Hauptkomponentenanalyse und explorative Faktorenanalyse

Sabrina J. Mayer

Inhalt

Zusammenfassung

Der Beitrag führt in die Grundlagen der Hauptkomponentenanalyse (PCA) und explorativen Faktorenanalyse (EFA) ein. Gemeinsam ist diesen Verfahren eine Reduktion von einer Menge von korrelierten Variablen auf wenige Komponenten mit den Zielen der Vereinfachung, der leichteren Interpretation und zur Darstellung von zugrunde liegenden latenten Variablen. Zunächst werden die mathematischen Grundlagen der PCA erörtert, bevor Kriterien zur Bestimmung der idealen Anzahl der Komponenten und Rotationsverfahren zur Vereinfachung der Interpretation vorgestellt werden. Anschließend werden zwei Anwendungsbeispiele aus der Politikwissenschaft diskutiert. Darauf folgend wird die PCA von der EFA abgegrenzt sowie die EFA als induktives Datenmodellierungsverfahren

Für wertvolle Hinweise bei der Überarbeitung gilt Dank an Bernd Schlipphak. Für Hilfe bei der Erstellung und kritische Diskussion gilt mein Dank Achim Goerres und David Johann sowie den Hilfskräften Jakob Kemper, Sebastian Krause und Erik Wenker.

S. J. Mayer (✉)
Institut für Politikwissenschaft, Universität Duisburg-Essen, Duisburg, Deutschland

Deutsches Zentrum für Integrations- und Migrationsforschung (DeZIM), Berlin, Deutschland
E-Mail: mayer@dezim-institut.de; sabrina.mayer@uni-due.de

C. Wagemann et al. (Hrsg.), *Handbuch Methoden der Politikwissenschaft*,
https://doi.org/10.1007/978-3-658-16936-7_31

eingeführt. Abschließend erfolgt eine kommentierte Darstellung von Einführungswerken und weiterführender Literatur.

Schlüsselwörter

Hauptkomponentenanalyse · Faktorenanalyse · Dimensionsreduktion · Principal Component Analysis (PCA) · Explorative Faktorenanalyse (EFA)

1 Einleitung

Das Ziel der Hauptkomponentenanalyse (*Principal Component Analysis*, kurz PCA) ist es *k* metrische, korrelierte Variablen durch eine kleinere Anzahl von unkorrelierten Komponenten zu ersetzen, die trotzdem noch einen großen Anteil der Informationen des ursprünglichen Variablensatzes enthalten. Diese Reduktion ermöglicht es, die Interpretation der Variablenstruktur zu vereinfachen, da es in der Regel einfacher ist, eine geringe Zahl von Komponenten zu überschauen als zahlreiche paarweise Korrelationen. In der politischen Einstellungs- und Verhaltensforschung finden wir oftmals Fragen, die Ähnliches messen und deren gleichzeitige Verwendung die Analysen unübersichtlich machen: Verwenden wir eine PCA, so lassen sich beispielsweise mehrere Einstellungsvariablen zur Europäischen Union auf eine gemeinsame Komponente „Haltung zur EU“ zurückführen, die für weitere Berechnungen herangezogen wird (Evans 2000). Das Verfahren der PCA ist dabei eines der ältesten multivariaten Verfahren, deren Idee gleichermaßen auf Pearson (1901) und Hotelling (1933) zurückgeht. Eng verwandt mit der PCA ist die explorative Faktorenanalyse (EFA), bei der *k* metrische, korrelierte Variablen auf einige wenige zugrunde liegende latente Faktoren zurückgeführt werden. PCA und EFA unterscheiden sich in der praktischen Ausführung in empirischen Analysen unter der Verwendung von computergestützter Datenanalyse kaum noch – beide Verfahren haben die Gemeinsamkeit, dass eine Menge von Variablen zur Vereinfachung der Interpretation reduziert werden soll. Beide Verfahren setzen metrische bzw. quasi-metrische Variablen voraus.[1] Für kategoriale Variablen kann das ähnliche Verfahren der Korrespondenzanalyse herangezogen werden.

Beiden Verfahren liegen jedoch unterschiedliche Annahmen zugrunde. Die PCA ist ein deskriptives Verfahren zur Datenreduktion und kommt ohne inferenzstatistische Grundannahmen aus. Solange die Variablen hinreichend miteinander korrelieren, lässt sich eine PCA durchführen. Die EFA gehört als Datenmodellierungstechnik in den Bereich der induktiven Statistik und setzt verschiedene Grundannahmen (u. a. die multivariate Normalverteilung der Variablen) voraus. Weiterführend ermöglicht das Verständnis als Modellbildung die Anwendung der konfirmatorischen Faktorenanalyse (CFA), bei der überprüft wird, ob eine Menge

[1] Allerdings gibt es Autoren, die die Durchführung einer Hauptkomponenten-/Faktorenanalyse auch für dichotome Variablen als möglich beurteilen (siehe beispielsweise Arminger 1979, S. 159). Siehe dazu auch Abschn. 2.1.2.

von Variablen einer vorher festgelegten Datenstruktur entspricht. Ähnlich verhält es sich bei Strukturgleichungsmodellen, die ein Messmodell entsprechend der CFA und ein Pfadmodell umfassen (siehe auch den Beitrag von Berning in diesem Band). Der Unterschied zwischen PCA und EFA markiert den Übergang von deskriptiver zu induktiver Statistik. Daher wird zuerst das Verfahren der PCA erörtert. Anschließend erfolgen die Abgrenzung von der explorativen Faktorenanalyse sowie eine kommentierte Darstellung von Einführungswerken und weiterführender Literatur.

2 Die Hauptkomponentenanalyse

2.1 Mathematische Grundlagen der Extraktion von Komponenten

Ziel der PCA ist das Ersetzen von k korrelierten Variablen durch wenige unkorrelierte Hauptkomponenten, die trotzdem noch so viele ursprüngliche Informationen wie möglich bereithalten, die eine Interpretation der Zusammenhänge erleichtern. Anstatt beispielsweise viele Antworten in einer Umfrage zum gleichen Thema wie Immigrant/innen in Deutschland einzeln zu verwenden, versucht man, diese mit wenigen neu gebildeten Variablen zu beschreiben. Was ist aber mit „so viel wie möglich“ gemeint?

Die zentrale Idee des Verfahrens basiert auf dem Anteil an der totalen Varianz/Gesamtvarianz, d. h. der Summe der Varianzen der ursprünglichen k korrelierten Variablen, der von jeder der Hauptkomponenten erklärt wird. Die PCA wandelt also eine Menge von korrelierten Variablen (x_1, x_2, x_3,. . ., x_k) in eine gleiche Menge von unkorrelierten Komponenten um (y_1, y_2, y_3,. . ., y_k). Dies geschieht unter der Maßgabe, dass die Hauptkomponenten in absteigender Reihenfolge den größten Anteil der totalen Varianz (erste Hauptkomponente y_1) erklären, dann den größten Anteil der verbliebenen Varianz (zweite Hauptkomponente y_2), dann den größten Anteil der noch übrigen Varianz (dritte Hauptkomponente y_3) und so weiter. Die Menge aller k Hauptkomponenten erklärt also die totale Varianz (Gl. 1). Dabei stellen i und j in Gl. 1 die Laufparameter für die Menge der Variablen (i) bzw. Hauptkomponenten (j) dar.

$$\sum_{j=1}^{k} var\left(y_j\right) = \sum_{i=1}^{k} var(x_i) \tag{1}$$

Die Menge an Hauptkomponenten entspricht also der Menge an Variablen.[2] Wenn sich nun zeigt, dass bereits die erste/n Hauptkomponente/n einen großen Anteil der totalen Varianz der ursprünglichen Variablen erklären, so ist es möglich, auf die weiteren Hauptkomponenten ohne einen allzu großen Informationsverlust zu

[2]Ausnahme ist der eher hypothetische Fall perfekt korrelierter Variablen, hier erklärt bereits die erste Komponente die Gesamtvarianz.

verzichten. Beispielsweise zeigt sich häufig in der Analyse von Vertrauen in bestimmte politische Institutionen wie Parteien, Parlament, Regierung etc., dass diese fast ausschließlich durch eine Hauptkomponente dargestellt werden können.

Vor der Durchführung der PCA werden die ursprünglichen Variablen (x_i) standardisiert[3], so dass jede der x-Variablen gleichermaßen zur totalen Varianz beiträgt. Alle neu erstellten Variablen haben eine Standardabweichung und Varianz von 1 sowie einen Mittelwert von 0. Somit ist die Summe der Varianz der Variablen gleich der Anzahl der Variablen (siehe Gl. 2). Im nächsten Abschnitt soll nun das Verfahren der PCA an einem reduzierten Beispiel mit zwei Variablen dargestellt werden.

$$\sum_{i=1}^{k} var(x_i) = k \tag{2}$$

2.1.1 Die Umformung durch die PCA am Beispiel von zwei Variablen

Mit einem einfachen Beispiel soll die Art und Weise dargestellt werden, mit der die Umformung der k Variablen in Hauptkomponenten gelingt.[4] Sicherlich ist das genaue Verständnis dieses Vorgangs für eine erfolgreiche Anwendung des Verfahrens in empirischen Arbeiten nicht unbedingt erforderlich; das Nachvollziehen erleichtert jedoch die spätere Interpretation der Ergebnisse.

Im ersten Schritt überlegen wir, welchen Sinn eine Umformung der Ausgangsvariablen hat. Nehmen wir an, dass uns zwei Variablen x_1 und x_2 vorliegen, die in der gleichen Befragung auf einer identischen 100-stufigen Ratingskala in Form eines Sympathie-Thermometers erhoben wurden. Nehmen wir außerdem an, dass Variable x_1 die Sympathie mit der Regierungschefin und Variable x_2 die Sympathie mit der Partei der Regierungschefin bezeichnet. Zudem weisen beide Variablen die gleiche Varianz auf. Beide Variablen sind dabei hochkorreliert (Pearsons r = 0,90). Bei einer grafischen Darstellung in einem Streudiagramm würden die einzelnen Beobachtungen daher hier nahe bei der 45-Grad-Gerade durch den Ursprung liegen, da beide Variablen die gleiche Skala und Varianz aufweisen und eine positive Korrelation besteht.

