Bachelorarbeit

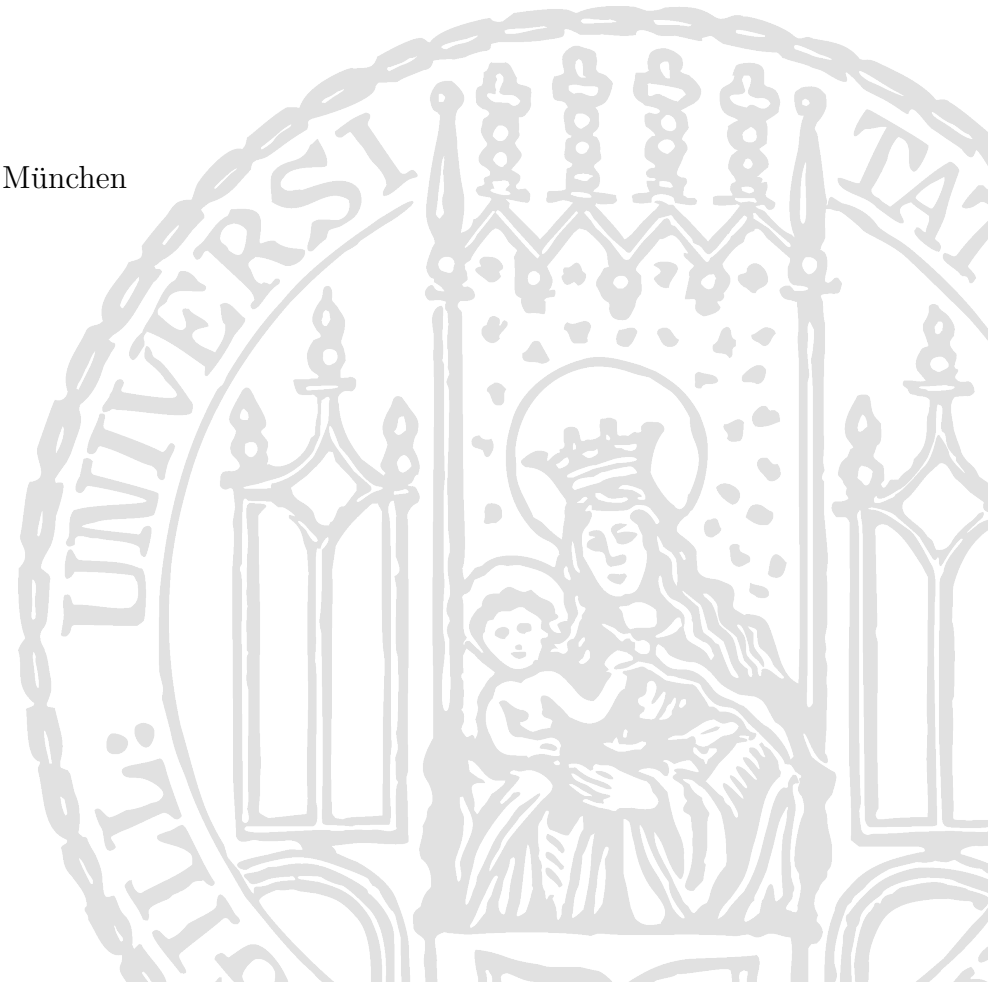
Clusteranalyse zur Gruppierung von Items:

Strategien zur Auffindung von Faktorstrukturen

Andreas Hölzl

Betreuer:
Prof. Dr. Helmut Küchenhoff
Stella Bollmann

23. Juni 2013
Institut für Statistik
Ludwig-Maximilians-Universität München



Abstract

In dieser Arbeit wird ein neues Qualitätsmerkmal zur Beurteilung von Clusterverfahren für ordinalskalierte Itemdaten vorgestellt, die sogenannte Faktorstrukturerkennung. Basierend auf einer theoretischen Faktorenstruktur gegeben durch Faktorladungen und Faktorkorrelationen wird untersucht, welche Clusterverfahren diese Faktorstruktur am besten wiederfinden. Neben den bekannten hierarchischen Verfahren werden drei neu entwickelte Clusterverfahren für ordinalskalierte Itemdaten, K-Means-MDS, K-Means-Koord und K-Means-Kor anhand dieses neu entwickelten Qualitätsmerkmals beurteilt. Alle diese Clusterverfahren werden auch auf das bereits bekannte Clusterqualitätsmerkmal der Strukturerhaltung getestet.

Inhaltsverzeichnis

1 Einführung

1.1 Motivation

In der Psychologie werden sehr oft Daten mithilfe von ordinalskalierten Fragebögen gewonnen. Für diese Art der Daten wird dann häufig auch untersucht, welche Items des Fragebogens sich untereinander ähnlich sind, also von den Personen ähnlich beantwortet worden sind. Hierfür werden die Items mithilfe von Clusterverfahren in Gruppen aufgeteilt. Es gibt nun verschiedene Clusterverfahren, die zu unterschiedlichen Ergebnissen führen. Allerdings ist es schwer zu sagen, welche dieser Clusterverfahren besser sind. In dieser Arbeit wird deshalb ein Ansatz vorgestellt, mit dem in diesem Fall verschiedene Clusterverfahren nach ihrer Güte beurteilt werden können.

1.2 Übersicht

In dieser Arbeit wird untersucht, welche Clusterverfahren sich am besten dafür eignen, Faktorstrukturen in ordinalen Itemdaten wiederzufinden. In Kapitel 2 wird darauf eingegangen, wie die Faktorstrukturen erzeugt werden, wie sie variiert werden können und wie man aus den Faktorstrukturen eine Korrelationsmatrix erzeugen kann. In Kapitel 3 und 4 werden dann verschiedene Clusterverfahren vorgestellt, die auf die Korrelationsmatrizen der gegebenen Faktorenstrukturen angewandt werden können. Im nächsten Kapitel wird untersucht, wie verschiedene Clusterungen miteinander verglichen werden können und in Kapitel 6 wird vorgestellt, wie beurteilt wird, welche Clusterverfahren die Faktorstruktur am besten wiederfinden. Es kann nämlich die Faktorstruktur auch als Clusterung interpretiert werden, indem man jedes Item dem Cluster mit der größten Hauptladung zuweist. Für die verschiedenen Clustervergleichsmethoden können nun die theoretische Clusterung der Faktorstruktur mit den sich ergebenden Clusterungen der verschiedenen Clusterverfahren verglichen werden. In Kapitel 7 wird zuerst etwas theoretische Vorarbeit geleistet. Die Ergebnisse werden im folgenden Kapitel dann untersucht. In Kapitel 9 werden die vorgestellten Clusterverfahren auch noch auf ein weiteres Qualitätsmerkmal untersucht, nämlich wie sehr bei ihnen die Clusterstruktur erhalten bleibt, wenn nicht der Gesamtdatensatz zum Bestimmen der Korrelationen verwendet wird, sondern nur eine Stichprobe.

2 Simulation der Faktoranalysedaten

2.1 Erklärung der Faktorenanalyse

2.1.1 Lineares Faktorenmodell

Die Faktoranalyse ist ein Verfahren, mit dem beobachtbare Variablen auf weniger zugrunde liegende Variablen zurückgeführt werden. Nach [Harman, 1976, Seite 15] werden die beobachtbaren Variablen z_j folgendermaßen linear durch latente Variablen dargestellt:

$$z_j = \sum_{p}^{m} a_{jp} F_p + u_j Y_j$$

oder in Matrixschreibweise:

$$Z = AF + uY \tag{2.1.1}$$

Die F_i stellen dabei die latenten Variablen dar und die a_{ji} geben den Einfluss der i-ten latenten Variable auf die j-te beobachtbare Variable an. Der Term $d_j U_j$ ist der Residuenfehler, der oftmals auch als ϵ_j bezeichnet wird und gibt die Abweichung des theoretischen Wertes der Faktorenanalyse $\sum_i^m a_{ji} F_i$ vom tatsächlich beobachteten Wert z_j an. Wir betrachten ordinalskalierte Itemdaten. z_{ij} gibt an, welchen der ordinalskalierten Werte Individuum i bei Item j hat. Die Faktoranalyse kann also nach [Harman, 1976, Seite 15] explizit geschrieben werden als:

$$z_{ij} = \sum_{p}^{m} a_{jp} F_{pi} + u_j Y_{ij}$$

Dabei ist F_{pi} die Ausprägung des Individuums i bei der latenten Variable p und a_{jp} gibt den Einfluss der latenten Variable p auf die beobachtbare Variable j an. Die a_{jp} werden als Faktorladungen bezeichnet. $u_j Y_{ij}$ ist der Fehlerterm bei Individuum i und Item j.

2.1.2 Bestimmen der Faktorladungen

Für die Untersuchungen dieser Arbeit ist die Anzahl der Faktoren immer schon vorgegeben, von daher wird nicht darauf eingegangen, wie sich die beste Anzahl an Faktoren berechnen lässt. Es soll hier erklärt werden, wie sich gegeben einer Anzahl an zu extrahierenden Faktoren die Faktorladungen bestimmen lassen. In der geometrischen Interpretation der Faktoranalyse nach [Harman, 1976, Teil 1 Kapitel 4.9] werden die gegebenen Variablenausprägungen als Vektoren in-

terpretiert, und zwar derartig, dass die Korrelation zwischen zwei Items den Winkel zwischen den diese Items repräsentierenden Vektoren bestimmt. In einem p-dimensionalen Raum können alle möglichen Korrelationsmatrizen der Items derartig dargestellt werden. Ziel der Faktoranalyse ist es nun, diese p-Vektoren so als Summen einer bestimmten Anzahl q mit $q < p$ an Vektoren darzustellen, dass die Winkel zwischen den sich so ergebenden Linearkombinationsvektoren möglichst gut die Korrelationen zwischen diesen derartig dargestellten Items darstellen.

2.1.3 Rotation

Wie in [Manhart und Hunger, 2008, Kapitel 2] erklärt wird, gibt es viele verschiedene Lösungen, die die Korrelationsmatrix alle gleich gut abbilden und die durch Rotation der Koordinatenachsen erzeugt werden können. Ersetzt man in 2.1.1 A durch $A* = AM$ und F durch $F* = M^T F$, wobei M eine orthogonale Matrix sei, so gilt:

$$Z* = A * F * + dU = AMM^T F + dU = AMM^{-1} F + dU = AF + dU = Z$$

Da M eine orthogonale Matrix ist, gilt: $M^T = M^{-1}$. Der systematische Teil, der die Daten erklärt, bleibt also gleich, obwohl anderen Ladungen vorliegen. Es gibt zwei Arten von Rotationen. Bei orthogonalen Transformationen stehen die Faktorvektoren, die die Variablen linear aufspannen, alle senkrecht aufeinander, sind also unkorreliert. Bei obliquen Rotationen können die Faktoren auch untereinander korreliert sein. Es gibt verschiedene Möglichkeiten, wie man nun die optimale Rotation bestimmt. In dieser Arbeit wurde die Promax-Rotation angewendet. Wie in [Manhart und Hunger, 2008, Abschnitt 5.4.2] beschrieben, ist das Ziel der Proxmax-Rotation, möglichst große Hauptladungen und möglichst kleine Nebenladungen zu erhalten. Dies ermöglicht eine leichtere Interpretation der Faktoranalyse.

2.2 Erzeugung der Faktorenanalysedaten

In diesem Abschnitt wird erklärt, wie die Faktorstrukturen, auf der dann weitere Analysen durchgeführt werden, erzeugt werden. Als Grundlage der Faktorstruktur werden Daten eines psychologische Fragebogens mit dem Themengebiet Persönlichkeit genommen. Dieser Fragebogen mit dem Namen NEO-PI-R enthält 240 Fragen zu 5 verschiedenen Persönlichkeitsfacetten mit je 6 Unterfacetten. Pro Unterfacette gibt es also 8 Fragen, die im Folgenden immer als Items bezeichnet werden. Die untersuchten Persönlichkeitsfacetten sind Neurotizismus, Extraversion, Offenheit für Erfahrung, Verträglichkeit und Gewissenhaftigkeit und jede dieser Persönlichkeitsfacetten hat wieder 6 Unterfacetten. So sind beispielsweise die Unterfacetten von Gewissenhaftigkeit: Kompetenz, Ordnungsliebe, Pflichtbewusstsein, Leistungsstreben, Selbstdisziplin und Besonnenheit. Weiteres zu dem Fragebogen kann in [Ostendorf und Angleitner, 2004] nachgelesen werden. Jedes Item wird auf einer ordinalen Skala von 1 bis 5 beantwortet und es liegen Daten von 11724

Personen [1] vor. Es wurden aus allen 5 verschiedenen Persönlichkeitsfacetten jeweils eine Unterfacette ausgewählt und diese 5 Unterfacetten wurden dann als Grundlage dieser Arbeit genommen. Die Faktoranalyse wird folglich auf 40 Items aus 5 Unterfacetten ausgeführt. Die Unterfacetten werden aus verschiedenen Hauptfacetten gewählt, da sie dann nicht so sehr korreliert sind. Für die Untersuchungen dieser Arbeit wurden folgenden Unterfacetten gewählt: N1, E2, O3, A4 und C5 (Ängstlichkeit aus der Facette Neurotizismus, Geselligkeit aus Extraversion, Gefühle aus Offenheit, Entgegenkommen aus Verträglichkeit und Selbstdisziplin aus Gewissenhaftigkeit).
Auf diesen Daten wird zunächst die Faktorenanalyse mit 5 zu extrahierenden Faktoren angewandt. Dieses Ergebnis wird als Grundlage der Untersuchung genommen. Für spätere Untersuchungen werden aber nur die Hauptladungen übernommen, sowie die Faktoren zunächst als unkorreliert angenommen.
Die Faktorladungen sehen ausschnittsweise so aus:

	$ML1$	$ML3$	$ML2$	$ML4$	$ML5$
V1	0.62	0.05	0.01	−0.02	0.00
V31	0.60	0.04	0.05	0.02	0.04
V61	0.75	0.04	0.03	0.05	0.05
V91	0.53	−0.06	−0.09	−0.03	−0.08
V121	0.53	0.01	−0.05	0.04	−0.01
V151	0.67	0.04	−0.01	−0.09	−0.03
V181	0.68	0.03	0.03	0.03	0.06

Dies bedeutet, dass beispielsweise die Variable $V1$ als folgende Linearkombination der Faktoren dargestellt werden kann:

$$V1 = 0.62ML1 + 0.05ML3 + 0.01ML2 - 0.023Ml4 + 0.00ML5$$

Für die Faktoranalyse, die wir verwenden, wird eine oblique Rotation durchgeführt, die Faktoren können also untereinander korreliert sein. Deshalb ist für die Beschreibung des Faktormodells auch noch die Korrelationen der Faktoren wichtig. Diese sieht bei den Daten ursprünglich so aus:

[1] Ursprünglich bestand der Datensatz aus 12003 befragten Personen, allerdings wurden die Patienten, bei denen eine psychische Krankheit identifiziert wurde, entfernt. Für derartige Patienten ist der Fragebogen nicht geeignet.

	$ML1$	$ML3$	$ML2$	$ML4$	$ML5$
$ML1$	1.00	−0.33	−0.14	0.19	−0.06
$ML3$	−0.33	1.00	0.01	−0.08	0.06
$ML2$	−0.14	0.01	1.00	0.28	0.10
$ML4$	0.19	−0.08	0.28	1.00	−0.02
$ML5$	−0.06	0.06	0.10	−0.02	1.00

Für die Datengeneration in dieser Arbeit wird allerdings die ursprüngliche Faktorstruktur nur als Grundlage genommen. Die Hauptladungen der Faktorladungen werden übernommen, die Korrelationen zweier verschiedener Faktoren und die Nebenladungen werden allerdings verändert. Die Hauptladungen sind dabei für jede Variable der Faktor, der die größte Ladung für diese Variable hat. Alle anderen Faktoren werden als Nebenladungen dieser Variable bezeichnet. Werden nun beispielsweise die Nebenladungen alle auf 0 gesetzt, so ergibt sich die folgende Faktorstruktur:

	$ML1$	$ML3$	$ML2$	$ML4$	$ML5$
V1	0.62	0	0	0	0
V31	0.60	0	0	0	0
V61	0.75	0	0	0	0
V91	0.53	0	0	0	0
V121	0.53	0	0	0	0
V151	0.67	0	0	0	0
V181	0.68	0	0	0	0

und die Faktorkorrelationen auch auf 0 gesetzt Dass die ersten Variablen alle die gleichen Hauptladungen haben, ist dadurch bedingt, wie die Variablen ausgewählt worden sind. Werden alle Faktorkorrelationen auf 0 gesetzt, so ergeben sich die folgenden Faktorkorrelationen:

	$ML1$	$ML3$	$ML2$	$ML4$	$ML5$
$ML1$	1.00	0	0	0	0
$ML3$	0	1.00	0	0	0
$ML2$	0	0	1.00	0	0
$ML4$	0	0	0	1.00	0
$ML5$	0	0	0	0	1.00

Die 0 als Wert für die Nebenladungen und Faktorkorrelationen ist allerdings nur ein Beispiel und die Nebenladungen und Faktorkorrelationen können auch anders gewählt werden. Für die Untersuchungen werden die Nebenladungen beispielsweise schrittweise immer weiter erhöhrt,

um zu sehen, welche Clusterverfahren die Faktorstruktur trotzdem noch am besten erkennen. Ähnlich kann auch mit den Faktorkorrelationen verfahren werden.

2.3 Clustern aus Faktorenanalysedaten

Jede Faktorstruktur kann auch als Clusterung interpretiert werden, indem die Faktoren als Cluster interpretiert werden und jede Variable dem Cluster zugewiesen wird, dessen entsprechender Faktor die Hauptladung dieses Items ist. Etwas formaler lässt sich dies schreiben als:

$$C_j = \{c_i : \max_j(|\gamma_{ij}|)\}$$

Den Items aus der Beispielsfaktorladung der obigen Tabelle würden also alle Cluster 1 zugewiesen werden.

2.4 Korrelationsmatrix aus Faktorenanalysedaten

2.4.1 Erzeugen der Korrelationsmatrix des Modells

Für eine bestimmte Faktorstruktur soll nun die Korrelationsmatrix der Items bestimmt werden. Jedes Item ist gegeben durch die Faktorstruktur, betrachten wir hier den systematischen Teil der zwei Items A und B, der durch die Faktorstruktur gegeben ist.

$$A = \sum_i^n a_i F_i$$

und

$$B = \sum_j^n b_j F_j$$

Für die Kovarianz von A_i und B_i gilt:

$$Cov(A, B) = Cov(\sum_i^n a_i F_i, \sum_j^n b_j F_j) = \sum_i \sum_j a_i b_j Cov(F_i, F_j)$$

Nach [Harman, 1976, Seite 16] sind allerdings sowohl alle Variablen als auch alle Faktoren entsprechend standardisiert, dass sie also eine Varianz von 1 haben. Somit entspricht die Korrelation der Kovarianz. Folglich:

$$Cor(A, B) = Cor(\sum_i^n a_i F_i, \sum_j^n b_j F_j) = \sum_i \sum_j a_i b_j Cor(F_i, F_j)$$

Nach dieser Formel kann also die Korrelation zwischen allen Items bestimmt werden. Somit kann eine Korrelationsmatrix der zugrunde liegenden Faktorstruktur für alle Variablen erzeugt werden. Diese sieht dann beispielsweise ausschnittsweise folgendermaßen aus:

	V1	V31	V61	V91	V121
V1	1.0000000	0.6214919	0.7167665	0.5783702	0.5809310
V31	0.6214919	1.0000000	0.6965368	0.5641386	0.5665884
V61	0.7167665	0.6965368	1.0000000	0.6447042	0.6477822
V91	0.5783702	0.5641386	0.6447042	1.0000000	0.5298398
V121	0.5809310	0.5665884	0.6477822	0.5298398	1.0000000

2.4.2 Berücksichtigung der Fehlerkorrelation

Es ist nun bekannt, wie die Korrelationsmatrix für eine theoretische Faktorstruktur berechnet werden kann. Für die Korrelationsmatrizen von realen Daten kommt aber immer auch noch ein Fehlerterm zu dieser theoretischen Faktorstruktur hinzu. Dies ist der Teil der Kovarianz der Daten, der nicht durch das Faktormodell beschrieben werden kann. In 2.1.1 wurde dieser Fehlerterm mit $u_j Y_{ij}$ bezeichnet. Für unsere Untersuchungen, die den Fehlerterm berücksichtigen, wird als Fehlerterm der Fehlerterm des ursprünglichen Faktormodells genommen. Wenn die theoretische Faktorstruktur des Modells verändert wird, indem beispielsweise die Nebenladungen auf 0.1 gesetzt werden, so wird aber immer noch die gleiche Fehlerkorrelation des ursprünglichen Faktormodells übernommen. Eine direkte Berechnung des Fehlerterms für diese Modelle ist nicht möglich, da es sich dabei dann um rein theoretische Modelle handelt ohne tatsächliche Daten auf denen sie basieren. Als Annäherung eines Fehlerterms, der nicht durch das Modell erklärt wird, wird also der Fehlerterm eines anderen Modells genommen. Wenn man den Fehlerterm hinzuaddiert, liegt allerdings keine Korrelationsmatrix, sondern eine Kovarianzmatrix vor. Aus dieser Kovarianzmatrix kann aber wieder die Korrelationsmatrix berechnet werden und diese Korrelationsmatrix wird für die Anwendung der Clusterverfahren verwendet.

3 Hierarchische Clusterverfahren

Hierarchische Clusterverfahren sind Clusterverfahren, die basierend auf Distanzen zwischen Items Cluster aus Items bestimmen. Diese sollten die Eigenschaft haben, dass die Distanzen zwischen Items in einem Cluster möglichst klein sind und die Distanzen zwischen Items in verschiedenen Clustern möglichst groß sind. Hier werden nur agglomerative Verfahren vorgestellt, also Verfahren, in denen zuerst jedes Item ein eigener Cluster ist. Die hierarchische Clusterung wird in [Xu und Wunsch, 2008, Kapitel 3] gut erklärt. Nach ihrer Ähnlichkeit zueinander werden diese Cluster dann schrittweise immer weiter zu größeren Clustern zusammengefasst. Im Folgenden werden alle nötigen Begriffe für dieses Verfahren vorgestellt.

3.1 Distanzmaße

Zunächst werden zwei Ähnlichkeitsmaße vorgestellt.

- Das erste Ähnlichkeitsmaß ist die Pearson-Korrelation zwischen zwei Items. Die vorliegenden Daten sind zwar Ordinaldaten, doch in der Psychologie ist es üblich, die Korrelationen von ordinalen Item-Daten auch durch die Pearson-Korrelation zu bestimmen. Die Formel zur Bestimmung des ersten Ähnlichkeitsmaßes zwischen zwei Items A und B lautet also:

$$s(a,b) = \rho(a,b) = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2(b_i - \bar{b})^2}}$$

  $\rho(a,b)$ ist also der Pearson-Korrelationskoeffizient zwischen den Items A und B.

