Vergleich von Methoden zur Strukturfindung in der Psychometrie mit Hilfe echter Daten

Stella Bollmann

München, 2014

Vergleich von Methoden zur Strukturfindung in der Psychometrie mit Hilfe echter Daten

Inauguraldissertation
zur Erlangung des Doktorgrades der Philosophie
an der Ludwig–Maximilians–Universität München

vorgelegt von
Stella Bollmann
aus Nürnberg

2015

Erstgutachter: Prof. Dr. Markus Bühner
Zweitgutachter: Prof. Dr. Moritz Heene
Datum der mündlichen Prüfung: 22.01.2015

Inhaltsverzeichnis

Tabellenverzeichnis

AIC Akaike Informationskriterium

BIC Bayesianisches Informationskrieterium

CA Clusteranalyse

CAAL Clusteranalyse mit *average linkage*

CACL Clusteranalyse mit *complete linkage*

CFA konfirmatorische Faktorenanalyse

DGP datengenerierender Prozess

EFA exploratorische Faktorenanalyse

FA Faktorenanalyse

K-means cor K-means Korrelation

k-means SDM K-means skaliertes Distanzmaß

MAP Test *Minimum Average Partial Test*

ML Maximum-Likelihood

PA Parallelanalyse

PA-PAF Parallelanalyse mit Hauptachsenanalyse

PA-PCA Parallelanalyse mit Hauptkomponentenanalyse

PAF Hauptachsenanalyse

PCA Hauptkomponentenanalyse

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Evaluation von strukturfindenden Methoden, die die Items psychologischer Fragebogendaten in homogene Gruppen von ähnlichen Items zusammenfassen. Ein wesentlicher Unterschied zwischen Methoden, die zu diesem Zweck verwendet werden, ist, ob sie ein zugrundeliegendes Messmodell annehmen oder ob sie nur eine möglichst brauchbare Gruppierung der Items anstreben.
Zum einen gibt es die modellbasierte Faktorenanalyse (FA), die auf dem Faktormodell basiert. Der mathematische Ansatz ist ähnlich der Hauptkomponentenanalyse, oder principal component analysis (PCA). In der FA wird im Unterschied zur PCA noch angenommen, dass die Antworten auf die Items kausal von zugrundeliegenden Faktoren plus einem einzigartigen Residualterm kausal erklärt werden. Und dieser spezifische Residualterm jedes Items wird als völlig unkorreliert zu allen anderen Items angenommen. Ein Verfahren, das keine Modellannahmen trifft, ist die Clusteranalyse (CA). Hier werden lediglich Objekte zusammengefügt, die sich auf einem bestimmten Kriterium ähnlicher sind als andere.
So wie man Methoden darin unterscheiden kann, ob sie ein zugrundeliegendes Modell annehmen oder nicht, kann man auch bei der Evaluation von Methoden diese Unterscheidung treffen. Eine Evaluationtechnik, die ein Modell annimmt, ist die Monte Carlo Simulation. Eine Technik, die nicht zwangsweise ein Modell zugrunde legt, ist das *Resampling*. Es werden Stichproben aus einem echten Datensatz gezogen und das Verhalten der Methode in diesen Stichproben wird untersucht.
In der ersten Studie wurde ein solches *Resampling*-Verfahren angewandt, das wir *Real World Simulation* nennen. Es soll das bestehende Problem der mangelnden Validität von Monte Carlo Studien zur FA beheben. Es wurde eine *Real World Simulation* an zwei großen Datensätzen durchgeführt und die Schätzer der Modellparameter aus dem echten Datensatz anschließend für die Monte Carlo Simulation als Modellparameter verwendet. So kann getestet werden, welchen Einfluss die spezifischen Datensatzcharakteristiken sowie kontrollierte Veränderungen von ihnen auf die Funktion der Methoden haben. Die Ergebnisse legen nahe, dass die Resultate von Simulationsstudien immer stark von bestimmten Spezifikationen des Modells und seiner Verletzungen abhängen und daher keine allgemeingültigen Aussagen getroffen werden können. Die Analyse echter Daten ist wichtig, um die Funktion verschiedener Methoden zu verstehen.
In der zweiten Studie wurde mit Hilfe dieser neuen Evaluationstechnik ein neues k-means Clusterungsverfahren zur Clusterung von Items getestet. Die zwei Verfahren, die vorgeschlagen wurden, sind: k-means skaliertes Distanzmaß (*k-means SDM*) und *k-means cor*.

In den Analysen zeigte sich, dass sich die neuen Verfahren besser eignen, Items zu Konstrukten zuzuordnen als die EFA. Lediglich bei der Bestimmung der Anzahl der zugrundeliegenden Konstrukte, waren die EFA-Verfahren genauso gut. Aus diesem Grund wird vorgeschlagen eine Kombination dieser beiden Verfahren zu verwenden. Ein großer Vorteil der neuen Methoden ist, dass sie das Problem der Unbestimmtheit der Faktorwerte in der EFA lösen können, da die Clusterwerte der Personen auf den Clustern eindeutig bestimmt werden können.
Am Ende der Arbeit wird auf die unterschiedlichen Evaluierungs- bzw. Validierungstechniken für modellbasierte und nicht-modellbasierte Verfahren eingegangen. Für die Zukunft wird vorgeschlagen, für die Evaluation des neuen k-means CA Verfahrens zur Clusterung von Items, *Real World Simulationen* sowie Validierungen der Clusterwerte mit Außenkriterien anzuwenden.

Kapitel 1

Der Vergleich strukturfindender Methoden: Eine Einleitung

1.1 Dimensionsreduzierung und das latente Faktormodell

Wenn man in der Psychometrie von „Dimensionsreduzierung“ spricht, ist im Allgemeinen die Reduzierung der Komplexität eines Itemsatzes gemeint. Viele nicht beobachtbare Eigenschaften werden in der Psychometrie über Fragebögen erfasst, und es wird davon ausgegangen, dass sich die Antworten auf all diese Fragen auf wenige gemessene Eigenschaften reduzieren lassen. Das Ziel ist also, nicht die Antworten auf alle Items interpretieren zu müssen, sondern diese Vielzahl an Antworten so zusammenzufassen, dass nur ein paar wenige Gesamtwerte für jede Person übrigbleiben, die als die Eigenschaften dieser Person interpretiert werden können. Da solche Gruppen von Items auch als Struktur eines Fragebogens gesehen werden können, spricht man auch von *strukturfindenden Methoden*. Es gilt nun also herauszufinden, wie viele verschiedene Konstrukte von den Items repräsentiert werden, welches Item welchen Konstrukte bzw. Eigenschaft repräsentiert und wie gut jedes einzelne Item die jeweilige Eigenschaft repräsentiert. Dabei gibt es verschiedene Ansätze, dieses Problem zu lösen und in dieser Arbeit sollen davon die folgenden drei besprochen werden:

Die *Hauptkomponentenanalyse* oder englisch *principal component analysis* (PCA)

Die *Faktorenanalyse* (FA)

Die *Clusteranalyse* (CA)

Ein grundlegener Unterschied dieser drei Verfahren ist, dass nur die FA ein zugrundeliegendes Modell annimmt und sowohl PCA als auch CA nicht. In der FA wird angenommen, dass die Antworten auf die Items von wenigen zugrundeliegenden Faktoren verursacht wurden. Die beobachteten Variablen, also die Items, sind hier die abhängigen Variablen, die

durch gemeinsame Faktoren sowie einen einzigartigen Residualterm erklärt werden. Diese Idee geht auf Spearman (1904) zurück, der annahm, dass sich die Ergebnisse von Intelligenztests alle auf einen generellen Intelligenzfaktor zurückführen lassen. Es wird hier also eine klare Kausalstruktur angenommen: Die latente Variable erklärt die Antworten auf die beobachteten Variablen. Die PCA hingegen ist eine Methode, mit der, ohne Annahme einer Kausalstruktur, Dimensionen reduziert werden. Es werden Komponenten angenommen, deren Anzahl deutlich geringer ist als die der beobachteten Variablen. Dabei werden die Komponenten als Linearkombinationen der Variablen dargestellt. In der PCA wird nicht angenommen, dass die Komponenten die Unterschiede auf den Itemantworten erklären. Die Komponenten werden vielmehr als Zusammensetzung der Itemantworten beschrieben. Die beobachteten Variablen stehen also auf der anderen Seite der Gleichung. Sie definieren bzw. bilden die Komponenten (Fahrmeir, Brachinger, Hamerle & Tutz, 1996, S. 661; Fabrigar, Wegener, MacCallum & Strahan, 1999; Rencher & Christensen, 2012, S. 475).

Die CA wird meistens für die Clusterung von Fällen, also Objekten oder Personen verwendet, die sich im Bezug auf bestimmte Eigenschaften möglichst ähnlich sind. Sie kann jedoch auch zur Clusterung von Items angewandt werden. Dann ist die Grundidee sehr ähnlich zu der der PCA. Die Items werden dabei zu homogenen Gruppen zusammengefügt, um die Komplexität des Datensatzes zu reduzieren (Tryon, 1935; Loevinger, Gleser & DuBois, 1953).

Mathematisch ist das Vorgehen in PCA und FA verwandt, wie im weiteren Verlauf noch gezeigt werden wird. Der Begriff der FA umfasst zwar verschiedene Techniken, sie haben jedoch alle das gleiche zugrundeliegenden mathematische Modell. Die CA hingegen ist vielmehr ein Sammelbegriff für eine Vielzahl verschiedener Methoden, die alle zum Ziel haben, p Objekte (in diesem Fall Items) zu m Clustern zusammenzuschließen, mit $m < p$. Dabei sollen die Abstände der Items innerhalb eines Clusters möglichst gering sein und im Falle von manchen Cluster-Algorithmen auch die Abstände der Items aus verschiedenen Clustern möglichst groß.

Die Arbeit wird sich im weiteren Verlauf in erster Linie mit den letzten beiden, der FA und der CA beschäftigen. Zunächst soll jedoch zuvor die PCA beschrieben werden, da dies essentiell dazu beiträgt, den Unterschied zwischen den anderen Methoden besser zu verstehen.

1.1.1 Hauptkomponentenanalyse

Bei der PCA, die auf Pearson (1901) und Hotelling (1933) zurück geht, werden die Hauptkomponenten durch Lineartransformation aus den beobachtbaren Variablen gebildet. Die Hauptkomponenten sind dabei unkorreliert und nach fallender Varianz geordnet. Dadurch wird erreicht, dass die ersten $m < p$ Hauptkomponenten (mit p = Anzahl der Variablen) ein Maximum an Gesamtvarianz auf sich vereinen (Fahrmeir et al., 1996, S. 661). Die beobachtete Korrelationsmatrix wird schließlich approximiert als das Produkt dieser Komponenten. Mathematisch lässt sich die PCA folgendermaßen darstellen: Die Ausgangsgröße

der PCA sind die Vektoren der Beobachtungen für jede Person

$$\boldsymbol{x_i^T} = (x_{i1}, \ldots, x_{ip}),$$

$i = 1, \ldots, n$ stellt dabei den Index der Personen und $j = 1, \ldots, p$ den Variablenindex dar. Das hochgestellte ^{T} steht für „transponiert“ und bedeutet, dass aus dem Spaltenvektor ein Zeilenvektor gemacht wird.

Es ist nun ein bestimmter Gewichtungsvektor

$$\boldsymbol{\alpha}_k^T = (\alpha_{k1}, \ldots, \alpha_{kp})$$

der k-ten Hauptkomponente gesucht. Dabei ist $k = 1, \ldots, m$ der Komponentenindex. Dieser Vektor muss die unten beschriebenen Eigenschaften erfüllen und mit ihm können die Beobachtungsvektoren mittels Linearkombination umgeschrieben werden zu:

$$z_{ik} = \boldsymbol{\alpha}_k^T \boldsymbol{x}_i \tag{1.1}$$

wobei z_{ik} der Wert der Person i auf der Komponente k ist. Für jede Komponente gibt es einen Vektor $\boldsymbol{\alpha}_k$ mit p Elementen, den Gewichtungen der Variablen. Es findet hier also eine Matrixmultiplikation statt, bei der die Datenmatrix $\mathbf{X}(n \times p)$ mit einer Gewichtungsmatrix $\mathbf{A}(p \times m)$ multipliziert wird, so dass eine Matrix $\mathbf{Z}(n \times m)$, der Werte der Personen auf den Komponenten, entsteht, die nun nur noch so viele Spalten hat, wie Komponenten extrahiert wurden. Wie bereits erwähnt, sollen diese Hauptkomponenten möglichst viel Varianz besitzen.

Zunächst soll die erste Hauptkomponente extrahiert werden. Dazu muss die Linear-Kombination

$$z_{i1} = \boldsymbol{\alpha}_1^T \boldsymbol{x_i}$$

mit dem Gewichtungsvektor $\boldsymbol{\alpha_1}$ gefunden werden, so dass

$$Var(\boldsymbol{z_1})$$

maximal wird. Für den allgemeinen Fall einer Komponente k, lässt sich diese Varianz der Beobachtungen $z_{1k}, \ldots, z_{nk}$, also der Werte der Personen auf der Komponente k berechnen über:

$$Var(z_k) = \frac{1}{n} \sum_{i=1}^{n} (z_{ik} - \bar{z}_k)^2 = \boldsymbol{\alpha}_k^T \boldsymbol{\Sigma}_x \boldsymbol{\alpha}_k$$

wobei $\bar{z}_k = \frac{1}{n} \sum_{i=1}^{n} z_{ik}$ ist und $\boldsymbol{\Sigma}_x = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ die empirische Kovarianzmatrix ist, die aus den Datenvektoren $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ berechnet wurde. Damit die Lösung für die erste Hauptkomponente eindeutig ist, muss folgende Nebenbedingung eingeführt werden:

$$||\boldsymbol{\alpha}_1||^2 = \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = 1$$

Das bedeutet, dass die Kovarianz der Komponenten $\boldsymbol{\Sigma}_z$ bei Extraktion von genau p Komponenten genau der Kovarianz der beobachteten Werte $\boldsymbol{\Sigma}_x$ entspricht. Würde diese Nebenbedingung nicht eingeführt, könnte die Varianz der Komponenten ohne Grenze erhöht

werden, indem die Elemente von $\boldsymbol{\alpha}_k$ vergrößert werden und es gäbe keine eindeutige Lösung. Im zweiten Schritt wird die zweite Hauptkomponente gesucht. Dies erfolgt analog zur Ermittlung der ersten Hauptkomponente durch Maximierung der Varianz der zweiten Linearkombination. Die Nebenbedingungen, die nun eingeführt werden müssen, sind:

1. $||\boldsymbol{\alpha}_2||^2 = \boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_2 = 1$
2. $Cov(\boldsymbol{z}_1, \boldsymbol{z}_2) = 0 \Leftrightarrow \boldsymbol{\alpha}_1 \perp \boldsymbol{\alpha}_2$

Die zweite Nebenbedingung bedeutet, dass die beiden Hauptkomponenten unkorreliert sein müssen. Ebenso wird auch mit den weiteren Komponenten verfahren. Unter Verwendung des Lagrange Multiplikators λ, kann gezeigt werden, dass das Maximierungsproblem dem Eigenwertproblem

$$\boldsymbol{\Sigma}_x \boldsymbol{\alpha}_1 = \lambda \boldsymbol{\alpha}_1$$

entspricht. Folglich ist der Eigenvektor $\boldsymbol{\alpha}_1$ zum größten Eigenwert λ von $\boldsymbol{\Sigma}_x$ die Lösung des Maximierungsproblems (R. E. Anderson, Hair, Tatham & Black, 2006). Äquivalent ergibt sich die 2. Hauptkomponente $\boldsymbol{\alpha}_2$ als der Eigenvektor zum zweitgrößten Eigenwert und so weiter. Mittels Spektralzerlegung kann die über die Komponenten reproduzierte Kovarianzmatrix $\boldsymbol{\Sigma}_x$ geschrieben werden als:

$$\boldsymbol{\Sigma}_x = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T \tag{1.2}$$

wobei $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)$ die Eigenvektoren von $\boldsymbol{\Sigma}_x$ sind und den gesuchten Gewichtungsvektoren entsprechen und $\boldsymbol{\Lambda}$ die Diagonalmatrix der Eigenwerte $\lambda_1 \geq \cdots \geq \lambda_p$ ist. In diesem Fall wurden also alle p Komponenten extrahiert. Deshalb kann die beobachtete Kovarianzmatrix $\boldsymbol{\Sigma}_x$ auch exakt reproduziert werden. Werden weniger als p Komponenten extrahiert, entspricht die reproduzierte Matrix im Regelfall nicht exakt der empirischen Kovarianzmatrix. Es wird aber immerhin die maximal durch $m < p$ Komponenten reproduzierbare Varianz, reproduziert, wenn man eine Eigenwertzerlegung dieser Matrix durchführt und nur die ersten m Komponenten extrahiert. Es lässt sich einfach zeigen, dass für die Kovarianzmatrix der Vektoren der Hauptkomponentenwerte $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ gilt:

$$\boldsymbol{\Sigma}_z = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})^T = \boldsymbol{\Lambda} \tag{1.3}$$

Die Hauptkomponenten sind also unkorreliert und ihre Varianz ist über die Eigenwerte gegeben (Fahrmeir et al., 1996, S. 662 ff.). Gleichung (1.2) wird als Ausgang für die Gleichung der FA verwendet. Dann wird üblicherweise folgende Terminologie verwendet, in der die Eigenvektorelemente als Ladungen der Items auf den Komponenten/Faktoren interpretiert werden:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T \tag{1.4}$$

wobei $\boldsymbol{\Lambda}$ die Ladungsmatrix der Variablen oder Items auf den Komponenten ist und $\boldsymbol{\Phi}$ die Diagonalmatrix der Kovarianzmatrix der Hauptkomponenten bzw. Faktoren.

1.1.2 Faktorenanalyse

Die FA wurde ursprünglich für die Analyse mentaler Testwerte entwickelt (Spearman, 1904; Thurstone, 1935). Inzwischen wird sie aber auf ein viel breiteres Spektrum an Situationen angewandt, wie z.B. Einstellungstests, Messungen physikalischer Größen und ökonomische Messwerte (Revelle, 2014).

Wie bereits erwähnt, liegt bei der FA im Gegensatz zur PCA ein Modell über das Zustandekommen der Messung, also der beobachteten Daten, zugrunde. Es wird angenommen, dass der Vektor der Beobachtungen zusammengesetzt ist aus einem systematischen Term und einem beobachteten, unsystematischen, also Residualterm. Die Elemente des Residualterms werden als unkorreliert, bzw. unabhängig, angesehen, wohingegen der systematische Teil als Linearkombination der relativ geringen Anzahl an latenten Faktorvariablen angesehen wird. Wie die beobachteten Variablen variieren auch die Faktoren von Person zu Person. Im Unterschied zu ihnen, können sie jedoch nicht beobachtet werden, sondern sind latent (Rencher & Christensen, 2012, S. 435). Das Ziel ist es also, die Variabilität der beobachteten Variablen über die latenten Faktoren, zum Beispiel Intelligenz, zu erklären. Eine solche Variable kann dabei ein einzelnes Item sein, oder auch ein ganzer Test, wenn das Faktormodell auf die Faktorisierung von Tests angewandt wird. Im Folgenden werden sowohl die Bezeichnung „beobachtete Variable“ als auch „Item“ gebraucht. Die zugrundeliegende Idee ist, dass sich Werte innerhalb einer Person auf verschiedenen Items, die das Gleiche messen, ähnlicher sein sollten als Werte zwischen verschiedenen Personen auf dem gleichen Item. Man geht also davon aus, dass Itemwerte von einer Person auf Items, die das Gleiche messen, miteinander zusammen hängen. Oder, um es wieder in Varianzen auszudrücken: Die Varianz eines Items lässt sich einteilen in einen Varianzanteil, der mit anderen Items, die das Gleiche messen, zusammenhängt, und einen Varianzanteil, der spezifisch für diese Messung ist. Der zweite Teil wird dabei als *Residuum* bezeichnet und ist das, was die FA von der PCA abgrenzt, wo dieses Residuum nicht enthalten ist. Der erste Teil ist, ähnlich wie in der PCA, eine Funktion der Varianz der *Faktorwerte* der Personen (T. W. Anderson, 1958, S. 569-570). Das Faktormodell soll im Folgenden genauer beschrieben werden.

Das Modell

Wir beginnen auf Ebene der Itemwerte, die durch zugrundeliegende Faktoren erklärt werden, bevor daraus die Zusammensetzung der Varianzen abgeleitet wird:

Der Vektor der Beobachtungen einer Person i kann geschrieben werden als:

$$\boldsymbol{x_i} = \boldsymbol{\Lambda f_i} + \boldsymbol{u_i} + \boldsymbol{\mu} \tag{1.5}$$

wobei $\boldsymbol{x_i}$ der Spaltenvektor der Beobachtungen der Person i ist; $\boldsymbol{u_i}$, und $\boldsymbol{\mu}$ sind Spaltenvektoren mit p Elementen für die p Items, $\boldsymbol{f_i}$ ist der Spaltenvektor der Faktorwerte der Person i mit $m \leq p$ Elementen, und $\boldsymbol{\Lambda}$ ist eine $p \times m$ Matrix.

Wenn es um psychometrische Tests geht, dann ist jedes Element x_{ij} ein Wert auf einem Item. Die zugehörige Komponente μ_j ist der Erwartungswert auf diesem Item in der

Population. Die Population ist die Grundgesamtheit aller Individuen, die bestimmte Merkmale gemeinsam haben, weshalb sie für die jeweilige Untersuchung interessant sind. Die Elemente f_{ik} sind die Faktorwerte auf den Faktoren, deren Linearkombinationen in den Itemwert einfließen. Die Koeffizienten dieser Linearkombinationen sind die Elemente λ_{jk}, die *Faktorladungen* genannt werden. Der Unterschied zu der Linearkombination der PCA (1.1) ist erstens, dass nicht die Komponente über die Variablen beschrieben wird, sondern umgekehrt die Faktoren die Variable erklären. Zweitens, erklärt die Linearkombination $\boldsymbol{\Lambda f}_i$ nun nur noch die systematische Abweichung von dem Populationsmittelwertvektor $\boldsymbol{\mu}$ und die restlichen, unsystematischen Abweichungen stecken in $\boldsymbol{u_i}$. Die Komponenten von $\boldsymbol{u_i}$ sind die Anteile des Items, die nicht von den Faktoren erklärt werden, die Residuen. Sie setzen sich jeweils zusammen aus einem *Messfehler* und einem *itemspezifischen Faktor*, den es nur in diesem Item gibt (T. W. Anderson, 1958, S. 569-570). Wenn man, wie in unserem Fall, nur einen Satz an Beobachtungen für jedes Individuum hat, kann man nicht zwischen diesen beiden unterscheiden, deshalb soll $\boldsymbol{u_i}$ hier als Messfehler, oder Residuum, bezeichnet werden.

Im Folgenden soll die Kovarianzmatrix $\boldsymbol{\Sigma}$ der Beobachtungen in Abhängigkeit von der Faktorkovarianzmatrix $\boldsymbol{\Phi}$ der Faktorvariablen $\boldsymbol{f_i}$ betrachtet werden. Es wird angenommen, dass $\boldsymbol{u_i}$ unabhängig von $\boldsymbol{f_i}$ verteilt ist mit Mittelwert 0 und Kovarianzmatrix $\mathbb{E}(\boldsymbol{u_i}, \boldsymbol{u_i^T}) = \mathbf{U}^2$. Wir nehmen weiterhin an, dass $\mathbb{E}(\boldsymbol{f_i}) = 0$ und $\mathbb{E}(\boldsymbol{f_i f_i^T}) = \boldsymbol{\Phi}$, wobei sowohl $\boldsymbol{f_i}$ als auch $\boldsymbol{u_i}$ normalverteilt sind.

Eine weitere Annahme der FA ist, dass die Komponenten von $\boldsymbol{u_i}$ über die Personen hinweg unkorreliert sind, da alle gemeinsame Varianz der beobachteten Variablen von den Faktoren erklärt wird. Das bedeutet, dass die Kovarianzmatrix der Residualterme $\mathbf{U^2}$ eine Diagonalmatrix ist. Aus allen bisherigen Annahmen ergibt sich die Kovarianzmatrix der beobachteten $\boldsymbol{x_i}$ durch:

$$\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^T = \mathbb{E}(\boldsymbol{\Lambda f_i} + \mathbf{u_i})(\boldsymbol{\Lambda f_i} + \mathbf{u_i})^T = \boldsymbol{\Lambda \Phi \Lambda}^T + \mathbf{U^2} \tag{1.6}$$

wobei $\boldsymbol{\Sigma}$ die $p \times p$ Korrelationsmatrix der p beobachteten Variablen ist, $\mathbf{U}^2$ die $p \times p$ Diagonalmatrix der Residualvarianzen, $\boldsymbol{\Lambda}$ ist die $p \times m$ Matrix der Faktorladungen auf den $m < p$ Faktoren und $\boldsymbol{\Phi}$ steht für die $m \times m$ Matrix der Faktorkorrelationen.

Diese Gleichung weist einige Ähnlichkeiten zu Gleichung (1.4) auf, wenn man die Varianz der Residualterme $\mathbf{U}^2$ vernachlässigt. Allerdings sollte beachtet werden, dass $\boldsymbol{\Sigma}$ in der PCA die reproduzierte Kovarianzmatrix war, die nur dann der beobachteten Kovarianzmatrix entspricht, wenn so viele Komponenten extrahiert werden wie es Items gibt. Man versucht nun mit dieser reproduzierten Kovarianzmatrix die beobachtete Kovarianzmatrix möglichst gut zu approximieren. In der FA geht man allerdings gar nicht davon aus, dass man die gesamte Varianz der beobachteten Variablen erklären kann. Deshalb wird hier der Residualterm $\mathbf{U}^2$ eingefügt, um eine Gleichung für die komplette Kovarianzmatrix der Beobachtungen zu bekommen.

Eine Möglichkeit, eine FA durchzuführen, ist es nun z.B. einfach eine PCA an einer reduzierten Matrix durchzuführen, in der in der Diagonale nur die aufklärbare Varianz steht. Dieses Vorgehen nennt sich *Hauptachsenanalyse* (PAF) (Revelle, 2014). Neben der PAF

gibt es noch weitere Methoden zur Parameterschätzung. In der vorliegenden Arbeit wurde die *Maximum-Likelihood* (ML)-Schätzung verwendet (für eine genauere Beschreibung siehe z.B. Fahrmeir et al., 1996, S.646 ff.).

Das Modell der FA wird sehr kontrovers diskutiert und ist für viele Statistiker keine legitime multivariate Methode (Rencher & Christensen, 2012, S. 441). Das liegt an den verschiedenen bekannten Problemen der Methode. Das erste Problem der FA ist, dass die Anzahl der Faktoren m nicht eindeutig festgelegt ist. Um dieses Problem zu lösen, gibt es verschiedene mögliche Methoden, die der eigentlichen FA zeitlich vorgeschaltet werden und die im Kapitel 2 genauer beschrieben werden. Es ist jedoch wichtig anzumerken, dass die FA nicht angibt, wie viele Faktoren zu extrahieren sind, also theoretisch $p - 1$ Faktoren möglich wären, von denen einige unter Umständen auch zu sehr unterschiedlichen Parameterschätzungen führen.

