformation

Um die Eigenschaften der Potenztransformation zu untersuchen, wurde eine Vielzahl von Simulationen durchgeführt. In diesem Abschnitt werden die Ergebnisse dieser Simulationen vorgestellt.

(a) Konvergenzverhalten des Transformationsparameters

Eine wichtige Eigenschaft einer Transformationsmethode ist die Fähigkeit, zu erkennen, ob eine Transformation notwendig ist oder nicht. In anderen Worten, die Methode sollte keine Daten transformieren, die bereits die richtige Form haben. Im Falle von Stichprobenziehungen aus einer Normalverteilung sollte der optimale Transformationsparameter im Durchschnitt gleich eins sein. Darüber hinaus ist zu erwarten, dass sich mit wachsendem Stichprobenumfang die unsystematischen, Stichproben-bedingten Abweichungen gegenseitig aufheben und dass die Streuung um den wahren Parameter abnimmt. Diese Eigenschaft des Algorithmus sollte auch für allgemeinere Werte gelten, d.h. für wachsendes n sollte der optimale Transformationsparameter $\hat{\theta}$ gegen den „wahren" Parameter stochastisch konvergieren. Da für die meisten Verteilungen nicht im Voraus klar ist, welcher Parameter optimal ist, und um das Verhalten der Potenztransformation näher zu untersuchen, wurden drei Reihen von Monte-Carlo Simulationen für wachsende Stichprobenumfänge durchgeführt, wobei für jeden Stichprobenumfang jeweils 500 Stichproben gezogen wurden und jeweils der Potenzparameter berechnet wurde. Auf diese Art und Weise ist es möglich, arithmetische Mittel und empirische Varianz der Potenzparameter für festen Stichprobenumfang zu berechnen.

Die untersuchten Verteilungen sind:

a1) Eine Normalverteilung $N(8, 1{,}44)$. Der Transformationsparameter sollte für ausreichend großes n im Durchschnitt eins sein, da, von stochastischen Effekten abgesehen, bereits normalverteilte Daten nicht transformiert werden müssen. Wie in Abbildung 5.5 veranschaulicht, konvergiert das arithmetische Mittel der Transformationsparameter asymptotisch gegen eins und die Varianz nimmt monoton ab.

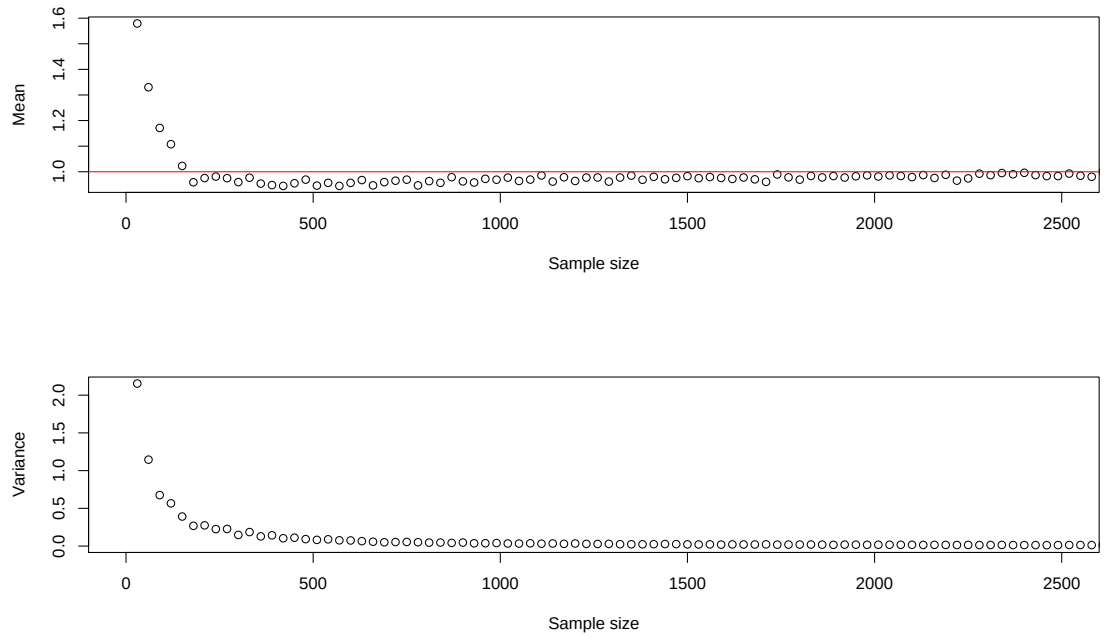

Abbildung 5.5: Entwicklung des arithmetischen Mittels und der Varianz des Transformationsparameters für wachsenden Stichprobenumfang. Die Stichproben wurden aus einer Normalverteilung $N(8, 1.44)$ gezogen.

a2) Die Verteilung einer transformierten Zufallsvariablen $Y := \psi(X) = X^3$.

Die Dichtefunktion dieser transformierten Zufallsvariablen Y kann mit Hilfe des Transformationssatzes bestimmt werden. Im Folgenden werden die notwendigen Schritte für die Anwendung dieses Satzes aufgeführt:

Es sei X eine normalverteilte Zufallsvariable, es gelte also $\mathrm{X} \sim \mathrm{N}(\mu, \sigma^2)$. Ferner sei folgende bijektive Abbildung definiert:

$$\psi : \mathbb{R} \to \mathbb{R};\ X \mapsto \psi(X) = X^3,$$

und die folgenden Größen gegeben:

$$\begin{aligned} \psi(x) &= x^3 =: y \\ \psi^{-1}(y) &= y^{\frac{1}{3}} \\ (\psi^{-1}(y))' &= \frac{1}{3} y^{\frac{-2}{3}}. \end{aligned}$$

Dann gilt

$$\begin{aligned} g_Y(y) &= f_X(y^{\frac{1}{3}}) \left| \frac{1}{3} y^{\frac{-2}{3}} \right| && \text{für } y \in \mathbb{R} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \left| \frac{1}{3} y^{\frac{-2}{3}} \right| \mathrm{e}^{\frac{-1}{2\sigma^2}(y^{1/3}-\mu)^2}. \end{aligned}$$

Obwohl es keinen i.i.d. Zufallszahlengenerator für diese Verteilung gibt, ist es mit den bereits behandelten MCMC-Methoden möglich, abhängige Realisationen aus dieser Verteilung zu

ziehen. Für die vorliegende Simulation wurde X als $N(8, 1{,}44)$ gewählt und ein Metropolis-Hastings-Algorithmus (siehe Abschnitt 3.3) als Zufallszahlengenerator eingesetzt.

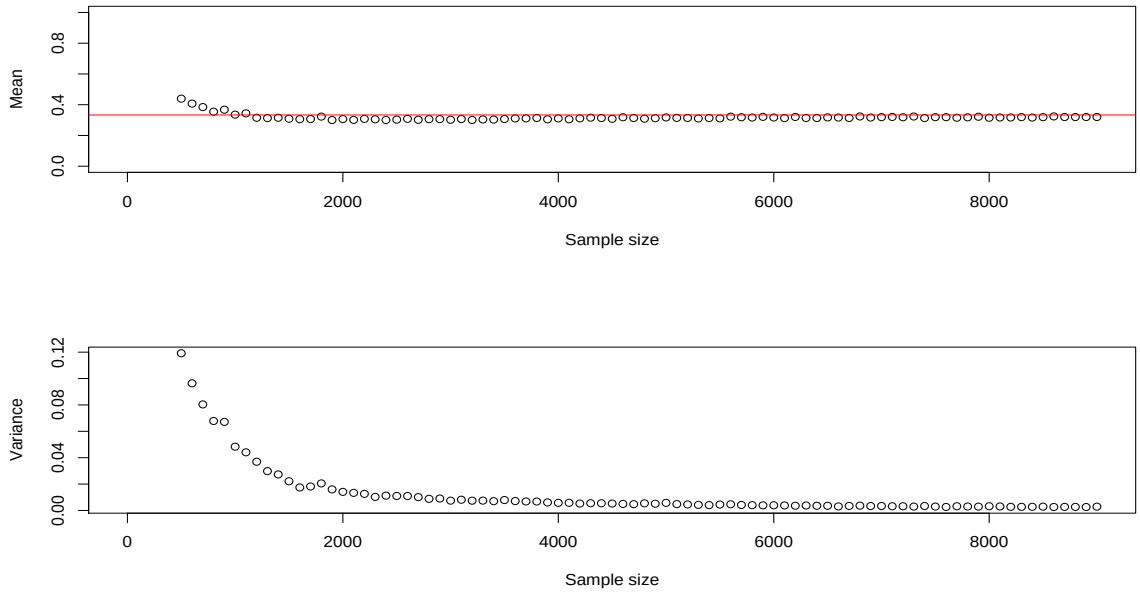

Abbildung 5.6: Entwicklung des arithmetischen Mittels und der Varianz des Transformationsparameters für wachsenden Stichprobenumfang. Stichproben aus einer transformierten Normalverteilung $N(8, 1.44)$. Die für die Simulation verwendeten transformierten Daten - $\psi(X) = X^3$, wobei $X \sim N(8, 1{,}44)$ - wurden mittels eines Metropolis-Hastings-Algorithmus gezogen.

Es ist offensichtlich, dass die approximative Normalverteiltheit bei einem Transformationsparameter von $\frac{1}{3}$ erreicht werden kann. Wie in Abbildung 5.6 zu sehen ist, konvergiert das arithmetische Mittel der Transformationsparameter asymptotisch gegen $\frac{1}{3}$ mit einer monoton abnehmenden Varianz. Es ist allerdings zu beachten, dass der Metropolis-Hastings-Algorithmus erheblich langsamer als ein i.i.d. Zufallszahlengenerator sein kann. Aufgrund dessen ist der Umfang der Simulation in diesem Fall geringer. Dies schlägt sich in der Gestalt der Punktfolge in beiden Abbildungen nieder.

a3) Eine Gamma-Verteilung, $X \sim \Gamma(4, 1)$. Im Gegensatz zu den anderen Fällen ist nicht im Voraus ersichtlich, welcher Wert für den Transformationsparameter optimal ist. *A priori* kann lediglich die Beobachtung gemacht werden, dass die Gamma-Verteilung eine positive Schiefe aufweist und somit der zu erwartende Parameter kleiner als eins sein sollte. Es ist jedoch *a posteriori* anhand statistischer Tests möglich zu überprüfen, ob die Transformationsparameter in der Lage sind, die Daten auf eine approximative Normalverteilung zu bringen.

Abbildung 5.7 zeigt die Entwicklung des arithmetischen Mittels, welches mit wachsendem Stichprobenumfang gegen einen festen Wert konvergiert, der ungefähr $\frac{1}{3}$ beträgt. Exemplarisch wurden die transformierten Daten bei einem Stichprobenumfang von 500 einem Shapiro-Wilk Test auf Normalverteiltheit unterzogen. Der p-Wert dieses Tests betrug 0,9031. Analog zu den anderen Fällen zeigt die Varianz einen monoton fallenden Verlauf.

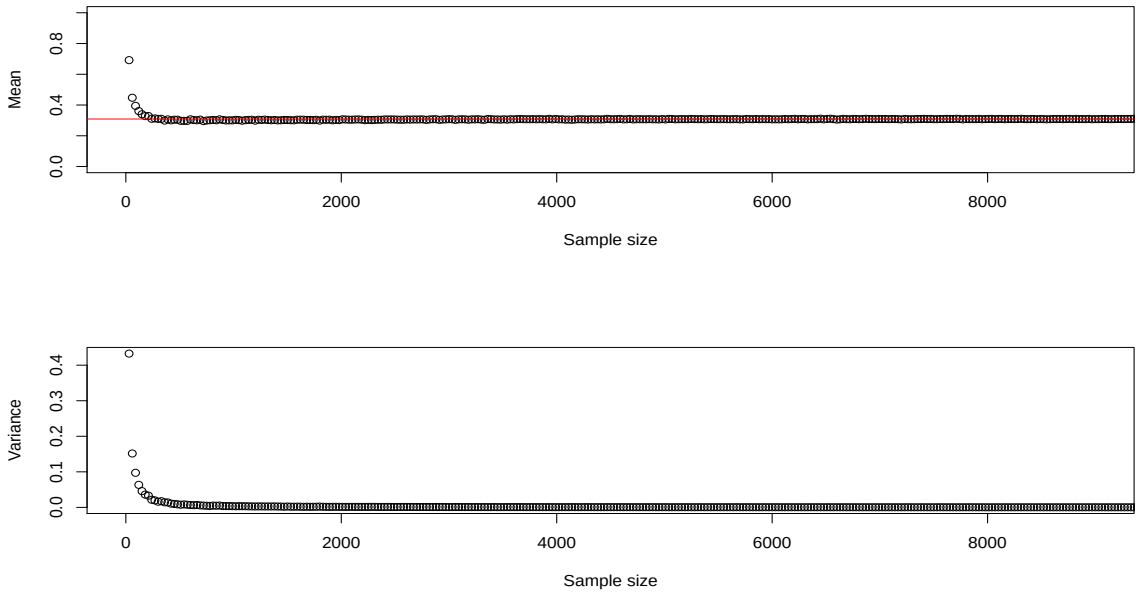

Abbildung 5.7: Entwicklung des arithmetischen Mittels und der Varianz des Transformationsparameters für wachsenden Stichprobenumfang. Stichproben aus einer $\Gamma(4,1)$ Verteilung.