- Das zweite Ähnlichkeitsmaß basiert auf der Idee, dass zwei Items sich ähnlich sind, wenn die Korrelationen dieser zwei Items zu allen anderen Items ähnlich sind. Hierfür wird die Korrelation der Korrelation zwischen zwei Items berechnet. Diese Idee wurde in [Bacon, 2001, Seite 400 f] zum ersten Mal vorgestellt. Die Formel lautet also:

$$s_{A,B} = \frac{\sum_{i=1}^{n}(\rho_{A,i} - \bar{\rho}_A)(\rho_{B,i} - \bar{\rho}_B)}{\sqrt{\sum_{i=1}^{n}(\rho_{A,i} - \bar{\rho}_A)^2(\rho_{B,i} - \bar{\rho}_B)^2}}$$

Dabei ist also $\rho_{A,i}$ die Pearson-Korrelation zwischen dem Item A und dem $i-ten$ Item. Aus diesem Ähnlichkeitsmaß sollte nun ein Distanzmaß erzeugt werden. Ein erster intuitiver Ansatz wäre es, die Distanz zu berechnen als $d(A,B) = 1 - s(A,b)$. Je größer die Korrelation ist, desto niedriger ist der Abstand zwischen zwei Punkten und bei einer Korrelation von 1 beträgt der Abstand 0. Dieses Distanzmaß ist allerdings nach [van Dongen und Enright, 2012] keine Metrik und es wäre vielleicht sinnvoll, auch ein metrisches Distanzmaß zu betrachten. In [van Dongen und Enright, 2012] findet sich der Beweis, dass $d(A,B) = \sqrt{0.5 - 0.5s(A,b)}$ eine Metrik ist und deshalb wird auch dieses Distanzmaß betrachtet. Da nun zwei Ähnlichkeitsmaße sowie zwei Distanzfunktionen vorgestellt wurden, haben wir somit insgesamt 4 verschiedene Distanzmaße.

3.2 Fusionierungsverfahren

Es wird so vorgegangen, dass die Items mit der größten Ähnlichkeit immer zusammengefasst werden. Es stellt sich die Frage, wie die Ähnlichkeit von einem Item zu einem Cluster oder von zwei Clustern bestimmt wird. Hierfür gibt es mehrere Möglichkeiten, von denen im Folgenden zwei vorgestellt werden:

- Beim Complete-Linkage wird als Abstand zweier Cluster der maximale Abstand zwischen Items dieser zwei Cluster gewählt. Es gilt also:

$$D_{com}(A,B) := \max_{a \in A, b \in B} \{d(a,b)\}$$

- Beim Average-Verfahren wird als Abstand zweier Cluster der durchschnittliche Abstand zwischen den Items dieser zwei Cluster gewählt. Es gilt also:

$$D_{avg}(A,B) := \frac{1}{|A||B|} \sum_{a \in A, b \in B} d(a,b)$$

3.3 Vorgehen beim hierarchischen Clustern

Es wurden zwei mögliche Ähnlichkeitsmaße vorgestellt, die Pearson-Korrelation und die Korrelation der Korrelation. Außerdem wurden zwei Arten vorgestellt, mit denen aus dem Ähnlichkeitsmaß das Distanzmaß berechnet werden kann, $d(A,B) = 1 - s(A,b)$ und $d(A,B) = \sqrt{0.5 - 0.5s(A,b)}$. Bei jedem dieser Distanzmaße wird nun so vorgegangen, dass immer die Items mit den geringsten Abständen zusammengeclustert werden. Um den Abstand zwischen einem Item und einem Cluster oder zwischen zwei Clustern zu berechnen, kann man entweder das Average-Linkage-Verfahren oder das Complete-Linkage-Verfahren verwenden. Derartig werden die Items nun so lange zusammengefasst, bis genau die gesuchte Anzahl an Clustern vorhanden ist. Da bei der Faktoranalyse 5 Faktoren extrahiert werden sollte die Clusterung auch 5 Cluster haben. In der

Grafik 3.1 sind diese Cluster durch die roten Kästchen gekennzeichnet. Auf der x-Achse sind die verschiedenen Items und die y-Achse bezeichnet die Distanz zwischen den Items/Clustern, bei denen sie zusammengefasst worden sind. Da die Ähnlichkeitsmaße, die Funktion wie aus dem Ähnlichkeitsmaß das Distanzmaße werden und die Fusionierungsverfahren beliebig kombiniert werden können, gibt es nun 8 Möglichkeiten, wie hierarchische Clusterungen durchgeführt werden kann. Zunächst beschränken wir uns bei den Untersuchungen auf 4 Möglichkeiten, da wir als Abstandsmaß immer $d(A, B) = \sqrt{0.5 - 0.5s(A, b)}$ verwenden. In späteren Untersuchungen wir dann auf den Vergleich der Abstandsmaße eingegangen.

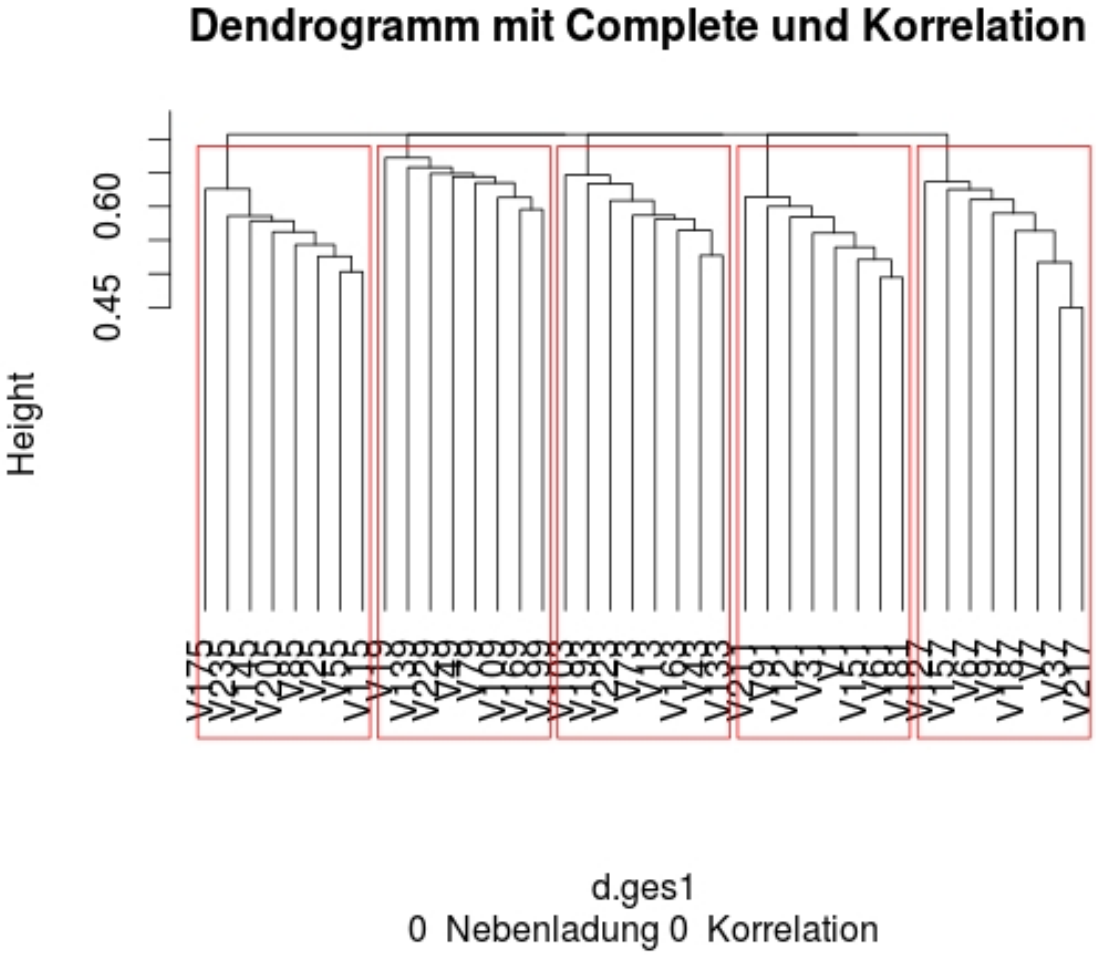


Abbildung 3.1: Beispielhaftes Dendrogramm

4 Clusterverfahren basierend auf K-Means

4.1 Allgemeines K-Means Clustering

Ein weiteres Clusterverfahren ist das sogenannte K-Means Clustering. Es wird im Gegensatz zu den hierarchischen Clusterverfahren nicht auf Distanzen angewendet, sondern auf Punkte im Koordinatensystem, auf denen eine Abstandsfunktion definiert ist. Diese sei im Folgenden immer die euklidische Distanzfunktion. Dafür, wie die Items in das Koordinatensystem überführt werden, gibt es verschiedene Möglichkeiten, die im Folgenden vorgestellt werden. Zunächst werde das K-Means-Verfahren orientiert an [Wu, 2012, 1.3] allgemein erklärt. Als Anfangszustand werden k zufällige Punkte der Menge jeweils einem Cluster zugeteilt, so dass jeder der k Cluster nun genau einen Punkt enthält. Dann wird folgendermaßen vorgegangen:

- Weise jeden Punkt dem Cluster zu, zu dessen Zentroiden, also Mittelpunkt aller Punkte in diesem Cluster, er am nächsten ist.
- Berechne die Zentroiden der neuen Cluster erneut.

Dies wird solange wiederholt, bis sich der Zustand nicht mehr ändert, dann ist die Clusterung in diesem Moment die sich ergebene Endclusterung. Das Ziel des K-Means Algorithmus ist es also, die quadrierten Abweichungen der Punkte eines Clusters von ihrem Mittelpunkt zu minimieren.

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

Es sei angemerkt, dass durch die Zufallsauswahl der ersten Cluster, bei gleichen Punkte-Koordinaten nicht immer die gleiche Clusterung als Ergebnis berechnet wird. Deshalb wird eine Clusterung immer öfter durchgeführt und dann dasjenige Ergebnis übernommen, das am öftesten auftritt.

4.2 K-Means Clustering der Korrelationen

In diesem Verfahren wird für jedes Item der Vektor der Korrelationen zu den anderen Items als Koordinaten dieses Items betrachtet. Mit diesen Koordinaten für die Punkte wird dann das K-Means Verfahren angewandt. Die Idee dieses Verfahrens ist es also, die Punkte danach zu clustern, wie ähnlich sich die Korrelationen sind. Betrachten wir beispielsweise die folgende Korrelationsmatrix:

	$V5$	$V35$	$V65$	$V95$
V5	1.000000000	0.08425798	0.30101154	0.13674140
V35	0.084257978	1.00000000	0.11323103	0.07252119
V65	0.301011538	0.11323103	1.00000000	0.20146350
V95	0.136741404	0.07252119	0.20146350	1.00000000

So wird Item V5 als Punkt $(1.000000000, 0.08425798, 0.30101154, 0.13674140)$ im kartesischen Koordinatensystem dargestellt. Item V35 ist der Punkt $(0.084257978, 1.00000000, 0.11323103, 0.07252119)$. Auf diesen Punkten im Koordinatensystem wird der K-Means Algorithmus angewandt. Der euklidische Abstand zweier Punkte in diesem Koordinatensystem zueinander lässt sich folgendermaßen berechnen:

$$d(A, B) = \sum_i (Korr(A, V_i) - Korr(B, V_i))^2$$

Je ähnlicher sich die Korrelationen zweier Items zu allen anderen Items sind, desto näher zusammen befinden sich die Items im Koordinatensystem und desto wahrscheinlicher ist es, dass die Items durch K-Means zusammen geclustert werden.

4.3 K-Means Clustering der Items in Koordinatensystem

4.3.1 Allgemeine Idee des Verfahren

Die Idee dieses Verfahrens ist es, dass zwei Items im Koordinatensystem näher zusammen sein sollten, wenn sie eine hohe Korrelation zueinander haben und dass sie weiter auseinander sein sollten, wenn sie eine niedrige Korrelation zueinander haben. Als erste Idee wäre es also sinnvoll, die Items so in ein Koordinatensystem zu überführen, dass die euklidischen Abstände zwischen den Punkten so groß sind wie die Distanzen $d(A, B) = 1 - Korr(A, B)$. Allerdings ist $d(A, B) = 1 - Korr(A, B)$ keine Metrik ([van Dongen und Enright, 2012]). Der euklidische Abstand ist allerdings eine Metrik und somit kann nicht jede Distanzmatrix mit diesem Abstandsmaß als euklidischer Abstand zwischen Punkten im Raum interpretiert werden. In [van Dongen und Enright, 2012] befindet sich allerdings auch der Beweis, dass $d(A, b) = \sqrt{0.5 - 0.5 * Korr(A, B)}$ eine Metrik ist und deshalb wird dieses Abstandmaß verwendet.
Mit diesem Abstandsmaß sollte es also möglich sein, die Items so in ein Koordinatensystem zu überführen, dass der Abstand zwischen zwei Items dem Abstand $d(A, b) = \sqrt{0.5 - 0.5 * Korr(A, B)}$ entspricht. Auf diesen Punkten kann dann K-Means ausgeführt werden.

4.3.2 Algorithmus für dieses Verfahren

Es gibt sehr viele mögliche Lösungen, mit denen bestimmte Abstände durch Punkte in einem Koordinatensystem dargestellt werden können. Es wird sich deshalb darauf festgelegt, dass die Lösung $n - 1$ Dimensionen haben soll, wenn n Items dargestellt werden sollen. Dann können

nämlich alle Abstände richtig abgebildet werden. Die Item-koordinaten haben die folgende Form, hier im Beispiel mit 4 Items und 3 Dimensionen:

Item A	a_1=0	a_2=0	a_3=0
Item B	b_1	b_2=0	b_3=0
Item C	c_1	c_2	c_3=0
Item D	d_1	d_2	d_3

Das erste Item wird immer als Punkt $(0, 0, ...)$ dargestellt. Für das zweite Item sind alle Koordinatenwerte, bis auf einen, auf 0 gesetzt. Jedes weitere Item hat einen weiteren Koordinatenpunkt ungleich 0. Für alle weiteren Items, die zu den vier im Beispiel hinzufügt werden, wird eine weitere Dimension hinzugefügt, deren Wert bei allen schon bestehenden Items 0 ist. Das neue Item $n + 1$ setzte sich zusammen aus n neuen Unbekannten ungleich 0. Für n Items haben wir also $(n-1)$ Dimensionen und $\sum_{i=0}^{n-1} i$ unbekannte Variablen. Da wir außerdem die Abstände von allen Items zueinander wissen, haben wir auch $\sum_{i=0}^{n-1} i$ Abstände gegeben. Wir können die Abstände nun in einem quadratischen Gleichungssystem der folgenden Form darstellen:

$$\sqrt{\sum_{i}^{n} (a_1 - b_1)^2} = d(A, B)$$

$$\sum_{i}^{n} (a_i - b_i)^2 = d(A, B)^2$$

Somit haben wir nun ein Gleichungssystem mit gleich vielen Gleichungen wie Unbekannten und können dieses lösen. Es handelt sich allerdings nicht um lineare Gleichungssysteme, sondern um quadratische Gleichungssysteme. Mit dem R-Paket rootSolve lassen sich derartige quadratische Gleichungssysteme lösen und man erhält als Lösung die Werte der Variablen. Diese wieder in die Koordinatenmatrix der obigen Form eingesetzt, ergibt die Koordinatenwerte für alle Punkte derartig, dass die euklidischen Abstände zweier Punkte den Distanzen der entsprechenden Items entsprechen. Auf diesen Punkten wird dann der K-Means Algorithmus angewendet. Dadurch, dass große Korrelationen zwischen den Items kleine Abstände zwischen zwei Items bedeuten, sind zwei Items mit hoher Korrelation wahrscheinlicher in einem Cluster als Items mit niedriger Korrelation.

4.3.3 Optimierung des Algorithmus

Da ein quadratisches Gleichungssystem gelöst werden muss, braucht das Lösen von Gleichungssystemen mit mehreren Unbekannten sehr viel Zeit. Eine Lösung kann allerdings leichter gefunden werden, wenn für jedes neue Item ein neues Gleichungssystem aufgestellt, das Ergebnis dann festgehalten wird und dann mit diesem festen Ergebnis erst der nächste Punkt als neues

Gleichungssystem hinzugefügt wird. Dies ist aufgrund der festen Form des Gleichungssystem möglich. Das Gleichungssystem für den ersten Punkt lautet beispielsweise:

$$b_1^2 = d(A, B)$$

Dann wird x_1 fest auf den sich ergebenden Wert gesetzt.
Das Gleichungssystem für das Hinzufügen des zweiten Punktes lautet dann:

$$\begin{aligned}(c_1 - b_1)^2 + c_2^2 &= d(B, C)\\ (c_1)^2 + c_2^2 &= d(A, C)\end{aligned}$$

Man beachte dabei, dass b_1 nun aber keine unbekannte Variable ist, sondern festgesetzt worden ist auf das Ergebnis des vorherigen Gleichungssystems. Auf diese Art und Weise stellt man nun für jeden hinzukommenden Punkt ein Gleichungssystem mit einer weiteren Variable und Gleichung auf. Kommt noch ein weiterer Punkt hinzu hat man nach der Form der Variablen aus der Tabelle folgendes Gleichungssystem zu lösen:

$$\begin{aligned}(d_1 - c1)^2 + (d_2 - c_2)^2 + d_3^2 &= d(C, D)\\ (d_1 - b_1)^2 + d_2^2 + d_3^2 &= d(B, D)\\ (d_1)^2 + d_2^2 + d_3^2 &= d(A, D)\end{aligned}$$

Dabei sind allerdings nur d_1, d_2 und d_3 Variablen, weil b_1,c_1,c_2 schon durch die vorherigen Gleichungssysteme gesetzt worden sind und $a_1, a_2, a_3, b_2, b_3, c_3$ nach Aufbau der Variablen in der Tabelle sowieso auf 0 gesetzt sind.
Mit insgesamt n Items hat das größte zu lösende Gleichungssystem also $n - 1$ Unbekannte und $n - 1$ Gleichungen. In dieser Form lässt sich das Gleichungssystem der Daten mit 39 Variablen in etwa 2 Minuten lösen.

4.4 K-Means Clustering der Items basierend auf Multidimensional Scaling

4.4.1 Allgemeine Idee

Dieses Verfahren basiert auf der gleichen Grundidee wie das vorherige Verfahren, nämlich die Items so als Punkte in einem Koordinatensystem darzustellen, dass die Abstände den gegebenen Abständen der Items entsprechen. Zur Berechnung wird hierbei allerdings das sogenannte Multidimensional Scaling verwendet. Das Multidimensional Scaling ist ein Verfahren, mit dem aus einer Distanzmatrix den Punkten in einer gegebenen Dimension Koordinaten zugewiesen werden, so dass die Abstände erhalten bleiben. Im Gegensatz zum obigen Verfahren kann hierbei also die Dimension spezifiziert werden und es wird eine möglichst gute Annäherung der Abstände in dieser Dimension erreicht. Das K-Means Clusterung der Items basierend auf Multidimensional

Scaling kann auch viel schneller berechnet werden als das Verfahren aus Kapitel 4.3. Bei der Herleitung sei auf [Borg und Groenen, 2005, 12.1] verwiesen.

4.4.2 Multidimensional Scaling

X bezeichnet die gesuchte Matrix der Koordinatenpunkte und D^2 die gegebene Matrix der Abstände der Punkte. X sei dabei so gewählt, dass die Matrix spaltenzentriert ist, die Summenwerte aller Spalten also 0 sind. Da gleiche Verschiebungen aller Punkte die Distanzen nicht verändern, stellt diese Annahme keine Einschränkung der darstellbaren Distanzen dar. Bevor nun erklärt wird, wie Multidimensional Scaling funktioniert, wird dafür etwas mathematische Vorarbeit geleistet.

Lemma 1. Wenn X spaltenzentriert ist, also die Spaltensummen alle 0 sind, so ist $B = XX^T$ spaltenzentriert und zeilenzentriert.

Beweis.

$$b_{ij} = \sum_{h}^{n} x_{ih}x_{jh}$$

Somit gilt für die Spaltensumme der $j - ten$ Spalte:

$$s_j = \sum_{i}^{n} b_{ij} = \sum_{i}^{n}\sum_{h}^{n} x_{ih}x_{jh} = \sum_{h}^{n}\sum_{i}^{n} x_{ih}x_{jh} = \sum_{h}^{n} x_{jh}\sum_{i}^{n} x_{ih} = \sum_{h}^{n} x_{jh}0 = 0$$

$\sum_i^n x_{ih}$ ist dabei immer 0, da die Spaltensummen von X immer 0 sind.
Für die Zeilensumme der $i - ten$ Zeile gilt:

$$s_i = \sum_{j}^{n} b_{ij} = \sum_{j}^{n}\sum_{h}^{n} x_{ih}x_{jh} = \sum_{h}^{n}\sum_{j}^{n} x_{ih}x_{jh} = \sum_{h}^{n} x_{ih}\sum_{j}^{n} x_{jh} = \sum_{h}^{n} x_{ih}0 = 0$$

$\sum_j^n x_{jh}$ ist dabei immer 0, da die Spaltensummen von X immer 0 sind.

□

Lemma 2. Die Distanzmatrix kann in Matrixschreibweise geschrieben werden als

$$D^2 = c1^T + 1c^T - 2XX^T$$

Dabei ist c der Vektor der Diagonalelemente von $B = XX^T$.

Beweis. Für das Element d_{ij} der Distanzmatrix gilt:

$$d_{ij} = \sum_{h}^{n} (x_{ih} - x_{jh})^2 = \sum_{h}^{n} x_{ih}^2 + \sum_{h}^{n} x_{jh}^2 - 2 * \sum_{h}^{n} (x_{ih}x_{jh})$$

Das Element b_{ij} der Matrix $B = XX^T$ hat den Wert $b_{ij} = \sum_h^n x_{ih}x_{jh}$. Für die Diagonalelemente c_{ii} von B gilt also $c_{ii} = \sum_h^n x_{ih}^2$ Entsprechend kann d_{ij} dargestellt werden als

$$d_{ij} = c_i + c_j - 2b_{ij}$$

In Matrixschreibweise kann die Distanzmatrix also folgendermaßen dargestellt werden:

$$D^2 = c1^T + 1c^T - 2XX^T$$

□

Lemma 3. Die Matrix $J = I_n - n^{-1}1_n1_n^T$ wird als Zentrierungsmatrix bezeichnet, da durch Multiplizierung mit ihr von links Matrizen spaltenzentriert werden und durch Multiplizierung mit ihr von rechts Matrizen zeilenzentriert werden.