Des Weiteren besteht für alle Modelle, die mehr als einen Faktor beinhalten, ein Rotationsproblem, das heißt, mehrere äquivalente Modelle können ineinander überführt werden und führen also zu der gleichen reproduzierten Kovarianzmatrix. Selbst wenn die Anzahl zu extrahierender Faktoren bekannt ist, ist die Ladungsmatrix noch nicht eindeutig. Sie kann durch Multiplikation mit einer orthogonalen Matrix in eine Ladungsmatrix überführt werden, die zu der gleichen reproduzierten Kovarianzmatrix führt. Diese Multiplikation wird aufgrund ihrer geometrischen Interpretation Rotation genannt. Für einen mathematische Beweis dieses Problems sei auf T. W. Anderson (1958, S. 571) verwiesen. Diese Problem wird im Allgemein dadurch gelöst, dass man die Ladungsmatrix sucht, die am ehesten eine *Einfachstruktur* aufweist, also hohe Ladungen auf einem Faktor und niedrige auf allen anderen Faktoren.

Aus dem genannten Grund gibt es viele verschiedene Rotationsmethoden, die verschiedene Optimierungskriterien für die Faktorstruktur ansetzen. Diese unterscheiden sich unter anderem dadurch, ob sie fordern, dass $\mathbf{\Phi} = \mathbf{I}$, was gleichbedeutend damit ist, dass die Faktoren *orthogonal* und auf 1 skaliert sind. Wenn $\mathbf{\Phi}$ nicht diagonal ist, dann spricht man von *obliquen* Faktoren. Wenn verlangt wird, dass $\mathbf{\Phi}$ diagonal ist, bedeutet das, dass die Komponenten von $\boldsymbol{f_i}$ unabhängig verteilt sind, sofern $\boldsymbol{f_i}$ normalverteilt ist (T. W. Anderson, 1958, S. 571).

Für den Fall, dass $\mathbb{E}(\mathbf{\Phi} = \mathbf{I})$, also die Faktoren als unkorreliert angenommen werden, ergibt sich folgende Kovarianzmatrix:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{U^2} \tag{1.7}$$

In dieser Arbeit wurde für die FA die Methode *Promax* verwendet, eine oblique Rotation (Hendrickson & White, 1964). Sie ist die am häufigsten verwendete oblique Rotationsmethode und ihr gutes Funktionieren konnte gezeigt werden (Gerbing & Hamilton, 1996).

Schließlich ergibt sich noch ein weiteres Problem der FA wenn schon alle Parameter des Modells mittels einer gewählten Methode geschätzt wurden: Die *Unbestimmtheit der Faktoren*. Als Ergebnis der FA erhält man eine Gleichung für jedes x_{ij}:

$$x_{ij} = \boldsymbol{\lambda}_j^T \boldsymbol{f}_i + u_{ij} + \mu_j \tag{1.8}$$

wobei x_{ij}, also der Wert der Person i auf dem Item j, und λ_{jk}, die Ladung des Items j auf dem zugehörigen Faktor k, das das Item messen soll, sowie μ_j bekannt sind. Von u_{ij} und $\boldsymbol{f}_i$ sind jedoch nur die Varianzen (und Kovarianzen) bekannt. Möchte man die Faktorwerte f_{ik} nun berechnen, gibt es unendlich viele Lösungen, mit welchen Faktorwerten und Residualtermen u_{ij} man genau diese Gleichung erfüllt. Für diese Gleichung gibt es nur dann eine eindeutige Lösung, wenn die Residualvarianzen auf null gesetzt werden, was der PCA entspricht. Schonemann und Steiger (1978) konnten zeigen, dass die Schätzung der Faktorwerte zwar möglich ist, aber die Lösung des Gleichungssystems eben nicht eindeutig. Sie können z.B. über die Regressionsmethode oder die Bartlett-Methode geschätzt werden (für eine Erklärung sei auf Revelle, 2014 verwiesen). Auf dieses Problem werden wir noch einmal zurück kommen, wenn im folgenden Kapitel auf die Unterschiede zwischen FA, PCA und CA in diesem Zusammenhang eingegangen wird.

1.1.3 Clusteranalyse

Im Gegensatz zu der PCA und FA, ist die CA, wie bereits erwähnt, in erster Linie dazu da, die Komplexität eines Datensatzes zu reduzieren, indem Objekte zu Gruppen zusammengefasst werden. Dies geschieht so, dass die Objekte innerhalb einer Gruppe sich ähnlicher sind als zwischen den Gruppen (Jain & Dubes, 1988). Die Verwendung der CA zur Clusterung von Variablen bzw. Items, anstatt Personen oder Objekten wurde in den 30er Jahren von Tryon (1939) in der Psychologie eingeführt. Während die CA in anderen Fachrichtungen, wie Biologie, Marketing bzw. zur Clusterung von Personen in der Familien- und klinischen Psychologie, viel Verwendung findet, wird sie in der Psychometrie zur Clusterung von Items kaum angewandt (Revelle, 2014). Auf der anderen Seite wurde die Verwendung der EFA schon vor einiger Zeit immer wieder kritisiert Tryon (1935); Loevinger et al. (1953). Zum Beispiel wurde das Problem angesprochen, dass die Grundannahme, jedes Item sei eine Linearkombination von mehreren Faktoren, nicht zu dem praktischen Problem passt, dass man jedes Item zu genau einem Untertest zuordnen möchte Loevinger et al. (1953).

Die CA hingegen beschäftigt sich in erster Linie mit genau diesem Problem, die Items zu Untertests zuzuordnen. Grundsätzlich können Clusteralgorithmen in zwei Gruppen eingeteilt werden: Die *hierarchischen* und die *nicht-hierarchischen*.

Hierarchische CA Ansätze

In der hierarchischen CA werden Cluster gebildet, die innerhalb von Clustern genestet sind. Das Ergebnis ist ein Baumdiagramm, das besser bekannt ist unter dem Namen *Dendrogram*, in dem die genestete Struktur gezeigt ist. Die hierarchische CA kann weiter unterteilt werden in *agglomerative* und *divisive* Cluster Algorithmen. Bei den divisiven wird mit einem großen Cluster gestartet, in dem sich alle Items befinden und dieses Cluster wird dann immer weiter aufgesplittet bis jedes Item ein eigenes Cluster bildet. Bei den agglomerativen Verfahren verläuft dieser Prozess in die andere Richtung: Als Startpunkt bildet jedes Item ein eigenes Cluster, die dann schrittweise zusammegefasst werden, bis alle Items nur noch ein Cluster bilden (Rencher & Christensen, 2012, S.505 ff.). Welche

Cluster zusammengefügt werden bzw. wo getrennt werden soll, wird über das *Distanzmaß* entschieden, das vorher festgelegt werden muss. Für den speziellen Fall der Clusterung von Items in der Psychometrie ist einer der prominentesten Vorschläge z.B. der von Revelle, das ICLUST Verfahren, in dem der *Reliabilitätskoeffizient beta* verwendet wird (Revelle, 1979), der laut Autor eine gute Alternative zu cronbachs alpha darstellt. Er basiert auf der geringstmöglichen split-half Reliabilität der Items.

Außerdem gibt es noch mehrere Möglichkeiten, auf welche Art und Weise anhand des Distanzmaßes bestimmt wird, wie ähnlich sich zwei Cluster sind. Sokal und Sneath (1963) zogen die Distanz zwischen zwei Clustern heran, die über den nächsten Nachbarn der beiden Cluster bestimmt wird (*single linkage*), sowie die Distanz, die über den entferntesten Nachbarn zweier Cluster bestimmt wird (*complete linkage*). Außerdem gibt es noch die Möglichkeit, den durchschnittlichen Abstand aller Elementenpaare zweier Cluster als deren Abstand zu definieren, was *average linkage* genannt wird. Bacon (2001) verglich verschiedene hierarchische CA Methoden mit der FA und fand z.B. heraus, dass *average linkage* im Allgemeinen passendere Ergebnisse liefert als *single linkage*.

Nicht-hierarchische CA Ansätze

Nicht-hierarchische Methoden für Skalenkonstruktionen zu verwenden wurde erstmals von Loevinger et al. (1953) vorgeschlagen. In dem von ihm eingeführten Algorithmus wird zunächst eine Ähnlichkeitsmatrix erstellt, dann werden die drei Items gesucht, die die größte Ähnlichkeit zueinander haben und daraus wird ein Anfangscluster gebildet. Es werden iterativ Items zu diesem Cluster hinzugefügt, die die Saturierung, definiert als das Verhältnis der Inter-Item-Kovarianz zu der Gesamtvarianz, erhöhen und Items entfernt, die die Saturierung verringert. So erhält man den ersten homogenen Subtest, also das erste Cluster. Anschließend wird mit der gleichen Vorgehensweise ein weiteres Cluster aus den verbleibenden Items gebildet. Dies wird so lange wiederholt, bis keine Items mehr übrig sind. Diese Cluster sind dann sowohl maximal homogen als auch maximal unabhängig.

Eine weitere Möglichkeit, nicht-hierarchisch zu clustern ist das k-means Clusterverfahren. Der Algorithmus wurde von Hartigan und Wong (1979) beschrieben und wurde zur Clusterung von Objekten, also Fällen, entwickelt. Die Idee der k-means Clusterung von Personen ist, dass die Personen Punkte im Koordinatensystem darstellen und der quadratische Abstand zwischen jeder Person eines Clusters und seinem Mittelpunkt iterativ minimiert wird. Eine Clusterung von Variablen mithilfe des k-means ist auch möglich und findet bislang vor allem im Zusammenhang mit Genexpressionsdaten Verwendung (Vigneau & Qannari, 2003; Chavent, Kuentz, Liquet & Saracco, 2011; Chavent, Genuer, Kuentz-Simonet, Liquet & Saracco, 2013). In dem R-package „psych“ (Revelle, 2011) wird eine Methode beschrieben, den k-means Algorithmus auf die Clusterung von Items eines Fragebogens anzuwenden. Dabei wird eine k-means Clusterung der Personen so transformiert, dass eine Zuordnung der Items zu Untertests entsteht. Auch dieser Algorithmus basiert auf dem ICLUST Verfahren (Revelle, 1979). Das k-means Verfahren wird hier also wie bekannt zur Clusterung von Personen verwendet, wohingegen die Clusterung der Items streng genommen keine k-means Clusterung ist. Insgesamt wird das Verfahren vom

Autor jedoch als eine Anwendung von k-means zur Clusterung von Items bezeichnet (für Details sei auf das Handbuch des R-Paketes „psych“ Revelle, 2011 verwiesen). Eine direkte k-means Clusterung der Items wurde bisher noch nicht verwendet und wird in dieser Arbeit in Kapitel 3 vorgestellt.

So wie in der FA (siehe Kapitel sec:FA) besteht auch bei Clusterverfahren das allgemeine Problem, dass die Anzahl der Cluster bestimmt werden muss. Bei hierarchischen Verfahren wird dies über Inspektion des Dendrograms gemacht, was ein relativ subjektives Kriterium ist. Bei nicht-hierarchischen Verfahren gibt es hierfür verschiedene Ansätze jedoch keine allgemeingültige Lösung (Milligan & Cooper, 1985).

Clusterwerte

Wie oben bereits beschrieben, gibt es in der FA ein Unbestimmtheitsproblem der Faktoren. Wenn man jedoch Werte von Personen auf Clustern bestimmen möchte, sind diese, genau wie bei der PCA, durchaus bestimmt. Sie sind einfach gewichtete Linearkombinationen der Variablen, wie in Gleichung (1.1) dargestellt. In Matrixschreibweise lässt sich die Matrix der Clusterwerte folgendermaßen darstellen:

$$\mathbf{C} = \mathbf{X}\mathbf{W}\mathbf{X}^T \tag{1.9}$$

wobei $\mathbf{C}(n \times m)$ die Matrix der Clusterwerte ist, $\mathbf{W}(p \times m)$ eine Gewichtungsmatrix und $\mathbf{X}(n \times p)$ die Datenmatrix. Wenn man die hierarchische Clusterung wählt, ist die Gewichtungsmatrix lediglich eine Matrix von 0en und 1en, die angeben, zu welchem Cluster das jeweilige Item gehört, da nicht angegeben werden kann, wie gut das Item das jeweilige Cluster repräsentiert, was in der PCA über die Ladungen gemacht wird (Revelle, 2014). Auch in der von Revelle (2011) vorgeschlagenen Methode, die Personen mit k-means zu clustern und auf dieser Basis eine Clusterung der Items vorzunehmen, sind die Gewichte nur 0 und 1. Wenn man jedoch die Items direkt mit einem k-means Algorithmus clustert, können äquivalent zu den Ladungen in der PCA, die Gewichte als Distanzen der Items zu dem Clustermittelpunkt angegeben werden.

In der vorliegenden Arbeit wird ein neues k-means Verfahren zur Clusterung von Items entwickelt und mit der FA verglichen. Im Folgenden soll nun erklärt werden, welche Möglichkeiten es gibt, diese Methoden zu vergleichen bzw. ihre Funktion zu untersuchen.

1.2 Vergleich von Methoden mit Hilfe von Simulationen und echten Daten

Wie im vorherigen Kapitel deutlich wurde, beschäftigen sich gerade Sozialwissenschaften oft mit relativ komplexen statistischen Modellen, z.B. Faktormodellen, die die Messung von nicht-beobachtbaren, latenten Variablen, beschreiben. Die zugehörigen Parameterschätzungen (der Ladungen und Faktorkorrelationen) werden meistens über ML-Schätzungen vorgenommen. Die asymptotischen Eigenschaften (also für für $n \longrightarrow \infty$)

dieser ML-Schätzer unter der Voraussetzung der Normalverteilungsannahme können zwar analytisch ermittelt werden, es ist jedoch nicht bekannt, wie sie sich bei endlicher Stichprobengröße und Verletzung der Normalverteilungsannahme verhalten (Boomsma, 2013). Außerdem gibt es bei der FA, wie bereits erwähnt, verschiedene konkurrierende Methoden z.B. zur Ermittlung der Faktoranzahl und für die Rotation, deren Effektivität verglichen werden muss. Aus diesem Grund, werden oftmals empirische Methoden verwendet, um die Eigenschaften der Schätzwerte der FA zu untersuchen. Diese Methoden werden *Monte-Carlo-Simulation* oder *Monte-Carlo-Studie* genannt. Eine Monte-Carlo-Studie bezeichnet einen computationalen Algorithmus, der zufällig verschiedene Stichprobendaten aus einer spezifizierten Population generiert, die auf einem datengenerierenden Prozess (DGP) basiert. (siehe z.B. Carsey & Harden, 2013, S. 4). Der DGP ist dabei der Mechanismus, der die Population charakterisiert, aus dem die simulierten Stichproben gezogen werden (Carsey & Harden, 2013, S. 2). Diese Studien bekamen in den letzten etwa vier Jahrzehnten eine immer größere Bedeutung, was besonders durch die rasante Weiterentwicklung der Computertechnologie, sowohl in Geschwindigkeit als auch Arbeitsspeicher, gefördert wurde.

Carsey und Harden (2013, S. 4) vergleichen Monte-Carlo-Simulationen mit Experimenten, die in einem Labor gemacht werden. Die Bedingungen können von dem Experimentator kontrolliert werden und das Ergebnis kann so leicht auf eine bestimmte Variation einer einzigen Variable zurückgeführt werden. Aus diesem Grund sind kausale Rückschlüsse in Monte-Carlo-Simulationen leicht zu treffen. Sie bringen jedoch auch den von Labor-Experimenten bekannten Nachteil mit sich, die mangelnde ökologische Validität. Auch bei Simulationen ist es fraglich, ob die Ergebnisse auf Situationen außerhalb der simulierten Welt übertragbar sind. Trotzdem bringen uns Simulationsstudien (genau wie Experimente) ein sehr mächtiges Werkzeug zur Überprüfung von Kausalzusammenhängen.

In diesem Kapitel wird zunächst die allgemeine Funktionsweise von Monte-Carlo-Simulationen erklärt und anschließend Eigenschaften von Schätzern vorgestellt, die mittels Simulationen ermittelt werden können. In Abschnitt 1.2.3 wird dann speziell auf die Simulation von Faktormodellen eingegangen und im letzten Abschnitt dieses Kapitels wird noch die Methode des *Resamplings*, eine Simulation aus echten Daten, erklärt.

1.2.1 Generierung zufälliger Zahlen

Die Idee, die Monte-Carlo-Simulationen zugrunde liegt, ist, dass es einen DGP gibt. Er besteht aus einem systematischen, sowie einem stochastischen Teil. Der systematische Teil ist ein Modell, das im Vorhinein spezifiziert wird. Es enthält entsprechende Modellgleichungen und einen Vektor der Modellparameter $\boldsymbol{\theta}$. Der stochastische Teil gibt die Eigenschaften der zufälligen Abweichungen vom Modell in den Stichproben an (Mooney, 1997, S. 5). Es wird also eine Verteilungsannahme gemacht. Es werden nun verschiedene Simulationsbedingungen spezifiziert, wobei sowohl das Modell verändert werden kann, als auch der stochastische Anteil. Diese verschiedenen Simulationsbedingungen sind die Zellen des experimentellen Designs. Aus der aus dem DGP spezifizierten Populationsverteilung werden Zufallszahlen gezogen. Meistens werden für jede Zelle des Designs Stichproben verschiedener Größe

generiert mit einer bestimmten Anzahl an Replikationen (meistens 1000 mal). Bei jeder Replikation wird das zu untersuchende Verfahren, bzw. der Schätzalgorithmus angewandt. Man erhält pro Replikation ein oder mehr Ergebnisse eines Verfahrens, die über alle Replikationen hinweg in einem Vektor betrachtet werden können (ein Ergebnis könnte beispielsweise die angezeigte Anzahl an Faktoren eines Verfahrens in einer Stichprobe sein). Außerdem kann man sich pro Replikation auch andere Werte ausgeben lassen, wie z.B. Standardfehler, Berechnungsdauern oder ob das Verfahren konvergiert ist oder nicht. Über die verschiedenen Stichproben hinweg kann man nun also die Eigenschaften der Verfahren untersuchen. Dazu muss ein Kriterium festgelegt werden, das etwas über die Funktionsweise bzw. die Qualität des Verfahrens aussagt. Um sich das zu überlegen, ist es sinnvoll, zunächst drei Schlüsseleigenschaften statistischer Schätzer zu betrahten: *Erwartungstreue*, *Effizienz* und *Konsistenz*. Für viele bekannte Parameterschätzer, wie zum Beispiel das arithmetische Mittel, können die genannten Eigenschaften auch analytisch überprüft werden. Die verwendeten Beweise gelten jedoch immer nur unter bestimmten Bedingungen in der Population, wie z.B. Normalverteilung. Zudem gibt es auch Schätzfunktionen, die so kompliziert sind, dass noch keine analytischen Verfahren gefunden wurden (und vielleicht auch nie gefunden werden), ihre Eigenschaften zu zeigen (Mooney, 1997, S. 1). Für solche Fälle werden Monte-Carlo-Simulationen verwendet, um die Eigenschaften zu überprüfen. Im Folgenden sollen diese drei wesentlichen Eigenschaften erklärt werden, sowie Möglichkeiten dafür aufgezeigt werden, sie mit Hilfe von Simulationsstudien zu zeigen.

1.2.2 Eigenschaften von Schätzern

Erwartungstreue

Ein statistischer Schätzer ist *erwartungstreu* oder *unverzerrt* wenn sein Erwartungswert gleich dem wahren Wert des zu schätzenden Parameters ist. Die Differenz zwischen Erwartungswert des Schätzers und wahrem Wert wird *Bias* genannt. Ist der Bias eines Schätzers nicht null, so wird er *verzerrt* genannt.

Oder formal ausgedrückt: Angenommen der Parameter θ soll über den Schätzer T aus gegebenen Daten geschätzt werden. Dann ist der *Bias* des Schätzers (relativ zu dem Parameter θ) definiert als:

$$Bias_\theta(T) = \mathbb{E}_\theta(T) - \theta = \mathbb{E}_\theta(T - \theta) \tag{1.10}$$

Ein Schätzer ist unverzerrt oder erwartungstreu wenn sein *Bias* für alle Werte des Parameters θ null ist. Wenn ein Schätzer erwartungstreu ist, bedeutet das nicht, dass jede einzelne Schätzung genau dem wahren Parameter entspricht, sondern lediglich, dass die Schätzer von wiederholten Schätzungen um den wahren Wert herumliegen (siehe z.B. Dekking, Kraaikamp, Lopuhaä & Meester, 2006, S. 288-290), also im Mittel den wahren Wert ergeben.

Da man in Monte-Carlo-Studien den wahren Wert des Parameters vorgibt, indem man mit ihm das Modell spezifiziert, hat man die vorteilhafte Situation, dass der wahre Wert des

Parameters bekannt ist. Man kann also für jede Simulationsbedingung, über alle wiederholten Stichproben hinweg, einen *Bias* berechnen, um die Erwartungstreue des Verfahrens oder Schätzers zu bestimmen.

Effizienz

Die Effizienz eines Schätzers bezieht sich auf seine Variabilität. Ein Schätzer ist effizient, wenn er von Stichprobe zu Stichprobe wenig variiert. Die Effizienz eines unverzerrten Schätzers T des Parameters θ ist definiert über seine kleinstmögliche Varianz dividiert durch seine tatsächliche Varianz:

$$e(T) = \frac{1/\mathbb{I}(\theta)}{Var(T)} \tag{1.11}$$

wobei $\mathbb{I}(\theta)$ die Fischerinformation der Stichprobe ist. Mit Hilfe der Cramer-Rao-Schranke kann gezeigt werden, dass $e(T) \leq 1$ (Rao, 1945; Cramér, 1946). Im Allgemeinen wird die Effizienz als nicht so wichtig erachtet, wie die Erwartungstreue, da man bei einem erwartungstreuen Schätzer nach vielen Schätzungen im Mittel den richtigen Wert bekommt. Jedoch sollte man berücksichtigen, dass die meisten Schätzungen in der Praxis auf nur einer Stichprobe basieren und bei geringer Effizienz die Wahrscheinlichkeit gering ist, dass ein einzelner Schätzer nahe bei dem wahren Wert liegt (Carsey & Harden, 2013, S.84-85).

In Simulationsstudien wird die Variabilität des Verfahrens bzw. des Schätzers ermittelt, indem man die Varianz bzw. Standardabweichung (SD) der Ergebnisse einer Simulationsbedingung über die Stichproben hinweg bestimmt.

Konsistenz

Für einen konsistenten Schätzer gilt, dass er in Erwartung immer näher am wahren Parameterwert liegt je größer der Stichprobenumfang. Formal entspricht die Konsistenz gewöhnlich der stochastischen Konvergenz: Eine Zufallsfolge von Schätzern T_n ist schwach konsistent, wenn gilt

$$T_n \xrightarrow{\mathbb{P}} \theta \tag{1.12}$$

Der Schätzer konvergiert also in Wahrscheinlichkeit gegen den wahren Parameter mit steigendem Stichprobenumfang n, was durch das $\mathbb{P}$ über dem Pfeil formalisiert ist. Ein konsistenter Schätzer kann erwartungstreu sein, muss es aber nicht. Das heißt, bei kleiner Stichprobe muss sein Erwartungswert nicht dem wahren Wert entsprechen. Außerdem muss seine Varianz mit steigendem Stichprobenumfang immer geringer werden. Bei unendlicher Stichprobengröße ist ein konsistenter Schätzer also unverzerrt mit Varianz null.

Durch die Verwendung unterschiedlich großer Stichproben kann in Monte-Carlo-Simulationen auch getestet werden, ob ein Schätzer mit steigender Stichprobengröße in Richtung wahrem Wert konvergiert.

Wie bereits erwähnt, kann es für Schätzer des Faktormodells sinnvoll sein, Monte-Carlo-Simulationen zu verwenden, um ihre Eigenschaften zu zeigen. Wie diese Simulation funktioniert, soll im Folgenden genauer beschrieben werden.

1.2.3 Simulation von Faktormodellen

Eine der wichtigsten Entscheidungen bei der Monte-Carlo-Simulation von Faktormodellen ist, die Wahl das Populationsmodells. Dabei ist die entscheidende Frage wieder die bezüglich der ökologischen Validität: Auf welche Art von Modellen sollen die Ergebnisse der Studie generalisierbar sein? Dazu empfehlen Paxton, Curran, Bollen, Kirby und Chen (2001), dass man Anwenderstudien aus verschiedenen Zeitschriften über mehrere Disziplinen hinweg prüft, auf die man seine Ergebnisse verallgemeinern möchte. Bei Faktormodellen können Ladungen, Faktorkorrelationen und Residualvarianzen spezifiziert werden. Natürlich ist es auch möglich Modellverletzungen zu spezifizieren, wie z.B. Residualkorrelationen, die auch Teil des simulierten Modells oder der Verteilungsannahme und damit des DGP's sind. Aus den Modellgleichungen und den zugehörigen Parametern, lässt sich dann direkt die Populationskovarianzstruktur ableiten. Diese wird als Kovarianzmatrix der multivariaten Populationsverteilung verwendet, aus der dann die zufälligen Stichproben gezogen werden. Normalerweise wird für die multivariate Populationsverteilung eine multivariate Normalverteilungsdichte verwendet, es sind aber auch diskrete, ordinale Variablen oder nicht-symmetrische kontinuierliche Zufallsvariablen möglich (siehe z.B. Boomsma, 1983).

Ergebnisse, die bei Monte-Carlo-Simulationen von FA's interessant sein können, sind zum Beispiel die Schätzer der Modellparameter, oder die Schätzer ihrer zugehörigen Standardfehler, deren Eigenschaften dann untersucht werden. Außerdem könnte eine interessierende Statistik auch eine Bernoulli Variable sein, die anzeigt, ob eine generierte Zufallsstichprobe zu einer konvergierenden Lösung führt oder nicht.

