Behandlung fehlender Daten in der Faktorenanalyse

- Diplomarbeit -

zur Erlangung des Grades eines
Diplomstatistikers

Institut für Statistik
Fakultät für Mathematik, Informatik und Statistik
Ludwig-Maximilians-Universität München

Autor: Stefan Sauer
Betreuer: PD Dr. Christian Heumann

Februar 2010

Vorwort

Diese Diplomarbeit wurde am Statistischen Institut der Ludwig-Maximilians-Universität München geschrieben.
Sie beschäftigt sich mit dem Problem fehlender Daten in statistischen Methoden. Insbesondere werden die Methoden zur Behandlung unvollständiger Daten beschrieben und auf die Faktorenanalyse angewendet.

Die Arbeit wurde mit dem Textsatzprogramm LaTeX erstellt. Die Berechnungen und Grafiken wurden mit dem statistischen Programmpaket *R* (R Development Core Team, 2009) erzeugt und mit der Funktion *Sweave* (Leisch, 2002) in den LaTeX-Code eingebunden. Eine Übersicht über die wichtigsten verwendeten Funktionen findet sich in Anhang A. Der Programmiercode liegt der Arbeit als CD bei.
Bei Aussagen, die aus Gründen des Umfangs der Arbeit nicht bewiesen werden können, wird für deren Beweise auf die entsprechende Literatur verwiesen. Für das Verständnis wichtige Methoden werden im Anhang genauer erklärt.

Mein besonderer Dank gilt Herrn PD Dr. Christian Heumann, der mit mir zusammen das Thema dieser Arbeit ausgearbeitet und im weiteren Verlauf meine Arbeit hervorragend betreut hat. Stets war er mir sehr hilfreich mit zahlreichen wertvollen inhaltlichen und methodischen Anmerkungen und Vorschlägen verbunden mit Diskussionen, in denen er sein umfassendes statistisches Wissen mit mir teilte.

Außerdem möchte ich mich bei meiner Familie und im Speziellen bei meinen Eltern bedanken, die mich stets mit Rat und Tat und nicht zuletzt auch finanziell während meines Studiums unterstützt haben.
Darüber hinaus gilt mein Dank auch meinen Korrekturlesern Jens Königer, Annette Sauer und Ines Sauer.

Inhaltsverzeichnis

1. Einführung

Sehr oft werden Statistiker in der Praxis bei der Analyse von Datensätzen mit dem Auftreten von fehlenden Daten konfrontiert. Speziell in umfangreichem und komplexem multivariaten Datenmaterial treten diese aus verschiedenen möglichen Gründen häufig auf. Problematisch ist dies, da die Theorie der meisten statistischen Analysemethoden nur auf den Idealfall vollständiger Datensätze ausgerichtet ist. Wendet man diese auf fehlende Daten an, so kann dies zu ineffizienten oder verzerrten Schätzern führen.

Daher ist es in der angewandten Statistik notwendig, sich auf der einen Seite mit den Auswirkungen von fehlenden Daten auf die Analyseergebnisse zu beschäftigen sowie auf der anderen Seite Methoden zum Umgang mit unvollständigen Datasets anzuwenden.

Aus diesen Gründen setzt sich die statistische Forschung sehr intensiv mit fehlenden Daten auseinander. Hervorragende Lektüren zu diesem Thema existieren vor allem von Little/Rubin (2002) sowie von Schafer (1997).
In ihren Werken geben sie einen ausführlichen Überblick zum Auftreten und zu den Auswirkungen von fehlenden Daten in der angewandten Statistik, sowie besonders zum Umgang mit ihnen.

1.1. Fragestellung der Arbeit

Diese Arbeit beschäftigt sich hauptsächlich mit der Behandlung fehlender Daten in der Faktorenanalyse, deren Zweck es ist die Dimensionen multivariater Datensätze zu reduzieren, indem man die Korrelationsstruktur der Variablen dazu nutzt, diese in Zusammenhang mit einer kleineren Anzahl von Faktoren zu bringen.

Es werden verschiedene mögliche Ansätze wie Imputationsmethoden und likelihoodbasierte Methoden vorgestellt. Außerdem wird der Frage nachgegangen, wie sich die durch das Auftreten von fehlenden Daten bedingten zusätzlichen Unsicherheiten der Schätzungen in der Faktorenanalyse quantifizieren lassen.

In einer Sensitivitätsanalyse wird verglichen, wie die Schätzungen mit den vorgeschlagenen Methoden bei verschiedenen Anteilen von fehlenden Daten innerhalb eines Datensatzes voneinander abweichen und inwieweit sich die Wahl der Faktorenanzahl in der Analyse sowie die Unsicherheiten der Schätzungen bei größerem Anteil von nicht beobachteten Daten verändern.

1.2. Aufbau der Arbeit

Im Folgenden wird zunächst eine Einführung in das Problem fehlender Daten gegeben. Dies geschieht durch einen Einblick in deren Ursachen und Mechanismen, die zum Fehlen führen, sowie in grundsätzliche Möglichkeiten zur Behandlung des Problems (Kapitel 2).

Darauf folgend präsentiert Kapitel 3 die theoretischen Konzepte und die Anwendung der Faktorenanalyse. Ein Beispiel aus der Praxis dient dabei zu deren Veranschaulichung.

Anschließend werden die beiden wichtigsten Ansätze zum Umgang mit fehlenden Daten in der Faktorenanalyse, der EM-Algorithmus und die multiple Imputation, mit ihrer Umsetzung in der Praxis präsentiert (Kapitel 4).

Kapitel 5 beschäftigt sich darüber hinaus mit dem Vergleich der verschiedenen Verfahren bei unterschiedlichen Anteilen von fehlenden Daten. Mittels einer Sensitivitätsanalyse werden die Unsicherheiten dieser Methoden untersucht und deren Ergebnisse diskutiert.

Abschließend fasst Kapitel 6 mit einem Überblick über die Kernaussage der Arbeit deren wichtigste Punkte zusammen und gibt einen Ausblick auf mögliche Erweiterungen der Konzepte unter anderen Modellannahmen.

2. Fehlende Daten

2.1. Einführung in das Problem fehlender Daten

Die meisten statistischen Analysemethoden basieren auf der Voraussetzung, dass die zu analysierenden Daten vollständig sind. In der Praxis und vor allem bei multivariaten Datensätzen ist dieser Idealfall jedoch zumeist nicht gegeben, und mehrere Merkmalsausprägungen sind nicht beobachtet.

$$X = \begin{pmatrix} x_{11} & \dots & & \dots & x_{1k} \\ \vdots & \ddots & \star & & \vdots \\ \star & & & \star & \vdots \\ \vdots & & & \ddots & \star \\ x_{n1} & \dots & & \dots & x_{nk} \end{pmatrix}$$

Es liegt also nach Erhebung der Daten eine Datenmatrix X mit n statistischen Einheiten und k Variablen vor, die Löcher enthält, welche hier in Anlehnung an Toutenburg (1992) durch ein $\star$-Zeichen dargestellt werden. In Datensätzen werden die fehlenden Werte auch häufig mit der Abkürzung *NA* (Not Available) bezeichnet.
Das Auftreten solcher fehlender Beobachtungen kann verschiedenste Gründe haben (Beispiele aus Toutenburg/Heumann (2007) und Little/Rubin (1987)):

- Zensierungen in der Lebensdaueranalyse oder Drop-Out bei Langzeitstudien,
- Versagen von Messgeräten,
- Individuen vergessen Angaben oder weigern sich Fragen wie etwa nach Einkommen, Sexualverhalten oder sonstigen persönlichen Merkmalen zu beantworten,
- Mangelndes Wissen oder unzureichende Antwortmotivation des Befragten,
- Geplantes Fehlen bei Teilerhebungen in einer Population, wenn beispielsweise Folgefragen nur gestellt werden, falls zuvor eine andere Frage mit "ja" beantwortet wurde,
- Codierungs- und Übertragungsfehler,
- Schlechtes Design einer Umfrage: Zum Beispiel liefert die Frage nach dem Alter der Kinder bei kinderlosen Personen einen fehlenden Wert.

Um einige dieser Quellen für fehlende Daten zu vermeiden ist daher bereits im Vorfeld einer Datenerhebung eine eingehende Überprüfung des Untersuchungsdesigns ratsam, ebenso wie die Sicherstellung, dass der Kreis der befragten Personen über die notwendige Kompetenz und Motivation zur Beantwortung der Fragen verfügt.

2.2. Fehlendmechanismen

Zuerst stellt sich nun im Falle von unvollständigen Daten die Frage nach deren Zustandekommen. Man sollte herausfinden, welche der Ursachen für das Fehlen zutrifft, und ob die fehlenden Werte einem unsystematischen oder einem systematischen *Fehlendmechanismus* entsprechen. Das heißt, ob die Daten zufällig oder nicht zufällig fehlen. Dies ist wichtig, da nur auf Basis der Kenntnis des Ausfallmechanismus eine angemessene Behandlung des Problems durchgeführt werden kann.

2.2.1. Unsystematische Fehlendmechanismen

Der wesentlich weniger problematische Fall liegt vor, wenn Daten nicht systematisch fehlen, das heißt, dass der Grund, der zum Fehlen führt nicht mit einem Merkmal oder mit einer Variable direkt zusammenhängt. Ein typisches Beispiel ist das Vergessen der Beantwortung einzelner Fragen aufgrund von Unkonzentriertheiten. Dann kann der Mechanismus, der das Fehlen erzeugt, meist vernachlässigt werden, da er nur zu einer geringen oder gar keiner Verzerrung des Untersuchungsergebnisses führt.
Ein Anzeichen für unsystematische Fehlendmechanismen kann die Streuung der Datenlöcher über die Datenmatrix sein, das heißt wenn sich fehlende Werte nicht auf einen bestimmten Bereich der Matrix konzentrieren. Exakte Definitionen für solche Mechanismen, die im Folgenden dargelegt werden, gehen auf Rubin (1976) zurück.
Dafür definiert man zuerst eine Indikatormatrix für die fehlenden Daten,

$$M = (m_{ij}), \quad \text{mit } m_{ij} = \begin{cases} 1, \text{ falls } x_{ij} \text{ fehlend} \\ 0, \text{ falls } x_{ij} \text{ beobachtet.} \end{cases} \tag{2.1}$$

Diese Indikatorvariablen behandelt man nun als Zufallsvariablen und ordnet ihnen eine Verteilung zu. Der Fehlendmechanismus lässt sich durch die bedingte Verteilung von M charakterisieren:

$$f(M|X, \Phi) = f(M|X_{obs}, X_{mis}, \Phi).$$

Dabei steht Φ für die unbekannten Parameter und X_{obs}, X_{mis} sind die beobachteten bzw. die fehlenden Daten des Datensatzes X.

Missing at Random (MAR)

Von *missing at random* (MAR) spricht man, wenn das Fehlen eines Wertes nicht von der Ausprägung des Wertes selbst abhängt, das heißt, wenn nicht beispielsweise besonders große Werte oder negative Werte mit höherer Wahrscheinlichkeit aus der Stichprobe ausgeschlossen werden.
Dann lässt sich die bedingte Verteilung der Zufallsvariable M unabhängig von den fehlenden Daten X_{mis} darstellen:

$$f(M|X_{obs}, X_{mis}, \Phi) = f(M|X_{obs}, \Phi) \quad \text{für alle } X_{mis}\,, \Phi. \tag{2.2}$$

Ein Beispiel hierfür wäre, wenn bei einer Befragung nach dem Einkommen einige Personen die Antwort verweigern. Hängt dieses Weigern nicht mit der Höhe des Einkommens zusammen, so liegt MAR vor. Weigern sich aber hauptsächlich Großverdiener die Höhe ihres Einkommens anzugeben, so ist die MAR-Annahme verletzt, da man so das durchschnittliche Einkommen der Grundgesamtheit unterschätzen würde.

Missing Completely at Random (MCAR)

Hängt die Antwortrate nicht von der Ausprägung anderer Werte ab, so nennt man den Fehlendmechanismus *observed at random* (OAR). Liegt der OAR-Mechanismus zusätzlich zu MAR vor, so sind die fehlenden Werte *missing completely at random* (MCAR).
Das heißt, dass in diesem Fall die bedingte Verteilung von M sogar komplett unabhängig von den Daten X ist:

$$f(M|X_{obs}, X_{mis}, \Phi) = f(M|\Phi) \quad \text{für alle } X_{obs}\,, X_{mis}\,, \Phi. \tag{2.3}$$

Bei Vorliegen von MCAR ist der Fehlendmechanismus ignorierbar, da die beobachteten Daten eine Zufallsstichprobe aus allen Daten darstellen.
Wenn im zuvor beschriebenen Einkommensbeispiel die fehlenden Werte in der Variable Einkommen weder von der Höhe des Einkommens noch von sonstigen Faktoren abhängen, so liegt MCAR vor. Hängt aber beispielsweise Alter oder Geschlecht mit dem Verweigern der Angabe zusammen, so ist die MCAR-Annahme verletzt.
Eine Aufzählung der unsystematischen Fehlendmechanismen findet sich in Tabelle 2.1.

2.2.2. Systematische Fehlendmechanismen

Kann man die Verteilung der Zuvallsvariablen M_{ij} nicht wie in den Formeln 2.2 oder 2.3 vereinfachen, so liegt ein *nicht zufälliger* oder *unsystematischer* Fehlendmechanismus vor, der auch als *not missing at random* (NMAR) bezeichnet wird.

MAR	Missing at random Fehlen ist unabhängig von der Ausprägung des Merkmals selbst.
OAR	Observed at random Fehlen ist unabhängig von anderen Variablen.
MCAR	Missing completely at random MAR + OAR

Tabelle 2.1.: Unsystematische Fehlendmechanismen

Dieser macht die statistische Analyse der Daten wesentlich komplizierter, da eine Nichtberücksichtigung des Mechanismus zu einer erheblichen Verzerrung der Ergebnisse führen kann. Analysiert man etwa die Höhe des Einkommens anhand des Mittelwerts, so unterschätzt man das durchschnittliche Einkommen, wenn vor allem Großverdiener nicht geantwortet haben. Es is allerdings notwendig die Systematik des Ausfallmechanismus zu erkennen, denn ist der Mechanismus bekannt, so lassen sich Annahmen über das Modell treffen. Zum Auffinden bedient man sich der drei grundlegenden Methoden der Statistik:

- *Deskriptive Analyse* durch Kennzahlen, anhand derer man das Verhältnis von beobachteten zu fehlenden Daten untersucht und unter Umständen Konzentrationen fehlender Werte innerhalb der Datenmatrix aufdeckt.
- *Explorative Analyse* durch Zusammenhangsanalyse innerhalb der Datenmatrix, um eventuelle Abhängigkeiten der fehlenden Werte aufzudecken.
- *Induktive Analyse* durch statistische Tests, zum Beispiel auf Konzentration fehlender Werte innerhalb der Datenmatrix.

Zur genaueren Beschreibung dieser Methoden sei unter anderem auf Bankhofer (1995, Kap.3) verwiesen.

2.3. Verfahren zum Umgang mit unvollständigen Daten

Die meisten statistischen Analyseverfahren basieren auf vollständigem Datenmaterial und sind daher nicht mehr ohne weiteres anwendbar, sobald fehlende Daten auftreten. Deshalb ist es notwendig, sich Abhilfe durch neue Verfahren zu schaffen. Im Folgenden werden die am meisten verbreiteten Ansätze zur Behandlung fehlender Daten in der statistischen Datenanalyse

vorgestellt. Diese dienen als Möglichkeit die Matrix mit Datenlöchern in eine reguläre vollständige Datenmatrix zu überführen und werden häufig auch als Ad-Hoc-Verfahren bezeichnet.

2.3.1. Complete Case und Available Case Analysis

Die *Complete Case Analysis* berücksichtigt zur Auswertung der Daten nur diejenigen m Individuen, bei denen alle Merkmale beobachtet wurden. Alle $n - m$ nicht vollständigen Zeilen der Datenmatrix werden demnach gelöscht.

Beispiel:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} \\ x_{41} & x_{42} & \star & x_{44} & \star \\ x_{51} & \star & x_{53} & x_{54} & x_{55} \end{pmatrix} \longrightarrow \tilde{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} \end{pmatrix}$$

Somit ergibt sich für eine Variable j folgender Mittelwert und Varianz:

$$\bar{x}_{cca} = \frac{1}{m}\sum_{i=1}^{m} x_{ij} \quad , \quad s^2_{cca} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)^2.$$

Die Annahme, dass das Problem der fehlenden Daten durch simples Ausschließen der nicht vollständigen Beobachtungen gelöst ist, stellt sich aber nahezu immer als falsch heraus. Denn diese dank ihrer Einfachheit sehr beliebte Vorgehensweise kann offensichtlich nur sinnvoll sein, wenn lediglich wenige fehlende Daten vorliegen. Ansonsten führt sie zu einer ineffizienten Analyse, da man einen Großteil der vorhandenen Informationen verliert. Im Extremfall, wenn in fast allen Zeilen Werte fehlen (m nahe an 0), würden kaum noch Daten zur Analyse übrigbleiben.

Eine Alternative dazu stellt die *Available Case Analysis* dar, die nicht sämtliche Zeilen mit fehlenden Werten löscht, sondern von Fall zu Fall entscheidet, ob die Zeilen die für die Analyse interessierenden Beobachtungen enthalten. Dies kann allerdings in der Praxis zu Problemen mit der Vergleichbarkeit der auf diese Art berechneten Statistiken von Variablen führen, da die Anzahl der Beobachtungen für unterschiedliche Variablen meist verschieden groß ist.
Beide bisher beschriebenen Eliminierungsverfahren sind allerdings nur unter Annahme des MCAR-Fehlendmechanismus uneingeschränkt anwendbar.

2.3.2. Imputationsmethoden

Imputationsmethoden versuchen die Löcher in der Datenmatrix durch geeignete Werte aufzufüllen, um anschließend die vollständige Matrix standardmäßig wie einen vollständigen

Datensatz statistisch analysieren zu können. Zum Finden dieser geeigneten Werte gibt es wiederum verschiedene Ansätze:

Cold deck imputation: In diesem Ansatz ersetzt man fehlende Beobachtungen einer statistischen Einheit durch Werte, die nicht aus dem Datensatz gewonnen werden, sondern aus externen Quellen kommen. Dies können zum Beispiel Daten aus früheren statistischen Erhebungen oder Expertenwissen sein.

Hot deck imputation: In der hot deck imputation erlangt man die eingesetzten Werte aus dem selben Datensatz, indem man Werte ähnlicher Individuen hernimmt. Die *nearest neighbour hot deck imputation* nimmt dasjenige Individuum als nächsten Nachbarn, das den geringsten metrischen Abstand zur Einheit mit dem fehlenden Wert besitzt. So würde man beispielsweise bei einem Datensatz mit den Variablen Geschlecht, Körpergröße und Gewicht für den fehlenden Wert in der Variable Gewicht einer 1,60m großen Frau, das Gewicht einer anderen Frau mit der gleichen oder der am geringsten entfernten Körpergröße einfügen.

Mean imputation: Hier ersetzt man den fehlenden Eintrag in der Datenmatrix durch den Mittelwert in der entsprechenden Variable. Dadurch ändert sich der Mittelwert in der jeweiligen Variable nicht, es tritt jedoch das Problem der Unterschätzung der Varianz auf. Bei m beobachteten und $n-m$ fehlenden Werten in einer Variable ergibt sich die Varianz als

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\left[\sum_{i=1}^{m}(x_i-\bar{x})^2 + \underbrace{\sum_{i=m+1}^{n}(\bar{x}-\bar{x})^2}_{=0}\right] \\ &< \frac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})^2. \end{aligned}$$

Regression imputation: Eine weitere Möglichkeit erstellt auf Grundlage der vollständigen Beobachtungen ein Regressionsmodell zur Schätzung der fehlenden Werte. Dazu nutzt man die Korrelationstruktur zwischen den Variablen aus und bestimmt den fehlenden Wert durch die beobachteten Variablen.

Multiple Imputation: Eine Weiterentwicklung dieser *single imputation Verfahren*, welche sich, wie oben dargelegt, mit dem Problem auseinandersetzen müssen, dass die Variabilität der Imputation nicht berücksichtigt wird, stellen multiple Imputationsmethoden dar. Statt nur einen Wert einzusetzen generiert man mehrere mögliche Imputationen (z.B. $m = 5$). Somit erhält man m komplette Datensätze, die man dann separat standardmäßig analysieren und anschließend zu einem Gesamtergebnis kombinieren kann.
Die multiple Imputation wird später noch ausführlich in Kapitel *4* präsentiert.

Diese Verfahren werden zwar in der Praxis gerne genutzt, bergen aber auch ihre Risiken in sich. So geben sie einerseits die Illusion einer kompletten Stichprobe und verursachen auf der anderen Seite oftmals Verzerrungen etwa durch Nichtberücksichtigung des Fehlendmechanismuses. Zudem gilt für sämtliche Imputationsmethoden die Voraussetzung des MCAR-Fehlendmechanismus.

2.4. Likelihood-basierende Behandlung fehlender Daten

Während die Imputations- und Eliminierungsverfahren die unvollständige Datenmatrix in eine vollständige umwandeln, basiert ein weiterer Ansatz zum Umgang mit fehlenden Daten auf der Maximum-Likelihood-Theorie. Deren Ansatz ist es die Parameterwerte zu finden, die am besten durch die Daten gestützt werden.
Als Voraussetzung für die Anwendung von Maximum-Likelihood bei fehlenden Daten muss der MAR-Fehlendmechanismus gelten, das heißt, der Fehlendmechanismus muss ignorierbar sein.
Die Grundlagen der Maximum-Likelihood-Theorie werden ausführlich in Anhang B beschrieben.

2.4.1. Likelihood-Funktion für unvollständige Daten

Die gemeinsame Dichte der beobachteten und fehlenden Daten wird wiederum dargestellt als

$$f(X|\theta) = f(X_{obs}, X_{mis}|\theta).$$

Daraus bestimmt man die marginale Dichte der beobachteten Werte durch Herausintegrieren der fehlenden Werte.

$$f(X_{obs}|\theta) = \int f(X_{obs}, X_{mis}|\theta) dX_{mis} \,. \tag{2.4}$$

Dann ist die Likelihoodfunktion von θ eine beliebige Funktion, die sich proportional zu $f(X_{obs}|\theta)$ verhält.

$$L(\theta|X_{obs}) \propto f(X_{obs}|\theta). \tag{2.5}$$

Allgemeiner lässt sich in das Modell wieder eine Indikatorvariable für das Fehlen einer Beobachtung einfügen, mit

$$m_{ij} = \begin{cases} 1, \text{ falls } x_{ij} \text{ fehlend} \\ 0, \text{ falls } x_{ij} \text{ beobachtet.} \end{cases}$$

Diese wird als Zufallsvariable gesehen. Die gemeinsame Verteilung der Daten und des Fehlendmechanismus ergibt sich nun aus dem Produkt der Verteilung von X und der bedingten

Verteilung von M gegeben X, die den Parametervektor Ψ hat:

$$f(X, M|\theta, \Psi) = f(X|\theta) \cdot f(M|X, \Psi). \tag{2.6}$$

Wiederum integriert man die fehlenden Daten aus dieser Dichte heraus, um die Dichte der beobachteten Daten zu erhalten.

$$f(X_{obs}, M|\theta, \Psi) = \int f(X_{obs}, X_{mis}|\theta) \cdot f(M|X_{obs}, X_{mis}, \Psi) dX_{mis}. \tag{2.7}$$

Und die Likelihoodfunktion von θ und Ψ kann wieder durch jede dazu proportionale Funktion dargestellt werden:

$$L(\theta, \Psi|X_{obs}, M) \propto f(X_{obs}, M|\theta, \Psi). \tag{2.8}$$

Die Inferenz lässt sich nach Little/Rubin (2002, S.119) auf Basis der einfacheren Likelihood aus Gleichung *2.5* durchführen, wenn MAR als Fehlendmechanismus vorliegt und die Parameter θ und Ψ voneinander unabhängig sind.

2.4.2. Der EM-Algorithmus

Wie bei normaler Maximum-Likelihood-Inferenz gilt es auch hier das Maximum dieser Funktion zu finden. Dies ist gleichbedeutend mit dem Lösen der Likelihood-Gleichung

$$S(\theta|X_{obs}) = \frac{\partial lnL(\theta|X_{obs})}{\partial \theta} = 0. \tag{2.9}$$

Dies lässt sich bei komplizierten Gleichungen üblicherweise durch den Newton-Raphson-Algorithmus oder das Fisher-Scoring (siehe Anhang C) lösen.
Im Falle von fehlenden Daten sind diese Verfahren aber im Allgemeinen nicht mehr praktikabel, da die benötigte Fisher-Informationsmatrix I, die sich aus der 2. Ableitung der Log-Likelihoodfunktion ergibt, bzw. die erwartete Fisher-Informationsmatrix J eine zu komplizierte und rechenaufwendige Struktur erhalten.
Deshalb greift man im Falle unvollständiger Daten auf den *EM-Algorithmus* (Dempster, Laird & Rubin, 1977) zurück, für den keine zweiten Ableitungen notwendig sind. Die Abkürzung *EM* steht hier für *Expectation Maximization*.
Das Grundprinzip dieses iterativen Algorithmus ist es, zunächst die fehlenden Werte durch Schätzungen zu ersetzen, damit eine Parameterschätzung durchzuführen, auf Basis derer wiederum die fehlenden Beobachtungen neu geschätzt werden. Die fehlenden Werte und die Parameter werden also so lange neu geschätzt bis es zur Konvergenz kommt. Dies lässt sich nach Little/Rubin praktizieren, wenn die Log-Likelihoodfunktion linear in X_{mis} ist. In jedem

Schritt müssen die suffizienten Statistiken und auch die Likelihoodfunktion selbst neu bestimmt werden.