Beweis.

$$JX = (I_n - n^{-1}1_n1_n^T)X = X - n^{-1}1_n1_n^TX = X - 1\bar{x}^T$$

Somit wird von jeder Variablen ihr Spaltenmittel subtrahiert und die herauskommende Matrix ist spaltenzentriert.

$$XJ = X(I_n - n^{-1}1_n1_n^T) = X - n^{-1}X1_n1_n^T$$

Analog wird bei der Multiplikation von rechts von jeder Variable ihr Zeilenmittel subtrahiert. Insbesondere gilt somit:

$$J1_n = (I_n - n^{-1}1_n1_n^T)1_n = 1_n - 1_n^- = 1_n - 1_n = 0$$

und

$$1_n{}^TJ = 1_n^T((I_n - n^{-1}1_n1_n^T)) = 1_n{}^T - n^{-1}1_n{}^T1_n1_n{}^T = 1_n - 1_n = 0$$

□

D^2 wird nun an beiden Seiten mit J multipliziert und zusätzlich noch mit dem Faktor $-1/2$

multipliziert. Dann ergibt sich:

$$\begin{aligned} -\frac{1}{2}JD^2J = \\ -\frac{1}{2}J(c1^T + 1c^T - 2XX^T)J = \\ -\frac{1}{2}Jc1^TJ - \frac{1}{2}J1c^TJ + J(2XX^T)J = \\ -\frac{1}{2}Jc0^T - \frac{1}{2}0c^TJ + J(XX^T)J = \\ J(XX^T)J = \\ XX^T \end{aligned} \tag{4.4.1}$$

Die ersten zwei Summanden sind dabei 0 wegen Lemma 3. Die Zentrierung von beiden Seiten verändert XX^T nicht, da XX^T nach Lemma 1 schon spalten- und zeilenzentriert ist. Auf diese Weise kann man also XX^T gegeben D^2 berechnen. Daraus kann man dann X durch Eigenwertszerlegung finden:

$$XX^T = Q\Lambda Q^T = (QA^{1/2})(QA^{1/2})^T$$

Dabei sind die Spalten von Q die Eigenvektoren von XX^T und in den Diagonalen von Λ stehen die Eigenwerte, der Rest sind $0en$. Folglich hat man die Koordinatenmatrix $X = QA^{1/2}$ gefunden. Es ist aber möglich, für $X = QA^{1/2}$ nur die q größten Eigenvektoren und Eigenwerte zu berücksichtigen. In diesem Fall hat Q die Dimensionen nxq und A die Dimensionen qxq. Somit ist dann X eine Matrix mit nxq Dimensionen, jeder der n-Punkte wird nun also durch q Koordinaten dargestellt. Entsprechend wird durch eine derartige Approximation auch XX^T und somit die Distanzmatrix auch nur approximiert. Damit werden die Punkte so in q Dimensionen verteilt, dass die Distanzen aber trotzdem noch approximativ erhalten bleiben.

5 Vergleich von Clusterungen

Es gibt mehrere Verfahren, mit denen die Ähnlichkeit von zwei verschiedenen Clusterungen bestimmt werden kann. Es sei hier vor allem auf [Wagner und Wagner, 2007] verwiesen. Im Folgenden wir aber nur auf eines dieser Verfahren eingegangen, den sogenannten Rand-Index, der in [Wagner und Wagner, 2007, Abschnitt 3.2.1] vorgestellt wird.

5.1 Paarmatrizen

Für die Berechnung der Clusterähnlichkeit wird zuerst einmal jede Clusterung ein-eindeutig in Matrixform dargestellt. Diese Matrix hat so viele Zeilen und Spalten wie es Items gibt, die geclustert werden sollen. Der Eintrag m_{ij} in der i-ten Zeile und in der j-ten Spalte gibt nun an, ob das Item i und das Item j zusammen in einem Cluster sind. Ist dieser Eintrag 0, so befinden sich Item i und j nicht in einem Cluster. Ist dieser Eintrag 1, so befinden sich die zwei Items in einem Cluster.

$$\begin{array}{c} \\ 1 \\ 2 \\ \\ 48 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & & 48 \end{array} \\ \begin{pmatrix} \text{N/A} & 0 & \cdots & 1 \\ 0 & \text{N/A} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & \text{N/A} \end{pmatrix} \end{array}$$

Diese Paarmatrix sagt nun beispielsweise aus, dass Item 1 und Item 2 nicht in einem Cluster sind, Item 1 und 48 aber schon in einem Cluster sind. Jede mögliche Clusterung kann eindeutig in diese Form überführt werden und aus jeder Paarmatrix lässt sich eindeutig die Clusterung bestimmen.

5.2 Vergleich von Paarmatrizen mit Rand-Index

Da wir nun die Clusterungen in Form von Paarmatrizen dargestellt haben, lassen sich nun zwei Clusterungen dadurch vergleichen, dass die Paarmatrizen miteinander verglichen werden. Dies

geschieht, in dem die Anzahl der übereinstimmenden Elemente der zwei Matrizen gezählt wird. Man zählt also $m_{F,ij} = m_{C,ij}$ mit $i \neq j$. m_F steht dabei für die Paarmatrix der sich aus der theoretischen Faktorstruktur ergebenden Clusterung, m_C für die sich aus dem Anwenden des jeweiligen Clusterverfahren auf die Korrelationsmatrix der Faktorstruktur ergebende Paarmatrix. Die Ähnlichkeit zweier Paarmatrizen berechnet sich also aus:

$$\text{Proximität} = \frac{\#\text{ gleicher Elemente}}{\#\text{ aller Elemente}}$$

6 Vorgehen zum Vergleichen der Clusterverfahren

6.1 Faktorstrukturerkennung als Vergleichskriterium der Clusterverfahren

In Kapitel 2 wurde vorgestellt, wie eine theoretische Faktorstruktur erzeugt werden kann und wie daraus die Korrelationsmatrix berechnet werden kann. In den nachfolgenden Kapiteln 3 und 4 wurden dann verschiedene Clusterverfahren vorgestellt, die auf einer Korrelationsmatrix durchgeführt werden können. Es ist nun Ziel dieser Arbeit, zu untersuchen, wie gut eine theoretische Faktorstruktur von verschiedenen Clusterverfahren erkannt wird. Die Faktorstruktur kann nach Kapitel 2.3 auch als Clusterung aufgefasst werden, indem die Items Clustern nach ihrer Hauptladung zugewiesen werden. Nun können die theoretische Clusterung einer Faktorstruktur nach Kapitel 2.3 und die sich ergebende Clusterung bei Anwenden des entsprechenden Clusterverfahrens auf die Korrelationsmatrix der Faktorstruktur miteinander verglichen werden, wobei die Ähnlichkeit zweier Clusterungen nach Kapitel 5 bestimmt wird. Für verschiedene theoretische Fakotrstrukturen, die sich in ihren Nebenladungen und Faktorkorrelationen unterscheiden, wird also für jedes Clusterverfahren die Ähnlichkeit der sich aus diesem Clusterverfahren ergebenden Clusterungen mit der theoretischen Faktorstrukturclusterung mithilfe des Rand-index berechnet. Das Clusterverfahren mit dem größten Rand-Index ist dann am besten dafür geeignet, theoretische Faktorstrukturen in der Clusterung wiederzufinden.

6.2 Strukturerhaltung als weiteres Vergleichskriterium der Clusterverfahren

Ein weiteres Qualitätsmerkmal für Clusterverfahren, das in dieser Arbeit auch verwendet wird, ist die Strukturerhaltung. Die Idee dahinter ist, dass man feststellen will, wie ähnlich sich für die verschiedenen Clustermethoden die Clusterung des Gesamtdatensatzes und die Clusterung eines Stichprobendatensatzes sind. Für den Stichprobendatensatz wurden die Stichproben aus den Personen, die die Items beantwortet haben genommen. Deshalb haben wir im Stichprobendatensatz genau so viele Items wie im Gesamtdatensatz. Allerdings sind die Korrelationen der Items untereinander im Stichprobendatensatz etwas verändert gegenüber dem Gesamtdatensatz. Es wird also für verschiedene Stichprobengrößen jeweils mehrere Male ein Stichprobendatensatz

gezogen und jede Clusterung dieses Stichprobendatensätze wird dann mit der Clusterung des Gesamtdatensatzes der gleichen Clustermethode verglichen. Die durchschnittliche Ähnlichkeit - bestimmt nach Kapitel 5 - der Gesamtdatensatzclusterung zu den Stichprobendatensatzclusterungn gilt dann als Maßzahl für die Strukturerhaltung des jeweils untersuchten Verfahren. Die Idee hinter diesem Qualitätsmerkmal ist, dass ein Clusterverfahren umso besser ist, desto stabiler die Clusterung erkannt wird, wenn nur eine Stichprobe der Personen berücksichtigt wird. Genaueres dazu kann man bei [Hölzl et al., 2013] nachlesen.

7 Einfluss der veränderten Faktorstruktur auf hierarchische Clusterungen

7.1 Theoretische Betrachtungen zu zunehmenden Nebenladungen

Für die Korrelation von zwei Variablen A und B, die durch eine Faktorstruktur gegeben sind, gilt nach 2.4.1 :

$$Cor(A,B) = \sum_{i}^{n} \sum_{j}^{n} a_i b_j Cor(F_i, F_j)$$

Dabei sei a_i die $i-te$ Nebenladung von A und analog a_j die $j-te$ Nebenladung von B. In diesem Kapitel werden wir nur betrachten, was passiert, wenn die Nebenladungen zunehmen, die Faktorkorrelationen zwischen verschiedenen Faktoren aber immer 0 sind. Folglich sind bis auf die Korrelationen zwischen den gleichen Faktoren, die 1 betragen, alle Faktorkorrelationen 0. Es gilt also in diesem Fall:

$$Cor(A,B) = \sum_{i}^{n} a_i b_i$$

Beachtet man nun zusätzlich, dass es Hauptladungen gibt und Nebenladungen, die wir alle konstant halten, so gibt es für die zwei Korrelationen zwei verschiedene Fälle. Der erste Fall ist, dass die zwei Variablen die gleiche Hauptladung haben. In diesem Fall gilt:

$$Cor(A,B) = HL_a HL_b + (n-1)NL^2$$

Dabei bezeichnet HL_1 den Wert der Hauptladung der ersten Variable, HL_2 den Wert der Hauptladung der zweiten Variable und NL den Wert der Nebenladungen. Haben die zwei Variablen verschiedene Nebenladungen, so gilt:

$$Cor(A,B) = HL_a NL + HL_b NL + (n-2)NL^2$$

Wie in 3 erklärt werden nun in der hierarchischen Clusterung nacheinander die Cluster mit der größten Ähnlichkeit zusammengefügt, bis alle Items sich in einem Cluster befinden. Betrachten wir im Folgenden erst einmal die Complete-Linkage-Clusterung, so werden immer diejenigen Cluster zusammengefasst, deren Elemente mit der größten Distanz sich am nächsten sind. Dann wird diejenige Clusterung aus dem Dendrogramm bestimmt, bei der es genau 5 verschiedene Cluster gibt. Im Idealfall werden alle Items mit der gleichen Hauptladung dem gleichen Cluster

zugeordnet. Es soll nun untersucht werden, unter welchen Umständen dies nicht geschieht und was dabei genau passiert. Betrachtet man im Folgenden den Cluster C_1, der bisher nur aus Elementen der Hauptladung HL_1 besteht, sowie den Cluster $C2$, bisher nur mit Elementen der Hauptladung HL_2. Für den Abstand zwischen $C1$ und $C2$ gilt dann bei Verwendung des Complete-Abstandsmaßes: $d(C1, C2) = 1 - (HL_1NL + HL_2NL + (n-2)NL^2)$. Dabei sind HL_1 und $HL2$ die Hauptladungen aus $C1$ bzw. $C2$, so dass der Abstand zwischen den zwei Elementen maximal ist. Als Distanzmaß wurde hier $d(A, B) = 1 - Cor(A, B)$ verwendet. Da aber beide Distanzmaße monoton sind, gelten die Aussagen für beide Distanzmaße.
Desweiteren betrachten wir den Abstand des Elementes $E3$ mit Hauptladung an der Stelle 3 zu den Cluster der bisher anderen geclusterten Elemente mit Hauptladung 3. Der Abstand berechnet sich folgendermaßen: $d(E, C3) = 1 - ((n-1)NL^2 + HL_eHL_3)$. Dabei bezeichnet HL_e die Hauptladung des Items e und HL_3 die Hauptladung des Elementes aus $C3$ mit der größten Distanz zum Item e.
Wir betrachten nun, wann die zwei Cluster $C1$ und $C2$ zusammengeclustert werden, bevor das Item e zu seinem Cluster hinzukommt:

$$1 - ((n-1)NL^2 + HL_eHL_3) > 1 - (HL_1NL + HL_2NL + (n-2)NL^2)$$

$$(n-1)NL^2 + HL_eHL_3 < HL_1NL + HL_2NL + (n-2)NL^2$$

$$NL^2 + (NL + x_e)(NL + x_3) < (NL + x_1)NL + (NL + x_2)NL$$

$$NL^2 + NL^2 + x_eNL + NLx_3 + x_1x_3 < NL^2 + x_1NL + NL^2 + x_2NL$$

$$NLx_3 + x_ex_3 < x_2NL$$

$$x_3(x_e + NL) < x_2NL$$

Für Elemente mit besonders kleiner Hauptladung $HL = NL + x_e$ kann es also geschehen, dass in den hierarchischen Verfahren zuerst zwei Cluster mit verschiedenen Hauptladungen fusioniert werden, bevor dieses Element mit der kleinen Hauptladung hinzukommt. Da allerdings die Clusteranzahl auf 5 festgesetzt ist, werden bei immer größer werdenden Nebenladungen entsprechend die Items mit großen aber verschiedenen Hauptladungen zusammengeclustert. Dafür bilden dann die Items mit sehr kleinen Hauptladungen extra Cluster. Dieses Verhalten kann man in Grafik 7.1 beobachten. Der erste Cluster von links im ersten Dendrogram werde mit $C1$ bezeichnet, der zweite mit $C2$, der dritte mit $C3$, etc. Im zweiten Dendrogram mit einer Nebenladung von 0.2 sieht man nun, dass $C4$ und $C5$ schon zusammenfusioniert wurden, bevor das Item $V19$ zu dem Cluster mit der gleichen Hauptladung hinzufusioniert wurde. Dieses Item bildet nun einen eigenen Cluster, da die Anzahl der Cluster auf 5 festgelegt ist.
Im dritten Dendrogram mit Nebenladung 0.25 wird nun auch $C1$ zu $C4$ und $C5$ fusioniert, bevor das Item 139 zu einem Cluster hinzugefügt wird und es gibt einen neuen aus einem Element bestehenden Cluster. Im dritten Dendrogramm gibt es dann noch einen großen Cluster und

vier kleinere Cluster die jeweils aus den Items mit sehr kleinen Hauptladungen bestehen. Es ist zu beachten, dass im Großen und Ganzen die ursprüngliche Clusterstruktur noch recht gut zu erkennen ist, die tatsächlich herauskommende Clusterung allerdings sehr verschieden ist von der sich ursprünglich ergebenden Clusterung.
Die obige Formel gilt allerdings nur für das Verfahren mit Complete-Linkage und Korrelation als Ähnlichkeitsmaß. Für Korrelation als Ähnlichkeitsmaß und Average-Linkage gilt aber analog:

$$\frac{1}{z}\sum_{h}^{z}(n-1)NL^2 + HL_{eh}HL_3 < \frac{1}{nm}\sum_{l}^{n}\sum_{h}^{m}HL_{1l}NL + HL_{2h}NL + (n-2)NL^2$$

Dabei wird die Ähnlichkeit nach dem Average-Linkage-Verfahren zwischen zwei Clustern berechnet. HL_{il} bezeichnet dabei die Hauptladung der $l-ten$ Variable mit Hauptladung im Cluster i. Dies kann in Grafik 7.2 beobachtet werden. Hier kann ein ähnliches Verhalten beobachtet werden, nämlich dass bei sich erhöhender Nebenladung das Clusterverfahren dazu neigt, alle Cluster zusammenzufügen und extra Cluster aus einzelnen Elementen mit niedrigen Hauptladungen zu erstellen. Wie man in den Grafiken 7.3 und 7.4 gut erkennen kann, tritt dies auch bei der Korrelation der Korrelation auf, allerdings etwas schwächer. Es fällt auf, dass die Abtrennung zwischen den Clustern bei der Korrelation der Korrelation als Ähnlichkeitsmaß eindeutiger ist und deshalb auch bei größeren Faktornebenladungen die Cluster noch besser erkannt werden.

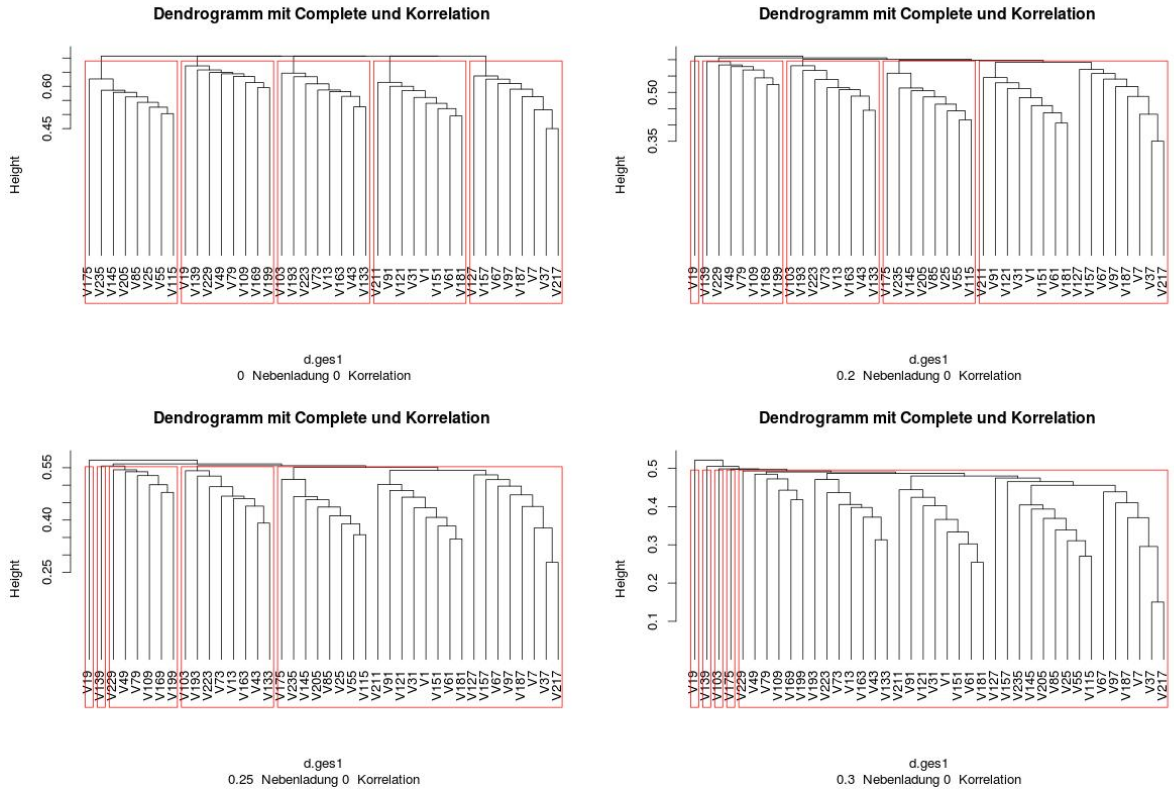


Abbildung 7.1: Dendrogramme mit zunehmender Nebenladung, Complete-Linkage und Korrelation als Ähnlichkeitsmaß

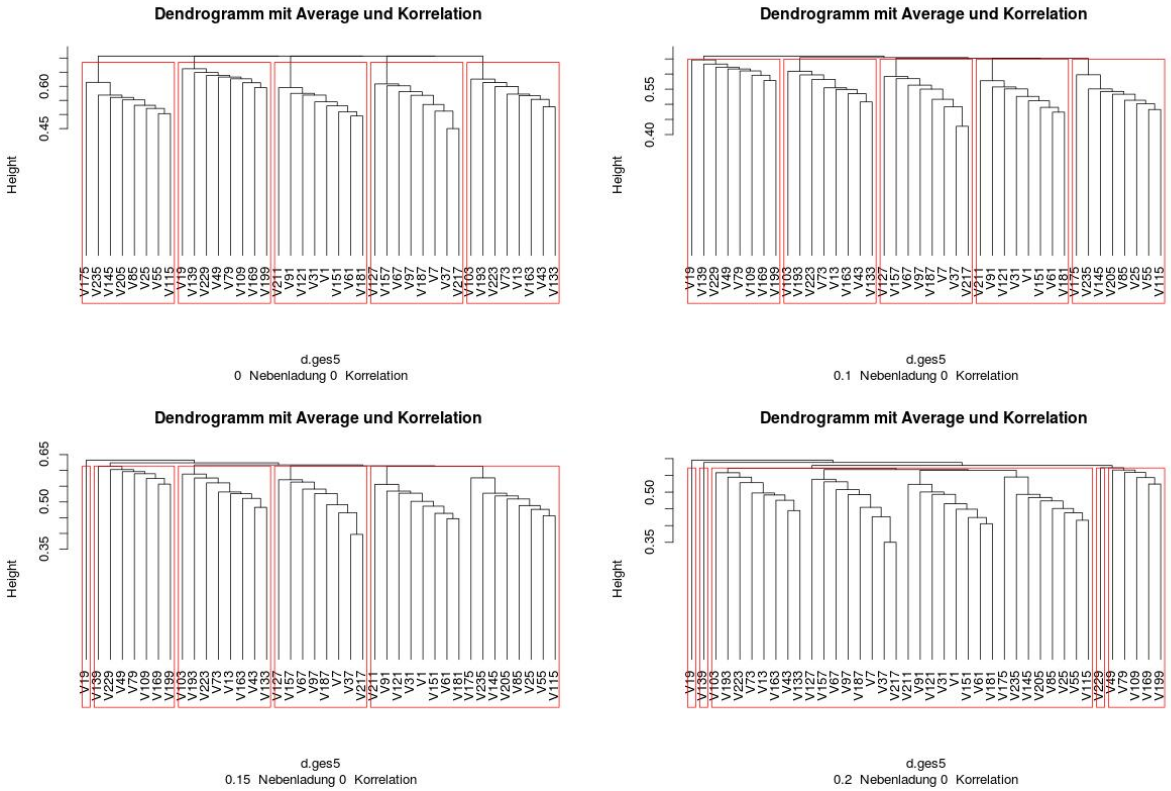


Abbildung 7.2: Dendrogramme mit zunehmender Nebenladung, Average-Linkage und Korrelation als Ähnlichkeitsmaß

