Explorative Faktorenanalyse: Einführung und Analyse mit R

Christina Werner

2009-11-12; aktualisiert 2012-03-21

1 Wozu verwendet man Faktorenanalysen?

Mit Hilfe von Faktorenanalysen kann untersucht werden, ob sich viele beobachtbare Variablen (z. B. die Antworten von Personen auf eine Menge Testitems) durch wenige dahinterstehende latente Variablen, sog. Faktoren erklären lassen (z. B. Persönlichkeitskonstrukte wie Extraversion). So eine Analyse macht Sinn, wenn die beobachtbaren (manifesten) Variablen miteinander zusammenhängen, so daß sie mehr oder weniger übereinstimmende Information enthalten. In diesem Fall kann man sich vorstellen, daß mehrere beobachtbare Variablen jeweils Indikatoren (Anzeichen, Operationalisierungen) der gleichen dahinterstehenden latenten Variable darstellen. Beispielsweise könnten dann viele Testitems, die auf den Faktor Extraversion zurückführbar wären, zur Messung dieses Persönlichkeitskonstrukts verwendet werden.

Formal wird dies durch die Bedingung der sog. *lokalen stochastischen Unabhängigkeit* beschrieben: Wenn die beobachteten Variablen nur deshalb korrelieren, *weil* sie Indikatoren für die gleiche, dahinterstehende latente Variable sind, so müßte ihr Zusammenhang verschwinden, wenn man die latente Variable auf einem Wert konstanthielte. Ein Beispiel: Bezogen auf die Gesamtpopulation sollten verschiedene Testitems zu Extraversion zusammenhängen – sie sollen ja das gleiche Merkmal messen. Würde man dagegen lauter gleich extravertierte Personen befragen (die latente Variable konstanthalten), sollte der Zusammenhang zwischen den verschiedenen Extraversions-Items verschwinden. Wäre dies nicht der Fall, könnte dies ein Hinweis darauf sein, daß die Variablen z. B. noch einen weiteren, unberücksichtigten Faktor messen, oder zwischen Variablen direkte Effekte bestehen (z. B. weil Antworten auf ein Item von Antworten auf ein anderes Item abgeschrieben werden können).

2 Arten von Faktorenanalysen

Grundsätzlich gibt es zwei unterschiedliche Klassen von Faktorenanalysen: explorative Faktorenanalysen (EFA) und konfirmatorische Faktorenanalysen (CFA, confirmatory factor analysis).

Explorative Faktorenanalysen können zum *Auffinden* einer in einem Datensatz möglicherweise existierenden Faktorenstruktur benutzt werden. Sie bieten aber keine oder nur unzureichende Möglichkeiten zum exakten Prüfen von Hypothesen. Explorative Faktorenanalysen werden in der Praxis trotzdem oft in Situationen angewendet, in denen man bereits Annahmen über die Anzahl der einem Datensatz zugrundeliegenden Faktoren, deren Interkorrelationen oder die Ladungen der manifesten Variablen auf den Faktoren hat. Dies ist grundsätzlich auch kein

Problem: Bestimmte Entscheidungen in explorativen Faktorenanalysen wären gerade dann besonders willkürlich, wenn man überhaupt keine Vermutungen über die Faktorenstruktur hätte.

Konformatorische Faktorenanalysen erlauben dagegen auch die Prüfung konkreter Hypothesen über die Faktorenstruktur nach objektiven Kriterien. Konfirmatorische Faktorenanalysen sind einfache Sonderfälle von Strukturgleichungsmodellen und können mit entsprechenden Programmen (z. B. LISREL, Mplus, EQS, AMOS, oder dem R-Paket OpenMx) durchgeführt werden. Im Folgenden geht es ausschließlich um explorative Faktorenanalysen.

Eine weitere Unterscheidung ist die zwischen Faktorenanalysen (im engeren Sinne) und Hauptkomponentenanalysen, die methodisch eng verwandt sind mit Faktorenanalysen, deren Ergebnisse aber nicht genau so interpretierbar sind wie die von Faktorenanalysen. Diese Unterscheidung ist besonders für meßfehlerbehaftete Daten (z. B. Antworten auf Testitems) relevant und wird im Abschnitt Faktorenextraktion weiter unten besprochen.

3 Voraussetzungen

Alle hier besprochenen Methoden basieren auf der Annahme linearer Zusammenhänge zwischen beobachteten Variablen (z. B. Items) und Faktoren. Ginge man beispielsweise von 2 Faktoren aus (hier bezeichnet als ξ_1, ξ_2), so sollte jedes Item x_i durch diese beiden Faktoren erklärbar sein, d. h. rechnerisch eine gewichtete Summe der Faktoren darstellen (zuzüglich einer Fehlervariable e_i). Die Gewichte $\lambda_{i1}, \lambda_{i2}$ sind die sog. **Faktorladungen** des i-ten Items auf Faktor ξ_1 bzw. ξ_2:

$$x_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + e_i$$

Die Ergebnisse sind daher nur dann sinnvoll interpretierbar, wenn die zu faktorisierenden Variablen zumindest annähernd als intervallskaliert betrachtet werden können.