Der E-Schritt und der M-Schritt des Algorithmus

In jeder Iteration wird sowohl ein E-Schritt (Erwartungsschritt) als auch ein M-Schritt (Maximierungsschritt) durchgeführt.
Der M-Schritt maximiert die Likelihood als ob keine fehlenden Werte vorliegen würden mit Standardverfahren wie Newton-Raphson.
Der E-Schritt bildet den Erwartungswert für die fehlenden Werte bedingt auf die gegebenen X_{obs} und die geschätzten Parameter. Dadurch erhält man dann für die t-te Iteration die erwartete Log-Likelihoodfunktion

$$E\left(l(\theta \mid X)\right) \equiv Q(\theta|\theta^{(t)}) = \int l(\theta|X) \cdot f(X_{mis}|X_{obs}, \theta = \theta^{(t)}) dX_{mis}. \tag{2.10}$$

Darauf sucht der M-Schritt wiederum durch Maximieren dieser Funktion die resultierenden Parameter $\theta^{(t)}$ mit

$$\theta^{(t+1)} = \max_{\theta} Q(\theta|\theta^{(t)}). \tag{2.11}$$

Dabei gilt in jedem Schritt

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}), \quad \text{für alle } \theta. \tag{2.12}$$

EM-Algorithmus für Exponentialfamilien

Noch einfacher interpretieren lässt sich der EM-Algorithmus, wenn sich die Verteilung der Daten als reguläre *Exponentialfamilie*

$$f(X|\theta) = h(X) \exp\left(\frac{T(X) \cdot \theta}{b(\theta)}\right) \tag{2.13}$$

darstellen lässt. Hierbei ist $T(X)$ der Vektor der suffizienten Statistiken, und h und b stehen für Funktionen von X beziehungsweise vom Parametervektor θ. Dies gilt unter anderem für die Normalverteilung und auch viele andere häufig gebrauchte Verteilungen, wie Binomial- oder Gammaverteilung.
Im E-Schritt schätzt man dann in jeder Iteration die suffizienten Statistiken der vollständigen Daten durch

$$T^{(t+1)} = E(T(X)|X_{obs}, \theta^{(t)}). \tag{2.14}$$

Der M-Schritt ersetzt in den Likelihood-Gleichungen der vollständigen Daten die suffizienten Statistiken $T(X)$ durch die im E-Schritt geschätzten $T^{(t+1)}$. Dies vereinfacht die Bestimmung der geschätzten Parameter θ erheblich.
Im folgenden Beispiel wird die Anwendung des EM-Algorithmus mit E-Schritt und M-Schritt insbesondere bei einer vorhandenen Exponentialfamilie demonstriert.

Beispiel: *Univariat normalverteilte Daten* (aus Little/Rubin (2002, S.168))
Die Daten x_i seien unabhängig und identisch verteilt $N(\mu, \sigma^2)$. Davon seien die Werte x_i mit $i = 1, ..., r$ beobachtet und $i = r+1, ..., n$ fehlend mit ignorierbarem Fehlendmechanismus. Der Erwartungswert für jeden fehlenden Wert x_i für gegebene beobachtete Werte und Parametervektor $\theta = (\mu, \sigma^2)$ ist μ. Die Log-Likelihood basierend auf allen Daten stellt sich dar als

$$l(\mu, \sigma^2|X) = -\frac{n}{2} ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2},$$

mit den suffizienten Statistiken $\sum_{i=1}^{n} x_i$ und $\sum_{i=1}^{n} x_i^2$.
Durch den E-Schritt folgen für die Parameterschätzer $\theta^{(t)} = (\mu^{(t)}, \sigma^{(t)})$im t-ten Schritt

$$E(\sum_{i=1}^{n} x_i|\theta^{(t)}, X_{obs}) = \sum_{i=1}^{r} x_i + (n-r)\mu^{(t)} \tag{2.15}$$

$$E(\sum_{i=1}^{n} x_i^2|\theta^{(t)}, X_{obs}) = \sum_{i=1}^{r} x_i^2 + (n-r)\left[(\mu^{(t)})^2 + (\sigma^{(t)})^2\right]. \tag{2.16}$$

Der M-Schritt benutzt jetzt die üblichen ML-Schätzer für den Erwartungswert $\hat{\mu} = \sum_{i=1}^{n} \frac{x_i}{n}$ sowie für die Varianz $\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{x_i^2}{n} - (\sum_{i=1}^{n} \frac{x_i^2}{n})^2$, um die neuen Parameterschätzer zu finden:

$$\mu^{(t+1)} = E\left(\sum_{i=1}^{n} x_i|\theta^{(t)}, X_{obs}\right)/n \tag{2.17}$$

$$\left(\sigma^{(t+1)}\right)^2 = E\left(\sum_{i=1}^{n} x_i^2|\theta^{(t)}, X_{obs}\right)/n - \left(\mu^{(t+1)}\right)^2. \tag{2.18}$$

Wie leicht zu sehen ist, konvergiert der Algorithmus in

$$\hat{\mu} = \sum_{i=1}^{r} \frac{x_i}{r} \quad und \quad \hat{\sigma}^2 = \sum_{i=1}^{r} \frac{x_i^2}{r} - \hat{\mu}^2. \tag{2.19}$$

In diesem Beispiel hätte man die Parameter natürlich auch direkt schätzen können, und der EM-Algorithmus wäre nicht notwendig gewesen. □

Beobachtete Information und Konvergenzverhalten

Der Vorteil des EM-Algorithmus ist, dass er meist einfach zu konstruieren ist und dass sich jeder Schritt leicht interpretieren lässt, da jede Iteration die Log-Likelihood vergrößert, die dann zu einem stationären Wert konvergiert.

Als Nachteil gilt dagegen, dass im Algorithmus nicht die Unsicherheit der Schätzung berücksichtigt wird. Um diese zu berechnen, benötigt man die beobachtete Fisher-Informationsmatrix $I(\theta \,|\, X_{obs})$, welche der negativen 2. Ableitung der Log-Likelihood entspricht. Diese ist aber im Allgemeinen durch direktes Ableiten analytisch nur schwer zu berechnen.
Eine Art, den Standardfehler für die Schätzung mit dem EM-Algorithmus zu berechnen, ist die *Formel von Louis* (Louis 1982). Dabei benutzt man, dass die beobachtete Information gleich der kompletten Information abzüglich der fehlenden Information ist. Dies ergibt sich nach Little/Rubin (2002, S.172) aus der Aufspaltung der Log-Likelihood in

$$l(\theta \,|\, X_{obs}) \;=\; l(\theta \,|\, X) \;-\; ln\, f(X_{mis} \,|\, X_{obs}, \theta)) \,. \tag{2.20}$$

Bildet man für beide Seiten den Erwartungswert, so erhält man mit der negativen 2. Ableitung die Beziehung

$$\underbrace{I(\theta \,|\, X_{obs})}_{\text{beobachtete Information}} = \underbrace{E\left[I(\theta \,|\, X) | \, X_{obs}, \theta\right]}_{\text{gesamte Information}} - \underbrace{\left[-\partial^2 ln\, f(X_{mis} \,|\, X_{obs}, \theta)/\partial\theta\partial\theta\right]}_{\text{fehlende Information}} \,. \tag{2.21}$$

Louis bewies in seiner Formel, dass sich bei $\theta = \hat{\theta}_{ML}$ die beobachtete Information durch folgende Gleichung berechnen lässt:

$$I(\theta \,|\, X_{obs}) \;=\; E\left[I(\theta \,|\, X) \,|\, X_{obs}, \theta\right] \;-\; E\left[S(\theta \,|\, X) S(\theta \,|\, X)^T | \, X_{obs}, \theta\right] \,. \tag{2.22}$$

Damit ist kein explizites Berechnen der beobachteten Informationsmatrix durch zweimaliges Ableiten mehr notwendig.

Ein weiteres Manko des EM-Algorithmus ist die unter Umständen sehr langsame Konvergenz. Während Verfahren wie Newton-Raphson im Allgemeinen eine quadratische Konvergenzgeschwindigkeit haben, ist die des EM-Algorithmus nur linear. Je höher der Anteil der fehlenden Daten ist desto länger braucht der Algorithmus, um zum Ende zu kommen. Des Weiteren besteht zudem die Gefahr einer möglichen Konvergenz in einem lokalen Maximum.
Die Konvergenzgeschwindigkeit ist eng verbunden mit den zuvor beschriebenen Informationen. Ist bei skalarem θ nämlich der aktuelle Schätzer $\theta^{(t)}$ nahe am Konvergenzwert θ^*, so gilt die Beziehung

$$|\theta^{(t+1)} - \theta^*| = \lambda |\theta^{(t)} - \theta^*| \,, \tag{2.23}$$

wobei λ den Anteil der fehlenden Information darstellt, $\lambda = \frac{\text{fehlende Information}}{\text{gesamte Information}}$.

Maximum-Likelihood-basierte Behandlung fehlender Daten wird später noch ausführlicher in Kapitel 4 für multivariat normalverteilte Daten und ignorierbaren Fehlendmechanismus MAR beschrieben.
Insbesondere wird für diesen Fall die Anwendung des EM-Algorithmus für die Faktorenanalyse erläutert.

3. Faktorenanalyse

In diesem Kapitel wird eine ausführliche Beschreibung der Zielsetzung und Durchführung der Faktorenanalyse geliefert. Ihre Anwendung wird in Abschnitt 3.7 anhand eines Beispiels aus der Praxis demonstriert.

3.1. Ausgangspunkt und Zielsetzung

Die *Faktorenanalyse* ist ein Gebiet der multivariaten Statistik, das zu den dimensionsreduzierenden Methoden zählt.
Ursprünglich wurde sie anfangs des 20. Jahrhunderts hauptsächlich von Psychologen wie Spearman und Thompson entwickelt und war vor allem zur Untersuchung der Korrelations- oder Kovarianzmatrizen bei kognitiven Tests bestimmt. Nach einigen Weiterentwicklungen der Theorie, unter anderem durch Thurstone (1935), Holzinger (1941) und Lawley seit Anfang der 40er Jahre, hat sich die Faktorenanalyse heute zu einer der meistangewandten Verfahren der multivariaten Statistik entwickelt.
Ausgangspunkt der Faktorenanalyse ist ein Datensatz X mit n Individuen und p Variablen, welche in der Regel nicht unabhängig, sondern korreliert sind. Ziel ist es nun latente Größen zu finden, "*to discover [...] those principal factors that are truly operative in producing the correlation coefficients*", wie es Thurstone (1931, S.421) formulierte. Man will also multivariate Zusammenhänge zwischen manifesten beobachtbaren Variablen durch eine geringere Anzahl latenter Variablen erklären. Dabei nimmt man an, dass die nicht beobachtbaren Faktoren für die Korrelationen zwischen den beobachteten Variablen verantwortlich sind.
Die beobachteten Variablen dienen also nur als Indikatoren für die Faktoren, die von Interesse sind, wie in folgenden Beispielen:

- Zehnkampf: Punkte für 10 Disziplinen → Eigenschaften: Ausdauer, Kraft, Technik, Dynamik
- Gesundheitliche Kennzahlen (Blutdruck, Körpertemperatur, Hautausschläge,...) → Gesundheitszustand von Patienten
- Schulnoten in Mathematik, Deutsch, ... → Intelligenz der Schüler

Ziel der Faktorenanalyse ist es, die Faktoren so zu bestimmen, dass sie gleichzeitig auch noch möglichst gute Interpretierbarkeit gewährleisten.

Grundsätzlich existieren zwei Herangehensweisen an die Faktorenanalyse: Zum einen die *konfirmatorische*, in der man versucht den Nachweis von vermuteten Faktoren zu erbringen und auf deren Grundlage die Parameter des Modells zu bestimmen. Zum anderen die *explorative*, in der man anhand der vorliegenden Daten mögliche Faktoren extrahieren will, die die Korrelationsstruktur zwischen den Variablen möglichst genau darstellen. Hierbei ist anfangs nichts über die genaue Art und Anzahl der Faktoren bekannt. Die konstruierten Faktoren müssen dann im Anschluss noch inhaltlich gedeutet werden.

In den folgenden Abschnitten wird das Modell der Faktorenanalyse aufgestellt und anschließend beschrieben, wie die Faktoren ausgehend von der empirischen Kovarianzmatrix geschätzt und interpretiert werden können.

3.2. Das lineare Faktorenmodell

3.2.1. Grundgleichung

Das klassische lineare Modell der Faktorenanalyse hat die Form

$$x = Lf + u \quad bzw. \quad X = LF + U\,. \tag{3.1}$$

Dabei ist $x = (x_1, ..., x_p)^T$ der Vektor der beobachteten Variablen, $f = (f_1, ..., f_k)^T$ der Vektor der *latenten Faktoren* und $u = (u_1, ..., u_k)^T$ der Vektor, der die latenten *spezifischen Faktoren* sowie Messfehler enthält. Die beobachteten Variablen sind also nicht vollständig durch die Faktoren determiniert, sondern sind auch noch anderen Einflüssen ausgesetzt. Die Parameter dieser linearen Funktion werden als Ladungen bezeichnet und $L = (l_{ir}) = (p \times k)$ ist die sogenannte *Ladungsmatrix*. Hierbei ist zu beachten, dass Gleichung 3.1 zwar die Form einer Regressionsgleichung hat, jedoch keine ist, da die Faktorenwerte nicht beobachtbar sind, sondern im Anschluss erst geschätzt werden müssen.

Es können verschiedene Arten von Faktoren unterschieden werden:
Sind in der Ladungsmatrix alle Ladungen $l_{1k}, ..., l_{pk}$ des k-ten Faktors F_k deutlich von 0 verschieden, so spricht man von einem allgemeinen Faktor. Weichen mindestens 2 seiner Ladungen deutlich von 0 ab, so nennt man ihn gemeinsamer Faktor. Allgemeine Faktoren sind somit Spezialfälle von gemeinsamen Faktoren. Faktoren, bei denen sich nur eine Ladung von 0 abhebt, werden als merkmalseigene Faktoren bezeichnet.
Tabelle 3.1 fasst nochmals alle Komponenten dieser Gleichung und ihre Bedeutungen sowie Dimensionen zusammen.

Symbol	Bedeutung	Dimension	Anmerkung
q	Anzahl der Faktoren	skalar	$q < p$
f_k	Faktorenwerte aller Beobachtungen für Faktor k	Vektor der Länge n	
F	Matrix der Faktorenwerte $F = (f_1 ... f_q)$	$n \times q - Matrix$	Faktorenwerte aller Beobachtungen
f_{ik}	Eintrag aus Matrix F	skalar	Faktorenwert von Beobachtung i bzgl Faktor k
k	Index der Faktoren	$k = 1, ..., q$	
L	Ladungsmatrix	$p \times q - Matrix$	
l_{jk}	Ladungskoeffizient von Variable j für Faktor k	skalar	
u_j	Fehlerterm	Vektor der Länge n	Restfehler für Variable j
U	Matrix der Fehlerterme $U = (u_1, ..., u_p)$	$n \times p - Matrix$	
u_{ij}	Eintrag aus Matrix U	skalar	Fehlerterm von Beobachtung i in der Variable j

Tabelle 3.1.: Notation für die Komponenten der Faktorenanalyse. Alle Elemente dieser Tabelle sind unbekannt und müssen geschätzt werden.

3.2.2. Grundannahmen und Fundamentaltheorem

Da die Ergebnisse einer Faktorenanalyse skalierungsunabhängig sind, wird die Datenmatrix $x_1, ..., x_p$ meist zuerst standardisiert, so dass $E(x_i) = 0$ und $Var(x_i) = 1$ für $i = 1, ..., p$ gelten.
Des Weiteren werden alle gemeinsamen und spezifischen Faktoren als unkorreliert

$$E(fu') = 0\,, \tag{3.2}$$

sowie die gemeinsamen Faktoren als standardisiert

$$E(F) = 0 \quad , \quad Cov(F) = I_k \tag{3.3}$$

vorausgesetzt.
Für die spezifischen Faktoren wird ebenfalls ein Erwartungswert von 0 angenommen und

außerdem, dass sie untereinander unkorreliert sind, aber beliebige Varianzen besitzen.

$$E(U) = 0 \quad , \quad Cov(U) = E(UU') = \Psi = diag(\Psi_1, ..., \Psi_p). \tag{3.4}$$

Mit den getroffenen Annahmen folgt für die Kovarianzmatrix Σ von x:

$$\begin{aligned} \Sigma &= E(xx') = E[(Lf + u)(Lf + u)'] \\ &= E(Lff'L') + E(Lfu') + E(uf'L') + E(uu') \\ &= LL' + 0 + 0 + \Psi = \\ \Sigma &= LL' + \Psi. \end{aligned} \tag{3.5}$$

Gleichung 3.5 wird als *Fundamentaltheorem der Faktorenanalyse* bezeichnet.
Die *Kommunalitäten* h_i^2, mit

$$h_i^2 := l_{i1}^2 + ... + l_{ik}^2$$

sind die Varianzen, die die Variable i über die gemeinsamen Faktoren mit den anderen Variablen teilt, das bedeutet, der Teil ihrer Varianz, der durch das Faktorenmodell beschreiben wird.
Die Komponente Ψ_i ist dagegen die Spezifität, das heißt, der Teil der Varianz, der nur aus der Variable stammt und nicht durch die Faktoren beschrieben wird.
Die Faktorenanalyse zerlegt also die Varianz der Stichprobe in einen allgemeinen und eine spezifischen Teil.
Je besser die Anpassung des Faktorenmodells ist, desto kleiner sind die Einträge in Ψ . Für die Schätzung der Parameter des Modells ist es wichtig, dass die Anzahl der Faktoren kleiner ist als die Anzahl der Variablen, da man ansonsten $\Sigma = L'L + 0$ betrachten würde.

3.2.3. Identifizierbarkeit der Parameter

Für $k = 1$ reduziert sich die Ladungsmatrix L zu einem p-dimensionalen Spaltenvektor. Wegen der Annahme der Standardisiertheit ist dieser Vektor bis auf das Vorzeichen eindeutig bestimmt.
Im Fall $k = 1$ ist das Faktorenmodell also eindeutig identifizierbar.
Im Fall $k > 1$ ist die Zerlegung allerdings nicht eindeutig.
Ersetzt man L durch $\tilde{L} = LR$, wobei $R = (k \times k)$ eine orthogonale Transformationsmatrix (*Rotationsmatrix*) ist, so ergibt sich

$$\tilde{L}\tilde{L}^T = L \underbrace{RR^T}_{=I} L^T = LL^T.$$

Somit verändert sich Gleichung 3.2 nicht und das Faktorenmodell 3.1 bleibt ebenfalls gleich, da $\tilde{f} = R^T f$ wiederum ein Vektor mit standardisierten unkorrelierten gemeinsamen Faktoren ist.

Man erhält demzufolge erneut

$$\tilde{L}\tilde{f} = L\underbrace{RR^T}_{=I}f = Lf\,.$$

Die Diagonalelemente von LL^T bleiben also bei jeglicher orthogonaler Transformation (*Rotation*) der Ladungsmatrix unverändert. Da sowohl L als auch $\tilde{L}$ gültige Lösungen sind, existiert eine unendliche Anzahl von Ladungsmatrizen.

Die Faktorenanalyse wird deshalb gewöhnlich in 4 Schritten durchgeführt. Im ersten werden mögliche Ladungen berechnet, die die geschätzten Varianzen und Kovarianzen ergeben, welche die beobachteten Variablen fitten.
Im Anschluss legt man anhand verschiedener Kriterien die Anzahl q der Faktoren im Modell fest.
Da die Ladungen unter Umständen nur sehr schwer interpretierbar sind oder nicht den Erwartungen entsprechen, werden sie in einem dritten Schritt rotiert, um zu einer Ladungsmatrix zu gelangen, die die Variablen genauso gut fittet, aber leichter zu interpretieren ist oder näher an den Erwartungen liegen.
Die einfachste Struktur von $\tilde{L}$ wäre gegeben, wenn in jeder Reihe nur eine Ladung von 0 verschieden wäre. Das würde bedeuten, dass jede Variable x_i nur durch einen gemeinsamen Faktor beeinflusst wird, und die Teilmenge der Variablen, auf die sich ein Faktor f_r mit$r = 1, ..., k$ auswirkt, würde eine natürliche Charakterisierung dieses Faktors erlauben.
Abschließend muss man dann nur noch die Faktorenwerte $f = (f_1, ..., f_k)'$ schätzen.
Im Folgenden werden nun zwei Methoden zur Berechnung einer Ladungsmatrix mit der jeweiligen Bestimmung der Faktorenzahl und danach die gängigsten Rotationsverfahren dargelegt.

3.3. Schätzung der Ladungsmatrix

Die beiden gängigsten Methoden zur Bestimmung der Ladungsmatrix sind zum einen die Maximum-Likelihood-Faktorenanalyse nach Lawley und Maxwell (1971), sowie zum anderen die Hauptkomponentenmethode, die hauptsächlich auf Hotelling (1936) zurückzuführen ist.

3.3.1. ML-Faktorenanalyse

In der auf der Maximum-Likelihood-Theorie (siehe Anhang C) basierenden Schätzung der Ladungsmatrix werden ausgehend von Gleichung 3.5 die Ladungsmatrix L mit $p \cdot q$ unbekannten Parametern sowie die p Diagonalelemente der spezifischen Varianz Ψ durch ML-Schätzer geschätzt. Außerdem lässt sich nach deren Schätzung mit Hilfe eines Likelihood-Quotiententests die passende Anzahl k der Faktorenwerte bestimmen.

Als Voraussetzung wird angenommen, dass es sich um (standartisierte) normalverteilte Daten handelt, mit

$$x \sim N(0, \Sigma) \quad , \quad f \sim N(0, I_k) \quad und \quad u \sim N(0, \Psi), \tag{3.6}$$

sowie zusätzlich zur Gewährleistung der Eindeutigkeit

$$L^T \Psi^{-1} L \qquad \text{diagonal.} \tag{3.7}$$

Im weiteren Verlauf berechnet man dann wie gewohnt durch Aufstellung und Maximierung der Likelihoodfunktion die Maximum-Likelihood-Schätzer.

Schätzung von L und Ψ

Als erwartungstreuen Schätzer für die Kovarianzmatrix Σ verwendet man die empirische Kovarianzmatrix S, mit

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'. \tag{3.8}$$

Diese besitzt nach Fahrmeir u.a. (1996, S.51) eine Wishart-Verteilung, $(n-1)S \sim W_p(\Sigma, n-1)$ mit Dichte

$$L(\Sigma|S) = const \cdot |S|^{\frac{n-p-2}{2}} |\Sigma|^{-\frac{n-1}{2}} exp\{-(\frac{n-1}{2}) tr(S\Sigma^{-1})\}. \tag{3.9}$$

Die gesuchten Schätzer $\hat{L}$ für L und $\hat{\Psi}_1, ..., \hat{\Psi}_p$ für Ψ erhält man nun durch Maximierung der logarithmierten Likelihoodfunktion von S:

$$l(\Sigma|S) = l(L, \Psi|S) = const + (\frac{n-p-2}{2}) \cdot ln|S| - (\frac{n-1}{2}) \cdot ln|\Sigma| - (\frac{n-1}{2}) \cdot tr(S\Sigma^{-1}). \tag{3.10}$$

Da es sich bei den ersten beiden Summanden um Konstanten handelt ist dies äquivalent zu einer Minimierung der Funktion

$$f(L, \Psi) = ln|\Sigma| + tr(S\Sigma^{-1}). \tag{3.11}$$

Durch Nullsetzen der partiellen Ableitungen dieser Funktion nach L und Ψ erhält man nach Lawley (1971, S.26) folgende Gleichungen zur Bestimung von $\hat{L}$ und $\hat{\Psi}$:

$$\begin{aligned} &1) \quad \hat{L} = S\hat{\Sigma}^{-1}\hat{L} \qquad \text{mit } \hat{\Sigma} = \hat{L}\hat{L}' + \hat{\Psi} \\ &2) \quad diag(\hat{\Sigma}) = diag(S) \qquad mit\ \hat{\Sigma} = \hat{L}\hat{L}' + \hat{\Psi} \\ &3) \quad \hat{L}\hat{\Psi}^{-1}\hat{L}' \quad \text{diagonal.} \end{aligned} \tag{3.12}$$

Diese Gleichungen sind nach Fahrmeir u.a. (1996, S.650) mit Wahrscheinlichkeit 1 eindeutig nach $\hat{L}$ und $\hat{\Psi}$ auflösbar, falls $\hat{\Sigma}$ vollen Rang p besitzt. $\hat{L}$ und $\hat{\Psi}$ sind dann unter der Normalverteilungsannahme aus Gleichung 3.6 konsistente und asymptotisch normalverteilte Schätzer für L und Ψ.
Ist hingegen die Normalverteilungsannahme für x verletzt, so spricht man bei den Schätzern aus dieser Methode von Quasi-Maximum-Likelihood-Schätzern, die aber, sofern die ersten beiden Momente von x existieren, weiterhin konsistent und asymptotisch normal sind.