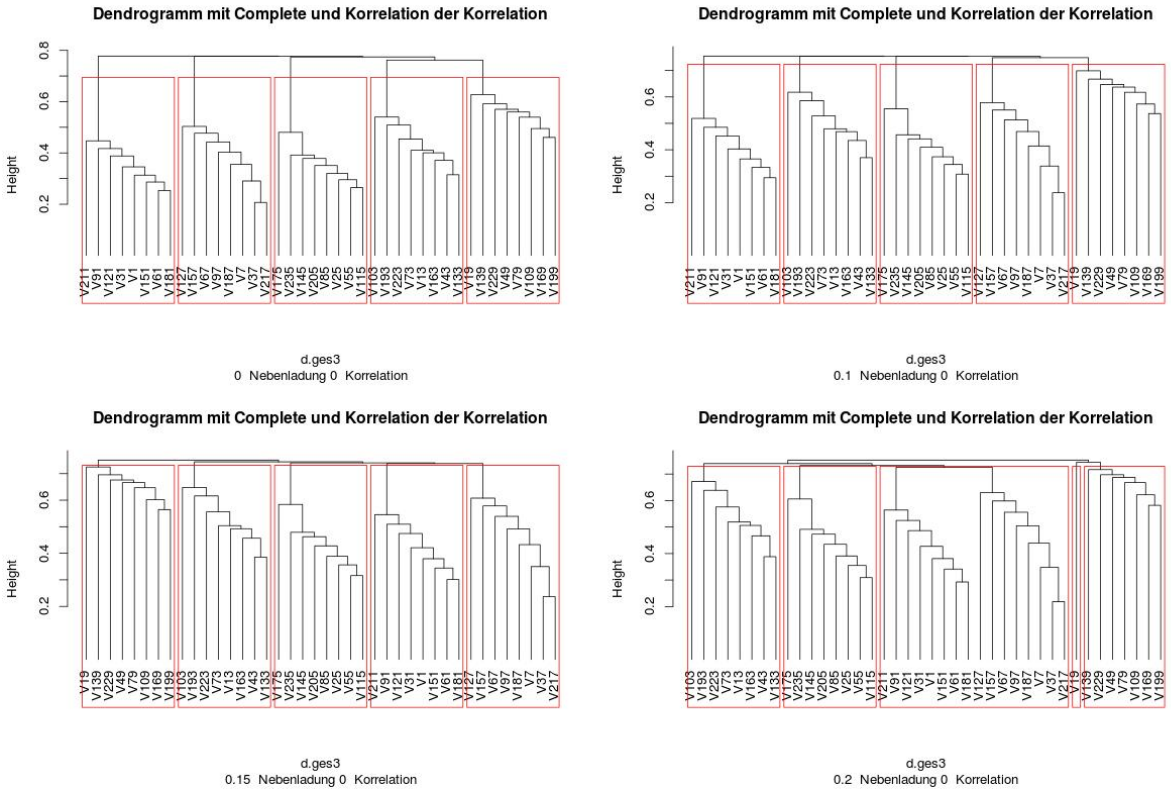


Abbildung 7.3: Dendrogramme mit zunehmender Nebenladung, Complete als Linkage-Verfahren und Korrelation der Korrelation als Ähnlichkeitsmaß

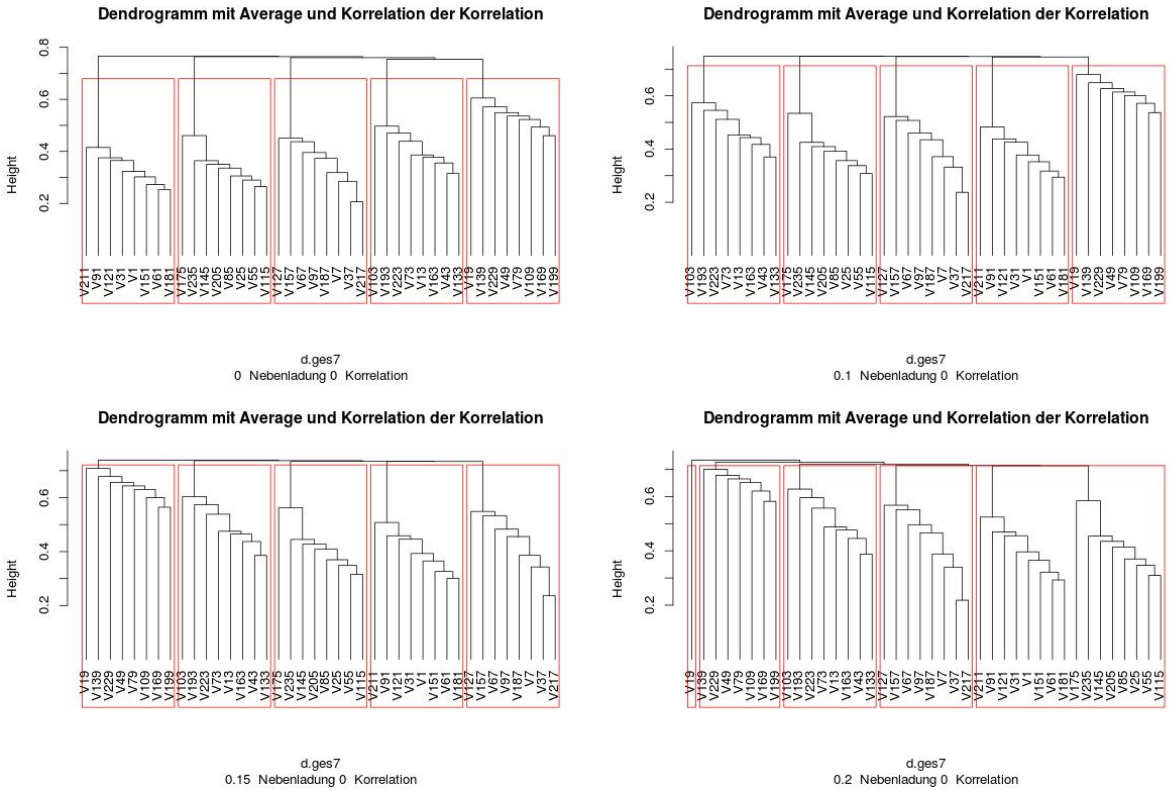


Abbildung 7.4: Dendrogramm mit zunehmender Nebenladung, Average-Linkage und Korrelation der Korrelation als Ähnlichkeitsmaß

7.2 Theoretische Betrachtungen zu zunehmenden Faktorkorrelationen

Analog zu 7.1 kann man sich nun auch anschauen, was passiert, wenn die Nebenladungen alle 0 sind und sich die Faktorkorrelationen immer weiter erhöhen. Für die Korrelation zwischen zwei Variablen gilt dann:

$$Cor(A,B) = \sum_{i}^{n}\sum_{j}^{n} a_i b_j Cor(F_i, F_j) = HL_a HL_b Cor(F_a, F_b)$$

Wenn dabei A und B die gleiche Hauptladung haben, so ist $Cor(F_a, F_b)$ die Faktorkorrellation eines Faktors mit sich selber und es gilt:

$$Cor(A,B) = \sum_{i}^{n}\sum_{j}^{n} a_i b_j Cor(F_i, F_j) = HL_a HL_b$$

Betrachten wir nun also genauso wie in 7.1 die Cluster $C1$, $C2$ und $C3$ beim Complete-Linkage-Verfahren. Dabei sind HL_1 und HL_2 die Hauptladungen aus $C1$ bzw. $C2$, so dass der Abstand

zwischen den zwei Elementen maximal ist. HL_e bezeichnet die Hauptladung des Items e und HL_3 die Hauptladung des Elementes aus $C3$ mit der größten Distanz zum Item e. Dann werden Cluster mit zwei verschiedenen Hauptladungen zusammengeclustert, bevor das Item $e3$ zu dem Cluster mit der gleichen Hauptladungen hinzukommt, wenn gilt:

$$HL_1 HL_2 Cor(F_1, F_2) < HL_3 HL_e$$

Dies ist wieder der Fall für besonders kleine Hauptladungen HL_e und deshalb werden diese wieder bei zunehmenden Hauptladungen als extra Cluster geclustert. Dieses Verhalten kann man gut in 7.5 beobachten.

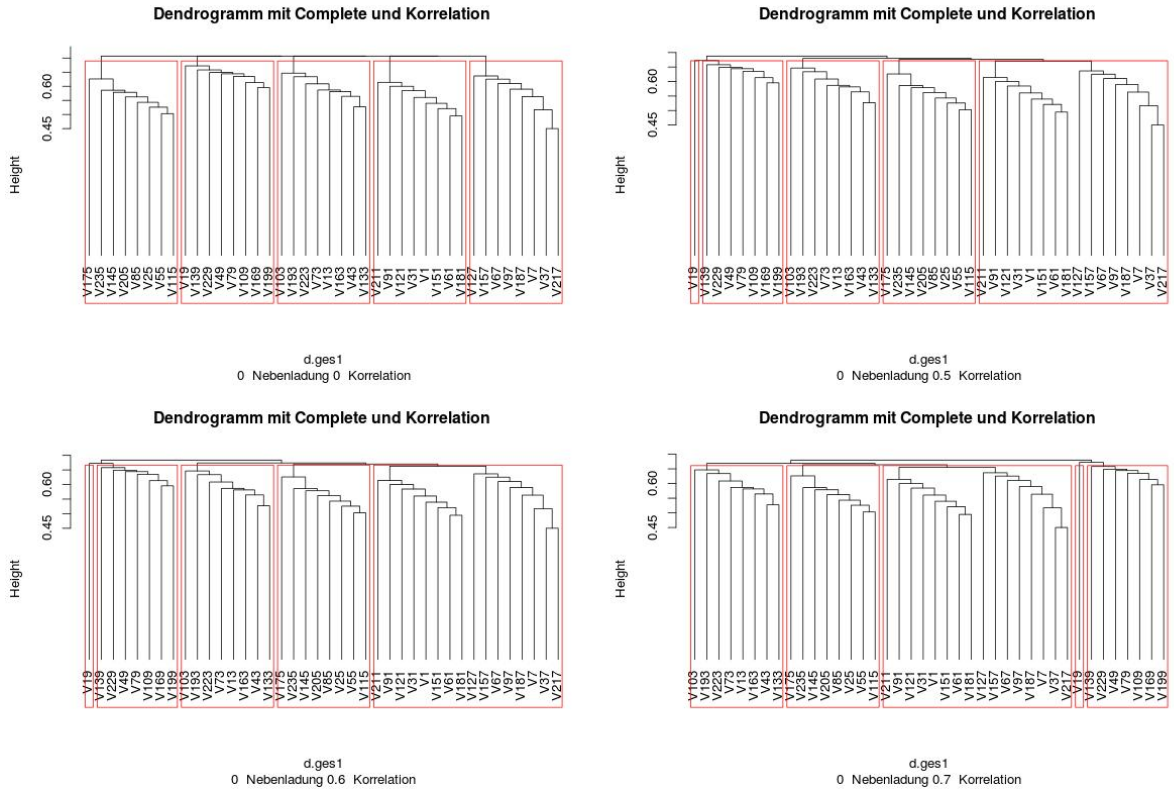


Abbildung 7.5: Dendrogramme mit zunehmender Korrelation, Complete-Lonkage und Korrelation als Ähnlichkeitsmaß

Für das Average-Linkage-Verfahren lautet die Formel analog:

$$\frac{1}{z}\sum_{h}^{z} HL_e HL_{h3} < \frac{1}{nm}\sum_{l}^{n}\sum_{h}^{m} HL_{l1} HL_{h2} Cor(F_1, F_2)$$

Auch beim Average-Linkage tritt dieses Verhalten auf, siehe 7.6

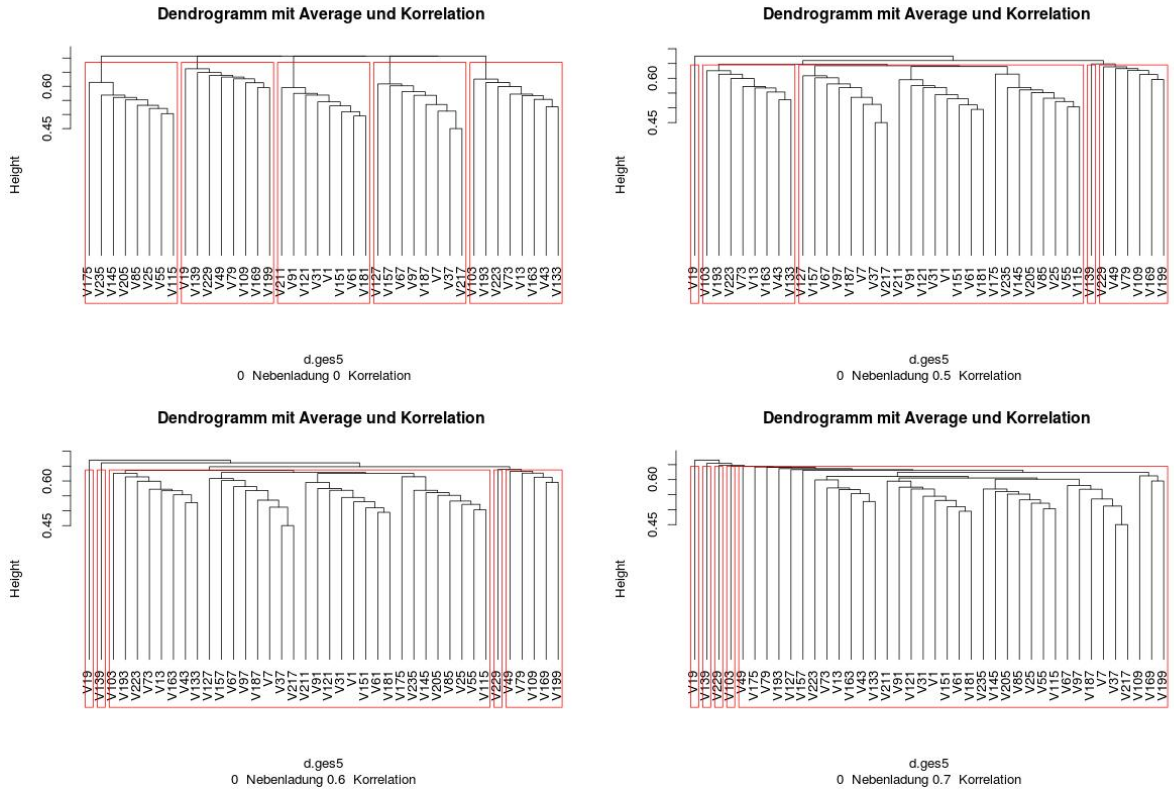


Abbildung 7.6: Dendrogramme mit zunehmender Korrelation, Average-Linkage und Korrelation als Ähnlichkeitsmaß

Bei der Korrelation der Korrelation kann man dieses Verhalten auch beobachten, siehe 7.7 und 7.8

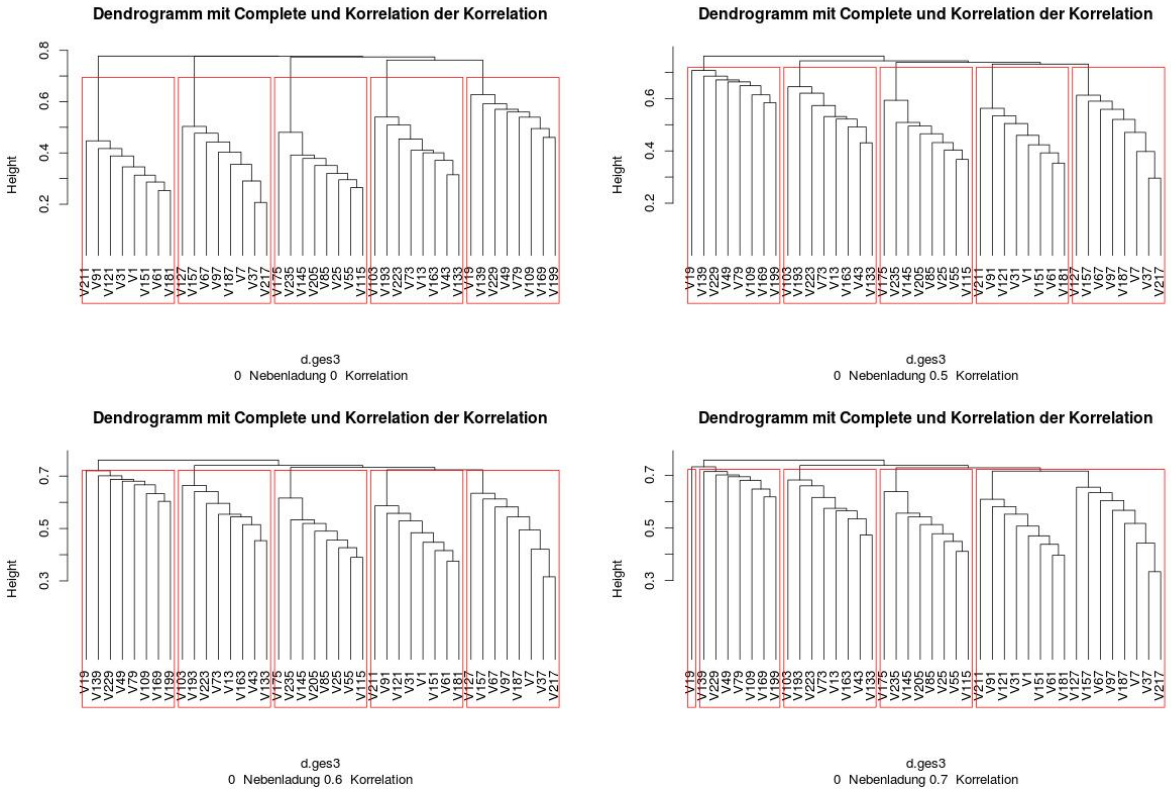


Abbildung 7.7: Dendrogramme mit zunehmender Korrelation, Complete-Linkage und Korrelation der Korrelation als Ähnlichkeitsmaß

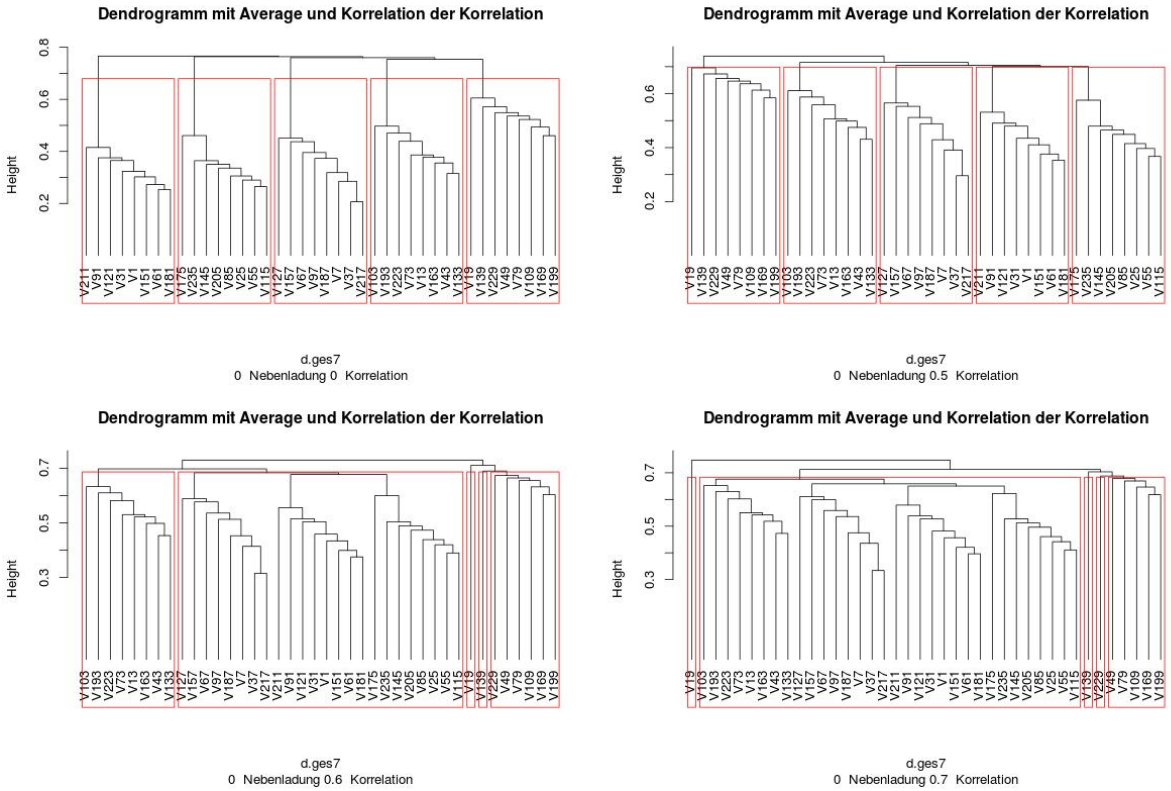


Abbildung 7.8: Dendrogramme mit zunehmender Korrelation, Average-Linkage und Korrelation der Korrelation als Ähnlichkeitsmaß

8 Ergebnisse zum Auffinden von Faktorstrukturen der Clustermethoden

In diesem Kapitel werden nun die tatsächlichen Ergebnisse vorgestellt, welche Clusterverfahren sich am besten dazu eignen, die Clusterung der zugrunde liegenden Faktorstruktur wiederzuerkennen, wenn sich auch die Nebenladungen und Faktorkorrelationen verändern. Im Folgenden bezieht sich der Begriff K-Means-MDS auf das in Kapitel 4.4, K-Means-Koord auf das in Kapitel 4.3 und K-Means-Kor auf das in Kapitel 4.2 vorgestellte Verfahren. Erst im letzten Abschnitt dieses Kapitels wird wie in 2.4.2 erklärt auch die Fehlerkorrelationsmatrix hinzuaddiert, in den anderen Kapiteln wird nur die Korrelationsmatrix der theoretischen Faktorstruktur betrachtet. Es ist zu beachten, dass die niedrigste Hauptladung in den untersuchten Daten 0.276 beträgt. Die maximale untersuchte Nebenladung beträgt allerdings 0.3. Zwar kann man dann nicht mehr von Nebenladung sprechen, da diese Nebenladung höher ist als die geringste vorhandene Hauptladung, diese Untersuchung wurde allerdings der Vollständigkeit halber trotzdem in der Arbeit gelassen. In den Grafiken bezeichnet averagecor das hierarchische Verfahren mit Average-Linkage und Korrelation als Ähnlichkeitsmaß, averagecorcor beschreibt das hierarchische Clusterverfahren mit Average-Linkage und Korrelation der Korrelation als Ähnlichkeitsmaß. Analog bezieht sich completecor auf Complete-Linkage mit Korrelations als Ähnlichkeitsmaß und completecorcor auf Complete-Linkage mit Korrelation der Korrelation als Ähnlichkeitsmaß. Das Verfahren K-Means-Koord wird in den Grafiken mit kmeanskoord bezeichnet, K-Means-MDS mit kmeansmds und K-Means-Kor mir kmeanscor.

8.1 Zunehmende Größe der Nebenladungen

In diesem Abschnitt werde untersucht, was passiert, wenn nur die Hauptladung aus dem Datensatz übernommen wird und die Nebenladung vergrößert wird. Die Ergebnisse sind visualisiert in den Grafiken 8.1 und 8.2.