(b) Simulative Untersuchung der Varianz des Potenzparameters als Funktion des Stichprobenumfangs

Einblicke in das Verhalten der Varianz liefert die Untersuchung der empirischen Varianz der Potenzparameter für die bereits vorgestellten Simulationen. Zu diesem Zweck wurde der Logarithmus der Varianz auf den Logarithmus des Stichprobenumfangs regressiert. Das unterstellte Modell ist:

$$\log\left(\mathrm{Var}(\hat{\theta})\right) = \beta_0 + \beta_1 \log(n) + \varepsilon.$$

Abbildung 5.8 zeigt die Anpassung des o.g. Modells auf die Varianz der in a1) vorgestellten Simulation. Die Signifikanz der geschätzten Parameter ist auf allen üblichen Niveaus gegeben.

Unter der Annahme der Repräsentativität der Monte-Carlo Ergebnisse gilt:

$$\begin{aligned}
&\mathrm{Var}(\hat{\theta}) \approx \frac{100}{n^{1,13}} \\
&\mathrm{Var}(\hat{\theta})\, n^{1,13} \approx 100 \\
&\mathrm{Var}(\hat{\theta})\, n^{1,13} \approx \mathrm{c} \\
&\mathrm{Var}(\hat{\theta}) \in \mathcal{O}(n^{-1,13}).
\end{aligned}$$

Die Varianz des Transformationsparameters in a1) ist approximativ von der Ordnung $n^{-1.13}$.

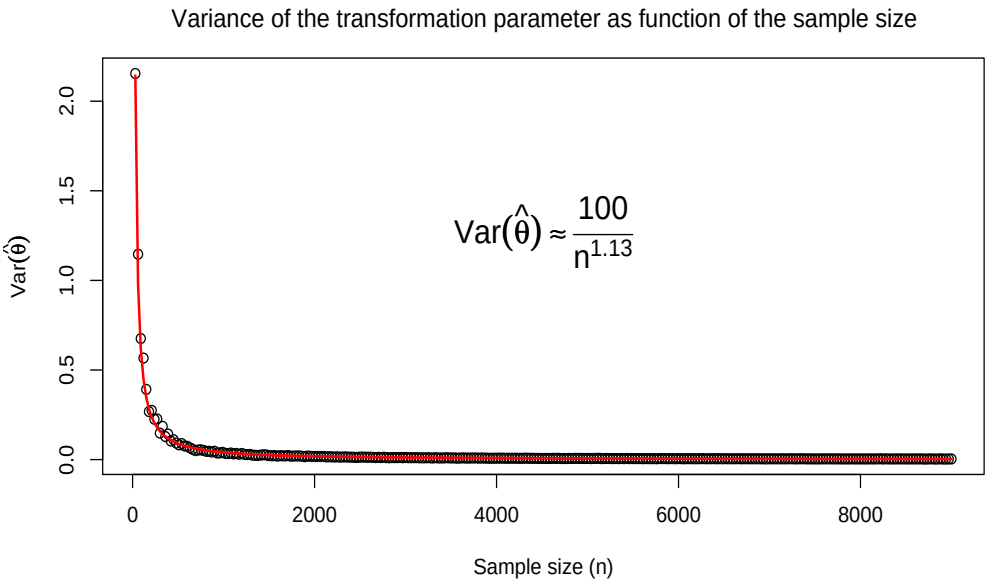

Abbildung 5.8: Simulative Untersuchung der Varianz des Transformationsparameters als Funktion des Stichprobenumfangs. Modell: $\log \mathrm{Var}(\hat{\theta}) = 4,57 - 1,13 \log(n)$.

(c) Simulative Untersuchung der Verteilung des Transformationsparameters

Die Ergebnisse der Simulationen in a1) und a3) wurden verwendet, um die Verteilung der Transformationsparameter für die unterschiedlichen Stichprobenumfänge zu untersuchen. Zu diesem Zweck wurde pro Stichprobenumfang ein Shapiro-Wilk Test auf Normalverteiltheit (zu einem Signifikanzniveau von 5%) durchgeführt. Die Nullhypothese dieses Tests unterstellt das Vorhandensein normalverteilter Daten. Jeder Stichprobe wurde ein Wert von eins bei Nicht-Ablehnung und von Null bei Ablehnung der Nullhypothese zugeordnet. Dieses Vorgehen ergibt somit eine Stichprobe von Ergebnissen der statistischen Tests. Darauffolgend wurde ein Fenster von 100 aufeinanderfolgenden Testergebnissen definiert und ein gleitender Durchschnitt für die ganze Stichprobe folgendermaßen konstruiert: Für die ersten 100 Ergebnisse wurde ein arithmetisches Mittel berechnet und anschließend das Fenster um ein Testergebnis verschoben. Um eine mögliche Verzerrung der Ergebnisse durch Verringerung des Stichprobenumfangs gegen Ende der Stichprobe zu vermeiden, wurde der gleitende Durchschnitt abgebrochen, als die übrig bleibenden Testergebnisse weniger als 50 wurden.

Abbildung 5.9 (obere Grafik) zeigt die Proportion an beibehaltenen Nullhypothesen für verschiedene Stichprobenumfänge, ausgedrückt durch die gleitenden Durchschnitte im Falle einer normalverteilten Grundgesamtheit. Es ist ersichtlich, dass diese Proportion approximativ monoton wächst und gegen das Nominalniveau des Tests konvergiert.

Die Ergebnisse für die Gamma-Verteilung sind in Abbildung 5.9 (untere Grafik) zu sehen. Trotz des monotonen Verlaufs ist in diesem Fall die Proportion an beibehaltenen Nullhypothesen für gleichen Stichprobenumfang geringer. Auch wenn die Verteilung des Transformationsparameters aufgrund der Konstruktion des Algorithmus nicht von der Varianz der zugrunde liegenden Verteilung abhängen kann, deutet Abbildung 5.9 (untere Grafik) auf eine gewisse Abhängigkeit vom dritten Moment hin. Dies ist daran erkennbar, dass die Konvergenz zum Nominalniveau

des Shapiro-Wilk Tests im Falle einer schiefen Verteilung deutlich größere Stichprobenumfänge benötigt.

Es sei an dieser Stelle erneut darauf hingewiesen, dass die Shapiro-Wilk Tests nicht die Fähigkeit der Potenztransformation überprüfen, eine Stichprobe auf eine approximative Normalverteilung zu bringen, sondern die Verteilung der Transformationsparameter für verschiedene Stichproben.

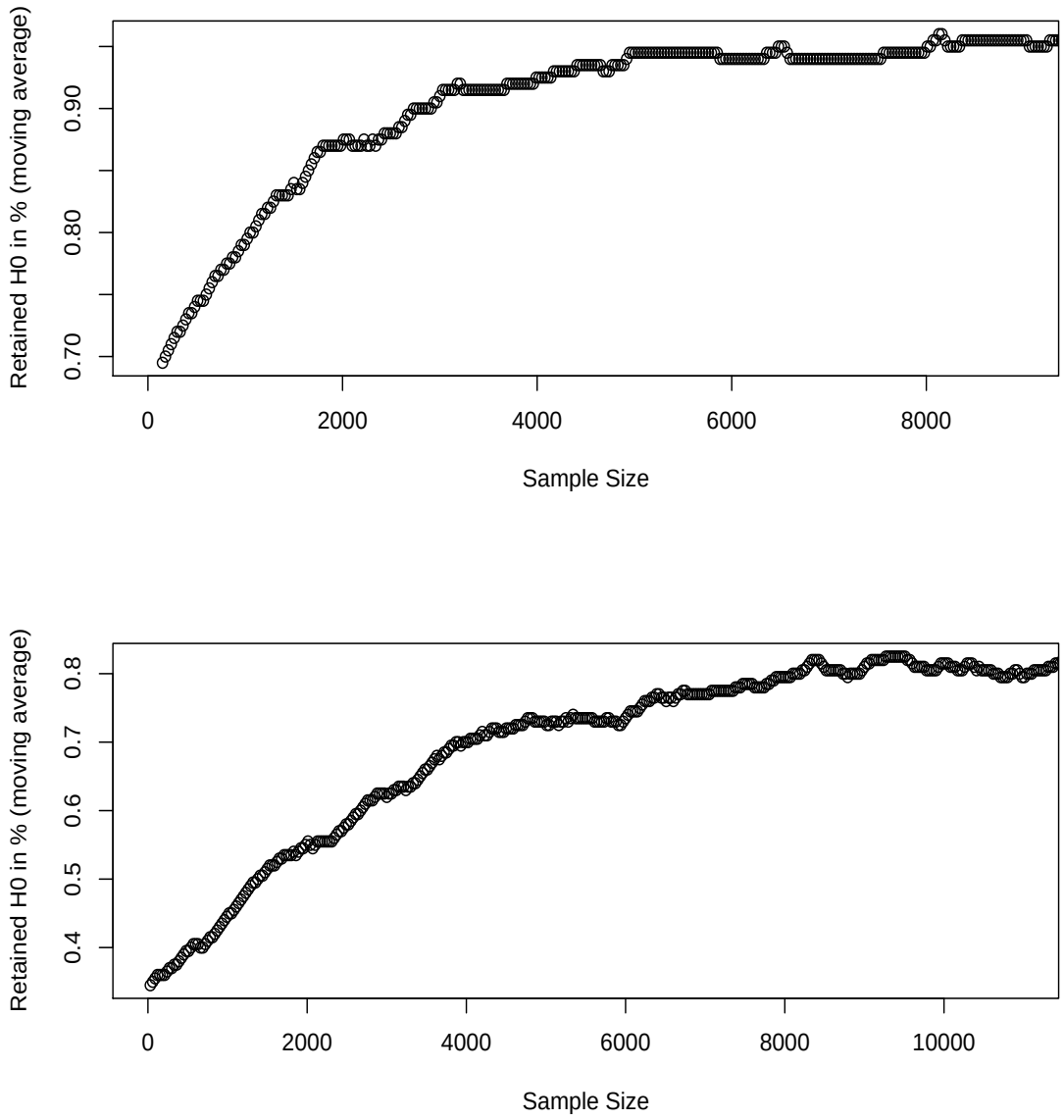

Abbildung 5.9: Untersuchung der empirischen Verteilung des Transformationsparameters für wachsenden Stichprobenumfang mittels eines Shapiro-Wilk Tests.

(d) Vergleich mit der Logarithmierung

Wie bereits in Abschnitt 5.2.1.1 erwähnt, ist die Logarithmierung eine sehr verbreitete, häufig effektive, nichtlineare und umkehrbare Transformation.

Da die vorgeschlagene Potenztransformation ein numerisches Optimierungsverfahren verwendet und insgesamt eine komplexere Struktur besitzt, ist es angebracht zu überprüfen, ob der zusätzliche Aufwand gerechtfertigt ist. Zu diesem Zweck wurden Stichproben aus verschiedenen Verteilungen und für drei Stichprobenumfänge gezogen und anschließend beiden Transformationen unterzogen.

d.1) Die Daten in Abbildung 5.10 wurden aus einer Exponentialverteilung mit Parameter $\lambda = 1$ gezogen. Es handelt sich also dabei um eine rechtsschiefe Verteilung. Es ist offensichtlich, dass die vorgeschlagene Potenztransformation eine bessere Annäherung an eine Normalverteilung als die Logarithmierung erzielt. Außerdem erhöht sich die Qualität der Annäherung mit wachsendem Stichprobenumfang.

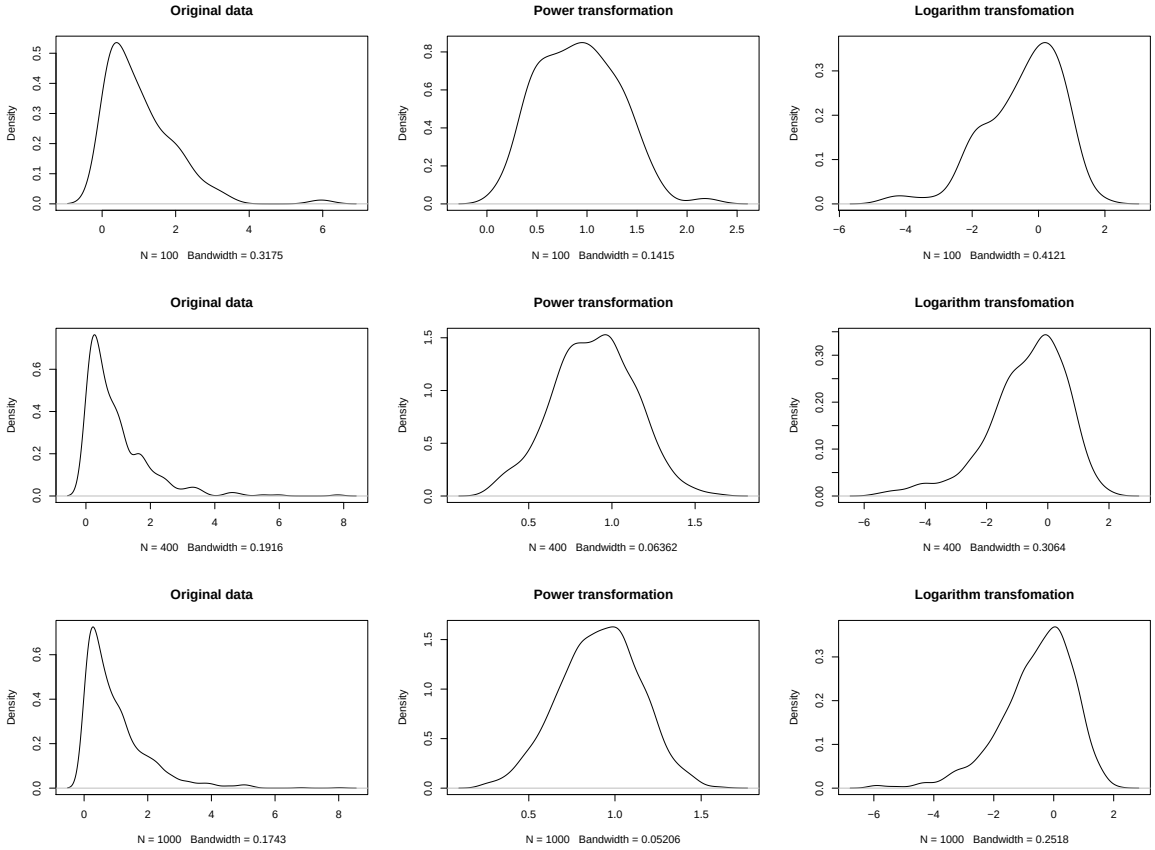

Abbildung 5.10: Vergleich zwischen Potenz- und Logarithmustransformation für exponential-verteilte Daten.