Da explorative Faktorenanalysen keine exakten Hypothesenprüfungen ermöglichen, erfordern die hier besprochenen Verfahren im Normalfall aber keine bestimmten Verteilungsannahmen (mit Ausnahme der Maximum Likelihood-Faktorenanalyse, s. u.).

4 Durchführung von Faktorenanalysen

4.1 Vorarbeiten

Nach Auswahl der Variablen, die faktorisiert werden sollen, sollte man sich zuerst mit Grafiken und Deskriptivstatistik einen Eindruck verschaffen, inwieweit die Daten plausibel sind und sich die Variablen tatsächlich zur Faktorisierung eignen dürften:

- Deskriptivstatistik (insbesondere Minima und Maxima, Häufigkeitstabellen der Werte) und/oder Grafiken wie Histogramme und Box-Whiskers-Plots: Damit läßt sich prüfen, daß keine Werte außerhalb des möglichen Bereichs oder sonstige unplausible Ausreißer auftreten (z. B. aufgrund von Dateneingabefehlern).

- Korrelationsmatrix der Variablen und paarweise bivariate Streudiagramme (Scatterplots) aller zu faktorisierenden Variablen: Faktorenanalysen gehen von linearen Zusammenhängen aus, insofern sollten zwischen Variablen, die mutmaßlich auf den gleichen Faktor zurückführbar sind, positive Korrelationen bestehen (bzw. negative im Falle invertierter Items) und in Streudiagrammen entsprechende Zusammenhänge erkennbar sein.

4.2 Extraktionsmethoden

Der erste Hauptschritt einer explorativen Faktorenanalyse ist die Faktorenextraktion, bei der die *Anzahl* der Faktoren bestimmt wird, auf die die manifesten Variablen zurückgeführt werden sollen, und abgeschätzt wird, wieviel Varianz der Variablen auf gemeinsame Faktoren zurückgeführt werden kann. Für die Extraktion gibt es verschiedene Methoden. Der Unterschied zwischen den beiden am häufigsten genutzten Methoden (Haupt*komponenten*analyse und Haupt*achsen*analyse) beruht auf den Varianzanteilen, die in der Varianz von beobachtbaren (manifesten) Variablen enthalten sein können:

- Gemeinsame Varianz: Eine beobachtbare Variable kann Varianz mit anderen beobachtbaren Variablen gemeinsam haben (z. B. ein Extraversions-Testitem mit anderen Testitems für Extraversion).
- Spezifische Varianz: Die beobachtbare Variable beinhaltet etwas systematisches, daß sie mit keiner anderen Variablen gemeinsam hat (z. B. erfaßt eine bestimmte Intelligenztest-Aufgabe neben allgemeiner Intelligenz möglicherweise auch noch ein spezifisches Wissensgebiet).
- Fehlervarianz: Psychologische Variablen unterliegen oftmals zufälligen Meßfehlereinflüssen. Testitems werden z. B. von Personen bei wiederholter Testdurchführung nicht immer genau gleich angekreuzt, auch wenn sich das Merkmal selbst zwischenzeitlich gar nicht verändert haben muß.

Den Anteil der gemeinsamen Varianz an der gesamten Varianz einer Variablen bezeichnet man als **Kommunalität** der Variablen. Die Anteile der gemeinsamen und spezifischen Varianz zusammen ergeben entsprechend die Reliabilität). Hier besteht der Unterschied zwischen den Methoden (vgl. Fabrigar, Wegener, MacCallum & Strahan, 1999; Preacher & MacCallum, 2003; Bandalos & Boehm-Kaufman, 2009):

- Bei einer **Hauptkomponentenanalyse** (Principal Components Analysis, PCA) wird versucht, die *gesamte* Varianz der manifesten Variablen auf Faktoren zurückzuführen. Da psychologische Variablen in aller Regel Fehlervarianz enthalten, ist dies für die meisten psychologischen Anwendungen *nicht* sinnvoll. Tendenziell führt der Versuch, die gesamte Varianz manifester Variablen zu erklären auch dazu, daß die (durch die Fehlervarianz beeinflußte) Faktorenlösung weniger stabil ist, d. h. das Faktorisieren der selben Variablen in einer anderen Stichprobe eher zu abweichenden Ergebnissen führt. Möchte man außerdem die Faktoren als latente Variablen im Sinne psychologischer Konstrukte auffassen, so sollten sie das *Gemeinsame* der beobachtbaren Variablen darstellen. Spezifische Varianz einzelner manifester Variablen oder gar zufällige Fehlervarianz sollte deshalb nicht auf Faktoren zurückgeführt werden.
- Bei einer **Hauptachsenanalyse** (Principal Axes Factoring, PAF) soll von vornherein nur die *gemeinsame* Varianz der Variablen durch Faktoren erklärt werden, was für psychologische Anwendungen in der Regel sinnvoll ist. Hierfür muß man aber die Kommunalitäten