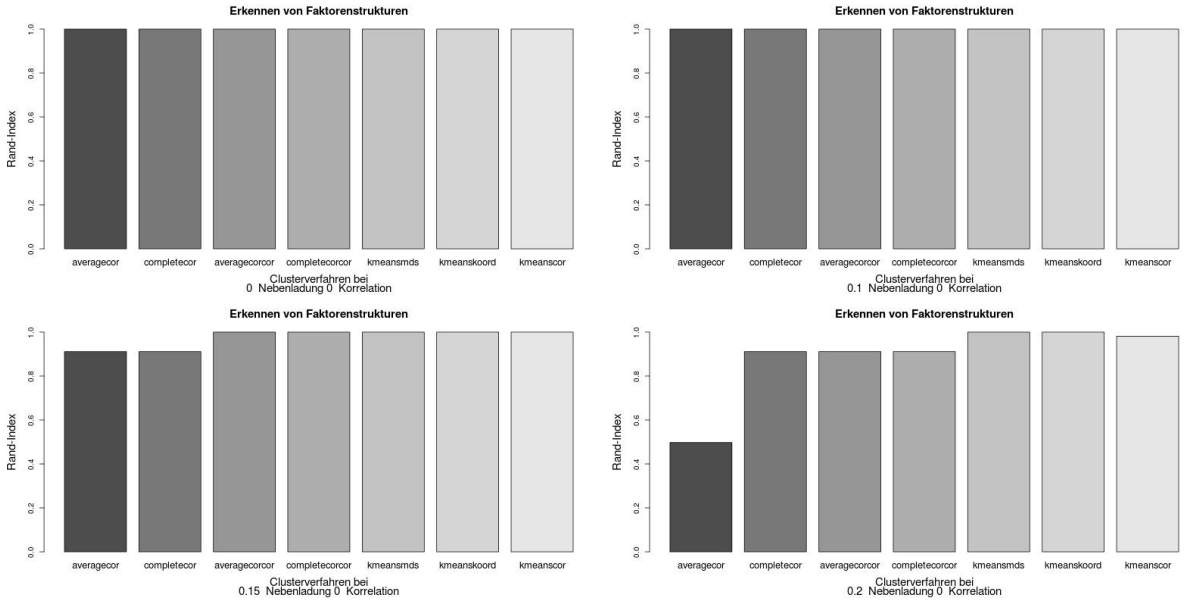


Abbildung 8.1: Vergleich der Clustermethoden bei zunehmenden Nebenladungen

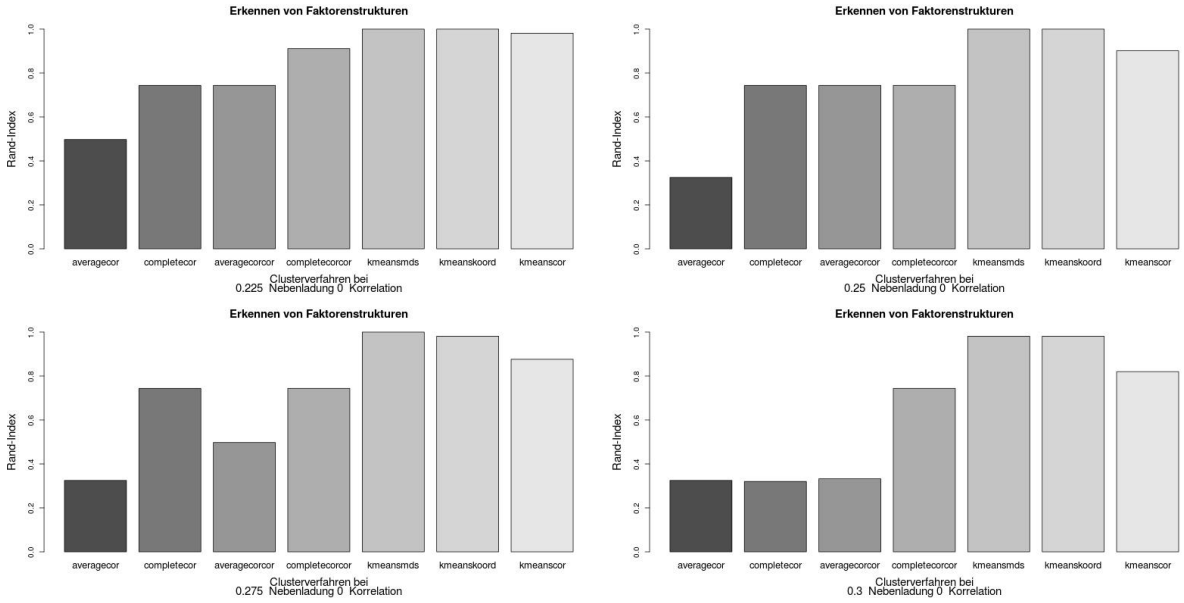


Abbildung 8.2: Vergleich der Clustermethoden bei noch größeren Nebenladungen

Bei einer Nebenladung von 0 erkennen alle Clustermethoden die Faktorstrukturen genau. Mit zunehmenden Nebenladungen werden die Faktorstrukturen immer schlechter erkannt. Bei den hierarchischen Methoden, die auf Average-Linkage basieren nimmt die Clusterähnlichkeit am schnellsten ab. Auch schneidet die Korrelation der Korrelation grundsätzlich besser ab als die Korrelation als Ähnlichkeitsmaß. Bei der höchsten betrachteten Nebenladung ist auch das Complete-Verfahren mit Korrelation der Korrelation als Abstandsmaß das beste Verfahren. Ge-

rade bei der größten Faktornebenladung ist allerdings ersichtlich, dass die auf K-Means basierenden Clustermethoden viel besser dafür geeignet sind, die Faktorstrukturen zu erkennen. Der Grund, warum die K-Means Methoden bei größeren Nebenladungen so viel besser abschneiden als die hierarchischen Methoden, scheint darin zu liegen, dass die hierarchischen Methoden, wie im vorherigen Kapitel 7.1 dargelegt wurde, bei wachsender Nebenladung sehr schnell die Variablen mit den kleinsten Hauptladungen als eigene Cluster clustern. Deshalb haben die hierarchischen Verfahren öfter genau die gleiche Ähnlichkeit zum Mustercluster, weil beispielsweise alle so geclustert sind, dass sich alle Elemente bis auf die vier Elemente mit den kleinsten Nebenladungen in einem Cluster befinden. Die Methoden, die auf K-Means basieren, haben dieses Problem nicht. Bei den K-Means Methoden schneiden K-Means-MDS und K-Mean-Koord eindeutig besser ab als K-Means-Kor. Dies wird bei den größeren Nebenladungen ersichtlich.

8.2 Nebenladungen aus Normalverteilung

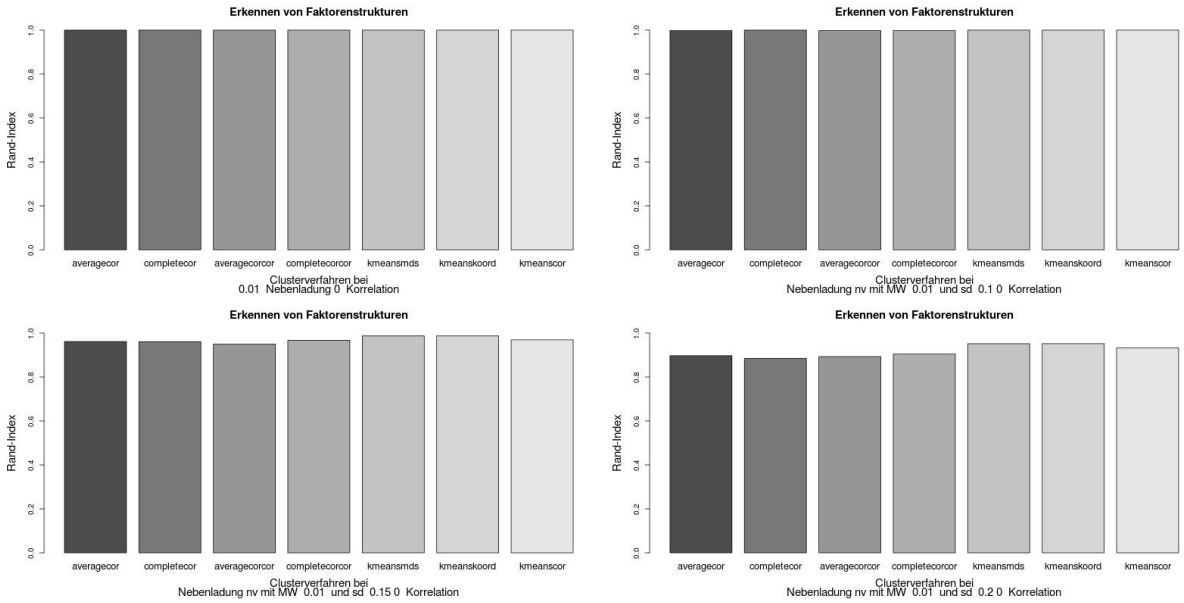


Abbildung 8.3: Vergleich der Clustermethoden bei normalverteilten Nebenladungen mit zunehmender Varianz

Nun wird auch wieder die Nebenladung variiert, allerdings werden die Nebenladungen nun nicht mehr immer nur größer, sondern die Nebenladungen entstammen einer Normalverteilung und die Varianz dieser wird immer größer gewählt. Die Ergebnisse sind in Grafik 8.3 abgebildet.
Es fällt auf, dass die Verfahren mit zunehmender Varianz der Normalverteilung schlechter werden. Allerdings sind die Unterschiede, um wie viel die verschiedenen Verfahren schlechter werden, eher gering. Die zwei Verfahren K-Means-MDS und K-Means-Koord schneiden aber etwas besser ab als die anderen Verfahren und K-Means-Kor liegt zwischen den hierarchischen Verfahren und

den anderen zwei auf K-Means basierten Verfahren. Eine höhere Nebenladung scheint also einen stärkeren Einfluss zu haben als variierende Nebenladungen.

8.3 Zunehmende Größe der Faktorenkorrelation

In diesem Abschnitt werden nun die Nebenladungen gleich gelassen und dafür die Faktorkorrelationen verändert. Die Ergebnisse werden in den Grafiken 8.4 und 8.5 gezeigt.

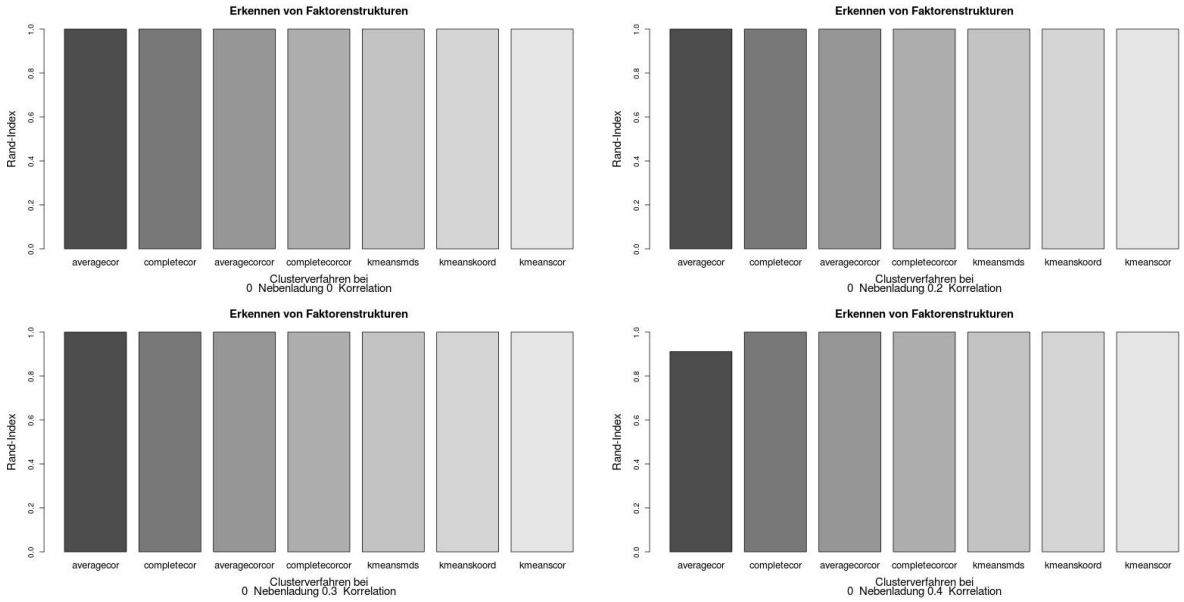


Abbildung 8.4: Faktorstrukturerkennung verschiedener Clustermethoden bei zunehmender Korrelation

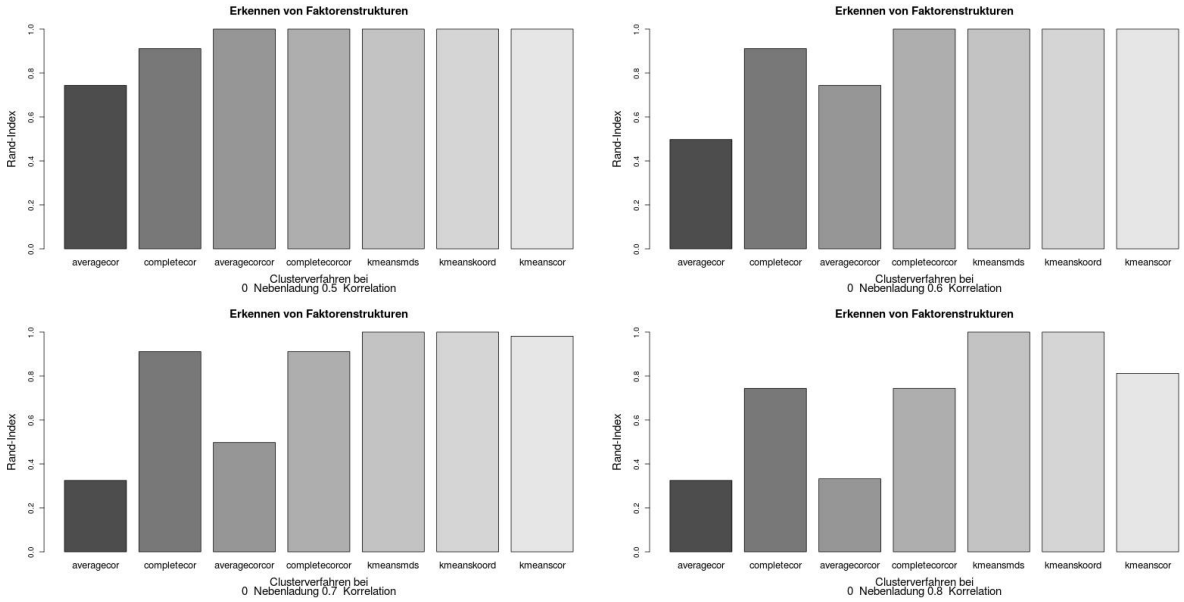


Abbildung 8.5: Faktorstrukturerkennung verschiedener Clustermethoden bei größerer zunehmender Korrelation

Bei den hierarchischen Verfahren zeigt sich zunächst bei den niedrigeren Faktorkorrelationen in Abbildung 8.4, dass eine Erhöhung der Faktorkorrelation von bis zu 0.4 bei allen Clustermethoden kaum eine Veränderung der Clusterung bewirkt. Die Erhöhung der Nebenladung ohne Veränderung der Faktorkorrelation scheint also einen stärkeren Einfluss zu haben. Bei den noch größeren Nebenladungen in Grafik 8.5 zeigt sich dann auch, dass die Complete-Verfahren besser abschneiden als die Average-Verfahren und die Korrelation der Korrelation als Abstandsmaß auch eine Verbesserung mit sich bringt. Besser als die hierarchischen Methoden schneiden allerdings die K-Means Methoden ab, vor allem K-Means-Koord und K-Means-MDS. Der Grund dafür liegt wieder darin, dass, wie in 7.2 dargelegt, die hierarchischen Clusterungen dazu neigen, Cluster für einzelne Items mit sehr geringen Hauptladungen zu erzeugen.

8.4 Faktorenkorrelation aus Normalverteilung

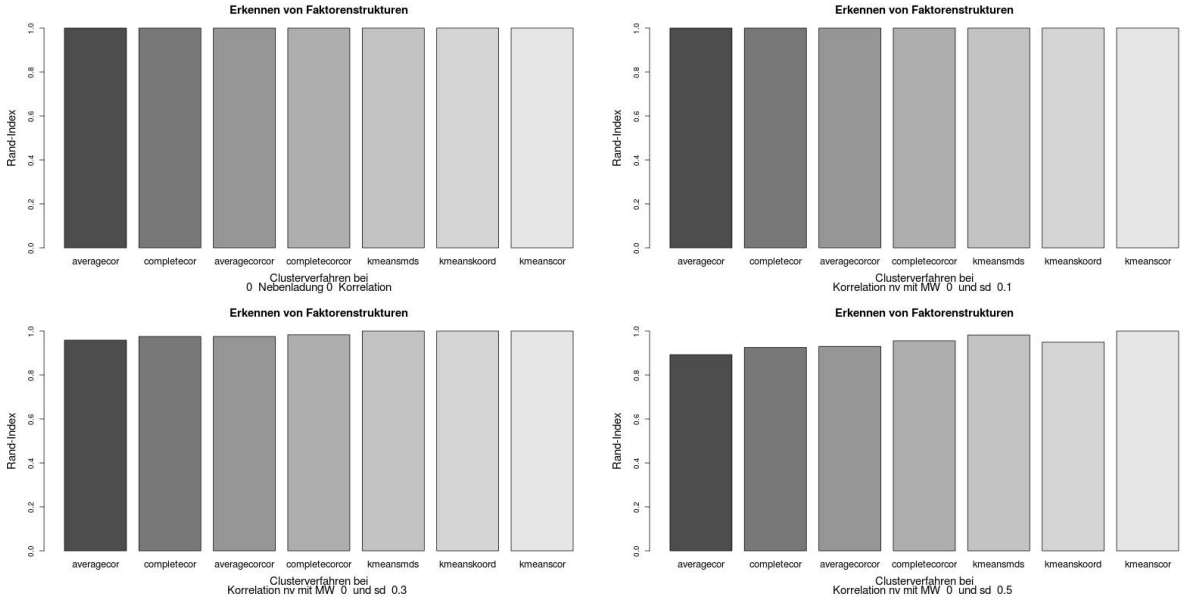


Abbildung 8.6: verschiedene Clustermethoden bei Faktorkorrelation aus NV

Grafik 8.6 zeigt, was passiert, wenn die Faktorkorrelation normalverteilt mit zunehmender Varianz ist und die Nebenladungen alle 0 bleiben. Auch hier schneidet bei den hierarchischen Methoden die Korrelation der Korrelation etwas besser ab als Ähnlichkeitsmaß und das Complete-Linkage-Verfahren ist immer minimal besser als das Average-Linkage-Verfahren. Von den K-Means-basierten Methoden sind K-Means-MDs und K-Means-Kor etwas besser und K-Means-Koord in etwa gleich gut wie die hierarchischen Verfahren, allerdings sind die Unterschiede zwischen den Verfahren alle eher gering.

8.5 Zunehmende Nebenladungen und Faktorenkorrelationen

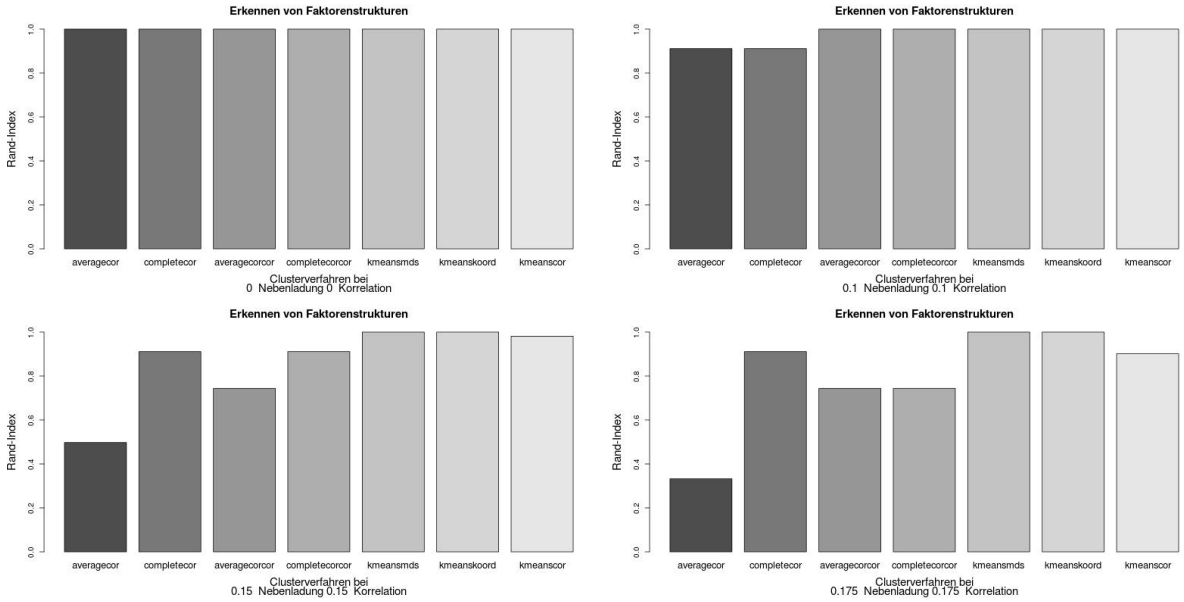


Abbildung 8.7: verschiedene Clustermethoden bei zunehmender Korrelation und Nebenladung

In Grafik 8.7 ist zu sehen, wie die Faktorstrukturen wiedererkannt werden, wenn sowohl die Nebenladung als auch die Faktorkorrelation erhöht werden. Im Gegensatz zu den vorherigen Ergebnissen, bei denen die besten hierarchischen Verfahren Complete-Linkage und Korrelation der Korrelation war, ist hier Complete-Linkage und Korrelation bei der größten Nebenladung und Faktorkorrelation das beste Verfahren. Bei kleineren Nebenladungen und Faktorkorrelationen schneidet allerdings die Korrelation der Korrelation als Abstandsmaß doch besser ab und Complete-Linkage schneidet besser ab als Average-Linkage. Ähnlich wie bei den Ergebnissen vorher sind auch hier die Verfahren mit K-Means-Koord und K-Means-MDS die besten Verfahren, während K-Means-Kor etwas schlechter ist als diese Verfahren.

8.6 Ergebnisse beim Hinzuaddieren der Fehlerkorrelationsmatrix

Hier wird betrachtet, was passiert, wenn auch der Fehlerkorrelationsterm hinzuaddiert wird. In 8.8 kann man sehen, wie gut verschiedenen Clustermethoden bei zunehmenden Nebenladungen die Faktorstruktur erkennen, wenn die Fehlerkorrelation nicht hinzuaddiert wird. In 8.9 kann das gleiche beobachtet werden, allerdings wird diesmal die Fehlerkorrelationsmatrix hinzuaddiert. Die wichtigsten Ergebnisse der vorigen Abschnitte, nämlich dass bei den hierarchischen Verfahren das Complete-Linkage sinnvoll ist und Korrelation der Korrelation oft besser ist als die Korrelation, aber die K-Means basierten Methoden insgesamt besser geeignet sind, bleiben auch bei Dazuaddieren der Fehlerkorrelationsmatrix erhalten. Die hierarchischen Verfahren schneiden bei Hinzuaddieren der Fehlerkorrelationsmatrix manchmal besser ab, manchmal schlechter. So

schneidet beispielsweise das Verfahren Complete mit Korrelation als Ähnlichkeitsmaß bei einer Nebenladung von 0.275 in 8.9 mit Fehlerkorrelationsmatrix schlechter ab als in 8.8 ohne Fehlerkorrelationsmatrix, aber bei einer Nebenladung von 0.3 ist es umgekehrt. Dies könnte daran liegen, dass sich durch die Addition eines Fehlerterms, der nicht durch das Modell erklärt ist, die Korrelationen zufällig verändern und dies manchmal sich auch positiv auf die Erkennung der Faktorstruktur auswirken kann, die sonst wegen der schon recht großen Nebenladungen nicht mehr erkannt worden wäre. Die bisher immer sehr guten Verfahren K-Means-Koord und K-Means-MDS werden aber bei der größten Nebenladung bei Hinzuaddieren der Fehlerkorrelationsmatrix schlechter. Durch das Hinzuaddieren eines zufälligen Fehlerterms, der nicht durch das Modell erklärt ist, werden sich die Verfahren im Bezug auf das Qualitätsmerkmal der Faktorstrukturerkennung also ähnlicher. Dies scheint sinnvoll, da bei rein zufälligen Korrelationen alle Clusterverfahren gleich gut sein müssten. Hier in der Arbeit sind nur die Ergebnisse für größere zunehmende Nebenladungen abgebildet, die anderen Ergebnisse können aber im Ordner ersteDaten/faktorstrukur/allsmall/ betrachtet werden, der Name der Grafiken, bei den die Fehlerkorrelationsmatrix hinzuaddiert wurden, endet auf Fehler.