d.2) Eine Zufallsvariable X besitzt dann eine Lognormal-Verteilung, wenn $\ln(X)$ normalverteilt ist. Die Daten für den Vergleich in Abbildung 5.11 wurden aus einer Lognormal-Verteilung gezogen und stellen somit den bestmöglichen Fall für die Logarithmierung dar. Das Ziel dieser Untersuchung ist es, beide Verfahren anhand einer Verteilung zu vergleichen, die möglichst günstig für die Logarithmierung ist. Um die Qualität beider Transformationen zu testen, wurden jeweils 100 Stichproben für verschiedene Umfänge gezogen und die Stichproben anschließend mit beiden Methoden transformiert. Die transformierten Daten wurden darauffolgend einem Shapiro-Wilk Test auf Normalverteiltheit unterzogen und die resultierenden p-Werte gespeichert. Zu einem Signifikanzniveau von 5% führt ein p-Wert kleiner gleich $0,05$ zu einer Ablehnung der Nullhypothese. Der Diskrepanz zwischen beiden Methoden, sowohl bei Ablehnung als auch bei Nicht-Ablehnung, wird ein Wert von Eins und übereinstimmenden Ergebnissen ein Wert von Null zugeordnet. Abbildung 5.11 veranschaulicht die Proportion an abweichenden Ergebnissen für verschiedene Stichprobenumfänge. Die linke Grafik zeigt die Ergebnisse in der ursprünglichen Skalierung, während die rechte Grafik eine Detailansicht bietet.

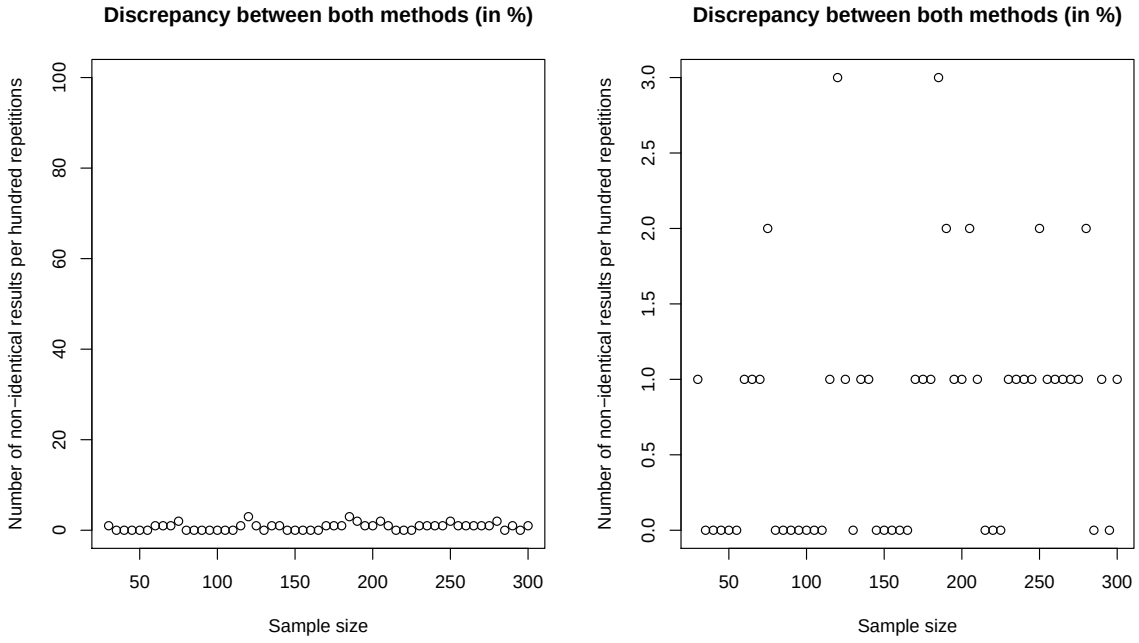

Abbildung 5.11: Vergleich zwischen Potenz- und Logarithmustransformation für lognormal-verteilte Daten. Diskrepanz statistischer Tests.

d.3) Der Vollständigkeit halber wird ein Beispiel vorgestellt, in dem die Potenztransformation keine zufriedenstellenden Ergebnisse erzielt. Die Daten wurden aus einer Gleichverteilung gezogen. Abbildung 5.12 zeigt, dass weder die Potenz- noch die Logarithmustransformation in der Lage sind, die Daten auf eine mit einer Normalverteilung kompatiblen Form zu bringen. Da gleichverteilte Daten mit Hilfe der „Inversen Transformationsmethode“ (vgl. Fishman (2006, S. 77)) nahezu jede beliebige Verteilungsform annehmen können, ist diese Schwachstelle der Potenztransformation mehr von theoretischer als von praktischer Relevanz.

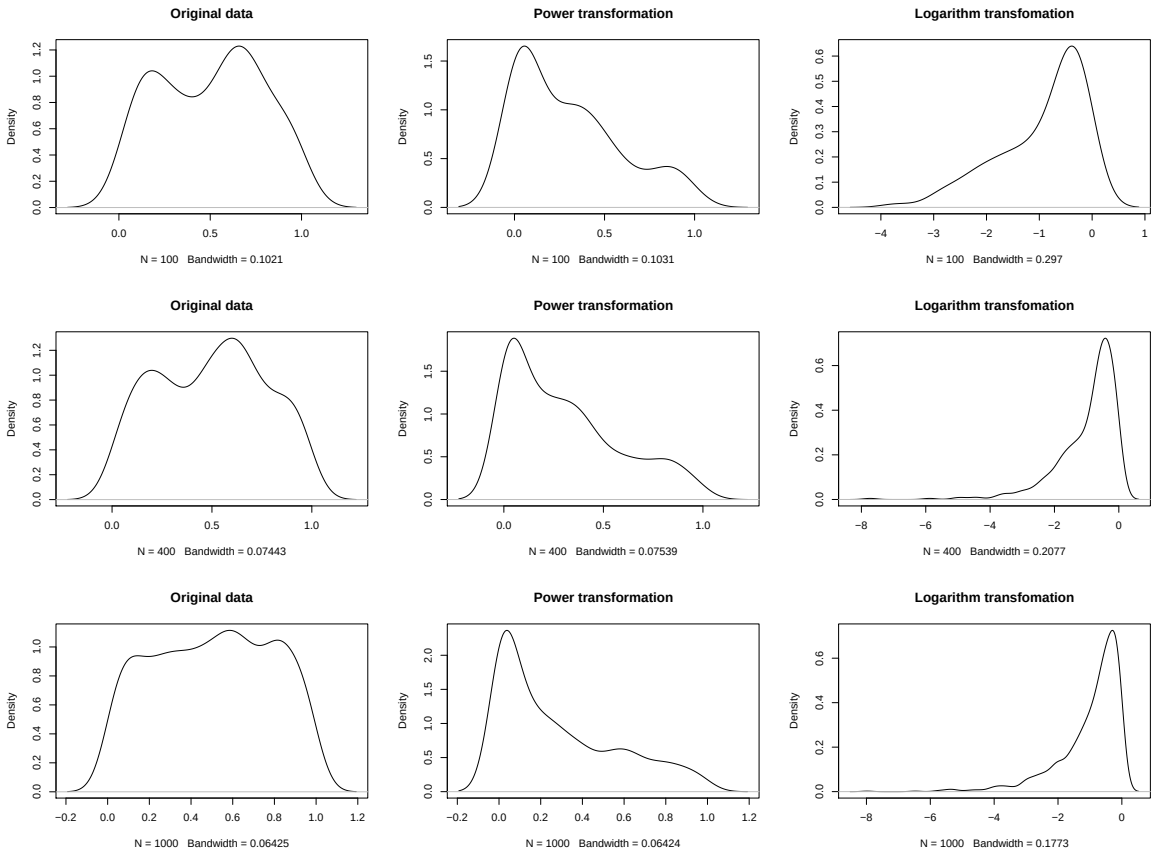

Abbildung 5.12: Potenz- und Logarithmustransformation im Falle gleichverteilter Daten.

(e) Beispiel aus dem KEI-Datensatz

Schließlich wird exemplarisch die Potenztransformation auf den KEI-Indikator A2a3 angewendet, welcher eine etwas für diese Indikatoren untypische Gestalt besitzt, und die Qualität der Transfomation mittels eines Shapiro-Wilk Tests untersucht. Die Daten werden anschließend zurück transformiert, um die Umkehrbarkeitseigenschaft der Potenztransformation grafisch darzustellen.

Die Ergebnisse beider Transformationen sind in Abbildung 5.13 zu sehen. Dieses Beispiel zeigt, dass die vorgeschlagene Methode ebenfalls für linksschiefe Verteilungen geeignet ist. Der Transformationsparameter $\hat{\theta}$ beträgt in diesem Fall 3,868.

Erwähnenswert ist ebenfalls die Tatsache, dass die transformierten Daten trotz des geringen Stichprobenumfangs eine annähernd normalverteilte Gestalt aufweisen. Während der Shapiro-Wilk Test für die untransformierten Daten die Nullhypothese auf allen üblichen Signifikanzniveaus ablehnt, beträgt der p-Wert für die transformierten Daten 0,790.

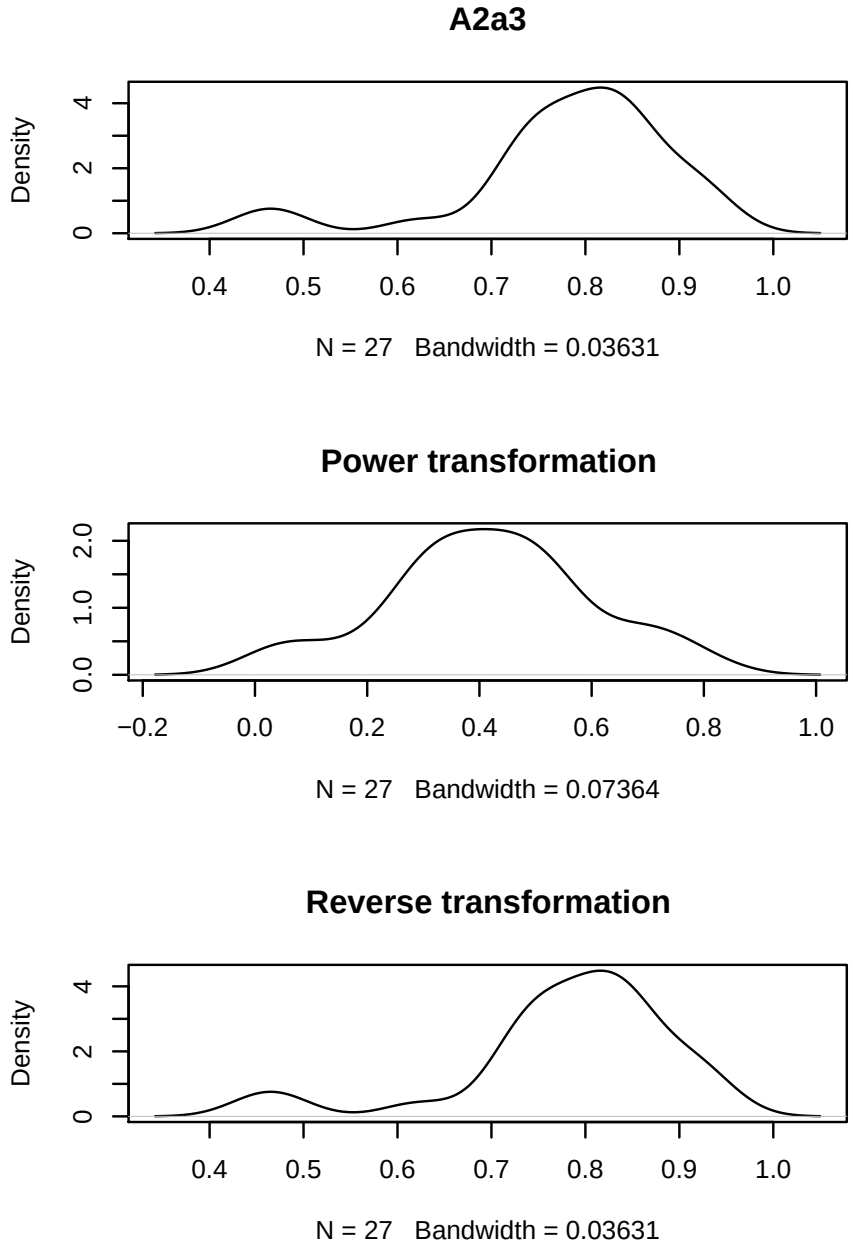

Abbildung 5.13: Inverse Transformation für Indikator A2a3. Der Indikator A2a3 (Werte für 2004) wurde mittels der Potenztransformation auf eine approximative Normalverteiltheit gebracht (der Transformationsparameter $\hat{\theta}$ beträgt 3,868.) und anschließend zurück transformiert. Der Shapiro-Wilk Test für die untransformierten Daten lehnt die Nullhypothese auf allen üblichen Signifikanzniveaus ab, während der p-Wert für die transformierten Daten 0,790 beträgt.

