Hauptkomponentenanalyse (PCA)

Multivariate Analyse

Mathematischer Hintergrund, durchgerechnetes Beispiel, Handhabung des Tools

Deutsch

Wolfgang Böhme

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

8. November 2026

Dokumentationen dieser Webseite wurden mit Unterstützung der KI Claude (Anthropic) entwickelt.

Inhaltsverzeichnis

1 Einleitung und Anwendungsgebiete **2**

2 Mathematischer Hintergrund **2**
2.1 Datenmatrix und Standardisierung . . . 2
2.2 Eigenwertzerlegung der Korrelationsmatrix . . . 2
2.3 Faktorladungen . . . 2
2.4 Erklärte Varianz und Kaiser-Kriterium . . . 3
2.5 Komponentenscores . . . 3
2.6 Kommunalitäten . . . 3
2.7 Varimax-Rotation . . . 3

3 Durchgerechnetes Beispiel: Bilanzkennzahlen **4**
3.1 Datenbasis ($n = 24$, $p = 6$) . . . 4
3.2 Schritt 1: Deskriptive Statistiken . . . 5
3.3 Schritt 2: Korrelationsmatrix $\mathbf{R}$ . . . 6
3.4 Schritt 3: Eigenwerte . . . 6
3.5 Schritt 4: Eigenvektoren und Ladungen . . . 6
3.6 Schritt 5: Komponentenscores (Auswahl) . . . 7

4 Handhabung des interaktiven PCA-Tools **8**
4.1 Dateneingabe . . . 8
4.2 Parameterauswahl (linke Seitenleiste) . . . 8
4.3 Ergebnistabs . . . 8
4.4 Typischer Arbeitsablauf . . . 9

5 Formelübersicht **9**

6 Weiterführende Literatur **9**

Einleitung und Anwendungsgebiete

Die **Hauptkomponentenanalyse** (engl. Principal Component Analysis, PCA) ist ein Verfahren der multivariaten Statistik zur **Dimensionsreduktion**. Ausgehend von p korrelierten Variablen erzeugt die PCA $k \leq p$ neue, unkorrelierte Variablen – die **Hauptkomponenten** – die einen maximalen Anteil der Gesamtvarianz erklaeren.

Typische Anwendungsgebiete:

- Indexbildung: Zusammenfassung mehrerer Indikatoren zu einem Gesamtindex
- Vorverarbeitung: Beseitigung von Multikollinearitaet vor Regressionsanalysen
- Visualisierung: Darstellung hochdimensionaler Daten im 2D-Biplot
- Datenkompression: Reduktion des Speicherbedarfs bei Signaldaten
- Exploration: Aufdeckung latenter Strukturen in Beobachtungsdaten

Mathematischer Hintergrund

Datenmatrix und Standardisierung

Gegeben sei eine Datenmatrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ mit n Beobachtungen und p Variablen. Bei unterschiedlichen Messskalen wird Z-Transformation empfohlen:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \qquad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \qquad s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \tag{1}$$

Nach Z-Transformation entspricht die Kovarianzmatrix $\mathbf{S}$ der Korrelationsmatrix $\mathbf{R}$.

Eigenwertzerlegung der Korrelationsmatrix

Die symmetrische, positiv semidefinite Korrelationsmatrix $\mathbf{R} \in \mathbb{R}^{p \times p}$ wird zerlegt gemaess:

$$\mathbf{R} = \mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}^{\top} \tag{2}$$

mit $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, und orthonormalen Eigenvektoren $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$.

Numerische Berechnung erfolgt im Tool mit dem **Jacobi-Verfahren** fuer symmetrische Matrizen (zyklische Variante). In jedem Sweep werden nacheinander alle Nichtdiagonalelemente durch Jacobi-Rotationen zu null gemacht, bis die Summe der quadrierten Nichtdiagonalelemente unter 10^{-12} liegt.

Faktorladungen

$$l_{jk} = v_{jk} \cdot \sqrt{\lambda_k} \tag{3}$$

l_{jk} ist die Korrelation zwischen der j-ten Originalvariable und der k-ten Hauptkomponente.