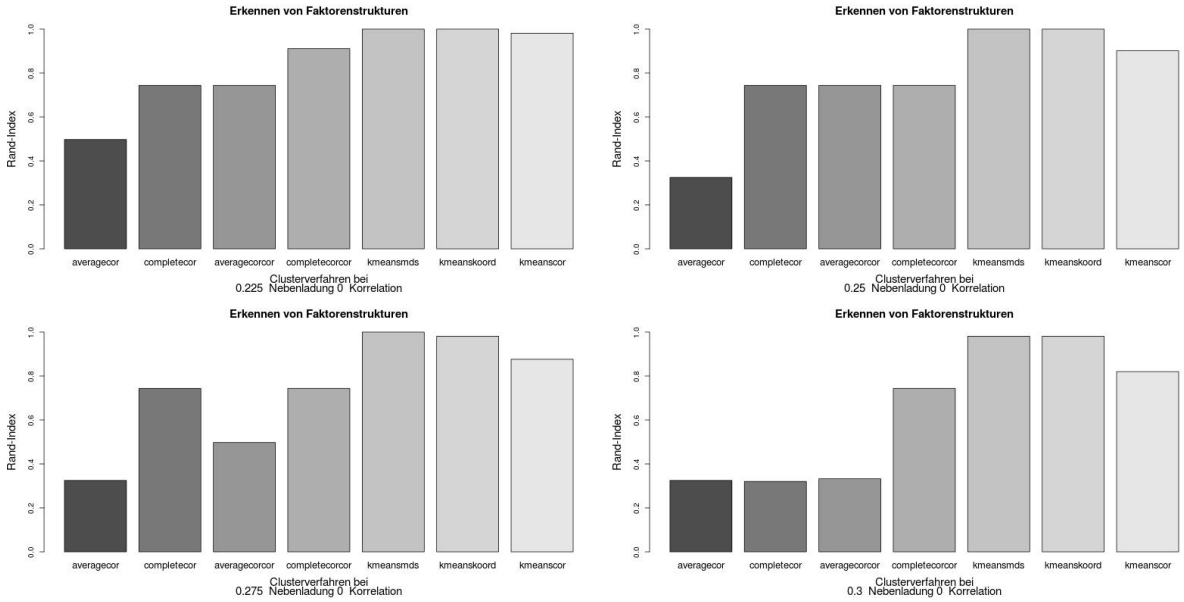


Abbildung 8.8: verschiedene Clustermethoden bei größerer zunehmenden Nebenladungen ohne Fehlerkorrelation

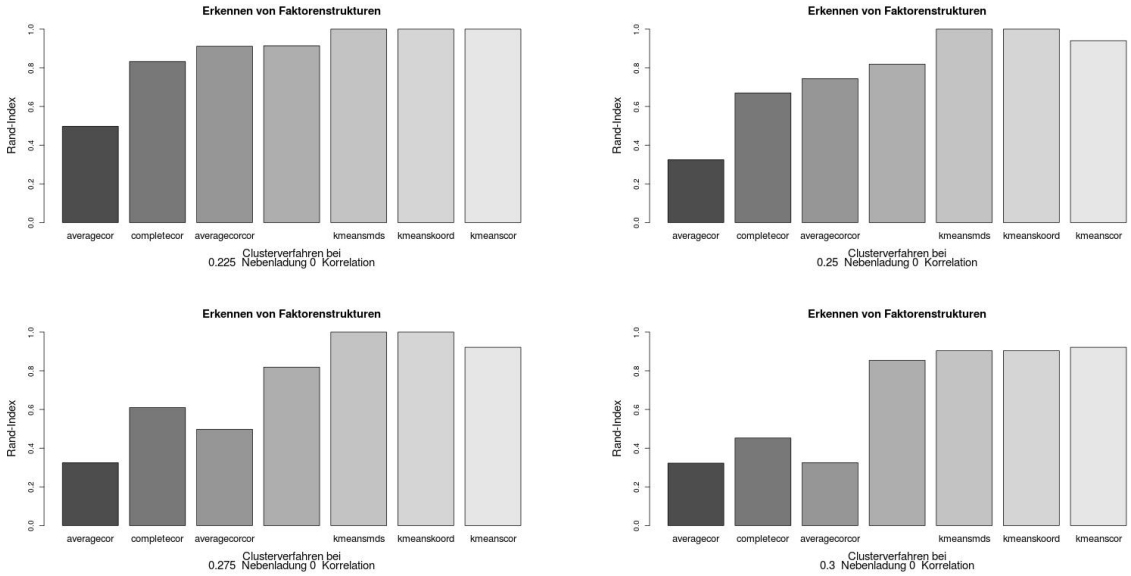


Abbildung 8.9: verschiedene Clustermethoden bei größerer zunehmenden Nebenladungen mit Fehlerkorrelation

9 Ergebnisse zur Strukturerhaltung der Clustermethoden

Zusätzlich zu diesen Untersuchungen zur Erkennung von Faktorstrukturen scheint es auch sinnvoll zu sein, die neu vorgestellten Clusterverfahren auch darauf zu testen, wie gut sie Clusterstrukturen in einem Stichprobendatensatz wiederfinden (siehe Kapitel 6.2)

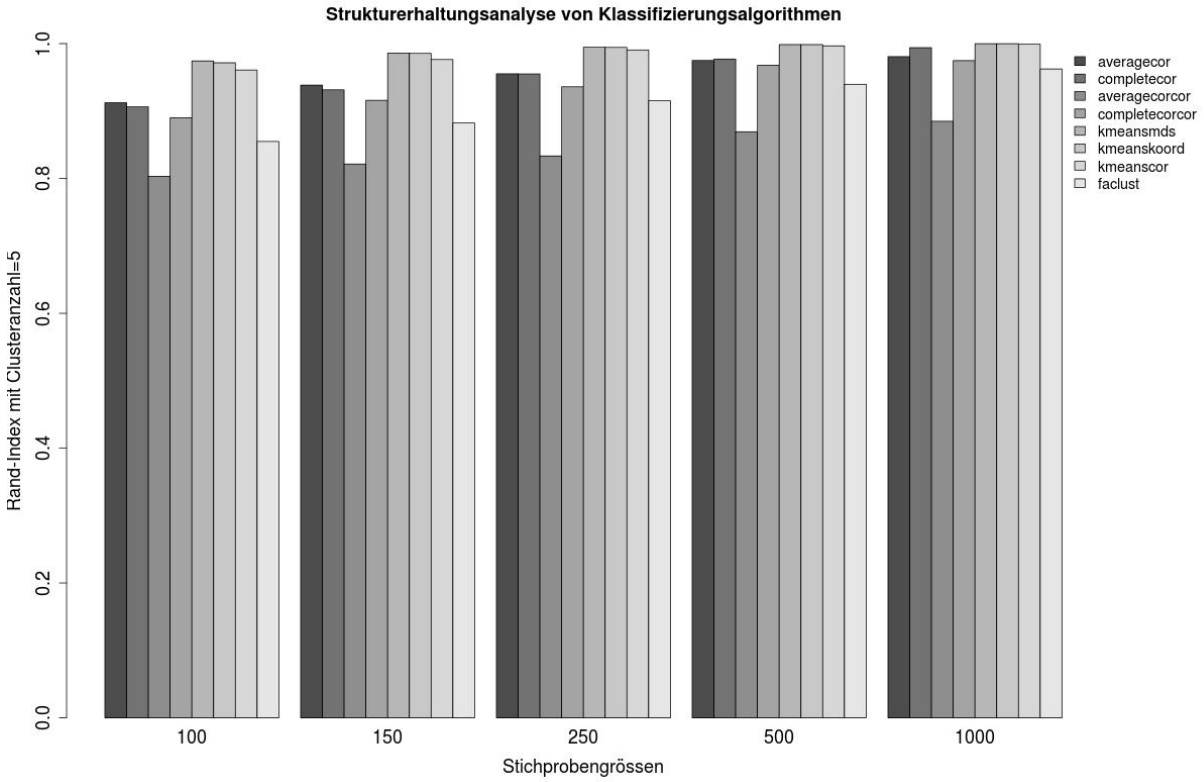


Abbildung 9.1: Strukturerhaltung der verschiedenen Clustermethoden

Bei den hierarchischen Verfahren schneidet das Verfahren mit Average-Linkage und Korrelation der Korrelation schlechter ab als die anderen Verfahren und ansonsten schneidet das Complete-Clustering etwas besser ab als das Average Clustering. Besser als diese alle Verfahren schneiden die zwei Verfahren K-Means-MDS und K-Means-Koord ab. Das K-Means-Kor Verfahren liegt ungefähr auf dem gleichen Niveau aber etwas darunter. Auch für die Strukturerhaltung scheint also das Anwenden des K-Means-Clustering sinnvoll zu sein, bei kleineren Stichprobengrößen

ist der Unterschied zu den hierarchischen Verfahren auch noch größer als bei größeren Stichprobengrößen. Zusätzlich zu den bereits bekannten Verfahren wurde in der Strukturerhaltung auch das in der Grafik als faclust bezeichnete Verfahren untersucht. Dieses Verfahren entspricht dem Verfahren, bei dem zuerst eine Faktoranalyse auf den Daten durchgeführt wird und dann die Items dem Cluster mit der größten Hautpladung zugewiesen werden. Dieses Verfahren schneidet bei der Strukturerhaltung aber nicht besser ab als die hierarchischen Verfahren.

10 Sonstige Ergebnisse

10.1 Abstandsmaß bei den hierarchischen Verfahren

Wie in Kapitel 3 erklärt wurde, gibt es neben den zwei Abstandsmaßen, Korrelation und Korrelation der Korrelationen und den zwei Linkage Verfahren, Average- und Complete-Linkage, auch noch zwei Möglichkeiten, wie man aus dem Abstandsmaß das Distanzmaß berechnen kann. Zum einen lässt sich das Distanzmaß berechnen aus $d(A,B) = 1 - Korr(A,B)$. Dieses Abstandsmaß scheint zunächst sehr intuitiv zu sein, allerdings ist es, wie in 3 beschrieben, keine Metrik. Das Abstandsmaß $d(A,B) = \sqrt{0.5 - 0.5s(A,b)}$ hingegen ist eine Metrik. Bei allen vorherigen Untersuchungen wurde für die hierarchischen Verfahren das Abstandsmaß $d(A,B) = \sqrt{0.5 - 0.5s(A,b)}$ verwendet, da für die Verfahren K-Means-Koord und K-Means-MDS nur dieses Abstandsmaß verwendet werden kann und die Verfahren mit gleichem Distanzmaß besser vergleichbar sind. In diesem Abschnitt soll allerdings untersucht werden, ob die Wahl des Distanzmaßes einen Unterschied macht für die Erkennung der Faktorstruktur und die Strukturerhaltung.

10.1.1 Faktorstruktur

Bei der Erkennung der Faktorstruktur macht die Wahl des Distanzmaßes in den von uns betrachteten Fallbeispielen der Erhöhung der Nebenladungen zunächst keinen Unterschied. Es sei hier nur die Grafik 10.1 für die Average-Verfahren für größere zunehmende Nebenladungen und 10.2 für die Complete-Verfahren ebenfalls bei größeren zunehmenden Nebenladungen aufgeführt, die restlichen Grafiken können aber im Ordner ersteDaten/faktorbasis/hierarchialaverage und ersteDaten/faktorbasis/hierarchialcomplete eingesehen werden. Die Verfahren die auf dem Abstandsmaß basieren, das keine Metrik ist, werden in den Grafiken mit der Endung -nom beschrieben.

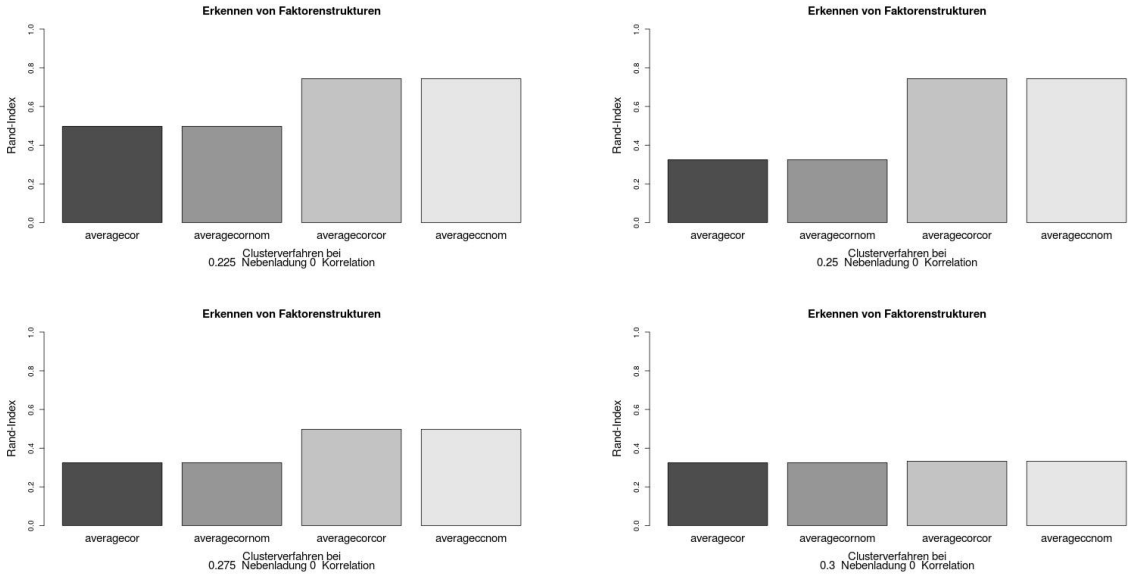


Abbildung 10.1: Erkennung der Faktorstruktur der hierarchischen Clustermethoden mit Average-Linkage

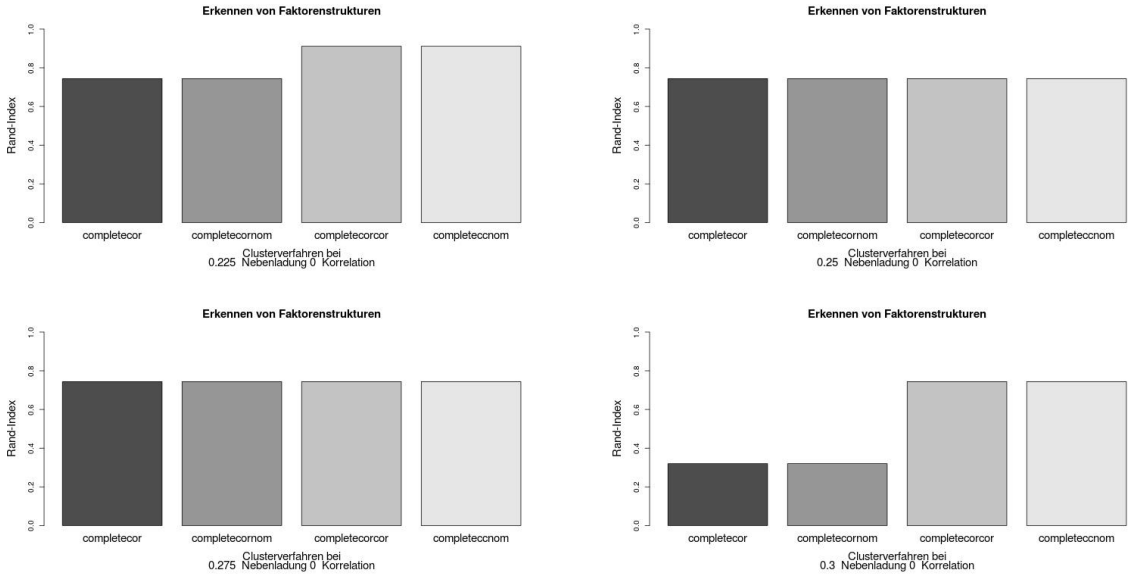


Abbildung 10.2: Erkennung der Faktorstruktur der hierarchischen Clustermethoden mit Complete-Linkage

10.1.2 Strukturerhaltung

Wenn es allerdings darum geht, wie gut die Clusterstruktur in Stichproben wiedererkannt wird, sind kleine Unterschiede feststellbar, wie in 10.3 gesehen werden kann. Das Verfahren Average-Linkage und Korrelation als Ähnlichkeitsmaß schneidet mit dem Abstandsmaß, das eine Metrik

ist, besser ab als bei dem Abstandsmaß, das keine Metrik ist. Bei den anderen Verfahren ist kein großer Unterschied feststellbar.

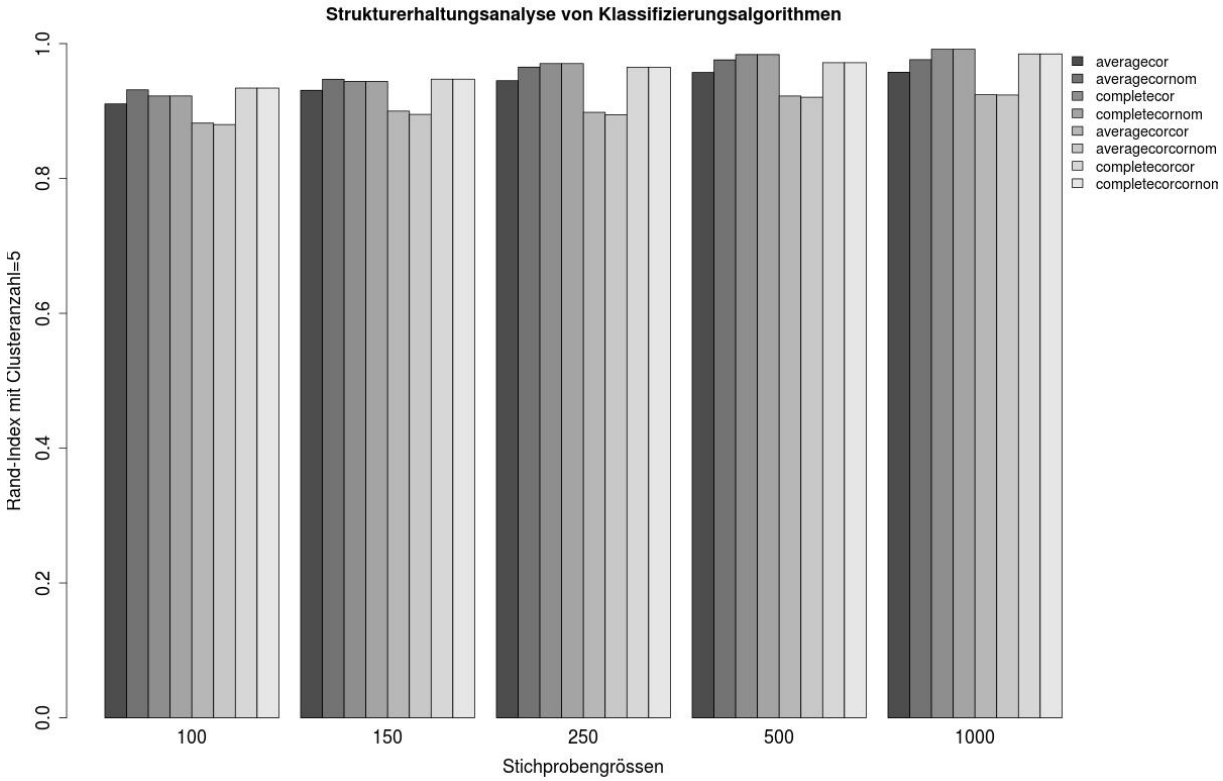


Abbildung 10.3: Strukturerhaltung der hierarchischen Clustermethoden

10.2 Benötigte Dimensionen beim Multidimensional Scaling

Für den Algorithmus des Multidimensional Scaling kann, wie in Kapitel 4.4 erklärt wurde, die Dimension der Items reduziert werden und zwar so, dass der euklidische Abstand zwischen den Items immer noch möglichst stark dem gewählten Abstandsmaß entspricht. Es soll in diesem Abschnitt nun untersucht werden, welche Dimension schon ausreicht, um die Faktorstruktur genügend gut zu erkennen und die Struktur in Stichprobendatensätzen möglichst gut wiederzufinden. Der Verfahren Dim1 bezieht sich auf das K-Means-MDS Verfahren in Dimension 1, Dim2 auf K-Means-MDS in Dimension 2,... und kmeansmds bezeichnet das K-Means-MDS Verfahren in der vollen Dimension von 39 (da 40 Items).

10.2.1 Faktorstruktur

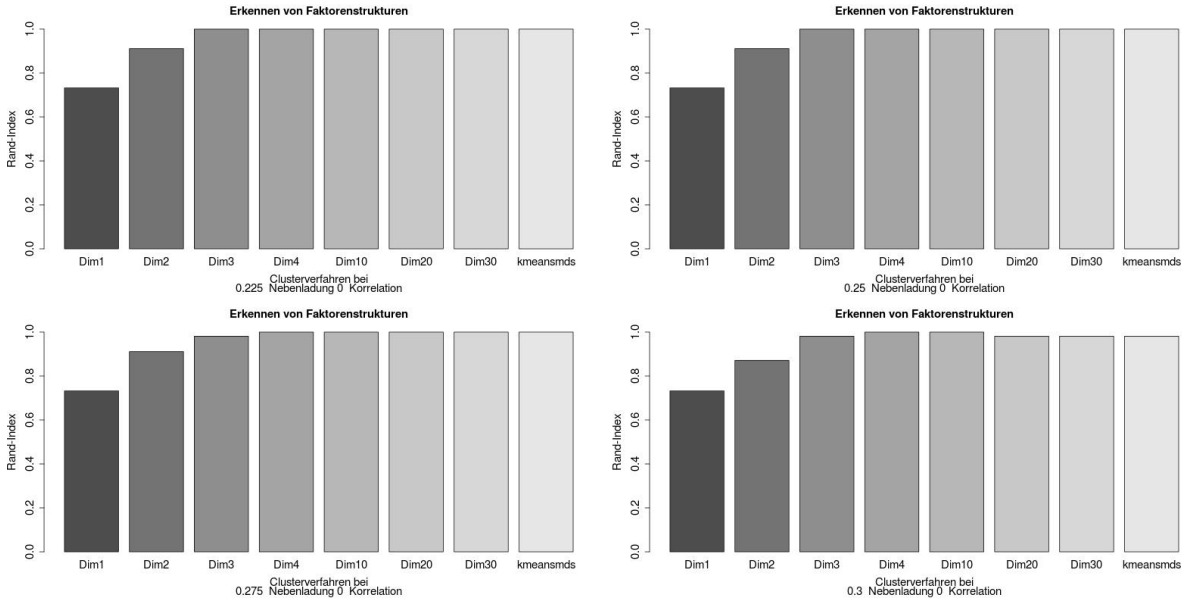


Abbildung 10.4: Faktorstrukturerkennung von Multidimensional Schaling verschiedener Dimensionen bei größeren zunehmenden Nebenladungen

Wie in Grafik 10.4 zu erkennen ist, werden bereits ab 4 Dimensionen die Faktorstrukturen in etwa so gut wiedererkannt wie bei der vollen Anzahl an Dimensionen, nämlich 39. Diese Grafik bezieht sich zwar nur auf den Fall, in dem die Nebenladungen größer werden, doch im Ordner ersteDaten/faktorbasis/dimensionMDS kann nachgeprüft werden, dass dies tatsächlich auch für die anderen hier untersuchten Veränderungen der Faktorstruktur gilt.