Eine explizite ML-Schätzung lässt sich in der Praxis jedoch nicht angeben, vielmehr müssen die Werte mit einem iterativen Algorithmus zur Maximierung der Likelihood geschätzt werden. Nach Festlegung von Startwerten werden abwechselnd die folgenden beiden Schritte ausgeführt und die Ergebnisse des einen als neue Startwerte für den anderen eingesetzt, bis ein Konvergenzkriterium erfüllt ist:

1. Für einen festen Wert Ψ_0 wird die Gleichung $f(L, \Psi = \Psi_0)$ minimiert. Die Lösung hierfür gewinnt man analytisch durch ein Eigenwertproblem. Dies ergibt den neuen Startwert L_0 für $\hat{L}$.
2. Für festes L_0 wird durch numerische Verfahren, wie etwa Newton-Raphson, die Funktion $f(L = L_0, \Psi)$ minimiert. Dies ergibt wiederum den neuen Startwert Ψ_0 für $\hat{\Psi}$.

Zur genaueren Beschreibung der numerischen Verfahren und des zu lösenden Eigenwertproblems siehe Fahrmeir u.a. (1996, S.651) oder Anderson (2003, S.580).
Die Konvergenz dieses Algorithmus muss nicht zwingend gewährleistet sein. Im Falle der Nichtkonvergenz müssen die Startwerte variiert werden.

Likelihood-Quotiententest zur Bestimmung der Faktorenzahl

Der Hauptvorteil der Maximum-Likelihood-Methode unter Normalverteilungsannahme ist die Möglichkeit mittels eines Likelihood-Quotiententests zu überprüfen, ob die Daten dem Faktorenmodell mit der gewählten Anzahl von q Faktoren entsprechen. Dies entspricht den Hypothesen

$$\begin{aligned} H_0: &\quad \Sigma = LL' + \Psi \quad L \in \mathbb{R}^{p\times q}, \quad \Psi > 0 \\ H_1: &\quad \Sigma \in \mathbb{R}^{p\times p} \quad \text{beliebig positiv definit}, \quad (H_0 \subset H_1). \end{aligned} \tag{3.13}$$

Der Test basiert auf dem Verhältnis L_0/L_1 der Likelihoodfunktion L_0 (Gleichung 3.10 unter H_0 mit den eingesetzten ML-Schätzern $\hat{\Sigma} = \hat{L}\hat{L}' + \hat{\Psi}$) zur ursprünglichen Likelihood L_1 unter H_1.

Teststatistik U_q ist das (-2)-fache des Logarithmus dieses Verhältnisses, was sich darstellt als

$$\begin{aligned} U_q &= -2 \cdot ln(\frac{L_0}{L_1}) = (n-1)[ln|\hat{\Sigma}| - ln|S| + \underbrace{tr|S\hat{\Sigma}^{-1}|}_{=p} - p] \\ &= (n-1)[ln|\hat{\Sigma}| - ln|S|]. \end{aligned} \tag{3.14}$$

Diese Prüfgröße besitzt eine asymptotische Verteilung $U_q \overset{a}{\sim} \chi^2(d_q)$. Die Anzahl d_q der Freiheitsgrade dieser Verteilung entspricht den freien Parametern im Fundamentaltheorem.

$$d_q = [\frac{p(p+1)}{2}] - [\frac{(pq + p - q(q-1)}{2}] = \frac{1}{2}[(p-q)^2 - (p+q)].$$

Das Modell mit q Faktoren wird also bei einem zuvor festgelegten Signifikanzniveau α abgelehnt, falls

$$U_q > \chi^2(d_q; 1-\alpha).$$

Diese Vorgehensweise entspricht dem Test, ob das Modell mit einer zuvor bestimmten Faktorenanzahl q die Daten beschreibt. Denkbar ist jedoch auch, dass diese Anzahl erst noch festgelegt werden muss. In diesem Fall benutzt man den gleichen Test, startend mit einer kleinen Faktorenanzahl q_0, zum Beispiel $q_0 = 1$. Lehnt der Test dieses Modell ab, so erhöht man die Anzahl sukzessive um jeweils *1*, bis nicht mehr abgelehnt wird. Auf diese Weise erhält man die passende Anzahl der Faktoren. Allerdings muss auch in jedem Schritt erst die komplette Maximum-Likelihhod-Schätzmethode für die Parameter L und Ψ durchgeführt werden.

3.3.2. Hauptkomponentenmethode

Die Hauptkomponentenmethode bietet eine weitere Möglichkeit eine Ladungsmatrix zu bestimmen. Sie basiert auf der Hauptkomponentenanalyse (*Principal Component Analysis*), die auf Pearson (1901) und Hotelling (1936) zurückgeht.
Bei der Hauptkomponentenmethode werden nicht wie bei der Maximum-Likelihood-Faktorenanalyse Verteilungsannahmen gestellt, sondern es wird versucht, die Daten direkt durch Komponenten möglichst gut anzunähern. Diese Komponenten sind Linearkombinationen der ursprünglichen Variablen. Mit möglichst wenigen solcher Komponenten, die in diesem Fall die Faktoren darstellen, soll einen möglichst großen Teil der Varianz erfasst werden. Die Eindeutigkeit der Lösung der Faktorenschätzung mit dieser Methode wird gewährleistet durch die Nebenbedingung, dass die Faktoren sukzessive weniger Varianz erklären. Dies erreicht man dadurch, dass man die Eigenwerte ihrer Größe nach ordnet.
Im Folgenden wird nun zunächst in einem kleinen Exkurs die Hauptkomponentenanalyse vorgestellt.

Hauptkomponentenanalyse

Ebenso wie die Faktorenanalyse ist die Hauptkomponentenanalyse (PCA) ein lineares Modell zwischen den Komponenten und Variablen. Sie dient ebenfalls zur Dimensionsreduktion und führt auch oft zu ähnlichen Ergebnissen.
Die PCA unterscheidet sich in einem Punkt wesentlich von den Methoden der Faktorenanalyse. Versucht man bei der Faktorenanalyse von vornherein q orthogonale Faktoren $F_1, ..., F_q$ zu extrahieren und die Varianz in einen gemeinsamen und einen spezifischen Teil zu zerlegen, werden bei der PCA zunächst p Komponenten konstruiert, aus denen dann q ausgewählt werden, die den größten Teil der Gesamtvarianz der ursprünglichen Daten auf sich vereinigen. Da mit p Komponenten die ursprünglichen Variablen komplett erfasst werden, benötigt man keine Fehlerterme. Diese kommen erst durch Weglassen von Komponenten hinzu. Die Hauptkomponentenanalyse konzentriert sich hauptsächlich auf die Varianzen der beobachteten Variablen, während bei der Faktorenanalyse die Kovarianzen die größere Bedeutung haben.
Ziel der PCA ist es in der Regel nicht, interpretierbare Faktoren zu konstruieren. Deshalb sieht man auch meist von der Rotation des Ergebnisses ab, und beschränkt sich darauf komplizierte Beziehungen in beobachteten Daten auf eine einfache Form zu reduzieren.
Die erste Komponente z_1 ergibt sich aus der Linearkombination

$$z_1 = a_{11}x_1 + a_{21}x_2 + ... + a_{p1}x_p = Xa_1, \tag{3.15}$$

$$mit \quad a_1 = (a_{11}, a_{21}, ..., a_{p1})^T \quad \text{unter der Nebenbedingung } \; a_1^T a_1 = 1$$

Die erste Komponente wird so bestimmt, dass die Varianz von z_1 maximal wird. Dabei gilt:

$$Var(z_1) = Var(Xa_1) = a_1^T \Sigma a_1 =: \lambda_1. \tag{3.16}$$

Dies entspricht einem Eigenwertproblem, bei dem das Gleichungssystem

$$(\Sigma - \lambda I_p)a_1 = 0 \tag{3.17}$$

gelöst werden muss. Lösungen können nur Eigenwerte $\lambda_1 > \lambda_2 > ... > \lambda_p \geq 0$ der Matrix Σ sein, die alle größer oder gleich 0 sind, da die Kovarianzmatrix positiv semidefinit ist. Zur Maximierung der Varianz wird als Lösung der größte Eigenwert λ_1 gewählt. a_1 ist dann der zugehörige Eigenvektor.
Die zweite Komponente $z_2 = Xa_2$ wird wiederum so konstruiert, dass ihre Varianz maximal ist mit den Nebenbedingungen $a_2^T a_2 = 1$ und $a_2^T a_1 = 0$. Letztere Bedingung besagt, dass die zweite Komponente orthogonal zur ersten steht, das heißt $Cov(z_1, z_2) = 0$.
Auf diese Weise werden Schritt für Schritt die weiteren Komponenten konstruiert, die jeweils die maximale verbliebene Varianz erklären.

Diese Vorgehensweise entspricht einer Spektralzerlegung der Kovarianzmatrix Σ in

$$\Sigma = A\Lambda A^T, \tag{3.18}$$

wobei die orthogonale $p \times p - Matrix$ A der Eigenvektoren genau die gesuchten Vektoren $a_1, ..., a_p$ enthält, und Λ eine $p \times p - Diagonalmatrix$ mit den gesuchten Eigenwerten $\lambda_1, ..., \lambda_p$ in absteigender Größe auf der Diagonalen ist. Diese können als Varianzen der Hauptkomponenten $z_1, ..., z_p$ interpretiert werden.
Aus den Gleichungen 3.16 und 3.17 folgt nun

$$\sum_{j=1}^{p} Var(z_j) = \sum_{j=1}^{p} \lambda_j = tr(\Lambda) = tr(A\Sigma A^T) = tr(\Sigma \underbrace{AA^T}_{=I}) = tr(\Sigma) = \sum_{j=1}^{p} Var(x_j). \tag{3.19}$$

Dies beweist, dass die Summe der Varianzen der ursprünglichen Variablen und der Hauptkomponenten gleich ist. Die k-te Komponente erklärt also den Anteil $\frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}$ der Gesamtvarianz.
Eine Hauptkomponente mit einem Eigenwert größer als 1 beschreibt mehr Varianz als eine durchschnittliche ursprüngliche Variable. Eine Hauptkomponente mit einem Eigenwert kleiner als 1 erklärt weniger Varianz als eine durchschnittliche ursprüngliche Variable. Beschreibt eine Anzahl von q Komponenten die Daten bereits gut, so wird angenommen, dass es höchstens q latente Faktoren gibt. In der Regel ist es das Ziel, die Dimensionen der Daten zu reduzieren und sich auf wenige Komponenten zu beschränken. Die Frage, wie viele Komponenten zu verwenden sind und ab wann die Daten gut genug beschrieben werden, ist hierbei subjektiv.
Es existieren verschiedene mögliche Kriterien zur exakten Festlegung der Anzahl q der Hauptkomponenten:

- *Kaiser-Kriterium*: Es werden so viele Hauptkomponenten verwendet, wie es Eigenwerte größer als 1 gibt. Das heißt es werden nur Komponenten benutzt, die mehr Varianz als eine durchschnittliche ursprüngliche Variable erklären, $q = \{max_j \mid \lambda_j \geq 1\}$.
- *Scree-Test* (Catell 1966): Man versucht zwei Gruppen von Komponenten zu finden, die sich deutlich voneinander abgrenzen lassen. Dies geschieht mit Hilfe eines sogenannten Scree-Plots, bei dem die Eigenwerte λ_j gegen den Index j angetragen werden. Besitzt dieser Plot einen deutlichen Knick, so können die Eigenwerte rechts davon als nur zufällig von 0 verschieden aufgefasst werden, und man kann diese Komponenten weglassen.
- Man verwendet so viele Hauptkomponenten, bis deren kumulierte Varianz mindestens einen vorgegebenen Anteil κ der Gesamtvarianz erreicht hat. Diese berechnet sich durch $\sum_{k=1}^{q} \lambda_k / \sum_{j=1}^{p} \lambda_j \geq \kappa$, mit zum Beispiel $\kappa = 0,8$. Die Festlegung von κ ist hierbei wieder subjektiv.
- Man setzt inhaltliches Vorwissen ein, um die Anzahl der Komponenten festzulegen.

Die hier dargelegten Verfahren können in der Praxis durchaus zu verschiedenen Ergebnissen für die Anzahl der Komponenten führen. In diesem Fall wählt man meist die kleinste Komponentenanzahl.

Schätzung der Ladungsmatrix mit der Hauptkomponentenmethode

Auf der Basis der q ausgewählten Komponenten kann jetzt eine Ladungsmatrix durch die Hauptkomponentenmethode geschätzt werden, indem man

$$\hat{L}_{(q)} = (a_1^*...a_q^*), \tag{3.20}$$

mit $a_j^* = a_j\sqrt{\lambda_j}$ zur Ladungsmatrix zusammensetzt. Dann ergibt sich wieder analog zum Fundamentaltheorem

$$\Sigma = \hat{L}_{(q)}\hat{L}_{(q)}^T + \hat{\Psi}_{(q)}, \tag{3.21}$$

wobei es sich hier aber bei der geschätzten Fehlermatrix $\hat{\Psi}_{(q)}$ nicht mehr um eine Diagonalmatrix handelt.

3.3.3. Überblick über weitere Verfahren

Neben den beiden bisher dargelegten, meistangewandten Methoden für die Bestimmung einer ersten Ladungsmatrix existieren auch noch zahlreiche weitere Möglichkeiten, über die in diesem Abschnitt ein kurzer Überblick präsentiert wird.
Die *Zentroidmethode* oder auch *Schwerpunktmethode* nach Thurstone (1931) bietet eine in der Praxis relativ leicht durchzuführende Möglichkeit, die allerdings nur eine Näherungslösung für die Ladungsmatrix L liefert. Sie verlangt, dass die Achse, die den ersten der q extrahierten Faktoren darstellt, den Schwerpunkt der p Merkmalspunkte im q-dimensionalen Raum schneidet. Die weiteren Achsen sollen orthogonal dazu sein. Zu einer genaueren Beschreibung der Theorie und Durchführung der Zentroidmethode siehe unter anderem Überla (1968, S.113).
Die auf Rao (1955) zurückgehende *kanonische Faktorenanalyse* versucht die q kanonischen Korrelationen zwischen den p Merkmalen und den q Faktoren zu maximieren. Ihr Vorteil ist, dass hier die Anzahl der Faktoren nicht vorher festgelegt werden muss. Bei gleicher Faktorenanzahl führt die kanonische Faktorenanalyse zur gleichen Schätzung der Ladungsmatrix wie bei der Maximum-Likelihood-Methode. Für weiteres siehe unter anderem Fahrmeir u.a. (1996, S.634).
Die *Jörgeskog-Methode* nach Jörgeskog (1963) schätzt die Werte der Matrix $\Psi = diag(\Psi_1, ..., \Psi_p)$ der merkmalsspezifischen Varianzen proportional zu den Werten der Diagonalelemente der inversen Kovarianzmatrix Σ^{-1}. Dies geschieht wiederum durch Lösen eines Eigenwertproblems. Siehe dazu auch unter anderem Hartung/Elpelt (1995, S.541).

3.4. Rotation der Faktoren

Die konstruierte Ladungsmatrix ist nun meist so beschaffen, dass sich die q Faktoren kaum interpretieren lassen. Durch Rotation soll sie nun so transformiert werden, dass das Koordinatensystem der Faktoren die beobachteten Merkmale möglichst einfach beschreibt.
Wie bereits zuvor in Abschnitt 3.2.3 beschrieben wird f durch eine Rotationsmatrix R in $\tilde{f} = R^T f$ und die Ladungsmatrix L in $\tilde{L} = LR$ transformiert, so dass gilt:

$$Lf = LRR^T f = \tilde{L}\tilde{f}$$

Für die Kovarianzmatrix Σ folgt nun:

$$\begin{aligned} \Sigma &= cov(Lf) + cov(u) = LL' + \Psi \\ &= cov(\tilde{L}\tilde{f}) + cov(u) = \tilde{L}\Phi\tilde{L}' + \Psi, \qquad mit \quad \Phi = cov(\tilde{f}) = RR^T. \end{aligned} \tag{3.22}$$

Der Faktor Φ entspricht also hier der Kovarianzmatrix der rotierten Faktoren.
Ziel ist es jetzt aus der unendlichen Anzahl möglicher Rotationen, diejenige zu finden, die die beste Interpretation erlaubt. Thurstone (1947) führte hierfür den Begriff der *Einfachstruktur* der Ladungsmatrix ein, die besagt, dass Faktoren möglichst in einigen Variablen hoch und in den anderen Faktoren niedrig geladen sein sollen. Dazu gibt es wiederum verschiedene Vorschläge, wie man diese Form erreichen kann.

Zunächst lassen sich *orthogonale* sowie *schiefwinklige Faktorrotationen* unterscheiden. Während die orthogonale Rotation die Orthogonalität, also die Unkorreliertheit der Faktoren, beibehält, gibt die schiefwinklige Rotation diese Voraussetzung auf und konstruiert schiefwinklige, das heißt miteinader korrelierte Faktoren. Die Ergebnisse einer orthogonalen Rotation sind zwar in der Regel aufgrund der Unabhängigkeit der Faktoren einfacher zu interpretieren, schiefwinklige Rotation liefert dafür eine deutlichere Zuweisung der Merkmale zu einzelnen Faktoren.

Aus den unzähligen in der Theorie vorhandenen Rotationsverfahren werden die bekanntesten nachfolgend aufgeführt und kurz beschrieben.

3.4.1. Orthogonale Rotationsmethoden

Eine orthogonale Rotation entspricht einer Drehung des Koordinatenkreuzes der Faktoren um einen bestimmten Winkel α, was wiederum zu neuen unabhängigen Faktoren führt, die die Unkorreliertheit erhalten, da die Achsen auch nach der Rotation noch aufeinander senkrecht stehen. Die geläufigsten Verfahren dazu sind die *Varimax-Methode* sowie die *Quartimax-Methode*, die im Folgenden kurz dargelegt werden.

Varimax-Methode

Die Varimaxmethode nach Kaiser (1958) maximiert als Maß für die Einfachstruktur die Summe der Varianzen der quadratischen Ladungen innerhalb jeder Spalte der Ladungsmatrix $\tilde{L}$, so dass das entsprechende zu minimierende Kriterium lautet:

$$k_{vari} = \sum_{r=1}^{q} \sum_{i=1}^{p} (\tilde{l}_{ir}^2 - \frac{d_r}{p})^2 \rightarrow max, \qquad mit \quad d_r = \sum_{i=1}^{p} \tilde{l}_{ir}^2. \tag{3.23}$$

Dies entspricht der quadrierten Abweichung der Ladungsquadrate von den jeweiligen Mittelwerten einer Spalte.
Zur Maximierung sind wiederum iterative Algorithmen notwendig, deren Durchführung unter anderem in Hartung/Elpelt (1995, S.550) oder bei Lawley/Maxwell (1971, S. 73) näher ausgeführt sind.

Quartimax-Methode

Einen sehr ähnlichen Ansatz verfolgt die Quartimax-Methode. Ebenfalls soll dabei die Summe der quadratischen Varianzen maximiert werden. Allerdings geschieht dies nicht wie bei Varimax innerhalb der Spalten, sondern stattdessen innerhalb der Zeilen der Ladungsmatrix $\tilde{L}$. Dies führt zu dem neuen Kriterium:

$$k_{quarti} = \sum_{i=1}^{p} \sum_{r=1}^{q} (\tilde{l}_{ir}^2 - \frac{d_i}{q})^2 \rightarrow max, \qquad mit \quad d_i = \sum_{r=1}^{q} \tilde{l}_{ir}^2. \tag{3.24}$$

Laut Harman (1976, S.290) besteht der Hauptvorteil der Quartimax-Methode in der Tendenz zu einem generellen Faktor, das heißt eine Spalte der Ladungsmatrix enthält einen Großteil der Summe der quadrierten Ladungen.

3.4.2. Schiefwinklige Rotationsmethoden

Bei der schiefwinkligen Rotation wird nun im Gegensatz zur orthogonalen nicht das ganze Koordinatenkreuz verschoben, sondern jeder Faktor $j = 1, ..., q$ separat um den Winkel α_j gedreht, so dass miteinander korrelierte Faktoren entstehen. Als meist verwendete Methoden sind hier die Rotation nach der *Promax-Methode* sowie nach dem *Oblimax-Kriterium* zu nennen.

Promax-Methode

Die von Hendrickson und White (1964) entwickelte Promax-Methode versucht die aus der orthogonalen Varimax-Methode resultierte Ladungsmatrix $\tilde{L}$ mittels einer weiteren, diesmal allerdings schiefwinkligen Rotation weiter zu verbessern. Die Einträge der Lösung sollen dann

noch näher an 0 beziehungsweise bei 1 liegen. Man kann ein bestimmtes 0,1-Faktorenmuster vorgeben, dem die rotierte Matrix durch Kleinste-Quadrate-Schätzung möglichst nahe angepasst werden soll. Zur genaueren Beschreibung siehe Fahrmeir u.a. (1996, S.681).

Oblimax-Kriterium

Das Oblimax-Kriterium nach Saunders (1954) ist das schiefwinklige Gegenstück zum orthogonalen Quartimax-Verfahren. Wiederum sollen die Summen der quadratischen Varianzen innerhalb der Zeilen von $\tilde{L}$ maximiert werden. Allerdings wird hier auf die Zusatzforderung nach orthogonalen Faktoren verzichtet.

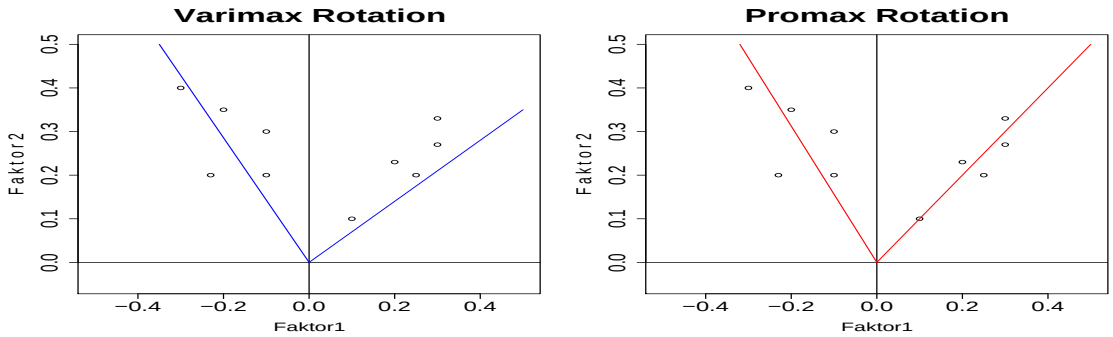

Abbildung 3.1.: Zweidimensionales grafisches Beispiel für das Rotieren von Ladungen. Bei Varimax (links) wird das gesamte Koordinatensystem verschoben und der rechte Winkel bleibt erhalten. Bei Promax (rechts) dagegen werden beide Achsen getrennt voneinander verschoben, um eine bessere Anpassung an die unrotierten Originalladungen zu bekommen.

Nach der Rotation interpretiert man die Zusammenhänge zwischen den ursprünglichen Variablen und den Faktoren anhand der rotierten Ladungsmatrix. Ist die Ladung zwischen einem Faktor und einer Variablen betragsmäßig hoch, so wird diese Variable dem Faktor zugeordnet. Hat eine Variable hohe Nebenladungen auf weiteren Faktoren, so bedeutet dies, dass sie noch einen zusätzlichen Faktor erfasst. Hat eine Variable dagegen auf allen Faktoren nur geringe Ladungen, so misst sie etwas anderes als die extrahierten Faktoren. Variablen, die dem gleichen Faktor zugeordnet werden, stehen untereinander in höherem Zusammenhang als solche, die verschiedenen Faktoren zugeordnet werden.

3.5. Schätzung der Faktorenwerte

Die letzten zu bestimmenden Größen des Faktorenmodells sind die Faktorenwerte für jedes Individuum selbst. Die beiden dazu am häufigsten verwendeten Methoden sind zum einen die

ML-Methode nach Bartlett (1937) sowie zum anderen die *Regressionsmethode* nach Thomson (1951).
Nehmen die Faktorenwerte einen hohen positiven Wert an, so steht das für eine hohe Ausprägung eines Individuums auf einem Faktor, sind sie dagegen negativ, so steht das für eine geringe Ausprägung auf dem Faktor. Der Wert 0 entspräche genau dem durchschnittlichen Wert aller Individuen ($E(f) = 0$).