5.2.3 Multivariate Transformation

Eine Möglichkeit der Behandlung von Datensätzen mit fehlenden Daten besteht in der univariaten Transformation der beobachteten Daten. Analog zur allgemeinen Imputationsproblematik und in Abhängigkeit vom Mechanismus, welcher die fehlenden Daten verursacht, können dadurch jedoch erhebliche Verzerrungen auftreten. Diese Verzerrungsmöglichkeit gewinnt zunehmend mit der Anzahl an fehlenden Daten an Bedeutung.

Als besonders gravierendes Beispiel sei der Fall einer rechtsschiefen Verteilung mit überwiegend fehlenden Daten in der rechten Flanke erwähnt. Die ausschließliche Betrachtung der beobachteten Daten könnte eine Umdrehung der Schiefe zur Folge haben. Die Transformationsmethode würde in diesem Fall die Gestalt der Verteilung verzerren und somit die Ausgangssituation verschlechtern.

Es erweist sich also als zweckmäßig, die Daten multivariat zu transformieren.

5.2.3.1 Vollständiger Datensatz

Der mit den bisher vorgestellten Momentenbedingungen ausgestattete Algorithmus ist lediglich in der Lage, univariate Daten auf eine approximative Normalverteilung zu bringen. Da die Normalverteiltheit der Randverteilungen eine notwendige, aber keine hinreichende Bedingung für das Vorhandensein einer multivariaten Normalverteilung ist, verlangt die Behandlung multivariater Daten die Berücksichtigung zusätzlicher Bedingungen. Eine mögliche Erweiterung des Transformationsalgorithmus wird jetzt vorgestellt, welche versucht, dieser Tatsache Rechnung zu tragen.

Es ist erwünscht, dass die hinzukommenden Bedingungen im Einklang mit der Tradition von Box und Cox stehen. Jedoch wird nur die erste der zwei übrig bleibenden Bedingungen, Homoskedastizität und Additivität, als relevant betrachtet. Dies lässt sich durch die unterschiedliche Zielsetzung beider Transfomationsmethoden rechtfertigen, welche ebenfalls die abweichende Behandlung der Variablen motiviert hat (siehe dazu die Ausführung in Abschnitt 5.2.1.2).

Zur Homoskedastizitätsbedingung wird ein bekannter Test der Ökonometrie herangezogen, der Breusch-Pagan Test. Das Prinzip hinter diesem Test besagt grundsätzlich, dass unter Homoskedastizität die Größe der Störterme keine Funktion der Regressoren sein darf. Dies lässt sich als Momentenbedingung folgendermaßen schreiben

$$E\left[x_i[(y-\beta' x_i)^2-\sigma^2]\right] = E\left[x_i[\varepsilon_i^2-\sigma^2]\right] = 0, \tag{5.16}$$

wobei x_i die i-te Zeile der Regressorenmatrix bezeichnet. Das empirische Pendant dieser Momentenbedingung in (5.16) lautet

$$\bar{m} = \frac{1}{n}\sum_{i}^{n} x_i(e_i^2 - s^2), \tag{5.17}$$

wobei e_i die KQ-Residuen und s^2 ihre empirische Varianz bezeichnen (vgl. Greene (2003, S. 506)). Diese letzte Bedingung kann als Orthogonalitätsbedingung zwischen Regressoren und Abweichungen der quadrierten Residuen von ihrem Erwartungswert aufgefasst werden.

5.2.3.2 Struktur des Algorithmus im multivariaten Fall

Es sei $v = (v_1, v_2, \ldots, v_p)$ ein allgemein verteilter multivariater Datensatz mit p Variablen, welche auf approximative multivariate Normalverteiltheit transformiert werden sollen. Die multivariate Version des Algorithmus funktioniert wie folgt:

1. Die Randverteilung der ersten Variablen v_1[12] wird mit dem univariaten Algorithmus transformiert. Diese univariate Transformation ergibt den Parameter $\hat{\theta}_1$.

2. Bei der Behandlung der zweiten Variablen v_2 wird der Tatsache Rechnung getragen, dass v_1 bereits in transformierter Form vorliegt. Um dies zu berücksichtigen, werden die Momentenbedingungen folgendermaßen erweitert:

   (a) **y-Bedingungen:** Zwei Momentenbedingungen für die Randverteilung, analog zum univariaten Algorithmus (i.e. $\overline{m}_1$ and $\overline{m}_2$).

   (b) **e-Bedingungen:**

      i Zwei Momentenbedingungen für die Verteilung der Residuen einer linearen Regression der zu transformierenden auf die bereits transformierte(n) Variable(n).

      ii Eine Momentenbedingung zur Homoskedastizität dieser Residuen.

   Die zweite Variable v_2 wird also derart transformiert, dass

   (a) die Randverteilung approximativ normal ist.

   (b) die Verteilung der Residuen einer linearen Regression auf die bereits transformierte Variable annähernd normal ist.

   (c) die Varianz der Residuen konstant ist.

3. In jedem zusätzlichen Iterationsschritt wird eine neue aus den $p-2$ übrig bleibenden Variablen $(v_3, \ldots, v_p)$ nach dem oben erläuterten Schema transformiert. Obwohl die Anzahl an Regressoren mit jedem Schritt steigt, bleibt die Anzahl an Bedingungen konstant. Dies ist eine wünschenswerte Eigenschaft einer allgemeinen Transformationsmethode, welche zu ihrer numerischen Stabilität beiträgt.

Die e-Bedingungen werden folgendermaßen konstruiert:

Eine lineare Regression von y auf X, wobei y die zu transformierende Variable und X die bereits transformierten Variablen bezeichnen, wird durchgeführt. Diese liefert die Residuen e

$$e = y^* - X\left(X'X\right)^{-1} X'y^*,$$

welche mittels einer z-Transformation standardisiert werden

$$z_e = \frac{e - \overline{e}}{\sigma_e} = \frac{e}{\sigma_e},$$

[12] Ohne Beschränkung der Allgemeinheit wurde die natürliche Reihenfolge gewählt.

wobei $\sigma_e = \sqrt{\frac{1}{n}\sum e^2}$. Die Residuen addieren sich per Definition zu Null auf. Unter der Annahme, dass $z_e \sim N(\mu_{z_e}, \sigma^2_{z_e}) = N(0,1)$ lauten die Momentenbedingungen der dritten und vierten zentrierten Momente dieser z_e:

$$\begin{aligned}
\overline{m}_3 &= \frac{1}{n}\sum_{i=1}^{n}(z_{ei} - \mu_{z_e})^3 = \frac{1}{n}\sum_{i=1}^{n}(z_{ei})^3 \\
\overline{m}_4 &= \frac{1}{n}\sum_{i=1}^{n}(z_{ei} - \mu_{z_e})^4 - 3\sigma^4_{z_e} = \frac{1}{n}\sum_{i=1}^{n}(z_{ei})^4 - 3. \\
\overline{m}_5 &= \frac{1}{n}\sum_{i}^{n} x_i({z_e}_i^2 - 1), \qquad (5.18)
\end{aligned}$$

wobei z_{ei} das i-te Element des Vektors z_e bezeichnet.

Um alle p Transformationsparameter zu erhalten, muss diese Minimierung $p-1$-Mal durchgeführt werden. In jeder Iteration wird eine neue Variable aus dem Datensatz als abhängige Variable y gewählt, während die letzte bereits transformierte Variable, d.h. die abhängige Variable der vorherigen Transformation, die Regressorenmatrix X erweitert. Durch diese Vorgehensweise wird die Unabhängigkeit der Anzahl an Momentenbedingungen von der Anzahl an zu transformierenden Variablen erreicht.

Kommentar zu den e-Bedingungen: Die den e-Bedingungen zugrunde liegende Überlegung ist sehr intuitiv. Da alle bedingten Verteilungen einer multivariaten Normalverteilung ebenfalls normal sind, stellen die Residuen einer multiplen Regression einer Variablen auf die restlichen Variablen unabhängige Ziehungen aus $N(0, \sigma_e^2)$ dar. Dabei hängt σ_e^2 nicht von den Regressoren ab. Somit sind diese Residuen normalverteilt, wie in Abbildung 5.14 veranschaulicht wird.

Die Momentenbedingungen für die Residuen wählen also den Potenzparameter $\hat{\theta} \in \Theta$, für den die Residuen approximativ normalverteilt sind und die Orthogonalitätsbedingung (5.16) möglichst genau erfüllt ist.

Auf diese Art und Weise wird eine approximative gemeinsame Normalverteilung mit Hilfe der Rand- und bedingten Verteilungen nach und nach konstruiert.

Beispiel (Fortsetzung):

Um die in Abschnitt 5.2.1 erläuterten Probleme zu beseitigen, werden die in Beispiel 5.2.1 benutzten Daten der Potenztransformation unterzogen. Die transformierten Daten werden mit dem EM-Algorithmus imputiert und der vervollständigte Datensatz zurück transformiert. Abbildung 5.15 zeigt vergleichend die Ergebnisse der Imputation ohne (linke Grafik) und mit Transformation (rechte Grafik). Es ist offensichtlich, dass die Imputation nach der Transformation die Form der Daten besser abbildet. Noch wichtiger ist die Tatsache, dass keine unzulässigen Werte imputiert werden.

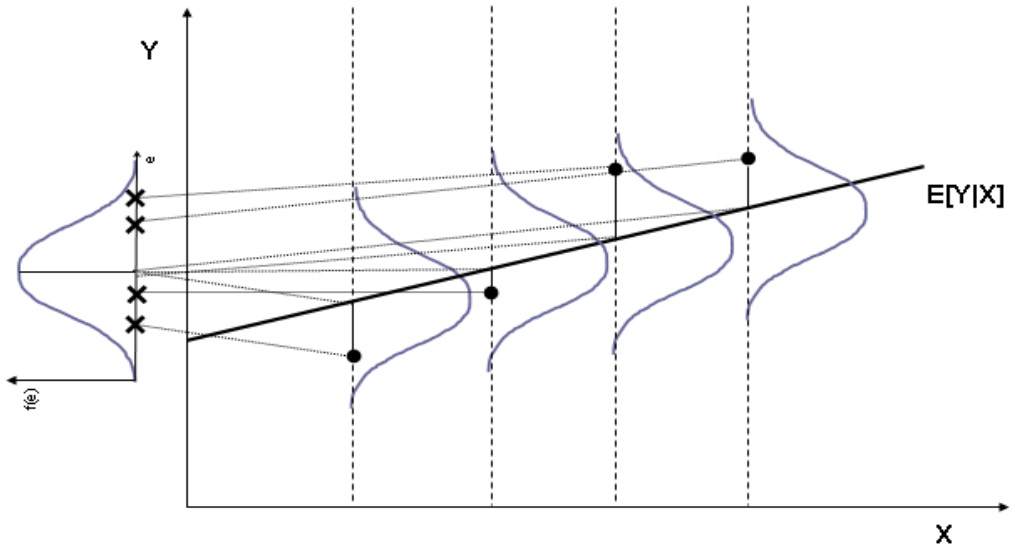

Abbildung 5.14: Die Residuen $e_i := y_i - E[Y|X = x_i]$ im Falle einer bivariaten Normalverteilung sind ebenfalls normalverteilt.

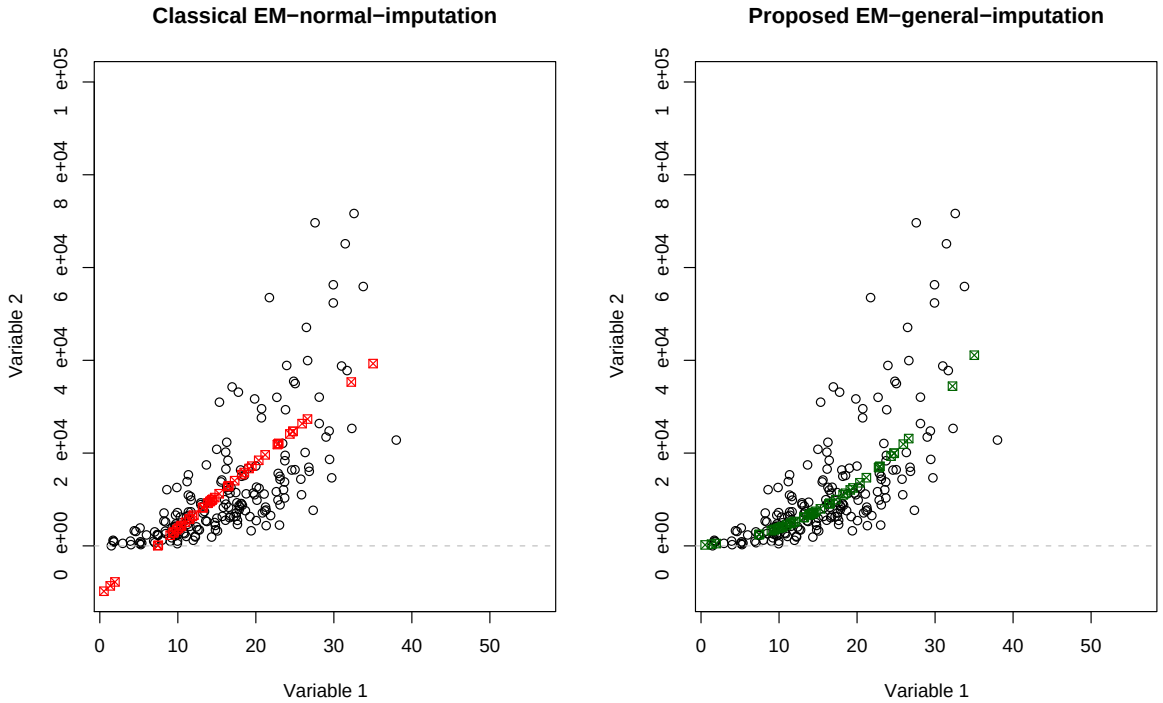

Abbildung 5.15: Vergleich der Imputationen. Die (EM-)Imputation der transformierten Daten weist keine unzulässigen Werte auf.