Die meisten bekannten Simulationsstudien zur EFA haben streng genommen nicht die Effektivität der FA untersucht, sondern die der Methoden, die erstellt wurden, um die Anzahl der zu extrahierenden Faktoren zu bestimmen (z.B. Tucker, Koopman & Linn, 1969; Zwick & Velicer, 1982; Hakstian, Rogers & Cattell, 1982; Zwick & Velicer, 1986; Gerbing & Hamilton, 1996; Crawford et al., 2010). Nach Kenntnisstande der Autorin ist die erste bekannte Studie dazu, die von Tucker et al. (1969), in der das Eigenwert-größer-1-Kriterium, zur Bestimmung der Faktorenanzahl, untersucht wurde. Eine sehr wichtige Frage in diesem Zusammenhang war die, wie groß eine Stichprobe sein muss, um eine akkurate Bestimmung der Faktorenanzahl zu erhalten (für einen Überblick sei auf MacCallum, Widaman, Zhang & Hong, 1999 verwiesen). Eine Übersicht über wichtige Entscheidungen, die bei solchen Studien getroffen werden, wird im folgenden Kapitel 2.1 gegeben. Eine umfassende Zusammenfassung aller bisherigen Monte-Carlo-Studien im Bereich der EFA fehlt leider noch, soll jedoch auch nicht Gegenstand dieser Arbeit sein. Ein möglicher Grund, weshalb es hierzu keine Zusammenfassung gibt, könnte auch sein, dass Monte-Carlo-Studien zur EFA allgemein nicht so viel Aufmerksamkeit geschenkt wurde, wohingegen sehr viel Arbeit in diesem Feld bisher in die Analyse der CFA bzw. von Strukturgleichungsmodellen gesteckt wurde. Einer Zusammenfassung von Boomsma (2013) ist zu entnehmen, dass seit der Gründung des Journals *Structural equation modeling* 31 % der Artikel reine Monte-Carlo-Studien waren. Für einen detaillierten Überblick über bisherige Monte-Carlo-Studien zu CFA und Strukturgleichungsmodellen sei auf die Studie von Gerbing und Anderson (1992) verwie-

sen, in der alle bedeutenden Monte-Carlo-Simulationen zu Strukturgleichungsmodellen bis 1992 zusammengefasst sind. Ein erheblicher Mangel besteht bei der Modulation von Modellfehlspezifikationen für EFA-Verfahren. Dies ist ein Thema, das gerade in den letzten Jahrzehnten in der Forschung zur CFA große Aufmerksamkeit gefunden hat (z.B. Marsh, Balla & McDonald, 1988; La Du & Tanaka, 1989; Bentler, 1990). Im Bereich der EFA gibt es allerdings nur eine einzige Studie, die sich diesem Thema widmete (MacCallum, Widaman, Preacher & Hong, 2001).

Es ist also festzustellen, dass es sich bei der Monte-Carlo-Simulation um ein modellbasiertes Verfahren handelt, das vor allem dann sinnvoll ist, wenn eine zugrundeliegende Modellstruktur angenommen wird. Es gibt jedoch bei empirischen Evaluationsverfahren statistischer Methoden, genau wie bei den Methoden selbst, die Möglichkeit, ohne Modell auszukommen. Eine solche Möglichkeit ist die *Resampling* Methode. Mit diesem Verfahren können zwar auch Modellparameter untersucht werden, es kann jedoch auch ohne Modellannahmen angewandt werden.

1.2.4 Resampling Methoden

Bei *Resampling* Methoden werden, wie in Monte-Carlo-Studien, mehrere Stichproben gezogen, allerdings nicht aus einer Verteilung, die aus einem spezifizierten theoretischen DGP bzw. Modell generiert wurde, sondern aus beobachteten Daten. Der Forscher weiß bei *Resampling* Methoden also nicht, welcher Prozess den Daten zugrundeliegt, bzw. was das datengenierende Modell ist oder ob es überhaupt ein solches gibt. Es wird unterschieden zwischen verschiedenen Techniken, wie zum Beispiel *Jackknifing* oder *Bootstrapping*. Beim *Jackknifing* werden neue Stichproben erstellt, indem iterativ immer eine Beobachtung oder eine Gruppe von Beobachtungen entfernt wird. *Bootstrapping* bedeutet, dass wiederholt mit Zurücklegen Beobachtungen aus dem Datensatz gezogen werden, wobei die Stichprobengröße des gezogenen Datensatzes genau der Stichprobe des Ausgangsdatensatzes entspricht (Carsey & Harden, 2013, p. 7).

Die Ermittlung der Gütekriterien erfolgt bei *Resampling* ähnlich wie in Monte-Carlo-Studien. Die Erwartungstreue kann im Allgemeinen jedoch nicht bestimmt werden, da der wahre Wert nicht bekannt ist. Dieses Problem kann umgangen werden, wenn man die Ergebnisse des Gesamtdatensatzes als wahre Werte betrachtet und die Abweichungen der Stichprobenwerte davon als Bias ansieht. Dabei besteht dann jedoch das Problem, dass stichprobengrößensensitive Methoden im Gesamtdatensatz ein verzerrtes Ergebnis liefern werden. Wie man diesen Effekt durch eine anschließende Monte-Carlo-Studie entdecken und kontrollieren kann, wird im folgenden Kapitel 2 genauer erklärt. Es kann im *Resampling* zwar auch angenommen werden, dass es ein Modell gibt, muss aber nicht. Der Vorteil ist jedoch, dass man kein willkürliches Modell spezifizert, sondern eine reale Datensituation hat, was auch zur Evaluation von modellbasierten Methoden, wie der EFA, die auch in realen Datensituationen bestehen können soll, sinnvoll ist.

Resampling mit EFA

In der Untersuchung der Funktionsweise der EFA wurden bislang ein paar wenige Monte-Carlo-Studien durchgeführt (siehe Kapitel 1.2.3), und in manchen Studien wurde an einem echten Datensatz getestet, wie plausibel oder sinnvoll das Ergebnis der FA ist. Es wurden jedoch selten systematische *Resampling* Methoden angewandt. *Resampling* in Verbindung mit EFA wird meist in dem Zusammenhang erwähnt, dass für EFA eine Kreuzvalidierung mittels CFA gefordert wird (Mulaik, 1987). Dazu werden Stichproben aus dem Ursprungsdatensatz gezogen, die EFA wird an einer Stichprobe durchgeführt und die CFA an einer anderen. Trotz der seltenen Anwendung des *Resamplings* in diesem Bereich, gibt es jedoch drei Studien, die sich mit der Frage der minimalen Stichprobengröße für EFA's beschäftigt haben und dafür echte Datensätze verwendeten, aus denen Stichproben gezogen wurden.

Barrett und Kline (1981) untersuchten zwei große empirische Datensätze: Der eine enthielt 491 Versuchspersonen, die Cattells 16 Persönlichkeitsfaktoren-Test (16 PF, Cattell & Eber, 1972) bearbeitet hatten. Der zweite bestand aus Maßen von 1.198 Versuchspersonen, die den Eysenck Personality Questionnaire (EPQ, Eysenck & Eysenck, 1975) beantwortet hatten. Aus jedem Datensatz zogen Barrett und Kline Unterstichproben verschiedener Größen und führten je eine FA durch. Sie verglichen die Ergebnisse der Unterstichproben mit den Ergebnissen, die anhand der gesamten Stichprobe gewonnen wurden. Eine ähnliche Studie wurde von Arrindell und Ende (1985) durchgeführt. Sie untersuchten zwei große Datensätze zu zwei verschiedenen Angstfragebögen. Auch sie zogen Unterstichproben und verglichen die Ergebnisse dieser Stichproben mit dem Ergebnis des Gesamtdatensatzes.MacCallum et al. (2001) verwendete in seiner Studie einen ähnlichen Ansatz. Sie zogen ebenfalls Unterstichproben aus einem großen Datensatz und verglichen die Ergebnisse der FA mit dem in der Gesamtstichprobe. Der ganzen *Resampling* Methode ging eine Monte-Carlo-Studie voraus. Keine dieser Studien entspricht dem klassischen *Resampling* Verfahren, wie es in der computationalen Statistik angewandt wird. So werden zum Beispiel weder die oben genannten Kriterien von *Jackknifing* noch von *Bootstrapping* komplett erfüllt. Das Verfahren ist jedoch dazu verwandt und wird daher als Beispiel genannt. In der ersten Studie der vorliegenden Arbeit, die in Kapitel 2 vorgestellt wird, wird eine *Bootstrapping Resampling* Methode angewandt, bei der jedoch die gezogenen Stichprobengrößen nicht der des Gesamtdatensatzes entsprechen, um den Effekt der Stichprobengröße auch untersuchen zu können. Dieses Verfahren wird *Real World Simulation* genannt.

In der vorliegenden Arbeit werden strukturfindende Methoden verglichen. Die modellbasiert FA wird mit einem neuen CA Ansatz verglichen, der keine Modellannahmen zugrunde legt. Der Vergleich basiert sowohl auf einer modellbasierten Monte-Carlo-Simulation als auch einem *Resampling* Verfahren ohne Modellannahmen, der *Real World Simulation*.

Kapitel 2

Was können echte Daten für Simulationsstudien leisten?

2.1 Einleitung

Im Rahmen psychologischer Fragebogendaten ist die Entdeckung latenter Strukturen eine wichtige Aufgabe. Hier müssen Items zu Gruppen zusammengefasst werden, so dass sich die Items innerhalb einer Gruppe möglichst ähnlich sind, um eindeutige Diagnosen erstellen zu können. Grundsätzlich stellt die *exploratorische Faktorenanalyse* (EFA) einen allgemein anerkannten Standard für eine solche Untersuchung einer Teststruktur und die Beschreibung von Items, die einen ähnlichen Inhalt haben, dar. Das zugrundeliende Verfahren entspricht dem der FA, so wie sie in Kapitel 1.1.2 erklärt wurde. Der erste Schritt in einem solche Strukturfindungsprozess ist es, die angemessene Anzahl zu extrahierender Faktoren zu bestimmen (*Dimensionsbestimmung*). Der zweite Schritt beinhaltet die Feststellung, welche Variable zu welchem Faktor gehört (*Variablenzuordnung*), was üblicherweise so gemacht wird, dass man jede Variable zu dem Faktor zuordnet, auf dem sie die höchste Ladung hat.

Für die Untersuchung von EFA-Verfahren, wurden in früheren Studien sowohl echte als auch simulierte Daten verwendet. In Monte-Carlo-Studien simulierte Daten werden, wie in Kapitel 1.2.3 beschrieben, generiert, indem man eine zugrundeliegende Kovarianzmatrix des Populationsmodells spezifiziert, die als Kovarianzmatrix der Verteilung verwendet wird, aus der zufällige Daten gezogen werden (Bacon, 2001; Beauducel, 2001; Crawford et al., 2010; D. A. Jackson, 1993; Ruscio & Roche, 2012; Smith, 1996; Velicer, Eaton & Fava, 2000; Velicer & Fava, 1998; Zwick & Velicer, 1986). Bei der Verwendung von echten Daten ist hingegen das wahre, zugrundeliegende Modell unbekannt und kann deshalb nicht getestet werden. Eine Möglichkeit ist es, die Ergebnisse von Unterstichproben mit dem zugehörigen Ergebnis der Gesamtstichprobe zu vergleichen, eine Methode, die auch für die EFA zur Analyse von Stichprobengrößeneffekten schon angewandt wurde (Arrindell & Ende, 1985; Barrett & Kline, 1981; MacCallum et al., 2001). Das genau Vorgehen in diesen Studien ist in Kapitel 1.2.4 beschrieben.

Bestehende EFA Monte-Carlo-Studien basieren auf relativ künstlichen Bedingungen

und die meisten von ihnen beruhen auf einfachen Populations-Faktor-Modellen. In der Regel werden überhaupt nur ein paar der möglichen Modellparameter in die Simulation integriert, während die anderen auf null gesetzt werden. Modellverletzungen sind fast nie einbezogen. Wie in Kapitel 1.2.3 bereits angemerkt, ist die Berücksichtigung von Modellverletzungen in Monte-Carlo-Simulationsstudien zur Untersuchung der *konfirmatorischen Faktorenanalyse* (CFA)durchaus weit verbreitet. Hier können Modellfehler explizit durch Residualkorrelationen, also Korrelationen der Residualterme, modelliert werden (z.B. Hu & Bentler, 1999). Ein Großteil der Simulationsstudien mit der EFA basieren jedoch auf dem Idealfall einer exakten Passung des gemeinsamen Faktormodells in der Population. Das stellt eine große Einschränkung dieser Studien dar, insbesondere da das Ergebnis der EFA stark von Modellabweichungen beeinflusst wird (MacCallum et al., 2001). Ganz allgemein fehlt es Studien, die auf simulierten Daten basieren, meist an einem Bezug zu Bedingungen, wie sie in echten Daten vorgefunden werden. Zum Beispiel wurden in Studien, die den Einfluss von unterschiedlichen Ladungshöhen und Kommunalitäten untersuchten, die Nebenladungen in der Regel auf null gesetzt (Gerbing & Hamilton, 1996; Mundfrom, Shaw & Ke, 2005; Velicer & Fava, 1998). Oder wenn Nebenladungen mit einbezogen wurden (Bacon, 2001; Sass & Schmitt, 2010; Sass, 2010; Zwick & Velicer, 1982), wurde meistens entweder ein Zweifaktorenmodell untersucht oder es wurde nur eine oder zwei hohe Nebenladungen pro Variable spezifiziert, während die anderen auf null fixiert wurden. Psychometrische Daten in realen Situationen hingegen beinhalten oft mehrere Faktoren und viele kleine Nebenladungen von allen Items auf allen Faktoren (z.B. Church & Burke, 1994; Haynes, Miles & Clements, 2000). In Studien, die Faktorkorrelationen untersuchten (Bacon, 2001; Crawford et al., 2010; Sass & Schmitt, 2010; Sass, 2010), wurden entweder zweifaktorielle Modelle untersucht oder alle Faktorkorrelationen gleichgesetzt. Daten aus den Sozialwissenschaften weisen jedoch meistens eine komplexe Modellstruktur mit vielen Nebenladungen und Faktorkorrelationen verschiedener Größe auf. Darüber hinaus treten in echten Daten auch teilweise große Modellverletzungen, wie z.B. Residualkorrelationen auf, da auch komplex spezifizierte Modelle nie exakt stimmen, selbst in der Population (Cudeck & Henly, 1991; MacCallum & Tucker, 1991). Modellverletzungen sind ein ernsthaftes Problem, das berücksichtigt werden muss, da sie die Bestimmung der Populationskorrelationmatrix erheblich erschweren können (MacCallum et al., 2001). Es konnte außerdem gezeigt werden, dass die Missachtung von vorhandenen Modellverletzungen zu verzerrten Schätzungen der Modellparameter führe kann (Kaplan, 1988; Yuan, Marshall & Bentler, 2003). Aus diesem Grund können Schlussfolgerungen aus Monte-Carlo-Simulationen, die keine realistischen Faktormodelle gemeinsam mit realistischen Modellverletzungen modellieren, nicht auf echte Daten übertragen werden. Dies entspricht auch den Ergebnissen einiger anderer Studien, die gezeigt haben, dass es traditionellen Simulationsstudien und Daumenregeln an Validität mangelt (Fan & Sivo, 2005; D. L. Jackson, 2007; MacCallum & Tucker, 1991; MacCallum et al., 2001). Sie haben gezeigt, dass einfache Daumenregeln für z.B. die minimale Stichprobengröße für die EFA irreführend sind. Um valide Aussagen aus Simulationsstudien ableiten zu können, müssen also die Charakteristiken des Modells und seine Verletzungen berücksichtigt werden. Diese Charakteristiken des Modells und seine Verletzungen in der Form, wie sie in einem bestimmten Datensatz auftreten, sollen im Folgenden als *Daten-*

satzcharakteristiken bezeichnet werden.

In Studien, in denen echte Daten, die Modellverletzungen aufweisen, untersucht werden, sind die Datensatzcharakteristiken unbekannt und werden nicht manipuliert. Der Vorteil der Verwendung von echten Daten ist, dass die Bedingungen in reelen Datensätzen beschrieben werden können. Allerdings können diese Ergebnisse kaum auf andere Datensätze übertragen werden, insofern, als man nicht weiß, welche Art von Datensatz das gleiche Verhalten zeigen wird und welche nicht. Wir halten es also für einen sinnvollen Ansatz, nicht nur die Funktionsweise der verschiedenen Methoden an echten Daten zu untersuchen, sondern auch die Datensatzcharakteristiken des jeweiligen Datensatzes zu untersuchen. Es sollten nicht nur alle Parameter des Modells, das in den Daten gefunden wurde, mit einbezogen werden, sondern auch die entsprechenden, aufgetretenen Modellverletzungen. Die Parameter können anschließend in einer Monte-Carlo-Studie kontrolliert modifiziert werden, um ihren spezifischen Einfluss zu bestimmen. In der vorliegenden Studie werden verschiedene Datensätze mit bestimmten Charakteristiken untersucht mit dem Ziel, solche Charakteristiken herauszufinden, die den Unterschied der Leistung der EFA in Simulationsstudien und echten Daten ausmachen. Wir nehmen Parameterschätzungen von echten Datensätzen und integrieren sie in Simulationsstudien, um ihren Einfluss zu vergleichen. Erst wenn diese Datensatzcharakteristiken alle berücksichtigt sind, soll der Einfluss der Stichprobengröße auf die Funktion der verschiedenen Methoden untersucht werden.

In einem ersten Schritt verwenden wir echte Datensätze mit mehreren tausend Fällen, um eine (endliche) Population zu definieren. Der gesamte Datensatz wird analysiert und das Ergebnis der untersuchten Methoden wird als Populationsmodell angesehen. Anschließend werden Stichproben mit Zurücklegen aus diesem Datensatz gezogen, um zu untersuchen, wie gut die gefundene Struktur in jeder dieser Unterstichproben reproduziert werden kann. Diese Analyse soll *Real World Simulation* genannt werden. Das Hauptziel dieses Prozesses ist es, einen tieferen Einblick in die Funktion der EFA unter realistischen Datenbedingungen zu bekommen, im Vergleich zu *traditionellen Simulationsbedingungen*. Im zweiten Schritt werden die Ergebnisse der *Real World Simulation* mit Ergebnissen des üblichen Simulationsstudien-Designs verglichen, in dem das gleiche hypothetische Faktormodell verwendet wird. Auf diese Art wird die Hypothese getestet, dass die meisten Methoden in *traditionellen Simulationsstudien* andere Funktionsniveaus zeigen als in der *Real World*. Die Idee ist, stichprobenbasierte Schätzer der Modellparameter aus den echten Daten als Basis für die Simulation der künstlichen Daten in der Simulationsstudie zu verwenden. Dabei suchen wir nach solchen Datensatzcharakteristiken, die für die Durchführung von *traditionellen Simulationsstudien* notwendig sind, um die *Real World* Bedingungen so gut wie möglich widerzuspiegeln. Ein Vorteil der anschließenden traditionellen Monte-Carlo-Simulation ist auch, dass hier zwar mit dem gleichen Modell wie in der *Real World Simulation* verglichen wird, aber das simulierte Modell frei von Effekten der Stichprobengröße ist. Sollte sich also in der *Real World Simulation* gezeigt haben, dass eine Methode anfällig für Stichprobengrößen ist, und aus diesem Grund, im Gesamtdatensatz andere Ergebnisse erzielte, kann das in der anschließenden Monte-Carlo-Studie noch einmal überprüft werden.

Wir verwenden zwei verschiedene psychometrische Datensätze und untersuchen fünf verschiedene EFA Methoden für die *Dimensionsbestimmung* und eine EFA Methode für

die *Variablenzuordnung*.

2.1.1 Dimensionsbestimmung

Die EFA basiert auf der Idee, dass latente Variablen die Ursache von Korrelationen zwischen Testitems sind. Folglich ist das Ziel der EFA die Reproduzierung der empirischen Korrelationsmatrix wie in Gleichung (1.6).

Es wurden verschiedene Methoden vorgeschlagen, um die Anzahl der Faktoren in der EFA zu bestimmen. Nach früheren Forschungsergebnissen scheint die *Parallelanalyse* (PA, Horn, 1965) die genaueste Methode für die Bestimmung der zu extrahierenden Anzahl an Faktoren zu sein (Fabrigar et al., 1999; Humphreys & Montanelli Jr., 1975; Lance, Butts & Michels, 2006; Patil, Singh, Mishra & Todd Donavan, 2008). Die PA basiert auf simulierten Datensätzen, die Zufallszahlen mit der gleichen Anzahl an Variablen und der gleichen Stichprobengröße beinhalten. Diese Variablen haben also keinen zugrundeliegenden gemeinsamen Faktor. Für jeden simulierten Datensatz wird eine *Hauptkomponentenanalyse* (PCA) oder eine *Hauptachsenanalyse* (PAF) durchgeführt und die gewonnenen Eigenwerte werden gespeichert. Man erhält also eine Stichprobenverteilung von Eigenwerten. Es werden die Faktoren extrahiert, deren Eigenwerte den Mittelwert der zugehörigen simulierten Faktoren bzw. deren 95%−Perzentil übersteigen. Da sich in früheren Studien herausstellte, dass das 95%−Perzentil-Kriterium bessere Ergebnisse lieferte und die Tendenz des Mittelwert-Kriterium, zu viele Faktoren vorzuschlagen, milderte (Crawford et al., 2010; Glorfeld, 1995; Hayton, Allen & Scarpello, 2004), wurde in der vorliegenden Studie dieses Kriterium gewählt.

Der *Minimum Average Partial Test* (MAP Test, Velicer, 1976) scheint die zweitbeste Methode zur *Dimensionsbestimmung* zu sein (Zwick & Velicer, 1986, 1982). Er basiert auf einer vorher durchgeführten PCA, bei der nur die erste Hauptkomponente extrahiert wird. Diese Komponenten werden dann aus den Korrelationen zwischen den Variablen auspartialisiert. Die Nebendiagonalen der resultierenden Matrix von Partialkorrelationen wird verwendet, um den *average squared coefficient*, die mittlere quadrierte Partialkorrelation, zu berechnen. Im nächsten Schritt werden die ersten zwei Hauptkomponenten aus der ursprünglichen Korrelationsmatrix auspartialisiert und es wird wieder der *average squared coefficient* berechnet. Diese Schritte werden so oft mit einer zusätzlichen Komponente pro Schritt wiederholt, bis der *average squared coeffient* nicht mehr weiter absinkt. An diesem Punkt, wenn der *average squared coeffient* minimal ist, wird angenommen, dass die Varianz in der partiellen Korrelationsmatrix keine systematische Varianz mehr beinhaltet. Die Anzahl der extrahierten Komponenten an dieser Stelle gibt dann an, wieviele Komponenten oder Faktoren extrahiert werden müssen.

Die Anzahl der zu extrahierenden Faktoren kann auch über Fit-Indizes bestimmt werden, die auf dem *Maximum-Likelihood* (ML) Ansatz beruhen, wie z.B. das *Akaike Informationskriterium* (AIC, Akaike, 1974) und das *Bayesianische Informationskriterium* (BIC, Schwarz, 1978). Für beide Indizes wird zuerst der Wert der Likelihoodfunktion L für das

geschätzte Modell maximiert. Dieser Wert wird dann auf folgende Weisen transformiert:

$$AIC = -2ln(L) + 2p \tag{2.1}$$

$$BIC = -2ln(L) + ln(n) \times p \tag{2.2}$$

wobei p die Anzahl der zu schätzenden Parameter ist und n die Stichprobengröße. Kleinere Werte zeigen einen besseren Fit an und folglich ist das Modell mit dem kleinsten AIC- oder BIC-Wert das Optimale.

Es gab eine lange Diskussion über die Frage, wie die Stichprobengröße die *Dimensionsbestimmung* in der EFA beeinflusst. Die Ergebnisse sind keineswegs eindeutig und verschiedene Forscher sind zu unterschiedlichen Ergebnissen gekommen. Wenn man ihre Ergebnisse interpretiert, müssen jedoch zwei Aspekte berücksichtigt werden. Erstens, sinkt der Schätzfehler mit steigender Stichprobengröße und deshalb werden die Parameterschätzer über die Stichproben hinweg stabiler. Zweitens wählen manche Methoden mit zunehmender Stichprobengröße komplexere Modelle aus (hier: mehr Faktoren). Der erste Aspekt wurde in einigen früheren Simulationsstudien untersucht (z.B. Beauducel, 2001; Browne, 1968; Crawford et al., 2010; Guadagnoli & Velicer, 1988; Velicer & Fava, 1998). Allerdings scheint der Effekt der besseren Schätzung der Faktoranzahl mit steigender Stichprobengröße zunehmend zu verschwinden, je mehr weitere Parameter in die Simulation eingefügt werden, wie z.B. Kommunalitäten und Anzahl an Variablen (Mundfrom et al., 2005; Pennell, 1968; Velicer & Fava, 1998). Interessanterweise konnte eine Studie, die echte Daten untersuchte, gar keinen Zusammenhang zwischen Faktorstabilität und Stichprobengröße finden (Arrindell & Ende, 1985). Der zweite Aspekt, dass die Faktorzahl mit steigender Stichprobengröße systematisch ansteigt, wurde von McDonald (1989) aufgezeigt. Er stellte eine starke Abhängigkeit von Modellkomplexität mit Stichprobengröße bei der Verwendung des AIC fest. Cudeck und Henly (1991) erklärten jedoch in ihrer Studie, dass die Abhängigkeit nicht unbedingt unerwünscht ist, in dem sie sagten, dass es besser sein kann, ein einfaches Modell in einer kleinen Stichprobe zu verwenden als eines, das zwar realistischerweise komplexer ist aber das nicht genau geschätzt werden kann. Es ist wichtig zu beachten, dass das Zielmodell des AIC sich mit der Stichprobengröße verändert, während der BIC annimmt, dass es ein wahres Modell gibt, das unabhängig von der Stichprobengröße n ist (Burnham & Anderson, 2004; Kuha, 2004). Der AIC bevorzugt Sparsamkeit weniger als der BIC, weshalb erwartet werden kann, dass der AIC komplexere Modelle eher bevorzugt als der BIC, besonders mit steigender Stichprobengröße (Vrieze, 2012). Andere Simulationsstudien zur Funktion des AIC (Homburg, 1991; Lubke & Neale, 2006) konnten keinen großen Einfluss von Stichprobengröße auf die Modellkomplexität feststellen. Sie fanden, dass der AIC die richtige Anzahl an Faktoren mit steigender Stichprobengröße immer eher findet. Diese heterogenen Ergebnisse legen nahe, dass die Charakteristiken des Datensatzes eine große Rolle spielen, wenn es um den Einfluss von Stichprobengröße auf die Funktion der Methoden geht.

2.1.2 Variablenzuordnung

Der erste Schritt, um mit der EFA-Variablen zu Faktoren zuordnen zu können, ist es, eine Ladungsmatrix zu schätzen. Dazu ist es wichtig, dass die jeweils verwendete Methode eine Ladungsmatrix schätzt, die am genauesten das zugrundeliegende Populations-Ladungsmuster wiedergibt. Die meisten der bisherigen Studien verglichen die Höhe aller Ladungen zwischen den Populations-Ladungsmatrizen und der reproduzierten Ladungsmatrix (Arrindell & Ende, 1985; Guadagnoli & Velicer, 1988; Mundfrom et al., 2005; Sass & Schmitt, 2010; Sass, 2010; Schmitt & Sass, 2011; Velicer & Fava, 1998). Nach unserem Kenntnisstand gibt es bisher nur eine Studie von Gerbing und Hamilton (1996), die nicht nur die Ladungsmatrix untersuchte, sondern auch die Zuordnung der Variablen zu Faktoren, indem sie jeden Indikator dem Faktor zuordneten, auf dem er die höchste Ladung aufwies. Sie haben jedoch keine Nebenladungen in ihre Simulation einbezogen. Um die genaue Zuordnung von Variablen zu Faktoren zu untersuchen, haben wir eine EFA mit PAF pro Datensatz durchgeführt. Die Variablen wurden zu dem Faktor zugeordnet, auf dem sie die höchste Ladung hatten.

