Principal Component Analysis (PCA)

Multivariate Analysis

Mathematical Background, Worked Example, Tool Usage

English

Wolfgang Boehme

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

8 November 2026

The documentation on this website was developed with the support of the AI Claude (Anthropic).

Table of Contents

1 Introduction and Application Areas **2**

2 Mathematical Background **2**
2.1 Data Matrix and Standardisation . . . 2
2.2 Eigendecomposition of the Correlation Matrix . . . 2
2.3 Factor Loadings . . . 2
2.4 Explained Variance and Kaiser Criterion . . . 3
2.5 Component Scores . . . 3
2.6 Communalities . . . 3
2.7 Varimax Rotation . . . 3

3 Worked Example: Balance-Sheet Ratios **4**
3.1 Data ($n = 24$, $p = 6$) . . . 4
3.2 Step 1: Descriptive Statistics . . . 5
3.3 Step 2: Correlation Matrix $\mathbf{R}$ . . . 6
3.4 Step 3: Eigenvalues . . . 6
3.5 Step 4: Eigenvectors and Loadings . . . 6
3.6 Step 5: Component Scores (Selection) . . . 7

4 Using the Interactive PCA Tool **8**
4.1 Data Input . . . 8
4.2 Parameter Selection (Left Sidebar) . . . 8
4.3 Results Tabs . . . 8
4.4 Typical Workflow . . . 9

5 Formula Reference **9**

6 Further Reading **9**

Introduction and Application Areas

Principal Component Analysis (PCA) is a multivariate statistical method for **dimensionality reduction**. Starting from p correlated variables, PCA produces $k \leq p$ new, uncorrelated variables – the **principal components** – that explain the maximum share of total variance.

Typical application areas:

- Index construction: combining several indicators into a composite index
- Pre-processing: eliminating multicollinearity before regression analyses
- Visualisation: representing high-dimensional data in a 2D biplot
- Data compression: reducing storage requirements for signal data
- Exploration: uncovering latent structures in observational data

Mathematical Background

Data Matrix and Standardisation

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a data matrix with n observations and p variables. When variables are measured on different scales, z-transformation is recommended:

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

After z-transformation the covariance matrix $\mathbf{S}$ equals the correlation matrix $\mathbf{R}$.

Eigendecomposition of the Correlation Matrix

The symmetric, positive semi-definite correlation matrix $\mathbf{R} \in \mathbb{R}^{p \times p}$ is decomposed as:

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

with $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and orthonormal eigenvectors $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$.

Numerical computation in the tool uses the **Jacobi method** for symmetric matrices (cyclic variant). In each sweep, all off-diagonal elements are set to zero one after another by Jacobi rotations until the sum of the squared off-diagonal elements falls below 10^{-12}.

Factor Loadings

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

l_{jk} is the correlation between the j-th original variable and the k-th principal component.

Explained Variance and Kaiser Criterion

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

The **Kaiser criterion** retains all components with $\lambda_k \geq 1$: for standardised data, a component with $\lambda_k < 1$ explains less variance than a single variable.

Component Scores

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

The scores are the projections of the standardised observations onto the principal component space. They are pairwise uncorrelated.

Communalities

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

$h_j^2 \in [0, 1]$ gives the share of variance of the j-th variable explained by K components.

Varimax Rotation

Varimax rotation maximises the variance of squared loadings per factor:

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

The rotation is orthogonal: the communalities and the total explained variance remain unchanged, but the variance shares are redistributed among the components. The loading structure becomes more interpretable (simple structure).