5.2.3.3 Unvollständige Daten

Der in Abschnitt 5.2.3.1 vorgestellte Algorithmus verwendet lineare Regressionen, um Residuen zu konstruieren, anhand derer die e-Bedingungen überprüft werden können. Um diese Regressionen zu ermöglichen, ist es daher notwendig, die fehlenden Daten *vor der Transformation* zu ergänzen.

Diese Notwendigkeit, Daten vor der Transformation zu ergänzen, bringt jedoch ein kreisförmiges Problem mit sich:

- Der EM-Algorithmus kann unvollständig beobachtete Daten manipulieren, setzt jedoch das Vorhandensein multivariat normalverteilter Variabler voraus.
- Der Transformationsalgorithmus kann die Daten auf eine approximative multivariate Normalverteilungsgestalt bringen, benötigt jedoch vollständig beobachtete Daten.

Diese gegenseitige Abhängigkeit legt ein iteratives Schema nahe, welches abwechselnd die Ergebnisse beider Algorithmen verbessert, bis ein Konvergenzkriterium erreicht wird.

Der neue Algorithmus, welcher als Verbesserungs- bzw. Ergänzungsvorschlag für den klassischen EM-Algortihmus für multivariat-normalverteilte Daten angesehen werden kann, läuft wie folgt ab:

1. Der EM-Algorithmus wird auf die untransformierten Daten angewendet, um einen vollständigen Datensatz zu erhalten.
2. Die ergänzten Daten werden einer *univariaten* Transformation unterzogen. Der resultierende Vektor der Transformationsparameter, $\hat{\theta}_0$, wird gespeichert.
3. Der EM-Algorithmus wird auf die mittels $\hat{\theta}_0$ transformierten Daten angewendet und liefert erneut einen ergänzten Datensatz.
4. Die ergänzten Daten werden einer *multivariaten* Transformation unterzogen. Das Ergebnis der neuen Transformation ist ein Parametervektor $\hat{\theta}_1$.
5. Der EM-Algorithmus wird auf die mit dem Produkt beider Parameter $(\hat{\theta}_0 \cdot \hat{\theta}_1)$ transformierten Daten angewendet, denn es gilt $(X^{\hat{\theta}_0})^{\hat{\theta}_1} = X^{\hat{\theta}_0 \cdot \hat{\theta}_1}$.
6. Die durch Schritte 4 und 5 definierte Iteration wird fortgeführt, bis sich das Produkt der Transformationsparameter mit zusätzlichen Iterationen nicht mehr ändert, d.h. es gilt

$$\prod_{i=1}^{k} \hat{\theta}_i \approx \prod_{i=1}^{k-1} \hat{\theta}_i. \tag{5.19}$$

Dieses Kriterium lässt sich dadurch erklären, dass die Bedingung in (5.19) äquivalent zu $\hat{\theta}_k \approx 1$ ist, d.h. die Form der Verteilung kann nicht durch weitere Iterationen verbessert werden. Da $\hat{\theta}_i$ i.d.R. eine vektorwertige Größe ist, konvergieren die Produkte

$$\prod_{i=1}^{k} \hat{\theta}_i^p, \quad p \in 1, \ldots, m \quad \text{und} \quad m : \text{Dimension des Vektors } \hat{\theta}$$

mit unterschiedlichen Geschwindigkeiten gegen die entsprechenden Elemente von $\hat{\theta}$. Es ist daher notwendig, ein Kriterium heranzuziehen, anhand dessen diese Bedingung skalarwertig überprüft werden kann. Die Wahl

$$\epsilon_i := \max_p \left\{ \prod_{i=1}^{k} \hat{\theta}_i^p - \prod_{i=1}^{k-1} \hat{\theta}_i^p \right\} < \text{Toleranz}$$

zum Abbruch der Iteration wurde für die Zwecke dieser Arbeit bevorzugt und implementiert.

Abbildung 5.16 veranschaulicht die resultierende Struktur des zusammengesetzten Algorithmus (vgl. Abbildung 4.7). Nach Konvergenz des EM + Transformationsalgorithmus werden sowohl die Parameter der gemeinsamen Verteilung als auch die optimalen Transformationsparameter von den MCMC-Methoden übernommen. Nach der multiplen Imputation werden die Daten mit $\hat{\theta}^{-1}$ zurück transformiert.

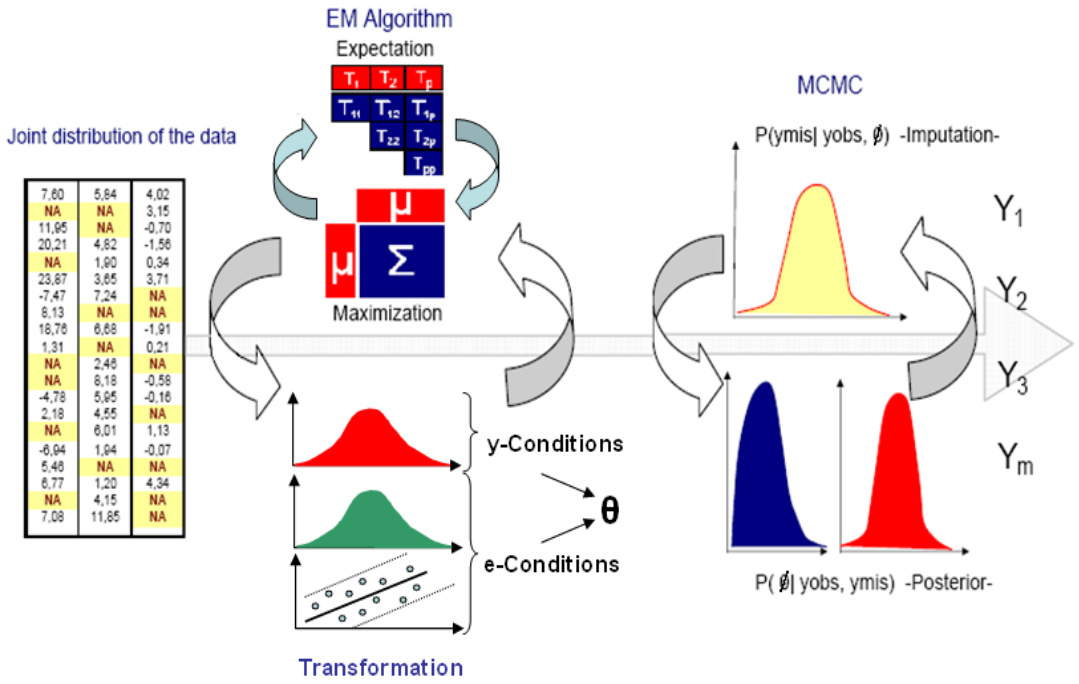

Abbildung 5.16: Zusammengesetzte Methode EM+Transformation.

Kommentare zum vorgeschlagenen Algorithmus: Aufgrund der Verschachtelung von jeweils iterativen Algorithmen ist der analytische Nachweis, dass es sich beim vorgeschlagenen Algorithmus um eine Fixpunktiteration handelt, von einer sehr hohen Komplexität. Die bisherigen Erfahrungen mit dem Algorithmus haben jedoch gezeigt, dass die Konvergenz relativ schnell erreicht wird. Der Fall der Nichtkonvergenz wurde trotz der extensiven Benutzung des Algorithmus im Rahmen des KEI-Projekts nicht beobachtet.

Umfangreiche Simulationen haben ebenfalls gezeigt, dass der vorgeschlagene Algorithmus zufriedenstellende Ergebnisse in praxisnahen Situationen erzielen kann und dass die resultierenden Imputationen akkurater als diejenigen der untransformierten Daten sind.

Auch die Anwendung des vorgeschlagenen Algorithmus im Rahmen des KEI-Projekts hat die o.g. Erfahrungen bestätigt. Die imputierten Daten wurden von allen Partner-Institutionen verwendet. Die Plausibilität der Ergebnisse, insbesondere im Hinblick auf die imputierten Werte, wurde anhand externer Daten von mehreren Forschungsinstitutionen überprüft. Es sind bisher keine Unstimmigkeiten in den Ergebnissen der Partner-Institutionen aufgrund der Imputation bekannt.

Nichtsdestotrotz ist weitere Arbeit notwendig, insbesondere hinsichtlich der Momentenbedingungen, um sowohl Konstellationen, in welchen der Algorithmus ein zuverlässiges Verhalten aufweist, als auch mögliche Gefahren zu identifizieren.

Schlussbemerkungen und Ausblick

Im Rahmen des KEI-Projekts wurde die Universität Tübingen u.a. mit der Aufgabe betraut, einen mit 42% fehlenden Daten behafteten Datensatz zu imputieren, welcher die Grundlage für die Untersuchungen der anderen Partner-Institutionen darstellte. Angesichts des hohen Anteils fehlender Werte wurde zur Imputation des Datensatzes ein Imputationsmodell (Basis-Imputationsmodell, siehe Kapitel 4) ausgewählt, das sich in der Fachliteratur aufgrund seiner hohen Leistungsfähigkeit einer großen Beliebtheit erfreut (vgl. Schafer (1997) bzw. Little und Rubin (2002)).

Untersuchungen des Datensatzes haben jedoch ergeben, dass unterschiedliche Verletzungen der in diesem Basis-Modell getroffenen Annahme einer multivariaten Normalverteilung vorlagen, welche die Qualität der Imputationen gefährden konnten. Die Tatsache, dass die Stichproben per Definition von einem geringen Umfang sind, hat die Lage zusätzlich erschwert.

Die Arbeit hat sich daher auf sechs beobachtete Probleme konzentriert und versucht, sie weitmölichst zu beheben:

1. Präsenz von Ausreißern z.B. aufgrund uneinheitlicher Definition der Indikatoren in den untersuchten Ländern.
2. Große Abweichungen der empirischen Verteilungen von der Normalverteiltheit.
3. Variable meist strikt positiv.
4. Nichtlinearität der Beziehungen zwischen Variablen.
5. Kleine Stichproben.
6. Hoher Anteil an fehlenden Werten (NA's).

Vorgeschlagene Lösungsmöglichkeiten:

1. Präsenz von Ausreißern: Die von Liu und Rubin (1995); Lange et al. (1989); Little (1988); Liu (1995) entwickelten Modelle t-Modell, adaptives-t-Modell und kontaminiertes-normales-Modell, welche als Weiterentwicklungen des Basis-Modells angesehen werden können, wurden implementiert und ihre Eigenschaften wurden untersucht. Diese Modelle sind imstande, mittels selektiver Gewichtung der Beobachtungen, robuste Schätzungen der Parameter einer multivariaten Normalverteilung zu liefern. Die Faktorisierung dieser gemenisamen Verteilung ermöglicht die Imputation der fehlenden Daten.

2.-3. Abweichungen von der Normalverteilung: Viele Variable weisen empirische Verteilungen auf, die mehr oder weniger stark von der im Basis-Modell angenommenen Normalverteilung abweichen. Viele von ihnen sind auf $\mathbb{R}+$ definiert, d.h. negative Werte sind unzulässig. Um diese Abweichungen von der Normalverteilung zu korrigieren bzw. um, sofern möglich, zu verhindern, dass unzulässige Werte imputiert werden, wurde der EM-Algorithmus um eine adaptive Potenztransformation erweitert, die auf der Theorie der verallgemeinerten Momentenmethode basiert. Diese Potenztransformation ist häufig in der Lage, beliebig verteilte Daten auf eine der Normalverteilung ähnliche Form zu bringen und somit eine bessere Schätzung des multivariaten Modells zu ermöglichen.

4. Nichtlinearität der Beziehungen zwischen den Variablen: Die in dieser Arbeit behandelten Algorithmen imputieren mittels linearer Funktionen, d.h. bei Abweichungen von der Linearität in den Beziehungen zwischen den Variablen können Verzerrungen entstehen. Um diesem Problem Rechnung zu tragen, wurde die Potenztransformation derart erweitert, dass eine möglichst gute Annäherung an eine *multivariate* Normalverteilung erzielt werden konnte. Da die Abhängigkeitsstruktur einer multivariaten Normalverteilung linear ist (die Normalverteilung ist elliptisch), sind ihre Regressionskurven (siehe Schaich und Münnich (2001, S. 78)) ebenfalls linear.

5. Kleine Stichproben: Aufgrund der Tatsache, dass jedes Land nur einen Wert pro Indikator liefert, sind die Stichproben von eingeschänktem Umfang. Es war daher wünschenswert, frühere Beobachtungen in die Schätzung mit einzubeziehen. Aufgrund dessen wurde eine einfache Erweiterung der verschiedenen Modelle auf eine Panel Struktur vorgenommen.

   Verschiedene Imputationsmethoden zur Behandlung von Panel-Daten werden von der Literatur vorgeschlagen (z.B. Nijman und Verbeek (1992)). Die modernsten basieren auf den sogenannten *Mixed Effects* Modellen (siehe Schafer und Yucel (2002)). Diese i.d.R. bayesianischen Methoden unterstellen das Vorhandensein einer gemeinsamen Normalverteilung, welche bei den KEI-Daten nicht annehmbar ist. Die Weiterentwicklung dieser Methoden, um Abweichungen von der Normalverteilung zu berücksichtigen, erweist sich als technisch sehr komplex und es ist fraglich, ob die kleinen Stichproben, die typisch für die KEI-Daten sind, diese Komplexität rechtfertigen können.