10.2.2 Strukturerhaltung

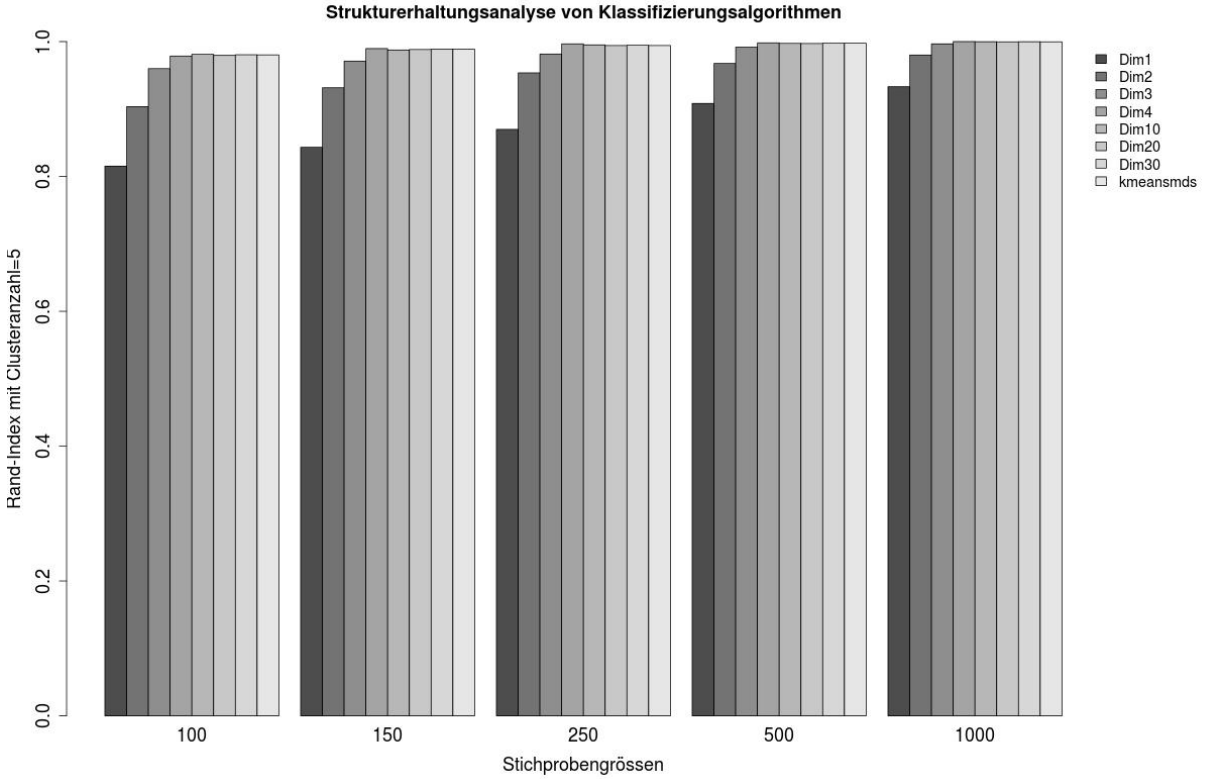


Abbildung 10.5: Strukturerhaltung von Multidimensional Schaling verschiedener Dimensionen

In Abbildung 10.5 erkennt man, dass ab einer Dimension von 4 auch die Struktur so gut erkannt wird wie bei der Verwendung der maximalen Dimension.
Diese Erkenntnis ist deshalb so interessant, weil es bedeutet, dass in Bezug auf die Clusterung ein Teil der in der Korrelationsmatrix enthaltenen Information ausreicht, um die Clusterung bezogen auf die zwei untersuchten Qualitätsmerkmale so gut durchzuführen wie wenn die gesamte Information der Korrelationsmatrix verwendet werden würde.

10.3 Korrelation der Korrelation als grundlegendes Abstandsmaß

Es scheint auch noch sinnvoll, für die auf K-Means basierenden Verfahren auszuprobieren, ob sie denn besser werden, wenn die Korrelation der Korrelation anstelle der Korrelation als Abstandsmaß verwendet wird. Bei K-Means-MDS bedeutet das, dass die Punkte so im Koordinatensystem platziert werden, dass die Abstände der Punkte abhängig sind von der Korrelation der Korrelation der entsprechenden Items. Für K-Means-Kor bedeutet das, dass die Koordinaten der Punkte den Korrelationen der Korrelationen zu den jeweils anderen Items entsprechend. Diese Ergebnisse können in 10.6 betrachtet werden. Für die Verfahren, für die die Korrelation der Kor-

relation als Abstandsmaß verwendet wird, wird ein -cor als Endung angehängt. Beim Verfahren K-Means-Cor ist eine Verbesserung feststellbar bei größeren Nebenladungen und Faktorkorrelationen. Über K-Means-MDS kann hier keine Aussage getroffen werden, da das Verfahren bei den verwendeten Nebenladungen und Faktorkorrelation auch schon ohne Verwendung der Korrelation der Korrelation die Faktorstruktur genau wiedererkennt. Da K-Means-MDS und K-Means-Koord bei der Erkennung der Faktorstruktur bisher immer gleich gut waren, wird hier nur K-Means-MDS untersucht. Höhere Nebenladungen und Faktorkorrelationen können nicht ausprobiert werden, da die dabei herauskommenden Korrelationsmatrizen irgendwann ungültige Werte größer als 1 als Korrelationen ergeben. Zusätzlich werden auch die hierarchischen Verfahren mit der Korrelation der Korrelation der Korrelation als Abstandsmaß gemessen. Die Grundidee ist dabei die gleiche wie bei der Korrelation der Korrelation als Abstandsmaß, nur dass das Verfahren zur Berechnung der Korrelation der Korrelation nun auf die Korrelation der Korrelation angewendet wird und nicht auf die Korrelation. Wie in 10.7 gesehen werden kann, bringt ein weiteres Verwenden der Korrelation der Korrelation für die hierarchischen Verfahren einen Zugewinn bei der Faktorstrukturerkennung. Die Ergebnisse für andere Nebenladungen und Faktorkorrelationen können in ersteDaten/faktorbasis/additionalCor betrachtet werden.

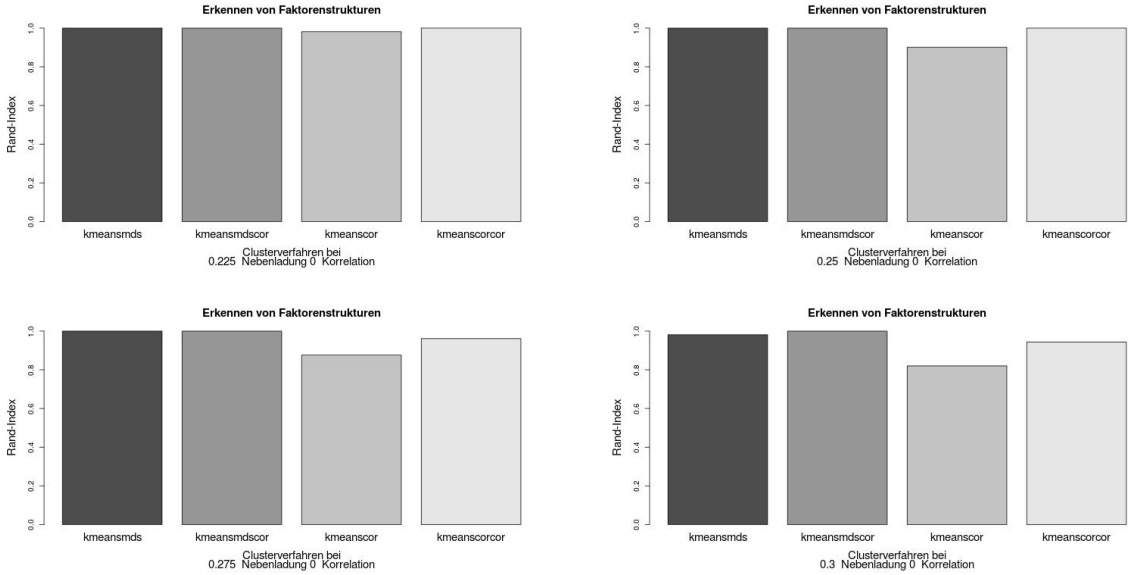


Abbildung 10.6: Faktorstrukturerkennung der Verfahren basierend auf der Korrelation der Korrelation bei zunehmenden Faktorkorrelationen und Nebenladungen

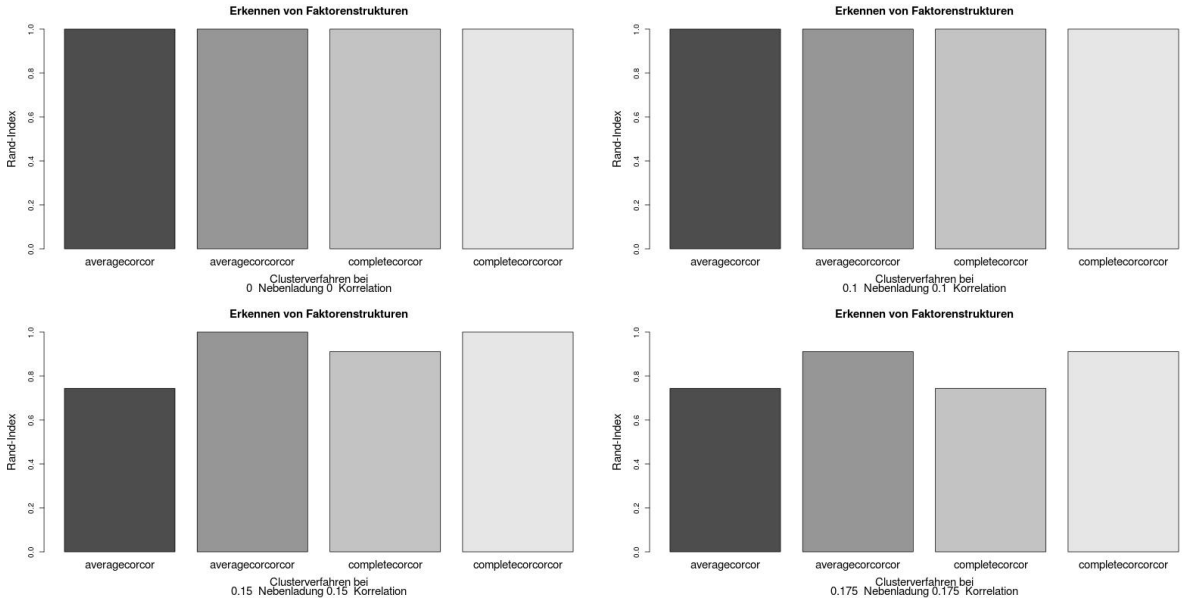


Abbildung 10.7: Faktorstrukturerkennung der Verfahren basierend auf der Korrelation der Korrelation bei zunehmenden Faktorkorrelationen und Nebenladungen

Einen Aufschluss darauf, warum die hierarchischen Verfahren bei einem weiteren Verwenden der Korrelation besser abschneiden, könnten die Dendrogramme liefern.

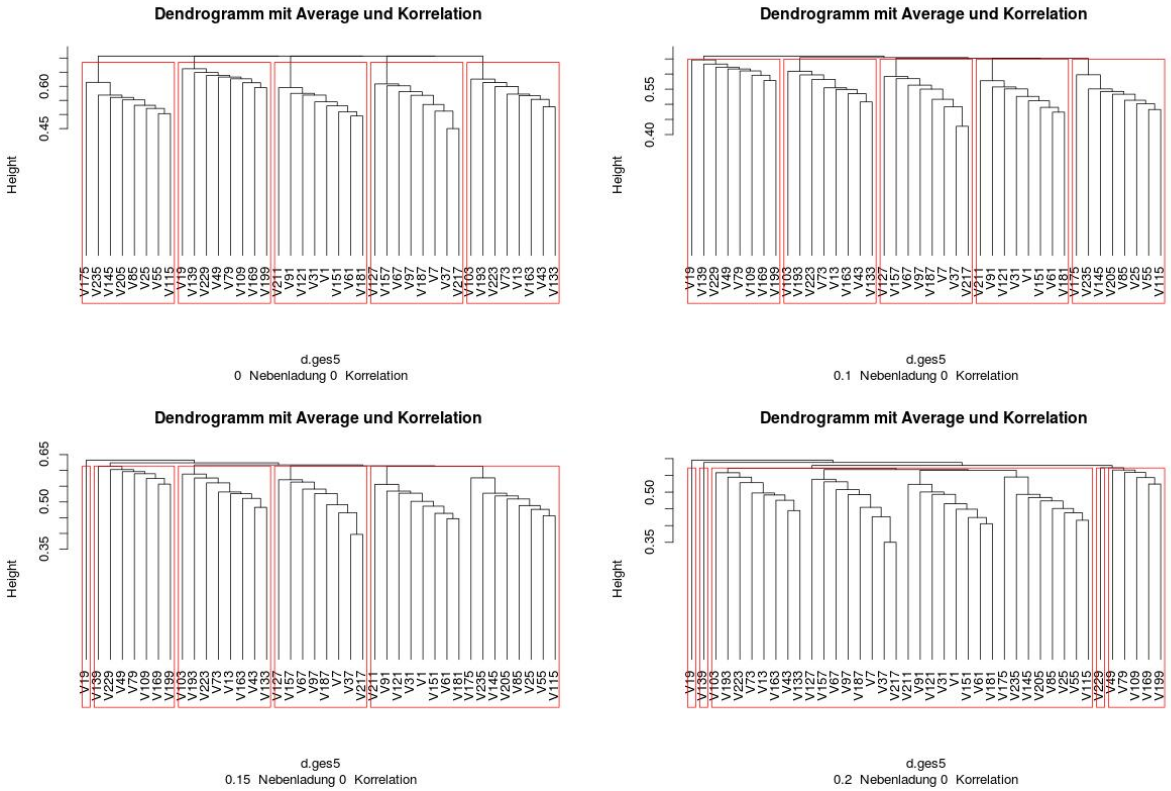


Abbildung 10.8: Dendrogram für Average-Linkage mit Korrelation als Ähnlichkeitsmaß

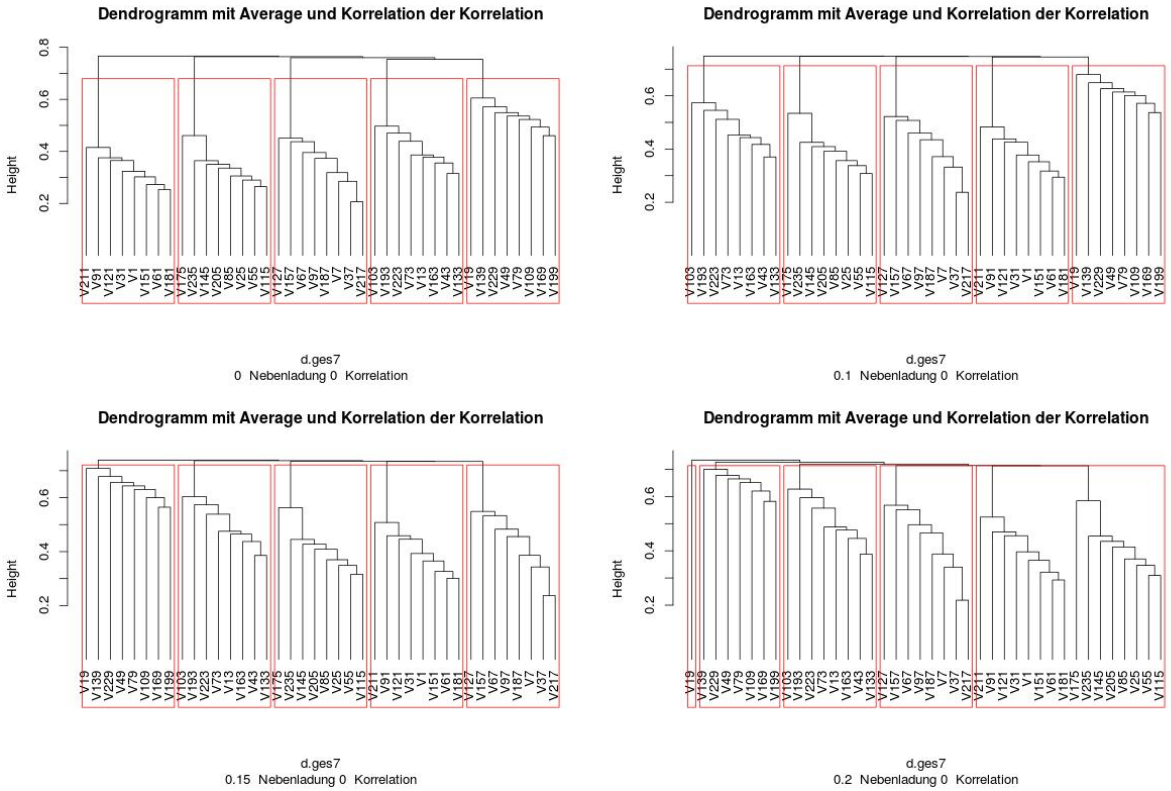


Abbildung 10.9: Dendrogramm für Average-Linkage mit Korrelation der Korrelation als Ähnlichkeitsmaß

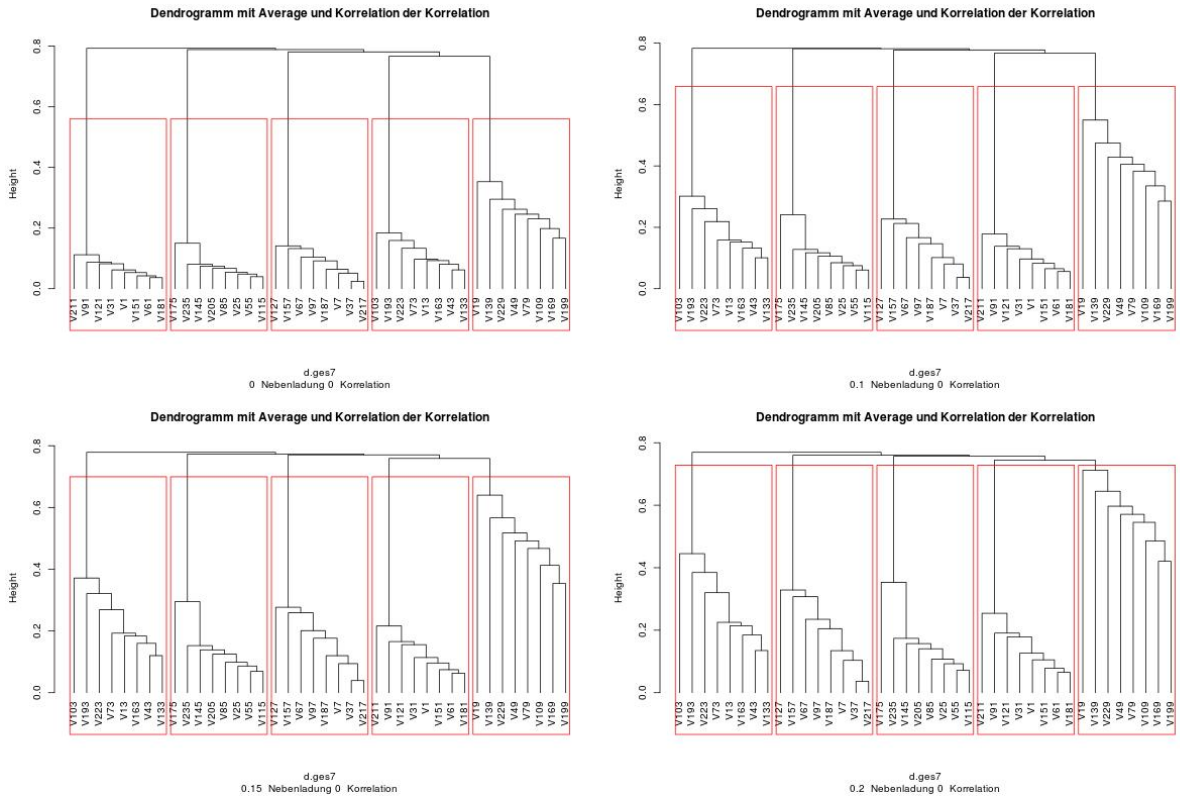


Abbildung 10.10: Dendrogramm für Average-Linkage mit Korrelation der Korrelation der Korrelation als Ähnlichkeitsmaß

In den drei Grafiken 10.8, 10.9 und 10.10 lässt sich erkennen, wie die Ähnlichkeit zwischen den Clustern bei Anwenden der Korrelation auf die bisherige Korrelationsmatrix gleich bleibt, aber die Ähnlichkeit innerhalb der Cluster immer größer wird. Leider kann dieses Verhalten noch nicht ausreichend mathematisch beschrieben werden. Diese neuen auf einer zusätzlichen Anwendung der Korrelation auf die bisherige Korrelationsmatrix schneiden allerdings auch bei der Strukturerhaltung schlechter ab als die anderen Verfahren, siehe Grafik 10.11.