der Variablen kennen – bzw. schätzen, da man sie im Vorhinein ja noch nicht kennt. Eine untere Grenze für die Kommunalität einer Variablen ist der höchste *einfache* Determinationskoeffizient, der zwischen dieser Variable und irgendeiner anderen zu faktorisierenden Variable auftritt (soviel Varianz hat sie schon mit einer einzigen anderen Variablen gemeinsam). Eine bessere Schätzung der Kommunalität ist der *multiple* Determinationskoeffizient bei Vorhersage dieser Variablen durch alle anderen zu faktorisierenden Variablen (diese Schätzung wird üblicherweise in der Hauptachsenanalyse verwendet).

Das eigentliche rechnerische Vorgehen der Extraktion ist bei Hauptkomponenten- und Hauptachsenanalysen ähnlich. Der Unterschied besteht darin, daß die Hauptkomponentenanalyse von einer gewöhnlichen Korrelationsmatrix der Variablen mit Einsen in der Diagonale ausgeht. Für die Hauptachsenanalyse werden dagegen die Einsen in der Diagonale durch die geschätzten Kommunalitäten ersetzt (sog. reduzierte Korrelationsmatrix).

Die Instabilität der Ergebnisse von Hauptkomponentenanalysen und die potentiellen Unterschiede zwischen Hauptkomponenten- und Hauptachsenanalysen sind dann besonders groß, wenn die Stichprobe eher klein ist, nur wenige Items pro Faktor vorhanden sind und die spezifischen Varianzanteile der Items groß sind. Für sehr große Stichproben und sehr große Itemanzahlen pro Faktor wird der Unterschied dagegen vernachlässigbar.

- Eine Alternative zur Hauptachsenanalyse ist die **Maximum Likelihood-Faktorenanalye**, die rechnerisch anders erfolgt. Sie geht (wie die Hauptachsenanalyse) von der Vorstellung aus, daß sich die beobachteten Variablen aus einem durch gemeinsame Faktoren erklärbaren Anteil und einem Fehleranteil zusammensetzen. Die Faktoren und die Fehler werden als normalverteilt angenommen. Mit dieser Verteilungsvoraussetzung können die Ladungen der Variablen auf den Faktoren so geschätzt werden, daß die Plausibilität (Likelihood) maximal wird, unter Annahme der geschätzten Parameter Daten mit den empirisch tatsächlich gefundenen Zusammenhängen zu beobachten.

  Maximum Likelihood-Faktorenanalysen haben den Vorteil einer inferenzstatistischen Testmöglichkeit für die Faktorenanzahl (s. u.). Bei erfüllter Verteilungsvoraussetzung und wenn die Modellannahme korrekt ist, d. h. sich die Variablen tatsächlich genau aus den angenommenen Faktoren und zufälligen Fehlern zusammensetzen, können Maximum Likelihood-Faktorenanalysen die präzisesten Ergebnisse liefern. Allerdings treten unter bestimmten Umständen häufig Schätzprobleme mit nicht sinnvoll interpretierbaren oder irreführenden Lösungen auf: Dies passiert besonders in kleinen Stichproben, bei Variablen mit niedrigen Faktorladungen, und in Fällen, in denen die Annahme über die Zusammensetzung der Variablen aus Faktoren und Zufallsfehler verletzt ist, z. B. wenn die Variablen Anteile eines nicht berücksichtigten Faktors enthalten (vgl. MacCallum & Tucker, 1991; Briggs & MacCallum, 2003). Leider ist dies in der Psychologie eine häufig anzutreffende Situation, so daß Maximum Likelihood-Faktorenanalysen hier nicht immer empfehlenswert sind.

4.3 Anzahl zu extrahierender Faktoren

Für das Festlegen der Anzahl zu extrahierender Faktoren gibt es verschiedene Möglichkeiten. Rein rechnerisch lassen sich maximal so viele Faktoren extrahieren, wie Variablen faktorisiert werden. Praktisch wird dagegen eine Rückführung auf *wenige* Faktoren angestrebt.

- Hat man vorher schon eine feste Vorstellung über die zu erwartenden (oder erhofften) Faktoren, so kann man die Anzahl direkt vorgeben.

