Factor Analysis

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 The Factor Model . . . 2
2.2 Communality and Uniqueness . . . 2
2.3 Reproduced Correlation Matrix . . . 2
2.4 Principal Axis Factoring (PAF) . . . 2
2.5 KMO Measure of Sampling Adequacy . . . 3
2.6 Bartlett's Test of Sphericity . . . 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: Correlation Matrix, KMO and Bartlett's Test . . . 4
3.3 Step 2: Eigenvalues (Kaiser Criterion) . . . 4
3.4 Step 3: PAF Extraction . . . 5

4 Using the Interactive Factor Analysis Tool **5**
4.1 Data Input . . . 5
4.2 Parameter Selection . . . 6
4.3 Results Tabs . . . 6
4.4 Typical Workflow . . . 6

5 Formula Reference **7**

6 Further Reading **7**

Introduction and Application Areas

Factor analysis is a method for identifying *latent factors* that underlie the observed variables. Unlike PCA, which merely seeks to maximise variance, factor analysis rests on an explicit measurement model: each variable is modelled as a linear combination of common factors and a variable-specific uniqueness.

Typical application areas:

- Scale construction in psychology and social sciences
- Validation of questionnaire items (e.g. Big Five)
- Indicator reduction for indices (e.g. quality of life)
- Identification of latent dimensions in market research data

Mathematical Background

The Factor Model

For p standardised variables $\mathbf{z} = (z_1, \ldots, z_p)^\top$:

$$z_j = l_{j1}F_1 + l_{j2}F_2 + \cdots + l_{jk}F_k + u_j, \qquad j = 1, \ldots, p \tag{1}$$

with common factors F_k ($\mathrm{Var}(F_k) = 1$, $\mathrm{Cov}(F_k, F_m) = 0$), factor loadings l_{jk} and specific variance u_j (uniqueness).

In matrix notation:

$$\mathbf{Z} = \mathbf{F}\mathbf{L}^\top + \mathbf{U} \tag{2}$$

Communality and Uniqueness

$$h_j^2 = \sum_{k=1}^{K} l_{jk}^2, \qquad u_j = 1 - h_j^2 \tag{3}$$

The **communality** h_j^2 is the share of variance of the j-th variable explained by the K factors; u_j is the unexplained residual share.

Reproduced Correlation Matrix

The factor model implies:

$$\mathbf{R} = \mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}, \qquad \mathbf{\Psi} = \mathrm{diag}(u_1, \ldots, u_p) \tag{4}$$

Extraction goal: minimise residuals $\mathbf{R} - \hat{\mathbf{R}}$.

Principal Axis Factoring (PAF)

The **Principal Axis Factoring** method is iterative:

1. Initial communality estimates: $h_j^{(0)} = \max_{m \neq j} |r_{jm}|^2$
2. Reduced correlation matrix: $\mathbf{R}^* = \mathbf{R}$ with r_{jj} replaced by h_j^2
3. Eigendecomposition of $\mathbf{R}^*$; extract the K largest positive eigenvalues and eigenvectors
4. Updated communalities: $h_j^{(t+1)} = \sum_{k=1}^{K} l_{jk}^2$
5. Repeat until convergence ($|h_j^{(t+1)} - h_j^{(t)}| < 10^{-6}$)

KMO Measure of Sampling Adequacy

The **Kaiser-Meyer-Olkin measure** assesses whether the correlation structure is suitable for factor analysis:

$$\text{KMO} = \frac{\sum_{j \neq m} r_{jm}^2}{\sum_{j \neq m} r_{jm}^2 + \sum_{j \neq m} p_{jm}^2} \tag{5}$$

where p_{jm} is the partial correlation between z_j and z_m.

Table 1: Interpretation of the KMO measure

KMO value	Interpretation
≥ 0.90	Marvellous
$0.80 - 0.89$	Meritorious (good)
$0.70 - 0.79$	Middling
$0.60 - 0.69$	Mediocre
< 0.60	Unacceptable

Bartlett's Test of Sphericity

H_0: The correlation matrix is an identity matrix ($\mathbf{R} = \mathbf{I}$).

$$\chi^2 \approx -\left(n - 1 - \frac{2p + 5}{6}\right) \ln |\mathbf{R}|, \quad \text{df} = \frac{p(p-1)}{2} \tag{6}$$

A significant result ($p < 0.05$) indicates that the correlation matrix differs from the identity matrix – a prerequisite for factor analysis.