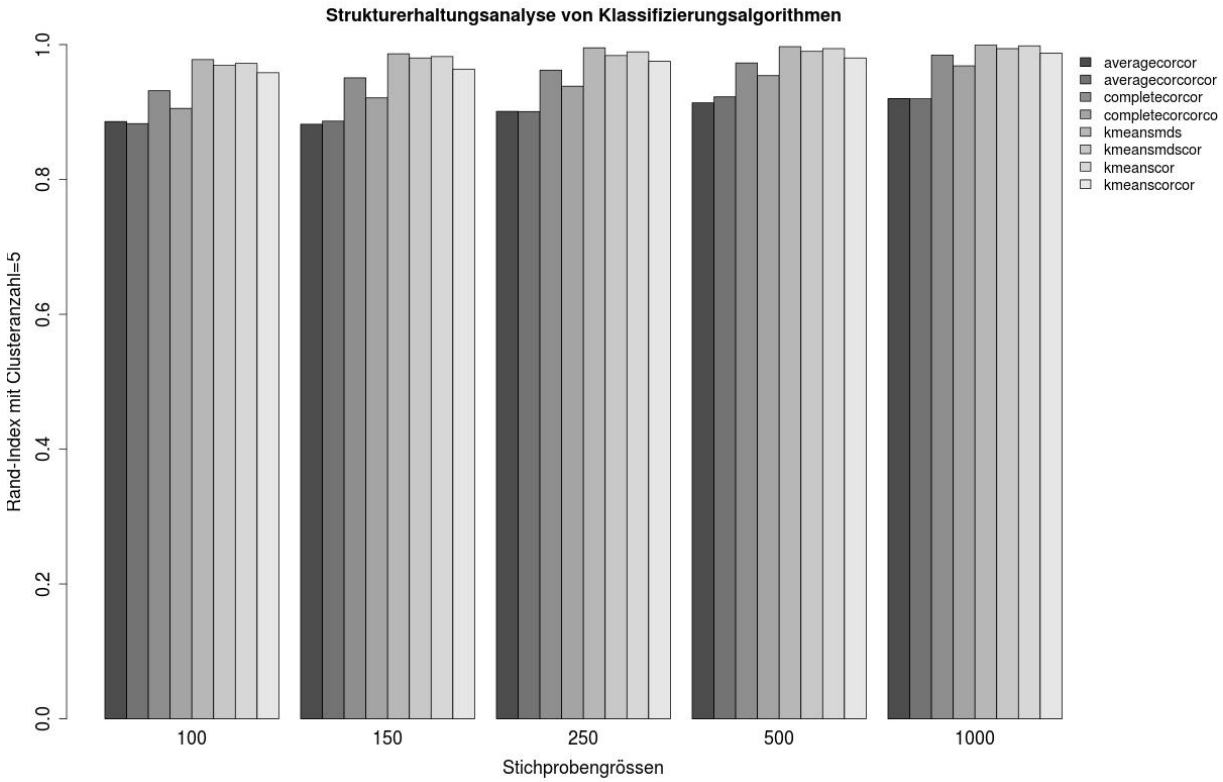


Abbildung 10.11: Vergleich der Strukturerhaltung der Verfahren basierend auf der Korrelation der Korrelation

11 Überprüfung der Ergebnisse anhand anderer Daten

In diesem Kapitel wird überprüft, ob die in den vorherigen Untersuchungen gefundenen Ergebnisse auch für andere Daten wiedergefunden werden können. Zwar wurden aus den Daten nur die Hautpladungen übernommen, allerdings haben auch diese einen Einfluss auf die Ergebnisse und deshalb werden nun andere Daten als Grundlage der Untersuchungen genommen. Während in den vorherigen Kapitel die Unterfacetten N1, E2, O3, A4 und C5 als Grundlage der Daten genommen wurden, sind es in diesem Kapitel die Unterfacetten N2, E3, O4,A5 und C6. Dies entspricht den Unterfacetten Reizbarkeit aus Neurotizismus, Durchsetzungsfähigkeit aus Extraversion, Offenheit für Handlungen aus Offenheit, Bescheidenheit aus Verträglichkeit und Besonnenheit aus Gewissenhaftigkeit. Die Grundlage der Faktorstruktur dieser Untersuchungen ist also die Faktoranalyse mit 5 Faktoren angewandt auf die Itemdaten aus diesen Unterfacetten. Allerdings werden wie in den vorherigen Kapiteln auch wieder nur die Hauptladungen übernommen und die Nebenladungen und Faktorkorrelationen variiert. Es werden bei diesen Daten nicht alle Untersuchungen noch einmal vorgenommen, sondern überprüft, ob die wichtigsten Erkenntnise auch auf diese anderen Daten zutreffen. Im Ordner /zweiteDaten können aber auch die nicht in dieser Arbeit aufgeführten Ergebnisse betrachtet werden.

11.1 Zunehmende Nebenladungen

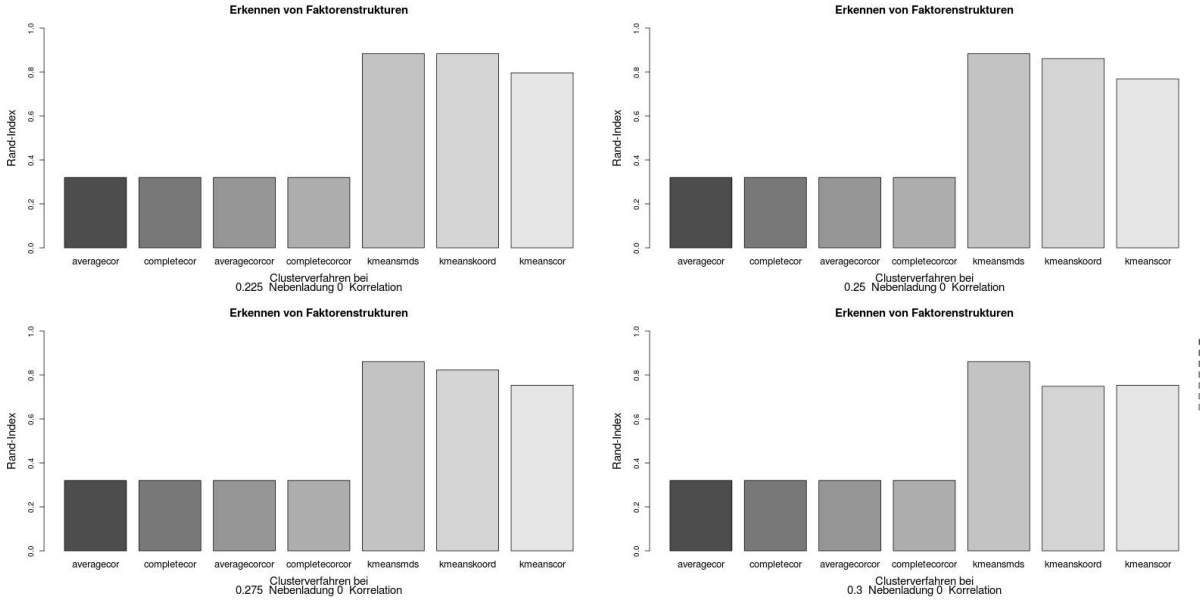


Abbildung 11.1: verschiedene Clustermethoden bei größerer zunehmenden Nebenladungen

Wie man in 11.1 sehen kann, schneiden auch bei diesen Daten die auf K-Means basierten Verfahren besser ab als die hierarchischen Verfahren und unter den K-Means basierten Verfahren schneidet K-Means-MDS am besten ab. Im Unterschied zu den vorherigen Daten schneiden aber die hierarchischen Verfahren noch einmal schlechter ab und es ist ein Unterschied feststellbar zwischen K-Means-MDS und K-Means-Koord. K-Means-MDS schneidet nämlich deutlich besser ab. Der Grund, warum die hierarchischen Verfahren noch etwas schlechter abschneiden als bisher könnte sein, dass in den neuen Daten die Hauptladungen etwas kleiner sind, die kleinste Hauptladung in den neuen Daten beträgt beispielsweise 0.182 im Gegensatz zu den 0.276 der alten Daten.

11.2 Zunehmende Korrelation

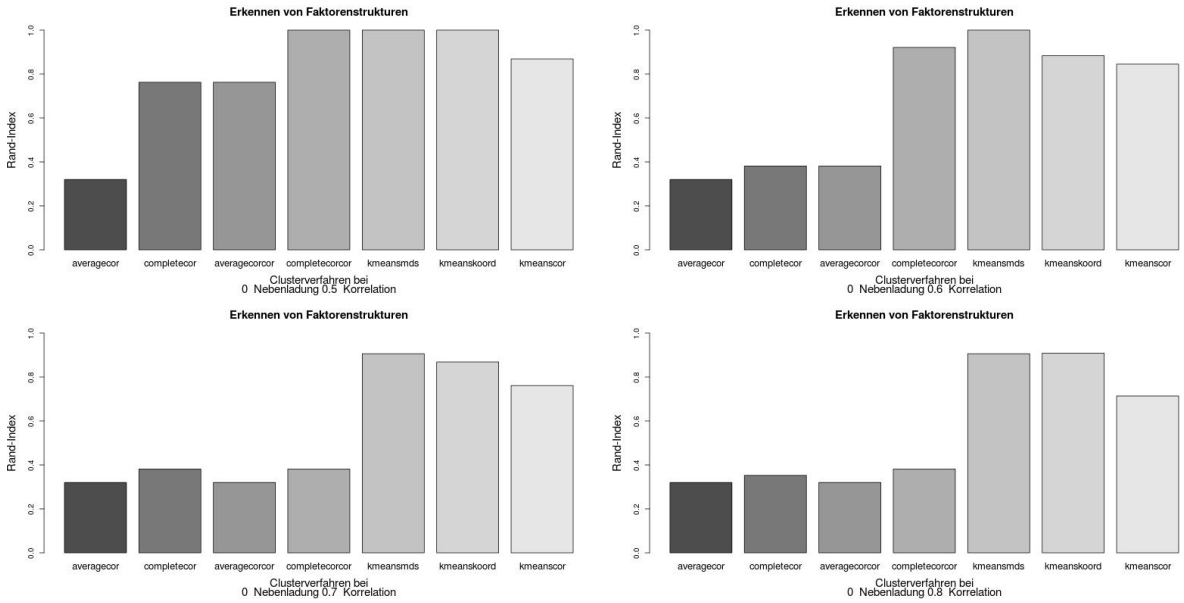


Abbildung 11.2: verschiedene Clustermethoden bei größereren zunehmenden Faktorkorrelationen

Bei der zunehmenden Faktorkorrelation und gleichbleibenden Nebenladungen in 11.2 ergibt sich auch wieder ein ähnliches Bild, nur dass auch hier die hierarchischen Verfahren noch etwas schlechter abschneiden als bei den alten Hauptladungen und auch wieder K-Means-MDS besser abschneidet als K-Means-Koord.

11.3 Strukturerhaltung

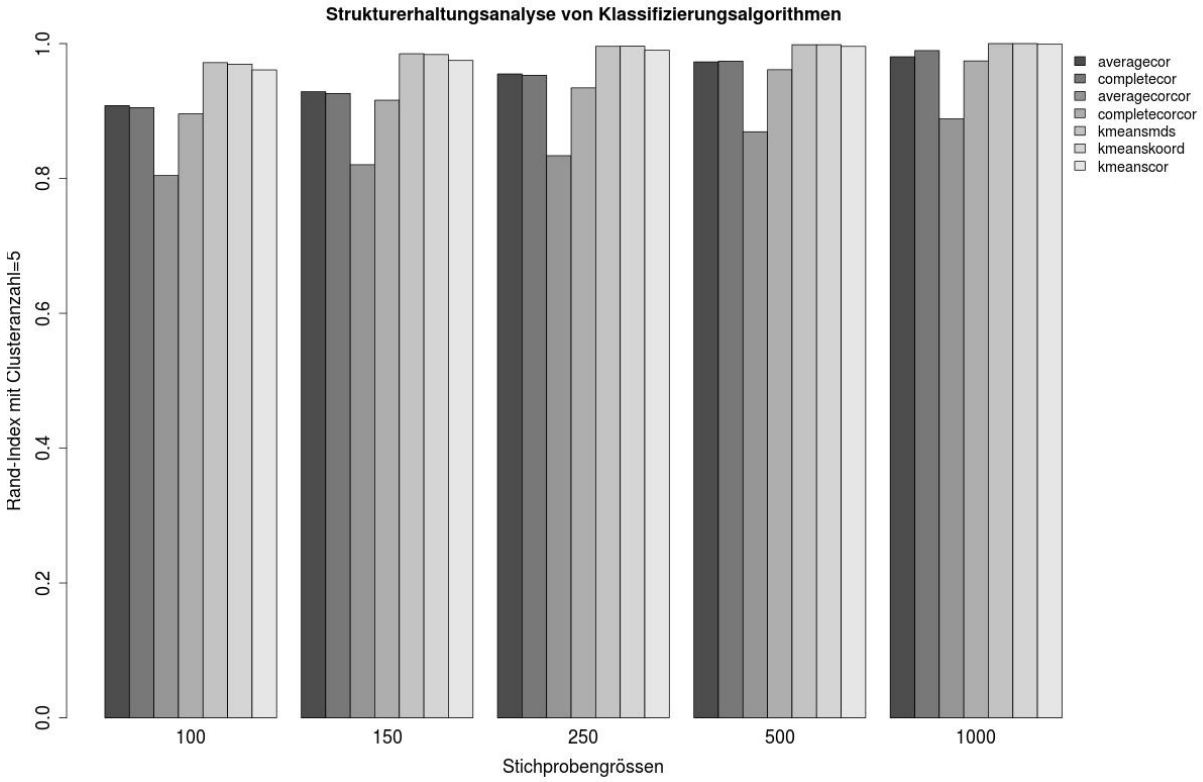


Abbildung 11.3: Strukturerhaltung verschiedene Clustermethoden

Auch die Ergebnisse der Strukturerhaltung in 11.3 sind sehr ähnlich zu den vorherigen Ergebnissen. Bei den hierarchischen Verfahren schneidet Average und Korrelation der Korrelation etwas schlechter ab, während die auf K-Means basierten Verfahren gerade bei kleineren Stichprobengrößen eindeutig besser abschneiden und K-Means-Kor tendenziell minimal schlechter ist als die anderen beiden auf K-Means basierten Verfahren.

11.4 Benötigte Dimensionen beim Multidimensional Scaling

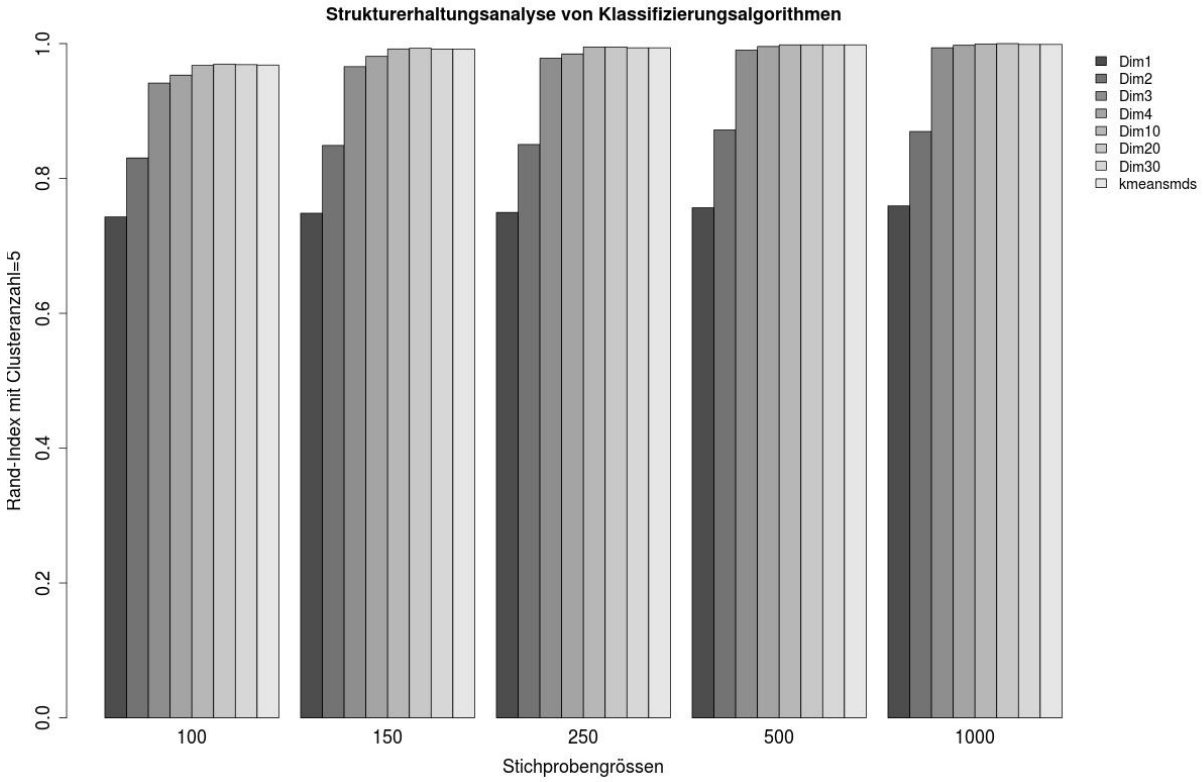


Abbildung 11.4: Strukturerhaltung von Multidimensional Schaling verschiedener Dimensionen

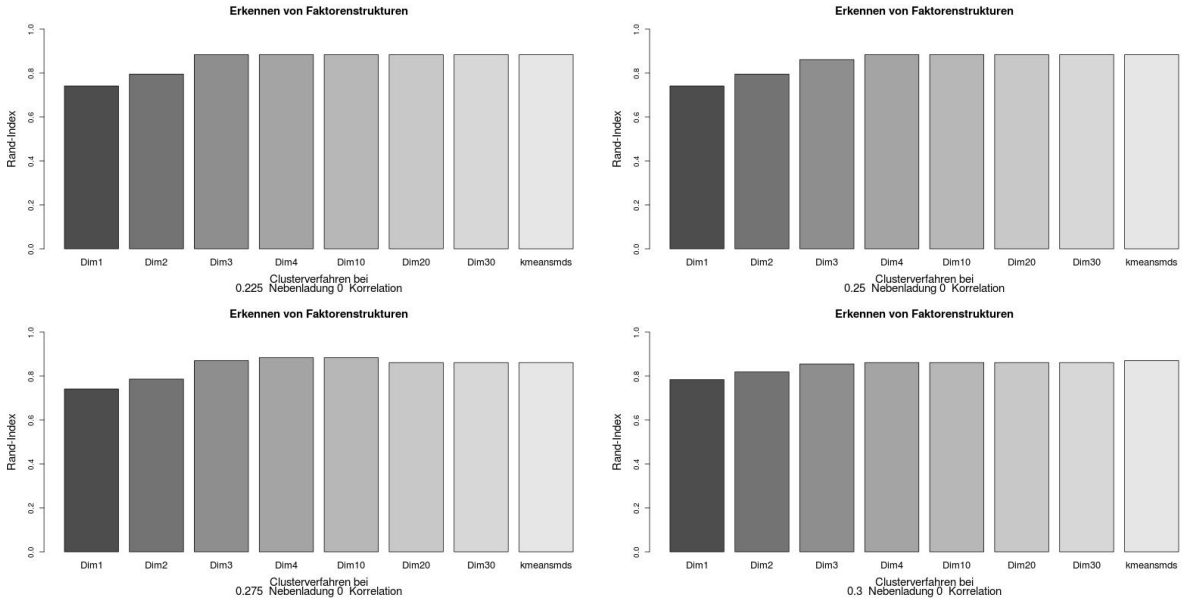


Abbildung 11.5: Faktorstrukturerkennung von Multidimensional Schaling verschiedener Dimensionen bei zunehmenden Nebenladungen

Wie man in 11.4 und 11.5 sehen kann, kann auch die Korrelationsmatrix der zweiten Daten so in niedrigerer Dimension dargestellt werden, dass die Faktorstrukturerkennung und Strukturerhaltung gleich gut ist. 5 Dimensionen sind ungefähr ausreichend um die gleiche Qualität zu erhalten wie bei Verwenden von allen 39 Dimensionen. Es wird bei den neuen Daten also eine Dimension mehr benötigt um die Korrelationsmatrix ausreichend gut zu beschreiben.

12 Fazit und Ausblick

In dieser Arbeit wurde ein neues Qualitätsmerkmal zur Beurteilung von Clusterverfahren, die Faktorstrukturerkennung, entwickelt. Zudem wurden neue Verfahren zum Clustern von ordinalskalierten Itemdaten entwickelt, das Kmeans-MDS Verfahren, das KMeans-Kor Verfahren und das KMeans-Koord Verfahren. In Bezug auf das Qualitätsmerkmal der Faktorstrukturerkennung schneiden alle diese drei Verfahren besser ab als die sonst üblichen hierarchischen Clusterverfahren. Über die hierarchischen Verfahren lässt sich sagen, dass grundsätzlich die Korrelation der Korrelation für die Faktorstrukturerkennung besser geeignet ist als Ähnlichkeitsmaß und das Complete-Linkage als Linkage-Verfahren. Bei dem bereits bekannten Merkmal zur Beurteilung von Clusterverfahren, der Strukturerhaltung, schneiden die auf K-Means basierten Verfahren auch wieder alle besser ab als die hierarchischen Verfahren. Besonders hervorzuheben sei dabei vor allem das KMeans-MDS Verfahren, bei dem zudem noch die Möglichkeit besteht, die Dimension zu reduzieren und somit die Komplexität der Korrelationsmatrix niedriger darzustellen. Eine Reduktion der Dimensionszahl auf 5 Dimensionen führt bei beiden Qualitätsmerkmalen zu keiner Verschlechterung mehr im Gegensatz zur vollen Dimension. Das K-Means-MDS Verfahren schneidet in der Regel sowohl bei der Faktorstrukturerkennung und Strukturerhaltung besser ab als K-Means-Kor und hat gegenüber K-Means-Koord den Vorteil, dass es viel schneller berechnet werden kann. Dieses Verfahren sollte deshalb weiter auf die Anwendbarkeit in der psychologischen Praxis getestet werden.

Digitaler Anhang

In einer der Bachelorarbeit beigelegten CD befinden sich die folgenden Daten:

Daten

Die Daten im Originalformat.

Code

Der R-Code für die Simulation

Dateneinlesen Code für das Einlesen der Daten

faktorensimulation Code für das Simulieren zur Faktorstrukturerkennung

faktorstruktur Code für das Wiederfinden der Faktorstruktur

hierarchialclustermethods die hierarchischen Clustermethoden

kmeansclustermethods die auf K-Means basierenden Clustermethoden

koorddendrograms Code zum Erzeugen der Dendrogram-Grafiken

strukturerhaltung Code zum Simulieren der Strukturerhaltung

Vergleichsverfahren Implementierung der Clustervergleichsmethoden

Grafiken

Hier befinden sich alle im Bericht befindlichen Grafiken

ersteDaten

die ursprünglich untersuchten Daten

faktorbasis die Untersuchungen zur Faktorstrukturerkennung

strukturerhaltung die Untersuchungen zur Strukturerhaltung

zweiteDaten

die Daten mit denen die Ergebnisse dann noch einmal überprüft werden

faktorbasis die Untersuchungen zur Faktorstrukturerkennung

strukturerhaltung die Untersuchungen zur Strukturerhaltung

Abbildungsverzeichnis

Literaturverzeichnis

[Bacon, 2001] Bacon, D. R. (2001). An evaluation of cluster analytic approaches to initial model specification. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(3):400 f.

[Borg und Groenen, 2005] Borg, I. und Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer.

[Harman, 1976] Harman, H. (1976). *Modern Factor Analysis*. University of Chicago Press.

[Hölzl et al., 2013] Hölzl, A., Käsdorf, A., Lamar, T., und Högerle, K. (2013). Vergleich von Faktoren- und Clusteranalyse.

[Manhart und Hunger, 2008] Manhart, J. und Hunger, M. (2008). Die Faktorenanalyse: Das Rotationsproblem / Extraktionskriterien für faktoren.

[Ostendorf und Angleitner, 2004] Ostendorf, F. und Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae: NEO-PI-R*. Hogrefe, Verlag für Psychologie.

[van Dongen und Enright, 2012] van Dongen, S. und Enright, A. J. (2012). Metric distances derived from cosine similiarity and pearson and spearman correlations.

[Wagner und Wagner, 2007] Wagner, S. und Wagner, D. (2007). Comparing clusterings - an overview.

[Wu, 2012] Wu, J. (2012). *Springer Theses: Advances in K-Means Clustering: a Data Mining Thinking*. Springer Theses, Recognizing Outstanding Ph.D. Research. Springer Berlin Heidelberg.

[Xu und Wunsch, 2008] Xu, R. und Wunsch, D. (2008). *Clustering*. IEEE Press Series on Computational Intelligence. Wiley.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Pörndorf, 23.06.2013

Andreas Hölzl