Zusammenfassend kann gesagt werden, dass die vorliegende Studie beabsichtigt, die Funktion der EFA in *Dimensionsbestimmung* und *Variablenzuordnung* zu untersuchen unter der Verwendung einer *Real World Simulation* und einer *traditionellen Simulation*. Wir erwarten, dass sich Unterschiede zwischen den beiden Simulationen ergeben und hoffen, sinnvolle Richtlinien zu finden, wie man Simulationsstudien mit der realen Datensituation vergleichbarer machen kann.

2.2 Methode

Wir implementierten einen neuen Ansatz für die Untersuchung der Funktion der EFA in *Dimensionsbestimmung* und *Variablenzuordnung*, die *Real World Simulation*. Dafür werden als Populationsdatensätze zuerst der Normdatensatz der deutschen Version des überarbeiteten NEO Persönlichkeitsinventars (NEO-PI-R, Ostendorf & Angleitner, 2004) verwendet und als zweites die zweite Ausgabe des mehrdimensionalen Intellingenz- Strukturtests IST-2000-R von Amthauer, Brocke, Liepmann und Beauducel (2007)[1]. Beides sind weit verbreitete diagnostische Instrumente in der psychologischen Praxis. Wir haben mit jeder Methode ein Populationsmodell im Populationsdatensatz spezifiziert und dann versucht, in einer kleineren Unterstichprobe, die aus dem Populationsdatensatz gezogen wurde, das jeweilige Modell wiederzufinden. In einem zweiten Schritt wurden diese Ergebnisse mit den Ergebnissen aus der traditionell simulierten Bedingung verglichen, bei denen die Schätzer aus dem Populationsdatensatz verwendet wurden, um Daten aus einer Verteilung zu simulieren.

[1]Wir möchten uns an dieser Stelle ganz herzlich bei Herrn Ostendorf und Herrn Beauducel für die Bereitstellung der Datensätze bedanken.

2.2.1 Real World Simulation

Der NEO-PI-R Datensatz

Der NEO-PI-R ist ein weit verbreitetes Persönlichkeitsinventar, das Persönlichkeit in fünf Hauptbereichen misst: Neurotizismus, Extraversion, Offenheit für Erfahrungen, Verträglichkeit und Gewissenhaftigkeit. Jede Hauptskala ist in sechs Facetten unterteilt und acht Items messen jede Facette. Der Fragebogen besteht also aus 240 Items. Für diese Studie wurde die Selbstbild-Version verwendet, in der Teilnehmer auf einer fünfstufigen Likert-Skala von 0=*starke Ablehnung* bis 4=*starke Zustimmung* typisches Verhalten und Reaktionen angeben. Validität und Reliabilität wurde für alle Fragebögen von Ostendorf und Angleitner (2004) gezeigt. Mittelwerte, Standardabweichungen und Schiefen der Facetten-Summenwerte sind in Tabelle 2.1 dargestellt.

Wie im Handbuch des NEO-PI-R berichtet, weist die Korrelationsmatrix der Facetten moderate bis hohe Interkorrelationen innerhalb der Faktoren und niedrige Interkorrelationen zwischen den Faktoren auf (weitere Angaben bei Ostendorf & Angleitner, 2004). Der Mittelwert der Faktorkorrelationen ist -0.05 und der Mittelwert des Betrags der Faktorkorrelationen ist 0.19 mit einer Spannweite von 0.02 bis 0.52. Hauptladungen liegen im Bereich von 0.37 bis 0.88 (siehe Tabelle 2.2) Diese Werte mögen zwar besonders klein erscheinen, sie sind aber eher normal bei Persönlichkeitsfragebögen. Peterson (2000) zeigte in einer Metaanalyse zu Faktorladungen in EFA's von Fragebogendaten, dass die mittlere Faktorladung 0.32 ist, wobei 25% der Faktorladungen unter 0.23 liegen und 25% über 0.37. Nebenladungen des Datensatzes weisen einen Mittelwert von 0.01 und eine Spannweite von -0.44 bis 0.47 auf. Alles in allem hat der NEO-PI-R Datensatz relativ hohe Nebenladungen und niedrige Faktorkorrelationen.

Der NEO-PI-R Datensatz besteht aus 11,724 Versuchspersonen. Das mittlere Alter der Stichprobe ist 29.92, und streut von 16 bis 91 mit 36% Männern und 64% Frauen. Die Summenwerte der Facetten wurden berechnet, wobei Versuchspersonen mit mindestens einem fehlenden Wert pro Facette ein *NA* anstatt eines Summenwertes erhielten, sie wurden also nicht in die weiteren Analysen für diese Facette eingschlossen.

Der IST-2000-R Datensatz

Das Basismodul des IST-2000-R misst Intelligenz in drei Hauptbereichen: verbale Intelligenz, numerische Intelligenz und räumliche Intelligenz, wovon jeder in drei Untertests eingeteilt ist. Der Test beinhaltet insgesamt 180 Fragen, die nur richtig oder falsch beantwortet werden können, also dichotom sind.

Mittelwerte, Standardabweichungen und Schiefen der Untertest-Summenwerte sind in Tabelle 2.3 gezeigt. Die Hauptladungen streuen von 0.47 bis 0.83 (siehe Tabelle 2.4). Validität und Reliabilität wurden von Amthauer et al. (2007) gezeigt. Die Nebenladungen streuen von -0.15 bis 0.20 mit einem Mittelwert von 0.01. Die drei Faktorkorrelationen haben die Werte 0.66, 0.49 und 0.43. Zusammenfassend kann gesagt werden, dass die Nebenladungen geringer sind als im NEO-PI-R Datensatz und die Faktorkorrelationen viel höher. Der Normdatensatz besteht aus $1,352$ Beobachtungen. Das mittlere Alter ist 19.09

Tabelle 2.1: *Mittelwerte, Standardabweichungen und Schiefen der Facetten-Summenwerte des NEO-PI-R Datensatzes*

Facette	M	SD	Schiefe
N1 Ängstlichkeit	16.64	5.70	0.04
N2 Reizbarkeit	14.50	4.83	0.25
N3 Depression	13.89	6.05	0.32
N4 Soziale Befangenheit	16.38	4.89	0.14
N5 Impulsivität	17.09	4.55	0.07
N6 Verletzlichkeit	12.59	4.95	0.49
E1 Herzlichkeit	21.94	4.02	-0.49
E2 Geselligkeit	18.28	5.41	-0.30
E3 Durchsetzungsfähigkeit	15.64	5.29	-0.03
E4 Aktivität	18.29	4.37	-0.02
E5 Erlebnishunger	15.01	5.00	0.03
E6 Frohsinn	21.34	5.27	-0.50
O1 Offenheit für Fantasie	20.39	5.44	-0.19
O2 Offenheit für Ästhetik	22.01	5.30	-0.60
O3 Offenheit für Gefühle	23.12	4.32	-0.41
O4 Offenheit für Handlungen	17.64	4.44	-0.07
O5 Offenheit für Ideen	19.54	5.20	-0.13
O6 Offenheit des Werte- und Normensystems	21.13	3.68	-0.03
A1 Vertrauen	18.30	4.62	-0.37
A2 Freimütigkeit	17.69	4.41	-0.17
A3 Altruismus	21.80	3.91	-0.42
A4 Entgegenkommen	16.27	4.30	-0.06
A5 Bescheidenheit	17.38	4.63	-0.10
A6 Gutherzigkeit	21.19	3.60	-0.48
C1 Kompetenz	20.46	3.74	-0.35
C2 Ordnungsliebe	18.27	4.75	-0.23
C3 Pflichtbewusstsein	21.65	4.19	-0.35
C4 Leistungsstreben	18.90	4.57	-0.11
C5 Selbstdisziplin	18.38	5.35	-0.30
C6 Besonnenheit	16.27	4.78	-0.12

Anmerkungen. M=Mittelwert; SD=Standardabweichung. Fälle mit fehlenden Werten wurden entfernt.

Tabelle 2.2: *Ladungsmatrix des Populationsdatensatzes NEO-PI-R*

Facette	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
N1	0.88		0.12		0.10
N2	0.76	−0.44			
N3	0.80			−0.15	0.12
N4	0.65	0.13		−0.17	
N5	0.37	−0.18	−0.23	0.41	0.12
N6	0.77		−0.16		
E1		0.47	0.10	0.81	
E2		0.15		0.76	−0.16
E3	−0.27	−0.37	0.21	0.37	
E4		−0.22	0.36	0.50	
E5		−0.23	−0.17	0.37	
E6	−0.16	0.14		0.64	0.18
O1		−0.10	−0.21		0.62
O2	0.15	0.13			0.76
O3	0.26		0.11	0.21	0.67
O4	−0.24		−0.16	0.16	0.38
O5	−0.19	−0.15	0.12	−0.25	0.65
O6	−0.16		−0.18		0.46
A1	−0.23	0.54		0.38	
A2		0.60			
A3		0.73	0.15	0.44	
A4	−0.14	0.73			
A5	0.21	0.55			−0.12
A6	0.21	0.52		0.26	0.25
C1	−0.32		0.55	0.12	0.14
C2	0.12		0.72		
C3		0.22	0.73		
C4	0.10	−0.19	0.74		0.18
C5	−0.13		0.76		
C6		0.13	0.48	−0.40	

Anmerkungen. N=Neurotizismus; E=Extraversion; O=Offenheit für Erfahrungen; A=Verträglichkeit; C=Gewissenhaftigkeit. Promax Rotation; ML-Schätzung. Ladungen unter 0.10 wurden unterdrückt.

Tabelle 2.3: *Mittelwerte, Standardabweichungen und Schiefen der Untertests des IST-2000-R*

Untertest	M	SD	Schiefe
V1 Satzergänzungen	11.36	3.40	-0.40
V2 Analogien	10.39	3.39	-0.26
V3 Gemeinsamkeiten	10.92	3.62	-0.38
N1 Rechenaufgaben	11.18	3.68	-0.10
N2 Zahlenreihen	11.70	4.37	-0.33
N3 Rechenzeichen	11.69	4.11	-0.15
F1 Figurenauswahl	10.72	3.36	-0.02
F2 Würfelaufgaben	10.94	3.78	-0.19
F3 Matrizen	8.11	3.30	0.71

Anmerkungen. M=Mittelwert; SD=Standardabweichung; V=Verbale Intelligenz; N=Numerische Intelligenz; F=Figurale Intelligenz. Fälle mit fehlenden Werten wurden entfernt.

und reicht von 16 bis 25 mit 44% Frauen und 56% Männern.

In beiden Datensätzen waren die Facetten (im NEO-PI-R), bzw. die Untertests (im IST-2000-R), die Grundlage unserer Berechnungen. Wir betrachteten sie wie Variablen unseres Datensatzes und ordneten sie den darüberliegenden Faktoren zu. Das wurde gemacht, um das Problem von kategorialen bzw. binären Daten zu vermeiden, das in zukünftiger Forschung behandelt werden soll.

Für die Bestimmung der Faktoranzahl verwendeten wir den MAP Test, PA-PAF, PA-PCA, den AIC, und den BIC. Der AIC und der BIC basierten auf einer PAF mit ML-Schätzung und Promax Rotation. Für die Zuordnung der Facetten wurde ebenfalls eine PAF mit ML-Schätzung und Promax Rotation durchgeführt. Van der Linden, Tsaousis und Petrides (2012) zeigten in einer Studie zu Faktorkorrelationen von verschiedenen Persönlichkeitsinventaren, dass die Faktoren üblicherweise korreliert sind. Ihre Interkorrelation schwankt zwischen 0.52 und 0.67 (im Betrag) und die mittlere Interkorrelation beträgt 0.60. Aus diesem Grund wurde die oblique Promax Rotation für unsere Studie verwendet. Die Facetten wurden den Faktoren zugeordnet, auf denen sie die höchste Ladung hatten.

Stichprobenziehungen

Stichproben der Größe 100, 200, 300, 400, 500 und 1000 wurden zufällig mit Zurücklegen aus dem gesamten Datensatz gezogen, mit jeweils 1000 Replikationen. Ziehen mit Zurücklegen wurde deshalb gewählt, weil anderenfalls eine größere Stichprobe automatisch eine größere Ähnlichkeit mit dem Populationsdatensatz und folglich eine größere Erfolgsrate (Anteil identischer Anzahlen an Faktoren wie im Gesamtdatensatz) bedeutet hätte. Durch das Ziehen mit Zurücklegen, beabsichtigten wir, besser vergleichbare Bedingungen zwischen verschiedenen Stichprobengrößen zu bekommen. Bei der *Dimensionsbestimmung* wurde für jede Stichprobe der MAP Test, PA-PAF und PA-PCA, sowie AIC und BIC durchgeführt. Anschließend wurde der Prozentsatz an korrekten Faktoranzahlen, also identisch zu der

Tabelle 2.4: *Ladungsmatrix des Populationsdatensatzes IST-2000-R*

Untertest	Faktor 1	Faktor 2	Faktor 3
V1		0.74	−0.11
V2		0.67	0.18
V3	0.18	0.53	
N1	0.47	0.12	0.18
N2	0.83		
N3	0.74		
F1			0.64
F2			0.49
F3	−0.11		0.51

Anmerkungen. V=Verbale Intelligenz; N=Numerische Intelligenz; F=Figurale Intelligenz. Promax Rotation; ML-Schätzung. Ladungen unter 0.10 wurden unterdrückt.

Anzahl, die im Gesamtdatensatz gefunden wurde, bestimmt und der Mittelwert der vorgeschlagenen Anzahl an Dimensionen. Folglich hatten wir 6×5 verschiedene Bedingungen in jedem Datensatz für die Bestimmung der Anzahl der Faktoren, da die fünf Methoden mit jeweils sechs Stichprobengrößen verglichen wurden.

Um die Ähnlichkeit der Faktorlösungen in der *Variablenzuordnung* anzugeben, setzten wir die Anzahl der Faktoren auf die theoretisch angenommene Anzahl: Fünf Faktoren bei dem NEO-PI-R und drei Faktoren bei dem IST-2000-R. Die Ähnlichkeit wurde dann über Berechnung des Randindexes (Rand, 1971) bestimmt, indem die Anzahl der korrekt klassifizierten Paare von Elementen gezählt wurden. Der Randindex ist also definiert als:

$$R(C, C') = \frac{2(n_{11} + n_{00})}{n(n-1)} \tag{2.3}$$

C ist die aktuelle Clusterlösung in dem Sample, C' ist die Clusterlösung im Gesamtdatensatz, n_{11} ist die Anzahl an Paaren, die in C und C' im gleichen Cluster sind und n_{00} ist die Anzahl an Paaren, die in C und C' in verschiedenen Clustern sind.

2.2.2 Traditionelle Simulation

Wir spezifizierten das NEO-PI-R Modell mit fünf Faktoren und jeweils sechs Variablen, sowie das IST-2000-R Modell mit drei Faktoren und jeweils drei Variablen. Um die *traditionelle Simulation* mit dem *Real World* Modell vergleichbar zu machen, wurden Schätzer der Hauptladungen und Residualvarianzen aus der EFA des Normdatensatzes für die Simulation verwendet. Das Design enthielt drei fixe Bedinugen: Stichprobengröße, Nebenladungen und Faktorkorrelationen. In der ersten Simulationsbedingung wurden die Nebenladungen und Faktorkorrelationen nicht manipuliert, sondern nahmen ihre Schätzwerte aus dem Populationsdatensatz (*Populationsmodell*). Darauf aufbauend variierten wir dann auf vier verschiedene Arten, die so bestimmt wurden, dass sie möglichst gut das Standardvorgehen

Tabelle 2.5: *Faktorkorrelationsmatrix des Populationsdatensatzes NEO-PI-R*

	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
Factor 1	1.0	−0.34	0.10	−0.26	−0.08
Faktor 2		1.00	0.04	−0.02	−0.08
Faktor 3			1.00	−0.30	−0.13
Faktor 4				1.00	0.52
Faktor 5					1.00

Anmerkungen. Promax Rotation; ML-Schätzung.

bei Monte-Carlo-Simulationen widerzuspiegeln. Zweitens wurden alle Nebenladungen und drittens alle Faktorkorrelationen auf null (*keine Nebenladungen / keine Faktorkorrelationen*) gesetzt. Viertens setzten wir eine Nebenladung pro Variable auf einen hohen Wert (*eine Nebenladung*), während wir die Faktorkorrelationen von den Schätzern aus dem echten Datesatz nahmen. Und fünftens wurden die Faktorkorrelationen auf Zufallswerte in der selben Spannweite gesetzt, wie sie im Populationsdatensatz aufgetreten ist (*variierende Faktorkorrelationen*) und die Nebenladungen aus dem Gesamtdatensatz genommen. Die Werte der Nebenladungen wurden bestimmt, indem die quadrierten Nebenladungen jeder Variable aus dem realen Datensatz aufsummiert wurden und daraus die Wurzel gezogen wurde. Dadurch blieb die Kommunalität pro Variable erhalten. Die Nebenladungen und/oder Faktorkorrelationen auf null zu setzen ist ein unrealistisches, aber in Simulationsstudien oft angewandtes Vorgehen. Zunächst wurden in allen Simulationsbedingungen die Korrelation der Residualterme auch auf null gesetzt. Die resultierenden Ladungsmatrizen und die NEO-PI-R Faktorkorrelationsmatrix können in den Tabellen 2.2, 2.4 und 2.5 eingesehen werden. Später fügten wir auch Residualkorrelationen hinzu, indem wir die Residualkorrelationen, die wir bei der EFA des Gesamtdatensatzes gefunden hatten, auf die Korrelationsmatrix des *Populationsmodells* aufaddierten. Residualkorrelationen im NEO-PI-R Datensatz waren zwischen -0.10 und 0.15 und im IST-2000-R Datensatz zwischen -0.04 und 0.06. Zusätzlich wurden noch verschiedene Stichprobengrößen (100, 200, 300, 400, 500 und 1000) verwendet. Es werden multivariat normalverteilte Daten aus einer Verteilung mit spezifizierter Populationskovarianzmatrix gezogen, die aus Ladungen, Faktorkorrelationen und Residualtermen berechnet wurde.

Für die *Dimensionsbestimmung* wurde der MAP Test, PA-PAF, PA-PCA und AIC und BIC angewandt und die oben beschriebenen Variationen getestet. Das Ergebnis ist ein vollständig gekreuztes $5 \times 6 \times 5$ Design. Jede der resultierenden Bedingungen wurde 1000 mal wiederholt, um reliable Ergebnisse aus diesen Simulationen zu bekommen. Für jede Methode wurden die Erfolgsraten und die Mittelwerte der angezeigten Faktoranzahl über die Stichproben hinweg aufgezeichnet. Für die *Variablenzuordnung* wurde ein vollständig gekreuztes 6×5 Design getestet (sechs verschiedene Stichprobengrößen und fünf Simulationsmodelle). Mittelwerte und Randindizes der FA wurden dann berechnet. Alle Berechnungen wurden in der open source Software R 0.94.110 programmiert unter Verwendung des Paketes „psych“ (Revelle, 2011).

2.3 Ergebnisse

2.3.1 Real World Simulation

Dimensionsbestimmung

Im Gesamtdatensatz des NEO-PI-R ergaben der MAP Test und die PA-PCA fünf Faktoren und die PA-PAF sechs Faktoren. Folglich wurden in den Unterstichproben fünf Faktoren bei dem MAP Test erwartet, fünf Faktoren bei der PA-PCA und sechs Faktoren bei der PA-PAF. AIC erreichte seinen geringsten Wert für die maximal mögliche Anzahl an Faktoren, nämlich 29. Der BIC schlug 22 Faktoren vor [2].

Die Erfolgsraten in den Unterstichproben sind in Tabelle 2.6 angezeigt. In kleineren Stichproben schlug PA-PAF öfter vor, fünf Faktoren zu extrahieren als PA-PCA, obwohl PA-PAF im Gesamtdatensatz sechs Faktoren vorgeschlagen hatte. Aus diesem Grund erreichte PA-PAF in den Unterstichproben vergleichsweise geringe Erfolgsraten, da das Vergleichskriterium hier sechs Faktoren war. Für Stichproben mit $n > 250$ stieg die Häufigkeit von sechs vorgeschlagenen Faktoren bei PA-PAF an und die Häufigkeit von fünf Faktoren nahm ab. Dieses Ergebnis zeigt an, dass PA-PAF, zumindest in dem Real World Datensatz, von der Stichprobengröße abhängt. Sowohl AIC als auch BIC schlugen in keiner der Unterstichproben die gleiche Anzahl an Faktoren vor, wie im Gesamtdatensatz. Das lag daran, dass der AIC, wie angenommen, stark von der Stichprobengröße abhängig war (siehe Tabelle 2.7) und deshalb im großen Gesamtdatensatz sein Minimum bei der maximalen Anzahl an Faktoren (29) erreichte. Entgegen unserer Erwartungen war auch der BIC in der *Real World Simulation* stichprobengrößenabhängig und stieg bis zu acht vorgeschlagenen Faktoren bei 1000 Beobachtungen an (siehe Tabelle 2.7)

Tabelle 2.6: *Erfolgsraten für die Dimensionsbestimmung in der Real World Simulation für verschiedene Stichprobengrößen - NEO-PI-R*

n	MAP 5 Faktoren	PA-PAF 6 Faktoren	PA-PCA 5 Faktoren
100	0.619	0.023	0.582
200	0.822	0.025	0.912
300	0.922	0.025	0.988
400	0.966	0.030	0.997
500	0.977	0.038	1.00
1000	0.999	0.100	1.00

Anmerkungen. MAP=MAP Test; PA-PCA=Parallelanalyse mit Hauptkomponentenanalyse; PA-PAF=Parallelanalyse mit Hauptachsenanalyse; n=Stichprobengröße. Facetten wurden zu Faktoren zugeordnet. Alle Erfolgsraten basieren auf 1000 Replikationen.

[2] Ab 15 extrahierten Faktoren war es nicht mehr möglich, die Kommunalitäten für die FA zu schätzen, weshalb 1en in die Diagonale eingesetzt wurden.

Tabelle 2.7: *Mittlere Anzahl angezeigter Faktoren für verschiedene Stichprobengrößen in der Real World Simulation, AIC und BIC - NEO-PI-R*

	n					
	100	200	300	400	500	1000
AIC	5.40	6.72	7.80	8.54	9.12	11.29
BIC	4.19	4.98	5.23	5.69	6.32	8.38

Anmerkungen. AIC=Akaike Informationskriterium; BIC=Bayesianisches Informationskriterium; n=Stichprobengröße. Facetten wurden zu Faktoren zugeordnet.

Tabelle 2.8: *Erfolgsraten für die Dimensionsbestimmung in der Real World Simulation für verschiedene Stichprobengrößen - IST-2000-R*

n	MAP	PA-PAF	PA-PCA	AIC	BIC
	1 Faktor	4 Faktoren	2 Faktoren	4 Faktoren	4 Faktoren
100	0.99	0.18	0.58	0.00	0.00
200	1.00	0.25	0.79	0.04	0.00
300	1.00	0.28	0.93	0.13	0.01
400	1.00	0.32	0.97	0.25	0.03
500	1.00	0.35	0.99	0.43	0.07
1000	1.00	0.64	1.00	0.95	0.50

Anmkerungen. MAP=MAP Test; PA-PAF=Parallelanalyse mit Hauptachsenanalyse; PA-PCA=Parallelanalyse mit Hauptkomponentenanalyse; AIC=Akaike Informationskriterium; BIC=Bayesianisches Informationskriterium; n=Stichprobengröße. Alle Erfolgsraten basieren auf 1000 Replikationen.

Im Gesamtdatensatz des IST-2000-R schlug keine der verwendeten Methoden die theoretisch angenommenen drei Faktoren vor. Der MAP Test fand einen und PA-PCA ergab zwei Faktoren, während alle anderen Methoden vier Faktoren anzeigten. Das waren dann auch die jeweiligen Faktoranzahlen, die in den Unterstichproben erwartetet wurden. Wie man in Tabelle 2.8 sieht, reproduzierte der MAP Test fast perfekt den einen Faktor, den er in der Gesamtstichprobe vorgeschlagen hatte. Auch die PA-PAF konnte relativ gut zwei Faktoren wiederfinden, für zumindest mittlere Stichprobengrößen. Demgegenüber schafften es die anderen Methoden nicht, die vier Faktoren wiederzufinden, die sie zuvor vorgeschlagen hatten, da sie alle für die meisten Stichprobengrößen auf kleinere Faktoranzahlen kamen. Wenn man sich die mittlere Anzahl, angezeigter Faktoren ansieht (siehe Tabelle 2.7), sieht man, dass in diesem Datensatz keine der Methoden sonderlich anfällig für die Stichprobengröße war, wohingegen alle Methoden (außer der MAP Test) zumindest eine leichte Tendenz dazu zeigten. Das könnte eventuell daran liegen, dass in dem Datensatz weniger Residualkorrelationen auftreten. Es kann aber auch daran liegen, dass die Gesamtanzahl möglicher Faktoren (also die Anzahl der Variablen) viel geringer war.

Tabelle 2.9: *Mittlere Anzahl angezeigter Faktoren für verschiedene Stichprobengrößen in den Real World Stichproben - IST-2000-R*

	n					
	100	200	300	400	500	1000
MAP	1.01	1.00	1.00	1.00	1.00	1.00
PA-PAF	3.02	3.25	3.28	3.32	3.35	3.64
PA-PCA	1.64	1.86	1.97	1.99	2.00	2.00
AIC	2.22	2.93	3.12	3.25	3.45	3.95
BIC	1.30	2.14	2.72	2.97	3.07	3.50

Anmerkungen. MAP=MAP Test; PA-PAF=Parallelanalyse mit Hauptachsenanalyse; PA-PCA=Parallelanalyse mit Hauptkomponentenanalyse; AIC=Akaike Informationskriterium; BIC=Bayesianisches Informationskriterium; n=Stichprobengröße.

Variablenzuordnung

Die Faktorladungen auf den fünf Faktoren für den Gesamtdatensatz des NEO-PI-R können Tabelle 2.2 entnommen werden. Wenn die Facetten zu dem Faktor zugeordnet wurden, auf dem sie die höchste Ladung haben, zeigte die EFA eine Gruppierung der Facetten gemäß der theoretischen Annahmen, außer dass Facette $N5$ *Impulsivität*, die zu den Facetten der Extraversionsfaktors zugeordnet wurde. Deshalb erwarteten wir für einen perfekten Fit (Randindex=1.0) in den Unterstichproben, dass die EFA $N5$ auch zum Faktor Extraversion zuordnen würde. In den Unterstichproben stieg der Randindex mit steigender Stichprobengröße von 0.891 (100 Beobachtungen) auf 0.967 (1000 Beobachtungen) an.