Als nächstes formen wir diese Ausgangsvariablen in zwei neue Variablen, $y_1 = (\frac{x_1+x_2}{\sqrt{2}})$ und $y_2 = (\frac{x_2-x_1}{\sqrt{2}})$, um. In unserem Beispiel sind beide x-Variablen hochkorreliert. Daher offenbart uns die Differenz zwischen beiden Variablen (y_2) wenig Information hinsichtlich der Variation zwischen den einzelnen Individuen, während die Summe beider Variablen (y_1) für uns wesentlich hilfreicher ist. Da die beiden Variablen x_1 und x_2 hochkorreliert sind, wird also die erste der neuen Variablen y_1 eine große Varianz und die zweite y_2 eine geringe Varianz aufweisen. Daher kann man sagen, dass y_1 die meisten Informationen der beiden neuen Variablen hinsichtlich der Varianz zwischen den Individuen enthält.

[3] In den gängigen Statistikprogrammen (R, SPSS, Stata) geschieht dies automatisch durch die Verwendung der Korrelationsmatrix als Ausgangspunkt der PCA.

[4] Das folgende Beispiel geht auf Bartholomew et al. (2002, S. 116–120) zurück.

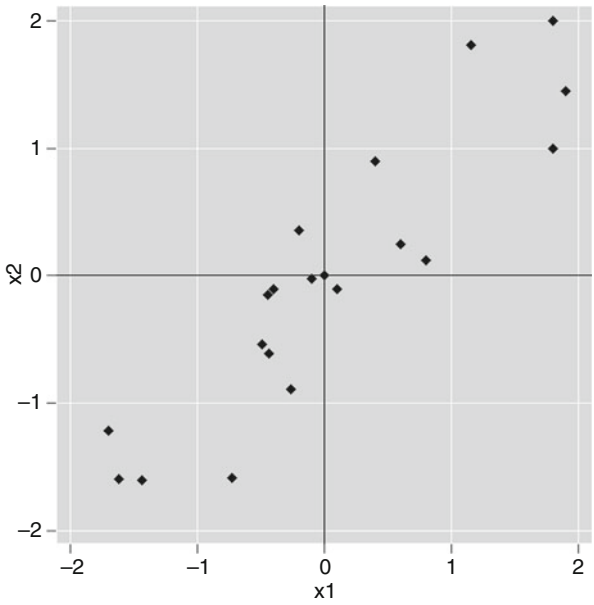

Abb. 1 Scatterplot zweier Variablen mit gleicher Varianz und einer Korrelation von r = 0,90

Auf Basis dieser Idee der Varianzverteilung können wir uns nun dem Verfahren der PCA mit zwei Variablen zuwenden. Dafür werden die beiden Variablen standardisiert, so dass sie einen Mittelwert von 0 und eine Varianz von 1 aufweisen. Abb. 1 zeigt den Scatterplot für 20 Beobachtungen.

Für das Finden der Hauptkomponenten beider Variablen ist eine orthogonale Rotation der Achsen notwendig, bei der beide Achsen weiterhin im rechten Winkel zueinanderstehen. Das bedeutet: Man versucht die neuen Komponenten so zu legen, dass die Werte der Merkmalsträger auf diesen Komponenten eine möglichst große Unterschiedlichkeit aufweisen. Die erste Hauptkomponente wird dabei in Richtung der größten Varianz liegen und durch die Minimierung der Summe der Abweichungsquadrate von den Beobachtungen ermittelt.[5] Sobald die Lage der ersten Hauptkomponente feststeht, ist die Lage der zweiten Hauptkomponente ebenfalls festgelegt, da diese ja durch das Fehlen jeglicher Korrelation zwischen den Hauptkomponenten in rechtem Winkel zur ersten Hauptkomponente liegen muss. Die beiden Hauptkomponenten sind in Abb. 2 durch gestrichelte Linien dargestellt. Da die Varianz von x_1 und x_2 gleich ist und zwischen beiden Variablen eine positive Korrelation besteht, liegt die erste Hauptkomponente im 45-Grad-Winkel zu den Ursprungsachsen x_1 und x_2. Wenn die Varianz ungleich wäre, würde die erste Hauptkomponente näher an der Achse mit der größeren Varianz liegen.

Das Beispiel zeigt, dass die Ermittlung der Hauptkomponenten von der Varianz der ursprünglichen x-Variablen abhängt. Je größer die Varianz einer Variablen, desto

[5]Siehe für eine detaillierte Darstellung der mathematischen Grundlagen der Transformation Bortz und Schuster (2010, S. 397–400).

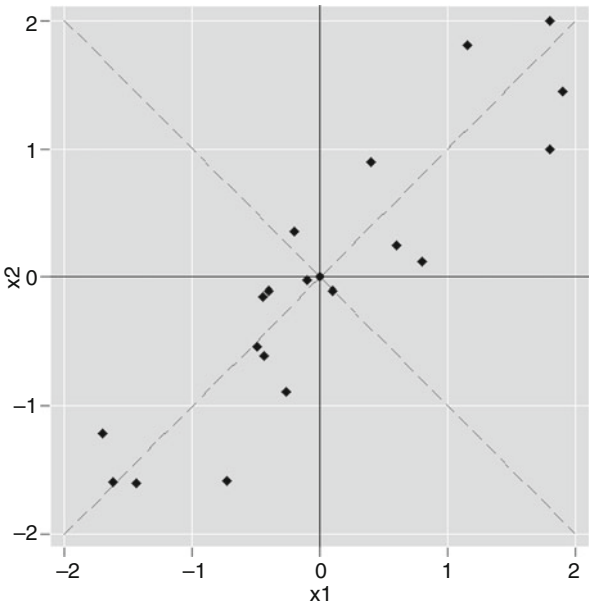

Abb. 2 Die Hauptkomponenten (y_1 und y_2) für zwei hochkorrelierte Variablen (r = 0,90) mit gleicher Varianz

wichtiger ist ihre Rolle bei der Identifikation der (ersten) Hauptkomponente. Sollten mehrere Variablen verwendet werden, die auf verschiedenen Skalen gemessen werden, ist es daher notwendig, dass diese vor der PCA standardisiert werden, so dass alle Variablen zum Finden der Komponentenlösung gleichermaßen beitragen. Als erster Schritt in der PCA wird daher oftmals die Korrelationsmatrix berechnet und für die Analysen herangezogen.

Im Beispiel beträgt die Varianz der ersten Hauptkomponente 1,90 und die der zweiten Hauptkomponente 0,10, d. h. beide Komponenten erklären zusammen die totale Varianz der beiden ursprünglichen Variablen, die bei 2 liegt. Dies zeigt, dass sich durch die PCA die totale Varianz der Variablen nicht verändert, sondern lediglich auf die Hauptkomponenten umverteilt wird. Die erste Komponente erklärt dabei 95 %, die zweite die restlichen 5 % der Gesamtvarianz. Von der Höhe der Korrelation zwischen beiden Variablen hängt auch ab, welchen Anteil der Varianz die erste Hauptkomponente erklärt – je stärker der Zusammenhang, desto höher ist auch der durch diese erklärte Anteil an der Gesamtvarianz und desto näher liegen die Beobachtungen an der ersten Hauptkomponente.

Auch bei mehr als zwei Variablen (dann mit mehr als zwei Hauptkomponenten) kann im mehrdimensionalen Raum ähnlich vorgegangen werden, was jedoch aus Gründen der schwierigeren Visualisierbarkeit hier nicht abgebildet ist.

Wie stellt man nun die Verbindung zwischen den *x*-Variablen und den Zielvariablen y her? In einer PCA gilt generell, dass eine Menge korrelierter *x*-Variablen in eine Menge unkorrelierter *y*-Komponenten umgeformt wird. Jede Hauptkomponente kann dabei als lineare Kombinationen der *x*-Variablen ausgedrückt werden (siehe Gl. 3).

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{21}x_2 + \ldots + a_{k1}x_k \\ y_2 &= a_{12}x_1 + a_{22}x_2 + \ldots + a_{k2}x_k \\ &\ldots \\ y_k &= a_{1k}x_1 + a_{1k}x_2 + \ldots + a_{kk}x_k \end{aligned} \tag{3}$$

Jede y-Komponente ist eine gewichtete Summe der x-Variablen. Es gibt keine Residualvariablen in der PCA, so dass die Unterschiede zwischen den x-Variablen vollständig auf die Unterschiede der Hauptkomponenten zurückgeführt werden. Die a_{ij}-Koeffizienten sind dabei die Gewichte für jede Variable i auf Komponente j. Das bedeutet, dass diese Koeffizienten die Ladungen der einzelnen Variablen auf die Hauptkomponente ausdrücken – je höher die Ladung auf die Komponente, desto höher ist der Zusammenhang zwischen der ursprünglichen Variablen und der extrahierten Hauptkomponente. Bei orthogonalen (also unkorrelierten) Hauptkomponenten können die Ladungen als Korrelation zwischen Variable und Hauptkomponente aufgefasst werden.

Die a_{ij}-Koeffizienten sind dabei beschränkt und können nicht beliebige Werte annehmen, da die verschiedenen y-Komponenten ja im rechten Winkel zueinanderstehen müssen. Daher müssen diese Koeffizienten die Bedingungen in Gl. 4 erfüllen.

$$\begin{aligned} &\sum_{i=1}^{k} a_{ij}^2 = 1 \qquad (j = 1,2,\ldots,k) \\ &\qquad\qquad \& \\ &\sum_{i=1}^{k} a_{ij}a_{im} = 1 \qquad (j \neq m; j = 1,2,\ldots,k; m = 1,2,\ldots,k) \end{aligned} \tag{4}$$

Im Umkehrschluss kann jede Variable auch wie folgt durch die Hauptkomponenten ausgedrückt werden (siehe Gl. 5).

$$x_k = \lambda_{k1}y_1 + \lambda_{k2}y_2 + \ldots + \lambda_{kk}y_k \tag{5}$$

Die Variable x_k ist also die Summe der jeweiligen Varianzen (λ_1, λ_2, …, λ_k) multipliziert mit den Hauptkomponenten (y_1, y_2,. . ., y_k).

Aus Gl. 1 ist bereits bekannt, dass die totale Varianz der y-Komponenten gleich der totalen Varianz der x-Variablen ist. Das heißt, dass sich durch die PCA die totale Varianz nicht verändert, sondern nur umverteilt wird. Die Komponenten werden in absteigender Reihenfolge der Wichtigkeit für die Erklärung extrahiert, so dass y_1 die maximale Varianz und den größten Anteil der totalen Varianz erklärt. Die zweite Komponente y_2 wird so extrahiert, dass sie die zweitgrößte Varianz aufweist und dabei den folgenden Einschränkungen unterliegt, so dass sie unkorreliert bzw. orthogonal zu y_1 ist, wie in Gl. 6 abgebildet.

$$\sum_{i=1}^{k} a_{i2}^2 = 1 \qquad \& \qquad \sum_{i=1}^{k} a_{i1}a_{i2} = 0 \tag{6}$$

Alle weiteren Hauptkomponenten werden gleichermaßen in absteigender Reihenfolge der erklärten Varianz ermittelt und müssen ebenfalls unkorreliert mit den vorhergehenden Komponenten sein.