3.5.1. Maximum-Likelihood-Methode

Folgen die beobachteten Daten einer Normalverteilung, so wird der Vektor $f = (f_1, ..., f_q)$ der Faktorenwerte nach Bestimmung der restlichen Parameter des Modells der Faktorenanalyse als unbekannter aber fester Verteilungsparameter angesehen, so dass der Vektor $(x-\mu)$ p-dimensional normalverteilt ist mit Erwartungswert Lf, Kovarianzmatrix Ψ und der Likelihood-Dichte

$$L(x - \mu \mid f) = \sqrt{2\pi^{-p}} \mid \Psi \mid^{-\frac{1}{2}} exp[-\frac{1}{2}[(x - \mu - Lf)'\Psi^{-1}(x - \mu - Lf)]]. \tag{3.25}$$

Den ML-Schätzer $\hat{f}$ von f erhält man durch Maximierung von L.
Äquivalent dazu ist nach Fahrmeir (1996, S.690) die Minimierung der durch Ψ standartisierten Abweichungsquadratsumme

$$(x - \mu - Lf)'\Psi^{-1}(x - \mu - Lf) = u'\Psi^{-1}u = \sum_{i=1}^{p}(\frac{u_i^2}{\Psi_i^2}) = y \ \rightarrow \ min. \tag{3.26}$$

Deshalb wird die ML-Methode auch häufig als *Kleinste-Quadrate-Methode* bezeichnet.
Nullsetzen der partiellen Ableitung liefert dann den Schätzer

$$\hat{f}_{ML} = (L'\Psi^{-1}L)^{-1}L'\Psi^{-1}(x - \mu). \tag{3.27}$$

Dies ist der beste unverzerrte Schätzer für f (siehe Fahrmeir u.a. 1996, S.691). Da in der KQ-Methode keine Momente von f vorkommen, ist der Schätzer auch invariant gegenüber einer Skalierung der Daten.

3.5.2. Regressionsmethode

Ausgehend von Gleichung 3.1 wäre bei einer normalen Regression f die bekannte Einflussgröße und die Elemente der Ladungsmatrix L die unbekannten Parameter, welche durch die Regression ermittelt werden. In diesem Fall verhält es sich genau umgekehrt, da L bereits bekannt ist, und f gesucht wird.
Die Idee ist nun, jedes f in linearer Abhängigkeit von μ und l multipliziert mit einer neuen Variablen a zu schreiben. Der Vektor f sei hierbei zufällig mit Erwartungswert 0 und

Kovarianz Φ. Die Variable a wiederum ist abhängig von der Ladungsmatrix L.

$$\tilde{f}_r = a_r(x-\mu) = (x-\mu)'a_r \quad , \quad r = 1, ..., q$$

mit

$$E(\tilde{f}_r - f_r)^2 = E((x-\mu)'a_r - f_r)^2 = y \;\rightarrow\; min, \quad r = 1, ..., q. \tag{3.28}$$

Nach partieller Differentiation von y nach a_r und anschließendem Nullsetzen erhält man nach Fahrmeir, u.a. (1996) den Regressionsschätzer

$$\tilde{f}_{Reg} = \Phi(I + L'\Psi^{-1}L\Phi)^{-1}L'\Psi^{-1}(x-\mu). \tag{3.29}$$

Dieser Schätzer ist zwar nicht unverzerrt, besitzt aber unter allen linearen Schätzern die geringste Fehlervarianz.

3.6. Standardfehler und Konfidenzintervalle

Bisher wurden die Parameter des Modells der Faktorenanalyse nur als Punktschätzer betrachtet. Durch Standardfehler und Konfidenzintervalle kann man nun die Unsicherheit dieser Schätzungen quantifizieren.
Von den vorhandenen Beobachtungen, auf die das Modell angewendet wurde, soll in der Inferenzstatistik auf die zugrundeliegende Grundgesamtheit geschlossen werden. Da die Auswahl der beobachteten Stichprobe aus der Grundgesamtheit schon einem Zufallsprozess entsprach, sind die aus dieser Stichprobe erhaltenen Punktschätzer des Faktorenmodells Zufallsvariablen.
Die Verteilung dieser Zufallsvariablen kann man ohne parametrische Annahmen direkt aus der Stichprobe mittels sogenannter *nonparametrischer Bootstrap Methoden* schätzen.
Bootstrap Verfahren wurden erstmals von Efron 1979 beschrieben. Sie kommen dann zum Einsatz, wenn statistische Kennzahlen geschätzt werden sollen, über die keine parametrischen Annahmen getroffen werden können oder sollen. Diese Schätzung der Parameter θ entspricht im Falle der Faktorenanalyse der Schätzung der Elemente der Ladungsmatrix l_{jk} oder eines Eigenwerts λ_j. Auf Grundlage der vorhandenen Stichprobe werden die Parameter $\hat{\theta}$ geschätzt und ihre Unsicherheit dann durch Bootstrap beziffert.
Dieses Verfahren lässt sich im Allgemeinen in drei Schritte gliedern:

1. Man zieht B Stichproben des Umfangs n aus den Daten der vorliegenden Stichprobe *mit Zurücklegen*. Das heißt in einer Stichprobe $b = 1, ..., B$, welche als Bootstrap-Stichprobe bezeichnet wird, kann eine Beobachtung zwischen *0* und n-mal vorhanden sein.

2. Aus jeder Bootstrap-Stichprobe schätzt man nun die interessierenden Parameter, so dass man B Schätzungen $\theta^* = (\theta_1^*, ..., \theta_B^*)$ erhält. Für das Faktorenmodell bedeutet dies, dass

B-mal das Faktorenmodell aufgestellt wird und die Eigenwerte und Ladungen bestimmt werden.

3. Aus diesen Bootstrap-Schätzern bestimmt man Standardfehler und Konfidenzintervalle der Punktschätzer.

Insgesamt sind $\binom{2n-1}{n}$ verschiedene Bootstrap-Stichproben möglich. Daher spricht man von einer idealen Bootstrap-Schätzung, wenn man B als die Anzahl der möglichen Stichproben wählt. Da dies in der Praxis häufig nicht möglich ist, beschränkt man sich auf die Faustregeln, dass zur Schätzung von Standardfehlern $B = 200$ und zur Schätzung von Konfidenzintervallen $B = 2000$ ausreichend sind.
Im Folgenden wird nun beschrieben, wie aus der Verteilung der Bootstrap-Schätzer θ^* auf die interessierenden Kennzahlen zur Quantifizierung der Unsicherheit des Schätzers $\hat{\theta}$ geschlossen werden kann.

Bootstrap-Schätzung des Standardfehlers

Den theoretischen Standardfehler $se(\theta)$ schätzt man durch

$$\hat{se}_B = \left(\frac{1}{B-1} \sum_{i=1}^{B} (\theta_b^* - \theta_{(.)}^*)^2 \right)^{\frac{1}{2}}, \tag{3.30}$$

wobei $\theta_{(.)}^*$ für das arithmetische Mittel der Bootstrap-Schätzer steht: $\theta_{(.)}^* = \frac{1}{B} \sum_{i=1}^{B} \theta_b^*$.

Bootstrap-Schätzung des Konfidenzintervalls

Zur Berechnung der Konfidenzintervalle für die Punktschätzer $\hat{\theta}$ gibt es mehrere Ansätze, von denen die gängigsten im Folgenden dargelegt werden:

Klassisches Konfidenzintervall: Am naheliegendsten ist es üblicherweise, ein klassisches zweiseitiges $(1-\alpha)$-Konfidenzintervall auf Basis der Bootstrap-Schätzung des Standard- fehlers zu konstruieren:

$$KI = \left[\hat{\theta} \pm t_{n-1,1-\alpha/2} \cdot \hat{se}_B \right], \tag{3.31}$$

mit dem $(1 - \alpha/2)$-Quantil einer t-Verteilung mit $n - 1$ Freiheitsgraden.
Allerdings kann man auf diese Weise nur symmetrische Intervalle bilden, die außerdem auch noch außerhalb des Definitionsbereichs der Parameter liegen können. So sind zum Beispiel alle Eigenwerte stets positiv, das Konfidenzintervall könnte hingegen auch negative Werte abdecken.

Bootstrap-Perzentil-Konfidenzintervall: Eine weitere Idee zur Konstruktion besteht darin, die Bootstrap-Schätzer θ_b^* ihrer Größe nach zu ordnen und dann das Konfidenzintervall direkt

aus der empirischen Verteilungsfunktion mit den Grenzen

$$KI_{BP} = \left[\theta^{*(\alpha/2)}; \theta^{*(1-\alpha/2)}\right] \tag{3.32}$$

zu bilden, wobei $\theta^{*(\alpha/2)}$ das $B \cdot \alpha/2$-te Element und $\theta^{*(1-\alpha/2)}$ das $B \cdot (1-\alpha/2)$-te Element der geordneten Bootstrap-Schätzer $(\theta_{[1]}, ..., \theta_{[B]})$ bezeichnet.
Die Konfidenzintervalle nach der Perzentilmethode liegen zwar auf jeden fall innerhalb des Definitionsbereichs, haben allerdings nach Efron (1993) häufig eine geringere als die Überdeckungswahrscheinlichkeit.

Bias-corrected and accelerated Konfidenzintervall: Eine Alternative schlagen Efron und Tibshirani (1993, S.184ff) durch das Einführen einer Korrektur für die Verzerrung (*bias*) und die Beschleunigung (*acceleration*) vor. Dadurch soll den zuvor beschriebenen Nachteilen der anderen Ansätze entgegengewirkt werden. Das $(1 - \alpha) - BC$ Konfidenzintervall stellt sich dann dar als

$$KI_{BC} = \left[\theta^{*(\alpha_1)}; \theta^{*(\alpha_2)}\right], \tag{3.33}$$

wobei wiederum $\theta^{*(\alpha_1)}$ bzw. $\theta^{*(\alpha_2)}$ das $B \cdot \alpha_1$-te bzw. das $B \cdot \alpha_2$-te Element der geordneten Bootstrap-Schätzer beschreiben, mit

$$\begin{aligned} \alpha_1 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/2)}}{1 - \hat{\alpha} \cdot (\hat{z}_0 + z^{(\alpha/2)})}\right) \\ \alpha_2 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/2)}}{1 - \hat{\alpha} \cdot (\hat{z}_0 + z^{(1-\alpha/2)})}\right). \end{aligned} \tag{3.34}$$

Dabei bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung und $z^{(\alpha)}$ ihr $100 \cdot \alpha$-Quantil ist. Für $\hat{z}_0 = \hat{\alpha}_0 = 0$ entspräche das BC-Intervall genau dem Konfidenzintervall nach der Perzentil-Methode.
Der Korrekturfaktor für die Verzerrung berechnet sich als

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#(\theta_b^* < \hat{\theta})}{B}\right). \tag{3.35}$$

Dies ist die Quantilsfunktion der Standardnormalverteilung für den Anteil der Bootstrap-Schätzer, die kleiner als der ursprüngliche Punktschätzer sind.
Den zweiten Korrekturfaktor schätzt man nach Efron (S.186) als

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^3}{6\left\{\sum_{i=1}^{n} (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2\right\}^{\frac{3}{2}}}, \tag{3.36}$$

wobei $\hat{\theta}_{(i)}$ die Schätzung für θ ohne die i-te Beobachtung der Stichprobe ist und $\hat{\theta}_{(.)} =$

$\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$ der Mittelwert über alle n derartigen Schätzer.
Mit diesen Faktoren lässt sich das Intervall aus Gleichung 3.33 berechnen, welches zwar etwas komplizierter zu konstruieren ist, im Allgemeinen aber durch seine theoretischen Vorteile (siehe Efron und Tibshirani, 1993, S.186-188) den anderen Intervallen vorzuziehen ist.

Die Bootstrap-Verfahren sind also eine gute Möglichkeit zur Berechnung von Standardfehlern und Konfidenzintervallen. Wichtig für Schlüsse von einer Stichprobe auf eine Grundgesamtheit ist allerdings, dass die Stichprobe diese Grundgesamtheit möglichst gut abbildet. Die einzelnen Elemente der Stichprobe werden dabei als unabhängig angenommen.
Für die Faktorenanalyse ist es demnach möglich, die Unsicherheiten der Punktschätzer wie etwa der Eigenwerte zu beziffern, wenn keine theoretische Verteilung bekannt ist. In der Praxis werden solche Unsicherheitsabschätzungen jedoch eher selten durchgeführt.

3.7. Anwendungsbeispiel

Anhand eines Beispiels aus der Praxis wird die Faktorenanalyse nun in den zuvor beschriebenen Schritten vorgeführt.
Dabei handelt es sich um einen Datensatz aus einem Zehnkampf, bei dem die männlichen Sportler für die erbrachten Leistungen in 10 leichtathletischen Disziplinen jeweils Punkte erhalten. Es handelt sich dabei um quantitative Variablen, die als normalverteilt angesehen werden können.
Über die Variablen und Disziplinen wird in Tabelle 3.2 eine Übersicht gegeben.

Variablenname	Disziplin
X100mP	100Meter-Sprint
LJP	Weitsprung (Long Jump)
ShotP	Kugelstoßen (Shot Put)
HJP	Hochsprung (High Jump)
X400mP	400Meter-Lauf
X110HP	110Meter-Hürdenlauf
DisP	Diskuswerfen
PVP	Stabhochsprung (Pole Vault)
JavP	Speerwerfen (Javelin)
X1500mP	1500Meter-Lauf

Tabelle 3.2.: Übersicht über die Disziplinen im Zehnkampf

Diese Punktzahlen sollen nun als Indikatoren für wenige latente Variablen dienen, die die Leistungen der Sportler bestimmen, das heißt man versucht $q \ll 10$ Faktoren, wie etwa Kraft, Ausdauer etc., auf Basis dieser Variablen zu konstruieren.

Dazu betrachtet man sich zuerst die Korrelationsmatrix (siehe Tabelle 3.3), um sich einen Überblick über die Abhängigkeiten zwischen den Punktezahlen in den einzelnen Disziplinen zu verschaffen.

	X100mP	LJP	ShotP	HJP	X400mP	X110HP	DisP	PVP	JavP	X1500mP
X100mP	1.000	0.424	0.338	0.286	0.348	0.306	0.114	0.129	-0.120	-0.104
LJP	0.424	1.000	0.104	0.467	0.315	0.496	-0.133	-0.264	0.130	0.201
ShotP	0.338	0.106	1.000	0.118	0.147	0.304	0.315	0.012	0.226	-0.066
HJP	0.286	0.467	0.118	1.000	0.246	0.315	-0.003	-0.063	0.041	0.025
X400mP	0.348	0.315	0.147	0.246	1.000	-0.052	0.224	0.043	-0.184	0.207
X110HP	0.306	0.496	0.304	0.315	-0.052	1.000	0.002	-0.107	0.064	-0.149
DisP	0.114	-0.133	0.315	-0.003	0.224	0.002	1.000	-0.044	-0.018	0.081
PVP	0.129	-0.264	0.012	-0.063	0.043	-0.107	-0.044	1.000	-0.159	-0.306
JavP	-0.120	0.130	0.226	0.041	-0.184	0.064	-0.018	-0.159	1.000	0.141
X1500mP	-0.104	0.201	-0.066	0.025	0.207	-0.149	0.081	-0.306	0.141	1.000

Tabelle 3.3.: Korrelationsmatrix der Variablen

Man kann erkennen, dass einige Variablen positiv und einige negativ miteinander korreliert sind. So ist etwa ein guter Weitspringer durch seine Sprungkraft offensichtlich auch meist ein guter Hochspringer sowie ebenfalls erfolgreich in den Sprintdisziplinen, wohingegen beispielsweise sprintstarke Athleten im Ausdauerbereich über 1500m im Durchschnitt etwas schwächere Leistungen bringen.

In der Faktorenanalyse versucht man nun, passende Faktoren zu extrahieren und die Varianzen in einen gemeinsamen und einen spezifischen Teil zu zerlegen.
Da hier über Art und Anzahl der Faktoren zunächst nichts bekannt ist, hat die durchgeführte Analyse einen explorativen Charakter.
Für die Analyse werden nicht die absolut erreichten Punktzahlen in den Disziplinen verwendet, sondern der Datensatz wird zuvor normiert, so dass man die relativen Stärken und Schwächen der Athleten vorliegen hat.

Bestimmung der Faktorenzahl

Im nächsten Schritt muss die Anzahl der Faktoren festgelegt werden. Für die Wahl dieser Anzahl gibt es keine objektiv beste Methode. Wie bereits zuvor beschrieben versucht man vielmehr unter anderem aus inhaltlichen Gesichtspunkten auf eine passende Anzahl zu schließen, die einen möglichst guten Kompromiss zwischen Modellkomplexität und Interpretierbarkeit, sowie der Anpassung an die Daten schafft.

Eine Möglichkeit liefert die Zerlegung der Kovarianzmatrix Σ in die Eigenwerte λ_j und deren zugehörige Eigenvektoren nach der Hauptkomponentenmethode. Dabei ergeben sich für die

Zehnkampfdaten die folgenden 10 Eigenwerte $\lambda_1, ..., \lambda_{10}$:

`2.482  1.570  1.394  1.312  0.842  0.687  0.540  0.476  0.460  0.237.`

Nach dem Kaiser-Kriterium würde man *4* Faktoren wählen, da es so viele Eigenwerte gibt, die größer als 1 sind. Der Screeplot (siehe Abbildung 3.2) weißt einen Knick nach dem 2. Eigenwert und einen noch deutlicheren nach dem 5. Eigenwert auf, was dementsprechend auf *5* Faktoren hindeutet. Als weiteres Kriterium kann man noch den Anteil der erklärten Varianz $\sum_{k=1}^{q} \lambda_k / \sum_{j=1}^{p} \lambda_j$ heranziehen. 3 Faktoren würden demnach 54,4% der Varianz der ursprünglichen Variablen erklären, für 4 Faktoren wären es 67,6% und bei 5 Faktoren 76,0% (siehe Stabdiagramm in Abbildung 3.2). Durch das Hinzunehmen von weiteren Faktoren ließen sich darüber hinaus jeweils nur noch wenige Prozente dazugewinnen.

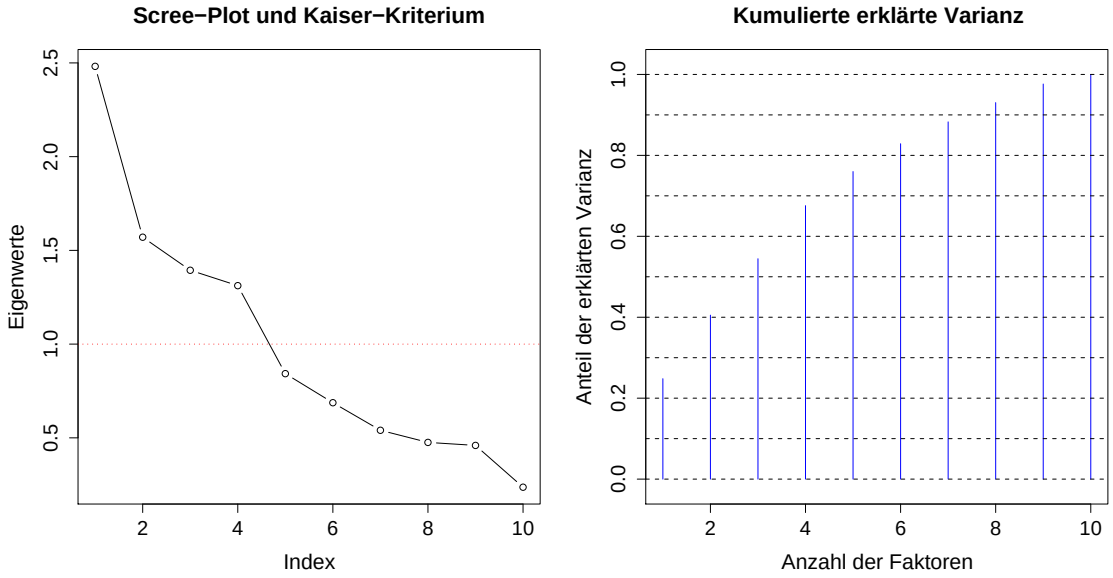

Abbildung 3.2.: Kriterien zur Bestimmung der Faktoranzahl

Wie man sieht, hat man bei der Durchführung der Faktorenanalyse nach der Hauptkomponentenmethode einen subjektiven Spielraum bei der Festlegung der Anzahl der Faktoren. Da die beschriebenen Kriterien hier zu unterschiedlichen Faktorenzahlen führen, wählt man aus Gründen der einfacheren Interpretierbarkeit die geringste Zahl, die hier dem 4-Faktorenmodell entspricht, welches das Kaiser-Kriterium empfiehlt.

Aufgrund der Normalverteilungsannahme der Daten kann man hier auch die Maximum-Likelihood-Methode zur Bestimmung der Faktorenzahl q durchführen.
Man testet dabei schrittweise, mit *q=0* beginnend und sich jeweils um einen Faktor steigernd, die

Hypothesen, dass die Daten einem Faktorenmodell mit der entsprechenden Anzahl an Faktoren entspricht. Dies geschieht solange, bis der Likelihood-Quotiententest die Hypothese auf einem Signifikanzniveau α nicht mehr ablehnt. In diesem Beispiel wird als Signifikanzniveau $\alpha = 0,05$ gewählt.
Die Hypothese, dass *0* Faktoren vorliegen, was bedeuten würde, dass die Variablen unabhängig sind und es daher nicht sinnvoll wäre eine Faktorenanalyse durchzuführen, wird ebenso abgelehnt wie die Hypothesen eines 1-, 2- und auch eines 3-Faktorenmodell, welche mit $5,45 \cdot 10^{-9}$, bzw. $2,18 \cdot 10^{-5}$ und 0.0063 nur sehr geringe p-Werte aufweisen.
Da im Gegensatz dazu das 4-Faktorenmodell einen hohen p-Wert von 0,766 erreicht und die Gegenhypothese damit nicht mehr signifikant ist, wird die Hypothese für dieses Modell nicht mehr abgelehnt, und man entscheidet sich für eine Analyse mit 4 Faktoren.

Schätzung und Rotation der Ladungsmatrix

Nach Bestimmung der Faktorenzahl wird das Faktorenmodell mit $q = 4$ Faktoren aufgestellt und die Ladungsmatrix L durch das Maximum-Likelihood-Verfahren geschätzt. Anschließend hat man die Möglichkeit diese Matrix durch orthogonale oder schiefwinklige Rotation auf eine besser zu interpretierende Form zu transformieren. Diese beiden Möglichkeiten werden in diesem Beispiel mit der Varimax- und der Promax-Methode angewendet.
In der Praxis lässt sich dies im Programmpaket *R* alles auf einmal mittels der Funktion *factanal* (siehe Anhang A) durchführen.
Tabelle 3.4 zeigt die Ladungsmatrizen nach der Varimax- bzw. nach der Promaxmethode. Darin tauchen nur die Werte auf, die betragsmäßig größer als *0,1* sind.
Für die Ladungsmatrix nach der Varimax-Rotation ist zu beachten, dass wieder orthogonale Faktoren entstehen, weswegen diese getrennt voneinander interpretiert werden können. Der Ladungskoeffizient l_{jk} gibt die Korrelation der j-ten Variable mit dem k-ten Faktor an. Man kann also sehen, dass der 1. Faktor vor allem mit dem Weitsprung korreliert ist. Relativ hohe Ladungen treten auch im Hochsprung sowie den Sprintdiszipinen auf, für die ebenfalls die für den Weitsprung notwendigen Sprung- bzw. Schnelligkeitsfähigkeiten vorteilhaft sind. Faktor *2* ist mit einer Ladung von 0.93 eng mit dem Kugelstoßen verbunden und besitzt zudem positive Ladungen in den weiteren Wurfdisziplinen sowie gering positive ladungen in den Sprintdisziplinen. Der 3. Faktor entspricht hauptsächlich den für den 400 Meterlauf nötigen Fähigkeiten und mit Faktor *4* laden vor allem die 1500 Meter.

Die Kommunalitäten geben an, welcher Anteil der Varianz einer Variable durch das Faktorenmodell beschrieben wird. Einige Variablen werden demnach fast vollständig durch das Modell erklärt, wohingegen andere Variablen, die jeweils nur geringere Ladungen mit den Faktoren aufweisen, niedrige Kommunalitäten haben, was einem größeren Wert in der Diagonalmatrix Ψ für die spezifische Varianz entspricht. Der Stabhochsprung beispielsweise weist nur mit dem 4. Faktor eine signifikant von *0* verschiedene Ladung auf, die in diesem

			Varimax		
Variable	Faktor1	Faktor2	Faktor3	Faktor4	Kommunalität
100m	0.57	0.21	0.22	-0.23	0.47
Weitsprung	0.90	-0.15	0.12	0.38	0.99
Kugelstoßen	0.28	0.93	0.12		0.96
Hochsprung	0.48		0.13		0.25
400m	0.26	0.18	0.84		0.81
110m-Hürden	0.61	0.11	-0.26		0.46
Diskuswerfen	-0.11	0.40	0.21		0.22
Stabhochsprung				-0.51	0.28
Speerwerfen		0.19	-0.28	0.38	0.27
1500m			0.26	0.61	0.45
			Promax		
Variable	Faktor1	Faktor2	Faktor3	Faktor4	
100m	0.40	0.18	0.25	-0.25	
Weitsprung	1.02	-0.13	0.11	0.30	
Kugelstoßen		0.99			
Hochsprung	0.47		0.14		
400m	0.15		0.88	0.14	
110m Hürden	0.56	0.18	-0.25	-0.13	
Diskuswerfen	-0.26	0.37	0.26		
Stabhochsprung	-0.24			-0.50	
Speerwerfen	0.10	0.28	-0.27	0.34	
1500m			0.26	0.67	

Tabelle 3.4.: Geschätzte Ladungen für das 4-Faktorenmodell mit Varimax- bzw. Promax-Rotation

Fall zudem noch negativ ist. Zur erfolgreichen Bewältigung dieser Disziplin sind demnach offensichtlich noch andere Einflüsse ausschlaggebend.