Worked Example: Balance-Sheet Ratios

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

This example is pre-loaded in all four calculators of the tool collection and is worked through in all four handbooks. It consists of **synthetic** balance-sheet ratios of 24 companies (U01–U24), all in per cent:

- **Stability:** X_1 equity ratio (equity/total assets), X_2 quick ratio ((cash + short-term receivables)/current liabilities), X_3 asset coverage II ((equity + long-term debt)/fixed assets)
- **Profitability:** X_4 return on sales, X_5 return on assets, X_6 sales growth over the previous year

The data were generated so that three company types of eight companies each emerge: solid asset-rich companies (rating A), highly profitable but more leveraged growth companies (rating B) and companies at risk (rating C). The rating is used as a group variable only in discriminant analysis (Handbook 4); PCA, factor analysis and cluster analysis work with the six ratios alone.

Goal of the PCA: condense the partly highly correlated ratios into a few uncorrelated principal components.

Table 1: Balance-sheet ratios of the 24 companies (in %, synthetic)

No.	Rating	X_1 Equity	X_2 Liqu.	X_3 AC II	X_4 ROS	X_5 ROA	X_6 Growth
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

Step 1: Descriptive Statistics

Table 2: Means and standard deviations

	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

The standard deviations differ by almost an order of magnitude (quick ratio vs. return on sales). Without z-transformation, liquidity and asset coverage would dominate the analysis – standardisation is therefore essential here. Example U01: $z_{11} = (46.6 - 35.754)/10.074 = 1.077$.

Step 2: Correlation Matrix R

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

The matrix splits into two blocks: within the stability ratios (X_1–X_3) and within the profitability ratios (X_4–X_6) the correlations range from 0.70 to 0.79, between the blocks they are at most 0.31. **Two** principal components are therefore to be expected.

Step 3: Eigenvalues

Table 3: Eigenvalues and explained variance

Component	Eigenvalue λ	Variance share	Cumulative	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 criterion: **2 components**, which together explain 83.01 % of the variance. The scree plot also shows a clear elbow after PC2 (2.010 $\rightarrow$ 0.397). The eigenvalues sum to $p = 6$; e. g. $2.971/6 = 49.52\,\%$.

Step 4: Eigenvectors and Loadings

First two eigenvectors (normalised):

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

Loadings via $l_{jk} = v_{jk}\sqrt{\lambda_k}$, e. g. $l_{11} = 0.416 \cdot \sqrt{2.971} = 0.717$ and $l_{12} = -0.384 \cdot \sqrt{2.010} = -0.544$.

Table 4: Loading matrix: unrotated and Varimax-rotated

	unrotated		Varimax			
Variable	PC1	PC2	RC1	RC2	h_j^2	$1-h_j^2$
X_1 Equity ratio	0.717	−0.544	**0.893**	0.115	0.810	0.190
X_2 Quick ratio	0.720	−0.555	**0.902**	0.109	0.826	0.174
X_3 Asset coverage	0.684	−0.617	**0.921**	0.040	0.850	0.150
X_4 Return on sales	0.762	0.513	0.184	**0.900**	0.844	0.156
X_5 Return on assets	0.671	0.581	0.071	**0.885**	0.789	0.211
X_6 Sales growth	0.662	0.651	0.015	**0.928**	0.862	0.138
Expl. variance	49.5 %	33.5 %	41.6 %	41.4 %	Σ 83.0 %	

Example communality: $h_1^2 = 0.717^2 + (-0.544)^2 = 0.810$ – identical for the rotated solution ($0.893^2 + 0.115^2 = 0.810$). The rotated solution is obtained in the tool with *Rotation* = **Varimax**.

Interpretation:

- **Unrotated:** PC1 loads positively on all six ratios (0.66–0.76) – a general factor of "overall financial strength". PC2 is bipolar: negative for the stability ratios, positive for the profitability ratios. It describes a company's "profile" between asset strength and returns.
- **Varimax:** After rotation a simple structure emerges. RC1 = **stability** (X_1–X_3, loadings 0.89–0.92), RC2 = **profitability** (X_4–X_6, loadings 0.89–0.93). The explained variance is distributed almost evenly (41.6 % and 41.4 %); the total of 83.0 % remains unchanged.