Das mathematische Problem der PCA besteht daher darin, die a_{ij}-Koeffizienten so zu bestimmen, dass diese die erforderlichen Eigenschaften aus Gl. 6 erfüllen. Hierfür sind jedoch in allen gängigen Software-Paketen entsprechende Standard-Algorithmen verfügbar, die sich auf Basis der Kovarianz- oder meist der Korrelationsmatrix dieser Suche annehmen. Die Varianzen der Hauptkomponenten werden dabei in der Regel mit dem griechischen Buchstaben lamda notiert: $\lambda_1, \lambda_2, \ldots, \lambda_k$.

2.1.2 Voraussetzungen: Skalenniveau und Vorhandensein von Zusammenhängen zwischen den Variablen

Nach der Betrachtung der mathematischen Grundlagen leuchtet auch sofort ein, warum Zusammenhänge zwischen den ursprünglichen *x*-Variablen zwingend notwendig sind: Wenn eine Menge von *k* korrelierten Variablen durch einige wenige Hauptkomponenten mit dem Ziel ersetzt werden sollen, die Interpretation der Zusammenhänge zwischen den *x*-Variablen zu vereinfachen, ist es notwendig, dass überhaupt Korrelationen zwischen den Ausgangsvariablen bestehen. Wenn man nur Variablen zur Analyse heranzieht, die gar nicht korrelieren, könnte keine Kombination von Hauptkomponenten eine reduzierte Verdichtung der Informationen hervorbringen. Für diese Untersuchung gibt es zwei häufig verwendete Tests: der *Sphärizitätstest* nach Bartlett und das *Kaiser-Meyer-Olkin-Kriterium* (Bartlett 1950; Dziuban und Shirkey 1974). Der Spherizitätstest nach Bartlett überprüft mittels eines χ^2-Tests (mit k(k-1)/2 Freiheitsgraden) die Unabhängigkeit der Variablen, indem er die berechnete Korrelationsmatrix mit einer Korrelationsmatrix vergleicht, bei der alle Korrelationen gleich 0 sind. Ziel ist es, die Nullhypothese, dass die Variablen unkorreliert sind, zu verwerfen, so dass substanzielle Zusammenhänge zwischen den Variablen gesichert sind, die die Voraussetzung für die Durchführung der PCA bilden. Da der Bartlett-Test eine multivariate Normalverteilung der Daten voraussetzt, dies jedoch keine Voraussetzung für die PCA (jedoch für die EFA) ist, muss eine Ablehnung der Nullhypothese nicht unbedingt heißen, dass keine PCA durchgeführt werden kann. Einige Autoren sehen diesen Test auch bei einer PCA als sinnvoll an (siehe beispielsweise Wolff und Bacher 2010), andere eher für die EFA (Eid et al. 2015). Sicherlich kann es nicht schaden, sich die Ergebnisse anzusehen und gegebenenfalls entsprechend für oder gegen die Durchführung einer PCA zu argumentieren.

Das Kaiser-Meyer-Olkin-Kriterium (KMO) variiert zwischen Null und Eins und wird auf Basis der Interkorrelationen der Variablen berechnet, also der Korrelationen jeder Variable mit jeder anderen Variablen. Je näher der Wert an 1 ist, desto besser sind die Variablen für die Durchführung einer PCA geeignet. In verschiedenen Handbüchern finden sich unterschiedliche Schwellenwerte, ab wann die Daten die Durchführung einer PCA zulassen. Dziuban und Shirkey (1974) empfehlen einen Wert des KMO-Kriteriums größer 0,60. Kaiser, Meyer und Olkin selbst raten zu einem Wert von mindestens 0,50 (Cureton und D'Agostino 2009).

Zu Beginn dieses Beitrags wurde darauf verwiesen, dass die PCA wie auch die EFA eine Menge von k metrischen Variablen transformiert. Metrische Variablen sind jedoch gerade in den Sozialwissenschaften selten, insbesondere in der Einstellungsforschung. Hier dominieren vielmehr Variablen, die auf mehrstufigen Ratingskalen erhoben werden. Abhängig von der Anzahl der Stufen, auf denen diese Fragen beurteilt werden können, gibt es Wissenschaftler/innen, die diesen Skalen quasi-metrische Eigenschaften unterstellen. Hier wird angenommen, dass von Befragten bei bspw. sechsstufigen Ratingskalen von 0–5 die Abstände zwischen den einzelnen Beurteilungspunkten als gleich wahrgenommen werden (siehe beispielsweise Kenny 1986, S. 407). Es gibt jedoch auch kritische Stimmen, die darauf verweisen, dass die Verwendung von Ratingskalen bei der EFA zu einer Überidentifikation von Dimensionen, d. h. zur Identifikation von mehr Dimensionen als eigentlich vorhanden sind, führen kann (van der Eijk und Rose 2015).

Auch bei dichotomen Variablen kann auf Basis der φ-Koeffizienten eine PCA durchgeführt werden, falls die Merkmalsalternativen nicht so stark asymmetrisch verteilt sind, dass φ nicht 1 werden kann (siehe Bortz und Schuster 2010, S. 397). Generell empfiehlt sich der Blick in Zeitschriftenartikel des jeweiligen Fachgebiets, um herauszufinden, welche Verfahren für dichotome und rating-skalierte Items verwendet werden.[6] In diesem Beitrag werden sechs- bis elfstufige Ratingskalen herangezogen, die als quasi-metrisch betrachtet werden.

Zur besseren Veranschaulichung werden Daten aus der 13. Welle des GESIS-Panels (August-Oktober 2015) verwendet, das jeweils für verschiedene Statements zu Wertorientierung auf einer 6-stufigen Ratingskala erhebt, wie ähnlich eine Person, die diese Werte vertritt, der eigenen Person ist (1 „Ist mir überhaupt nicht ähnlich“ bis 6 „Ist mir sehr ähnlich“) (siehe auch den Beitrag von Schlipphak und Isani in diesem Band). Die Korrelationsmatrix für die paarweisen Korrelationen zwischen den verschiedenen Wertorientierungen ist in Tab. 1 dargestellt. Es zeigt sich, dass die Korrelationen zwischen den einzelnen Aussagen stark und zwischen –0,03 (Toleranz und starker Staat) und 0,42 (Toleranz und Gleichbehandlung) variieren. Der Bartlett-Test auf Spherizität[7] ergibt, dass die Nullhypothese mit sehr hoher Sicherheit verworfen werden kann, d. h. die Variablen sind nicht vollständig unkorreliert (χ^2 (10) = 1821,7, p <0,001). Das KMO-Kriterium liegt bei 0,61 und damit gerade über den in der Literatur vorgeschlagenen Grenzwerten. Eine PCA kann daher durchgeführt werden.

[6]Dabei stehen noch andere Berechnungsverfahren zur Verfügung: Für ordinale Variablen kann beispielsweise auch das Verfahren der polychorischen PCA bzw. EFA angewendet werden, das beispielsweise in Stata 14 über das Ado „polychoricpca“ verfügbar ist. Bei dichotomen Variablen kann auf die logistische PCA zurückgegriffen werden, die beispielsweise als Package für R zur Verfügung steht (Landgraf und Lee 2015).

[7]Dieser kann standardmäßig bei SPSS 22 über/PRINT=KMO angefordert werden, für Stata 14 ist das Ado „factortest“ verfügbar.

Tab. 1 Paarweise Korrelationen zwischen den Wertorientierungen (Pearsons r)

	Toleranz	Gleich-behandlung	Gesetze befolgen	Tradition	Starker Staat
Toleranz gegenüber vielen verschiedenen Menschen	1				
Dass alle Menschen gerecht behandelt werden	0,42	1			
Alle Gesetze befolgen	0,12	0,25	1		
Traditionelle Werte und Überzeugungen bewahren	0,07	0,08	0,27	1	
In einem starken Staat leben	−0,03	0,15	0,33	0,30	1

ZA5665, GESIS-Panel 2015 (13. Welle), N = 3496, „Wir beschreiben Ihnen nun kurz verschiedene Personen. Bitte lesen Sie jede Beschreibung durch und denken Sie darüber nach, inwieweit Ihnen die Person ähnlich oder nicht ähnlich ist. Bitte geben Sie an, wie ähnlich Ihnen die beschriebene Person ist." Einzelne Variablen eingeleitet durch „Es ist ihm/ihr wichtig, dass . . ."

2.2 Bestimmung der Komponentenanzahl

Ziel der PCA ist die Vereinfachung der Interpretation der Zusammenhänge zwischen den *x*-Variablen durch eine Interpretation der Datenstruktur auf Basis der wichtigsten Hauptkomponenten. Die Hauptkomponenten sollten dabei auf bestimmte Inhalte zurückgeführt werden können, denn nur dann sind die Hauptkomponenten sinnvoll zu interpretieren. Tab. 2 enthält die Ladungen der einzelnen Variablen auf die fünf Hauptkomponenten aus unserem Beispiel zu Wertorientierungen. Um die einfachere Untersuchung der Zusammenhänge zu erreichen, ist es jetzt notwendig, die Zahl der Hauptkomponenten zu reduzieren. Dabei stehen die Ziele Sparsamkeit (so wenige Komponenten wie möglich) und Erklärungskraft (so viele Informationen wie möglich) miteinander in Konflikt – es sollen also so wenige Komponenten wie nötig unter Beibehaltung so vieler Informationen wie möglich erhalten bleiben. Die aus Gl. 5 bekannte Formel für die vollständige Erklärung der einzelnen Variablen durch die Hauptkomponenten wird durch eine solche Reduktion angepasst (siehe Gl. 7). Der Term R bezeichnet nun die Varianz, die durch die Hauptkomponenten erklärt wird, die nicht mehr Teil der Lösung sind. R ist also das Residual. Es benennt die Varianz der Hauptkomponenten, die man nicht weiterverfolgt.

$$x_k = \lambda_{k1} y_1 + \lambda_{k2} y_2 + \ldots + R \tag{7}$$

Es gibt verschiedenen Kriterien, nach denen die Anzahl der Hauptkomponenten ausgewählt werden kann. Diese Kriterien unterscheiden sich in ihrer Komplexität und ihrer Verbindung zur inhaltlichen Interpretation. Dabei werden in der Forschung am häufigsten das *Kaiser-Kriterium* und der *Screetest* herangezogen, auch wenn beide Maße mit Kritik behaftet sind (siehe beispielsweise Fabrigar et al. 1999). Oftmals ist in der Praxis eine Kombination beider Maße verbreitet, so dass die Reduktion auf Basis von Kaiser-Kriterium und Screetest erfolgt.