Erklärte Varianz und Kaiser-Kriterium

$$\text{ErkVar}_k = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j} \cdot 100\% \tag{4}$$

Das **Kaiser-Kriterium** behalt alle Komponenten mit $\lambda_k \geq 1$: Bei standardisierten Daten erklaert eine Komponente mit $\lambda_k < 1$ weniger Varianz als eine einzelne Variable.

Komponentenscores

$$\mathbf{T} = \mathbf{Z}\,\mathbf{V}_k, \qquad \mathbf{V}_k \in \mathbb{R}^{p \times k} \tag{5}$$

Die Scores sind die Projektionen der standardisierten Beobachtungen auf den Hauptkomponentenraum. Sie sind paarweise unkorreliert.

Kommunalitäten

$$h_j^2 = \sum_{k=1}^{K} l_{jk}^2 \tag{6}$$

$h_j^2 \in [0, 1]$ gibt den durch K Komponenten erklaerten Varianzanteil der j-ten Variable an.

Varimax-Rotation

Die Varimax-Rotation maximiert die Varianz der quadrierten Ladungen je Faktor:

$$\max \sum_{k=1}^{K} \left[\frac{1}{p} \sum_{j=1}^{p} l_{jk}^4 - \left(\frac{1}{p} \sum_{j=1}^{p} l_{jk}^2 \right)^2 \right] \tag{7}$$

Die Rotation ist orthogonal: Die Kommunalitaeten und die insgesamt erklaerte Varianz bleiben unveraendert, die Varianzanteile verteilen sich jedoch neu auf die Komponenten. Die Ladungsstruktur wird dadurch interpretierbarer (einfache Struktur).

Durchgerechnetes Beispiel: Bilanzkennzahlen

Datenbasis ($n = 24$, $p = 6$)

Das Beispiel ist in allen vier Rechnern der Toolsammlung vorgeladen und wird in allen vier Handbüchern durchgerechnet. Es handelt sich um **synthetische** Bilanzkennzahlen von 24 Unternehmen (U01–U24), alle in Prozent:

- **Stabilität:** X_1 Eigenkapitalquote (EK/Bilanzsumme), X_2 Liquidität 2. Grades ((flüssige Mittel + kurzfristige Forderungen)/kurzfristige Verbindlichkeiten), X_3 Anlagendeckung II ((EK + langfristiges FK)/Anlagevermögen)
- **Ertragskraft:** X_4 Umsatzrendite, X_5 Gesamtkapitalrendite, X_6 Umsatzwachstum gegenüber dem Vorjahr

Die Daten wurden so erzeugt, dass drei Unternehmenstypen zu je acht Unternehmen entstehen: solide Substanzwerte (Rating A), ertragsstarke, höher verschuldete Wachstumsunternehmen (Rating B) und gefährdete Unternehmen (Rating C). Das Rating wird nur in der Diskriminanzanalyse (Handbuch 4) als Gruppenvariable verwendet; PCA, Faktoranalyse und Clusteranalyse arbeiten allein mit den sechs Kennzahlen.

Ziel der PCA: Die teils eng korrelierten Kennzahlen sollen auf wenige, unkorrelierte Hauptkomponenten verdichtet werden.

Tabelle 1: Bilanzkennzahlen der 24 Unternehmen (in %, synthetisch)

Nr.	Rating	X_1 EK-Q.	X_2 Liqu.	X_3 AD II	X_4 UR	X_5 GKR	X_6 Wachst.
U01	A	46.6	127	124	4.3	8.1	9.0
U02	A	49.4	141	127	3.4	7.0	5.5
U03	A	32.0	109	100	4.8	5.0	4.5
U04	A	33.7	134	129	5.7	4.8	2.7
U05	A	51.1	120	122	6.4	5.8	2.7
U06	A	41.2	143	144	3.1	9.3	5.0
U07	A	50.5	117	125	8.8	9.5	10.1
U08	A	56.6	130	130	5.6	8.6	2.1
U09	B	38.3	119	86	7.7	12.2	4.3
U10	B	40.8	114	98	8.5	10.7	15.0
U11	B	35.5	102	99	7.0	10.9	8.7
U12	B	39.0	128	108	11.8	11.9	13.5
U13	B	20.2	65	93	8.1	11.9	18.0
U14	B	33.2	107	97	8.3	11.9	11.1
U15	B	31.5	82	100	7.7	12.9	11.0
U16	B	34.5	78	93	9.3	6.9	12.1
U17	C	16.5	86	81	1.1	8.0	5.6
U18	C	26.6	74	103	1.8	4.7	-3.0
U19	C	33.1	82	98	1.7	2.5	-3.4
U20	C	23.5	80	72	1.6	7.9	-2.1
U21	C	32.5	58	82	2.3	2.3	2.7
U22	C	23.2	96	85	2.8	3.2	-3.6
U23	C	34.3	101	95	1.6	5.6	-1.9
U24	C	34.3	98	95	2.1	2.6	2.8