Varimax Rotation

Identical to PCA (see Handbook 1). Objective: simple structure – each variable loads highly on as few factors as possible.

Worked Example: Balance-Sheet Ratios

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

The same synthetic dataset as in Handbook 1 is used (full data table there): 24 companies with three stability ratios (X_1 equity ratio, X_2 quick ratio, X_3 asset coverage II) and three profitability ratios (X_4 return on sales, X_5 return on assets, X_6 sales growth), all in per cent. The dataset is pre-loaded in the tool.

Goal of the factor analysis: identify the latent factors behind the ratios. Unlike PCA, factor analysis explains only the *common* variance of the ratios; the ratio-specific remainder enters the uniqueness u_j. Settings: PAF, Kaiser criterion, Varimax (tool defaults).

Step 1: Correlation Matrix, KMO and Bartlett's Test

The correlation matrix (see Handbook 1, Step 2) shows two blocks with correlations of 0.70–0.79 within and at most 0.31 between the blocks.

Table 2: Quality criteria

Criterion	Value	Assessment
KMO	0.678	mediocre (threshold 0.6 met)
Bartlett χ^2	80.68 (df = 15)	$p < 0.001$, significant

Bartlett calculation: $|\mathbf{R}| = 0.0183$, hence

$$\chi^2 = -\left(24 - 1 - \tfrac{2\cdot 6+5}{6}\right) \ln 0.0183 = -20.167 \cdot (-4.001) = 80.68, \qquad \mathrm{df} = \tfrac{6\cdot 5}{2} = 15.$$

The correlation matrix differs highly significantly from the identity matrix. The KMO value is only mediocre: each block consists of just three ratios, so the partial correlations within the blocks remain considerable. The minimum requirement KMO ≥ 0.6 is nevertheless met – factor analysis is admissible.

Step 2: Eigenvalues (Kaiser Criterion)

Table 3: Eigenvalues of the correlation matrix

Factor	Eigenvalue	Variance share	Cumulative
F1	2.971	49.52 %	49.52 %
F2	2.010	33.50 %	83.01 %
F3	0.397	6.61 %	89.62 %
F4–F6	< 0.3	–	–

Kaiser criterion: **2 factors** ($\lambda \geq 1$).

Step 3: PAF Extraction

Initial communalities (largest squared correlation per row): $h_1^{(0)} = 0.743^2 = 0.553$, $h_2^{(0)} = h_3^{(0)} = 0.771^2 = 0.595$, $h_4^{(0)} = h_6^{(0)} = 0.793^2 = 0.628$, $h_5^{(0)} = 0.727^2 = 0.529$. The diagonal of $\mathbf{R}$ is replaced by these values, the reduced matrix is decomposed, and the procedure is repeated until convergence.

Table 4: PAF loading matrix: unrotated and Varimax-rotated

	unrotated		Varimax			
Variable	F1	F2	F1	F2	h_j^2	u_j
X_1 Equity ratio	0.673	−0.497	**0.827**	0.122	0.699	0.301
X_2 Quick ratio	0.683	−0.522	**0.853**	0.112	0.739	0.261
X_3 Asset coverage	0.664	−0.595	**0.890**	0.047	0.795	0.205
X_4 Return on sales	0.741	0.486	0.183	**0.867**	0.785	0.215
X_5 Return on assets	0.618	0.506	0.082	**0.795**	0.638	0.362
X_6 Sales growth	0.654	0.631	0.019	**0.909**	0.827	0.173
Expl. variance	45.3 %	29.4 %	37.4 %	37.3 %	Σ 74.7 %	

Example: $h_1^2 = 0.827^2 + 0.122^2 = 0.699$, $u_1 = 1 - 0.699 = 0.301$. Together the factors explain 74.7 % of the total variance – less than the two principal components of the PCA (83.0 %), because PAF models only the common variance.