- Um abzuschätzen, wieviele der potentiell extrahierbaren Faktoren vermutlich überzufällig Varianz der Variablen erklären können, ist der **Scree-Plot** hilfreich. Hierbei werden die Eigenwerte der extrahierbaren Faktoren (rechnerisch die Eigenwerte der Korrelationsmatrix) grafisch veranschaulicht. Der **Eigenwert** eines Faktors entspricht der von ihm bei allen Variablen zusammen erklärten Varianz. Prinzipbedingt ist dabei der Eigenwert des ersten Faktors der höchste, der Eigenwert des zweiten Faktors der zweithöchste, etc. Trägt man die Faktornummer auf der horizontalen Achse, den jeweiligen Eigenwert auf der vertikalen Achse ab, so zeigt die entstehende Kurve meist einen mehr oder weniger deutlichen Knick (vgl. die Beispiel-Abbildung in Abschnitt 5.2). Man extrahiert dann alle Faktoren, deren Eigenwerte vor diesem Knick liegen, da man annimmt, daß diese tatsächlich überzufällige Gemeinsamkeiten der faktorisierten Variablen widerspiegeln, während die Faktoren nach dem Knick mutmaßlich nur geringe, zufällige Zusammenhänge erklären. Dieses grafische Vorgehen kann über die Logik der Parallelanalyse von Zufallsvariablen auch statistisch begründet werden.

- Bei **Parallelanalysen** werden zum Vergleich Zufallsvariablen herangezogen, die in der Population unkorreliert sind. Erzeugt man solche Zufallsdaten für die jeweilige Anzahl von Variablen und Personen, treten dabei in jeder Stichprobe üblicherweise trotzdem geringe, zufällige Korrelationen der Variablen auf. Diese Zufallsdaten kann man dann zum Vergleich ebenfalls faktorisieren, und erhält dabei sehr geringe Eigenwerte (in der Regel um 1 herum), die in etwa auf einer langsam abfallenden Gerade liegen. Vergleicht man nun die empirischen Eigenwerte der eigenen Daten mit denen aus der Parallelanalyse von Zufallsdaten, so würde man die Faktoren beibehalten, deren Eigenwerte größer sind als die entsprechenden aus der Parallenanalyse, da man davon ausgehen kann, daß diese überzufällige Zusammenhänge erklären.

- Oftmals wird empfohlen, keine Faktoren mit einem Eigenwert kleiner 1 zu extrahieren (**Kaiser-Kriterium**). Begründet wird dies damit, daß Faktorenanalysen im Regelfall von standardisierten Variablen ausgehen (*Korrelations*matrix), somit ein Eigenwert von 1 nur der Varianz einer einzigen Variablen entspricht und Faktoren mit noch geringerem Eigenwert keinen Erklärungswert hätten, da man die Variablen ja auf *wenige* Faktoren zurückführen möchte.

  Dies ist so aber allenfalls für Hauptkomponentenanalysen korrekt, bei denen ein Eigenwert von 1 tatsächlich der zu erklärenden Varianz einer Variablen entspräche. Bei Faktorenanalysen (im engeren Sinne, also zur Erklärung *gemeinsamer* Varianz der Variablen) ist dies nicht richtig: Hängen die zu faktorisierenden Variablen systematisch, aber nur niedrig zusammen, so ist die Kommunalität jeder Variablen unter Umständen viel kleiner als 1, so daß auch ein Faktor mit Eigenwert kleiner 1 substantiell gemeinsame Varianz mehrerer Variablen erklären kann.

  In der Praxis ist das Kaiser-Kriterium damit bestenfalls für die Anzahl sinnvoll zu extrahierender Komponenten in Hauptkomponentenanalysen anwendbar, wobei hier jedoch oftmals weniger, manchmal aber auch mehr Komponenten inhaltlich angemesen wären. Für Faktorenanalysen im engeren Sinne ist das Kaiser-Kriterium logisch nicht begründbar und sollte daher nicht angewandt werden (vgl. Fabrigar et al., 1999; Preacher & MacCallum, 2003; Bandalos & Boehm-Kaufman, 2009).

- Im Rahmen von Maximum Likelihood-Faktorenanalysen kann ausgehend von der Modellannahme normalverteilter Faktoren und Fehlervariablen geprüft werden, ob die beobachteten Zusammenhänge der Daten durch die angenommene Anzahl Faktoren erklärbar sind. Hierfür gibt es eine χ^2**-basierte Teststatistik**, wobei ein signifikanter Wert auf Abweichungen zwischen der angenommenen Struktur und den Daten hinweist, also nicht erstrebenswert ist. Die Faktorenanzahl könnte dann ggf. weiter erhöht werden, bis keine signifikante Abweichung mehr besteht, die Zusammenhänge der Daten also mit der dann angenommenen Anzahl Faktoren erklärt werden können.

4.4 Rotationsmethoden

Bei der Extraktion werden die Faktoren so bestimmt, daß der erste Faktor möglichst viel (gemeinsame) Varianz der Variablen erklärt, der nächste Faktor von der verbleibenden Varianz der Variablen wiederum so viel wie möglich, etc., wobei die Faktoren voneinander unabhängig sind. Die so entstehende Lösung ist bei mehr als einem Faktor in der Regel inhaltlich nicht interpretierbar, da dann alle Variablen mehr oder weniger hoch auf dem ersten Faktor laden, etwas niedriger auf dem zweiten, etc.