Wenn man eine EFA mit dem Gesamtdatensatz des IST-2000-R mit drei Faktoren durchführt, zeigen alle Untertests die höchste Ladung auf dem Faktor zu dem sie theoretisch zugeordnet wurden. Das war dann also das Zuordnungsmuster, das für die Unterstichproben erwartete wurde. Dort stieg der Randindex von 0.83 für 100 Beobachtungen auf 1.00 für 1000 Beobachtungen an. Obwohl der Randindex für 100 Beobachtungen etwas geringer war als im NEO-PI-R Datensatz, war er für den IST-2000-R bei allen anderen Stichprobengrößen höher.

2.3.2 Traditionelle Simulation

Dimensionsbestimmung

NEO-PI-R: Wenn man die *keine Nebenladungen* Bedingung und die *Populationsmodell* Bedingung des NEO-PI-R vergleicht, zeigten alle EFA Methoden eine etwa gleich gute Wiedererkennungsleistung in beiden Bedingungen (siehe Tabelle 2.10). Der MAP Test und der AIC erzielten die beste Erfolgsraten. Der BIC hatte die schlechtesten Ergebnisse in beiden Bedingungen, konvergierte jedoch zu fünf Faktoren in allen *traditionellen Simulationsbedingungen* im Gegensatz zu den Ergebnissen der *Real World Simulation*. In der Bedingung, in der nur eine Nebenladung pro Variable auf einen hohen Wert gesetzt wurde, fiel die

Tabelle 2.10: *Erfolgsraten und Mittelwerte der Faktorenzahlen über alle Stichprobengrößen hinweg für alle traditionellen Simulationsbedingungen - NEO-PI-R*

	MAP	PA-PAF	PA-PCA	AIC	BIC
		Keine Nebenladungen			
Erfolgsrate	0.98	0.95	0.91	0.98	0.80
M	4.99	4.95	4.91	4.98	4.77
		Eine Nebenladung			
Erfolgsrate	0.97	0.88	0.76	0.96	0.77
M	4.97	4.88	4.76	4.97	4.77
		Keine Faktorkorrelationen			
Erfolgsrate	0.99	1.00	1.00	0.94	0.68
M	4.99	5.00	5.00	5.05	4.51
		Variierende Faktorkorrelationen			
Erfolgsrate	0.99	0.93	0.88	0.98	0.77
M	4.99	4.98	4.95	5.00	4.91
		Populationsmodell			
Erfolgsrate	0.99	0.96	0.92	0.99	0.83
M	4.99	4.96	4.92	4.99	4.83
		Populationsmodell + Residualkorrelationen			
Erfolgsrate	0.91	0.93	0.92	0.14	0.44
M	5.03	4.98	4.92	7.88	8.25
		Real World			
Erfolgsrate	0.88	0.91	0.91	0.13	0.45
M	5.06	4.99	4.92	7.83	5.65

Anmerkungen. MAP= MAP Test; PA-PCA=Parallelanalyse mit Hauptkomponentenanalyse; PA-PAF=Parallelanalyse mit Hauptachsenanalyse; AIC=Akaike Informationskriterium; BIC=Bayesianisches Informationskriterium; M=Mittelwert. Facetten wurden zu Faktoren zugeordnet. Alle Erfolgsraten basieren auf 1000 Replikationen. In den *keine Nebenladungen* und *eine Nebenladung* Bedingungen wurde die Faktorkorrelation auf die Werte des Populationsmodells gesetzt, In den *keine Faktorkorrelation* und *variierende Faktorkorrelationen* Bedingungen wurden die Nebenladungen auf die Werte des Populationsmodells gesetzt; Stichprobengrößen= 100, 200, 300, 400, 500, 1000.

Leistung des MAP Tests und beider PA Methoden auf eine Erfolgsrate von unter 0.90 ab. Der AIC und der BIC erreichten fast das gleiche Ergebnise wie in der *keine Nebenladungen* Bedingung. Wenn alle Faktorkorrelationen auf null gesetzt wurden, zeigten alle Methoden ihre besten Erfolgsraten. Das legt nahe, dass die Funktion dieser Methoden mehr von den Populationsfaktorkorrelationen beeinflusst wird, als von den Nebenladungen, obwohl die Faktorkorrelationen viel geringer waren als in anderen Persönlichkeitsfragebögen (Van der Linden et al., 2012). Wenn zufällige Werte in der Spannweite des Populationsdatensatzes ausgewählt wurden, waren die Ergebnisse beinahe genau so wie im *Populationsmodell*. Dies scheint also eine vernünftige Art sein, Faktorkorrelationen zu simulieren.

Sowohl PA-PAF als auch AIC waren nur in der Real World Bedingung für die Stichprobengröße anfällig jedoch in keiner der anderen Simulationsbedingungen, wo die Erfolgsraten mit steigender Stichprobengröße zunahmen. Ausgehend von der Tatsache, dass der einzige Parameter, der in der *traditionellen Simulation* nicht in die Berechnung der Ausgangs-Korrelationsmatrix mit einbezogen wurde, die Residualkorrelationen sind, kann man davon ausgehen, dass die von null verschiedenen Residualkorrelationen, die nur in der *Real World Simulation* vorliegen, diesen Effekt verursachten. Um diese Annahme zu testen, führten wir eine weitere Simulation durch, in der die Residualkorrelationen aus dem *Real World* Datensatz in das *Populationsmodell* miteinbezogen wurden. Die Ergebnisse dieser Simulation sind in Tabelle 2.10 dargestellt. Wie erwartet waren die Erfolgsraten aller Methoden fast identisch zu denen der *Real World Simulation* und daher für AIC und BIC wesentlich schlechter, als wenn die Populationsparameter ohne Residualkorrelationen simuliert wurden. AIC scheint, im Gegensatz zu den PA Methoden und dem MAP Test, nicht in der Lage zu sein, eine brauchbare Anzahl an Faktoren anzuzeigen, wenn Residualkorrelationen vorhanden sind, obwohl er in allen anderen Simulationsbedingungen die fünf Faktoren sehr genau findet.

Die PA-PCA hatte große Schwierigkeiten mit den vergleichsweise hohen Faktorkorrelationen des IST-2000-R Datensatzes (siehe Tabelle 2.11). Wenn diese auf null gesetzt wurden, sprang PA-PCA zurück auf fast 100% Erfolgsrate. PA-PCA scheint nur in Anwesenheit von hohen Faktorkorrelationen schlecht zu funktionieren. Der MAP Test jedoch findet die richtige Anzahl an Faktoren auch nicht, wenn die Faktorkorrelationen auf null gesetzt sind. Mit den im IST-2000-R vorliegenden Ladungs- und Faktorkorrelationsmustern scheint der MAP Test immer einen Faktor zu finden, egal welche spezielle Simulation verwendet wird. Über alle Simulationsbedingungen hinweg zeigt PA-PAF die besten Erfolgsraten über alle Simulationsbedingungen hinweg. Obwohl die PA-PAF im NEO-PI-R Datensatz eine leichte Abhängigkeit von der Stichprobengröße zeigte und aus diesem Grund keine guten Ergebnisse in der *Real World Simulation* erzielen konnte, scheint diese Methode die konsistenteste über alle Simulationsbedingungen hinweg zu sein. Sie ist die einzige Methode, die nicht so große Einbußen aufzuweisen hat, wie alle anderen Methoden, wenn Faktorkorrelationen auftreten. Sowohl AIC als auch BIC waren relativ gut in fast allen Simulationsbedingungen, obwohl sie insgesamt nicht so gut funktionierten wie im NEO-PI-R Datensatz. Der Leistungsabfall vom *Populationsmodell* ohne Residualkorrelationen zu dem *Populationsmodell + Residualkorrelationen* war jedoch lange nicht so schlimm wie im NEO-PI-R Datensatz, aufgrund der Tatsache, dass die Residualkorrelationen viel ge-

Tabelle 2.11: *Erfolgsraten und Mittelwerte der Faktorenzahlen über alle Stichprobengrößen hinweg für alle Simulationsbedingungen - IST-2000-R*

	MAP	PA-PAF	PA-PCA	AIC	BIC
		Keine Nebenladungen			
Erfolgsrate	0.00	0.93	0.09	0.81	0.56
M	1.00	3.00	1.94	2.78	2.31
		Eine Nebenladung			
Erfolgsrate	0.00	0.88	0.00	0.76	0.47
M	1.00	2.90	1.26	2.72	2.15
		Keine Faktorkorrelationen			
Erfolgsrate	0.06	0.99	1.00	0.98	0.90
M	1.47	3.01	3.00	2.98	2.90
		Variierende Faktorkorrelationen			
Erfolgsrate	0.00	0.76	0.08	1.00	0.99
M	1.00	2.99	1.96	2.81	2.41
		Populationsmodell			
Erfolgsrate	0.00	0.93	0.04	0.82	0.03
M	1.00	2.99	1.93	2.80	2.44
		Populationsmodell + Residualkorrelationen			
Erfolgsrate	0.00	0.59	0.04	0.56	0.68
M	1.00	3.36	1.95	3.16	2.50
		Real World			
Erfolgsrate	1.00	0.27	0.85	0.30	0.50
M	1.00	3.24	1.89	3.15	2.61

Anmerkungen. MAP=MAP Test; PA-PCA=Parallelanalyse mit Hauptkomponentenanalyse; PA-PAF=Parallelanalyse mit Hauptachsenanalyse; AIC=Akaike Informationskriterium; BIC=Bayesianisches Informationskriterium; M=Mittelwert. Alle Erfolgsraten basieren auf 1000 Replikationen. In den *keine Nebenladungen* und *eine Nebenladung* Bedingungen wurde die Faktorkorrelation auf die Werte des *Populationsmodell*s gesetzt. In den *keine Faktorkorrelation* und *variierende Faktorkorrelationen* Bedingungen wurden die Nebenladungen auf die Werte des *Populationsmodell*s gesetzt. Stichprobengrößen=100, 200, 300, 400, 500, 1000.

ringer waren. Aus dem gleichen Grund funktionierten sie auch besser in der *Real World Simulation*. Über alle Bedingungen hinweg hatte der BIC die geringste Erfolgsrate von allen Kriterien in beiden Datensätzen.

Vorherige Ergebnisse, dass der AIC dazu neigt, bei großen Stichproben zu viele Faktoren zu extrahieren (Cudeck & Henly, 1991; McDonald & Marsh, 1990; Vrieze, 2012), konnten nur im *Real World* Datensatz des NEO-PI-R bestätigt werden und scheint an den dort vorliegenden mittleren Residualkorrelationen (-0.10 -0.15) zu liegen. Dieser Effekt tritt nicht bei kleinen Residualkorrelationen, wie im IST-2000-R (-0.04 -0.06) auf.

Alle untersuchten Methoden zur *Dimensionsbestimmung* schienen mehr von Faktorkorrelationen als von Nebenladungen in der Höhe wie sie in diesen Datensätzen auftreten, beeinträchtigt zu werden, selbst im NEO-PI-R Datensatz, wo Faktorkorrelationen vergleichsweise gering sind. Diese Abhängigkeit führte zu einem enormen Abfall in der Funktion beim IST-2000-R Datensatz, der fast komplett verschwand, wenn die Faktorkorrelationen auf null gesetzt wurden. Bei der *Dimensionsbestimmung* fanden wir substantielle Unterschiede zwischen den Methoden im Bezug auf die Anwesenheit oder Abwesenheit von Populations-Nebenladungen, wobei fast alle Methoden eine Erfolgrate von 1.00 hatten, wenn Faktorkorrelationen auf null gesetzt wurden. Diese Ergebnisse passen zu anderen, früheren Ergebnissen (Bacon, 2001; Beauducel, 2001).

Variablenzuordnung

Bei der *traditionellen Simulation* wurden die Faktorlösung mit fünf Faktoren vewendet. Die Ergebnisse sind in Tabelle 2.12 angegeben. Im NEO-PI-R Datensatz wies die EFA ihre besten Ergebnisse in der *keine Nebenladung* Bedingung auf, mit fast 100% korrekten Zuordnungen und ihr schlechtestes Ergebnis in der *eine Nebenladung* Bedingung, wo der Randindex sogar geringer war als in der *Real World Simulation*. Im IST-2000-R Datensatz war der Randindex am besten wenn alle Faktorkorrelationen auf null gesetzt wurden und am schlechtesten bei *variierende Faktorkorrelationen*, während die *eine Nebenladung* Bedingung immer noch ein schlechtes Ergebnis bekam. Dass im IST-2000-R Datensatz die Faktorkorrelation auch bei der Zuordnung der Facetten den größten Einfluss hatte, könnte an ihrem hohen Niveau in diesem Datensatz liegen. Unsere Ergebnisse bezüglich der *Variablenzuordnung* konnten frühere Ergebnisse von Gerbing und Hamilton (1996) bestätigen: Bei geringen Faktorkorrelationen und keinen Nebenladungen kann fast perfekte Strukturfindung erlangt werden. Nur wenn Nebenladungen einbezogen werden und besonders wenn sowohl Nebenladungen als auch Faktorkorrelationen vorhanden sind, sinkt die Leistung der EFA substantiell (Randindizes von 0.95 oder weniger).

2.4 Diskussion

Die vorliegende Studie untersuchte Datensatzcharakteristiken von echten Datensätzen und deren Einfluss auf die Funktion von Faktorstrukturfindungsmethoden in der Psychometrie. Als Beispiel wurden zwei große Datensätze untersucht und die Konsistenz der EFA Metho-

Tabelle 2.12: *Randindizes über alle Stichprobengrößen hinweg für Real World Simulation und traditionelle Simulation im NEO-PI-R Datensatz und dem IST-2000-R Datensatz*

	Datensatz	
Simulationsbedingung	NEO-PI-R	IST-2000-R
Keine Nebenladungen	1.00	0.98
Eine Nebenladung	0.93	0.95
keine Faktorkorrelation	0.96	1.00
Variierende Faktorkorrelationen	0.95	0.93
Populationsmodell	0.95	0.96
Populationsmodell + Residualkorrelationen	0.94	0.94
Real World	0.93	0.94

Anmerkungen. NEO-PI-R Datensatz: Facetten wurden zu Faktoren zugeordnet. Alle Randindizes basieren auf 1000 Replikationen. In den *keine Nebenladungen* und *eine Nebenladung* Bedingungen wurde die Faktorkorrelation auf die Werte des *Populationsmodell*s gesetzt. In den *keine Faktorkorrelation* und *variierende Faktorkorrelationen* Bedingungen wurden die Nebenladungen auf die Werte des *Populationsmodell*s gesetzt; Stichprobengrößen=100, 200, 300, 400, 500, 1000.

den in Unterstichproben verschiedenener Größen analysiert. Diese neue *Real World Simulation* kann das Problem der mangelnden Validität von *traditionellen Simulationsstudien*, also Monte-Carlo-Studien, überwinden, indem sie entscheidende Datensatzcharakteristiken aus echten Datensätzen findet und diese anschließend in Monte-Carlo-Simulationen systematisch manipuliert. Sie gibt damit Wissenschaftlern die Möglichkeit, viel über Datensätze im psychologischen Kontext zu verstehen. Es werden Konstellationen simuliert, die die echte Welt im psychologischen Daten repräsentieren und daher liefern die Ergebnisse dieser Simulationen praxisrelevante Informationen. Wenn unrealistische Bedingungen in Monte-Carlo-Studien angewandt werden, sind die produzierten Ergebnisse oft irreführend. Wir haben festgestellt, dass bei Spezifizierung einer hohen Nebenladung pro Variable, was ein übliches Vorgehen in *traditionellen Simulationsstudien* ist (z.B. Zwick & Velicer, 1986), alle Methoden schlechter zu funktionieren scheinen als in den anderen Simulationsbedingungen, in denen die gleiche Kommunalität pro Variable angewandt wurde. Diese Ergebnisse legen nahe, dass die Art und Weise, wie Nebenladungen manipuliert werden eine viel größere Rolle spielt, als die Höhe der Nebenladung. Aber die Ergebnisse, die hier berichtet werden, zeigen nicht nur, dass die Funktion von bestimmten Methoden von den Charakteristiken des Modells und seinen Verletzungen im jeweiligen Datensatz abhängt, sondern sich auch die Reihenfolge der Methoden verändern kann. Zum Beispiel konnten wir für den NEO-PI-R Datensatz eine enormen Abfall an Funktionsleistung von MAP Test und AIC zeigen, die ehemals die besten Methoden waren, wenn kleine Residualkorrelationen einbezogen werden, in der Höhe, wie sie im Populationdatensatz vorlagen, wobei die PA Methoden kaum davon betroffen waren.

Ein spezieller Befund dieser exemplarischen *Real World Simulation* ist, dass es keinen Sinn ergibt, generelle Daumenregeln über eine grudsätzliche minimale Stichprobengröße oder eine minimale Stichprobengröße, die nur von den Kommunalitäten abhängt,

aufzustellen. Der Einfluss der Stichprobengröße auf die Strukturfindung hängt von vielen Parametern ab. Die Ergebnisse dieser Studie zeigen, dass ein wichtiger Parameter die Residualkorrelationen in kleinem Ausmaß sind, wie sie in beinahe jedem Datensatz auftreten. Sobald sie in die simulierte Korrelationsmatrix einbezogen werden, werden manche Methoden stark von der Stichprobengröße abhängig und ihre Funktion sinkt daher mit steigender Stichprobengröße. Dieser Effekt war am stärksten für den AIC und am schwächsten für beide PA Methoden. Da der AIC gar nicht annimmt, dass es ein wahres Modell gibt, sondern nur den Vorhersagefehler minimieren will (Vrieze, 2012), ist seine Anfälligkeit für die Stichprobengröße in Anwesenheit von Residualkorrelationen nicht überraschend. Je größer die Stichprobe ist, desto mehr Residualkorrelationen können von weiteren Unterfaktoren erklärt werden und dadurch die Vorhersage optimiert werden. Man darf jedoch nicht außer Acht lassen, dass Residualkorrelationen in der Höhe wie sie im NEO-PI-R Datensatz vorliegen eine nicht-diagonale $\mathbf{U}^2-$Matrix anzeigen und Modellverletzungen implizieren, die anzeigen können, dass ein Modell-Missfit vorliegt. Folglich ist eine faire Schlussfolgerung, dass im Falle des NEO-PI-R, das 5-Faktoren Modell möglicherweise nicht komplett gilt, auch wenn es vielleicht das praktisch sinnvollste ist. Mit den vorliegenden Residualkorrelationen gibt es also viele kleine zusätzliche Faktoren, die berücksichtigt werden müssen, wenn man die beobachtete Korrelationsmatrix $\mathbf{\Sigma}(p \times p)$ erklären will. In solchen Fällen ist die Datenstruktur in obliquer EFA dann:

$$\mathbf{\Sigma} = \mathbf{\Lambda \Phi \Lambda}^T + \mathbf{M M}^T + \mathbf{U}^2 \tag{2.4}$$

wobei $\mathbf{\Sigma}$ die $p \times p$ Korrelationsmatrix der p beobachteten Variablen ist, $\mathbf{U}^2$ die $p \times p$ Diagonalmatrix der Residualvarianzen, $\mathbf{\Lambda}$ ist die $p \times m$ Matrix der Faktorladungen auf den $m \leq p$ Faktoren und $\mathbf{\Phi}$ steht für die $m \times m$ Matrix der Faktorkorrelationen. $\mathbf{M}$ ist nun die $p \times l$ Matrix der Ladungen der p Variablen auf die l Unterfaktoren.

Das kann auch der Grund sein, warum auch BIC im *Real World* Datensatz für Stichprobengrößen anfällig ist. Man kann auch vom BIC nicht erwarten, dass er konvergiert, wenn das wahre Modell nicht unter den vorgeschlagenen ist. Die Anzeige von mehr Faktoren mit steigender Stichprobengröße ist bei Daten mit Residualkorrelationen nicht unbedingt unerwünscht. Diese Entscheidung hängt jedoch immer noch von dem jeweiligen Ziel der Studie ab. Vermutlich waren alle bisher gefundenen Zusammenhänge zwischen Stichprobengröße und Modellerkennung auch von dem jeweiligen Modell und der Höhe der Modellverletzungen in den Daten, also den Datensatzcharakteristiken abhängig. Das ist eine mögliche Erklärung dafür, warum Forscher in manchen Studien einen Stichprobengrößeneffekt gefunden haben und in manchen nicht, je nach spezifiziertem Modell in der Simulation (z.B. Browne, 1968; Guadagnoli & Velicer, 1988) und den Datensatzcharakteristiken in den echten Daten (Arrindell & Ende, 1985; Barrett & Kline, 1981).

Es sollte erwähnt werden, dass es mit den Parametern, die in dieser Monte-Carlo-Studie berücksichtigt wurden, nicht möglich war, alle Variationen in der Funktion der Methoden im echten Datensatz erklären zu können. Was zum Beispiel noch erklärt werden muss, ist dass der Funktionsabfall des MAP Tests im IST Datensatz, auch nachdem die Faktorkorrelationen auf null gesetzt wurden, nicht reduziert werden konnte. Wenn man zusätzlich

noch die Anzahl der Variablen pro Faktor, oder die Anzahl der Faktoren variiert, könnte diese Problem behoben werden. Das zeigt jedoch wieder, wie wichtig die Verwendung von echten Daten ist, um Hinweise über wichtige Datensatzcharakteristiken zu bekommen.

2.4.1 Praktische Auswirkungen

Für die Anwendung der EFA in der Praxis sollte angemerkt werden, dass keine eindeutige Aussage darüber gemacht werden kann, welche Methode besser ist, auch nicht in Abhängigkeit von der Stichprobengröße. Die Funktionsunterschiede der Methoden hängen stark von Datensatzcharakteristiken und dem jeweiligen Ziel der Studie ab. Insbesondere zeigt diese Studie, dass die Ergebnisse der EFA extrem instabil gegenüber bereits kleinen Variationen bestimmter Parameter, wie z.B. Residualkorrelationen sind. Für eine umfassende Analyse der zugrundeliegenden Struktur sollten verschiedene Methoden in Kombination verwendet werden. Zum Beispiel scheinen Faktorkorrelationen das entscheidende Problem bei der Auswahl von Methoden zur *Dimensionsbestimmung* zu sein. PA-PAF scheint in Anwesenheit von Faktorkorrelationen eine gute Wahl zu sein. Außerdem sollte Anwendern geraten werden, Residualkorrelationen genauer zu untersuchen, wenn der MAP Test oder die PA verwendet werden, da diese nicht für solche Modellverletzungen anfällig sind. Sobald der AIC oder der BIC mehr Faktoren vorschlagen, als die PA Methoden und der MAP Test, sind korrelierte Residualvarianzen wahrscheinlich und Unterfaktoren sollten berücksichtigt werden. Auf jeden Fall sollten vor jeder Analyse die jeweiligen Datensatzcharakteristiken des vorliegenden Datensatzes genau untersucht werden. Vielleicht ist es letztlich sogar die beste Lösung, vor jeder Analyse eine eigene Simulation durchzuführen, um die beste Methode für den jeweiligen Datensatz auszuwählen (Muthén & Muthén, 2002).

Für die vergleichende Untersuchung der Funktion verschiedener statistischer Methoden in Monte-Carlo-Simulationen, legt die vorliegende Studie nahe, dass Charakteristiken von echten Datensätzen eingehend untersucht werden sollten, bevor die Simulationsstudie geplant wird. Auf der Basis von einzelnen Aspekten, wie Stichprobengrößen oder Anzahl an Variablen pro Faktor, können keine endgültigen Aussagen getroffen werden, solange der Einfluss der Interaktion mit anderen Datensatzcharakteristiken, wie etwa Ladungshöhen, Residualkorrelationen oder Faktorkorrelationen außer Acht gelassen wird. Mit der *Real World Simulationsmethode* können Schlüsseldatensatzcharakteristiken für die Funktion von bestimmten Methoden identifiziert werden, mit Hilfe von echten Datensätzen. Der Zusammenhang von Datensatzcharakteristiken mit speziellen Funktionsunterschieden kann anschließend für die *traditionelle Simulation* verwendet werden. Dort werden dann diese Hauptcharakteristiken systematisch variiert und anschließend kann ihre ursächliche Beziehung getestet werden. Zukünftige Arbeit sollte die folgenden Punkte behandeln: Datensatzcharkateristiken von großen Datensätzen sammeln, die entscheidend für die Funktion verschiedener Methoden sind und die Untersuchung der Funktion dieser Methoden in systematischen Simulationsstudien, wo diese realistischen Variationen getestet werden.

Kapitel 3

Ein neuer K-means Ansatz zur explorativen Clusterung von Items

3.1 Einleitung

Wenn man von explorativer Strukturfindung in psychometrischen Daten spricht, meint man meistens die Findung von homogenen Itemgruppen in Fragebögen ohne Vorinformation über die Teststruktur. Homogen bedeutet dabei in der Regel, dass die Items innerhalb einer Gruppe stark zusammenhängen. Die am häufigsten verwendete Methode für diesen Prozess ist die *exploratorische Faktorenanalyse* (EFA). Das zugrundeliende Verfahren entspricht dem der FA, so wie sie in Kapitel 1.1.2 erklärt wurde. Sie basiert auf der Idee, dass latente Variablen Zusammenhänge zwischen Testitems verursachen und dass die Zusammenhänge zwischen diesen p Items folglich von weniger als p zugrundeliegenden Faktoren erklärt werden können. Das Ziel der EFA ist also das Auffinden des „wahren“ Faktormodells wie in Gleichung (1.6) dargestellt.

Die explorative Strukturfindung mit der EFA beinhaltet zwei Schritte: Erstens, die *Dimensionsbestimmung*, also die Bestimmung der Anzahl der Faktoren m und zweitens die *Variablenzuordnung* zu den Faktoren. Der zweite Schritt, die *Variablenzuordnung* basiert direkt auf der Schätzung der Ladungsmatrix der FA wie in Kapitel 1.1.2 beschrieben, für den Fall dass die Anzahl der Faktoren m vorgegeben ist. In verschiedenen Studien mit simulierten und echten Daten hat sich diese als zufriedenstellend funktionierende Methode herausgestellt (Arrindell & Ende, 1985; Guadagnoli & Velicer, 1988; Mundfrom et al., 2005; Sass & Schmitt, 2010; Sass, 2010; Schmitt & Sass, 2011; Velicer & Fava, 1998). Zur *Dimensionsbestimmung* werden verschiedene Methoden verwendet, die der eigentlichen EFA vorgeschaltet werden. Zum Beispiel wurden zu diesem Zweck die *Parallelanalyse* (PA, Horn, 1965) und der *Minimum Average Partial Test* (MAP Test, Velicer, 1976) vorgeschlagen, deren Effektivität in mehreren Studien gezeigt wurde (Fabrigar et al., 1999; Patil et al., 2008; Velicer et al., 2000; Zwick & Velicer, 1986).

Dennoch steht die EFA wegen einiger ungelöster Probleme in der Kritik. Ein großer mathematischer Nachteil ist das Problem der *Unbestimmtheit der Faktoren*. Auch wenn das

Faktormodell auf der Korrelationsebene bestimmt ist, ist es auf Datenebene nicht bestimmt (z.B. R. E. Anderson et al., 2006). Bei der Schätzung der Faktorwerte wird es immer mehr Parameter als beobachtete Datenpunkte geben. Die Schätzung der Faktorwerte ist zwar möglich (Schonemann & Steiger, 1978), sie hat jedoch keine eindeutig definierte Lösung (für eine ausführlichere Beschreibung des Problems siehe Kapitel 1.1.2).