   Die hier implementierte Panel Struktur kommt durch eine einfache Erweiterung der in Schafer (1997); Little und Rubin (2002) vorgeschlagenen Modelle zustande, welche die Robustheit der Methoden, nämlich die Möglichkeit, mittels selektiver Gewichtung Ausreißer zu neutralisieren, nach Bewältigung einiger technischer Probleme, beibehalten kann. Dabei wurde pro Jahr eine *Dummy*-Variable eingefügt, die als hinzu kommende Variable aufgefasst und mit modelliert wurde.

6. Hoher Anteil an NA's: Es ist bei den vorhandenen Daten häufig vorgekommen, dass ein Land keinen Wert für gewisse Indikatoren im berücksichtigten Zeitintervall geliefert hat (eine ausführlichere Beschreibung befindet sich in Anhang B auf Seite 155). Außerdem gab es ganze Jahre, für die kein Land Werte für einen Indikator geliefert hat. Diese Problematik verhindert bzw. erschwert die Anwendung bekannter Imputationsmodelle, z.B. Modelle des Typs „Repeated measures models“ (siehe Schafer (1997, S. 379:380)), welche die einzelnen Beobachtungen im Laufe der Zeit individuell betrachten. Aufgrund dessen wurden

Modelle unter der (nicht unumstrittenen) i.i.d. Form gewählt, um sicher zu stellen, dass Imputationen immer möglich sind.

Probleme

Die Vorgeschlagenen Lösungsansätze versuchen das Basis-Imputationsmodell auf die besonderen Eigenschaften der Daten auszurichten. Dennoch haben diese Verfahren eigene Probleme, die an dieser Stelle erwähnt werden müssen:

1. Die i.i.d. Annahme ist bei Makrodaten wenig realistisch. Länder weisen in der Regel räumliche Abhängigkeiten auf, d.h. die geografische Lage hat Informationsgehalt bezüglich gewisser, für das KEI-Projekt relevanter, Eigenschaften der Länder (man denke z.B. an eine Einteilung in Nordwest-, Südwest- und Osteuropa).
2. Die Transformationsmethode basiert auf der Schätzung von höheren Momenten einer Verteilung und ist daher eher für große Stichproben geeignet (vgl. Hayashi (2000, S. 215)). Die zur Konstruktion von zusammengesetzten Indikatoren benutzten Datensätze haben per Definition einen kleinen Umfang.
3. Die Mischverteilungsmodelle (siehe Abschnitt 5.1) erlauben keine selektive Gewichtung der Elemente einer Zeile. Dies ist insbesondere dann problematisch, wenn ein Land, aufgrund einer abweichenden Definition **eines** Indikators, als Ausreißer identifiziert wird. **Alle** Indikatoren für dieses Land gehen dann in die Schätzung der Modellparameter mit einem geringen Gewicht ein.

Ausblick

Wie im vorherigen Abschnitt erläutert wurde, mussten im KEI-Projekt bei einer sehr schlechten Datenlage die verschiedenen Untersuchungen durchgeführt werden. Die Frage, ob ein Datensatz mit 42% fehlenden Werten zur Grundlage für die Untersuchung sowohl methodisch als auch politisch relevanter Fragestellungen genommen werden kann, ist durchaus gerechtfertigt. Trotz der Vielzahl an vorgeschlagenen Methoden zur Verbesserung der Datenlage und der ermutigenden Ergebnisse zahlreicher Simulationen zur Überprüfung ihrer Funktionsfähigkeit darf diese Ausgangssituation bei Vorhandensein ergänzter Daten nicht vergessen werden.

Andererseits gibt es in der Praxis zahlreiche Situationen, in denen die Möglichkeit, eine Arbeit einzustellen bzw. zu warten bis verlässlichere Daten vorhanden sind, nicht in Frage kommt oder mit erheblichen Kosten verbunden ist. Im konkreten Fall des KEI-Projekts war die Möglichkeit „die Daten in Ruhe zu lassen“, wie ein namhafter deutscher Ökonometriker dem Verfasser dieser Arbeit bei Betrachtung der Datenlage geraten hatte, überhaupt nicht vorhanden und jede Verzögerung hätte die Gefahr mit sich gebracht, dass das Projekt nicht rechtzeitig hätte abgeschlossen werden können.

Die Unvereinbarkeit beider Positionen und der sich ergebende Zwiespalt insbesondere im Fall methodisch ausgebildeter Entscheidungsträger hat auch zahlreiche wissenschaftliche Arbeiten motiviert (siehe z.B. Klaxton (1999) für eine interessante entscheidungstheoretische Diskussion bzgl. der „Irrelevanz der Signifikanz“).

Bei der Implementierung bzw. Weiterentwicklung der in dieser Arbeit betrachteten Methoden wurde immer Wert darauf gelegt, einen Kompromiss zwischen beiden Positionen zu finden. Die betrachteten Methoden besitzen eine solide statistische Basis, in deren Rahmen diese Verfahren weiterentwickelt wurden. Der Schwerpunkt der Arbeit liegt jedoch in der praktischen Anwendbarkeit der untersuchten Verbesserungsmöglichkeiten unter besonderer Berücksichtigung der Rahmenbedingungen des KEI-Projekts.

Es ist allerdings zu erwarten, dass das Projekt einen Beitrag dazu leisten wird, die Kluft zwischen beiden Positionen zu verringern und die zukünftige Datenlage aus den folgenden Gründen zu verbessern:

- Die mit dem Projekt verbundene Problematik der fehlenden Werte wird die statistischen Ämter der verschiedenen Länder für die Notwendigkeit der rechtzeitigen Messung und Lieferung der Indikatoren sensibilisieren.
- Da sogar die einzelnen Indikatoren oft aus einem Aggregationsverfahren resultieren, könnte die oben genannte Sensibilisierung dazu führen, dass die Imputation in Zukunft auf einem *micro level* vorgenommen wird. Solch eine Imputation (auch mit den vorgeschlagenen Methoden) würde aufgrund zweier Tatsachen viel bessere Schätzungen ergeben:
  1. Die Verschiebung der statistischen Einheiten von „Land“ auf eine niedrigere Ebene, z.B. „Stadt“, macht die i.i.d. Annahme plausibler.
  2. Die Stichprobenumfänge sind auf einem *micro level* i.d.R. von einem erheblich größeren Umfang.
- Aufgrund der Vereinheitlichung in der Definition der einzelnen Indikatoren ist mit weniger Ausreißern zu rechnen.

Es ist deshalb zu erwarten, dass die konsequente Anwendung der *Knowledge Economy Indicators* zu einer Erhöhung der Qualität der Indikatoren, einer Harmonisierung ihrer Konstruktion und Messung und einer Verringerung der Anzahl an fehlenden Werten führen wird.

Anhang A

Grundlagen zur Theorie der Markov-Ketten

MCMC-Methoden konstruieren eine Markov-Kette, deren Verteilung gegen eine vorgegebene Zielverteilung konvergiert. Um sicher zu stellen, dass die erzeugten Markov-Ketten tatsächlich eine invariante (stationäre) Verteilung besitzen bzw. dass die Voraussetzungen für den *Ergodensatz* erfüllt sind, müssen diese gewisse Eigenschaften besitzen. Ziel dieses Abschnitts ist es, Ergebnisse der Theorie der Markov-Ketten bereit zu stellen, um die folgenden beiden Fragen zu beantworten:

1. Mit welchen Eigenschaften müssen Markov-Ketten ausgestattet sein, damit sie *eine eindeutige* invariante Verteilung besitzen?
2. Welche Ergebnisse ermöglichen die *Konstruktion* von Markov-Ketten, die die gewünschte Verteilung als invariantes Maß besitzen?

Zu diesem Zweck werden im Folgenden die wichtigsten Ergebnisse der Theorie der Markov-Ketten zusammengestellt, unter besonderer Berücksichtigung ihrer Relevanz für Markov-Chain-Monte-Carlo-Methoden.

Auch wenn die meisten mittels MCMC-Methoden zu simulierenden Verteilungen einen überabzählbaren Wertebereich besitzen, können ohne Beschränkung der Allgemeinheit die meisten Eigenschaften der zugrunde liegenden Markov-Ketten für den Fall eines endlichen Zustandsraums bewiesen werden. Das liegt daran, dass der Übergang in der Regel keine konzeptionellen Änderungen mit sich bringt, die Darstellung jedoch deutlich technischer macht (vgl. Gamerman (1997, S. 93)). Aufgrund des Bestrebens, eine anschauliche und gleichzeitig mathematisch korrekte Darstellung anzubieten, wird ein endlicher Zustandraum gewählt.

A.1 Allgemeine Eigenschaften von Markov-Ketten

Definition: Zufallsvariable

Es sei $(\Omega, \mathfrak{F}, \mathrm{P})$ ein Wahrscheinlichkeitsraum, S eine abzählbare Menge und $\mathfrak{P}$ die Potenzmenge von S. Dann heißt die Abbildung

$$X : \Omega \to S,\ \omega \mapsto X(\omega)$$

eine Zufallsvariable auf dem von S induzierten Wahrscheinlichkeitraum, falls X $\mathfrak{F}$-$\mathfrak{P}$-messbar ist.

■

Definition: Markov-Eigenschaft

Es sei $(X_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen mit abzählbarem Zustandsraum S. Dann ist $(X_n)_{n\in\mathbb{N}}$ eine Markov-Kette, d.h. $(X_n)_{n\in\mathbb{N}}$ besitzt die Markov-Eigenschaft, falls für alle $n \in \mathbb{N}$ und für alle $i_0, \ldots, i_{n+1}$[1] mit $\mathrm{P}[X_0 = i_0, \ldots, X_n = i_n] > 0$ gilt:

$$\mathrm{P}[X_{n+1} = i_{n+1} | X_0 = i_0, \ldots, X_n = i_n] = \mathrm{P}[X_{n+1} = i_{n+1} | X_n = i_n].$$

■

Bemerkungen:

- Die Markov-Eigenschaft besagt, dass die Zukunft, d.h. der Zustand zum Zeitpunkt $t = n+1$, nicht von der kompletten Vergangenheit, sondern nur von der Gegenwart, d.h. vom Zustand zum Zeitpunkt $t = n$ abhängt.
- Eine Markov-Kette heißt *zeitlich homogen*, falls zusätzlich für alle $i, j \in S$

  $$\mathrm{P}[X_{n+1} = j | X_n = i]$$

  nicht von n abhängt.

Definition: Stochastische Matrix

Die Matrix $K = (K(i,j))_{i,j\in S}$ heißt *stochastische Matrix*, falls gilt:

- $K(i,j) \geq 0$ für alle $i, j \in S$ und
- $\sum_{j\in S} K(i,j) = 1$ für alle i aus S (d.h. alle Zeilensummen von K sind gleich 1).

■

[1] Im Folgenden werden die Zustände einer Markov-Kette mit abzählbarem Zustandsraum mit $i, j \in S$, wenn nur zwei Zustände und mit i_n, $i \in S$ und $n \in \mathbb{N}_0$, wenn mehr als zwei Zustände betrachtet werden, gekennzeichnet.

Diese Definitionen können folgendermaßen miteinander kombiniert werden:

Proposition: *Es sei $(X_n)_{n\in\mathbb{N}_0}$ eine homogene Markov-Kette. Dann existiert eine stochastische Matrix K mit der Eigenschaft, dass für alle $n \in \mathbb{N}$ und $i_0, \ldots, i_n \in S$ gilt*

$$\mathrm{P}[X_0 = i_0, \ldots, X_n = i_n] = \mathrm{P}[X_0 = i_0]\, K(i_0, i_1)\, K(i_1, i_2) \cdots K(i_{n-1}, i_n)$$

Beweis: Siehe Resnick (2002)

□

Interpretation: Der zeitliche Verlauf der Markov-Kette lässt sich anhand des Anfangzustands und der stochastischen Matrix darstellen. Diese stochastische Matrix beinhaltet die Wahrscheinlichkeiten dafür, dass der Prozeß den Zustand j im Zeitpunkt $n+1$ annimmt, wenn er sich zum Zeitpunkt n in i befindet, für alle $i, j \in S$. Aufgrund dessen wird sie *Übergangsmatrix* genannt. Das überabzählbare Pendant einer Übergangsmatrix wird als *Übergangskern* bezeichnet.

Random-Walk-Markov-Kette

Im Folgenden wird eine Darstellung einer Markov-Kette behandelt, die sich aus verschiedenen Gründen als sehr nützlich erweist. Einerseits ist es möglich, auf einfache Art und Weise mit Hilfe dieser Darstellung gewisse Eigenschaften der Markov-Ketten zu beweisen. Andererseits verwenden tasächlich bekannte MCMC-Methoden wie der *Random-Walk Metropolis-Hastings*-Algorithmus Markov-Ketten, welche diese Struktur aufweisen.

Satz: *Es seien die Zufallsvariable X_0 und die Folge von Zufallsvariablen $(Y_n)_{n\in\mathbb{N}}$ definiert. Dabei seien $X_0, Y_1, Y_2, \ldots$ unabhängig und $(Y_n)_{n\in\mathbb{N}}$ identisch verteilt. Ferner sei eine messbare Funktion f und die folgende Rekursion für alle $n \in \mathbb{N}$ definiert:*

$$X_{n+1} := f(X_n, Y_{n+1}).$$

Dann ist $(X_n)_{n\in\mathbb{N}_0}$ eine homogene Markov-Kette.