Mit der schiefwinkligen Promax-Rotation wird versucht, eine bessere Annäherung der Ladungsmatrix an die Einfachstruktur zu erreichen. Da dazu aber die Forderung nach Orthogonalität der Faktoren aufgegeben wird, ist zu beachten, dass die Faktoren nicht mehr getrennt voneinander interpretiert werden können. Die Angabe von Kommunalitäten macht daher keinen Sinn; diese könnten außerdem auch negativ werden bzw. größer als *1*.
Die Ladungsmatrix unterscheidet sich in diesem Beispiel nicht wesentlich von der zuvor interpretierten orthogonal rotierten Matrix.
Die Korreliertheit der Faktoren miteinander macht die Interpretation der schiefwinklig rotierten Ladungsmatrix etwas schwieriger. Sie ist hier aber aus inhaltlicher Gründen sinnvoll, da die körperlichen Fähigkeiten aus sportwissenschaftlicher Sicht durchaus nicht unabhängig zu sein scheinen.

Zum Abschluss der Faktorenanalyse muss noch die $n \times 4$-Faktormatrix F geschätzt werden. Dies lässt sich durch die zuvor beschriebene ML-Methode oder die Regressionsmethode praktizieren, was in der Praxis im Programmpaket *R* ebenfalls von der Funktion *factanal* durchgeführt wird.

Für einen Zehnkämpfer wäre es in diesem Beispiel vorteilhaft, möglichst in allen 4 Faktoren hohe Werte zu erzielen, da hohe Werte für eine hohe Ausprägung des Faktors bei einem Athleten stehen. Werte nahe bei *0* entsprechen dem Durchschnitt und negative Werte sprechen für unterdurchschnittliche Ausprägung des Faktors bei einem der Sportler.

4. Faktorenanalyse für nicht vollständige Daten

Bisher wurde die Faktorenanalyse nur für vollständige Datensätze dargelegt. Liegen jedoch fehlende Beobachtungen innerhalb der Daten vor, lässt sich die Faktorenanalyse nicht mehr ohne Weiteres durchführen. Man muss deshalb Methoden anwenden, die die Löcher in der nicht kompletten Datenmatrix X auffüllen.
Im Folgenden werden mögliche Vorgehensweisen dafür beschrieben, wenn fehlende Daten in allen Variablen vorliegen.
Die Arbeit beschränkt sich auf den speziellen Fall, in dem die Daten als multivariat normalverteilt vorausgesetzt werden können, was auch eine Voraussetzung für die ML-Faktorenanalyse ist. Zudem soll als Fehlendmechanismus Missing at Random, das heißt, ein ignorierbarer Ausfallmechanismus, vorliegen.

Als einfachste Methode ließen sich wieder Eliminierungsverfahren, wie Complete Case Analysis, anwenden, die die Faktorenanalyse nur anhand der vollständigen Beobachtungen durchführt. Diese sind jedoch nur bei einem sehr geringen Anteil fehlender Daten sowie bei Vorliegen von Missing Completely at Random sinnvoll. Daher wird hier auf diese Methoden nicht näher eingegangen, sondern die Konzentration auf *likelihoodbasierte Verfahren* und *multiple Imputationsverfahren* gerichtet.

4.1. Likelihoodbasierte Behandlung der Daten mit dem EM-Algorithmus

Einen grundlegenden Ansatz stellen die likelihoodbasierten Verfahren dar, in die schon in Abschnitt 2.4 eine Einführung gegeben wurde. Als wichtigstes dieser Verfahren gilt der EM-Algorithmus, der im Folgenden für multivariat normalverteilte Daten und ignorierbaren Fehlendmechanismus beschrieben wird.

4.1.1. Vorbereitende Definitionen

Vor der Erläuterung des EM-Algorithmus werden zum Verständnis seines Sinns und seiner Durchführung noch zwei wichtige Definitionen vorausgeschickt.

Monotone Patterns als spezielles Fehlendmuster

In unvollständigen Datenmatrizen können verschiedene *Fehlendmuster*, sogenannte *Patterns*, vorliegen. Diese sagen aus, wo innerhalb der Matrix die Datenlöcher auftreten oder wie deren Struktur beschaffen ist.
Als wichtigstes Beispiel dafür gelten *monotone Patterns*, wie in Abbildung 4.1 in einem Beispiel dargestellt. Wie man sieht, nehmen die beobachteten Daten mit größerem Variablenindex immer mehr ab. Ein solches Muster tritt beispielsweise häufig in klinischen Langzeitstudien mit Drop-Out auf. Dort können nach einer gewissen Zeit die Daten von immer weniger Individuen beobachtet werden, da viele im Laufe der Studie ausscheiden.

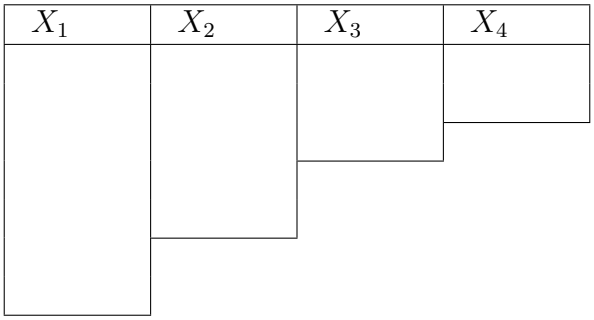

Abbildung 4.1.: Beispiel für monotones Pattern mit 4 Variablen

Viele Likelihood-Methoden für fehlende Daten basieren auf solchen monotonen Patterns. Die beobachteten Daten können in diesem Fall für p Variablen und MAR-Fehlendmechanismus folgendermaßen faktorisiert werden.

$$\prod_{i=1}^{n} f(x_{i1}, ..., x_{ik}|\Phi_1) = \prod_{i=1}^{n} f(x_{i1}|\Phi_2) \cdot \prod_{i=1}^{m_2} f(x_{i2}|\Phi_2) \cdot ... \cdot \prod_{i=1}^{m_k} f(x_{ik}|\Phi_k)\,. \tag{4.1}$$

Dabei ist $f(x_{ij}|x_{i1}, ..., x_{i,j-1}, \Phi_j)$ die bedingte Verteilung von x_{ij} für gegebene Daten $x_{i1}, ..., x_{i,j-1}$, sowie den Parameter Φ_j.

Für multivariat normalverteilte Daten mit einem monotonen Pattern ist diese bedingte Verteilung wiederum normalverteilt mit Erwartungswert linear in $x_{i1}, ..., x_{i,j-1}$ und mit konstanter Kovarianzmatrix.

Nach Little & Rubin (2002, S.144) berechnet sich die Likelihood für einen MVN-Datensatz dann nach folgendem Schema:
1. Schritt: Berechne den Erwartungswertvektor und die Kovarianzmatrix für Block 1 der vollständig beobachteten Variablen.
2. Schritt: Berechne die multivariate lineare Regression von Block 2 auf Block 1, der nächsthäufigst beobachteten Variablen. Dazu verwendet man alle bis dahin beobachteten Variablen.

$\vdots$

k. Schritt: Berechne die multivariate lineare Regression von Block k auf Block $k-1$. Dazu werden die Beobachtungen aus allen Variablen verwendet.

Die meisten multivariaten Datensätzen haben jedoch kein monotones Pattern und können auch nicht durch Zeilen- und Spaltentransformation in ein solches transformiert werden. Stattdessen sind die Datenlöcher oftmals zufällig in der Datenmatrix verteilt ohne erkennbares Fehlendmuster. Dann können ML-Schätzer nicht mehr so einfach durch Faktorisierung berechnet werden.
Für diesen Fall ist der EM-Algorithmus die passende Lösung, da man für diesen weder die Likelihood der beobachteten Daten noch deren Ableitungen explizit berechnen muss. Sein großer Vorteil ist demnach, dass er kein monotones Fehlendmuster voraussetzt, sondern auch auf ein generelles Pattern anwendbar ist.

Der Sweep-Operator

Als zweite Vorbereitung auf den EM-Algorithmus wird in diesem Abschnitt der sogenannte *Sweep-Operator* eingeführt. Diesen benötigt man zur Ausführung des E-Schritts, in dem die ML-Schätzer für die fehlenden Daten mittels linearer Regression berechnet werden.
Nach Little/Rubin (2002, S.148) transformiert der Sweep-Operator die Elemente einer symmetrischen $p \times p$-Matrix G auf folgende Weise, so dass eine neue Matrix $H = SWP[k]G$ entsteht:

$$\begin{aligned} h_{kk} &= -\frac{1}{g_{kk}}\,, \\ h_{jk} &= h_{kj} = \frac{g_{jk}}{g_{kk}}\,, \quad \text{für} \quad k \neq j\,, \\ h_{jl} &= = h_{lj} = g_{jl} - \frac{g_{jk}g_{kl}}{g_{kk}}\,, \quad \text{für} \quad k \neq j\,,\ k \neq l\,. \end{aligned} \tag{4.2}$$

Als Beispiel zur Veranschaulichung betrachte man eine symmetrische 3×3-Matrix G (aus Little & Rubin (2002, S.149)):

$$G = \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{12} & g_{22} & g_{23} \\ g_{13} & g_{23} & g_{33} \end{pmatrix}.$$

$$H = SWP[1]G = \begin{pmatrix} -1/g_{11} & g_{12}/g_{11} & g_{13}/g_{11} \\ g_{12}/g_{11} & g_{22} - g_{12}^2/g_{11} & g_{23} - g_{13}g_{12}/g_{11} \\ g_{13}/g_{11} & g_{23} - g_{13}g_{12}/g_{11} & g_{33} - g_{13}^2/g_{11} \end{pmatrix}.$$

In diesem Beispiel gilt trivialerweise $k = 1$. Für größere Matrizen ist auch die schrittweise Ausführung des Sweep-Operators möglich. In diesem Fall verwendet man die Bezeichnung $H = SWP[k_1, k_2, ..., k_t]G$. Die Reihenfolge spielt hierbei keine große Rolle, da der Sweep-Operator kommutativ ist, und somit jede Permutation von $k_1, ..., k_t$ zum gleichen Ergebnis führt. Einzig hinsichtlich der Komplexität der Berechnung können manche Permutationen besser geeignet sein.

Der Nutzen des Sweep-Operators für den EM-Algorithmus liegt in der Verbundenheit zur linearen Regression. Wird er auf normalverteilte Daten angewendet, so verwandelt er dabei abhängige Variablen in Prädiktorvariablen.

In einer 2×2 Kovarianzmatrix G für zwei Variablen X_1 und X_2 ist in $H = SWP[1]G$ das Element h_{12} gleichbedeutend mit dem Regressionskoeffizienten von X_1 aus einer Regression von X_2 auf X_1. Das Element h_{22} entspricht der Residuenvarianz von X_2 (Little & Rubin 2002, S.149).
Für den allgemeinen Fall eines Datensatzes mit n Beobachtungen von p Variablen wendet man nach Little/Rubin den Operator an auf die $(p+1) \times (p+1)$-Matrix der skalierten Kreuzprodukte

$$G = \begin{pmatrix} 1 & \bar{x}_1 & \dots & \bar{x}_1 & \dots & \bar{x}_p \\ \bar{x}_1 & \frac{1}{n}\sum x_1^2 & \dots & & \dots & \frac{1}{n}\sum x_1^2 x_p^2 \\ \vdots & \vdots & \ddots & & & \vdots \\ \bar{x}_k & & & \sum x_j^2 x_k^2 & & \\ \vdots & \vdots & & & \ddots & \vdots \\ \bar{x}_p & \frac{1}{n}\sum x_1^2 x_p^2 & \dots & & \dots & \frac{1}{n}\sum x_p^2 \end{pmatrix}.$$

Zum besseren Verständnis laufen hier die Indizes der Zeilen und Spalten von *0* bis p, damit die Variable X_j jeweils auch in Reihe j und Spalte j auftritt.

Dann führt die Anwendung des Sweep-Operators $SWP[0]$ zu einer Matrix

$$SWP[0]G = \begin{pmatrix} -1 & \bar{x}_1 & \dots & \bar{x}_1 & \dots & \bar{x}_p \\ \bar{x}_1 & s_{11} & \dots & & \dots & s_{p1} \\ \vdots & \vdots & \ddots & & & \vdots \\ \bar{x}_k & & & s_{jk} & & \\ \vdots & \vdots & & & \ddots & \vdots \\ \bar{x}_p & s_{1p} & \dots & & \dots & s_{pp} \end{pmatrix}.$$

Für eine lineare Regression würde dies bedeuten, dass die Mittelwerte in der ersten Zeile und Spalte den Regressionskoeffizienten von $X_1, ..., X_p$ für die Konstante $X_0 = 1$ entsprechen. Der Eintrag s_{jk} ist die Residuenkovarianzmatrix der Regression.
Diese Umwandlung wird nach Little/Rubin (2002, S.150) *Sweepen auf den konstanten Term* genannt, und die Matrix danach wird als *gesteigerte Kovarianzmatrix* der Variablen $X_1, ..., X_p$ bezeichnet.
Sweept man nun auf Zeile und Spalte 1, so erhält man die Matrix

$$\begin{aligned} SWP[0,1]G &= \begin{pmatrix} -(1+\bar{x}_1^2/s_{11}) & \bar{x}_1/s_{11} & \bar{x}_2 - (s_{12}/s_{11})\bar{x}_1 & \dots & \bar{x}_p - (s_{1p}/s_{11})\bar{x}_1 \\ \bar{x}_1/s_{11} & -1/s_{11} & s_{12}/s_{11} & \dots & s_{1p}/s_{11} \\ \vdots & & s_{22} - s_{12}^2/s_{11} & \dots & s_{2p} - s_{1p}s_{12}/s_{11} \\ & & & \vdots & \\ \bar{x}_p - (s_{1p}/s_{11})\bar{y}_1 & \dots & & \dots & s_{pp} - s_{1p}^2/s_{11} \end{pmatrix} \\ &= \begin{pmatrix} -A & B \\ B^T & C \end{pmatrix}, \quad \text{mit} \quad A = 2 \times 2\ ,\ ; B = 2 \times (p-1)\ ,\ C = (p-1) \times (p-1)\,. \end{aligned}$$

Diese Matrix stellt die Ergebnisse für die multivariate Regression von $X_2, ..., X_p$ auf X_1 dar. Die j-te Spalte von B enthält den Intercept und die Steigung der Regression von X_{j+1} auf X_1. Die Matrix C beschreibt die Residuenkovarianzmatrix von $X_2, ..., X_p$ für gegebenes X_1.
Im Laufe der Berechnung werden demnach die zuvor abhängigen Variablen der Regression in Prädiktorvariablen umgewandelt. Durch Sweepen auf immer mehr Zeilen und Spalten der ursprünglichen Kreuzproduktmatrix G kann man so die Parameter der multivariaten Regression für alle Variablen berechnen.

Schlussendlich ist es noch nützlich, sich einen sogenannten *Reverse Sweep-Operator* $H = RSW[k]G$ zu definieren, der die transformierte Matrix wieder in ihren ursprünglichen Zustand

zurückführen kann. Dieser ist gegeben durch

$$\begin{aligned} h_{kk} &= -\frac{1}{g_{kk}}\,, \\ h_{jk} &= h_{kj} = -\frac{g_{jk}}{g_{kk}}\,, \quad k \neq j\,, \\ h_{jl} &= h_{lj} = g_{jl} - \frac{g_{jk}g_{kl}}{g_{kk}}\,, \quad k \neq j\ ,\ k \neq l\,. \end{aligned} \tag{4.3}$$

Der einzige Unterschied zum normalen Sweep-Operator besteht demzufolge im Vorzeichen bei der Berechnung von h_{jk}. Der RSW-Operator besitzt ebenfalls die Eigenschaft der Kommutativität.

Beispiel: Sweep für multivariat normalverteilte Daten: (aus Little/Rubin (2002, S.153)

Zur Veranschaulichung des Sweep-Operators und des RSW-Operators wird ein Beispiel gegeben, wie die ML-Schätzer für Erwartungswerte und Kovarianzmatrix bei MVN-Daten mit monotonem Pattern und 3 Blöcken von Variablen gefunden werden können.

Man geht dazu in folgenden Schritten vor:

1. Man berechnet die ML-Schätzer $\hat{\mu}$ und $\hat{\Sigma}_{11}$ für den Block der vollständig beobachteten Variablen.
2. Durch Sweepen der Variablen des ersten Blocks der Variablen aus der gesteigerten Kovarianzmatrix können die ML-Schätzer für $\hat{\beta}_{20.1}$ (Intercept), $\hat{\beta}_{21.1}$ (Regressionskoeffizienten) und $\hat{\beta}_{22.1}$ (Residuenkovarianzmatrix) der Regression von X_2 auf X_1 gefunden werden. Zu dieser Schätzung verwendet man alle beobachteten Daten der ersten beiden Variablenblöcke.
3. Die ML-Schätzer der Regression von X_3 auf X_1 und X_2 finden sich durch Sweepen von X_1 und X_2 aus der gesteigerten Kovarianzmatrix. Diese Schätzung basiert auf allen beobachteten Werten.
4. Berechnung der Matrix

$$A = SWP[1]\begin{pmatrix} -1 & \hat{\mu}_1^T \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & A_{22} \end{pmatrix}.$$

5. Berechnung der Matrix

$$B = SWP[2]\begin{pmatrix} a_{11} & a_{12} & \hat{\beta}_{20.1}^T \\ a_{12} & A_{22} & \hat{\beta}_{21.1}^T \\ \hat{\beta}_{20.1}^T & \hat{\beta}_{21.1}^T & \hat{\beta}_{22.1}^T \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{12} & c_{22} & c_{23} \\ c_{13} & c_{23} & c_{33} \end{pmatrix}.$$

6. Anwendung des RSW-Operators zur Berechnung der Matrix

$$\begin{pmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{pmatrix} = RSW[1,2] \begin{pmatrix} c_{11} & c_{12} & c_{13} & \hat{\beta}_{30.12}^T \\ c_{12} & c_{22} & c_{23} & \hat{\beta}_{31.12}^T \\ c_{13} & c_{23} & c_{33} & \hat{\beta}_{32.12}^T \\ \hat{\beta}_{30.12}^T & \hat{\beta}_{31.12}^T & \hat{\beta}_{32.12}^T & \hat{\beta}_{33.12}^T \end{pmatrix}.$$

Diese Matrix enthält alle ML-Schätzer der Erwartungswerte und Kovarianzen.

4.1.2. Der EM-Algorithmus für MVN-Daten

Im Folgenden wird der EM-Algorithmus zur Lösung des Problems der Schätzung des Erwartungswertvektors und Kovarianzmatrix der Variablen einer MVN-verteilten Datenmatrix mit fehlenden Werten und generellem Fehlendmuster präsentiert.

Der Zufallsvektor $X = (X_1, ..., X_p)$ habe eine p-variate Normalverteilung mit Erwartungswertvektor $\mu = (\mu_1, ..., \mu_p)$ und eine symmetrisch und positiv definite Kovarianzmatrix Σ, mit

$$\Sigma = \begin{pmatrix} \sigma_{11} & ... & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & ... & \sigma_{pp} \end{pmatrix}.$$

Der beobachtete Datensatz $X = (X_{obs}, X_{mis})$ entspricht n unabhängigen Realisierungen dieses Zufallsvektors mit fehlenden Werten X_{mis}. Es wird implizit vorausgesetzt, dass keine Reihe komplett nicht beobachtet ist. Ein solcher Fall sollte vor Beginn des Algorithmus aus der Matrix ausgeschlossen werden, da er keine Information für die Likelihood beiträgt, aber die Konvergenz des Algorithmus verlangsamt.
Dabei seien die beobachteten Werte

$$X_{obs} = (x_{obs,1}; ...; x_{obs,n}) ,$$

bei denen $x_{obs,i}$ für die beobachteten Werte in den Fällen $i = 1, ..., n$ steht.
Dann ergibt sich nach Little/Rubin (2002, S.224) die Loglikelihoodfunktion für die Parameter μ und Σ basierend auf den beobachteten Daten

$$l(\mu, \Sigma | X_{obs}) = const - \frac{1}{2} \sum_{i=1}^{n} ln|\Sigma_{obs,i}| - \frac{1}{2} \sum_{i=1}^{n} (x_{obs,i} - \mu_{obs,i})^T \Sigma_{obs,i}^{-1} (x_{obs,i} - \mu_{obs,i}). \quad (4.4)$$

Hier bezeichnen $\mu_{obs,i}$ und $\Sigma_{obs,i}$ den Mittelwert bzw. die Kovarianzmatrix für die vorhandenen Werte aus Beobachtung i.

Wie bereits in Abschnitt 2.4 beschrieben, wird durch das Auftreten der fehlenden Daten die Struktur der Fisher-Informationsmatrix zu komplex. Deshalb können Newton-Raphson und Fisher-Scoring nicht zum Maximieren dieser Funktion verwendet werden. In den meisten Fällen könnten auch die Schätzer aus einer Complete Case Analyse als Startwerte verwendet werden. Stattdessen greift man auf den EM-Algorithmus zurück. Für diesen ist es von Vorteil, dass die Daten aus einer regulären Exponentialfamilie stammen mit den suffizienten Statistiken

$$T(X) = \left(\sum_{i=1}^{n} x_{ij} \; ; \; \sum_{i=1}^{n} x_{ij} x_{ik} \right), \quad \text{mit } j,k = 1,...,p. \tag{4.5}$$

Nun werden E-Schritt und M-Schritt des Algorithmus wieder abwechselnd bis zur Konvergenz durchgeführt. Dabei bezeichnet $\theta^{(t)} = \left(\mu^{(t)}, \Sigma^{(t)}\right)$ die aktuellen Parameterschätzungen in der t-ten Iteration.
Als Startwerte $\theta^{(0)} = \left(\mu^{(0)}, \Sigma^{(0)}\right)$ können zum Beispiel die Schätzer aus einer Datenmatrix genommen werden, die mit einem der Single-Imputationsverfahren aus Abschnitt 2.3 aufgefüllt wurde.

Der E-Schritt besteht aus der Schätzung der beiden suffizienten Statistiken

$$E\left(\sum_{i=1}^{n} x_{ij} \,|\, X_{obs}, \theta^{(t)} \right) = \sum_{i=1}^{n} x_{ij}^{(t)}, \quad j = 1,...,p, \tag{4.6}$$

sowie

$$E\left(\sum_{i=1}^{n} x_{ij} x_{jk} \,|\, X_{obs}, \theta^{(t)} \right) = \sum_{i=1}^{n} \left(x_{ij}^{(t)} x_{jk}^{(t)} + c_{ijk}^{(t)} \right), \quad j,k = 1,...,p. \tag{4.7}$$

Dabei bedeuten die Elemente

$$x_{ij}^{(t)} = \begin{cases} x_{ij}, & \text{falls } x_{ij} \text{ beobachtet.} \\ E\left(x_{ij} | x_{obs,i}, \theta^{(t)}\right), & \text{falls } x_{ij} \text{ nicht beobachtet.} \end{cases}$$

und

$$c_{ijk}^{(t)} = \begin{cases} 0, & \text{falls } x_{ij} \text{ oder } x_{ik} \text{ beobachtet.} \\ Cov\left(x_{ij}, x_{ik} | x_{obs,i}, \theta^{(t)}\right), & \text{falls } x_{ij} \text{ und } x_{ik} \text{ fehlen.} \end{cases}$$

Diese Größen werden nun mittels linearer Regression berechnet. Zu deren Durchführung bedarf es der Anwendung des Sweep-Operators auf die gesteigerte Kovarianzmatrix wie im vorigen Abschnitt beschrieben. Dann stellen die beobachteten Daten $x_{obs,i}$ die Prädiktoren der Regressiongleichungen dar, und die übrigen Variablen sind diejenigen, die es zu bestimmen gilt.