Einige Autoren (siehe beispielsweise Bartholomew et al. 2002, S. 122) schlagen vor, die ersten n Komponenten beizubehalten, die einen hohen Anteil der Gesamt-

Tab. 2 Komponentenladungen der einzelnen Items auf die Hauptkomponenten

	Hauptkomponente				
	1	2	3	4	5
Toleranz gegenüber verschiedenen Menschen	0,35	0,64	0,31	0,06	0,61
Dass alle Menschen gerecht behandelt werden	0,47	0,49	−0,25	0,21	−0,66
Alle Gesetze befolgen	0,53	−0,18	−0,32	−0,75	0,13
Traditionelle Werte und Überzeugungen bewahren	0,42	−0,37	0,79	0,01	−0,26
In einem starken Staat leben	0,45	−0,44	−0,35	0,62	0,32

ZA5665, GESIS-Panel 2015 (13. Welle), N = 3496

varianz von etwa 70–80 Prozent erklären. Das ist nur eine grobe Daumenregel, aber einfach handzuhaben. Der Forscher oder die Forscherin bestimmt eine bestimmte notwendige Mindestvarianz und sucht dann die Anzahl von Hauptkomponenten, die zusammen diese Varianz erklären. Eine solche Vorgehensweise ist jedoch abzulehnen, da sie nur auf einer ungefähren Daumenregel basiert und der oftmals geäußerten Kritik an der PCA, dass sie einen großen Ermessensspielraum lasse, entspricht.

Häufig werden Kriterien herangezogen, die auf den Eigenwerten der Hauptkomponenten (also deren Anteil an der erklärten Gesamtvarianz) basieren. Das Kaiser-Kriterium (Kaiser und Dickman 1959, nicht zu verwechseln mit dem Kaiser-Meyer-Olkin-Kriterium aus Abschn. 2.1) besagt, dass nur die Hauptkomponenten verwendet werden sollen, deren Eigenwert bei über 1 liegt oder gleich 1 ist. Bei einem Wert größer/gleich 1 weist eine Hauptkomponente mindestens genauso viel Varianz auf wie eine der einzelnen Variablen, die zur Analyse ja in der Korrelationsmatrix standardisiert wurden und somit eine Varianz von 1 aufweisen. Hauptkomponenten müssen bezüglich ihrer Varianz also genau so viel „können" wie die ursprünglichen standardisierten Variablen.

Wenige andere Autoren setzen diesen Wert mit 0,7 etwas niedriger an. Deren Kritik richtete sich gegen die starre Grenze von 1, da so Komponenten mit einem Eigenwert von 1,0 berücksichtigt, solche mit einem Eigenwert von 0,99 jedoch nicht beibehalten würden, was willkürlich erscheinen mag. Auch führt die Verwendung des Kaiser-Kriteriums bei Analysen mit sehr vielen x-Variablen dazu, dass das Kriterium der Sparsamkeit nicht mehr erfüllt ist, da mehr Komponenten beibehalten werden als nötig (siehe für eine überblicksartige Darstellung Fabrigar et al. 1999, S. 278).

In unserem Beispiel in Tab. 3 weisen die ersten beiden Komponenten einen Eigenwert von größer/gleich 1 auf. Die erste Komponente verfügt über einen Eigenwert von 1,80, d. h. sie weist etwa so viel Varianz wie 1,8 Variablen auf. Diese ersten beiden Komponenten erklären knapp 61 % der Gesamtvarianz.

Ein weiteres Auswahlverfahren basiert ebenfalls auf den Eigenwerten der Komponenten: Der Scree-Test, der auf Cattell (1966) zurückgeht, stellt grafisch den Verlauf der Eigenwerte dar. Da die Komponenten ja in absteigender Reihenfolge des von ihnen erklärten Anteils an der Gesamtvarianz extrahiert werden, verläuft der Screeplot ebenfalls absteigend. Ausgewählt werden die Faktoren, die sich von den anderen Faktoren deutlich abheben. Sie sollen also vor dem Faktor liegen, an dem die Eigenwerte langsamer abnehmen. Bildlich gesprochen sucht man also nach einem „Knick" im Plot und wählt die Anzahl der Komponenten aus, die vor dem

Tab. 3 Eigenwerte und erklärte Varianz der vollständigen Hauptkomponentenlösung für die Wertorientierungen

Komponente	Eigenwert (Varianz)	Erklärte Varianz (%)	Kumulativ (%)
1	1,80	0,36	0,36
2	1,26	0,25	0,61
3	0,76	0,15	0,76
4	0,65	0,13	0,89
5	0,53	0,11	1

ZA5665, GESIS-Panel 2015 (13. Welle), N = 3496

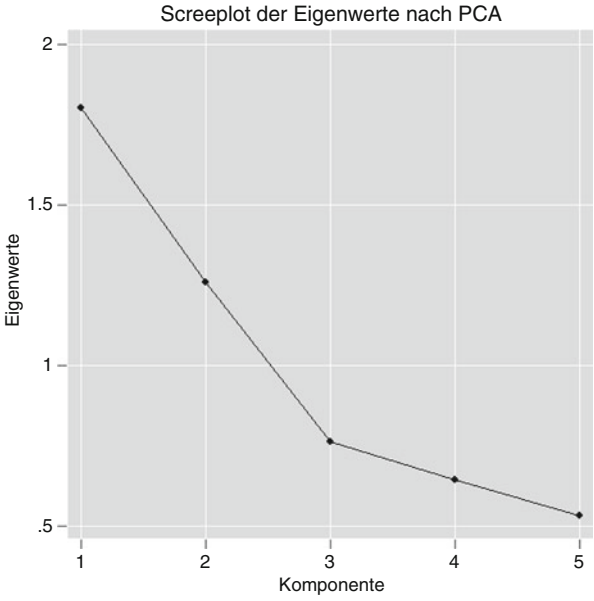

Abb. 3 Screeplot der Eigenwerte für die PCA der Wertorientierungen

Knick liegen. Andere Begriffe, die zur Beschreibung des Kurvenverlaufs verwendet werden, sind „Ellenbogen“ oder „Felsschutt“/„Geröll“ gemäß dem Namen des Tests: Man stellt sich vor, dass es sich bei der steil absteigenden Linie im Plot um eine Felswand handelt, an deren Fuß sich der „Schutt“ ansammelt. Kritisiert wird an diesem Kriterium, dass nicht immer ganz eindeutig ist, wo der Knick im Screeplot liegt und dass es so von subjektiven Einschätzungen abhängen kann, welche Anzahl von Hauptkomponenten beibehalten wird (siehe Fabrigar et al. 1999, S. 278).

In Abb. 3 ist der Screeplot für unser Beispiel mit den Wertorientierungen dargestellt. Hier ist klar zu sehen, dass ab Komponente 3 der Verlauf der Eigenwerte weniger steil ist als zuvor. Wir wählen also zwei Hauptkomponenten, da Komponente 2 die letzte vor dem Knick ist. Auf Basis der Eigenwerte und des Screeplots behalten wir zwei Hauptkomponenten bei, die zusammen etwa 3/5 der Gesamtvarianz erklären. Folglich bleiben durch diese Reduktion etwa 40 % der Gesamtvarianz

Tab. 4 Eigenwerte und erklärte Varianz der reduzierten Hauptkomponentenlösung für die Wertorientierungen

	Komponente		
	1	2	Unerklärte Varianz
Toleranz gegenüber verschiedenen Menschen	0,35	0,64	0,28
Dass alle Menschen gerecht behandelt werden	0,47	0,49	0,31
Alle Gesetze befolgen	0,53	−0,18	0,45
Traditionelle Werte und Überzeugungen bewahren	0,42	−0,37	0,51
In einem starken Staat leben	0,45	−0,44	0,39

ZA5665, GESIS-Panel 2015 (13. Welle), N = 3496

unerklärt. Es ist möglich, sich auch die unerklärte Varianz für jede der x-Variablen ausgeben zu lassen.

In Tab. 4 sind die reduzierte Lösung und die unerklärte Varianz abgebildet. Dabei ist zu sehen, dass sich die unerklärte Varianz nicht über alle Variablen gleichverteilt, sondern besonders die Aussage zu den traditionellen Werten betrifft, während sie bei der Aussage zur Toleranz nur etwa halb so hoch ist. Hier könnte man durchaus argumentieren, dass sich eine dritte Hauptkomponente rechtfertigen lässt, die die unerklärte Varianz für diese Aussage senkt. Solange aber etwa die Hälfte der Varianz der jeweiligen Variablen erklärt wird und Screeplot sowie Kaiser-Kriterium die gleiche Anzahl stützen, wird bevorzugt mit zwei Komponenten weitergearbeitet. Die Reduktion der Komponenten in unserem Beispiel zeigt bereits gut auf, warum der PCA und auch der EFA oftmals vorgeworfen wird, eine subjektive Methode zu sein. Daher ist es wichtig, die Bestimmung der Komponentenanzahl transparent zu halten, hierfür mehr als ein Maß zu verwenden (empfohlen werden Screetest + Kaiser-Kriterium) und mögliche andere Mengen für die Anzahl der Komponenten zumindest zu diskutieren, wie es im Abschnitt darüber bereits geschehen ist.

2.3 Rotation der Komponenten zur besseren Interpretation

Die Komponentenladungen beschreiben die Zusammenhänge zwischen den *x*-Variablen und der jeweiligen Hauptkomponente.[8] In Tab. 5 betrachten wir zuerst die unrotierte Ladungsmatrix in Spalte 2 und 3. Alle fünf Variablen laden dabei

[8]Bartholomew et al. (2002, S. 167) verweisen darauf, dass bei einer PCA die erste Hauptkomponente immer die gleichen Ladungen aufweist, egal, wie viele Komponenten beibehalten werden, während dies bei einer Faktorenanalyse (siehe Abschn. 4) nicht der Fall ist. Daher solle die Rotation der Ladungen für eine PCA nicht vorgenommen werden, da diese mit einem Informationsverlust verbunden ist, während dies bei der EFA nicht zutrifft. Dennoch erscheint der Zugewinn durch die erleichterte Interpretation, die ja das Ziel der PCA ist, ausreichend, um eine Rotation der Komponenten durchzuführen, wie dies auch durch andere Autoren vorgeschlagen wird.