Beweis: Es gilt

$$\mathrm{P}[X_{n+1} = i_{n+1} | X_0 = i_0, \ldots, X_n = i_n] = \frac{\mathrm{P}[X_0 = i_0, \ldots, X_n = i_n, f(i_n, Y_{n+1}) = i_{n+1}]}{\mathrm{P}[X_0 = i_0, \ldots, X_n = i_n]}.$$

Da $X_0 = i_0, \ldots, X_n = i_n \in \sigma(X_0, Y_0, \ldots, Y_n)$[2] und $f(i_n, Y_{n+1}) \in \sigma(Y_{n+1})$ sind diese Ereignisse unabhängig. Daraus folgt:

$$\mathrm{P}[X_{n+1} = i_{n+1} | X_0 = i_0, \ldots, X_n = i_n] = \mathrm{P}[f(i_n, Y_{n+1} = i_{n+1})].$$

[2] Mit $\sigma(X)$ wird die kleinste σ-Algebra gekennzeichnet, bezüglich der die Ergebnisse in X messbar sind. Sie wird als die von X *generierte* σ-Algebra bezeichnet.

Andererseits gilt:

$$\begin{aligned}\mathrm{P}[X_{n+1} = i_{n+1} | X_n = i_n] &= \frac{\mathrm{P}[X_n = i_n, f(i_n, Y_{n+1} = i_{n+1}]}{\mathrm{P}[X_n = i_n]} \\ &= \mathrm{P}[f(i_n, Y_{n+1} = i_{n+1})]\end{aligned}$$

aufgrund der Abhängigkeit beider Ereignisse.

□

Proposition (Chapman-Kolmogorov Gleichung): *Gegeben sei $(X_n)_{n\in\mathbb{N}_0}$ eine Markov-Kette mit Zustansdraum S und $t \leq u \leq v$. Dann gilt für alle $i, j \in S$:*

$$\mathrm{P}[X_v = j | X_t = i] = \sum_k \mathrm{P}[X_v = j | X_u = k] \, \mathrm{P}[X_u = k | X_t = i]. \tag{A.1}$$

Beweis: Es gilt:

$$\mathrm{P}[X_v = j | X_t = i] = \frac{\mathrm{P}[X_t = i, X_v = j]}{\mathrm{P}[X_t = i]}.$$

Anwendung des Gesetzes der totalen Wahrscheinlichkeit ergibt

$$\begin{aligned}\mathrm{P}[X_v = j | X_t = i] &= \frac{\sum_k \mathrm{P}[X_v = j, X_t = i | X_u = k] \, \mathrm{P}[X_u = k]}{\mathrm{P}[X_t = i]} \\ &= \frac{\sum_k \mathrm{P}[X_v = j, X_u = k, X_t = i]}{\mathrm{P}[X_t = i]} \\ &= \frac{\sum_k \mathrm{P}[X_v = j | X_u = k, X_t = i] \mathrm{P}[X_u = k, X_t = i]}{\mathrm{P}[X_t = i]}.\end{aligned}$$

Aufgrund der Markov-Eigenschaft ist dieser Ausdruck äquivalent zu

$$\begin{aligned}\mathrm{P}[X_v = j | X_t = i] &= \frac{\sum_k \mathrm{P}[X_v = j | X_u = k] \, \mathrm{P}[X_u = k, X_t = i]}{\mathrm{P}[X_t = i]} \\ &= \frac{\sum_k \mathrm{P}[X_v = j | X_u = k] \, \mathrm{P}[X_u = k | X_0 = i] \, \mathrm{P}[X_t = i]}{\mathrm{P}[X_t = i]} \\ &= \sum_k \mathrm{P}[X_v = j | X_u = k] \, \mathrm{P}[X_u = k | X_t = i].\end{aligned}$$

□

Interpretation: Die Chapman-Kolmogorov Gleichung besagt, dass eine Bewegung von einem Zustand i in einen Zustand j in $v - t$ Schritten folgendermaßen bewerkstelligt werden kann: Vom Zustand i bewegt sich der Prozeß in $u - t$ Schritten zu einem dazwischenliegenden Zustand k. Dies geschieht mit Wahrscheinlichkeit $p_{i,k}^n$, wobei $p_{i,k}^n$ das Element in der i-ten Zeile und k-ten Spalte der n-Schritt Übergangsmatrix K^n darstellt. Anschließend bewegt sich der Prozeß in $v - u$ Schritten von k nach j mit Wahrscheinlichkeit $p_{k,j}^m$. Die Markov-Eigenschaft stellt sicher, dass die Wahrscheinlichkeit der Bewegungen nicht von der Vorgeschichte bis zur Erreichung des Zustands k abhängt. Aufgrund des Gesetzes der totalen Wahrscheinlichkeit erfolgt die Berechnung der $(v - t)$-Schritt-Übergangswahrscheinlichkeit durch Summation über alle k.

A.2 Einige Grundbegriffe der Stochastik

Definition: Filtration

Sei Ω eine nichtleere Menge und $(\mathfrak{F}_n)_{n\in\mathbb{N}_0}$ eine Folge von $\sigma-$Algebren auf Ω mit der Eigenschaft, dass für alle n aus $\mathbb{N}_0$ gilt: $\mathfrak{F}_0 \subseteq \mathfrak{F}_1 \subseteq \mathfrak{F}_2 \subseteq \ldots \subseteq \mathfrak{F}$ (monoton steigende Folge). Dann heißt $(\mathfrak{F}_n)_{n\geq 0}$ Filtration in $\mathfrak{F}$.

■

Interpretation: Unter einer Filtration $\mathfrak{F}_n,\ n \in \mathbb{N}_0$, wird in der Stochastik eine Folge von σ-Algebren auf einem Grundraum Ω verstanden, die einen steigenden Fluss an Informationen mathematisch modellieren. Üblicherweise enthalten Filtrationen Informationen über den Verlauf eines stochastischen Prozesses. Zu jedem Zeitpunkt $n \in \mathbb{N}_0$ geben die Mengen der σ-Algebra $\mathfrak{F}_n$ an, wie viele Informationen bekannt sind und ermöglichen somit die eindeutige Kenntnis über das Eintreffen eines Ereignisses ω aus diesen Mengen. Da diese Folge von σ-Algebren monoton steigend ist, gehen die schon hinein geflossenen Informationen nie verloren. Eine Filtration kann zu jedem Zeitpunkt ausschließlich Informationen über den Verlauf eines stochastischen Prozesses oder zusätzliche Informationen enthalten. Im ersten Fall wird sie als *natürliche Filtration* oder *vom stochastischen Prozeß generierter Filtration* bezeichnet. Im zweiten Fall enthält die Filtration zusätzliche Informationen und sie wird als größer als die natürliche Filtration bezeichnet.

Definition: Stoppzeit

Gegeben sei eine Filtration $(\mathfrak{F}_n)_{n\in\mathbb{N}_0}$ auf Ω . Die Abbildung $\tau : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ heißt Stoppzeit bezüglich der Filtration $(\mathfrak{F}_n)_{n\in\mathbb{N}_0}$, wenn es für alle n aus $\mathbb{N}_0 \cup \{\infty\}$ gilt:

$$\{\omega : \tau(\omega) \leq n\} \in \mathfrak{F}_n. \tag{A.2}$$

■

Bemerkungen:

a) Bedingung (A.2) besagt, dass das Ereignis $\tau(\omega) \leq n$ in der σ-Algebra $\mathfrak{F}_n$ enthalten ist. Ereignisse, die in einer Filtration zum Zeitpunkt n enthalten sind, werden als $\mathfrak{F}_n$-messbar bezeichnet.

b) Es ist leicht ersichtlich, dass

$$\mathfrak{F}_\infty = \left(\bigcup_{n\in\mathbb{N}_0} \mathfrak{F}_n \right)$$

gilt.

c) Ohne Beschränkung der Allgemeinheit kann $\{\tau \leq n\}$ durch $\{\tau = n\}$ ersetzt werden, denn es gilt

$$\{\tau = n\} = \underbrace{\{\tau \leq n\}}_{\in\mathfrak{F}_n} \setminus \underbrace{\{\tau \leq n-1\}}_{\in\mathfrak{F}_{n-1}} \in \mathfrak{F}_n.$$

d) Die Ereignisse $\{\tau = n\}, \{\tau \leq n-1\}$ und $\{\tau \leq n\}$ sind also $\mathfrak{F}_n$-messbar.

Interpretation: Eine Stoppzeit ist einerseits eine Zufallsvariable bezüglich $(\Omega, \mathfrak{F})$, da $\mathfrak{F}_n \in \mathfrak{F}$. Andererseits ist sie bezüglich $(X_n)_{n\in\mathbb{N}_0}$ eine Entscheidungsregel, die den Prozeß stoppt, wenn ein gewisses Ereignis eintrifft. Ob das Ereignis eingetroffen ist oder nicht muss zum Zeitpunkt n anhand der Beobachtungen des Prozesses $x_0, x_1, \ldots, x_n$ entscheidbar sein. Insbesondere darf die Stoppregel nicht von zukünftigen Ereignissen abhängen.

Definition: Eintrittszeit eines Prozesses

Es sei $(X_n)_{n\in\mathbb{N}_0}$ eine Folge von reellwertigen Zufallsvariablen. Ferner sei mit B eine Ereignismenge aus der Borel-σ-Algebra $\mathfrak{B}(\mathbb{R})$ bezeichnet. Dann ist

$$\tau_B := \inf\{n \in \mathbb{N}_0 : X_n \in B\}$$

eine Stoppzeit bezüglich $(X_n)_{n\in\mathbb{N}_0}$.

■

Interpretation: Diese Stoppzeit stellt den kleinsten Zeitpunkt dar, in dem der Prozeß Werte aus B annimmt.

A.3 Spezielle Eigenschaften von Markov-Ketten

Definition: Erreichbarkeit

Ein Zustand j heißt von einem Zustand i erreichbar ($i \rightsquigarrow j$), wenn gilt:

$$P_i(X_n = j \text{ für ein } n \in \mathbb{N}_0) > 0.$$

Dabei bezeichnet $P_i(\cdot)$ die Wahrscheinlichkeit eines Ereignisses, gegeben dass die Markov-Kette im Zustand i gestartet ist.

■

Satz: *Gegeben seien zwei Zustände $i, j \in S$ mit $i \neq j$. Dann sind folgende Aussagen äquivalent:*

a) $i \rightsquigarrow j$.

b) Es gibt eine Folge von Zuständen $i_0 = i, i_1, \ldots, i_{n-1}, i_n = j$ mit Wahrscheinlichkeiten $p_{i_0,i_1}, p_{i_1,i_2}, \ldots, p_{i_{n-1},i_n} > 0$.

c) Es gibt ein $n \in \mathbb{N}$ mit $p^n_{i,j} > 0$.

d) $P_i[\tau_j < \infty] > 0$.

Beweise [a, b, c]: Siehe Waldmann und Stocker (2004, S. 24).

Beweis [d]: Siehe Resnick (2002, S. 78).

□

Definition: Verbundenheit

Ist j von i und i von j erreichbar, so sind i und j verbunden. Dies wird mit $i \sim j$ bezeichnet.

■

Proposition: *Die Relation $\sim$ erfüllt die Eigenschaften einer Äquivalenzrelation und führt zu einer Zerlegung von S in disjunkte Teilmengen verbundener Zustände (siehe Waldmann und Stocker (2004)).*

Beweis: Um zu beweisen, dass eine Äquivalenzrelation auf einer Menge vorliegt, müssen drei Eigenschaften überprüft werden (vgl. Lehmann und Schulz (2004)):

1. Reflexivität
2. Symmetrie
3. Transitivität

Reflexivität: Ist erfüllt, denn es gilt offensichtlich $i \sim i$.

Symmetrie: Ist erfüllt, denn laut Definition $i \sim j \Longrightarrow j \sim i$.

Transitivität: Zu zeigen gilt, dass aus $i \sim j$ und $j \sim k$, $i \sim k$ folgt.

Aufgrund der Verbundenheit von i, j und j, k gibt es $n, m \in \mathbb{N}_0$ mit der Eigenschaft, dass $K^n(i,j) > 0$ und $K^m(j,k) > 0$ gilt. Ferner ist

$$K^{n+m}(i,j) = (K^n K^m)(i,k).$$

Für ein allgemeines Element j^* gilt

$$\begin{aligned}(K^n K^m)(i,k) &= \sum_{j^* \in S} K^n(i,j^*) K^m(j^*,k) \\ &\geq K^n(i,j) K^m(j,k) > 0.\end{aligned}$$

Daraus folgt die Behauptung $i \sim k$.

□

Definition:

Eine Markov-Kette mit Übergangsmatrix K heißt irreduzibel, wenn der Zustandsraum S lediglich aus einer Äquivalenzklasse besteht, d. h. für alle $i, j \in S$ gilt $i \sim j$.

■

Definition: Rekurrenz und Transienz

Gegeben sei eine Markov-Kette $(X_n)_{n\in\mathbb{N}_0}$ mit abzählbarem Zustandsraum S. Ein Zustand $i \in S$ heißt rekurrent, falls $P_i(X_n = i \text{ für unendlich viele } n) = 1$ gilt. Der Zustand heißt transient, falls $P_i(X_n = i \text{ für unendlich viele } n) = 0$ gilt.

■

Bemerkung: Ein rekurrenter Zustand wird im Laufe der Zeit unendlich oft angenommen, ein transienter Zustand nur endlich oft. Da die Rekurrenz eine Äquivalenzklasse bezüglich der Relation $\sim$ definiert, sind alle Zustände einer irreduziblen Markov-Kette entweder rekurrent oder transient und man spricht von einer rekurrenten bzw. transienten Markov-Kette.

Definition: Periode eines Zustands

Die Periode eines Zustands i, bezeichnet als d_i, ist definiert als der größte gemeinsame Teiler (ggT) der Menge $D(i) \subseteq \mathbb{N}_0$, welche wie folgt definiert ist:

$$D(i) := \{n \in \mathbb{N}_0 : K^n(i,i) > 0\}.$$

Dabei soll $d(i) = \infty \quad \text{falls} \quad D(i) = \{0\}$.