Im M-Schritt müssen nur aus den zuvor geschätzten suffizienten Statistiken die neuen

Parameterwerte bestimmt werden.

$$\mu_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^{(t)}\,, \tag{4.8}$$

$$\sigma_{jk}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[(x_{ij}^{(t)} - \mu_j^{(t+1)})(x_{ik}^{(t)} - \mu_k^{(t+1)}) + c_{ijk}^{(t)}\right]\,. \tag{4.9}$$

Nach Schafer (1997, S.166) lässt sich eine Iteration des EM-Algorithmus zusammenfassen in der Formel

$$\theta^{(t+1)} = SWP[0]n^{-1}E(T|Y_{obs},\theta^{(t)})\,. \tag{4.10}$$

4.1.3. Faktorenanalyse nach Anwendung des EM-Algorithmus

Da nach Durchführung des EM-Algorithmus die ML-Schätzer für die Parameter der multivariaten Normalverteilung vorliegen, kann man anhand der geschätzten Kovarianzmatrix nun eine explorative Faktorenanalyse durchführen.
Wie in Kapitel 3 beschrieben spaltet man dabei die geschätzten Kovarianzen in einen gemeinsamen und einen merkmalseigenen Teil auf.
Die Schätzung der Ladungsmatrix L und der Matrix Ψ der spezifischen Varianzen, sowie die Rotation der Ladungen und die Berechnung der Faktorenmatrix verläuft hier vollkommen analog zum Fall mit vollständig beobachteten Daten.
Die Anwendung des EM-Algorithmus auf einen Datensatz aus der Praxis mit anschließender Faktorenanalyse wird später in Abschnitt 4.3 noch ausführlich demonstriert.

4.2. Anwendung von multipler Imputation

Einen zweiten wichtigen Ansatz stellt die Multiple Imputation dar.
Die Grundidee von Imputationsverfahren ist es, Datenlöcher durch plausible Werte aufzufüllen. Naives Auffüllen kann jedoch zu Verzerrung von Punktschätzern und Standardfehlern führen und somit mehr Probleme schaffen als lösen.
Deshalb wurden multiple Imputationsverfahren (MI), die zuerst von Rubin vorgeschlagen und 1987 in seinem Buch ausgearbeitet wurden, als Weiterentwicklung der in Kapitel 2.3 beschriebenen Single Imputation konzipiert.
Durch das Einführen von Zufallsfehlern in den Imputationsprozess ist es möglich, approximativ unverzerrte Schätzer zu produzieren. Da die MI zudem, im Gegensatz zu den zuvor bekannten Methoden, auch eine Möglichkeit bietet das Problem des Unterschätzens der Unsicherheit der Schätzungen zu lösen, spricht Rubin von *" proper imputation"* (saubere Imputation) gegenüber der einfachen *" improper imputation"* (unsaubere Imputation).

Als Voraussetzung für MI ist ein vernachlässigbarer Fehlendmechanismus MAR nötig.
Im Allgemeinen hat sich MI aber auch als robust gegenüber Abweichungen von der Voraussetzung des MAR-Mechanismus oder gegenüber einem ungenauen Parametermodell herausgestellt.

4.2.1. Grundkonzepte der multiplen Imputation

MI zählt zu den Markov Chain Monte Carlo-Techniken (siehe Anhang D). Dabei werden die fehlenden Werte durch $m > 1$ simulierte Werte ersetzt, wodurch m vollständige Datensätze entstehen.
Jeder der m kompletten Datensätze kann nun standardmäßig statistisch ausgewertet werden. Die Ergebnisse müssen anschließend kombiniert werden, um Punktschätzer und Konfidenzintervalle zu erhalten, die die durch die Ersetzung fehlender Werte zusätzlich implizierte Unsicherheit berücksichtigen.

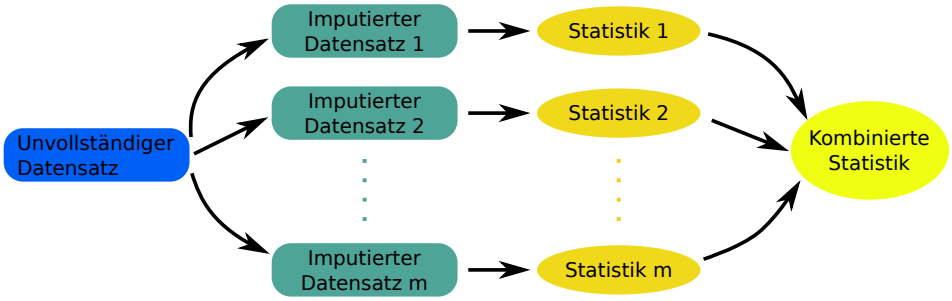

Abbildung 4.2.: Grundkonzept der multiplen Imputation mit den Schritten Imputation, Analyse und Kombination. Die Statistiken sind üblicherweise Mittelwertsschätzer und Standardfehler.

Imputation durch Data Augmentation

Die Werte, mit denen die Datenlöcher aufgefüllt werden, müssen aus einer Wahrscheinlichkeitsverteilung gezogen werden.
Dabei handelt es sich im Falle der MI um Ziehungen aus einer Posteriori-Verteilung der fehlenden Daten, $P(X_{mis} \mid X_{obs}, \theta)$. Diese resultiert aus der Annahme eines Modells für die gesamten Daten und einer Priori-Verteilung. Die Parameterwerte θ dieser bedingten Verteilung werden als nicht bekannt angenommen.
Da diese bedingte Verteilung aufgrund ihrer komplexen Struktur wiederum sehr schwierig ausgewertet werden kann, wird sie mit Hilfe der in Anhang D beschriebenen MCMC-Verfahren approximiert.

Das Problem wird nach Wahl von geeigneten Startwerten für die Parameter θ, welche etwa durch den EM-Algorithmus bestimmt werden können, in zwei sich fortwährend abwechselnde Schritte zerlegt:

1. **Imputation-Step (I-Step):** Fehlende Werte werden aus der Posteriori-Verteilung gezogen, für gegebene beobachtete Daten und Parameterwerte.

$$X_{mis}^{(t+1)} \sim P\left(X_{mis} \,|\, X_{obs}, \theta^{(t)}\right).$$

2. **Posterior-Step (P-Step):** Die Modellparameter werden simuliert, für gegebene beobachtete und zuvor gezogene fehlende Werte.

$$\theta^{(t+1)} \sim P\left(\theta \,|\, X_{obs}, X_{mis}^{(t)}\right).$$

Diese beiden Schritte, die auch als *Data-Augmentation* bezeichnet werden, wechseln sich iterativ bis zur Konvergenz der so entstehenden Markovkette ab. Die Parameterwerte konvergieren nicht gegen feste Werte, sondern gegen eine Posteriori-Verteilung $P(X_{mis}, \theta \,|\, X_{obs})$, aus der der Imputationswert gezogen werden kann.
Um gültige Inferenzschlüsse ziehen zu können, müssen die Realisationen aus $P(X_{mis} \,|\, X_{obs})$ unabhängig sein. Zur Gewährleistung dessen, simuliert man für die MI m unabhängige Ketten der Länge k und verwendet jeweils deren letzten Wert $Y_{mis}^{(k)}$. Dabei muss k groß genug sein, um Unabhängigkeit zu garantieren. Insgesamt sind demzufolge $k \cdot m$ Iterationen notwendig.
Eine andere, etwas weniger aufwändige Möglichkeit die Unabhängigkeit der gezogenen Daten sicherzustellen, ist es nur eine Markovkette zu simulieren und aus dieser Werte mit genügend großem Abstand k (z.B. $k = 50$) zu nehmen. Zur Kontrolle kann man einen Plot der Autokorrelationen der Werte betrachten, der mit größerem Abstand gegen 0 tendieren sollte. Ebenfalls als Indikator dienen sogenannte Traceplots, in denen die simulierten Werte gegen den Index abgetragen werden. Ist kein systematisches Muster zu erkennen, so kann man von Unabhängigkeit ausgehen.

Kombination und Analyse der vollständigen Datensätze

An jedem der m auf diese Art erschaffenen vollständigen Datensätze kann nun standardmäßig eine statistische Analyse durchgeführt werden. Deren Ergebnisse, wie Punktschätzer und Standardfehler für Mittelwerte oder Regressionskoeffizienten, sollen anschließend zu einem einzelnen kombiniert werden.
$\hat{Q}_j$ sei der Schätzer für die interessierende Größe aus dem imputierten Datensatz j, mit $j=(1,...,m)$, und U_j dessen zugehöriger Standardfehler. Dann werden von Rubin (1987) folgende Methoden zur Kombination der Ergebnisse vorgeschlagen: Als Gesamtschätzer wird

das arithmetische Mittel der einzelnen Schätzer herangezogen.

$$\bar{Q} = \frac{1}{m} \sum_{j=1}^{m} \hat{Q}_j. \tag{4.11}$$

Um den Gesamtstandardfehler zu erhalten müssen zunächst die Abweichungen innerhalb der Imputationen

$$\bar{U} = \frac{1}{m} \sum_{j=1}^{m} U_j \tag{4.12}$$

und die Abweichungen zwischen den Imputationen

$$B = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{Q}_j - \hat{Q})^2 \tag{4.13}$$

berechnet werden.
Als Gesamtvarianz ergibt sich dann

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B. \tag{4.14}$$

Der Gesamtstandardfehler ist folglich die Quadratwurzel aus T. Konfidenzintervalle können damit berechnet werden als

$$KI = \bar{U} \pm t_{\alpha/2,df} \cdot \sqrt{T}. \tag{4.15}$$

Dabei ist t das Quantil einer t-Verteilung mit df Freiheitsgraden,

$$df = (m-1) \left(1 + \frac{m\bar{U}}{(m+1)B}\right)^2. \tag{4.16}$$

Dieses Konzept lässt sich auch enfach auf multidimensionale Schätzer und ihre zugehörigen Kovarianzmatrizen erweitern, welche auf die gleiche Art berechnet werden.

Wahl der Anzahl m an Imputationen

Im Gegensatz zu anderen MCMC Verfahren reichen üblicherweise schon kleine Anzahlen m von *3-10* Wiederholungen für präzise Schätzungen aus. In statistischen Programmpaketen wählt man meist als Standardwert $m = 5$. Die Effizienzrate einer auf m Imputationen basierenden Schätzung beträgt nach Rubin (1987, S.114)

$$Eff = \left(1 + \frac{\gamma}{m}\right)^{-1}. \tag{4.17}$$

Dabei bezeichnet γ die von Rubin vorgeschlagene Rate der fehlenden Information, mit

$$\gamma = \frac{r + 2/(df + 3)}{r + 1}, \quad \text{wobei} \quad r = \frac{(T - \bar{U})}{\bar{U}} \; . \tag{4.18}$$

Diese dient als Maß dafür, wie stark die geschätzte Größe von den fehlenden Daten beeinflusst wird. Die enthaltene Größe r entspricht dem relativen Anstieg der Varianz bedingt durch die fehlenden Daten.
Tabelle 4.1 zeigt die erzielten Effizienzen für unterschiedliche Anzahl an Imputationen und verschiedenen Anteil an fehlender Information.
Wie man sieht ist eine geringe Anzahl an Imputationen in den meisten Fällen vollkommen

		γ				
		0.1	0.3	0.5	0.7	0.9
	3	97	91	86	81	77
m	5	98	94	91	88	85
	10	99	97	95	93	92
	20	100	99	98	97	96

Tabelle 4.1.: Effizienz (in %) von multipler Imputation für verschiedene Imputationsanzahlen und Raten für fehlende Information (aus Rubin 1987, S.115)

ausreichend. Nur bei einer sehr hohen Rate an fehlender Information lässt sich durch das Erhöhen dieser Anzahl ein entscheidender Effizienzgewinn erzielen.

4.2.2. Multiple Imputation für multivariat normalverteilte Daten

In diesem Abschnitt wird das Imputationsmodell explizit für den in dieser Arbeit betrachteten Fall multivariat normalverteilter Daten erläutert.
Ein p-dimensionaler Zufallsvektor $X = (X_1, ..., X_p)^T$ habe den Erwartungswertvektor $\mu = (\mu_1, ..., \mu_p)^T$ und die Kovarianzmatrix $\Sigma = (\sigma_{jk})$. Dann heißt eine Beobachtung $x = (x_1, ..., x_p)^T$ multivariat normalverteilt (MVN), wenn gilt

$$f(x \,|\, \mu, \Sigma) \propto |\Sigma|^{-1/2} \cdot \exp\left\{ -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right\}. \tag{4.19}$$

Daraus ergibt sich bei unbekanntem Parametervektor $\theta = (\mu, \Sigma)$ die Likelihoodfunktion für komplette Daten

$$L(\theta \mid X) \propto |\Sigma|^{-n/2} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right\} \tag{4.20}$$

$$= |\Sigma|^{-n/2} \cdot \exp\left\{-\frac{1}{2}tr(\Sigma^{-1}S_0)\right\}. \tag{4.21}$$

Dabei ist $S_0 = \sum_{i=1}^{n}(x_i-\mu)^T(x_i-\mu)$. Zum Beweis der Umformung von Formel 4.10 zu 4.11 siehe unter anderem Fahrmeir u.a. (1996, S.57)

Für unbekanntes μ und Σ verwendet man in der Regel als konjugierte Priori-Verteilung eine *normal- inverse-Wishart-Verteilung.*

Exkurs: Die inverse Wishart-Verteilung:
Sei X eine $m \times p$-Matrix mit Verteilung $X \sim MNV_p(0, \Lambda)$. Dann hat die Matrix der Quadratsummen $A = X^TX$ eine Wishart-Verteilung, $A \sim W(m, \Lambda)$. Dabei steht m für die Freiheitsgrade der Verteilung.
In diesem Zusammenhang ergibt sich für $B = A^{-1}$ die inverse Wishart-Verteilung, mit $B \sim W^{-1}(m, \Lambda)$.
Die Dichte dieser Größe stellt sich dar als

$$f(B \mid m, \Lambda) \propto |B|^{-(\frac{m+p+1}{2})} \cdot \exp\left\{-\frac{1}{2}tr(\Lambda^{-1}B^{-1}\right\}, \quad \text{für } B > 0. \tag{4.22}$$

Die Verteilung besitzt die Eigenschaften

$$E(B \mid m, \Lambda) = \frac{1}{m-p-1}\Lambda^{-1} \quad \text{und} \tag{4.23}$$

$$mod(B \mid m, \Lambda) = \frac{1}{m+p+1}\Lambda^{-1}. \tag{4.24}$$

Auf das Problem der Priori-Verteilung zurückkommend erhält man damit für

$$\mu|\Sigma \sim MVN(\mu_0, \kappa_0^{-1}\Sigma) \quad \text{und} \quad \Sigma \sim W^{-1}(m, \Lambda_0^{-1}) \tag{4.25}$$

die gemeinsame normal-inverse-Wishart-Verteilung

$$\theta = (\mu, \Sigma) \sim MVN - W^{-1}\left(\mu_0, \frac{1}{\kappa_0}; m_0, \Lambda_0^{-1}\right). \tag{4.26}$$

Daraus ergibt sich die gemeinsame Prioridichte

$$p(\mu, \Sigma) \propto |B|^{-(\frac{m_0+p+1}{2})} \cdot \exp\left\{-\frac{1}{2}\left(tr(\Lambda_0\Sigma^{-1}) - \kappa_0(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right)\right\}. \tag{4.27}$$

Dabei steht m_0 wieder für die Anzahl der Priori-Freiheitsgrade und κ_0 für die Priori-Anzahl an Messungen auf der Σ-Skala.
Damit gelangt man zur Posteriori-Verteilung von $\theta = (\mu, \Sigma)$:

$$\theta \,|\, x \propto \; MVN - W^{-1}(\mu_n, \kappa_n^{-1}\Lambda; m_n, \Lambda_n^{-1}). \tag{4.28}$$

Diese hat nach Gelman (2006, S.87) die Eigenschaften

$$\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{x}\,, \tag{4.29}$$

$$\kappa_n = \kappa_0 + n\,, \tag{4.30}$$

$$m_n = m_0 + n\,, \tag{4.31}$$

$$\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{x}-\mu_0)(\bar{x}-\mu_0)^T\,. \tag{4.32}$$

Der Posteriori-Erwartungswert ist ein gewichtetes Mittel aus Stichprobenmittelwert und Priori-Erwartungswert. Die Posteriori-Kovarianzmatrix lässt sich in drei Aspekte aufteilen: In Priori-Kovarianz, empirische Streuungsmatrix und Streuung zwischen Priori-Erwartungswert und Stichprobenmittel.

Ebenfalls möglich ist es, als noninformative Priori eine multivariate Version der Jeffrey's-Priori (siehe Anhang D) zu wählen,

$$p(\mu, \Sigma) \propto |\Sigma|^{-(p+1)/2}. \tag{4.33}$$

Diese entspricht dem Fall $\kappa_0 \to 0,\ m_0 \to -1,$ und $|\Lambda_0| \to 0$. Damit kann die entsprechende Posteriori-Verteilung geschrieben werden als

$$\Sigma|x \propto W^{-1}(n-1, S)\,, \tag{4.34}$$

$$\mu|\Sigma, x \propto N(\bar{x}, \Sigma/n)\,. \tag{4.35}$$

Aus der Posteriori-Verteilung sollen nun die Werte für die multiple Imputation gezogen werden. Dies geschieht mit Hilfe des zuvor beschriebenen Data-Augmentation-Algorithmus.
Dabei nimmt man als Startwerte für die unbekannten Parameter $\theta = (\mu, \Sigma)$ der Verteilung die ML-Schätzer aus den beobachteten Daten. Anschließend führt man iterativ den I-Schritt, mit Ziehen der unbeobachteten Werte bedingt auf die beobachteten Werte und die aktuellen Parameterschätzer, und den P-Schritt, mit neuem Schätzen der Parameterwerte der

Posterioriverteilung, bis zur Konvergenz aus. Den letzten Wert der Markovkette für X_{mis} nimmt man dann als Imputationswert für die fehlenden Werte.
Nach m-maligem Ausführen dieses Algorithmus erhält man somit m vollständige Datensätze, an denen sich die weitere statistische Analyse durchführen lässt.

In der Realität passen Daten nur selten genau in die Annahme einer Normalverteilung. Daher ist das Modell, das die imputierten Daten generiert, oftmals nur approximativ gültig.
MI hat sich jedoch zumindest bei geringem Anteil an fehlender Information als sehr robust gegenüber Abweichungen vom Imputationsmodell erwiesen.
Beispielsweise lässt sich die Normalverteilungsannahme auch für binäre oder ordinale Daten anwenden, indem man anschließend die imputierten Werte zur nähesten Kategorie rundet. Eine andere Möglichkeit wäre zudem, Variablen zu transformieren, wie etwa durch Logarithmieren, um eine Normalverteilung zu approximieren, und nach der Imputation wieder zurückzutransformieren.

4.2.3. Kombination der Datensätze zu einer Faktorenanalyse

Nachdem man m vollständige Datensätze auf die zuvor beschriebene Art erhalten hat, kann man aus jedem die zugehörigen Erwartungswerte und die Kovarianzmatrix schätzen. Diese müssen im Anschluss wie zuvor beschrieben zu einer gemeinsamen Statistik kombiniert werden. Der Vorteil dieser kombinierten Statistiken ist es, dass man zusätzlich noch die Unsicherheiten der Schätzer erhält.
Die Kombination ergibt als Schätzer $\hat{\mu}_{MI}$ für den Erwartungswertvektor den Mittelwert der einzelnen Erwartungswertvektoren $\hat{\mu}^{(j)}$

$$\hat{\mu}_{MI} \;=\; (\hat{\mu}_{1,MI}\,;...;\,\hat{\mu}_{p,MI}) \;=\; \frac{1}{m}\sum_{j=1}^{m}\hat{\mu}^{(j)}\,. \tag{4.36}$$

Die Kovarianzmatrixschätzung $\hat{\Sigma}_{MI}$ setzt sich zusammen aus einer Kombination des Mittelwerts der einzelnen geschätzten Kovarianzmatrizen und der durch die Imputation erzeugten zusätzlichen Varianz zwischen den imputierten Datensätzen,

$$\hat{\Sigma}_{MI} \;=\; \bar{S} + (1 + m^{-1})B\,. \tag{4.37}$$

Dabei entsprechen die Größen

$$\bar{S} \;=\; \frac{1}{m}\sum_{j=1}^{m}\Sigma^{(j)} \quad \text{und} \quad B \;=\; \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\mu}^{(j)} - \hat{\mu}_{MI})(\hat{\mu}^{(j)} - \hat{\mu}_{MI})^T\,. \tag{4.38}$$

Mit der kombinierten Kovarianzmatrix $\hat{\Sigma}_{MI}$ kann nun wieder analog zum Fall mit vollständigen

Daten eine Faktorenanalyse ausgeführt werden.
Dies wird anschließend ebenfalls noch an einem ausführlichen Beispiel demonstriert.

4.3. Anwendungsbeispiel

Zur Veranschaulichung der praktischen Ausführung einer Faktorenanalyse bei Vorliegen von fehlenden Daten wird wieder der zuvor schon analysierte Zehnkampfdatensatz herbeigezogen.
Es wurden hierfür nun 10% der Daten zufällig aus der Datenmatrix entfernt, so dass der Fehlendmechanismus ignorierbar ist.

Zuerst kann man versuchen, die unvollständigen Daten durch Complete Case Analyse zu untersuchen. Da man dabei allerdings alle Reihen eliminiert, die fehlende Werten enthalten, bleiben in diesem Fall nur noch 29 der ursprünglich 96 Beobachtungen übrig. Daraus resultiert ein enormer Informationsverlust.
Eine Faktorenanalyse, bei der nur die 29 vollständig beobachteten Reihen berücksichtigt werden, führt mit der ML-Methode zu einem Modell mit 2 Faktoren, was einen erheblichen Unterschied zum 4-Faktorenmodell für den ursprünglichen Datensatz darstellt.

Für die Durchführung des EM-Algorithmus ist es in der Praxis hilfreich, zunächst die unvollständige Datenmatrix zu transformieren. Im Programmpaket *R* lässt sich dies mit der Funktion *prelim.norm* durchführen, welche die Zeilen nach gleichem Fehlendmuster ordnet, sowie die Daten standardisiert.
Darauf berechnet man iterativ die E- und M-Schritte des Algorithmus bis zur Konvergenz. Dies vollzieht die Funktion *em.norm*, die für den unvollständigen Zehnkampfdatensatz nach 14 Iterationen zu einem Ergebnis konvergiert.
Mit den dadurch erhaltenen ML-Schätzern für die fehlenden Werte, sowie den ML-Schätzern für die Verteilungsparameter führt man nun ganz analog zum Vorgehen im Praxisbeispiel aus Abschnitt 3.7 die Faktorenanalyse durch.
Wiederum ergibt diese mit der ML-Methode ein 4-Faktorenmodell, welches mit einem p-Wert von 0.857 ganz deutlich nicht mehr abgelehnt werden kann. Das 3-Faktorenmodell hingegen wurde aufgrund seines p-Werts von 0.0284 abgelehnt, da dieser unter dem Signifikanzniveau von 0.05 liegt.
Die nach der Varimaxmethode rotierte Ladungsmatrix findet sich in Tabelle 4.2. Darin kann man lediglich geringe Unterschiede zur Rotationsmatrix aus dem Beispiel mit vollständigen Daten erkennen.

Der gleiche Zehnkampfdatensatz mit 10% fehlenden Daten kann auch durch multiple Imputation in eine Form gebracht werden, mit der eine Faktorenanalyse praktiziert werden kann.
Zunächst wendet man den Data-Augmentation-Algorithmus an, um Realisationen für die

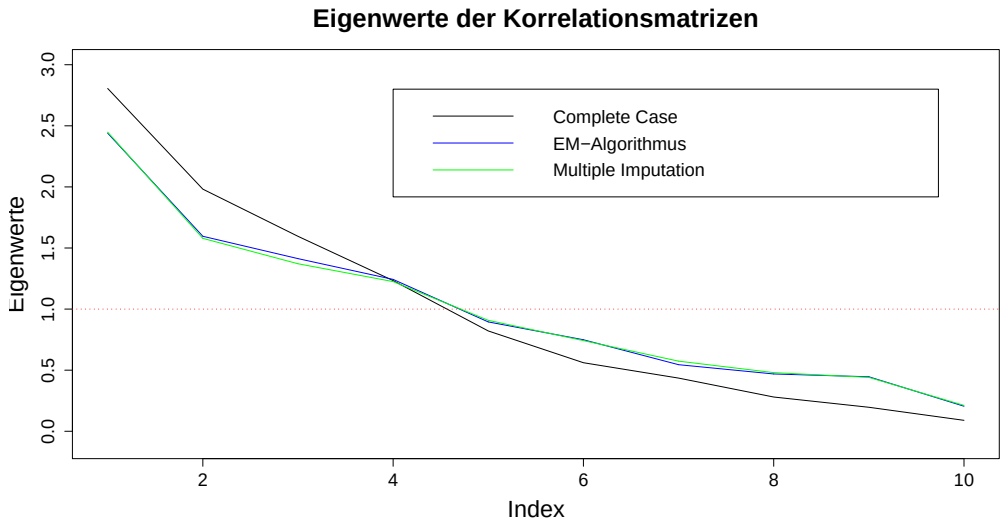

Abbildung 4.3.: Screeplots der Eigenwerte für die unterschiedlichen Herangehensweisen

fehlenden Werte, sowie Schätzer für die Parameter der multivariaten Normalverteilung der Daten zu erhalten.
Bei diesem noch relativ geringen Anteil an fehlender Information genügen $m = 5$ Imputationen. Man simuliert demzufolge 5 Markovketten einer Länge k und verwendet deren letzte Werte als Realisation für die fehlenden Beobachtungen.
Als Startwerte für den Data-Augmentation-Algorithmus können die durch den EM-Algorithmus erhaltenen Parameterschätzer genommen werden.
Die Imputation kann im Programmpaket *R* beispielsweise mit der Funktion *amelia* ausgeführt werden, welche die gewünschte Anzahl an imputierten Datensätzen erzeugt.