Tab. 5 Komponentenladungen vor und nach Rotation für die reduzierte Zwei-Komponenten-Lösung

	Unrotierte Lösung		Varimax-Rotation (orthogonal)		Promax-Rotation (schiefwinklig, p = 3)	
	1	2	1	2	1	2
Toleranz gegenüber verschiedenen Menschen	0,35	0,64	−0,09	0,72	−0,10	0,72
Dass alle Menschen gerecht behandelt werden	0,47	0,49	0,09	0,67	0,08	0,67
Alle Gesetze befolgen	0,53	−0,18	0,53	0,17	0,53	0,16
Traditionelle Werte und Überzeugungen bewahren	0,42	−0,37	0,56	−0,05	0,56	−0,05
In einem starken Staat leben	0,45	−0,44	0,62	−0,09	0,62	−0,10
Eigenwerte	1,80	1,26	1,61	1,45	1,60	1,46

ZA5665, GESIS-Panel 2015 (13. Welle), N = 3496

positiv auf die erste Hauptkomponente – die erste Hauptkomponente scheint daher etwas zu repräsentieren, das allen Variablen zugrunde liegt, während die zweite Hauptkomponente Unterschiede zwischen den *x*-Variablen verdeutlicht: Die ersten beiden Variablen besitzen positive, die anderen drei Variablen negative Ladungen.

Wenn mehr als eine Hauptkomponente beibehalten wird und die Komponenten in inhaltlicher Weise interpretiert werden sollen, ist es sinnvoll eine Rotation der Koordinatenachsen durchzuführen. Dabei werden zur besseren Deutung die Komponenten und ihre Ladungen so transformiert, dass sie bestimmten Kriterien entsprechen. Diese Rotation ist möglich, da die erhaltene unrotierte Lösung eine Anfangslösung darstellt, jedoch weitere Lösungen möglich sind, die die Struktur der Daten mit identischer Anpassungsgüte wiedergeben und inhaltlich sinnvoller erscheinen. Die Rotation zielt in der Regel auf eine Einfachstruktur nach Thurstone (1947) ab, bei der die *x*-Variablen möglichst nur auf eine Komponente hoch und auf die andere/n Komponente/n möglichst niedrig laden sollen, d. h. eine Variable kann am besten immer nur durch eine Komponente ersetzt werden. Es gibt sehr viele unterschiedliche Rotationsverfahren, die sich auf zwei grundlegende Ideen verteilen: orthogonale Rotation und schiefwinklige Rotation (oblique). Während bei der orthogonalen Rotation die Komponenten weiterhin unkorreliert sind (und im rechten Winkel zueinander stehen), lockert die oblique Rotation das Vorgehen der PCA und erlaubt die Korrelation der Komponenten. Die Auswahl des Verfahrens basiert dabei in der Regel auf der Interpretation der Komponenten: Geht man davon aus, dass diese nicht zusammenhängen? In diesem Fall wird das Verfahren der orthogonalen Rotation wesentlich häufiger eingesetzt. Auf die schiefwinklige Rotation wird meist nur zurückgegriffen, wenn theoretische Überlegungen, v. a. bei der explorativen Faktorenanalyse (siehe Abschn. 4), eine Kor-

relation der Komponenten/Faktoren wahrscheinlich erscheinen lassen.[9] Es gibt keinen Test, der bei der Auswahl des Rotationsverfahrens unterstützt, sodass die Auswahl hier in der Hand des Forschenden liegt.

Im Folgenden werden die zwei bekanntesten Rotationsverfahren der Varimax-Rotation (orthogonal) und Promax (oblique) dargestellt, die in jedem gängigen Statistikprogramm enthalten sind.[10] Das Varimax-Rotationsverfahren ist sicherlich das am weitesten verbreitete und geht auf Kaiser (1958) zurück. Hier wird nach einer Einfachstruktur für die Komponentenladungen gesucht, indem die Varianz der quadrierten Ladungen maximiert wird. Die varimax-rotierte Lösung ist in Tab. 5, Spalte 4 und 5 zu sehen. Durch die Rotation haben sich die Eigenwerte der Komponenten, nicht aber die Summe der Eigenwerte, geändert. Das Ziel der Einfachstruktur wurde dabei erreicht: Jetzt laden die letzten drei Variablen auf die erste Komponente und kaum auf die zweite, während die ersten beiden Variablen hoch auf die zweite Komponente und kaum auf die erste Komponente laden. Dies lässt sich gut anhand der Variable „In einem starken Staat leben" demonstrieren, die in der Anfangslösung noch gleichermaßen mit der ersten und zweiten Hauptkomponente korreliert war, jetzt jedoch nur noch hauptsächlich durch die erste Komponente erklärt wird.

Bei der Promax-Rotation wird die Bedingung der Orthogonalität der Hauptkomponenten gelockert und Korrelationen zwischen den Komponenten, die von Null verschieden sind, sind nun zulässig. Dabei ist es jedoch wichtig darauf zu achten, dass die Korrelationen zwischen den Komponenten nicht zu hoch sind, da sich die Komponenten sonst überlappen und eine sinnvolle Interpretation schwierig wird.[11] Die Ergebnisse der Promax-Rotation sind in Tab. 5, Spalte 6 und 7 zu sehen. Auch hier wurde eine Einfachstruktur erreicht. Die Komponentenladungen unterscheiden sich zudem kaum von der Lösung für das Varimax-Verfahren.

Für die Deutung der Komponenten werden nun die Ladungen herangezogen und von den Gemeinsamkeiten der Variablen, die jeweils hoch auf eine Komponente laden, auf die inhaltliche Interpretation der Komponente abstrahiert. Als Faustregel, ab wann von hohen Ladungen gesprochen werden kann, gibt es verschiedene Größen: Einige Autoren beginnen bei 0,3, andere erst bei 0,5. Wolff und Bacher (2010, S. 346) empfehlen, dass die Variablen, die zur Benennung herangezogen

[9]Für die praktische Durchführung empfiehlt sich bei der PCA zuerst die Durchführung der orthogonalen Rotation. Anschließend werden die Komponenten interpretiert, indem sie benannt werden. Sollte nach der Benennung eine Korrelation der Komponenten wahrscheinlich erscheinen, sollte die Rotation erneut mit einem obliquen Verfahren durchgeführt werden, das dann angemessener wäre.

[10]Für eine weiterführende Darstellung wird auf Eid et al. (2015) verwiesen.

[11]Bei schiefwinkligen Rotationen kann der Nutzer in der Regel einen Wert p für die zulässige Höhe der Korrelationen zwischen den Komponenten angeben, der nach Empfehlung einiger Autoren (z. B. Lawley und Maxwell 1971) nicht über 4 liegen sollte, da die Komponenten sonst zu stark korreliert sind. Der Standardwert für das Promax-Verfahren ist bei SPSS p=4 und bei Stata p=3. Beim Promax-Verfahren erhält der Nutzer zuerst eine Struktur- und dann eine Mustermatrix, wobei letztere die eigentlich zu interpretierenden Komponentenladungen enthält.

werden, auf keinen Fall mit höher als 0,3 auf eine andere Komponente laden sollten. Diese inhaltliche Deutung ist ein wichtiger Schritt zur Nutzung der Ergebnisse. Sie ist jedoch auch der Schritt, der oftmals Kritik an der PCA auslöst: Da die inhaltliche Deutung durch den Nutzer selbst vorgenommen wird, besteht hier ein Interpretationsspielraum des Nutzers. Dabei ist es notwendig, dass die Interpretation der Komponenten transparent erfolgt und sorgfältig argumentiert wird.

Die inhaltliche Interpretation in unserem Beispiel nehmen wir auf Basis der Ergebnisse nach der Rotation vor, da diese die bestmögliche Lösung zur Interpretation darstellt. Von allen möglichen Lösungen, die es für die Variablenladungen gibt, eignet sich die rotierte Lösung durch ihre Annäherung an die Einfachstruktur (d. h. eine Variable lädt möglichst nur auf eine Komponente) am ehesten zur Interpretation. Die erste Hauptkomponente umfasst Aussagen zu Recht, Ordnung und Tradition, während die zweite Hauptkomponente Aussagen zusammenfasst, die sich auf Freiheit und Gerechtigkeit beziehen. Man könnte Komponente 1 daher beispielsweise mit „Konservative Werte“ und Komponente 2 mit „Progressive Werte“ bezeichnen. Alternativ wäre auch eine andere Benennung denkbar, beispielsweise „Recht und Ordnung“ für Komponente 1 und „Toleranz“ für Komponente 2. Dabei lädt die Aussage „Toleranz gegenüber verschiedenen Menschen“ am höchsten auf die zweite, die Aussage „In einem starken Staat leben“ am höchsten auf die erste Komponente. Es gibt keine Aussage, die auf beide Komponenten lädt. Anschließend ist es möglich, sich für jeden Befragten den Komponentenwert (*component score*) für die einzelnen Komponenten ausgegeben und abspeichern zu lassen. Bei der PCA wird er als gewichtete Summe der x-Variablen nach Gl. 3 berechnet.

3 Politikwissenschaftliche Anwendungen

Auf Basis der Informationen aus Abschn. 2 werden nun zwei beispielhafte Anwendungen dargestellt, die das Analysepotenzial der PCA verdeutlichen sollen.

In Beispiel 1 wird eine Sympathiebatterie in Bezug auf Parteien aus der *German Longitudinal Election Study* (ZA5701) verwendet. Für sechs Parteien wurde auf 11-stufigen Sympathieskalometern erhoben, was die Befragten von der jeweiligen Partei halten. Die Korrelationsmatrix für die paarweisen Korrelationen ist in Tab. 6 dargestellt. Es zeigt sich, dass die Korrelationen zwischen den Sympathien für die einzelnen Parteien stark und von 0,35 (CDU und LINKE) bis 0,86 (CDU und CSU) variieren.

Der Bartlett-Test auf Spherizität ergibt, dass die Nullhypothese mit sehr hoher Sicherheit verworfen werden kann, d. h. die Variablen sind nicht vollständig unkorreliert ($\chi^2(15) = 3981{,}2$, $p < 0{,}001$). Das KMO-Kriterium liegt bei 0,66 und damit über den in der Literatur vorgeschlagenen Werten. Eine PCA kann daher durchgeführt werden (Tab. 7).