Nach der Extraktion folgt daher der Schritt der Faktorenrotation. Die Rotation dient dazu, die Faktoren inhaltlich interpretierbar zu machen. Das Ziel ist dabei eine sog. **Einfachstruktur**, bei der die Ladungen der Variablen auf den Faktoren entweder betraglich möglichst hoch oder aber um Null sind, so daß die Variablen klar einzelnen Faktoren zugeordnet und die Faktoren anhand der zugeordneten Variablen inhaltlich interpretiert werden können. Die rechnerische Veränderung bei der Rotation betrifft nur die Faktorladungen, also die Frage, welcher Faktor wieviel Varianz bei einer Variable erklärt. Es ändert sich jedoch nichts daran, wieviel Varianz jeder Variablen *insgesamt* durch die Faktoren erklärt wird, d. h. die Kommunalitäten bleiben gleich.

Für die Rotation gibt zwei verschiedene Klassen von Methoden:

- Bei **orthogonalen** Rotationen bleiben die verschiedenen Faktoren unkorreliert (nach der Extraktion sind sie es prinzipbedingt). Am häufigsten wird die **Varimax**-Rotation durchgeführt, bei der die Varianz der quadrierten Faktorladungen *pro Faktor* maximiert wird (d. h. man möchte pro Faktor möglichst viele betraglich hohe oder um Null liegende Ladungen erhalten), was die Interpretation der Faktoren erleichtert: Man kann den Faktor als das Gemeinsame der hoch auf ihm ladenden Variablen auffassen (ein Gegenstück hierzu ist **Quartimax**, bei der versucht wird, *pro Variable* möglichst viele hohe oder Null-Ladungen zu erhalten, um die Variablen klarer den Faktoren zuordnen zu können).

  Bei der Anwendung orthogonaler Rotationsmethoden setzt man voraus, daß die den Variablen zugrundeliegenden Faktoren tatsächlich unkorreliert sind. Falls die Faktoren aber tatsächlich korreliert sind, führen orthogonale Rotationen zu verzerrten Ergebnissen und liefern dann u. U. auch keine klare Einfachstruktur.

- Bei **obliquen** Rotationen werden Korrelationen zwischen Faktoren zugelassen. Psychologische Konstrukte dürften nur selten wirklich unkorreliert sein, daher sind oblique Rotationen für viele psychologische Fragestellungen angemessener als orthogonale (Fabrigar et al., 1999; Preacher & MacCallum, 2003; Bandalos & Boehm-Kaufman, 2009). Sie können u. U. auch dann gut interpretierbare Ergebnisse liefern, wenn orthogonale Rotationsmethoden zu keiner klaren Einfachstruktur führen. Sollten die Faktoren tatsächlich unabhängig sein, ist die Verwendung einer obliquen Rotationsmethode auch kein Problem, da diese zwar Korrelationen zwischen den Faktoren erlaubt, aber nicht erfordert oder erzwingt.

  Verbreitete oblique Rotationsmethoden sind Promax und Oblimin. **Promax** ist eine sog. Target-Rotationsmethode, d. h. es kann ein gewünschtes Kriteriums-Ladungsmuster vorgegeben werden. Wird kein bestimmtes Muster vorgegeben, rotiert Promax so, daß möglichst klar voneinander getrennte Item-Gruppen entstehen (ähnlich Varimax, aber eben mit potentiell korrelierten Faktoren). Direct **Oblimin** versucht, die Kovarianzen zwischen den quadrierten Faktorladungen aller Paare von Faktoren zu minimieren.

Schließlich kann man im Rahmen von Faktorenanalysen bei Bedarf auch noch **Faktorwerte** (factor scores) bestimmen lassen, d. h. geschätzte Ausprägungen der Personen auf den latenten Variablen. Solche Werte sind nicht völlig eindeutig (außer bei Hauptkomponentenanalysen), es stehen verschiedene Schätzverfahren dafür zur Verfügung.

5 Durchführung von Faktorenanalysen mit R

Für praktische Übungen zu diesem Abschnitt findet sich im Netz eine R-Beispielsyntax mit ausführlichen Kommentaren, sowie ein passender Beispieldatensatz:
`http://www.psychologie.uzh.ch/fachrichtungen/methoden/team/christinawerner/faktorenanalyse.html`

5.1 Vorarbeiten

Für die deskriptivstatistische Inspektion der Daten eignet sich die Basis-Funktion `summary()`, und/oder die Funktion `describe()` aus dem `psych`-Paket. Eine Korrelationsmatrix der Variablen erhält man mit `cor()`. Zur grafischen Inspektion ließe sich die Funktion `hist()` für Histogramme verwenden, die aber nur auf einzelne Variablen angewendet werden kann. Ebenso könnte man `plot()` für jeweils zwei Variablen verwenden, um einzelne bivariate Streudiagramme zu erhalten.