Andere Kritikpunkte sind eher praktischer Art. So ergeben sich beispielsweise Probleme bei kleineren Stichproben von $n \leq 150$ (Guadagnoli & Velicer, 1988; Sass & Schmitt, 2010; Velicer & Fava, 1998), wobei diese in der psychologischen Forschung nicht unüblich sind (Fabrigar et al., 1999). Tatsächlich zeigten Fabrigar et al. (1999) in einer Zusammenfassung sogar, dass 30% der damals aktuellen Studien Stichprobengrößen von $n = 100$ oder weniger aufweisen. Außerdem können mittlere bis hohe Nebenladungen einige Probleme mit sich bringen, wie zum Beispiel verzerrte Parameterschätzungen des Rests des Modells abhängig von dem gewählten Rotationskriterium (Asparouhov & Muthén, 2009; Sass & Schmitt, 2010). Andere Einwände sind eher theoretischer Natur. So wurde zum Beispiel angemerkt, es sei eine sehr vereinfachte und den reellen Umständen nicht gerecht werdende Annahme, dass Messwerte linear zusammengesetzt sind aus einem systematischen Teil, der durch wenigen Faktoren erzeugt wird plus einem Residuum, das unkorreliert sein muss (z.B. Tryon, 1935). Schon Cattell (1987, S. 101 ff.) war bewusst, dass es in Wirklichkeit in jedem Datensatz mehr zugrundeliegende, systematische Faktoren gibt als Variablen. Bis heute weisen Wissenschaftler immer wieder darauf hin, dass kein Faktormodell komplett wahr ist, auch nicht in der Population (Cudeck & Henly, 1991; MacCallum & Tucker, 1991). Während die meisten Forscher trotzdem weiterhin die EFA verwenden um zumindest das Modell zu finden, das am besten auf die Daten passt, stellt sich die Frage, ob es nicht eine andere Möglichkeit gibt. Wünschenswert wäre es die Komplexität eines Datensatzes zu reduzieren und Items zu Untertests zuzuordnen, ohne ein komplexes zugrundeliegendes Modell über die Entstehung der Messwerte zu brauchen. Für diesen Zweck wurde vor mehr als 70 Jahren die Clusteranalyse (CA) zur Clusterung von Items vorgeschlagen (Tryon, 1939). Seither besteht eine kontroverse Debatte darüber, ob die CA eine Alternative zur EFA sein könnte. Ein bekanntest Zitat in diesem Zusammenhang, ist das von Tryon und Bailey (1970): "Cluster Analysis is a poor man's factor analysis". Im Gegensatz dazu haben einige Forscher immer wieder vorgeschlagen, den Prozess der Zuordnung von Items zu Untertests mit der CA zu ergänzen oder die EFA sogar durch die CA zu ersetzen. Verschiedene *hierarchische* und *nicht-hierarchische* CA Methoden wurden vorgeschlagen (Bacon, 2001; Hunter, 1973; Revelle, 1979; Schweizer, 1991). Hunter (1973) bemerkte zum Beispiel, dass die EFA kein geeignetes Instrument für die Bestimmung der Homogenität eines Itemsatzes sei und schlug daher vor, eine CA an jede EFA anzuschließen. Für diesen Zweck schlug er den *similarity coefficient*, also *Ähnlichkeitskoeffizienten* vor. Revelle (1979) schlug für die Bestimmung der Clusteranzahlen mit psychometrischen Motiven das sogenannte ICLUST-Verfahren vor, in dem der *Reliabilitätskoeffizient beta* verwendet wird. Dieser Koeffizient stellt eine Alternative zu cronbachs alpha dar und basiert auf der geringsten aller möglichen split-half Reliabilitäten einer Skala. Das Verfahren lieferte in seinen Daten sinnvollere Ergebnisse als die EFA. Schweizer (1991) verwendete die hierarchische CA mit disaggregierten Korrelationen und Bacon (2001) entwickelte die Korrelation der Korrelation als Distanz-

maß. Loevinger et al. (1953) waren die ersten, die einen nicht-hierarchischen Ansatz in der Psychometrie verfolgten. Sie versuchten, die Homogenität innerhalb eines Clusters, ähnlich wie in der Reliabilitätstheorie, zu maximieren (siehe Kapitel 1.1.3). Niemand hat bisher versucht, k-means zur Clusterung von Items zu verwenden. Die vorliegende Studie hat zum Ziel, einen Algorithmus zu finden, der Items clustert und dabei direkt den k-means Ansatz verwendet.

Genau wie in der EFA muss in der CA vor der *Variablenzuordnung* noch eine *Dimensionsbestimmung* stattfinden. Wenn eine hierarchische CA durchgeführt wird, basiert diese Entscheidung gewöhnlich auf Informationen, die aus dem Dendrogram abgeleitet werden. Das ist jedoch nicht möglich, wenn nicht-hierarchische Verfahren verwendet werden. Hier muss der Anwender üblicherweise eine Anzahl von Clustern spezifizieren, bevor eine Clusterung durchgeführt werden kann (Milligan & Cooper, 1985). In anderen Disziplinen, die nicht-hierarchische Clusterung verwenden, wie z.B. der Genom-Forschung, wird die *Dimensionsbestimmung* normalerweise über visuelle Inspektion und Vorwissen durchgeführt, während die Passung mit den Daten eher seltener untersucht wird (Handl, Knowles & Kell, 2005). Handl et al. (2005) geben einen Überblick über Clustervalidierungsverfahren, die dazu genutzt werden können, die am besten passende Clusteranzahl zu bestimmen und manche von ihnen sind in dem R-Paket „clValid“ (Brock, Pihur, Datta & Datta, 2008) implementiert. In der vorliegenden Studie werden wir eine dieser Methoden zur *Dimensionsbestimmung* verwenden.

In diesem Artikel werden zunächst zwei neue k-means Ansätze vorgestellt, die für die Clusterung von Items entwickelt wurden. Zweitens soll ein vorhandener Ansatz zur Clusterung von Variablen erklärt werden, der bisher noch nicht im psychometrischen Kontext verwendet wurde.

3.1.1 K-means Clusterung von Items

Die Idee der k-means Clusterung von Items ist, dass die Items Punkte im Koordinatensystem darstellen und der quadratische Abstand zwischen jedem Item eines Clusters und seinem Mittelpunkt iterativ minimiert wird.

$$\underset{s}{argmin} \sum_{i}^{k} \sum_{x_j} (x_j - \mu_j)^2 \tag{3.1}$$

Als Ausgansbedingung werden k zufällige ausgewählte Items als zu jeweils einem Cluster zugehörig definiert, so dass jedes Cluster aus genau einem Item besteht. Dann werden die folgenden Schritte wiederholt bis sich das Ergebnis nicht mehr stärker verändert als ein vorgegebenes Kriterium:

1. Ordne jeden Punkt dem Cluster zu, zu dessen zugehörigem Zentroid es den geringsten Abstand hat. Dieser Zentroid ist der Mittelpukt von allen Items eines Clusters.

2. Berechne den Mittelpunkt des neuen Clusters.

Da die Anfangspunkte des Clusters zufällig gewählt werden, sind mit diesem Algorithmus verschiedene Ausgänge möglich. Aus diesem Gund wird der Algorithmus mehrmals ausgeführt und das am häufigsten vorkommende Ergebnis verwendet. Zuvor muss jedoch ein Weg gefunden werden, wie die Items in einem Koordinatensystem repräsentiert werden können. Wenn Items geclustert werden, ist normalerweise das Hauptziel die Items zusammen zu clustern, die die höchste Korrelation haben. Aus diesem Grund basieren beide hier implementierte Methoden auf Korrelationen.

3.1.2 K-means skaliertes Distanzmaß (SDM)

In dem ersten hier verwendeten Ansatz ist die Idee, Koordinatenpunkte zu erstellen, so dass die Distanz zwischen diesen Punkten direkt auf ihren Korrelationen basiert, oder um es präziser auszudrücken auf 1 minus der Korrelation (*1-cor*). Folglich sind die Items nahe beieinander, die hoch korreliert sind. Allerdings ist *1-cor* keine Metrik (van Dongen & Enright, 2012), weshalb das Distanzmaß nicht als euklidische Distanz zwischen zwei Punkten im Raum interpretiert werden kann. Von van Dongen und Enright (2012) konnte jedoch gezeigt werden, dass

$$d(A, B) = \sqrt{0.5 - 0.5 \cdot Cor(A, B)} \tag{3.2}$$

eine Metrik darstellt und deshalb soll dieses Distanzmaß hier verwendet werden. Um die Items so in einem Koordinatensystem anzuordnen, dass die Distanz zwischen zwei Items genau $d(A, B) = \sqrt{0.5 - 0.5 \cdot Cor(A, B)}$ entspricht, wurde ein Skalierungsverfahren verwendet, das auf der Idee des *multidimensional scaling* basiert. Wenn wir n Items betrachten, soll die Lösung $n - 1$ Dimensionen haben. *Multidimensional scaling* ist eine Methode, die eine Matrix von Koordinatenpunkten mit vorgegebener Anzahl an Dimensionen (D^2) aus einer Distanzmatrix (X) schätzt. X wird so gewählt, dass es eine spaltenzentrierte Matrix ist, wobei die Spaltensummen 0 ergeben. Wenn die Distanz jeder der n Punkte zueinander gegeben ist, hat man $n \cdot (n - 1)$ Gleichungen für die Distanzen. Vorausgesetzt, dass jeder Punkt eine Dimension von $n - 1$ hat, erhält man $n \cdot (n - 1)$ unbekannte Koordinaten und folglich ist die Lösung eindeutig. Wie die Skalierung im Detail funktioniert, kann bei Borg und Groenen (2005) nachgelesen werden.

3.1.3 K-means Korrelation (cor)

Die Idee des zweiten Ansatzes ist, dass die Korrelationen zwischen den Items direkt als Koordinaten der Items herangezogen werden. Der Vektor der Korrelationen eines Items zu allen anderen Items wird dargestellt als Koordinaten dieses Items. Der Abstand zwischen zwei Items ist dann

$$d(A, B) = \sum_i (Cor(A, V_i) - Cor(B, V_i))^2 \tag{3.3}$$

Alle k-means Ansätze haben einen Vorteil gegenüber hierarchischen Verfahren: Sie erlauben es, Items um einen Mittelpunkt herum zu clustern. Dieser Mittelpunkt wird von Vigneau

und Qannari (2003) *synthetische Variable* genannt und kann auch als das zu messende Konstrukt verstanden werden. Dadurch erhalten wir Distanzen zwischen Items und ihrem Mittelpunkt, also dem Konstrukt. Diese Distanzen können äquivalent zu den Ladungen in der EFA betrachtet werden. Es wird jedoch kein Messmodell aufgestellt, das die kritische Annahme unkorrelierter Residualvarianzen macht.

3.1.4 ClustOfVar

ClustOfVar (Chavent et al., 2011) ist ein Clusterverfahren, das in erster Linie zur Dimensionsreduzierung und Variablenselektion für Mischungen aus quantitativen und qualitativen Variablen entwickelt wurde (Chavent et al., 2011) und das auf Genexpressionsdaten angewandt wurde (Chavent et al., 2013). Auch hier ist die Idee, eine synthetische Variable als Mittelpunkt des Clusters zu definieren. Im Falle metrischer Variablen, ist die synthetische Variable die Variable, deren Summe der quadrierten Korrelationen zu allen Variablen des Clusters maximal ist. Es kann gezeigt werden (Vigneau & Qannari, 2003), dass diese synthetische Variable die erste Hauptkomponente aller Variablen des Clusters ist. Die Summme der quadrierten Korrelationen aller Variablen zu der synthetischen Variable wird *Homogenität* (H) genannt. Basierend auf diesen beiden Definitionen, werden zwei Clusteralgorithmen definiert:

1. Eine hierarchische Clustermethode (*ClustOfVar I*)
2. Eine k-means Clustermethode (*ClustOfVar II*)

Die erste ist eine agglomerative hierarchische Clustermethode mit dem Dinstanzmaß:

$$d(A, B) = H(A) + H(B) - H(A \cup B) \tag{3.4}$$

bei dem in jedem Schritt die zwei Cluster vereint werden, die das kleinste d haben. Die Clusteranzahl kann z.B. über Untersuchung des *Dendrograms* bestimmt werden.

In dem k-means Ansatz wird eine Ausgangsclusterung gewählt, die entweder zufällig sein kann oder aus dem hierarchischen Ansatz gewonnen wird. Basierend auf dieser Ausgansclusterung wird für jedes Cluster eine synthetische Variable, die erste Hauptkomponente der Korrelationsmatrix aller Variablen dieses Clusters, berechnet. Dann werden die Variablen zu dem Cluster zugeordnet, zu dessen Mittelpunkt sie die höchste Korrelation haben. Dieser Prozess wird iterativ optimiert bis Konvergenz erreicht ist.

Um die Anzahl der Cluster zu bestimmen, schlugen Chavent et al. (2011) vor, *Bootstrap* Stichproben aus Trainingsdaten zu ziehen und die Stabilität der resultierenden Cluster mit Hilfe des Randindexes zu untersuchen. Es wird dann die Clusteranzahl aus allen resultierenden Clusterlösungen gewählt, die die größte Ähnlichkeit aufweist. Dieser Ansatz bringt jedoch den Nachteil, dass er sehr rechenintensiv ist. Wie oben bereits erwähnt wurde, gibt es keinen allgemeinen Standard dafür, welchen Algorithmus man für die Bestimmung der Clusteranzahl in der CA verwenden soll. Brock et al. (2008) implementierten ein R Paket für verschiedene Clustervalidierungsmethoden, das auch zur Clusternanzahlsbestimmung

genutzt werden kann. Eine Gruppe von Validierungsmethoden, die sie verwenden, sind die *internen Maße*, die interne Validität messen, so dass sie die Kompaktheit, Verbindung und Trennung widerspiegeln (Brock et al., 2008). Eines dieser Maße ist die *Silhouette*-Breite.

3.1.5 Silhouette

Die *Silhouette*-Breite basiert auf dem *Silhouette*-Wert jedes Objekt, der folgendermaßen definiert ist:

$$S(j) = \frac{b_j - a_j}{max(b_j, a_j)} \tag{3.5}$$

wobei a_j der mittlere Abstand zwischen j und allen anderen Objekte des selben Clusters ist und b_j der mittlere Abstand zwischen j und allen Objekten in dem *Nächste-Nachbarn* Cluster.

Wenn Items geclustert werden, wird ein *Silhouette*-Wert für jedes Item ermittelt und diese Werte werden dann alle aufsummiert, um die *Silhouette*-Breite zu erhalten. Wenn für die vorgeschlagene Clusterlösung jeder möglichen Anzahl an Clustern eine *Silhouette*-Breite berechnet wird, kann man die Clusteranzahl auswählen, bei der die *Silhouette*-Breite am größten ist. Diese Methode soll im Folgenden *Silhouette* genannt werden.

Das Ziel dieser Studie war es zunächst, neue k-means Ansätze zu untersuchen, die noch nicht für die Clusterung von Variablen verwendet wurden. Diese Ansätze sind *k-means SDM* und *k-means cor*. Außerdem sollen zwei weitere Variablenclusterverfahren, die bisher nur für genomische Daten verwendet wurden, auf psychometrische Daten angewandt werden, *ClustOfVar I*, eine hierarchische Methode, und *ClustOfVar II*, eine k-means Methode. Deren Funktion soll, unter Verwendung von einer Simulationsstudie sowie echten Daten, mit der EFA und zwei traditionellen CA Methoden verglichen werden. Zweitens soll eine Formel zur Berechnung von Clustervalidität, *Silhouette*, als Abbruchkriterium verwendet werden und ihre Effizienz untersucht werden. Ihre Effizienz in *Dimensionsbestimmung* wird mit der Effizienz von drei bekannten Methoden der *Dimensionsbestimmung* aus der EFA verglichen: PA mit Hauptkomponentenanalyse (PA-PCA) und mit Hauptachsenanalyse (PA-PAF) nach Horn (1965) und der *Minimum Average Partial Test* (*MAP Test*) von Velicer (1976). Diese Methoden wurden ausgewählt, da ihre Genauigkeit im Vergleich zu anderen Methoden in früheren Studien gezeigt werden konnte und sie für die Verwendung in Simulationsstudien leicht automatisiert werden können (Fabrigar et al., 1999; Patil et al., 2008; Velicer et al., 2000; Zwick & Velicer, 1986).

Um einen umfassenden Eindruck der Funktion dieser Methoden zu bekommen, werden drei verschiedene Arten von Simulationen verwendet. Als erstes wird die *Real World Simulation* angewandt, ein neuer Ansatz unter Verwendung echter Daten. Dann wollen wir eine *traditionelle Simulationsstudie* durchführen, in der eine Populationsverteilung simuliert wird und verschieden Stichprobengrößen gezogen werden. Und drittens soll eine Kreuzvalidierung mit einer *konfirmatorischen Faktorenanalyse* (CFA) durchgeführt werden. Der Fokus soll jedoch auf der *Real World Simulation* liegen, da sie kein Faktormodell annimmt und damit am ehesten die oben erwähnten Vorteile der CA berücksichtigt.

Unser Ziel ist es, neue und reliable Clusteranalysenmethoden zu finden, um den explorativen Schritt der Strukturfindung in psychometrischen Daten zu unterstützen. Außerdem soll ihre Effektivität in realen Datensituationen untersucht werden. Wir gehen also davon aus, dass die neuen Clusteransätze und das Abbruchkriterium *Silhouette* dazu in der Lage sein werden, eine simulierte Teststruktur genauso gut zu finden, wie die EFA und dass sie ein Modell vorschlagen, das von der CFA akzeptiert wird. Zudem nehmen wir an, dass diese Methoden besser funktionieren als die traditionellen hierarchischen Methoden. Im Hinblick auf die typischen Charakteristiken von psychometrischen Daten, interessiert uns auch ihre Fähigkeit, mit den Problemen von psychometrischen Daten umzugehen, wie z.B. fehlende Eindimensionalität.

3.2 Methode

Wir untersuchten die Effektivität von zwei neuen k-means Clusteralgorithmen für die Clusterung von Items. Einer der beiden basiert auf dem *skalierten Distanzmaß 1-cor* (*k-means SDM*) und der andere ermittelt die Koordinatenpunkte direkt aus den Korrelationen zwischen den Items (*k-means cor*). Diese Methoden werden mit zwei bestehenden CA Methoden zur Clusterung von Variablen verglichen *ClustOfVar I* und *ClustOfVar II*, sowie mit zwei traditionellen hierarchischen Methoden und der EFA.

3.2.1 Dimensionsbestimmung

Für die Bestimmung der Anzahl der Faktoren bei der EFA waren die zu analysierenden Methoden die PA-PCA, die PA-PAF und der MAP Test. Für alle CA Methoden wurde die Clusteranzahl mit *Silhouette* bestimmt.

3.2.2 Variablenzuordnung

Die zwei neuen k-means Ansätze, die oben vorgestellt wurden, *k-means SDM* und *k-means cor*, wurden mit beiden *ClustOfVar* Methoden verglichen. Das erste basiert auf einer hierarchischen CA und das zweite auf einer k-means CA. Der k-means Ansatz verwendet als Ausgangsclusterung die Clusterlösung der hierarchischen *ClustOfVar* Methode. Außerdem wurden sie mit zwei hierarchischen Clustermethoden verglichen, einem mit *complete linkage* (*CACL*) und einem mit *average linkage* (*CAAL*), beide unter Verwendung des Distanzmaßes:

$$d(A, B) = 1 - cor(A, B) \tag{3.6}$$

Zusätzlich wurde eine FA durchgeführt, mit Pearson Korrelationen, Promax Rotation und ML-Schätzung. Van der Linden et al. (2012) zeigten in einer Studie zu Faktorkorrelationen verschiedener Persönlichkeitsinventare, dass Faktoren üblicherweise korreliert sind. Ihre Interkorrelationen reichen von 0.52 bis 0.67 (im Betrag) und die mittlere Korrelation ist 0.60. Aus diesem Grund wurde die Promax Rotation für diese Studie verwendet. Die Items wurden jeweils zu dem Faktor zugeordnet, auf dem sie die größten Ladung hatten.

3.2.3 Real World Simulation

Um die Effektivität dieser Methoden zu vergleichen wurde zuerst eine *Resampling* Methode angewandt, die *Real World Simulation* genannt wird. Das Ergebnis jeder Methode für einen echte, großen Datensatz wird als Populationsmodell angesehen. Stichproben der Größen 100, 200, 500 und 1000 werden dann mit Zurücklegen aus diesem Datensatz gezogen. Die Ergebnisse von jeder Methode in der Unterstichprobe werden mit dem Populationsmodell verglichen. Jede Stichprobengröße wird 1000 mal repliziert. Wir hatten also 4×4 verschiedene Bedingungen um die Anzahl der Faktoren zu bestimmen, da 4 verschiedene Methoden (*Silhouette, PA-PAF, PA-PCA* und der *MAP Test*) verglichen werden und 7×4 für die *Variablenzuordnung*. Die 7 *Variablenzuordnungs*-Methoden, die wir verglichen, waren: *k-means SDM, k-means cor, ClustOfVar I, ClustOfVar II, CACL, CAAL* und EFA.

1. NEO-PI-R Daten

Das NEO-PI-R ist ein weit verbreitetes Persönlichkeitsinventar, das Persönlichkeit in fünf Hauptskalen misst (Ostendorf & Angleitner, 2004): Neurotizismus, Extraversion, Offenheit für Erfahrungen, Verträglichkeit und Gewissenhaftigkeit. Jede Hauptskala ist in sechs Facetten, also Untertest, unterteilt und acht Items messen jede Facette. Folglich besteht der Fragebogen aus 240 Items. Für diese Studie wurde die Selbstbeurteilungsversion verwendet, bei der die Teilnehmer Selbsteinschätzungen über typisches Verhalten und Reaktionen auf einer fünfstufigen Likert-Skala, die von 0=*starke Ablehnung* bis 4=*starke Zustimmung* geht, abgeben. Untersuchungen von Validität und Reliabilität wurden von Ostendorf und Angleitner (2004) für alle Hauptskalen durchgeführt. Für die vorliegende Studie wurden pro Faktor aus einer untergeordneten Facette alle Items genommen. Wir ordneten die entstandenen 40 Items den darüber stehenden fünf Subfacetten zu. Die folgenden fünf Facetten aus jedem der fünf Faktoren wurden ausgewählt: N1 (Ängstlichkeit), E2 (Geselligkeit), O3 (Offenheit für Gefühle), A4 (Entgegenkommen) und C5 (Selbstdisziplin). Wie vom Handbuch des NEO-PI-R angezeigt, weist die Korrelationsmatrix der Items mittlere Interkorrelationen der Items innerhalb der Facetten auf (von 0.18 bis 0.36) und niedrige Interkorrelationen zwischen den Faktoren (für detailliertere Angaben siehe Ostendorf & Angleitner, 2004). Mittelwerte, Standardabweichungen und Schiefen der Items sind in Tabelle 3.1 dargestellt.

Wenn das theoretisch angenommene Faktormodell spezifizert wird, ist die mittlere Faktorkorrelation 0.00 und der Betrag der mittleren Faktorkorrelation ist 0.13, mit einer Spannweite von 0.01 bis 0.33 (siehe Tabelle 3.2). Die Hauptladungen reichen von 0.28 bis 0.79 (siehe Tabelle 3.3). Diese Werte erscheinen zwar sehr klein, sind aber typisch für Persönlichkeitsfragebögen. Peterson (2000) zeigte in einer Metaanalyse zu Faktorladungen in EFA's von Fragebogendaten, dass die mittlere Faktorladung 0.32 ist, wobei 25% der Faktorladungen kleiner sind als 0.23 und 25% größer als 0.37. Nebenladungen in diesem Datensatz weisen einen Mittelwert von 0.00, und eine Spannweite von -0.21 bis 0.16 auf. Zusammenfassend weist der NEO-PI-R Datensatz unter Verwendung dieses Modells relativ geringe Nebenladungen und geringe Faktorkorrelationen auf.

Tabelle 3.1: *Mittelwerte, Standardabweichungen und Schiefen der Items des NEO-PI-R Datensatzes*

	N1		E2		O3		A4		C5	
M	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
	1.85	2.26	1.62	3.02	2.58	3.07	1.15	2.97	2.05	2.87
SD	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
	0.95	1.14	0.88	1.19	0.75	1.03	0.90	1.17	0.81	1.15
Schiefe	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
	-0.30	0.19	-1.01	0.33	-0.99	-0.52	-0.93	1.06	-0.98	-0.09

Anmerkungen. M=Mittelwert; SD=Standardabweichung; N1=Ängstlichkeit; E2=Geselligkeit; O3=Offenheit für Gefühle; A4=Entgegenkommen; C5=Selbstdisziplin. Minima (Min) und Maxima (Max) der acht Items jeder Facette sind angegeben. Fälle mit fehlenden Werten wurden entfernt.

Der NEO-PI-R Normdatensatz besteht aus 11,724 Teilnehmern. Das mittlere Alter der Stichprobe ist 29.92 mit einer Spannweite von 16 bis 91, wobei 36% männlich und 64% weiblich sind.

2. IST-2000-R Daten

Das Grundmodul des IST-2000-R misst Intelligenz in drei Hauptdomänen: verbale Intelligenz, numerische Intelligenz und figurale Intelligenz, wobei jede davon in drei Untertests eingeteilt ist (Amthauer et al., 2007). Der Test beinhaltet insgesamt 180 Fragen, die nur richtig oder falsch beantwortet werden können.

Die Untertests waren die Grundlage unserer Berechnungen. Wir behandelten sie wie die Variablen des Datensatzes und ordneten sie darüberliegenden Faktoren zu. Dieses Vorgehen wurde gewählt, um das Problem der binären Daten zu vermeiden, das in späteren Studien adressiert werden soll. Es wurde der gleich Datensatz mit dem gleichen Modell verwendet wie in Kapitel 2. Mittelwerte, Standardabweichungen und Schiefen der Untertests können aus Tabelle 2.3 entnommen werden. Die Hauptladungen gehen von 0.47 bis 0.83 (siehe

Tabelle 3.2: *Faktorkorrelationsmatrix des Populationsdatensatzes NEO-PI-R - Zuordnung von Items zu Facetten*

	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
Factor 1	1.0	−0.33	−0.14	0.19	−0.06
Faktor 2		1.00	0.01	−0.08	0.07
Faktor 3			1.00	0.28	0.01
Faktor 4				1.00	−0.03
Faktor 5					1.00

Anmerkungen. Promax Rotation; ML-Schätzung.