Auf Basis der Eigenwerte wählen wir die ersten beiden Komponenten aus. Auch der Screeplot (nicht abgebildet) hat einen „Ellenbogen“ oder Knick nach der zweiten Komponente. Anders als im Beispiel aus Abschn. 2 laden nicht alle Variablen mit dem gleichen Vorzeichen auf die erste Komponente, d. h. bereits ohne Rotation

Tab. 6 Paarweise Korrelationen zwischen den Parteisympathien (Pearsons r)

		CDU	CSU	SPD	FDP	DIE LINKE	GRÜNE
Sympathie mit	*CDU*	1,00					
	CSU	0,86	1,00				
	SPD	0,07	−0,05	1,00			
	FDP	0,52	0,56	0,08	1,00		
	DIE LINKE	−0,35	−0,35	0,21	−0,12	1,00	
	GRÜNE	−0,05	−0,11	0,45	0,06	0,33	1,00

ZA5701, GLES Nachwahlbefragung 2013, N = 1908, Was halten Sie von dieser Partei? −5 „Halte überhaupt nichts“ bis +5 „Halte sehr viel von dieser Partei“

Tab. 7 Komponentenladungen der einzelnen Items auf die Hauptkomponenten und Eigenwerte, Parteisympathien

		Hauptkomponente					
		1	2	3	4	5	6
Sympathie mit	*CDU*	0,570	0,149	−0,055	0,009	0,417	−0,690
	CSU	0,583	0,084	0,071	−0,030	0,366	0,716
	SPD	−0,066	0,616	−0,504	0,592	−0,062	0,089
	FDP	0,437	0,266	0,444	0,081	−0,729	−0,056
	DIE LINKE	−0,346	0,356	0,728	0,256	0,396	−0,024
	GRÜNE	−0,138	0,628	−0,103	−0,759	−0,003	0,018
Eigenwerte		2,50	1,63	0,72	0,53	0,49	0,14
Erklärte Varianz (%)		0,42	0,27	0,12	0,09	0,08	0,02

ZA5701, GLES Nachwahlbefragung 2013

besteht hier eher eine Einfachstruktur mit Ausnahme der Ladungen der Sympathien für die FDP und DIE LINKE. Anschließend führen wir eine Varimax-Rotation durch, da wir annehmen, dass die Komponenten nicht korreliert sind (siehe Tab. 8). Hierdurch erreichen wir eine Annäherung an eine Einfachstruktur, bei der nur noch DIE LINKE eine Ladung von nahe 0,30 auf eine weitere Komponente hat. Im Gegensatz zur unrotierten Lösung jedoch weist die Sympathie mit DIE LINKE jetzt einen höheren Zusammenhang mit der zweiten Hauptkomponente auf. Eine erste Interpretation könnte darin bestehen, dass sich die erste Komponente mit Sympathien für die bürgerlich-liberalen Parteien und die zweite Komponente mit Sympathien für die Mitte-Linken-Parteien beschreiben lässt. Je nach Vorstellung, welche ideologische Position die Parteien vertreten, lässt sich natürlich auch eine andere Beschreibung finden. Dabei besteht besonders für Variable „Sympathie mit DIE LINKE“ eine negative Ladung auf die erste Hauptkomponente. Das heißt, dass die Sympathie mit DIE LINKE negativ mit den Sympathien für das bürgerlich-liberale politische Lager korreliert. Diese Analyse der Konfiguration des deutschen Parteiensystems nach den subjektiven Parteiensympathien der Wähler zeigt also Anzeichen, dass sich die großen Parteien in zwei Lager strukturieren lassen.

Tab. 8 Rotierte Komponentenladungen für das Zwei-Hauptkomponenten-Modell, Parteisympathien

	Varimax-Rotation Komponente	
	1	2
CDU	0,59	0,01
CSU	0,59	−0,05
SPD	0,08	0,61
FDP	0,49	0,16
DIE LINKE	−0,25	0,43
GRÜNE	0,01	0,64

ZA5701, GLES Nachwahlbefragung 2013

Für Beispiel 2 wird auf Daten des ALLBUS 2016 (ZA5250) zurückgegriffen. Hier wurden verschiedene Aussagen zur Frage erhoben, welche Eigenschaften notwendig sind, damit eine ausländische Person als „echte/r" Deutsche/r gelten kann. Es konnte auf 7-stufigen-Ratingskalen (1 „überhaupt nicht wichtig" bis 7 „sehr wichtig") beurteilt werden.

Für die Analyse wurden die fünf Items „Verbundenheit zu Deutschland", „Gut Deutsch sprechen", „Westliche Werte teilen", „Mindestens ein Elternteil muss Deutsch sein" und „In Deutschland geboren sein" ausgewählt. Der Bartlett-Test auf Spherizität ergibt, dass die Daten nicht unkorreliert sind ($\chi^2(10) = 1447{,}7$, $p < 0{,}001$). Das KMO-Kriterium liegt über 0,58. Eine PCA kann daher durchgeführt werden.

Auf Basis der Eigenwerte und des Screetests (nicht abgebildet) werden zwei Hauptkomponenten ausgewählt, die zusammen 68 % der gesamten Varianz erklären. Um eine Einfachstruktur und zwei unkorrelierte Komponenten zu erhalten, werden die Daten orthogonal rotiert. Die Ladungen für die unrotierte und die rotierte (Varimax-)Lösung sind in Tab. 9 abgebildet. Dabei lassen sich nach der Rotation zwei klare Muster identifizieren: Die beiden Variablen, die eine eher ethnische Deutung der Bedingungen erkennen lassen, zu denen jemand als „echte/r" Deutsche/r wahrgenommen wird, laden hoch auf die zweite Komponente („Mindestens ein Elternteil Deutsch" und „In Deutschland geboren"). Die anderen drei Variablen, die eher eine staatsbürgerliche/civic bzw. prozedurale Deutung des Deutschseins abbilden („Verbundenheit zu Deutschland", „Westliche Werte teilen" und „Gut Deutsch sprechen") laden hoch auf die erste Komponente. Wir könnten Komponente 2 also mit „ethnisches Konzept" des Deutschseins und Komponente 1 mit „civic Konzept" des Deutschseins beschreiben.

Für eine weiterführende Analyse lassen wir für jede/n Befragte/n den Wert für die Komponente berechnen, der sich nach Gl. 3 ergibt.[12] Dabei werden die rotierten Ladungen zur Berechnung des Komponentenwertes verwendet. Im Unterschied zu einer Mittelwert- oder Summenskala, bei der wir beispielsweise für das ethnische

[12]Dies erfolgt in Stata 14 mit dem Befehlt predict comp1 comp2, score.

Tab. 9 Unrotierte und rotierte Komponentenladungen für das Zwei-Hauptkomponenten-Modell, Identitätsdimensionen

	Unrotierte Lösung		Varimax-Rotation	
	1	2	1	2
Verbundenheit zu Deutschland	0,47	−0,35	0,58	−0,03
Gut Deutsch sprechen	0,48	−0,36	0,59	−0,03
Westliche Werte teilen	0,50	−0,25	0,55	0,07
Mind. 1 Elternteil Deutsch	0,38	0,60	−0,02	0,71
In Deutschland geboren	0,40	0,57	0,01	0,70
Eigenwert	1,82	1,58	1,75	1,66
Erklärte Varianz	0,35	0,33	0,36	0,32

ZA52050, ALLBUS 2016, N = 1512; zufälliger Split für die Hälfte der Befragten

Konzept des Deutschseins die Variablen 4 und 5 aufaddieren würden, enthält der Komponentenwert zusätzlich noch die (geringen) negativen Ladungen der anderen Variablen in der Analyse auf diese Komponente (die Korrelation zwischen Komponentenwert und Summenskala ist dabei mit r = 0,99 sehr hoch). Die beiden so erhaltenen Variablen *civic* und *ethnic* können wir nun für weiterführende Analysen heranziehen. So können wir beispielsweise untersuchen, wie eine Unterstützung für das ethnische Konzept mit des Deutschseins zusammenhängt: Der Komponentenwert für die zweite Komponente (*ethnic*) ist negativ mit der Unterstützung für ein generelles kommunales Wahlrecht für Ausländer/in moderatem Maße korreliert (r = −0,25, p <0,001). Wer über eine ethnische Konzeption des Deutschseins verfügt, die im Wesentlichen auf der Tatsache beruht, ob eine Person selbst oder ihre Eltern in Deutschland geboren sein sollten, lehnt ein kommunales Wahlrecht für Ausländer eher ab. Die Unterstützung eines *civic*-Konzeptes der anderen Fragen ist ebenfalls negativ mit dem kommunalen Wahlrecht korreliert, wenn auch in geringerem Ausmaß, (r = −0,15, p <0,001).

4 Die explorative Faktorenanalyse

Die EFA gehört zu einer Familie von Verfahren, die sich mit latenten Variablen beschäftigen. Sozialwissenschaftliche Beispiele für latente Variablen, die wir nicht direkt messen können, sind Prestige, Macht, Rechtsextremismus oder politische Einstellungen. Wenn latente Variablen gemessen werden sollen, erfolgt dies indirekt über manifeste Variablen, auch Indikatorvariablen genannt. Oftmals wird die EFA mit dem Vorgehen bei der Regressionsanalyse (siehe auch den Beitrag von Seng in diesem Band) verglichen: Im Regressionsmodell wird eine abhängige Variable auf eine oder mehrere unabhängige Variablen regressiert, d. h. zurückgeführt und durch diese erklärt. Ähnlich verhält es sich auch bei der EFA. Hier besteht die Beziehung zwischen manifesten Variablen und latenten Faktoren: Es wird versucht eine Menge von manifesten Variablen auf nicht direkt beobachtbare latente Faktoren zurückzuführen.

Wenn im Vorfeld bereits Annahmen darüber bestehen, welche Beziehungen zwischen Faktoren und manifesten Variablen bestehen, so bietet sich statt der EFA sofort das Verfahren der konfirmatorischen Faktorenanalyse (CFA) an, bei dem ein vorher festgelegtes Messmodell mit den bestehenden Daten abgeglichen werden kann. Wenn diese Annahmen im Vorfeld nicht bestehen oder zumindest offen analysiert werden soll, auf welche zugrundeliegenden latenten Faktoren eine Menge von x manifesten Variablen zurückgeführt werden kann, bietet sich das Verfahren der EFA an.

4.1 Gemeinsamkeiten und Unterschiede von Hauptkomponentenanalyse und explorativer Faktorenanalyse

Die Begriffe PCA und EFA werden oftmals synonym benutzt. Dennoch unterscheiden sich beide Verfahren in ihren Grundannahmen: Die PCA ist eine deskriptive Analysetechnik, bei der eine Datenmatrix (meist eine Korrelationsmatrix) durch einige wenige Komponenten zusammengefasst wird und ohne weitere Grundannahmen auskommt. Die EFA hingegen ist eine inferenzstatistische Datenmodellierungstechnik, bei der bestimmte Annahmen erfüllt sein müssen und die eine Menge von x manifesten Variablen auf einige wenige latente Variablen (Faktoren y) zurückführt. Auch die EFA basiert auf der Erklärung der Korrelationen zwischen den Variablen. Hier stellt sich die Frage, ob diese Korrelationen auf gemeinsame zugrundeliegende Dimensionen zurückgeführt werden können, wobei die Ergebnisse anschließend oftmals zur Hypothesengenerierung verwendet werden.