Anhand der durch Kombination der einzelnen Datensätze entstandenen Kovarianzmatrix führt man nun die weiteren Schritte der Faktorenanalyse durch.
Wie schon nach Betrachtung der nahezu identischen Eigenwerte (Abbildung 4.3) zu vermuten ist, weicht das Ergebnis dieser Analyse nur geringfügig von demjenigen der EM-Algorithmus-Methode ab.

Tabelle 4.3 stellt die resultierenden Ladungsmatrizen nach Varimaxrotation gegenüber. In den ersten beiden Fällen gibt es für die einzelnen Ladungskoeffizienten nur in wenigen Fällen geringe Unterschiede. Dabei werden beide Male die Disziplinen Weitsprung, Kugelstoßen und 400m-Lauf fast komplett vom Modell beschrieben und haben demnach nur noch geringfügigen spezifische Varianzen. Andere Disziplinen, wie Diskuswerfen, Stabhochsprung und Speerwerfen laden dagegen nur wenig mit den extrahierten Faktoren, und werden demzufolge noch von

	EM-Algorithmus					Multiple Imputation					CCA		
Var.	Fa1	Fa2	Fa3	Fa4	Kom	Fa1	Fa2	Fa3	Fa4	Kom	Fa1	Fa2	Kom
100m	0.53	0.14	0.22		0.34	0.51	0.14	0.25		0.34	0.35		0.13
Weit	0.84	0.10	-0.11	0.51	0.99	0.84			0.52	0.99	0.61	0.79	0.99
Kugel	0.30	-0.10	0.95		0.99	0.32		0.94		0.99	0.33		0.11
Hoch	0.52	0.14			0.29	0.53	0.13			0.29	0.68	0.32	0.56
400m	0.22	0.97	0.12		0.99	0.22	0.97			0.99		0.74	0.55
110mH	0.70	-0.29	0.12		0.60	0.70	-0.26	0.14		0.58	0.99	-0.14	0.99
Diskus		0.20	0.35	-0.13	0.19		0.22	0.32	-0.12	0.17	-0.42		0.17
Stab			0.12	-0.42	0.21		0.11	0.11	-0.43	0.11		-0.22	0.05
Speer		-0.25	0.22	0.39	0.26		-0.20	0.25	0.33	0.11	0.11	0.26	0.08
1500m		0.19		0.61	0.42		0.17		0.63	0.43	-0.22	0.63	0.44

Tabelle 4.2.: Ladungsmatrizen nach der Varimax-Rotation für die verschiedenen Herangehensweisen. Nur Werte, die betragsmäßig größer sind als 0.10, wurden berücksichtigt.

anderen Einflüssen mitbestimmt.

Anhand der Ergebnisse kann vermuten, dass die beiden Methoden für solch einen relativ geringen Anteil fehlender Daten noch stets zu ähnlichen Ergebnissen führen.
Lediglich das Resultat aus der Complete Case Analysis weicht beträchtlich von denen der anderen Methoden ab, da aufgrund der geringen Anzahl verbliebener Beobachtungen schon das 2-Faktorenmodell mit einem p-Wert von 0.149 nicht mehr als signifikant abgelehnt werden konnte.

Bis zu welchem Anteil an fehlenden Daten sich die Ergebnisse von Faktorenanalysen nach Anwendung des EM-Algorithmus bzw. nach multipler Imputation so ähnlich verhalten, und wie sich die Streuung der Ergebnisse verhält, wird nun im Anschluss in Kapitel 5 untersucht.

5. Sensitivitätsanalysen

In diesem Kapitel soll untersucht werden, inwieweit sich die Ergebnisse von Faktorenanalysen nach Anwendung der beschriebenen Methoden voneinander unterscheiden und wie sich die Unsicherheiten der Ergebnisse verhalten. Diese Analyse wird für verschiedene Anteile von fehlenden Daten durchgeführt.

Im Folgenden werden nun die Auswirkungen steigender Anteile an fehlenden Daten auf die Eigenwerte der Korrelationsmatrizen analysiert. Sowohl für den EM-Algorithmus als auch für multiple Imputation soll dabei gezeigt werden, wie sich die Standardabweichung der Eigenwerte, die ein wichtiger Indikator für die Faktorenanalyse sind, verändert. Dies geschieht erneut anhand des zuvor bereits beschriebenen Zehnkampfdatensatzes.

5.1. Sensitivität der Eigenwerte

Die entscheidende Größe für die Faktorenanalyse ist die Kovarianzmatrix, welche nach dem Fundamentaltheorem der Faktorenanalyse in einen gemeinsamen und einen spezifischen Teil zerlegt werden soll. Als Indikator für eine Faktorenanalyse gelten daher die Eigenwerte der Korrelationsmatrizen, die auch die entscheidende Rolle in der Hauptkomponentenmethode (Abschnitt 3.3.2) spielen.
Deshalb soll nun die Auswirkung von verschiedenen Anteilen an fehlenden Werten auf die Eigenwerte untersucht werden.

5.1.1. Auswirkungen des EM-Algorithmus auf die Eigenwerte

Zur Untersuchung des Verlaufs der Eigenwerte wird wieder der bereits bekannte Zehnkampfdatensatz herbeigenommen, welcher als multivariat normalverteilt angesehen werden kann.

Aus dem in Abschnitt 3.7 analysierten vollständigen Datensatz werden nun mit verschiedenen Wahrscheinlichkeiten Werte zufällig eliminiert. Dies geschieht jeweils 100mal mit den

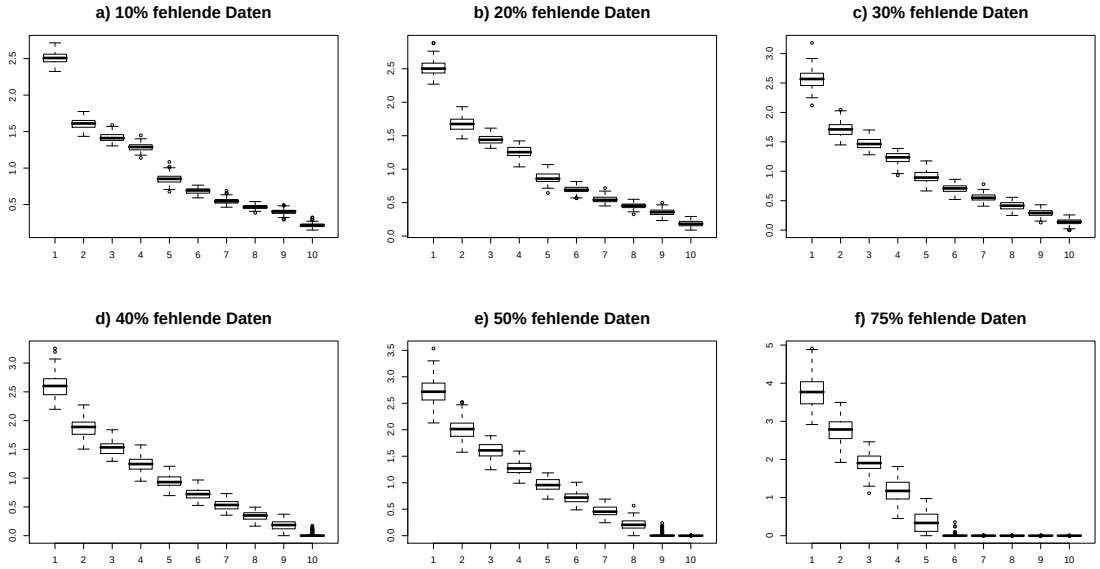

Abbildung 5.1.: Vergleich der Entwicklung der Boxplots der Korrelationsmatrizen-Eigenwerte nach Anwendung des EM-Algorithmus für verschiedene Anteile von fehlenden Daten.

Wahrscheinlichkeiten 10%, 20%, 30%, 40%, 50% und 75%, so dass für die 6 unterschiedlichen Fehlendanteile je 100 unvollständige Zehnkampfdatensätze entstehen.
Auf jeden dieser Datensätze wird nun der EM-Algorithmus zur Schätzung der Verteilungsparameter angewandt. Aus den erhaltenen ML-Schätzern können dann jeweils die 10 Eigenwerte der Korrelationsmatrix berechnet werden.
Für die 6 verschiedenen Fälle sind in Abbildung 5.1 jeweils die Boxplots für alle Eigenwerte dargestellt. Mittelwerte und Standardabweichungen der jeweils 3 größten Eigenwerte pro Kategorie sind in Tabelle 5.1 aufgelistet.

Dabei lässt für 10% fehlende Daten anhand der kleinen Boxen erkennen, dass in allen Fällen die resultierenden Eigenwerte nahezu gleich sind, mit wenigen geringfügigen Ausreißern. Der Eigenwert etwa streut nur sehr wenig um den Mittelwert 2.509. Mit steigendem Anteil fehlender Daten steigt aber auch diese Streuung. Während bei einem Anteil von 20% noch fast die gleiche Struktur der Eigenwerte festzustellen ist, werden schon ab 30% die ersten Eigenwerte im Mittel immer größer mit ebenfalls deutlich gestiegenem Standardfehler. Die letzten Eigenwerte dagegen verkleinern sich nach und nach. Dieser Trend setzt sich für ansteigenden Fehlendanteil immer weiter fort. Für 75% schließlich unterscheiden sich nur noch 5 der Eigenwerte von 0.

Die Konsequenz für die Faktorenanalyse ist demnach, dass sie für mehr fehlende Werte deutlich

	1.Eigenwert		2.Eigenwert		3.Eigenwert	
	mean	std.error	mean	std.error	mean	std.error
10% fehlend	2.509	0.074	1.608	0.065	1.419	0.055
20% fehlend	2.517	0.116	1.678	0.099	1.441	0.063
30% fehlend	2.561	0.159	1.724	0.125	1.472	0.090
40% fehlend	2.609	0.208	1.880	0.162	1.533	0.129
50% fehlend	2.735	0.246	2.010	0.186	1.604	0.142
75% fehlend	3.766	0.437	2.766	0.310	1.920	0.269

Tabelle 5.1.: Mittelwerte und Standardfehler der jeweils 3 größten Eigenwerte nach Anwendung des EM-Algorithmus.

größeren Unsicherheiten ausgesetzt ist.
Zudem ist noch eine deutliche Tendenz zu einer kleineren Faktorenzahl zu erkennen, da bereits die ersten Eigenwerte bei großem Fehlendanteil einen Großteil der Varianz auf sich vereinigen. Dementsprechend führt die Faktorenanalyse bei höheren Anteilen an fehlenden Werten aufgrund der verlorenen Information zu verschiedenen Ergebnissen.

5.1.2. Auswirkungen von multipler Imputation auf die Eigenwerte

Nach dem gleichen Konzept zur Untersuchung der Eigenwerte wird nun nochmals für multiple Imputation vorgegangen. Wiederum werden für die 6 verscgiedenen Fehlendwahrscheinlichkeiten jeweils 100 unvollständige Datensätze erzeugt. Diese werden alle mit multipler Imputation aufgefüllt, analysiert und anschließend zu einem Ergebnis kombiniert. Dann betrachtet man sich wieder die Eigenwerte der entstandenen 100 Korrelationsmatrizen pro Kategorie.

Abbildung 5.2 zeigt dabei wieder die Boxplots der Eigenwerte und Tabelle 5.2 stellt die Mittelwerte und Standardfehler der jeweils 3 größten Eigenwerte gegenüber.
Auch hier streuen die Eigenwerte für die Kategorie mit der niedrigsten Fehlendwahrscheinlichkeit von 10% nur geringfügig um die Mittelwerte. Bei Anstieg dieser Wahrscheinlichkeit ist wiederum eine Erhöhung der Standardfehler festzustellen. Allerdings bleibt die Vergrößerung der ersten Eigenwerte aus. Die Mittelwerte der Eigenwerte unterscheiden sich über alle Kategorien kaum.

Im Vergleich zu den Ergebnissen des EM-Algorithmus sind zunächst für 10% fehlender Daten weder bei den Mittelwerten noch bei den Standardfehlern Unterschiede zu erkennen.
In den weiteren Kategorien steigen für beide Methoden die Standardfehler der Eigenwerte

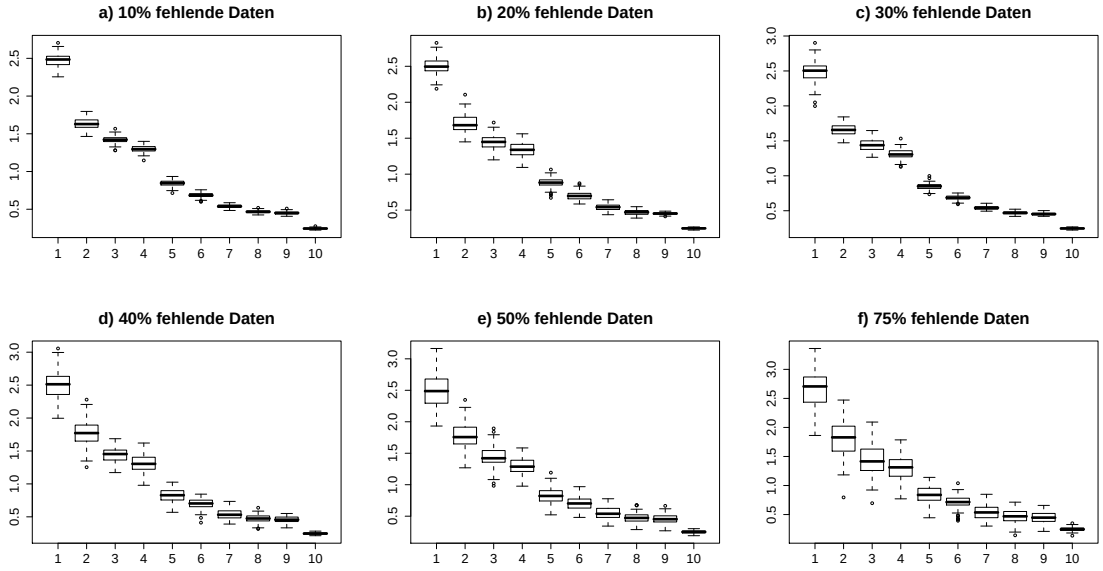

Abbildung 5.2.: Vergleich der Entwicklung der Boxplots der Korrelationsmatrizen-Eigenwerte nach Anwendung von multipler Imputation für verschiedene Anteile von fehlenden Daten.

nahezu gleichmäßig an. Der Unterschied der Resultate für die beiden Methoden liegt dafür in den Mittelwerten. Während für den EM-Algorithmus eine Tendenz zu größeren ersten Eigenwerten festzustellen ist, hält die multiple Imputation die Mittelwerte der Eigenwerte auch für größere Fehlendwahrscheinlichkeiten nahezu konstant.

Anhand dieser Ergebnisse lässt sich sagen, dass der EM-Algorithmus nur bis zu einer Fehlendwahrscheinlichkeit von ca. 30% angewandt werden sollte. Für größere Anteile an fehlenden Werten liefert nur noch die multiple Imputation konsistentere Ergebnisse.
Allerdings sei darauf hingewiesen, dass auch für multiple Imputation der Standardfehler stark ansteigt. Diese können jedoch durch Erhöhen der Anzahl m an Imputationen noch etwas verringert werden.
Die Unsicherheiten in der Schätzung der Kovarianzmatrix und der daraus resultierenden Eigenwerte führen demzufolge auch zu größeren Unsicherheiten in der Durchführung der Faktorenanalyse.

	1.Eigenwert		2.Eigenwert		3.Eigenwert	
	mean	std.error	mean	std.error	mean	std.error
10% fehlend	2.492	0.088	1.619	0.071	1.423	0.054
20% fehlend	2.494	0.113	1.690	0.117	1.437	0.080
30% fehlend	2.469	0.178	1.670	0.122	1.428	0.102
40% fehlend	2.507	0.180	1.752	0.173	1.430	0.118
50% fehlend	2.530	0.228	1.733	0.216	1.414	0.129
75% fehlend	2.523	0.289	1.778	0.294	1.399	0.154

Tabelle 5.2.: Mittelwerte und Standardfehler der jeweils 3 größten Eigenwerte nach Anwendung von multipler Imputation.

5.2. Fazit der Analyse

Als Fazit der Sensitivitätsanalysen lässt sich sagen, dass sich die multiple Imputation als wesentlich robuster gegenüber größeren Anteilen von fehlenden werten erweist.
Während das Vorgehen nach der Complete Case Analyse schon bei geringer Anzahl an fehlenden Werten einem erheblichen Informationsverlust ausgesetzt ist und damit abweichende Ergebnisse produziert, liefert das Vorgehen nach dem EM-Algorithmus zumindest bis circa 30% fehlender Daten brauchbare Ergebnisse. Für größere Prozentzahlen ist allerdings zu erkennen, dass sich die Eigenwerte dahingehend verändern, dass die ersten Werte immer größer werden und die kleineren Werte im Gegensatz dazu gegen 0 tendieren.

Multiple Imputation liefert im Mittel auch für große Fehlendanteile noch konsistente Ergebnisse. Der höhere Anteil an fehlenden Daten schlägt sich hier lediglich in einem größeren Standardfehler der Ergebnisse nieder.
Demzufolge kann die multiple Imputation als das stabilste Verfahren für die Faktorenanalyse im Umgang mit fehlenden Werten angesehen werden.

6. Zusammenfassung und Ausblick

Im Folgenden sind in einem Überblick die wichtigsten Punkte dieser Arbeit noch einmal aufgeführt.
Abschließend wird dann noch ein Ausblick gegeben über Erweiterungen dieser Ergebnisse auf nicht-vernachlässigbare Fehlendmechanismen oder nicht-normalverteiltes Datenmaterial.

6.1. Kernaussagen der Arbeit

- Die Faktorenanalyse ist eine beliebte Methode zur Dimensionsreduzierung in multivariaten Datensätzen. Dabei versucht man anhand der Korrelationsstruktur der Daten eine kleine Anzahl latenter Variablen, sogenannte Faktoren, zu finden, die die beobachteten Variablen erklären.
- Die Kovarianzmatrix wird dabei in einen gemeinsamen und einen variablenspezifischen Teil zerlegt. Die sogenannte Ladungsmatrix gibt an, wie stark die einzelnen Variablen mit den Faktoren korrelieren. Die Kommunalitäten sagen dagegen aus, welcher Anteil der Varianz einer Variable durch das Faktorenmodell erklärt wird. Der restliche Teil wird durch die spezifische Varianz beschrieben.
- Zur besseren Interpretation wird die Ladungsmatrix im Anschluss noch rotiert, um bessere Zuordnung der Variablen zu einzelnen Faktoren zu ermöglichen. Da es dafür eine unendliche Anzahl möglicher Rotationen gibt, und die Wahl der Rotation daher Heuristiken enthält, gilt dies bei vielen Wissenschaftlern als Hauptkritikpunkt an der Faktorenanalyse.
- Bei Vorliegen von fehlenden Daten kann die Faktorenanalyse nicht mehr ohne Weiteres durchgeführt werden. Man muss zuvor Methoden anwenden, welche auf geeignete Art und Weise die Verteilungsparameter der Daten schätzen. Dabei ist für die Faktorenanalyse vor allem die Kovarianzmatrix wichtig.
- Der EM-Algorithmus ist ein likelihoodbasiertes Verfahren, das iterativ bis zur Konvergenz anhand der Likelihood der beobachteten Daten Schätzer für die Verteilungsparameter sowie für die fehlenden Werte berechnet.

- Multiple Imputationsverfahren generieren m vollständige Datensätze durch mehrfaches Auffüllen der Löcher in der Datenmatrix. Die meist angewandte Methode dafür ist der sogenannte Data Augmentation Algorithmus. Dieser zieht iterativ bis zur Konvergenz Realisationen für die fehlenden Daten aus einer Posteriori-Verteilung und schätzt damit die Parameter dieser Verteilung neu.
  Die m kompletten Datensätze können im Anschluss mit statistischen Standardverfahren einzeln analysiert und zu einem Gesamtergebnis kombiniert werden. Auf diese Weise berücksichtigt man auch die durch Imputation zusätzlich induzierte Unsicherheit der Parameterschätzungen.
- Die Ergebnisse der Faktorenanalyse für die beiden verschiedenen Herangehensweisen sind bei einem geringen Anteil an fehlenden Werten ($\leq 20\%$) noch relativ ähnlich.
- Für größeren Anteil fehlender Werte ist nach Anwendung des EM-Algorithmus ein Trend zu größeren ersten Eigenwerten zu erkennen. Dies bedeutet für die Faktorenanalyse, dass schon wenige Faktoren einen Großteil der Varianz auf sich vereinigen. Mit steigendem Fehlendanteil steigen ebenfalls die Standardfehler der Eigenwerte, was größere Unsicherheiten für die Faktorenanalyse mit sich bringt. Für multiple Imputation liegt dieser Anstieg der Unsicherheit in gleichem Maße vor. Allerdings bleiben die Mittelwerte der Eigenwerte auch für größere Anteile an fehlenden Daten nahezu konstant.
- Die multiple Imputation erweist sich als das stabilste Verfahren zur Behandlung fehlender Daten in der Faktorenanalyse. Der EM-Algorithmus findet allerdings dabei noch in der Form seine Verwendung, dass die aus ihm resultierenden Parameterschätzer als Startwerte für den Data Augmenation Algorithmus genommen werden können.

6.2. Mögliche Erweiterungen der Ergebnisse

Die Untersuchungen zu den fehlenden Daten bezogen sich in dieser Arbeit nur auf multivariat normalverteiltes Datenmaterial mit ignorierbarem Fehlendmechanismus MAR. Für Abweichungen von diesem relativ speziellen Fall müssten die Ergebnisse noch erweitert und verallgemeinert werden.

Sind die Daten nicht mehr normalverteilt, so sollten sich der EM-Algorithmus und auch multiple Imputation dennoch meist noch anwenden lassen.
Der EM-Algorithmus kann völlig analog zum Normalverteilungsfall auch für andere Verteilungen durchgeführt werden. Speziell für Verteilungen, die zu einer Exponentialfamilie gehören, ist er meist sehr vorteilhaft.
Die multiple Imputation für normalverteilte Daten generiert bei Verletzung der

Normalverteilungsannahme nach Schafer (1997, S.217) meist zumindest noch approximativ gültige Modelle. Demnach hat sie sich als robust gegenüber Abweichungen vom Imputationsmodell erwiesen, solange der Anteil an fehlenden Daten nicht sehr hoch ist. So kann man sie etwa auch bei binären oder ordinalen Variablen anwenden, indem man die imputierten Werte zur nähesten Kategorie rundet.
Die Vorgehensweise mit robuster Schätzung für nicht normalverteilte Daten wird unter anderem beschrieben in Little/Rubin (2002, Kap.12+14).

Die etwas schwierigere Erweiterung des Modells ist nötig für systemmatische, und damit nicht-ignorierbare Fehlendmechanismen. Durch einfaches naives Imputieren von Werten, ohne Berücksichtigung des Ausfallmechanismus, würden die Ergebnisse der Analysen, wie etwa einer Faktorenanalyse, unter Umständen stark verzerrt werden.
In diesem Fall muss also zuerst der Fehlendmechanismus genau untersucht werden, um konkrete Annahmen über diesen aufstellen zu können. Diese Annahmen müssen dann anschließend in das Imputationsmodell aufgenommen werden. Je mehr Annahmen allerdings getroffen werden müssen, desto unsicherer wird das Modell.

Bei einer Faktorenanalyse für unvollständige Daten mit nicht-ignorierbarem Fehlendmechanismus kann demnach mit erheblichen Unsicherheiten in den Ergebnissen gerechnet werden, die mit steigendem Anteil fehlender Werte noch stärker zunehmen.
Zum Vorgehen bei Fehlendmechanismus not missing at random werden zum Beispiel bei Little/Rubin (2002, Kap.15) verschiedene Modelle, wie etwa spezielle Likelihoodmodelle, vorgeschlagen.