Tabelle 3.3: *Ladungsmatrix des Populationsdatensatzes NEO-PI-R*

	N1	E2	O3	A4	C5
V1	0.62				
V31	0.59				
V61	0.75				
V91	0.52				
V121	0.53				
V151	0.67				
V181	0.68				
V211	0.47		0.13	−0.11	
V7		0.61			
V37		0.79			
V67		0.50		0.14	
V97		0.51	0.11		
V127		0.43		0.13	
V157		0.43	−0.21		
V187		0.59			
V217		0.75		−0.15	
V13			0.56	−0.10	
V43			0.62		
V73			0.56		
V103			0.39		
V133			0.71		
V163			0.58		
V193			0.42		
V223			0.46		
V19			0.14	0.28	
V49	0.11		−0.13	0.41	0.14
V79		−0.18		0.42	
V109				0.46	
V139				0.35	
V169		−0.13		0.54	
V199				0.54	
V229	−0.17			0.38	
V25					0.64
V55					0.72
V85	0.10				0.63
V115					0.69
V145					0.57
V175	−0.17		0.11		0.39
V205					0.58
V235					0.55

Anmerkungen. N1=Ängstlichkeit; E2=Geselligkeit; O3=Offenheit für Gefühle; A4=Entgegenkommen; C5=Selbstdisziplin. Promax Rotation; ML-Schätzung. Ladungen unter 0.10 wurden unterdrückt.

Tabelle 2.4). Die Validität und Reliabilität aller Untertests wurden von Amthauer et al. (2007) gezeigt. Die Nebenladungen reichen von -0.11 bis 0.18 (siehe Tabelle 2.4) mit einem Mittelwert von 0.01. Die drei Faktorkorrelationen haben die Werte 0.66, 0.49 und 0.43. Man kann also zusammenfassen, dass die Nebenladungen in ihrer Höhe vergleichbar sind mit denen aus dem NEO-PI-R Datensatz und die Faktorkorrelationen viel höher sind. Der Normdatensatz besteht aus 1,352 Beobachtungen. Das mittlere Alter ist 19.09 und reicht von 16 bis 25 mit 44% Frauen und 56% Männern.

Für die *Dimensionsbestimmung* wurde für jede Stichprobengröße die Erfolgsrate berichtet, das heißt der Prozentsatz identischer Faktorenanzahlen zur Gesamtstichprobe. Und für die *Variablenzuordnung* setzten wir die Anzahl der Faktoren auf die theoretisch angenommene: Fünf Faktoren beim NEO-PI-R und drei Faktoren beim IST-2000-R. Die Ähnlichkeit wurde dann über den Randindex (Rand, 1971) ermittelt. Dieser wird durch Ermittlung der richtig klassifizierten Elementenpaare berechnet. Der Randindex ist definiert über:

$$R(C, C') = \frac{2(n_{11} + n_{00})}{n(n-1)} \tag{3.7}$$

wobei C die jeweilige Clusterlösung in der Stichprobe ist, C' ist die Clusterlösung im Populationsdatensatz, n_{11} ist die Anzahl der Paare, die unter C und C' im selben Cluster sind und n_{00} ist die Anzahl der Paare, die in C und C' in verschiedenen Clustern sind.

3.2.4 Traditionelle Simulation

Wir spezifizierten das Faktormodell, das die EFA in dem Populationsdatensatz gefunden hatte. Um die Simulation vergleichbar mit dem *Real World* Modell zu machen, wurden die Schätzer der Haupt- und Nebenladungen, Faktorkorrelationen und Residualvarianzen aus der EFA des Normdatensatzes genommen und anschließend für die Simulation genutzt (*Populationsmodell*). In der ersten Simulationsbedingung wurden alle Residualkorrelationen auf null gesetzt, was heißt, dass ein perfekt passendes Modell simuliert wurde. Und in der zweiten Simulationsbedingung wurden auch die Residualkorrelationen aus dem Populationsdatensatz genommen und in die Simulation eingefügt.

Zusätzlich wurden verschiedene Stichprobengrößen verwendet (100, 500, 1000) mit jeweils 1000 Replikationen. Die Erfolgsraten in den Unterstichproben wurden für die *Dimensionsbestimmung* berichtet und die Randindizes für die *Variablenzuordnung*. Wir hatten also $5 \times 3 \times 2$ verschiedene Bedingungen um die Faktorenanzahl zu bestimmen, da fünf verschiedene Methoden mit drei Stichprobengrößen verglichen wurden und $7 \times 3 \times 2$ für die *Variablenzuordnung*.

Es werden multivariat normalverteilte Daten aus einer spezifizierten Verteilung mit Populationskovarianzmatrix, die aus Ladungen, Faktorkorrelationen und Residualtermen berechnet wurde, gezogen.

3.2.5 CFA Kreuzvalidierung

Wir spezifizierten die Faktormodelle, die sich aus den Kombinationen an Methoden zur *Dimensionsbestimmung* und *Variablenzuordnung* ergeben auf Unterstichproben des Datensatzes mit $n = 100, 500$ und 1000 mit jeweils 1000 Replikationen. Es werden keine Ladungshöhen spezifiziert, weshalb auch für die CA Verfahren ein Faktormodell spezifiziert werden kann. Für diese Modelle wird dann eine CFA im Gesamtdatensatz gerechnet und der BIC (*Bayesianisches Informationskriterium*) angegeben.
Alle Berechnungen wurden in der open source software R 0.94.110 programmiert, unter Verwendung des Paketes "psych"(Revelle, 2011).

3.3 Ergebnisse

3.3.1 Real World Simulation

Dimensionalitätsbestimmung

Die Tabelle 3.4 fasst die Ergebnisse der *Real World Simulation Dimensionsbestimmung* zusammen, zuerst für den NEO-PI-R Datensatz und dann für den IST-2000-R Datensatz. Die Spalte „*# pop data*“ zeigt, wie viele Faktoren oder Cluster mit der jeweiligen Methode im Populationsdatensatz gefunden wurden. Die Spalte „SD“ gibt die Standardabweichungen gemittelt über alle Stichproben an. Die letzten Spalten zeigen für die verschiedenen Stichprobengrößen die Erfolgsraten, also jeweils den Prozentsatz an Stichproben, in denen die gleiche Anzahlen an Dimensionen angezeigt wurde wie im Gesamtdatensatz. Die Verfahren mit den höchsten Erfolgsraten sind in beiden Datensätzen die EFA-Methode PA-PCA, der MAP Test und *k-means cor* mit *Silhouette*. Diese drei wiesen auch vergleichsweise hohe Effizienz auf, also geringe SD's. Die hohe Erfolgsrate des MAP Tests, sowie seine hohe Effizienz, gehen im IST-2000-R Datensatz allerdings auf Kosten einer sehr unpraktischen Faktorlösung von einem gemeinsamen Faktor. Die hierarchischen CA Methoden (CAAL und CACL) hatten die geringsten Erfolgsraten über alle Datensätze hinweg. Grundsätzlich zeigen im IST-2000-R Datensatz nur die PA's eine Sensitivität für die Stichprobengrößen, wie schon in der vorherigen Studie (siehe Tabelle 2.7) jedoch nicht die CA Methoden.

Variablenzuordnung

Als nächstes zeigt Tabelle 3.5 den mittleren Anteil an Variablenzuordnungen, die identisch sind mit der Variablenzuordnung der jeweiligen Methode im Gesamtdatensatz (Randindex) für verschiedene Stichprobengrößen für beide Datensätze getrennt. Die neue *k-means SDM* weist in beiden Datensätzen den höchsten Randindex von allen Methoden auf. Besonders für kleine Stichprobengrößen von 100 oder 200 übertraf es die EFA. Auch *ClustOfVar II* erreichte in beiden Datensätzen mindestens so gute Randindizes wie die EFA. Die geringsten Anteile an identischen Lösungen wurde von den traditionellen hierarchischen Methoden erreicht.

Tabelle 3.4: *Erfolgsraten für die Dimensionsbestimmung in der Real World Simulation für verschiedene Stichprobengrößen für beide Datensätze*

				n			
Methode		# Pop data	SD	100	200	500	1000
			NEO-PI-R				
K-means SDM	Silhouette	6	0.88	0.28	0.35	0.41	0.60
K-means cor	Silhouette	5	0.45	0.58	0.83	0.98	1.00
ClustOfVar I	Silhouette	5	0.39	0.65	0.84	0.96	0.99
ClustOfVar II	Silhouette	5	0.44	0.69	0.89	0.96	1.00
CAAL	Silhouette	6	0.96	0.23	0.31	0.49	0.60
CACL	Silhouette	5	0.93	0.14	0.37	0.66	0.88
EFA	PA-PAF	8	0.56	0.02	0.00	0.00	0.01
	PA-PCA	6	0.46	0.26	0.34	0.53	0.89
	MAP	5	0.43	0.56	0.73	0.90	0.96
			IST-2000-R				
K-means SDM	Silhouette	2	0.67	0.65	0.80	0.88	0.98
K-means cor	Silhouette	2	0.51	0.71	0.80	0.93	0.96
ClustOfVar I	Silhouette	2	0.63	0.64	0.76	0.90	0.96
ClustOfVar II	Silhouette	2	0.61	0.67	0.75	0.87	0.97
CAAL	Silhouette	2	0.66	0.68	0.72	0.82	0.89
CACL	Silhouette	2	0.67	0.58	0.68	0.86	0.93
EFA	PA-PAF	4	0.55	0.18	0.27	0.39	0.64
	PA-PCA	2	0.34	0.58	0.80	0.95	0.99
	MAP	1	0.02	0.99	1.00	1.00	1.00

Anmerkungen. K-means SDM=k-means *skaliertes Distanzmaß*; *k-means cor*=k-means Korrelation; *ClustOfVar I*=*ClustOfVar* mit hierarchischer Clusteranalyse; *ClustOfVar II*=*ClustOfVar* mit k-means Clusteranalyse; CAAL=hierarchische Clusteranalyse mit average linkage; CACL=hierarchische Clusteranalyse mit complete linkage; PA-PAF=Parallelanalyse mit Hauptachsenanalyse; PA-PCA=Parallelanalyse mit Hauptkomponentenanalyse; MAP=MAP Test; n=Stichprobengröße; SD=mittlere Standardabweichung über alle Stichprobengrößen; # Pop data=Anzahl der Faktoren im Populationsdatensatz. NEO-PI-R: Items wurden zu Facetten zugeordnet. Alle Erfolgsraten basieren auf 1000 Replikationen.

Tabelle 3.5: *Randindizes in der Real World Simulation für verschiedene Variablenzuordnungsmethoden für verschiedene Stichprobengrößen für beide Datensätze*

	K-means SDM	K-means cor	ClustOfVar I	ClustOfVar II	CAAL	CACL	EFA
			NEO-PI-R				
100	0.96	0.88	0.94	0.95	0.86	0.88	0.94
200	0.99	0.89	0.97	0.97	0.90	0.93	0.98
500	1.00	0.89	0.99	0.98	0.93	0.97	1.00
1000	1.00	0.89	0.99	0.99	0.95	0.98	1.00
			IST-2000-R				
100	0.84	0.78	0.83	0.83	0.72	0.77	0.78
200	0.87	0.81	0.86	0.86	0.73	0.77	0.83
500	0.91	0.83	0.90	0.90	0.72	0.74	0.88
1000	0.91	0.85	0.91	0.91	0.72	0.73	0.89

Anmerkungen. K-means SDM=k-means skaliertes Distanzmaß; *k-means cor*=k-means Korrelation; *ClustOfVar I=ClustOfVar* mit hierarchischer Clusteranalyse; *ClustOfVar II=ClustOfVar* mit k-means Clusteranalyse; CAAL=hierarchische Clusteranalyse mit average linkage; CACL=hierarchische Clusteranalyse mit complete linkage; EFA=exploratorische Faktorenanayse. NEO-PI-R: Items wurden zu Facetten zugeordnet. Alle Randindizes basieren auf 1000 Replikationen.

3.3.2 Traditionelle Simulation

Dimensionsbestimmung

Die Tabelle 3.6 gibt einen Überblick über die Ergebnisse der *traditionellen Simulation* für *Dimensionsbestimmung* in den zwei verschiedenen Simulationsbedingungen. Die Erfolgsraten und mittlere Anzahlen an angezeigten Dimensionen für die Ergebnisse aus den simulierten Stichproben im Vergleich zu dem simulierten Modell werden berichtet. Bei Simulation des *Populationsmodells* ohne Residualkorrelationen, zeigten die EFA Methoden die besten Erfolgsraten von allen Methoden. Wenn die Residualkorrelationen hinzugenommen werden jedoch sinkt deren Leistung erheblich ab und andere Methoden, wie *k-means cor*, CACL und *ClustOfVar* sind besser. Im IST-2000-R Datensatz konnte fast keine der Methoden die simulierten drei Faktoren wiederfinden. Am häufigsten wurden hier zwei Faktoren vorgeschlagen, was für die CA Methoden die geringst-mögliche Anzahl war. Diese Ergebnisse sind vermutlich auf die hohen Faktorkorrelationen in diesem Datensatz zurückzuführen, die mitsimuliert wurden. Die PA-PAF war die einzige Methode, die die drei Faktoren noch einigermaßen oft finden konnte.

Variablenzuordnung

Alle Methoden hatten in beiden Simulationsbedingungen etwas höhere Randindizes als in der *Real World Simulation* (siehe Tabelle 3.7). Die höchsten Randindizes wurden jedoch

Tabelle 3.6: *Erfolgsraten und Mittelwerte der angezeigten Anzahlen an Dimensionen über alle Stichprobengrößen hinweg für die traditionelle Simulation für beide Datensätze*

Methode	Dimensionsbestimmung	Populationsmodell		Populationsmodell + res cor	
		Erfolgsrate	M	Erfolgsrate	M
		NEO-PI-R			
K-means SDM	Silhouette	0.48	7.65	0.44	7.77
K-means cor	Silhouette	0.89	3.83	0.86	3.86
ClustOfVar I	Silhouette	0.90	4.69	0.89	4.83
ClustOfVar II	Silhouette	0.92	4.57	0.91	4.75
CAAL	Silhouette	0.40	7.59	0.37	7.64
CACL	Silhouette	0.93	8.40	0.98	8.52
EFA	PA-PAF	0.99	5.36	0.33	5.94
	PA-PCA	0.98	5.05	0.56	5.39
	MAP	0.97	3.10	0.85	3.08
		IST-2000-R			
K-means SDM	Silhouette	0.05	2.66	0.44	2.68
K-means cor	Silhouette	0.05	2.31	0.05	2.29
ClustOfVar I	Silhouette	0.07	2.69	0.04	2.68
ClustOfVar II	Silhouette	0.07	2.70	0.06	2.69
CAAL	Silhouette	0.04	2.54	0.06	2.59
CACL	Silhouette	0.07	2.90	0.09	2.95
EFA	PA-PAF	0.92	3.17	0.57	3.30
	PA-PCA	0.04	1.65	0.04	1.66
	MAP	0.00	1.00	0.00	1.00

Anmerkungen. K-means SDM=k-means *skaliertes Distanzmaß*; *k-means cor*=k-means Korrelation; *ClustOfVar I*=*ClustOfVar* mit hierarchischer Clusteranalyse; *ClustOfVar II*=*ClustOfVar* mit k-means Clusteranalyse; CAAL=hierarchische Clusteranalyse mit average linkage; CACL=hierarchische Clusteranalyse mit complete linkage; PA-PAF=Parallelanalyse mit Hauptachsenanalyse; PA-PCA=Parallelanalyse mit Hauptkomponentenanalyse; MAP=MAP Test; M=Mittelwert; res cor=Residualkorrelationen. Simulierte Anzahl an Faktoren: NEO-PI-R:5, IST-2000-R:3. NEO-PI-R: Items wurden zu Facetten zugeordnet. Alle Erfolgsraten basieren auf 1000 Replikationen. Stichprobengrößen=100, 500, 1000.

von *k-means SDM* erzielt, gefolgt von der EFA und *ClustOfVar*. Die EFA war hierbei nur im IST-2000-R Datensatz schlechter als *k-means SDM*. Im NEO-PI-R Datensatz erreichten sowohl *k-means SDM* als auch die EFA eine Rate von fast 100 % korrekten Spezifikationen. Die traditionellen Methoden wiesen auch hier wieder die geringsten Randindizes auf.

Tabelle 3.7: *Randindizes über alle Stichprobengrößen hinweg für die traditionelle Simulation für beide Datensätze*

	K-means SDM	*K-means* cor	*ClustOfVar I*	*ClustOfVar II*	CAAL	CACL	EFA
			NEO-PI-R				
Populations-modell	0.99	0.97	0.98	0.99	0.93	0.96	0.99
Populations-modell + res cor	0.99	0.96	0.97	0.98	0.92	0.95	0.99
			IST-2000-R				
Populations-modell	0.96	0.91	0.95	0.95	0.70	0.77	0.95
Populations-modell + res cor	0.96	0.90	0.94	0.94	0.70	0.74	0.93

Anmerkungen. K-means SDM=k-means *skaliertes Distanzmaß*; *k-means cor*=k-means Korrelation; *ClustOfVar I*=*ClustOfVar* mit hierarchischer Clusteranalyse; *ClustOfVar II*=*ClustOfVar* mit k-means Clusteranalyse; CAAL=hierarchische Clusteranalyse mit average linkage; CACL=hierarchische Clusteranalyse mit complete linkage; EFA=exploratorische Faktorenanalyse. res cor=Residualkorrelationen. NEO-PI-R: Items wurden zu Facetten zugeordnet. Alle Randindizes basieren auf 1000 Replikationen. Stichprobengrößen=100, 500, 1000.

Wenn man alle Ergebnisse aus der *Real World Simulation* und der *traditionellen Simulation* vergleicht, lieferten die PA-PCA und *k-means cor* mit *Silhouette* die besten Ergebnisse bei der *Dimensionsbestimmung* über alle Bedingungen hinweg. Für die *Variablenzuordnung* hat allerdings der neue k-means Ansatz *k-means SDM* die besten Ergebnisse erzielt, besonders für kleine Stichprobengrößen. Ausgehend von diesen Ergebnissen könnte man erwarten, dass *k-means SDM* und die PA-PCA die beste Kombination für die explorative Strukturfindung darstellt. Die vorgeschlagenen Modelle für alle Kombinationen von *Dimensionsbestimmung* und *Variablenzuordnung* wurden in einer *CFA Kreuzvalidierungsstudie* getestet.

3.3.3 CFA Kreuzvalidierung

Die BIC's der *CFA Kreuzvalidierung* im NEO-PI-R lagen alle über 930,000 und im IST-2000-R über 62,000. Daher werden in Tabelle 3.8 alle BIC Werte minus 930,000 bzw minus 62,000. Gemäß unseren Erwartungen erreicht die Kombination aus *k-means SDM* und

Tabelle 3.8: *BIC für die CFA Kreuzvalidierung mit verschiedenen Kombinationen von Methoden der Dimensionsbestimmung und Variablenzuordnung für beide Datensätze*

	K-means SDM	*K-means* cor	*ClustOfVar I*	*ClustOfVar II*	CAAL	CACL	EFA
			NEO-PI-R				
Silhouette	5477	12580	7382	6614	6053	7060	-
PA-PAF	5406	6659	5870	5372	11384	10193	6200
PA-PCA	5828	7795	6494	5913	13708	11579	6636
MAP	6136	8370	6882	6287	14568	12078	6936
			IST-2000-R				
Silhouette	1857	1832	1854	1854	1911	1884	-
PA-PAF	1936	1941	1934	1934	2050	2077	1884
PA-PCA	1810	1816	1810	1811	1873	1838	1830
MAP	1964	1964	1964	1964	1964	1964	1964

Anmerkungen. K-means SDM=k-means *skaliertes Distanzmaß*; *k-means cor*=k-means Korrelation; *ClustOfVar I*=*ClustOfVar* mit hierarchischer Clusteranalyse; *ClustOfVar II*=*ClustOfVar* mit k-means Clusteranalyse; CAAL=hierarchische Clusteranalyse mit average linkage; CACL=hierarchische Clusteranalyse mit complete linkage; PA-PAF=Parallelanalyse mit Hauptachsenanalyse; PA-PCA=Parallelanalyse mit Hauptkomponentenanalyse; MAP=MAP Test. NEO-PI-R: Items wurden zu Facetten zugeordnet. Anzahl der Replikationen=1000. Es sind die BIC-Werte minus 930,000 für NEO-PI-R und minus 62,000 für IST-2000-R angegeben. Stichprobengrößen=100, 500, 1000.

PA-PCA den geringsten BIC im IST-2000-R, gefolgt von den nicht-hierarchischen CA Methoden zusammen mit der PA-PCA. IM NEO-PI-R jedoch zeigte die PA-PAF zusammen mit *ClustOfVar II* und *k-means SDM* die besten Ergebnisse. Im NEO-PI-R hatte auch *Silhouette* mit *k-means SDM* einen sehr niedrigen BIC. Insgesamt schlugen die PA's kombiniert mit entweder *ClustOfVar II* oder *k-means SDM* die am besten passenden Modell gemäß der CFA vor. Im Falle hoher Faktorkorrelationen, wie im IST-2000-R Datensatz, stellte sich die PA-PCA als nützlicher heraus, wohingegen im NEO-PI-R Datensatz die PA-PAF die besseren Ergebnisse lieferte. Der MAP Test lieferte immer den gleichen BIC, was daran liegt, dass er immer einen Faktor anzeigte. Eine weitere Möglichkeit für ein gutes Gesamtergebnis wäre ausgehend von den vorherigen Ergebnissen für eine Kombination von *k-means cor* mit *Silhouette* zur *Dimensionsbestimmung* und *k-means SDM* für die *Variablenzuordnung* zu erwarten.

3.4 Diskussion

Die oben genannten Ergebnisse der *Real World Simulation*, *traditionellen Simulation* und der *CFA Kreuzvalidierung* zeigen, dass die zwei neuen k-means Ansätze *k-means SDM* und *k-means cor* hinsichtlich der folgenden Aspekte der EFA zu bevorzugen sind:

1. *K-means SDM* zeigt genauere Ergebnisse für die *Variablenzuordnung* von Items zu Faktoren, vor allem bei kleinen Stichproben.

2. Die beiden Verfahren beinhalten keine Modellannahmen über die Entstehung der Varianzen und Kovarianzen der beobachteten Variablen.

3. Es ist auch bei der k-means Clusterung möglich, Clusterwerte zu berechnen, wenn man den Clustermittelpunkt als übergeordnetes Konstrukt ansieht und diese Clusterwerte sind eindeutig.

Tatsächlich könnte die k-means CA eines der Probleme der hierarchischen CA überwinden und dabei den Vorteil der CA gegenüber der EFA beibehalten, so dass keine kritischen Annahmen über unkorrelierte Residuen gemacht werden müssen. Das Problem der hierarchischen CA ist, dass sie nur auf Distanzmaßen zwischen Items basiert und keine Aussagen über das zu messende Konstrukt trifft. In der k-means Clusterung kann der Zentroid eines Clusters als die darüberliegende Konstrukt angesehen werden. Es sind nicht nur seine Koordinaten bekannt, sondern es können auch die Distanzen zwischen dem Konstrukt und den Items berechnet werden. Da diese Distanzen direkt aus den Korrelationen ermittelt werden können, können sie auch direkt in Korrelationen zurücktransformiert werden. Folglich können Korrelationsmatrizen und auch partielle Korrelationsmatrizen der Items mit den Konstrukten, äquivalent zu den Ladungsmatrizen in der EFA, ermittelt werden. Aus diesen Distanzen oder Ladungen können wiederum die Werte der Personen auf den Konstrukten berechnet werden, die *Clusterwerte*. Genau wie in der PCA und im Gegensatz zur FA sind diese Clusterwerte eindeutig, da keine Residualvarianz in die Gleichung einfließt. Die Gleichung der Linearkombinationen der Clusterwerte wird genauso gebildet, wie die für die Komponentenwerte der PCA und wurde in Kapitel 1.1.3 angegeben (siehe Gleichung (1.9)). Diese Information ist besonders für Anwender von psychologischen Tests von Bedeutung. Auf die mathematischen Hintergründe soll in dieser Studie jedoch nicht ausführlicher eingegangen werden.

Was die *Dimensionsbestimmung* betrifft, zeigte die EFA immer noch vergleichbar gute Ergebnisse wie die Kombination *k-means cor* und *Silhouette*. Dieser Effekt zeigte sich in beiden Simulationen, der *Real World Simulation* und der *traditionellen Simulation*. Die PA-PAF stellte sich als stichprobengrößensensitiv heraus und zeigte in der *Real World Simulation* im Gesamtdatensatz systematisch mehr Faktoren an, als in den Unterstichproben. Wenn jedoch die PA-PAF für die *Dimensionsbestimmung* verwendet wurde und anschließend die Items mittels des neuen k-means Ansatzes *k-means SDM* zugeordnet wurden, wurde in der Kreuzvalidierung für den NEO-PI-R Datensatz das am besten passende CFA Modell gefunden. Im IST-2000-R Datensatz war die PA-PCA besser als die PA-PAF. Dieses Ergebnis legt nahe, dass es eine sinnvolle Möglichkeit wäre, eine Kombination aus der PA für die *Dimensionsbestimmung* und *k-means SDM* für die *Variablenzuordnung* zu verwenden, um so in der Praxis explorativ Strukturfindung zu betreiben. Genauer gesagt, ist die PA-PCA zu empfehlen, wenn die Dimension des Datensatzes auf möglichst wenige Komponenten reduziert werden soll, die möglichst viel Varianz aufklären. Diese Kombination ist außerdem von Vorteil wenn von der Abwesenheit eines Messmodells profitiert

werden soll, das unkorrelierte Residualvarianzen annimmt. Eine weiter Möglichkeit, deren Kombination noch getestet werden muss, wäre *k-means cor* mit *Silhouette* für die *Dimensionsbestimmung* zu verwenden und anschließend *k-means SDM* für die *Variablenzuordnung*.

Ob das neue k-means Verfahren mit den in der Einleitung angesprochenen praktischen Probleme der EFA, z.B. dass sie Schwierigkeiten mit hohen Nebenladungen hat, umgehen kann, muss in der vorliegenden Studie nicht extra getestet werden. Durch die Verwendung der Parameter aus echten Daten wurde sicher gestellt, dass praktisch bedeutsame Bedingungen getestet wurden.

Aufbauend auf diesen Ergebnissen empfehlen wir, *k-means SDM* für die *Variablenzuordnung* zu verwenden und PA-PCA für die *Dimensionsbestimmung*, besonders wenn kleine Stichproben vorliegen, oder wenn man sich nicht sicher ist, ob ein zugrundeliegendes Messmodell mit unkorrelierten Residuen gerechtfertigt ist. Es sollte hier jedoch betont werden, dass der CFA-Fit nur ein mögliches Kriterium ist, das wieder auf dem Faktormodell basiert. Wie bereits erwähnt, ist das Faktormodell für die CA nicht notwendig und daher sollte sich die Evaluation von k-means Ansätzen zur Clusterung von Items in der Zukunft mehr auf die prädiktive Qualität der Clusterwerte, die aus den Analysen gewonnen werden, konzentrieren.