Während die PCA lediglich die Reduktion der bestehenden Daten zum Ziel hat, soll bei der EFA versucht werden, die zugrunde liegenden latenten Dimensionen zu bestimmen. Möchten wir beispielsweise rechtsextreme Einstellungen messen, so können wir eine EFA mit einer Menge von Variablen durchführen, von denen wir annehmen, dass sie Rechtsextremismus messen. Es ist möglich, dass wir nur einen gemeinsamen oder mehrere latente Faktoren erhalten, sodass es sich um ein sogenanntes multidimensionales Konzept handelt. Dieses Vorgehen hat den Vorteil, dass wir am Ende nur noch wenige Dimensionen von Rechtsextremismus haben und nicht mehr eine große Menge von verschiedenen Items. Die Ergebnisse können daher auch zur Instrumentenreduktion verwendet werden, so dass nur die Items ausgewählt werden, die hoch und/oder eindeutig auf nur einen Faktor laden. Damit alle relevanten Dimensionen identifiziert werden können, ist es notwendig, dass manifeste Variablen vorhanden sind, die diese Dimensionen abbilden. Als Richtwert wird in der Literatur oftmals genannt, dass mindestens drei oder vier manifeste Variablen vorhanden sein müssen, um eine Dimension gut zu identifizieren (siehe etwa Kim und Mueller 1978).

Wie die PCA (siehe Gl. 5) basiert auch die EFA auf einer Zerlegung von p manifesten Variablen in eine Linearkombination von q gewichteten Faktoren (siehe Gl. 8). Dabei bezeichnen $y_1,\ldots, y_q$ die latenten Faktoren und $\alpha_{i1},\ldots, \alpha_{kq}$ die

Ladungen der Faktoren. α_{k0} bezeichnet den Achsenabschnitt, der in der Praxis für die Interpretation der Ladungen nicht relevant ist.

$$x_k = \alpha_{k0+}\alpha_{k1}y_1 + \alpha_{k2}y_2 + \ldots + \alpha_{kq}y_q + \varepsilon_i \qquad (i = 1, \ldots, p) \tag{8}$$

Die wesentliche Neuerung im Vergleich zur PCA ist dabei die Residualvariable ε_i. Während in der PCA noch davon ausgegangen wird, dass die gesamte Varianz der Variablen durch die (vollständige!) Hauptkomponentenlösung erklärt werden kann, zeigt sich hier, warum die EFA zu den Datenmodellierungsverfahren gezählt wird: Die Annahme der EFA ist, dass es sich um ein Messmodell handelt, bei dem jede manifeste Variable sowohl auf latente gemeinsame Faktoren als auch auf spezifische Faktoren zurückgeführt werden kann. Letztere setzen sich aus einem Messfehler bei Erhebung der Daten oder aus anderen Eigenschaften zusammen, die spezifisch für diese eine Variable sind und durch das Modell nicht erklärt werden können. Nimmt man an, dass die spezifischen Faktoren gleich 0 sind, d. h. die gesamte Varianz lässt sich durch die gemeinsamen Faktoren erklären, würde sich eine PCA aus der EFA ergeben.

Welche Konsequenzen hat dies nun für die praktische Analyse? Die Anteile an der Streuung der Variablen, die jeweils auf gemeinsame und spezifische Faktoren zurückgehen, müssen geschätzt werden. Der Begriff der Kommunalität (*commonality*)[13] einer (standardisierten) Variablen bezeichnet den Anteil der Varianz der Variablen, die durch die gemeinsamen latenten Faktoren erklärt wird. Je höher die Kommunalität, desto eher lässt sich die Variable mit den in der EFA erhaltenen Faktoren erklären und desto niedriger ist die Residualvariable ε_i, die durch das Modell nicht erklärt werden kann. Wenn wir Faktorwerte (entsprechend der Komponentenwerte aus Abschn. 3) berechnen, können diese im Gegensatz zu den Komponentenwerten, die sich aus der Komponentenlösung vollständig berechnen lassen, nur geschätzt werden. Welches Verfahren dabei verwendet wird, sollte grundsätzlich vom Ziel der Analyse abhängen: Soll als Ziel eine reduzierte Lösung zur einfacheren Interpretation stehen, so kann die PCA herangezogen werden. Wenn man hingegen von der Existenz latenter Konstrukte ausgeht und an der Ermittlung latenter Faktoren interessiert ist, beispielsweise zur Instrumentenreduktion, so sollte die EFA angewendet werden.[14]

Es gibt verschiedene Methoden zur Berechnung der EFA, von denen die Maximum-Likelihood (ML)-Faktorenanalyse und die Hauptachsenanalyse[15] am bekanntesten sind. Bei der ML-Analyse erfolgt die Schätzung der Ladungen und Residualvariablen anhand der ML-Methode. Notwendig ist hierfür eine multivariate

[13]In manchen Statistikprogrammen, so beispielsweise Stata, erhält man statt der Kommunalität die Uniqueness (1 – Kommunalität) einer Variablen, also den Teil der Informationen der Indikatorvariablen, der nicht über die Faktoren dargestellt werden kann.

[14]Dabei werden beide Verfahren nicht selten vermischt, siehe beispielsweise Ray (2007), der eine PCA durchführt, aber mit EFA benennt.

[15]Die Berechnung verläuft dabei ähnlich wie bei der PCA, es wird jedoch eine Residualvariable berücksichtigt, siehe weiterführend Eid et al. (2015).

Normalverteilung der manifesten x-Variablen, die mittels des Bartlett-Tests auf Sphärizität aus Abschn. 2.1.2 überprüft werden kann. Zudem ist gefordert, dass die Faktoren unkorreliert sind (zumindest bei orthogonaler Rotation) und jeweils einen Mittelwert von 0 und eine Varianz von 1 aufweisen. Notwendig ist auch, dass die Residualvariablen ε_i untereinander unkorreliert sind und einen Mittelwert von 0 aufweisen (Bartholomew et al. 2002, siehe für eine Darstellung der mathematischen Grundlagen Eid et al. 2015).

4.2 Anwendungsbeispiel

In der generellen Vorgehensweise und der Reihenfolge „1. Schätzung der Lösung, 2. Reduktion der Variablen, und 3. Rotation der Faktoren" unterscheidet sich die EFA kaum von der PCA, sodass hier analog zu Abschn. 2 vorgegangen werden kann. Beide Verfahren führen zu ähnlichen Ergebnissen, wenn die Residualvariablen ε_i niedrig sind. Auf Grund der Residualvariablen ε_i fallen bei der EFA sowohl die Faktorenladungen als auch die erklärte Varianz der Faktoren niedriger als bei der PCA aus.

Um dies zu veranschaulichen, wird das Beispiel der Analyse von Wertorientierungen aus Abschn. 2 als EFA durchgeführt. Hier wollen wir untersuchen, auf welche generellen Wertedimensionen sich die fünf Wertorientierungen zurückführen lassen. Bei einer Analyse der Eigenwerte fällt bereits auf, dass diese für die ersten beiden Faktoren mit 1,17 und 0,67 niedriger ausfallen. Hier würden wir also nur einen Faktor extrahieren. Dies kann darauf zurückgeführt werden, dass wir mit insgesamt nur fünf Variablen, die vermutlich zwei Faktoren abdecken, nicht genügend manifeste Variablen mit einbeziehen.

Die Ergebnisse beider Verfahren nach Extraktion zweier Faktoren und Varimax-Rotation sind in Tab. 10 abgebildet. Für die PCA enthält die vierte Spalte den Anteil an der erklärten Varianz der jeweiligen Variablen durch die beiden Komponenten nach Reduktion der anderen drei Hauptkomponenten. Wenn wir nun die Ladungen

Tab. 10 Reduzierte Zwei-Komponenten- und Zwei-Faktorenlösung, Wertorientierungen

	PCA (Varimax-Rotation)			EFA (ML, Varimax-Rotation)		
	1	2	Erklärte Varianz	1	2	Kommunalität
Toleranz gegenüber verschiedenen Menschen	−0,09	0,72	0,73	−0,03	0,66	0,44
Dass alle Menschen gerecht behandelt werden	0,09	0,67	0,69	0,19	0,63	0,43
Alle Gesetze befolgen	0,53	0,17	0,55	0,54	0,22	0,34
Traditionelle Werte und Überzeugungen bewahren	0,56	−0,05	0,49	0,46	0,07	0,21
In einem starken Staat leben	0,62	−0,09	0,361	0,64	0,01	0,41

ZA5665, GESIS-Panel 2015 (13. Welle), N = 3496

vergleichen, sehen wir, dass die Ladungen für die EFA im Schnitt etwas geringer ausfallen. Bei der EFA unterstellen wir im Gegensatz zur PCA ein Messmodell, das die Erklärung der Varianz nicht nur auf die gemeinsamen Faktoren 1 und 2 zurückführt, sondern auch auf weitere spezifische Faktoren. In der Spalte „Kommunalitäten" sehen wir, dass dabei für keine der fünf Variablen mehr als die Hälfte der Varianz durch die beiden Faktoren erklärt wird. Gerade bei der Aussage „Traditionelle Werte und Überzeugungen bewahren", die nur zu 21 Prozent durch die beiden Faktoren erklärt wird, können wir annehmen, dass wir bei Einbezug weiterer Statements, die sich auf Traditionen beziehen, einen eigenen latenten Faktor „Tradition" identifizieren könnten.

5 Abschließende Hinweise für die eigene Anwendung

Für die eigene Anwendung sollte zuerst überlegt werden, für welchen Zweck ein dimensionsreduzierendes Verfahren eingesetzt werden soll. Ist man lediglich daran interessiert, die Interpretation einer großen Menge von Variablen zu vereinfachen? Dann kann die PCA eingesetzt werden. Möchte man die eigene Analyse bereits im Framework der Modellierung latenter Konstrukte verorten und zugrunde liegende latente Faktoren identifizieren, so sollte auf die Faktorenanalyse zurückgegriffen werden. Soll dabei überprüft werden, ob die vorhandenen Daten einer bestimmten vorgegebenen Struktur entsprechen, sollte das Verfahren der CFA gewählt werden. Wenn vorher nicht bekannt ist, welche Struktur den Daten zugrunde liegt, sollte das Verfahren der EFA eingesetzt werden. Dabei ist für die sichere Identifizierung latenter Faktoren ein Datensatz in einer in der Politikwissenschaft üblichen Größe von ca. 1000 Befragten mit etwa 3–4 Variablen pro Faktor notwendig. Andere Kennzahlen bei kleineren Datensätzen zur Generalisierung sind beispielsweise bei Bortz und Schuster (2010, S. 396) zu finden.