■

Proposition: *Die Periode ist eine Klasseneigenschaft bezüglich der Relation $\sim$, d.h. aus $i \sim j$ folgt $d(i) = d(j)$.*

Beweis: Siehe Gamerman (1997, S. 104).

□

Mit Hilfe von $d(i)$ kann nun die *Periodizität bzw. Aperiodizität* definiert werden.

Definition: Aperiodizität

Eine Übergangsmatrix K heißt aperiodisch, falls für alle $i \in S$ gilt $\quad d(i) = 1$.

■

Bemerkung: Obwohl die Aperiodizität keine notwendige Bedingung für die Existenz einer stationären Verteilung darstellt, ist sie relevant für die Konvergenz der Übergangswahrscheinlichkeiten.

Proposition: *Gegeben sei eine irreduzible und positiv rekurrente Markov-Kette $(X_n)_{n\in\mathbb{N}_0}$ mit Übergangsmatrix K. Dann gilt*

(a) Wenn $(X_n)_{n\in\mathbb{N}_0}$ aperiodisch ist, dann ist $\lim_{n\to\infty} K^n(i,j) = \pi(j)$ für alle $i, j \in S$.

(b) Wenn $(X_n)_{n\in\mathbb{N}_0}$ periodisch mit Periode d ist, dann gibt es für alle $i, j \in S$ eine Zahl $r \in \mathbb{N}_0$ mit $0 \leq r < d$ und der Eigenschaft, dass $K^r(i,j) = 0$.

Beweis: Siehe Gamerman (1997, S. 104)

□

Existenz eines invarianten Maßes

Definition: Stationarität

Eine Folge $(X_n)_{n\in\mathbb{N}_0}$ von Zufallsvariablen auf $(\Omega, \mathfrak{F}, \mathrm{P})$ heißt *stationär*, falls die Verteilung $\mathrm{P}_{(X_{n+k})_{n\in\mathbb{N}_0}}$ von $(X_{n+k})_{n\in\mathbb{N}_0}$ nicht von k abhängt, d.h. für alle $n, k \in \mathbb{N}_0$ gilt:

$$\mathrm{P}_{(X_0,X_1,\ldots,X_n)} = \mathrm{P}_{(X_k,X_{k+1}\ldots,X_{k+n})}.$$

■

Definition: Invariantes Maß[3]

Gegeben seien ein abzählbarer Zustandsraum S und eine stochastische Matrix K. Ein Maß π auf $(S, \mathfrak{P}(S))$ heißt *invariant* (oder *stationär*) für K, falls

$$\pi K = \pi.$$

D.h., für alle $j \in S$ gilt

$$(\pi K)(j) := \sum_{i\in S} \pi(i)K(i,j) = \pi(j).$$

■

Bemerkung: Das Maß π ist ein Linkseigenvektor von K zum Eigenwert 1.

Mit Hilfe der obigen Definition kann nun folgende wichtige Proposition angegeben werden:

Proposition: *Sei S ein endlicher Zustandsraum und K eine irreduzible Übergangsmatrix. Dann gibt es genau ein invariantes Wahrscheinlichkeitsmaß π für K. Ist S abzählbar, so muss die zusätzliche Bedingung der Rekurrenz gefordert werden.*

Beweis: Siehe Gamerman (1997, S. 104) bzw. Waldmann und Stocker (2004, Satz 2.17).

□

[3] Der Begriff stationäres Maß ist ebenfalls gebräuchlich.

Anhang B

Allgemeine Beschreibung des KEI-Datensatzes

Der KEI Datensatz besteht aus 116 Indikatoren (mehrere ursprünglich ausgewählte Indikatoren mussten außer Acht gelassen werden, da sie bisher in keinem Land erhoben wurden), 4 Jahren und 29 Ländern bzw. Gruppen von Ländern (25 europäischen Ländern, Japan, US, EU_{15}, EU_{25}). Diese Länder stellen die verfügbare „Stichprobe“[1] für ein bestimmtes Jahr dar.

Den Dimensionen dieses Arrays ist sofort zu entnehmen, dass die maximale Anzahl an Beobachtungen pro Indikator für ein bestimmtes Land 4 beträgt. Dieser Wert wird nun als Referenzzahl genommen

B.1 Indikatoren

16 Indikatoren wurden vollständig beobachtet.

67 Indikatoren haben eine durchschnittliche Anzahl an Beobachtungen, die kleiner als 3 ist, d.h. im Durchschnitt wurden weniger als drei Werte pro Land erhoben.

Die durchschnittliche Anzahl an Beobachtungen pro Indikator ist 2,35.

28 Indikatoren haben im Durchschnitt weniger als einen beobachteten Wert pro Land.

[1] Die Betrachtung dieser Daten als Stichprobe stellt einen Kunstgriff dar, um die Anwendung von Imputationsmodellen zu ermöglichen.

B.2 Länder

Kein einziges Land hat die Daten vollständig erhoben.

Die höchste durchschnittliche Anzahl an Beobachtungen pro Indikator liegt bei 2,89 für Finnland. Das Schlusslicht bilden US, Japan und Malta mit jeweils 1,66, 1,6 und 1,5 Werten pro Indikator. Insgesamt 7 Länder haben im Durchschnitt weniger als 2 Beobachtungen pro Indikator geliefert.

Insgesamt wurden im Datensatz 42 % aller Werte nicht beobachtet, jedoch mit großen Unterschieden zwischen Indikatoren. Dies hat die Imputation deutlich erschwert.

Literaturverzeichnis

Auer, T. und Sturz, W. (2007). *ABC der Wissensgesellschaft*. Doculine Verlag.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B*, 36: 192–236.

Box, G. E. und Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26: 211–252.

Bremaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer.

Brooks, S. P. (1998). Markov Chain Monte Carlo and Its Application. *The Statistician*, 47(1): 69–100.

Cappe, O. und Robert, C. P. (2000). Markov Chain Monte Carlo: 10 Years and Still Running! *Journal of the American Statistical Association*, 95(452): 1282–1286.

Casella, G. und George, E. L. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3): 167–174.

Chatfield, C. (1997). *The Analysis of Time Series*. Chapman & Hall.

Chib, S. und Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4): 327–335.

Cochrane, J. H. (2005). *Asset Pricing*. Princeton University Press.

Dempster, A. P. (1987). The Calculation of Posterior Distributions by Data Augmentation. Comment. *Journal of the American Statistical Association*, 82: 541.

Dempster, A. P., Laird, N. M., und Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38.

Drossos, Constantine A. Philippou, A. (1980). A Note on Minimum Distance Estimates. *Annals of the Institute of Statistical Mathematics*, 32: 121–123.

Euractiv (2004). Die strategie von lissabon. In *http://www.euractiv.com/innovation/strategie-lissabon/article-103671.*

EUws (2005). European union web site. In *http://eur-lex.europa.eu*. The Commision's report.

Fahrmeir, L. und Tutz, G. (2001). *Multivariate Statistical Modeling based on Generalized Linear Models*. Springer.

Fishman, G. S. (2006). *A first Course in Monte Carlo*. Thomson.

Gamerman, D. (1997). *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference*. Chapman & Hall.

Gelfand, A. E. und Smith, A. F. M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410): 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., und Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall, second edition.

Geman, S. und Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 6: 71–741.

Geyer, C. (1992). Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4): 473–511.

Gilks, W. R., Richardson, S., und Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Gilks, W. R. und Wild, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, 41: 337–348.

Givens, G. H. und Hoeting, J. A. (2005). *Computational Statistics*. Wiley Interscience.

Goodnight, J. (1979). A Tutorial on the SWEEP Operator. *The American Statistician*, 33: 149–158.

Greene, W. H. (2003). *Econometric Analysis*. Prentice & Hall.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments. *Econometrica*, 50(4): 1029–1054.

Hartley, H. O. (1958). Maximum Likelihood Estimation from Incomplete Data. *Biometrics*, 14: 174–194.

Hastings, W. K. (1970). Monte Carlo Sampling Methods using Markov Chains and its Applications. *Biometrika*, 57: 97–109.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Irwin, J. O. (1963). The Place of Mathematics in Medical and Biological Statistics. *Journal of the Royal Statistical Society, Series B*, 126: 1–45.

KEI (2004). Knowledge Economy Indicators: Development of Innovative and Reliable Indicator Systems. In *http://kei.publicstatistics.net/*. Sixth Framework Programme of the European Commission.

Kent, K. T., Tyler, D. E., und Vardi, Y. (1994). A curious Likelihood identity for the multivariate t-distribution. *Commun. Statist. B-Simulation and Computation*, 23: 441–453.

Klaxton, C. (1999). The irrelevance of inference: A decision making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics*, 18: 341–364.

Königsberger, K. (2000). *Analysis 1.* Springer.

Lange, K. L., Little, R. J. A., und Taylor, J. M. G. (1989). Robust Statistical Modeling Using the t Distribution. *Journal of the American Statistical Association*, 84(408): 881–896.

Lehmann, I. und Schulz, W. (2004). *Mengen- Relationen-Funktionen.* Teubner.

Little, R. J. A. (1988). Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values. *Applied Statistics*, 3(1): 23–38.

Little, R. J. A. und Rubin, D. (2002). *Statistical Analysis with Missing Data.* Wiley Interscience.

Liu, C. (1995). Missing Data Imputation Using the Multivariate t Distribution. *Journal of Multivariate Analysis*, 53: 139–158.

Liu, C. und Rubin, D. (1995). ML Estimation of the t Distribution using EM and its Extensions, ECM and ECME. *Statistica Sinica*, 5: 19–39.

Liu, C., Rubin, D., und Wu, Y. N. (1998). Parameter Expansion to Accelerate EM: The PX-EM Algorithm. *Biometrika*, 85(4): 755–770.

Liu, J. S. und Rubin, D. B. (1994). The ECME algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence. *Biometrika*, 81(4): 633–648.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, 2: 49–55.

McKendrick, A. G. (1926). Applications of mathematics to medical problems. *Proc. Edinburgh Math. Society*, 44: 98–130.

McLachlan, G. J. und Peel, D. (2000). *Finite Mixture Models.* John Wiley.

Mendenhall, W., Scheaffer, R. L., und Ott, R. L. (2006). *Elementary Survey Sampling.* Thomson, sixth edition.

Meng, X.-L. (2000). Missing Data: Dial M for ?? *Journal of the American Statistical Association*, 95(452): 1325–1330.

Meng, X. L. und Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM Algorithm. *Journal of the American Statistical Association*, 86: 899–909.

Meng, X. L. und Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80: 267–278.

Merz, M. und Wüthrich, M. V. (2008). *Stochastic Claims Reserving Methods in Insurance.* John Wiley & Sons Ltd, Finance Series, Chichester.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, M. N., und Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21: 1087– 1092.

Mittelhammer, R. C. (1996). *Mathematical Statistics for Economics and Business.* Springer.

Nijman, T. und Verbeek, M. (1992). Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function. *Journal of Applied Econometrics*, 7(3): 243:257.

Pipkin, A. C. (1991). *A course on integral equations*. Springer.

Rall, L. (1969). *Computational Solution of Non-linear Operator Equations*. John Wiley.

Resnick, S. (2002). *Adventures in Stochastic Processes*.

Ritter, C. und Tanner, M. A. (1992). Facilitating the Gibbs Sampler: The Gibbs-Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, 87(419): 861–868.

Rizzo, M. L. (2008). *Statistical Computing with R*. Chapman & Hall/ CRC.

Roberts, G. O. (1996). *Markov Chain Concepts related to Sampling Algorithms in Markov Chain Monte Carlo in Practice*, chapter 3, S. 45–58. Chapman & Hall.

Roberts, G. O. und Smith, A. F. M. (1994). Simple conditions for the convergence of the Gibbs Sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49: 207–216.

Rodrigues, M. J. (2003). *European Policies for a Knowledge Economy*. Edward Elgar.

Rubin, D. (1978). Multiple imputation in sample surveys. *Proceedings of the Survey Research Methodological Section. American Statistical Society*, S. 20–34.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63: 581–592.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Schafer, J. L. (1999). Multiple Imputation: a primer. *Statistical Methods in Medical Research*, 8: 3–15.

Schafer, J. L. und Ghosh-Dastidar, B. (2006). Outlier Detection and Editing Procedures for Continuous Multivariate Data. *Journal of Official Statistics*, 22(3): 487–506.

Schafer, J. L. und Yucel, R. M. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values. *Journal of Computational and Graphical Statistics*, 11(2): 437–457.

Schaich, E. und Münnich, R. (2001). *Mathematische Statistik für Ökonomen*. Vahlen.

Schürle, J. (2004). *Zusammenführung von Daten auf Basis des Modells von Fellegi and Sunter*. Diss., Wirtschaftswissenschaftliche Fakultät Universität Tübingen.

Schumpeter, J. A. (1912). *Theorie der Wirtschaftlichen Entwicklung*. Drucker und Humblot. Nachdr. 2006.

Tanner, M. A. (1991). *Tools for Statistical Inference. Observed Data and Data Augmentation Methods*. Springer.

Tanner, M. A. und Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398): 528–540.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4): 1701–1728.

van Buuren, S. (2008). Multiple Imputation. In *http://www.multiple-imputation.com*.

van Dülmen, R. und Rauschenbach, S. H. (2004). *Macht des Wissens. Die Entstehung der modernen Wissensgesellschaft*.

Waldmann, K.-H. und Stocker, U. M. (2004). *Stochastische Modelle*.

Wolfowitz, J. (1957). The Minimum Distance Method. *The Annals of Mathematical Statistics*, 28: 75–88.