Eine nützliche Hilfe für größere Variablenmengen in Form von Matrizen oder data frames ist die Funktion `pairs()`, die direkt auf Daten-Matrizen und data frames angewandt werden kann und als Voreinstellung alle paarweisen Streudiagramme der Variablen in einer einzigen, matrixförmigen Abbildung erzeugt. Die Funktion `pairs.panels()` aus dem `psych()`-Paket benutzt dies, um neben den Streudiagrammen auch alle univariaten Histogramme mit Dichteschätzung sowie die (Pearson-)Interkorrelationen der Variablen darzustellen (für Abwandlungen davon siehe die Hilfe zu `pairs()`). Mit einer einzigen Funktion erhält man so einen guten Überblick auch über größere Variablenmengen. Damit die Ausgabe nicht unübersichtlich klein wird, ist bei mehr als ca. 10-15 Variablen ein schrittweises Vorgehen sinnvoll, z. B. die Anwendung auf jeweils alle Variablen, die mutmaßlich zum selben oder zu einem von jeweils zwei Faktoren gehören.

5.2 Anzahl zu extrahierender Faktoren

Steht die Anzahl zu extrahierender Faktoren nicht schon theoriegeleitet fest, so bietet das `psych`-Paket zur Abschätzung der Anzahl der Faktoren die Funktionen `VSS.scree()` für Scree-Plots und `fa.parallel()` für Scree-Plots mit integrierten Ergebnissen von Parallel-Analysen. Beide Funktionen können direkt auf die Daten der zu faktorisierenden Variablen/Items angewandt werden und damit vor der eigentlichen Faktorenextraktion eingesetzt werden.

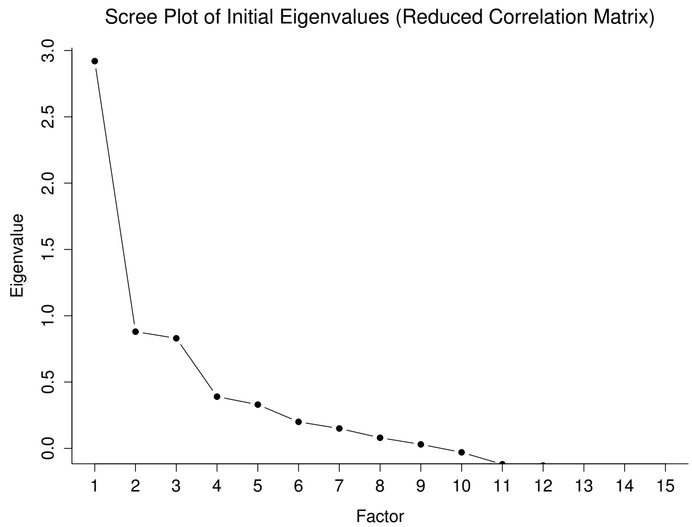
Ebensogut kann man aber die Eigenwerte einer Korrelationsmatrix auch mit der Basis-Funktion `eigen()` direkt bestimmen und mit `plot()` grafisch darstellen:

```
> plot(eigen(cor(items))$values)  # ggf. mit weiteren plot()-Optionen
```

Für Faktorenanalysen im engeren Sinne (also mit geschätzten Kommunalitäten auf der Diagonale der Korrelationsmatrix) erhält man die quadrierten multiplen Korrelationen je eines Items mit allen übrigen durch die Funktion `smc()` des `psych`-Pakets. Damit läßt sich die sog. reduzierte Korrelationsmatrix direkt angeben, so daß man auch deren Eigenwerte grafisch darstellen kann (was manche Statistik-Pakete nicht ohne weiteres beherrschen). Beispielsyntax:

```
> items.cors.reduced <- cor(items)
> diag(items.cors.reduced) <- smc(items)
> plot(eigen(items.cors.reduced)$values)  # ggf. mit weiteren plot()-Optionen
```

Wie in Abschnitt 4.3 erläutert, würde man alle Faktoren extrahieren, die im Eigenwerteverlauf vor dem „Knick“ zu einer langsam abfallen Gerade liegen, im folgenden Beispiel am ehesten drei:



5.3 Faktorenanalyse-Funktionen

Für die Faktorenextraktion stehen in R verschiedene Funktionen zur Verfügung. Aus dem Basis-Paket `stats` wären die Funktionen `factanal()` für Maximum Likelihood-Faktorenanalysen sowie `princomp()` für Hauptkomponentenanalysen nutzbar. Deutlich umfangreicher sind aber die hier besprochenen Funktionen des `psych`-Pakets:

- `fa()` für Faktorenanalysen mit verschiedenen Extraktionsmethoden, darunter
  - Hauptachsenanalyse (Principal Axes Factoring; Option `fm="pa"`),
  - Maximum Likelihood-Faktorenanalyse (`fm="ml"`),
  - Minimum Residual-Faktorenanalyse (`fm="minres"`),
- `principal()` für Hauptkomponentenanalysen

Für die anschließende Rotation werden z. T. Funktionen aus dem Paket `GPArotation` automatisch geladen, womit beispielsweise folgende Rotationsmethoden zur Verfügung stehen:

- Oblique: Promax, Oblimin
- Orthogonal: Varimax, Quartimax

Ein Beispiel für eine Hauptachsenanalyse mit Extraktion von 3 Faktoren, Kommunalitätenschätzung durch quadrierte multiple Korrelationen (squared multiple correlations, SMC) und anschließender Promax-Rotation sähe so aus:

```
> items.pa.promax <- fa(items,
                        nfactors=3,
                        SMC=TRUE,
                        fm="pa",
                        rotate="promax"
                        )
```

Hier werden die Rohdaten der zu faktorisierenden Variablen (`items`) verwendet, alternativ wäre auch eine Korrelationsmatrix möglich, ggf. mit Angabe der Anzahl der Personen (`n.obs=...`).

Die Ausgabe der so als Objekt `items.pa.promax` gespeicherten Faktorenanalyse-Ergebnisse beginnt mit einer Wiederholung aller Optionen der Analyse:

```
> print(items.pa.promax)
Factor Analysis using method =  pa
Call: fa(r = items, nfactors = 3, rotate = "promax", SMC = TRUE,
    max.iter = 100, fm = "pa")
```

Es folgt die **Faktorladungsmatrix** (nach der Rotation, es sei denn, daß `rotate="none"` gewählt wurde). Die Ladungsmatrix zeigt idealerweise eine Einfachstruktur, so daß man die Faktoren (hier: `PA1`, `PA2`, `PA3`) über das Gemeinsame der hoch auf ihnen ladenden Variablen (hier: `V1`, . . . , `V15`) interpretieren kann bzw. die Variablen eindeutig den Faktoren zuordnen kann:

```
    item  PA3  PA1  PA2   h2   u2
V1     1      0.45      0.26 0.74
V2     2      0.84      0.52 0.48
V3     3      0.56      0.33 0.67
V4     4      0.36      0.29 0.71
V5     5      0.36      0.38 0.62
V6     6 0.62           0.31 0.69
V7     7 0.48           0.21 0.79
V8     8 0.57           0.23 0.77
V9     9 0.58           0.38 0.62
V10   10 0.68           0.38 0.62
V11   11           0.65 0.32 0.68
V12   12           0.47 0.32 0.68
V13   13           0.64 0.31 0.69
V14   14           0.31 0.17 0.83
V15   15           0.42 0.22 0.78
```

In der Spalte `h2` (eigentlich: h^2) stehen die Kommunalitäten, also die von allen Faktoren zusammen bei der jeweiligen Variablen erklärten Varianzanteile. Die Spalte `u2` (u^2) enthält die Uniquenesses der Variablen, d. h. die Varianzanteile, die nicht durch die gemeinsamen Faktoren erklärt werden (dies beinhaltet spezifische Varianzanteile der Variablen und Fehlervarianz). Folglich ist für jede (standardisierte) Variable $h^2 + u^2 = 1$.

Als Voreinstellung werden Ladungen betraglich $\leq$ .30 nicht angezeigt. Für mehr Details kann man die Ladungsmatrix einzeln ausgeben lassen (hier z. B. mit allen Ladungen bis auf die betraglich sehr kleinen unter .10; mit `cutoff=0` würden alle ausgegeben):

```
> print(items.pa.promax$loadings, cutoff=.1)
Loadings:
    PA3    PA1    PA2
V1   0.169  0.451 -0.132
V2  -0.173  0.843 -0.130
V3          0.564
V4          0.361  0.213
V5   0.263  0.364  0.101
V6   0.624 -0.185
V7   0.483
V8   0.573 -0.157
V9   0.583
V10  0.676        -0.149
V11        -0.156  0.653
V12                0.470
V13 -0.129         0.639
V14  0.152         0.314
V15                0.415
```

Es folgen Angaben über die von den Faktoren über alle Variablen zusammen erklärte Varianz, d. h. die **Eigenwerte** nach der Rotation (`SS loadings`; sum of squared loadings). Außer den Absolutwerten sind auch die Varianzanteile (`Proportion Var`) und kumulativen Varianzanteile (`Cumulative Var`) aufgeführt. Hier werden also insgesamt 31% der Varianz aller Variablen durch die drei Faktoren zusammen erklärt:

```
                PA3  PA1  PA2
SS loadings    1.82 1.49 1.33
Proportion Var 0.12 0.10 0.09
Cumulative Var 0.12 0.22 0.31
```