Kapitel 4

Diskussion

Das Ziel dieser Arbeit lag in der Entwicklung einer validen Methode zur Evaluation strukturfindender Methoden sowie der Entwicklung eines *Clusteranalyse* (CA) Verfahrens zur Clusterung von Items, das mit eben dieser Methode evaluiert wurde. In der ersten Studie wurde die neue Methode *Real World Simulation* vorgestellt und exemplarisch an zwei Datensätzen zur Evaluation der EFA angewandt. Bei dieser Methode wird die zu untersuchende Methode an einem großen Datensatz angewandt und anschließend werden Unterstichproben verschiedener Größe aus diesem Datensatz gezogen und getestet, wie zuverlässig die Methode in den Unterstichproben zu dem gleichen Ergebnis kommt, wie im Gesamtdatensatz. Mit Hilfe dieser neuen Methode konnte gezeigt werden, dass es sinnvoll ist, echte Daten zu verwenden, um die Validität von Simulationsstudien zu verbessern. Es wurde deutlich, dass künstlich erstellte Simulationsstudien die Wirklichkeit nicht ausreichend widerspiegeln können und ihre Einschätzung der Funktion einer Methode daher teilweise sehr verfälscht sein kann. Die Ergebnisse zeigten, dass schon geringe Modellverletzungen ausreichen, um die Ergebnisse der EFA-Methoden erheblich zu verändern. In der zweiten Studie wurde dieses neu entwickelte Verfahren verwendet, um die Effektivität zwei neuer k-means Verfahren zur Clusterung von Items zu testen. Indem dieses Verfahren mittels der neuen Real World Simulation untersucht wurde, wurde nicht nur untersucht, wie das Verfahren in künstlichen Daten funktioniert, sondern auch, wie es sich unter reellen Datenbedingungen verhält. Die neuen k-means Verfahren wurden im Vergleich zu der klassischerweise angewandten EFA getestet und konnten ihr gegenüber einige Vorteile aufweisen. Für die *Variablenzuordnung* funktionierten sie besser als das klassische EFA-Verfahren, nicht jedoch zur *Dimensionsbestimmung*. Aus diesem Grund wurde vorgeschlagen, die PA-PCA mit einer anschließenden k-means Clusterung zu verbinden. Außerdem wurde aufgezeigt, dass diese k-means Clusterverfahren den großen Vorteil mit sich bringen, dass sie keine Modellannahmen machen und somit keinen Zufallsfehler beinhalten. Aufgrund des nicht vorhandenen Residual- oder Fehlerterms ist es mit der k-means Clusterung von Items auch möglich die *Clusterwerte* der Personen auf den Clustern eindeutig zu bestimmen, was bei der FA nicht möglich ist. Dieser Vorteil wurde zwar in der Vergangenheit schon allgemein für die Clusteranalyse angesprochen (z.B. Revelle, 2014), muss jedoch mathematisch noch für den speziellen Fall der zwei neuen Verfahren gezeigt werden.

Die angesprochenen Probleme der EFA sind in der Forschung bereits mehrfach thematisiert worden. Das Problem der Unbestimmtheit der Faktoren in der FA ist schon lange bekannt, es wurde erstmals von Wilson (1928) beschrieben. Auch das Problem der Annahme eines Faktormodells mit unkorrelierten Fehlern, wurde immer wieder angesprochen (z.B. Cudeck & Henly, 1991; Loevinger et al., 1953; MacCallum & Tucker, 1991; Tryon, 1935). Und dass die Option, die CA als Alternative zur EFA zu nutzen immer noch aktuell ist, kann man daran sehen, dass Revelle (2014) erst kürzlich die Vorzüge der beiden Verfahren gegenüberstellte. Sein Vorschlag ist es, eine Zuordnung der Items zu Gruppen über hierarchische Clusterung oder Faktorenanalyse vorzunehmen und basierend darauf, die Kennwerte des Tests zu berechnen (Revelle, 2011). Mit diesem Ansatz ist es jedoch nicht möglich, einen gewichteten Clusterwert zu berechnen. In dem neu entwickelten k-means Verfahren werden nun jedoch direkt die Items über k-means geclustert, man erhält also die Distanzen der Items zu dem Clustermittelpunkt und kann darüber die Gewichte für den Clusterwert ermitteln.

Aus dieser eher mathematischen bzw. theoretischen Perspektive bietet die k-means Clusterung der Items also einige bestechende Vorteile gegenüber der FA. Letztlich ist es jedoch eine grundsätzliche Frage, ob man es für sinnvoll erachtet für die Gruppierung von Items ein zugrundeliegendes Modell anzunehmen oder nicht.

Ob es sinnvoll ist, anzunehmen, dass es ein wahres Modell in der Population gibt, ist eher eine wissenschatstheoretische Frage. In der computationalen Statistik herrscht Uneinigkeit darüber ob es so etwas wie natürliche Klassifikationen in der Natur überhaupt gibt (z.B. Kleinberg, 2003; Buhmann, 2010; Gilmour & Walters, 1964; Jain, 2010; Lakoff, 1987). Wenn man davon ausgeht, dass es die zu findenden Kategorien in der Natur wirklich gibt, ist es auch sinnvoll eine Ursache dieser naturgegebenen Klassifizierung anzunehmen, was zu der Annahme eines Modells führt, wie z.B. dem Faktormodell [1]. Die andere Möglichkeit ist, sich lediglich zum Ziel zu setzen, sinnvolle Gruppierungen der Items im Datensatz zu finden. Diese Gruppierungen werden dann vielmehr anhand praktischer Kriterien wie z.B. Homogenität bzw. Varianzaufklärung vorgenommen. Unter dieser Annahme ist es sinnvoller ein Verfahren wie die PCA oder CA zu verwenden. In solchen Fällen muss der Forscher sich überlegen, was das Ziel der Clusterung sein soll, woran man also erkennen kann, ob die Methode gut funktioniert. Es wurde in der Literatur schon mehrfach betont, dass nicht beurteilt werden kann ob eine Clusterung gut oder schlecht ist ohne mit einzubeziehen, wofür man die ermittelte Clusterung anwenden möchte. (für eine Übersicht siehe v. Luxburg, Williamson & Guyon, 2012).

Diese Problematik entspricht im Grunde der alten Frage, ob man die *Konstrukt-* oder *Kriteriumsvalidität* eines Tests für wichtiger erachtet. Die Konstruktvalidität gibt an, wie gut ein Test das misst, was er zu messen beansprucht, wohingegen die Kriteriumsvalidität angibt, wie stark der Test mit einem oder mehreren Kriterien zusammenhängt (Buehner, 2011, S. 63) bzw. wie gut er diese vorhersagt. Steht die Konstruktvalidität im Vorder-

[1] Es sollte noch darauf hingewiesen werden, dass das Faktormodell nicht das einzige in der Psychometrie gebräuchliche Modell zur Modellierung von Fragebogendaten ist. Die Gruppe der Modelle aus der *Item Response Theorie* (IRT) z.B. stellt auch gebräuchliche Modelle dieser Art zur Verfügung, die in der vorliegenden Arbeit jedoch nicht thematisiert wurden (siehe z.B. Rost, 2004).

grund, so ist es wichtig, mittels FA zu testen, welche Struktur bzw. welches Messmodell dem Test zugrundeliegt. Wenn man jedoch primär möchte, dass ein Test ein bestimmtes Außenkriterium gut vorhersagen kann, so ist die Modellstruktur nicht besonders entscheidend. In solchen Fällen ist es eher entscheidend, dass die gebildeten Komponenten- oder Clusterwerte der Personen möglichst gut ein gewähltes Kriterium vorhersagen können. Es ist daher wichtig, eine möglichst sinnvolle Gruppierung der Items zu bekommen, was zum Beispiel anhand der Homogenität innerhalb einer Itemgruppe, bzw. der insgesamt aufgeklärte Varianz festgestellt wird. Dazu wird entweder eine PCA oder CA angewandt. Bei dieser Art von Tests spricht man auch von formativen Modellen, da die Konstrukte oder Cluster von den Items geformt werden und nicht umgekehrt die Faktoren die Varianz der Items erklären wie im Faktormodell (Buehner, 2011, S. 93).

Bei dem neu entwickelten k-means Verfahren zur Clusterung von Items handelt es sich im Grunde auch um ein formatives Verfahren, da die Clusterwerte aus den Items gebildet werden. Wie bereits in Kapitel 3 angemerkt wurde, basiert sowohl die in dieser Arbeit angewendete *traditionelle Simulation* als auch die *CFA Kreuzvalidierung* auf der Annahme, dass den Daten ein wahres Faktormodell als DGP zugrundeliegt, das von der CA gefunden werden muss. Man könnte jedoch berechtigterweise auch andere Kriterien aufstellen, die nicht auf dem Faktormodell basieren und der modellfreien Natur der k-means CA gerechter werden. Die folgenden zwei Kriterien könnten speziell für die Evaluation des neuen k-means Verfahrens zur Clusterung von Items von Bedeutung sein.

1. Die *Real World Simulation*
2. Die Validierung an einem Außenkriterium, also die Ermittlung der Vorhersagegüte der *Clusterwerte*

Der erste Punkt wurde in der zweiten Studie in Kapitel 3 bereits vorgestellt und auf die Clusterung von Items angewandt. Das verwendete Real World Vefahren ist insofern modellfrei, als vom Anwender kein Modell vorgegeben wird, sondern jedes Verfahren für den Gesamtdatensatz ein eigenes Modell vorschlägt. Dieses Modell muss dann auch nicht dem Faktormodell entsprechen, da keine Ladungen oder Residualvarianzen spezifiert werden, sondern beinhaltet lediglich die Zuordnung der Items zu Gruppen. Dieses Verfahren entspricht auch dem, was bei v. Luxburg et al. (2012) vorgeschlagen wird, wie CA Verfahren zu evaluieren sind. Hier wird ganz analog zu unserem Vorgehen in der *Real World Simulation* vorgeschlagen, an echten Daten zu ermitteln, wie stabil das Verfahren ist und ob es konvergiert. Wie stabil ein Verfahren ist, ist analog zur *Effizienz* zu verstehen, wie sie in Kapitel 1.2.2 als Gütekriterium eines Schätzers definiert wurde. In unserer *Real World Simulation* wurde das in der zweiten Studie (Kapitel 3) geprüft, indem die Standardabweichung (SD) über verschiedene Stichproben hinweg angegeben wurde. Ob ein Verfahren konvergiert, kann über die *Konsistenz* eines Schätzers, wie in Kapitel 1.2.2 definiert, ermittelt werden. Diese wurde in der Studie getestet durch die Ziehung verschieden großer Stichproben und durch die Prüfung, ob der Clusteralgorithmus gegen eine Version konvergiert. Die Bestimmung der Erwartungstreue bzw. des Bias tritt bei fehlender Modellannahme hingegen in den Hintergrund.

Der zweite Punkt, die Validierung an einem Außenkriterium, entspricht dem, was bei v. Luxburg et al. (2012) mit Abschätzen der Nützlichkeit eines Verfahrens gemeint ist. Das Ziel ist es, Clusterwerte zu bilden, die möglichst gut ein externes Kriterium vorhersagen können. Bei v. Luxburg et al. (2012) wird die Nützlichkeit der Clusterung als eines der wichtigsten Ziele beschrieben. Es soll daher ebenfalls ein Ziel der weiteren Forschung in diesem Bereich sein, an verschiedenen echten Datensätzen zu überprüfen, wie gut Clusterwerte, die über die neuen Verfahren gebildet wurden, externe Kriterien vorhersagen können. Sollte dadurch auch noch gezeigt werden können, dass die CA in der Nützlichkeit der EFA überlegen ist, so ist sie in der Anwendung auf psychometrische Daten der EFA vorzuziehen. Man darf aber nicht vergessen, dass man in der Psychometrie mit gemessenen Konstrukten nicht nur etwas vorhersagen können will, sondern auch immer inhaltlich sinnvoll Eigenschaften messen möchte. Das heißt, man möchte auch wissen, was man mit der Gruppe an Items misst, die zusammengefasst wurde. Dies ist aber eine Herausforderung, die auch bei Anwendung der FA letztlich über eine sorgsame Entwicklung der Items und eine anschließende inhaltliche Interpretation der gefundenen Item-Gruppen entschieden werden muss. Der Anwender muss darüber entscheiden, was die inhaltliche Bedeutung der gefundenen Gruppen ist und ob die gefundene Struktur inhaltlich sinnvoll interpretierbar ist.

In der vorliegenden Arbeit wurde aufgezeigt, welche Probleme die Annahme von Modellen in realen Datensätzen mit sich bringen kann und wie die *Real World Simulation* dabei helfen kann, echte Daten besser zu verstehen. Außerdem wurde eine Alternative zu der modellbasierten FA zur Strukturfindung in der Psychometrie vorgestellt, die ihr gegenüber einige Vorteile aufweist. Gegenüber bisherigen CA Verfahren zur Clusterung von Items bringt sie den Vorteil mit, dass gewichtete Clusterwerte berechnet werden können.

Literaturverzeichnis

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19* (6), 716–723.

Amthauer, R., Brocke, R., Liepmann, D. & Beauducel, A. (2007). *Intelligenz-struktur-test 2000 r - IST 2000-r* (2. erw. Aufl.). Göttingen: Hogrefe.

Anderson, R. E., Hair, J. F., Tatham, R. L. & Black, W. C. (2006). *Multivariate data analysis*. Pearson.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Arrindell, W. A. & Ende, J. v. d. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement*, *9* (2), 165–178.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438.

Bacon, D. R. (2001). An evaluation of cluster analytic approaches to initial model specification. *Structural Equation Modeling*, *8* (3), 397–429.

Barrett, P. T. & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study & Group Behaviour*, *1* (1), 23–33.

Beauducel, A. (2001). Problems with parallel analysis in data sets with oblique simple structure. *Methods of Psychological Research Online*, *6* (2), 141–157.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107* (2), 238–246.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Unveröffentlichte Dissertation, University of Groningen, Amsterdam: Sociometric Research Fonudation.

Boomsma, A. (2013). Reporting monte carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *20* (3), 518–540.

Borg, I. & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Brock, G., Pihur, V., Datta, S. & Datta, S. (2008). *clValid: An r package for cluster validation* (Bd. 25).

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika*, *33* (3), 267–334.

Buehner, M. (2011). *Einführung in die test- und fragebogenkonstruktion*. Pearson Deutschland GmbH.

Buhmann, J. M. (2010). Information theoretic model validation for clustering. *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 1398–1402.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, *33* (2), 261–304.

Carsey, T. M. & Harden, J. J. (2013). *Monte carlo simulation and resampling methods for social science* (Auflage: 1 Aufl.). Los Angeles: Sage Pubn Inc.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action: Its structure, growth and action*. Elsevier.

Cattell, R. B. & Eber, H. W. (1972). *The sixteen personality factor questionnaire (16pf)*. Champaign, Illinois, USA: Institute for Personality and Ability Testing.

Chavent, M., Genuer, R., Kuentz-Simonet, V., Liquet, B. & Saracco, J. (2013). *ClustOfVar: an r package for dimension reduction via clustering of variables. application in supervised classification and variable selection in gene expressions data*. Amsterdam.

Chavent, M., Kuentz, V., Liquet, B. & Saracco, L. (2011). *ClustOfVar: An r package for the clustering of variables*.

Church, A. T. & Burke, P. J. (1994). Exploratory and confirmatory tests of the big five and tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology*, *66* (1), 93–114.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton Univ. Press.

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D. & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, *70* (6), 885–901.

Cudeck, R. & Henly, S. J. (1991). Model selection in covariance structures analysis and the problem of sample size: A clarification. *Psychological Bulletin*, *109* (3), 512–519.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P. & Meester, L. E. (2006). *A modern introduction to probability and statistics: Understanding why and how*. Springer Science & Business Media.

Eysenck, H. J. & Eysenck, S. B. G. (1975). *Manual of the eysenck personality questionnaire (junior and adult)*. Hodder and Stoughton.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4* (3), 272.

Fahrmeir, L., Brachinger, W., Hamerle, A. & Tutz, G. (1996). *Multivariate statistische verfahren*. Walter de Gruyter.

Fan, X. & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12* (3), 343–367.

Gerbing, D. W. & Anderson, J. C. (1992). Monte carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods & Research*, *21* (2), 132–160.

Gerbing, D. W. & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3* (1), 62–72.

Gilmour, J. & Walters, S. (1964). Philosophy and classification. In W. Turrill (Hrsg.), *Vistas in botany* (S. 1–22). Oxford: Pergamon Press.

Glorfeld, L. (1995). An improvement on horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and psychological measurement*, *55* (3), 377–793.

Guadagnoli, E. & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological bulletin*, *103* (2), 265.

Hakstian, A. R., Rogers, W. T. & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research*, *17* (2), 193–219.

Handl, J., Knowles, J. & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, *21* (15), 3201–3212.

Hartigan, J. A. & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28* (1), 100–108.

Haynes, C. A., Miles, J. N. V. & Clements, K. (2000). A confirmatory factor analysis of two models of sensation seeking. *Personality and Individual Differences*, *29* (5), 823–839.

Hayton, J. C., Allen, D. G. & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, *7* (2), 191–205.

Hendrickson, A. E. & White, P. O. (1964). Promax: a quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*, 65–70.

Homburg, C. (1991). Cross-validation and information criteria in causal modeling. *Journal of Marketing Research*, 137–144.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30* (2), 179–185.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24* (6), 417–441.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6* (1), 1–55.

Humphreys, L. G. & Montanelli Jr., R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, *10* (2), 193–205.

Hunter, J. E. (1973). Methods of reordering the correlation matrix to facilitate visual inspection and preliminary cluster analysis. *Journal of Educational Measurement*, *10* (1), 51–61.

Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 2204–2214.

Jackson, D. L. (2007). The effect of the number of observations per parameter in misspecified confirmatory factor analytic models. *Structural Equation Modeling*, *14* (1), 48–76.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, *31* (8), 651–666.

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of strucutural equation models. *Multivariate Behavioral Research*, *23* (1), 69–86.

Kleinberg, J. (2003). An impossibility theorem for clustering. In S. Becker, S. Thrun & K. Obermayer (Hrsg.), *Advances in neural information processing systems.* Cambridge, MA: MIT Press.

Kuha, J. (2004). AIC and BIC comparisons of assumptions and performance. *Sociological Methods & Research*, *33* (2), 188–229.

La Du, T. J. & Tanaka, J. S. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology*, *74* (4), 625–635. doi: 10.1037/0021-9010.74.4.625

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind* (Bd. 29). Chicago, IL: The University of Chicago Press.

Lance, C. E., Butts, M. M. & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria what did they really say? *Organizational Research Methods*, *9* (2), 202–220.

Loevinger, J., Gleser, G. C. & DuBois, P. H. (1953). Maximizing the discriminating power of a multiple-score test. *Psychometrika*, *18* (4), 309–317.

Lubke, G. & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, *41* (4), 499–532.

MacCallum, R. & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109* (3), 502–511.

MacCallum, R., Widaman, K. F., Preacher, K. J. & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, *36* (4), 611–637.

MacCallum, R., Widaman, K. F., Zhang, S. & Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, *4*, 84–99.

Marsh, H. W., Balla, J. R. & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103* (3), 391–410.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6* (1), 97–103.

McDonald, R. P. & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, *107* (2), 247–255.

Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50* (2), 159–179.

Mooney, C. Z. (1997). *Monte carlo simulation*. Thousand Oaks, CA: Sage.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research*, *22* (3), 267–305.

Mundfrom, D. J., Shaw, D. G. & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5* (2), 159–168.

Muthén, L. K. & Muthén, B. O. (2002). How to use a monte carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9* (4), 599–620.

Ostendorf, F. & Angleitner, A. (2004). *NEO-persönlichkeitsinventar nach costa und McCrae, revidierte fassung*. Göttingen: Hogrefe.

Patil, V. H., Singh, S. N., Mishra, S. & Todd Donavan, D. (2008). Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one'criterion. *Journal of Business Research*, *61* (2), 162–170.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J. & Chen, F. (2001). Monte carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8* (2), 287–312.

Pearson, K. (1901). LIII. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, *2* (11), 559–572.

Pennell, R. (1968). The influence of communality and on the sampling distributions of factor loadings. *Psychometrika*, *33* (4), 423–439.

Peterson, R. A. (2000). A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing Letters*, *11* (3), 261–275.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66* (336), 846–850.

Rao, R. C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, *37* (3), 81–91.

Rencher, A. C. & Christensen, W. F. (2012). *Methods of multivariate analysis*. John Wiley & Sons.

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14* (1), 57–74.

Revelle, W. (2011). *An overview of the psych package*. Citeseer.

Revelle, W. (2014). *The personality project [book in preperation]*. Zugriff auf `https://personality-project.org/`

Rost, J. (2004). *Lehrbuch testtheorie - testkonstruktion* (2. Aufl.). Bern: Verlag Hans Huber.

Ruscio, J. & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological assessment*, *24* (2), 282.

Sass, D. A. (2010). Factor loading estimation error and stability using exploratory factor analysis. *Educational and Psychological Measurement*, *70* (4), 557–577.

Sass, D. A. & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45* (1), 73–103.

Schmitt, T. A. & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement*, *71* (1), 95–113.

Schonemann, P. H. & Steiger, J. H. (1978). On the validity of indeterminate factor scores. *Bulletin of the Psychonomic Society*, *12* (4), 287–290.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6* (2), 461–464.

Schweizer, K. (1991). Classifying variables on the basis of disaggregate correlations. *Multivariate behavioral research*, *26* (3), 435–455.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *3* (1), 25–40.

Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*.

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, *15* (2), 201–292.

Thurstone, L. L. (1935). *The vectors of mind*. Chicago, IL: University of Chicago.

Tryon, R. C. (1935). A theory of psychological components–an alternative to" mathematical factors.". *Psychological Review*, *42* (5), 425–454.

Tryon, R. C. (1939). *Cluster analysis: Correlation profile and orthometric factor analysis for the isolation of unities in mind and personality*. Ann Arbor: Edward Brothers.

Tryon, R. C. & Bailey, D. E. (1970). *Cluster analysis*. McGraw-Hill.

Tucker, L. R., Koopman, R. F. & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34* (4), 421–459.

Van der Linden, D., Tsaousis, I. & Petrides, K. V. (2012). Overlap between general factors of personality in the big five, giant three, and trait emotional intelligence. *Personality and Individual Differences*, *53* (3), 175–179.

van Dongen, S. & Enright, A. J. (2012). Metric distances derived from cosine similarity and pearson and spearman correlations. *arXiv:1208.3145 [cs, stat]*. (arXiv: 1208.3145)

Velicer, W. F. (1976). The relation between factor score estimates, image scores, and principal component scores. *Educational and Psychological Measurement*, *36* (1), 149–159.

Velicer, W. F., Eaton, C. A. & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In *Problems and solutions in human assessment* (S. 41–71). Springer.

Velicer, W. F. & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological methods*, *3* (2), 231.

Vigneau, E. & Qannari, E. M. (2003). Clustering of variables around latent components. *Communications in Statistics - Simulation and Computation*, *32* (4), 1131–1150.

v. Luxburg, U., Williamson, B. & Guyon, I. (2012). Clustering: Science or art? *Journal of Machine Learning Research Workshop and Conference Proceedings*, *27*, 65–80.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the akaike information criterion (AIC) and the bayesian information criterion (BIC). *Psychological methods*, *17* (2), 228.

Wilson, E. B. (1928). On hierarchical correlation system. *Proceedings of the National Academy of Sciences*, *14*, 283–291.

Yuan, K., Marshall, L. L. & Bentler, P. M. (2003, Januar). Assessing the effect of model misspecifications on parameter estimates in structural equation models. *Sociological Methodology*, *33* (1), 241–265. Zugriff am 2014-10-15 auf `http://onlinelibrary.wiley.com/doi/10.1111/j.0081-1750.2003.00132.x/abstract` doi: 10.1111/j.0081-1750.2003.00132.x

Zwick, W. R. & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, *17* (2), 253–269.

Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, *99* (3), 432.

Danksagung

Vielen Menschen habe ich zu danken, die mich in dieser aufregenden Phase meiner akademischen Laufbahn begleitet haben. Zu besonderem Dank bin ich meinen beiden Betreuern verpflichtet. An erster Stelle möchte ich mich bei meinem Betreuer Prof. Dr. Markus Bühner bedanken, der immer an mich geglaubt hat, auch wenn ich selbst den Glauben völlig verloren hatte und der mir jede Freiheit gegeben hat, die ich gebraucht habe und gleichzeitig an richtiger Stelle den nötigen Druck gemacht hat, was ihn zum besten Chef macht, den man sich wünschen kann. Darüber hinaus hat er mir die richtige Mischung aus wissenschaftlicher Begeisterungsfähigkeit und Pragmatismus vermittelt. Ein sehr großer Dank gilt Herrn Prof. Dr. Moritz Heene für den inhaltlich sehr bereichernden Austausch, aber auch für die vielen interessanten und hilfreichen „Mentorengespräche“, die mir sehr ans Herz gewachsen sind. Er hat mir die kritische Betrachtung der Wissenschaft aber auch die Demut angesichts der eigenen Unwissenheit gelehrt. Can Gürer möchte ich danken, da er mir nicht nur fachlich hoch kompetent zur Seit stand und auf beinahe jede Frage eine Antwort wusste, sondern auch weil er mir in den letzten Jahren zu einem sehr guten Freund geworden ist. Darüber hinaus möchte ich mich bei Herrn Prof. Dr. Küchenhoff für seine statistische Unterstützung danken, die es ermöglicht hat, der Arbeit den nötigen mathematischen Anstrich zu verleihen, was ihr einen großen Qualitätszuwachs gebracht hat. Eng damit verbunden ist der Dank an meinen hoch geschätzten Kollegen und Mitarbeiter Andreas Hölzl, der mir gerade in computationalen und statistischen Fragen immer zuverlässig zur Seite stand und das oft noch in letzter Minute vor einem Kongressvortrag. Für die ständige administrative Meisterleistung und moralische Unterstützung möchte ich mich außerdem ganz herzlich bei Cora Laugs bedanken, die immer für mich da war und jedes Problem sofort gelöst hat. Mein Dank geht auch an Florian Pargent und Julia Gsottschneider für das Korrekturlesen und die hilfreichen Kommentare. Nicht zuletzt möchte ich mich noch bei meiner Mutter Juliane Bollmann sowie bei Friederike Dietrich für die menschliche Unterstützung während dieser drei Jahre und ganz besonders in der harten Endphase und anderen Krisenzeiten bedanken.

Ebenso möchte ich mich für hilfreiche Anmerkungen und Beiträge, die diese Arbeit bereichert haben, bedanken bei Dr. Clemens Draxler, Prof. Dr. Gerhard Tutz, Christine Anderl, Florian Niedermann und Prof. Dr. Bernd Bischl.

Hiermit erkläre ich, Stella Bollmann, dass ich die vorliegende Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe. Die vorliegende Dissertation wurde in dieser oder anderer Form noch nicht als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt.

München, den 9.Oktober 2014 Stella Bollmann