Schritt 1: Deskriptive Statistiken

Tabelle 2: Mittelwerte und Standardabweichungen

	X_1	X_2	X_3	X_4	X_5	X_6
$\bar{x}$	35.754	103.792	103.583	5.229	7.675	5.517
s	10.074	24.270	18.547	3.084	3.379	6.083

Die Streuungen unterscheiden sich um fast eine Zehnerpotenz (Liquidität vs. Umsatzrendite). Ohne Z-Transformation würden Liquidität und Anlagendeckung die Analyse dominieren – deshalb ist die Standardisierung hier zwingend. Beispiel U01: $z_{11} = (46.6 - 35.754)/10.074 = 1.077$.

Schritt 2: Korrelationsmatrix R

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.710 & 0.743 & 0.311 & 0.122 & 0.113 \\ 0.710 & 1.000 & 0.771 & 0.233 & 0.225 & 0.079 \\ 0.743 & 0.771 & 1.000 & 0.172 & 0.087 & 0.108 \\ 0.311 & 0.233 & 0.172 & 1.000 & 0.699 & 0.793 \\ 0.122 & 0.225 & 0.087 & 0.699 & 1.000 & 0.727 \\ 0.113 & 0.079 & 0.108 & 0.793 & 0.727 & 1.000 \end{pmatrix} \tag{8}$$

Die Matrix zerfällt in zwei Blöcke: Innerhalb der Stabilitätskennzahlen (X_1–X_3) und innerhalb der Ertragskennzahlen (X_4–X_6) liegen die Korrelationen zwischen 0.70 und 0.79, zwischen den Blöcken höchstens bei 0.31. Zu erwarten sind daher **zwei** Hauptkomponenten.

Schritt 3: Eigenwerte

Tabelle 3: Eigenwerte und erklärte Varianz

Komponente	Eigenwert λ	Varianzanteil	Kumuliert	Kaiser
PC1	2.971	49.52 %	49.52 %	≥ 1
PC2	2.010	33.50 %	83.01 %	≥ 1
PC3	0.397	6.61 %	89.62 %	< 1
PC4	0.289	4.81 %	94.43 %	< 1
PC5	0.200	3.34 %	97.77 %	< 1
PC6	0.134	2.23 %	100.00 %	< 1

Kaiser-Kriterium: **2 Komponenten**, die zusammen 83.01 % der Varianz erklären. Auch der Scree-Plot zeigt nach PC2 einen deutlichen Knick (2.010 $\rightarrow$ 0.397). Die Summe der Eigenwerte ist $p = 6$; z. B. $2.971/6 = 49.52\,\%$.

Schritt 4: Eigenvektoren und Ladungen

Erste beiden Eigenvektoren (normiert):

$$\mathbf{v}_1 = (\ 0.416,\ \ 0.418,\ \ 0.397,\ 0.442,\ 0.389,\ 0.384)^\top$$
$$\mathbf{v}_2 = (-0.384,\ -0.391,\ -0.436,\ 0.362,\ 0.410,\ 0.459)^\top$$

Ladungen nach $l_{jk} = v_{jk}\sqrt{\lambda_k}$, z. B. $l_{11} = 0.416 \cdot \sqrt{2.971} = 0.717$ und $l_{12} = -0.384 \cdot \sqrt{2.010} = -0.544$.