Step 5: Component Scores (Selection)

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

Table 5: Scores of selected companies (unrotated solution)

Company	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

In the biplot (PC1 vs. PC2) the three rating groups form clearly recognisable clouds: C companies all have negative PC1 scores (low overall strength), A companies lie mostly in the region of negative and B companies in the region of positive PC2 scores. Two A companies stand out: **U03**, with $t_1 = -0.576$, lies closer to the C companies, and

U07, with its high returns, almost has a B profile. Both reappear as borderline cases in cluster and discriminant analysis (Handbooks 3 and 4).

Using the Interactive PCA Tool

Data Input

Tab "Manual"
Set the number of rows and columns at the top. Enter column names in the header row; enter measurements in the number fields. Existing input is preserved when dimensions are changed.

Tab "CSV"
Load a CSV file by clicking or drag-and-drop. The first row must contain column names. Comma, semicolon and tab delimiters are detected automatically.

Tab "Example"
Four prepared datasets are available: Balance-sheet ratios (24 companies, 6 variables, pre-loaded), Iris (15 obs., 4 variables), Economic indicators (12 countries, 5 variables), Job satisfaction (20 persons, 6 items). For each example, an explanatory box above the input table describes the data, the goal of the analysis and the expected result; it disappears when "Clear" is clicked or a CSV file is loaded.

Parameter Selection (Left Sidebar)

Standardisation
Z-transformation (recommended): required when variables have different units. *None:* use the covariance matrix (only meaningful when all units are identical).

Number of components
Automatic: Kaiser criterion ($\lambda \geq 1$). Alternatively: select a fixed number from 2 to 5.

Rotation
None: unrotated solution. *Varimax:* orthogonal rotation recommended for $k \geq 2$ components.

Loading threshold
Loadings $\geq |0.3|$ are highlighted (default). The highlighting can be switched off via a checkbox.

Results Tabs

Overview
Summary cards (n, p, number of components, explained variance) + complete eigenvalue table with Kaiser badge.

Loadings
Factor loading matrix with colour highlighting, communalities and explained variance per component.

Charts
Scree plot: eigenvalues with Kaiser threshold line (dashed). *Cumulative variance:*

combined bar and line chart. *Biplot:* scores (PC1 vs. PC2) with loading arrows.

Scores
Table of all component scores per observation.

Interpretation
Automatically generated assessment text with traffic-light indicator.

Export
CSV: loadings, scores, eigenvalues. PNG: scree plot, biplot.

Typical Workflow

1. Enter data or load a sample dataset.
2. Check standardisation (z-transformation for variables with different units).
3. Leave the number of components on *Automatic*.
4. Click **"Calculate PCA"**.
5. Check the scree plot: identify the elbow.
6. Interpret the loading table: which variables define which component?
7. For $k \geq 2$: activate Varimax rotation and compare loading structures.
8. Export scores as CSV if needed.

Important notes:
- At least $n = 3$ complete observations and $p = 2$ variables are required.
- Rows with missing values are excluded automatically.
- PCA is not robust to outliers – please check beforehand.
- Rule of thumb: $n \geq 5 \cdot p$ for stable results.
- The biplot requires at least $k = 2$ extracted components.

Formula Reference

Further Reading

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

Table 6: PCA formulae at a glance

Name	**Formula**
Z-transformation	$z_{ij} = (x_{ij} - \bar{x}_j)/s_j$
Correlation coeff.	$r_{jk} = \frac{1}{n-1} \sum_i z_{ij} z_{ik}$
Eigenvalue problem	$\mathbf{Rv} = \lambda \mathbf{v}$
Factor loading	$l_{jk} = v_{jk} \sqrt{\lambda_k}$
Explained variance	$\lambda_k / \sum_j \lambda_j \cdot 100\,\%$
Communality	$h_j^2 = \sum_{k=1}^{K} l_{jk}^2$
Component score	$t_{ik} = \sum_{j=1}^{p} z_{ij} v_{jk}$
Kaiser criterion	$\lambda_k \geq 1$