A. Statistische Software und benutzte Funktionen

Die Anwendungsbeispiele und Simulationen in dieser Arbeit wurden mit dem statistischen Programmpaket R (R Development Core Team, 2009) programmiert. Der zugehörige Code liegt der Arbeit auf CD bei.
In diesem Kapitel des Anhangs werden die wichtigsten implementierten Funktionen für die Faktorenanalyse bzw. für die Behandlung fehlender Daten beschrieben.
Ausführliche Beschreibungen findet man des Weiteren noch in den entsprechenden Online-Hilfen (Befehl: *help( )*). In geschweiften Klammern ist das Paket, in welchem die Funktion enthalten ist, angegeben.

factanal{stat}: Mit Hilfe dieser Funktion kann eine explorative Faktorenanalyse durchgeführt werden. Neben einem Datensatz oder einer Kovarianzmatrix muss dabei die Anzahl der zu schätzenden Faktoren übergeben werden. Die Ladungsmatrix wird nach der Maximum-Likelihood-Methode geschätzt und das Ergebnis kann mit $rotation$ = "$varimax$" (default) nach der Varimax-Methode und mit $rotation$ = " $promax$ " schiefwinklig rotiert werden. Standardmäßig werden keine Faktorenwerte für die Beobachtungenen geschätzt. Mit der Einstellung $scores$ = "$Bartlett$" werden die Faktorwerte nach der Maximum-Likelihood-Methode geschätzt, mit $scores$ = " $regression$ " nach der Regressionsmethode. Teilweise tritt die Fehlermeldung auf, dass Optimierung bei den vorgegebenen Startwerten fehlschlägt. In diesem Fall müssen mit der Einstellung $control = list(nstart = x)$ neue Startwerte ausprobiert werden.

amelia{Amelia}: Die Funktion führt multiple Imputation durch. Der unvollständige Datensatz wird in die Funktion eingesetzt, welche dann m vollständige Datensätze (Standard: $m = 5$) ausgibt.

em.norm{norm}: Berechnet die Maximum-Likelihood-Schätzungen der Parameterwerte von unvollständigen Datensätzen mit dem EM-Algorithmus.

imp.norm{norm}: Führt multiple Imputation unter der Annahme einer multivariaten Normalverteilung aus.

prelim.norm{norm}: Führt für den EM-Algorithmus oder für Imputation vorbereitende Transformationen der unvollständigen Datenmatrix durch. Unter anderem werden die beobachteten Daten standartisiert. Des Weiteren werden die Zeilen nach ihrem Fehlendmuster sortiert und der Code der fehlenden Daten wird zu NA geändert, falls noch nötig.

mi.inference{norm, mice}: Kombiniert die Analyseergebnisse aus den m imputierten vollständig Datensätzen zu einem Gesamtergebnis, wie in Abschnitt 4.1.1 beschrieben.

Zur Behandlung fehlender Daten in *R* sind vor allem die Pakete *Amelia* und *norm* empfehlenswert. Amelia (Honaker, Joseph, King, Scheve and Singh, 1998-2002) eignet sich besonders zur multiplen Imputation und im Paket norm (Schafer, 2002) sind der EM-Algorithmus, sowie MCMC-Algorithmen und ebenfalls multiple Imputation implementiert. Weitere *R*-Pakete, die zum Umgang mit unvollständigem Datenmaterial hilfreich sein können, sind zum Beispiel *cat, mice, mix* oder *pan*.

B. Abkürzungsverzeichnis

$\#$	*Anzahl*
Σ	*Kovarianzmatrix*
CCA	*Complete Case Analysis*
det	*Determinante*
EM	*Expectation Maximization*
I_p	*Indikatormatrix der Dimension p*
KI	*Konfidenzintervall*
L	*Likelihoodfunktion*
l	*Log − Likelihoodfunktion*
MAR	*Missing at Random*
$MCAR$	*Missing Completely at Random*
$MCMC$	*Markov Chain Monte Carlo*
MI	*Multiple Imputation*
ML	*Maximum Likelihood*
MVN	*Multivariate Normalverteilung*
NA	*not available, Bezeichnung für fehlenden Wert*
$NMAR$	*Not Missing at Random*
OAR	*Observed at Random*
S	*Scorefunktion*

C. Maximum-Likelihood Inferenz

Die *Maximum-Likelihood-Inferenz* wurde von R.A. Fisher ab den 1920er Jahren entwickelt und ist eine der zentralen Methoden der statistischen Inferenz. Wie das Wort Likelihood, was im Deutschen soviel wie Plausibilität oder höchste Wahrscheinlichkeit aussagt, andeutet, ist ihre Idee diejenigen Parameterschätzer zu finden, die durch die vorhandenen Daten am besten gestützt werden.

Likelihood-Funktion

Die Komponenten x_i des Vektors der Beobachtungen $x = (x_1, ..., x_n)$ seien unabhängig und identisch verteilte Realisationen einer Zufallsvariable X. Die Parameter der Verteilung werden durch $\theta = (\theta_1, ..., \theta_n)$ symbolisiert und seien bekannt. Dann lässt sich die Dichte aufgrund der Unabhängigkeit faktorisieren zu

$$f(x_1, ..., x_n|\theta) = f(x_1|\theta) \cdot ... \cdot f(x_n|\theta). \tag{C.1}$$

Die *Likelihood-Funktion* ist eine Funktion von θ bei gegebenen Beobachtungen x.

$$L(\theta|x) = f(x_1;\theta) \cdot ... \cdot f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{C.2}$$

Je höher der Wert der Funktion ist, umso wahrscheinlicher ist der Parametervektor θ unter den gegebenen Daten.
Das Maximum der Funktion ist also der plausibelste Wert aus dem Parameterraum und wird als *Maximum-Likelihood-Schätzer* bezeichnet.

$$\hat{\theta}_{ML} = \underset{\theta}{argmax}\, L(\theta|x). \tag{C.3}$$

Die Likelihood-Funktion ist *invariant* gegenüber eineindeutigen Transformationen von $\theta = \theta(\psi)$. Insbesondere gilt für die Likelihood-Schätzer $\hat{\psi}_{ML}$ und $\hat{\theta}_{ML}$

$$\hat{\theta}_{ML} = \theta(\hat{\psi}_{ML}). \tag{C.4}$$

Besitzen zwei Beobachtungen x und $\tilde{x}$ zueinander proportionale Likelihood-Funktionen, so müssen sie zu den selben Schlüssen führen. Diese Forderung wird auch als das *Likelihood-Prinzip* bezeichnet.

Charakteristika der Likelihood-Funktion

Meist ist es numerisch einfacher statt der Likelihood selbst den Logarithmus der Funktion, die *Log-Likelihood*, zu maximieren.
Ist sie differenzierbar, beschränkt und besitzt ein eindeutiges Maximum, so geschieht dies durch Nullsetzen der 1. Ableitung der Loglikelihood, der sogenannten *Score-Funktion* $S(\theta)$:

$$S(\theta) = \frac{dlnL(\theta)}{d\theta} \stackrel{!}{=} 0 \tag{C.5}$$

Dies gilt für den Fall, dass θ skalar ist, für vektorielles $\theta = (\theta_1, ..., \theta_n)$ erhält man ein n-dimensionales Gleichungssystem der partiellen Ableitungen, das es zu lösen gilt.
Die negative Krümmung $I(\hat{\theta})$ der Log-Likelihood an ihrem Maximum, die sogenannte *Fisher-Information*, gibt die Präzision des Schätzers an, wobei

$$I(\theta) = -\frac{d^2 lnL(\theta)}{d\theta^2} = -\frac{dS(\theta)}{d\theta}. \tag{C.6}$$

Der Standardfehler der ML-Schätzung entspricht asymptotisch dem Inversen der Fisher-Information:

$$se(\hat{\theta}_{ML}) = [I(\hat{\theta}_{ML})]^{-1}. \tag{C.7}$$

Beispiel: $X_1, ..., X_n \stackrel{iid}{\sim} N(\theta, \sigma^2)$ mit bekannter Varianz. Dann ergeben sich Likelihood, Log-Likelihood und Scorefunktion wie folgt:

$$L(\theta) = (2\pi\sigma)^{-\frac{n}{2}} \cdot \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\},$$

$$l(\theta) = lnL(\theta) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2,$$

$$S(\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta).$$

Nullsetzen der Scoregleichung ergibt den ML-Schätzer $\hat{\theta}_{ML} = \bar{x}$. Des weiteren ergibt sich die Fisher-Information

$$I(\theta) = \frac{n}{\sigma^2}.$$

Diese entspricht genau dem Inversen der Varianz des ML-Schätzers $Var(\hat{\theta}_{ML}) = Var(\bar{X}) = \frac{\sigma^2}{n}$.

Frequentistische Eigenschaften der Likelihood

Betrachtet man Scorefunktion, Fisher-Information und ML-Schätzer als Zufallsvariablen so lassen sich einige Aussagen über ihre Verteilungen treffen.
Da die Fisher-Information $I(\theta)$ als negative 2. Ableitung der Log-Likelihood oftmals von den Daten x abhängt, betrachtet man häufig die *erwartete Fisher-Information*, um sich von dieser Abhängigkeit zu lösen.

$$J(\theta) = E[I(\theta)] \tag{C.8}$$

Für die Verteilung der Scorefunktion ergibt sich unter Regularitätsbedingungen:

$$S(\theta) \overset{a}{\sim} N(0, J(\theta)) \tag{C.9}$$

Für die Verteilung des ML-Schätzers selbst ergibt sich:

$$\hat{\theta}_{ML} \overset{a}{\sim} N(\theta, [J(\theta)]^{-1}) \tag{C.10}$$

Den Beweis dieser Verteilungseigenschaften findet man unter anderem bei Lehmann (1998, S.449).
Anhand dieser Eigenschaften lässt sich dann die weitere Inferenz mit Konfidenzintervallen und Tests betreiben.

Iterative Verfahren zur Maximierung der Likelihood

Da es oftmals sehr kompliziert sein kann die Scoregleichung zu lösen, behilft man sich mit iterativen Algorithmen. Eines der bekanntesten dieser Verfahren stellt die *Newton-Raphson-Methode* dar. Sie ist definiert durch die Gleichung

$$\theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta^{(t)}|X)S(\theta^{(t)}|X). \tag{C.11}$$

Dies entspricht einer Taylorentwicklung, das heißt einer Linearisierung der Funktion, die in der ML-Schätzung konvergiert.
Eine weitere Variante stellt das *Fisher-Scoring* dar, welches statt der Fisher-Information $I(\theta)$ die erwartete Fisher-Information $J(\theta)$ verwendet.

$$\theta^{(t+1)} = \theta^{(t)} + J^{-1}(\theta^{(t)})S(\theta^{(t)}|X) \tag{C.12}$$

Im statistischen Programmpaket *R* steht zum Maximieren von Funktionen der implementierte Befehl *optim( )* zur Verfügung.

D. Bayes-Inferenz und Markov Chain Monte Carlo

In diesem Abschnitt wird eine kurze Einführung in das Konzept der Bayes-Inferenz gegeben, welche die zweite große Inferenzart in der Statistik neben der Likelihood-Inferenz darstellt. Näher eingegangen wird vor allem auf die Markov Chain Monte Carlo Methoden (MCMC), zu denen auch die Multiple Imputation zu zählen ist.

Grundlagen der Bayes-Inferenz

In der Bayes-Inferenz werden die zu schätzenden Größen, wie Parameter oder auch fehlende Daten, als Realisierungen von Zufallsvariablen gesehen.
Ihr Prinzip ist es eine gemeinsame Wahrscheinlichkeitsverteilung der beobachtbaren und nicht beobachtbaren Größen aufzustellen. Anschließend bildet man eine auf die beobachteten Größen bedingte *Posteriori-Verteilung*, aufgrund derer sich Aussagen über die interessierende Größe θ treffen lassen.

Zunächst spezifiziert man sich anhand von Vorwissen über θ eine *Priori-Verteilung* $p(\theta)$. Durch Beobachten der Daten $x = (x_1, ..., x_n)$ aus einer unabhängig und identisch verteilten Zufallsstichprobe und deren Likelihooddichte $L(x|\theta)$ lernt man mehr über die Verteilung der Parameter.
Nach dem Satz von Bayes erhält man dann die auf die beobachteten Daten bedingte Posterioriverteilung

$$f(\theta|x) = \frac{L(x|\theta) \cdot p(\theta)}{f(x)}. \tag{D.1}$$

Meist reicht für die Posteriori-Verteilung schon die unnormierte Beziehung

$$f(\theta|x) \propto L(x|\theta) \cdot p(\theta). \tag{D.2}$$

Man kombiniert also die Vorab-Information mit der Information aus einer Stichprobe zu einer neuen Dichte.

Als *Maximum a posteriori* Schätzer für θ nimmt man meist den Modus dieser Verteilung.
Zudem lassen sich Intervalle oder Regionen bestimmen, in denen θ mit Wahrscheinlichkeit $1-\alpha$ liegt.
Für die meisten Verteilungen gibt es sogenannte *konjugierte Priori-Verteilungen.* Das bedeutet, dass diese Priori kombiniert mit der Likelihood eine Posteriori-Verteilung ergibt, die wiederum aus der gleichen Verteilungsfamilie stammt wie die Priori. Beispielsweise wählt man die Betaverteilung als konjugierte Priori, wenn man die Erfolgswahrscheinlichkeit einer Binomialverteilung schätzen will. Dann ergibt sich als Posteriori wieder eine Betaverteilung.

Beispiel: (aus Fahrmeir 2006, S.384)
Man betrachtet n unabhängige Beobachtungen aus einer Verteilung $N(\mu, \sigma^2)$ mit unbekanntem μ und bekanntem σ^2. Als konjugierte Priori-Dichte für μ wird eine $N(\nu, \tau^2)$-Verteilung herangezogen, also

$$f(\mu) = (2\pi\tau^2)^{\frac{12}{\cdot}} \exp\left\{-\frac{(\mu-\nu)^2}{2\tau^2}\right\}. \tag{D.3}$$

Diese sagt aus, dass μ um den Erwartungswert ν mit der Varianz τ^2 streut. Je kleiner diese Streuung ist, desto genauer ist das Vorwissen.
Daraus ergibt sich die unnormierte Posteriori-Verteilung

$$f(\mu|x_1, ..., x_n) \propto L(\mu, \sigma^2) \cdot f(\mu). \tag{D.4}$$

Man erhält wiederum eine Normalverteilung $N(\tilde{\mu}, \tilde{\sigma}^2)$, mit Erwartungswert

$$\tilde{\mu} = \frac{n\,\tau^2}{n\tau^2 + \sigma^2}\bar{x} + \frac{\sigma^2}{\tau^2 + \sigma^2}\nu. \tag{D.5}$$

Der Erwartungswert ist ein gewichtetes Mittel aus Mittelwert der Stichprobe und Priori-Erwartungswert. Je weniger Vorwissen miteinfließt, das heißt, je größer die Priori-Varianz ist, desto mehr stützt man sich auf die Information aus der Stichprobe. Man steuert also mit τ^2 den Kompromiss zwischen der Information aus der Stichprobe und der subjektiven a priori Information (Fahrmeir 2006, S.385).
Als Posteriori-Varianz erhält man

$$\tilde{\sigma}^2 = \frac{\sigma^2}{n + \sigma^2/\tau^2}\ . \tag{D.6}$$

Wenn gar kein Vorwissen über die Datenstruktur vorliegt, ist es schwierig eine geeignete Priori zu finden. Deshalb kann in solchen Fällen eine sogenannte *nichtinformative Priori* zum Einsatz kommen. Diese in der Regel flache Verteilung lässt so wenig Vorwissen wie möglich einfließen und soll somit die Daten für sich sprechen lassen. Nichtinformative Prioris können auch *improper prioris* sein. Das heißt, dass sie ein infinites Integral haben, statt sich zu *1* zu integrieren.

Eines der bekanntesten Beispiele für nichtinformative Prioris ist die sogenannte *Jeffrey's Priori*. Diese wird proportional zur Wurzel der Fisher-Information gewählt,

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{D.7}$$

Sie ist invariant gegenüber bivariaten Transformationen und meist ein uneigentlicher Grenzfall der konjugierten Prioris.

Markov Chain Mote Carlo Methoden (MCMC)

In komplexen statistischen Modellen kann die Berechnung der Posteriori-Verteilung analytisch und numerisch zu komplex werden. Dies ist hauptsächlich der Fall, wenn hochdimensionale Integrale berechnet werden müssen, die in der Posteriori-Verteilung im Nenner vorkommen.
Mit Hilfe von MCMC Methoden kann jedoch eine sehr gute Näherung für diese Verteilung gefunden werden. Dies sind Simulationsverfahren, bei denen Stichproben aus Wahrscheinlichkeitsverteilungen gezogen werden. Dies geschieht durch die Konstruktion von Markovketten ($X_0, ..., X_n$) mit einer vorgegebenen stationären Verteilung π und einer sogenannten *Vorschlagsdichte* P. Das heißt, die Verteilung der gezogenen Zufallszahlen konvergiert gegen die interessierende Verteilung.
Die praktische Umsetzung erfolgt nach dem Muster

- Erzeuge $X_0 \sim \pi_0$
- Erzeuge $X_1 \sim P(\cdot \,|\, x_0)$

  $\vdots$
- Erzeuge $X_n \sim P(\cdot \,|\, x_{n-1})$

(D.8)

Unter Regularitätsbedingungen erzeugen die Realisierungen für genügend große n eine Verteilung nahe der stationären Verteilung und können deshalb als Sequenz aus der Zielverteilung π angesehen werden.
In der Praxis bedeutet dies, dass man für ein ausreichend groß gewähltes $m < n$ die Realisierungen $x_m, ..., x_n$ als Stichprobe der stationären Verteilung betrachtet. Man bezeichnet m als *burnim*. Alle vorherigen Realisierungen, bei denen sich die Markovkette noch nicht in ihrer stationären Verteilung befand, werden einfach gelöscht. Aus diesen Werten können dann die Verteilungseigenschaften geschätzt werden.
Im Folgenden werden die beiden geläufigsten MCMC Verfahren zur Generierung einer solchen Markovkette, der *Metropolis-Hastings-Algorithmus* und der *Gibbs-Sampler*, kurz erläutert.

Der Metropolis-Hastings-Algorithmus:
Bei gegebener stationärer Verteilung π simuliert man die Glieder der Markovkette wie zuvor

beschrieben nach dem Muster

$$X_i \sim P(\cdot \,|\, x_{i-1}).$$

Idee des MH-Algorithmus ist es die so erzielten Werte nur mit einer Akzeptanzwahrscheinlichkeit $\alpha(x_{i-1}, x_i)$ anzunehmen, wobei

$$\alpha(x_{i-1}, x_i) = min\left(1, \frac{\pi(x_i) \cdot P(x_{i-1} \,|\, x_i)}{\pi(x_{i-1}) \cdot P(x_i \,|\, x_{i-1})}\right). \tag{D.9}$$

Man zieht in der Praxis einen Wert U aus der Standardgleichverteilung. Erhält man $U < \alpha(x_{i-1}, x_i)$, so wird die Realisierung x_i akzeptiert. Ansonsten setzt man $x_i = x_{i-1}$ und behält den vorherigen Wert bei.

Gibbs-Sampling:
Beim Gibbs-Sampling wird der Übergangskern der Markovkette mit Hilfe der vollständig bedingten Verteilungen (*full conditionals*) gebildet. Dafür definiert man sich für den interessierenden Vektor $\theta = (\theta_1, ..., \theta_d)'$ die zugehörigen Vektoren aus $\mathbb{R}^{d-1}$, bei denen jeweils ein Element weggelassen wird:

$$\theta_{-i} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_d)', \quad \text{für } i = 1, ..., d. \tag{D.10}$$

Dann lässt sich die Markovkette mit der interessierenden Dichte $P(\theta)$ durch den Gibbs-Sampler wie folgt bilden:
Man wählt sich geeignete Startwerte $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})'$.
Dann erzeugt man iterativ neue Realisierungen dieses Vektors nach dem Muster

$$\theta_i^{(j)} = P\left(\theta_i^{(j-1)} | \theta_{-i}^{(j-1)}\right), \quad \text{für } i = 1, ..., d. \tag{D.11}$$

Dies wiederholt man solange, bis sich die Markovkette stationär verhält.

Die Anwendung des Gibbs-Samplers ist jedoch nur möglich, wenn sich aus allen bedingten Verteilungen $P\left(\theta_i | \theta_{-i}\right)$ zufällige Realisierungen ziehen lassen. Da die Dichten oftmals nur bis auf eine Proportionalitätskonstante bekannt sind, muss in diesen Fällen wieder auf den Metropolis-Hastings-Algorithmus zurückgegriffen werden.

Literaturverzeichnis

[1] **Allison, P.** 2000: *Multiple Imputation for Missing Data: A Cautionary Tale*, Sociological Methods and Research 28, S. 301-309.

[2] **Anderson, T.** (2003): *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics, 3. Auflage.

[3] **Backhaus, K. / Erichson, B. / Plinke, W. / Weiber, R.** (2006): *Multivariate Analysemethoden*, Springer, 11. Auflage.

[4] **Bankhofer, U.** (1995): *Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse*, Josef Eul Verlag.

[5] **Dellaert, F.** (2002): *The Expectation Maximization Algorithm*, College of Computing, Georgia Institute of Technology, Technical Report no. GIT-GVU-02-20.

[6] **Dempster, A. / Laird, N. / Rubin, D.** 1977: *Maximum-Likelihood from Incomplete Data via the EM-Algorithm*, Journal of the Royal Statistical Society, Series B 39, 1-38.

[7] **Efron, B. / Tibshirani, R.** 1993: *An Introduction to the Bootstrap*, Chapman & Hall.

[8] **Everitt, B.** 1984: *An Introduction to Latent Variable Models*, Chapman & Hall.

[9] **Fahrmeir, L. / Hamerle, A. / Tutz G.** (1996): *Multivariate statistische Verfahren*, de Gruyter, 2. Auflage.

[10] **Fahrmeir, L. / Künstler, R. / Pigeot, I. / Tutz, G.** (2006): *Statistik - Der Weg zur Datenanalyse*, Springer Verlag, 6. Auflage.

[11] **Fahrmeir, L. / Heumann, C.** (2008): *Testen und Schätzen I - Skript zur Vorlesung*, von *http://www.stat.uni-muenchen.de/ semwiso/schaetzentesten1-ws0809/vorlesung.html* .

[12] **Gelman, A. / Carlin, J. / Stern, C. / Rubin, D.** (2004): *Bayesian Data Analysis*, Chapman & Hall, 2. Auflage.

[13] **Ghahramani, Z. / Hinton, G.** (1996): *The EM-Algorithm for Mixtures of Factor Analysis*, Technical Report CRG-TR-96-1, University of Toronto.

[14] **Gilks, W. /Richardson, S. / Spiegelhalter, D.** (1996): *Markov Chain Monte Carlo in Practice*, Chapman & Hall.

[15] **Grimmet, G. / Stirzeker, D.** (2001): *Probability and Random Processes*, Oxford University Press.

[16] **Harman, H.** (1968): *Modern Factor Analysis*, The University of Chicago Press, 2. Auflage.

[17] **Hartung, J. / Elpelt, B.** (1995): *Multivariate Statistik*, Oldenbourg, 5. Auflage.

[18] **Heumann, C.** (2003): *Monte Carlo Methods for Missing Data in Generalized Linear and Generalized Linear Mixed Models*, Habilitationsschrift, Ludwig-Maximilians-Universität München.

[19] **Honaker, J. / Joseph, A. / King, G. / Scheve, K. / Singh, N.** (1998-2002): *Amelia: A Program for Missing Data*, *http://gking.harvard.edu/amelia* .

[20] **Hotelling, H.** (1936): *Relations between two Sets of Variates*, Biometrika 28 (1936), S. 321 - 377.

[21] **Lawley, D. / Maxwell, M.** (1971): *Factor Analysis as a Statistical Method*, London Butterworths.

[22] **Lehmann, E. / Casella, G.** (1998): *Theory of Point Estimation*, Springer, 2.Auflage.

[23] **Leisch, F.** (2002): *Sweave, Part I, Mixing R and LaTeX* , R News 2 (2002), December Nr.3, S.28-31, *http://CRAN.R-project.org/doc/Rnews* .

[24] **Little, R. / Rubin, D.** (2002): *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, 2. Auflage.

[25] **Louis, T.** (1982): *Finding the Observed Information Matrix when Using the EM Algorithm*, Journal of the Royal Statistical Society, Vol.44, No.2, S.226-233.

[26] **Rubin, D.** (1982): *EM-Algorithms for Factor Analysis*, Psychometrika Vol.47, No.1, S.68-77.

[27] **Rubin, D.** (1987) *Multiple Imputation for Nonresponse Survey*, Wiley Series in Probability and Statistics.

[28] **Schafer, J.** (1997): *Analysis of Incomplete Multivariate Data*, Chapman & Hall.

[29] **Schafer, J. / Graham, J.** 2002: *Missing Data: Our View of the State of the Art*, Psychological Methods Vol.7, No.2, S.147-177.

[30] **Schwab, G.** (1991): *Fehlende Werte in der angewandten Statistik*, Deutscher Universitäts-Verlag.

[31] **Thurstone, L.** (1947): *Multiple Factor Analysis*, Cambridge University Press.

[32] **Toutenburg, H.** (1992): *Lineare Modelle*, Physica Verlag.

[33] **Toutenburg, H. / Heumann, C.** (2007): *Induktive Statistik*, Springer, 4. Auflage.

[34] **Überla, K.** (1968): *Faktorenanalyse*, Springer Verlag.

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.
Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, den 04.02.2010

Unterschrift