Tabelle 4: Ladungsmatrix: unrotiert und Varimax-rotiert

	unrotiert		Varimax			
Variable	PC1	PC2	RC1	RC2	h_j^2	$1-h_j^2$
X_1 EK-Quote	0.717	−0.544	**0.893**	0.115	0.810	0.190
X_2 Liquidität	0.720	−0.555	**0.902**	0.109	0.826	0.174
X_3 Anlagendeckung	0.684	−0.617	**0.921**	0.040	0.850	0.150
X_4 Umsatzrendite	0.762	0.513	0.184	**0.900**	0.844	0.156
X_5 GK-Rendite	0.671	0.581	0.071	**0.885**	0.789	0.211
X_6 Umsatzwachstum	0.662	0.651	0.015	**0.928**	0.862	0.138
Erkl. Varianz	49.5 %	33.5 %	41.6 %	41.4 %	Σ 83.0 %	

Beispiel Kommunalität: $h_1^2 = 0.717^2 + (-0.544)^2 = 0.810$ – identisch für die rotierte Lösung ($0.893^2 + 0.115^2 = 0.810$). Die rotierte Lösung erhält man im Tool mit *Rotation* = **Varimax**.

Interpretation:

- **Unrotiert:** PC1 lädt auf allen sechs Kennzahlen positiv (0.66–0.76) – ein allgemeiner Faktor „finanzielle Gesamtstärke“. PC2 ist bipolar: negativ für die Stabilitäts-, positiv für die Ertragskennzahlen. Er beschreibt das „Profil“ eines Unternehmens zwischen Substanz und Rendite.
- **Varimax:** Nach der Rotation ergibt sich eine einfache Struktur. RC1 = **Stabilität** (X_1–X_3, Ladungen 0.89–0.92), RC2 = **Ertragskraft** (X_4–X_6, Ladungen 0.89–0.93). Die erklärte Varianz verteilt sich fast gleichmäßig (41.6 % bzw. 41.4 %), die Summe von 83.0 % bleibt unverändert.

Schritt 5: Komponentenscores (Auswahl)

$$t_{i1} = 0.416\, z_{i1} + 0.418\, z_{i2} + 0.397\, z_{i3} + 0.442\, z_{i4} + 0.389\, z_{i5} + 0.384\, z_{i6}$$

Tabelle 5: Scores ausgewählter Unternehmen (unrotierte Lösung)

Unternehmen	Rating	PC1-Score	PC2-Score
U08	A	1.822	−1.939
U07	A	2.307	−0.291
U03	A	−0.576	−0.309
U12	B	2.578	1.268
U13	B	0.150	3.258
U21	C	−2.602	0.161
U19	C	−2.270	−1.131

Im Biplot (PC1 vs. PC2) bilden die drei Ratinggruppen gut erkennbare Wolken: Die C-Unternehmen haben durchweg negative PC1-Scores (geringe Gesamtstärke), A-Unternehmen liegen überwiegend im Bereich negativer, B-Unternehmen im Bereich positiver PC2-Scores. Auffällig sind zwei A-Unternehmen: **U03** liegt mit $t_1 = -0.576$ näher bei den C-Unternehmen, **U07** ist mit hoher Rendite fast ein B-Profil. Beide

tauchen in der Cluster- und der Diskriminanzanalyse (Handbücher 3 und 4) wieder als Grenzfälle auf.

Handhabung des interaktiven PCA-Tools

Dateneingabe

Tab »Manuell«
Zeilen- und Spaltenanzahl oben einstellen. Spaltennamen in der Kopfzeile eintragen; Messwerte in die Zahlenfelder. Beim Aendern der Dimensionen bleibt vorhandene Eingabe erhalten.

Tab »CSV«
CSV-Datei per Klick oder Drag & Drop laden. Erste Zeile muss Spaltennamen enthalten. Trennzeichen Komma, Semikolon und Tab werden automatisch erkannt.

Tab »Beispiel«
Vier vorbereitete Datensätze stehen bereit: Bilanzkennzahlen (24 Unternehmen, 6 Vars., vorgeladen), Iris (15 Obs., 4 Vars.), Wirtschaftsindikatoren (12 Länder, 5 Vars.), Arbeitszufriedenheit (20 Pers., 6 Items). Zu jedem Beispiel erscheint über der Eingabetabelle ein Erläuterungskasten mit den Daten, dem Ziel der Analyse und dem erwarteten Ergebnis; er verschwindet bei »Leer« oder beim Laden einer CSV-Datei.