Reproduced correlation: the factor model reproduces, e. g., $\hat{r}_{12} = 0.827 \cdot 0.853 + 0.122 \cdot 0.112 = 0.719$ (observed: 0.710, residual −0.009) and $\hat{r}_{14} = 0.827 \cdot 0.183 + 0.122 \cdot 0.867 = 0.257$ (observed: 0.311). The small residuals confirm the two-factor solution.

Interpretation:

- **Factor 1 – financial stability:** equity ratio, quick ratio and asset coverage (loadings 0.83–0.89). It measures a company's asset strength and resilience to crises.
- **Factor 2 – profitability:** return on sales, return on assets and sales growth (loadings 0.80–0.91). It measures the ability to generate profits and to grow.
- All cross-loadings are below 0.2: the Varimax solution meets the simple-structure criterion very well. Return on assets has the lowest communality (0.638) and is thus explained least well by the factors, but it is clearly above the 0.4 threshold.

Using the Interactive Factor Analysis Tool

Data Input

Identical to PCA (Manual, CSV, Example). The balance-sheet ratios of the 24 companies are pre-loaded ($n = 24$, $p = 6$). Further examples: Job satisfaction ($n = 20$, $p = 6$), Personality scale (Big Five, $n = 25$, $p = 8$) and Consumer behaviour ($n = 18$, $p = 5$). 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

Extraction method
Principal Axis Factoring (PAF): Recommended; models latent factors correctly through iterative communality estimation.
Principal component method: Fast, but also explains specific variance.

Number of factors
Automatic: Kaiser criterion applied to the correlation matrix. Alternatively: specify 1–5 factors manually.

Rotation
Varimax (default): Orthogonal rotation; factors remain uncorrelated.
None: Unrotated solution for direct comparability with PCA.

Loading threshold
Highlight from $|l| \geq 0.4$ (adjustable). Values ≥ 0.5 are considered substantial.

KMO and Bartlett
Checkbox activates computation of the prerequisite checks.

Results Tabs

Overview
Key figures + eigenvalue table with Kaiser badge.

Loadings
Factor loading matrix with colour highlighting, communalities and uniqueness per variable.

Quality
KMO traffic light, Bartlett test with p-value, communality table with quality assessment.

Charts
Scree plot, variance bar chart with cumulative line, loading plot (F1 vs. F2) with variable labels.

Interpretation
Automatic assessment text covering KMO, Bartlett and rotation notes.

Export
CSV: loadings, communalities, eigenvalues. PNG: scree plot, loading plot.

Typical Workflow

1. Load data and select extraction method PAF.
2. Leave the KMO checkbox active.
3. Click **"Calculate Factor Analysis"**.
4. Quality tab: verify KMO ≥ 0.6 and Bartlett significant.
5. Scree plot: adjust the number of factors manually if needed.
6. Loadings tab: interpret the factors substantively.

7. Consider removing variables with $h_j^2 < 0.4$ and recalculating.

Important notes:

- At least $n = 5$ complete observations are required.
- Rule of thumb: $n \geq 10 \cdot p$ for reliable results.
- Variables with $h_j^2 < 0.3$ are poorly explained – consider removing them.
- When cross-loadings occur ($|l_{jk}| \geq 0.3$ on multiple factors): consider oblique rotation (Oblimin).

Formula Reference

Table 5: Factor analysis formulae

Name	Formula
Factor model	$z_j = \sum_k l_{jk} F_k + u_j$
Reproduced correlation	$\mathbf{R} = \mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}$
Communality	$h_j^2 = \sum_k l_{jk}^2$
Uniqueness	$u_j = 1 - h_j^2$
KMO measure	$\text{KMO} = \sum r_{jm}^2 / (\sum r_{jm}^2 + \sum p_{jm}^2)$
Bartlett χ^2	$-(n - 1 - \frac{2p+5}{6}) \ln \lvert\mathbf{R}\rvert$, $\text{df} = p(p-1)/2$

Further Reading

- Comrey, A. L. & Lee, H. B. (1992). *A First Course in Factor Analysis*. LEA.
- Backhaus, K. et al. (2021). *Multivariate Analysemethoden*, 16th ed. Springer Gabler.
- Fabrigar, L. R. et al. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299.