Lädt eine Variable selbst nach Rotation (zum Finden einer Einfachstruktur) nicht ausschließlich auf eine/n Komponente/Faktor hoch, so ist es möglich, weitere Variablen in die Analyse aufzunehmen. Dahinter steckt der Gedanke, dass diese bei der Identifikation einer/eines weiteren Komponente/Faktors helfen, die/der bis jetzt nur durch diese Variable vertreten ist und daher nicht identifiziert werden konnte. Da jedoch oft bereits alle verfügbaren Variablen Teil der Analyse sind, kann eine solche mehrfachladende Variable alternativ von der weiteren Analyse ausgeschlossen werden. Dafür können die in Abschn. 2.3 genannten Werte von 0,3 oder 0,5 herangezogen werden, um so beispielsweise Variablen auszuschließen, die mit mehr als 0,3 oder 0,5 auf mehrere Komponenten/Faktoren laden.

Wichtig für die Anwendung ist, dass PCA und EFA einigen Interpretationsspielraum lassen, der von Kritikern negativ beurteilt wird. Daher ist es notwendig, an den entsprechenden Stellen das eigene Vorgehen und die Entscheidungsgrundlagen klar darzulegen. Bei der Bestimmung der Anzahl der Komponenten/Faktoren (siehe Abschn. 2.2) sollte daher nicht nur ein Maß, sondern wenn möglich sowohl Screetest als auch Kaiser-Kritierium herangezogen und die Ergebnisse diskutiert werden. Für die Wahl des Rotationsverfahrens bietet es sich an, zuerst auf die

ursprüngliche Variante der orthogonalen Rotation zurückzugreifen und nach der ersten inhaltlichen Deutung der Komponenten/Faktoren anschließend zu überlegen, ob eine Korrelation der Komponenten/Faktoren theoretisch Sinn ergibt. Um zu zeigen, dass die Wahl des Rotationsverfahrens nicht auf einer möglichst passenden Komponenten-/Faktorenlösung beruht, kann zusätzlich in einer Fußnote berichtet werden, dass die Verwendung des jeweils anderen Rotationsverfahrens zu ähnlichen Ergebnissen führt. Auch bei der inhaltlichen Deutung der Komponenten/Faktoren sollte dem/der Forschenden stets bewusst sein, dass sich eine Benennung auf Basis der zugrunde liegenden Variablen begründen lassen muss.

6 Fazit

Auch wenn die PCA und EFA ältere Verfahren sind, werden sie noch heute in der Forschung verwendet. Eine Anwendung lohnt sich immer dann, wenn die Interpretation einer großen Anzahl von manifesten Variablen durch wenige Komponenten vereinfacht werden soll oder zugrunde liegende Faktoren identifiziert werden sollen, für die noch keine Modellstruktur angenommen wird. Der PCA und auch der EFA wird oftmals vorgeworfen eher subjektiv-geprägte Verfahren zu sein, da hier durchaus Interpretationsspielräume hinsichtlich des anzuwendenden Rotationsverfahrens (orthogonal oder schiefwinklig), der Anzahl der zu extrahierenden Komponenten/Faktoren sowie der inhaltlichen Deutung der Komponenten/Faktoren bestehen. PCA und EFA sind jedoch ausgezeichnete Reduktionstechniken, die einen schnellen Überblick über eine große Menge von Variablen ermöglichen und als Ausgangsbasis für weitere Berechnungen dienen können. Wichtig ist dabei, dass die eigenen Auswahlprozesse im Text stets transparent gemacht und begründet werden.

7 Kommentierte Literaturhinweise

Für einen Einblick in die tatsächliche Anwendung der Verfahren in der Forschung bieten sich beispielsweise Netjes und Binnema (2007) an, die eine PCA verschiedener Datenquellen zu Parteipositionen durchführen, Evans (2000), der eine PCA für verschiedene euroskeptische Einstellungen berechnet oder Jackman und Miller (1996), die eine PCA auf verschiedene Indikatoren zur institutionellen Performanz von italienischen Regionalregierungen anwenden. Die Ergebnisse einer EFA liegen auch der Arbeit von Gabel und Huber (2000) zur Validität der Einstufung von Parteien auf Basis von Daten der Party Manifesto Group zugrunde.

Da beide Verfahren seit längerem etabliert sind, findet sich eine breite Auswahl von Werken, die sich einführend und tiefergehend mit den mathematischen Grundlagen und der Interpretation auseinandersetzen. Gute Einführungswerke sind nach wie vor Arminger (1979) oder Kim und Mueller (1978), für eine mathematischere Einführung auch Cureton und D'Agostino (2009). Eine ausführliche Darstellung der explorativen Faktorenanalyse mit SPSS findet sich beispielsweise in Backhaus et al. (2016) oder in

Field (2013), der in Field et al. (2013) selbige auch für R vollzieht. Cleff (2015) führt eine explorative Faktorenanalyse (Hauptachsenanalyse) mit Stata durch.

Literatur

Arminger, Gerhard. 1979. *Faktorenanalyse*. Stuttgart: Teubner.

Backhaus, Klaus, Bernd Erichson, Rolf Weiber, und Wulff Plinke, Hrsg. 2016. Faktorenanalyse. In *Multivariate Analysemethoden*, 385–452. Berlin/Heidelberg: Springer.

Bartholomew, David J., Fiona Steele, Irini Moustaki, und Jane I. Galbraith. 2002. *The analysis and interpration of multivariate data for social scientists*. Boca Raton: Chapman & Hall/CRC.

Bartlett, Maurice Stevenson. 1950. Tests of significance in factor analysis. *British Journal of Statistical Psychology* 3(2): 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x.

Bortz, Jürgen, und Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. Berlin/Heidelberg: Springer.

Cattell, R. B. 1966. The scree test for the number of factors. *Multivariate Behavioral Research* 1(2): 245–276. https://doi.org/10.1207/s15327906mbr0102_10.

Cleff, Thomas, Hrsg. 2015. Faktorenanalyse. In *Deskriptive Statistik und Explorative Datenanalyse: Eine computergestützte Einführung mit Excel, SPSS und STATA*, 217–234. Wiesbaden: Gabler Verlag.

Cureton, Edward E., und Ralph B. D'Agostino. 2009. *Factor analysis: An applied approach*. New York: Psychology Press.

Dziuban, Charles D., und Edwin C. Shirkey. 1974. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin* 81(6): 358.

Eid, Michael, Mario Gollwitzer, und Manfred Schmitt. 2015. *Statistik und Forschungsmethoden: Mit Online-Materialien.* 4., überarb. u. erw. Aufl. Weinheim/Basel: Beltz.

Eijk, Cees van der, und Jonathan Rose. 2015. „Risky business: Factor analysis of survey data – Assessing the probability of incorrect dimensionalisation." *PLoS One* 10(3): e0118900. https://doi.org/10.1371/journal.pone.0118900.

Evans, Jocelyn A. J. 2000. „Contrasting attitudinal bases to Euroscepticism amongst the French electorate." *Electoral Studies* 19(4): 539–561. http://www.sciencedirect.com/science/article/pii/S0261379499000293. Zugegriffen am 04.08.2017.

Fabrigar, Leandre R., Duane T. Wegener, Robert C. MacCallum, und Erin J. Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 4(3): 272–299.

Field, Andy. 2013. *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock ‚n' roll*, 4. Aufl MobileStudy. Los Angeles/London/New Delhi: Sage.

Field, Andy, Jeremy Miles, und Zoë Field. 2013. *Discovering statistics using R*. Los Angeles: Sage. (Reprint).

Gabel, Matthew J., und John D. Huber. 2000. Putting parties in their place: Inferring party left-right ideological positions from party manifestos data. *American Journal of Political Science* 44(1): 94–103.

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

Jackman, Robert W., und Ross A. Miller. 1996. A renaissance of political culture? *American Journal of Political Science* 40(3): 632–659. https://doi.org/10.2307/2111787.

Kaiser, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrica* 23: 187–200.

Kaiser, H. F., und K. Dickman. 1959. Analytic determination of common factors. *American Psychological Reports* 14:425–438.

Kenny, Graham K. 1986. The metric properties of rating scales employed in evaluation research. *Evaluation Review* 10(3): 397–408. https://doi.org/10.1177/0193841X8601000309.

Kim, Jae-On, und Charles W. Mueller. 1978. *Factor analysis. Statistical methods and pratical issues*. Newbury Park: Sage.

Landgraf, Andrew J., und Yoonkyung Lee. 2015. Dimensionality reduction for binary data through the projection of natural parameters. arXiv:1510.06112.

Lawley, D. N., und A. E. Maxwell. 1971. *Factor analysis as a statistical method*, 2. Aufl. New York: American Elsevier Pub. Co.

Netjes, Catherine E., und Harmen A. Binnema. 2007. The salience of the European integration issue: Three data sources compared. *Electoral Studies* 26(1): 39–49. https://doi.org/10.1016/j.electstud.2006.04.007.

Pearson, Karl. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2:559–572.

Ray, Leonard. 2007. Validity of measured party positions on European integration: Assumptions, approaches, and a comparison of alternative measures. *Electoral Studies* 26(1): 11–22. https://doi.org/10.1016/j.electstud.2006.03.008.

Thurstone, Louis Leon. 1947. *Multiple factor analysis*. Chicago: University of Chicago Press.

Wolff, Hans-Georg, und Johann Bacher. 2010. Hauptkomponentenanalyse und explorative Faktorenanalyse. In *Handbuch der sozialwissenschaftlichen Datenanalyse*, Hrsg. Christof Wolf, 1. Aufl., 333–365. Wiesbaden: VS Verlag für Sozialwissenschaften.

Verwendete Datensätze und Stata-Ados

Bischof, Daniel. 2017. New graphic schemes for stata: Plotplain & plottig. *Stata Journal* 17(3): 748–759.

GESIS. 2017a. *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2016.* Version 2.1.0: GESIS Datenarchiv, Köln. https://doi.org/10.4232/1.12796.

GESIS. 2017b. *GESIS Panel – Standard Edition*. ZA5665 Data file Version 20.0.0: GESIS Datenarchiv, Köln. https://doi.org/10.4232/1.12766.

Rattinger, Hans, Sigrid Roßteutscher, Rüdiger Schmitt-Beck, Bernhard Weßels, Christof Wolf, Aiko Wagner, und Heiko Giebler. 2017. *Post-election Cross Section (GLES 2013).* Version 3.0.0: GESIS Data Archive, Cologne. https://doi.org/10.4232/1.12809.