Parameterauswahl (linke Seitenleiste)

Standardisierung
Z-Transformation (empfohlen): Pflicht bei unterschiedlichen Einheiten.
Keine: Kovarianzmatrix verwenden (nur bei gleichen Einheiten sinnvoll).

Anzahl Komponenten
Automatisch: Kaiser-Kriterium ($\lambda \geq 1$). Alternativ: feste Anzahl 2–5 waehlen.

Rotation
Keine: unrotierte Loesung. *Varimax:* orthogonale Rotation ab $k \geq 2$ Komponenten empfohlen.

Ladungsschwelle
Ladungen $\geq |0.3|$ werden hervorgehoben (Standard). Die Hervorhebung laesst sich per Kontrollkaestchen abschalten.

Ergebnistabs

Uebersicht
Kennzahlenkarten (n, p, Komponentenanzahl, erklaerte Varianz) + vollstaendige Eigenwert-Tabelle mit Kaiser-Badge.

Ladungen
Faktorladungsmatrix mit Farbhervorhebung, Kommunalitaeten und erklaerte Varianz je Komponente.

Grafiken
Scree-Plot: Eigenwerte mit Kaiser-Grenzlinie (gestrichelt). *Kumul. Varianz:* Bal-

ken und Linie kombiniert. *Biplot:* Scores (PC1 vs. PC2) mit Ladungspfeilen.

Scores
Tabelle aller Komponentenscores je Beobachtung.

Interpretation
Automatisch erzeugter Bewertungstext mit Gueteampel.

Export
CSV: Ladungen, Scores, Eigenwerte. PNG: Scree-Plot, Biplot.

Typischer Arbeitsablauf

1. Daten eingeben oder Beispieldatensatz laden
2. Standardisierung pruefen (Z-Transformation bei unterschiedlichen Einheiten)
3. Faktoranzahl auf *Automatisch* belassen
4. »**PCA berechnen**« klicken
5. Scree-Plot pruefen: Ellbogen identifizieren
6. Ladungstabelle interpretieren: welche Variablen definieren welche Komponente?
7. Bei $k \geq 2$: Varimax-Rotation aktivieren und Ladungsstruktur vergleichen
8. Scores bei Bedarf als CSV exportieren

Wichtige Hinweise:
- Mindestens $n = 3$ vollstaendige Beobachtungen und $p = 2$ Variablen erforderlich.
- Zeilen mit fehlenden Werten werden automatisch ausgeschlossen.
- Die PCA ist nicht robust gegenueber Ausreissern – bitte vorher pruefen.
- Faustformel: $n \geq 5 \cdot p$ fuer stabile Ergebnisse.
- Der Biplot erfordert mindestens $k = 2$ extrahierte Komponenten.

Formelübersicht

Weiterführende Literatur

- Jolliffe, I. T. (2002). *Principal Component Analysis*, 2. Aufl. Springer.
- Backhaus, K. et al. (2021). *Multivariate Analysemethoden*, 16. Aufl. Springer Gabler.
- Bortz, J. & Schuster, C. (2010). *Statistik fuer Human- und Sozialwissenschaftler*. Springer.

Tabelle 6: Formeln der PCA im Ueberblick

Bezeichnung	**Formel**
Z-Transformation	$z_{ij} = (x_{ij} - \bar{x}_j)/s_j$
Korrelationskoeff.	$r_{jk} = \frac{1}{n-1}\sum_i z_{ij} z_{ik}$
Eigenwertproblem	$\mathbf{R}\mathbf{v} = \lambda \mathbf{v}$
Faktorladung	$l_{jk} = v_{jk}\sqrt{\lambda_k}$
Erklaerte Varianz	$\lambda_k / \sum_j \lambda_j \cdot 100\,\%$
Kommunalitaet	$h_j^2 = \sum_{k=1}^{K} l_{jk}^2$
Komponentenscore	$t_{ik} = \sum_{j=1}^{p} z_{ij} v_{jk}$
Kaiser-Kriterium	$\lambda_k \geq 1$