Die Eigenwerte *vor* der Rotation werden nicht automatisch mit ausgegeben. Sie unterscheiden sich in der Regel nur minimal von den initialen Eigenwerten der reduzierten Korrelationsmatrix vor Durchführung der Analyse (die Werte, die ggf. zu Beginn im Scree-Plot dargestellt wurden). Bei Interesse kann man sie aber ausgeben lassen (vgl. den Scree-Plot weiter oben):

```
> print(items.pa.promax$values)
 [1]  2.92  0.88  0.83  0.39  0.33  0.20  0.15  0.08  0.03 -0.03 -0.12 -0.13
[13] -0.21 -0.29 -0.39
```

Bei der Extraktion klärt prinzipbedingt stets der erste Faktor die meiste Varianz auf, und die folgenden Faktoren jeweils weniger. Mit der Rotation ändert sich dies in der Regel so, daß die Eigenwerte der rotierten Faktoren dann ähnlicher zueinander werden. Die *Summe* der Eigenwerte der extrahierten Faktoren vor und nach der Rotation bleibt dabei gleich – es ändert sich lediglich, welcher Faktor wieviel Varianz erklärt, aber nicht, wieviel alle zusammen erklären (ebenso wie sich auch die Kommunalitäten der Items durch die Rotation nicht ändern).

Eine inhaltliche Interpretation der Eigenwerte im Sinne einer Wichtigkeit von Faktoren ist dennoch problematisch, da dies davon abhängt, welche Variablen in die Faktorenanalyse eingegangen sind: Faktorisiert man z. B. Items eines mehrdimensionalen Tests, so kann man nur dann gleich hohe Eigenwerte der Faktoren erwarten, wenn alle Testskalen gleich viele Items enthalten, und diese auch noch gleich reliabel sind.

Bei Verwendung einer obliquen Rotationsmethode folgen die **Faktor-Interkorrelationen:**

```
With factor correlations of
     PA3  PA1  PA2
PA3 1.00 0.55 0.50
PA1 0.55 1.00 0.46
PA2 0.50 0.46 1.00
```

Mit Werten zwischen .46 und .55 sind die Faktoren hier deutlich korreliert. Eine orthogonale Rotation (statt der benutzten Promax) wäre vermutlich nicht angemessen gewesen.

Schließlich wird noch ein **χ^2-Test auf Abweichung zwischen Modell und Daten** angegeben:

```
Test of the hypothesis that 3 factors are sufficient.

The degrees of freedom for the model is 63 and the objective function was 0.7
The number of observations was 100 with Chi Square = 63.95 with prob < 0.44
```

Dieser ist von entsprechenden Testmöglichkeiten für Maximum Likelihood-Faktorenanalysen abgeleitet und basiert auf dem dort zugrundegelegten Modell mit Normalverteilungsannahme für Faktoren und Fehlervariablen. Bei Nichterfüllung dieser Voraussetzungen sollte das Ergebnis – wenn überhaupt – nur mit Vorsicht interpretiert werden. Für explorative Faktorenanalysen sollte im Zweifelsfall die inhaltliche Interpretierbarkeit und der damit gegebene wissenschaftliche Nutzen der Ergebnisse ausschlaggebend sein.

6 Schlußfolgerungen

Bei der Durchführung und Ergebnisinterpretation explorativer Faktorenanlysen gibt es bei einer Reihe methodischer Entscheidungen erheblichen Ermessensspielraum, z. B. bei der Festlegung der Faktorenanzahl und der Frage, inwieweit das Ergebnis eine Einfachstruktur der faktorisierten Variablen aufweist. Damit besteht mehr Spielraum für Willkürlichkeit als bei vielen inferenzstatistischen Verfahren. Der Versuch, mit explorativen Faktorenanalysen allein etwas „beweisen“ zu wollen, kann daher kaum zum Erfolg führen (siehe z. B. die Debatten über „wahre“ Persönlichkeitsfaktoren in der Geschichte der Differentiellen Psychologie).

Inhaltlich aussagekräftig sind Faktorenanalysen in dem Maße, wie die faktorisierten Variablen inhaltliche Bedeutung im Rahmen von Theorien über das zugrundeliegende Konstrukt haben, und sie nicht bloß willkürliche Itemsammlungen darstellen. Gerade für die Validierung (explorativ-) faktorenanalytisch abgeleiteter latenter Variablen bieten sich außerdem konfirmatorische Faktorenanalysen an, sowie vollständige Strukturgleichungsmodelle, in die dann auch die Vorhersage von Außenkriterien einbezogen werden kann.

Literatur

Bandalos, D. L. & Boehm-Kaufman, M. R. (2009). Four common misconceptions in exploratory factor analysis. In C. E. Lance & R. J. Vandenberg (Hrsg.), *Statistical and methodological myths and urban legends* (S. 61–87). New York: Routledge.

Briggs, N. E. & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research*, *38*, 25–56.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

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

Preacher, K. J. & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, *2*, 13–43.