PCA & EFA: A Visual Guide

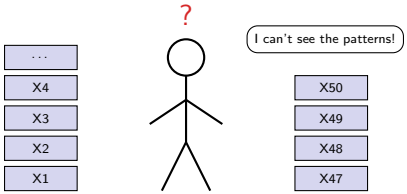
A Formula-Free Introduction

Statistical Data Analysis

Lesson 3

March 13, 2026

Can You See the Patterns?

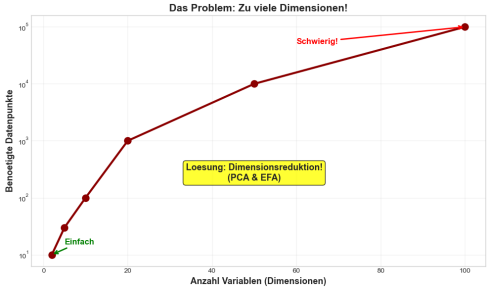


When you have 50 variables, where do you even start?

Dimensionality is the central challenge – PCA and EFA are your tools to tame it.

Why Do Dimensions Overwhelm Us?

- Every added variable increases noise and makes visualization impossible
- The “curse of dimensionality” – distances become meaningless in high dimensions



High-dimensional data needs reduction before analysis can begin.

What Will You Learn Today?

1. **Compress data** without losing insight (PCA)
2. **Discover hidden causes** behind correlations (EFA)
3. **Know when** to use PCA vs EFA

Three skills that turn overwhelming data into clear structure.

Does Compression Actually Work?

- 50 variables reduced to just 3 principal components
- Over 90% of the original information is preserved

PCA gives you a massive reduction with minimal information loss.

What Is PCA in One Sentence?

PCA = *"Find the camera angle that shows the most spread."*

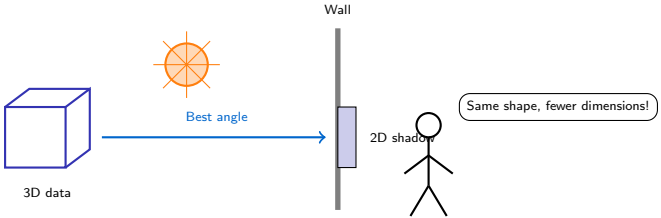
Three ways to think about it:

- **Photographer:** Rotate the camera until you see the widest panorama
- **Shadow:** Shine a light on a 3D object and find the shadow that keeps the most shape
- **Headline:** Summarize a long article in one sentence that captures the main point

---

PCA finds the directions of maximum variation in your data.

How Does Projection Work?

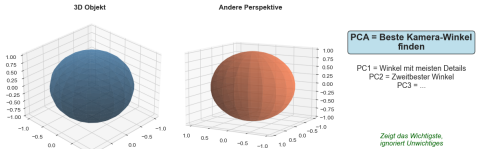


PCA finds the wall angle that preserves the most shape.

Projection reduces dimensions while keeping the essential structure.

What Does the Best Angle Look Like?

- PC1 = the direction of maximum spread in your data
- PC2 = the next best direction, perpendicular to PC1



Each principal component captures progressively less variance.

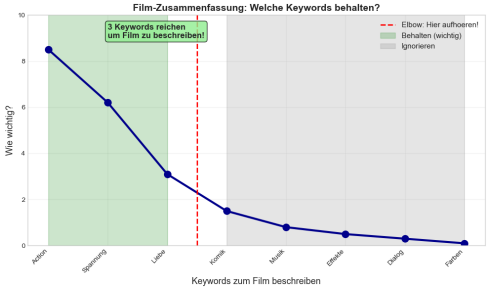
How Do You Read a Scree Plot?

- **Bars** = importance of each principal component (tallest = most variance)
- **Look for the "elbow"** = the point where bars flatten out
- **Keep PCs above the elbow** – everything below adds noise, not signal

The scree plot is your primary tool for deciding how many components to keep.

Where Is the Elbow?

- The keyword analogy makes the “elbow” concept intuitive
- Kaiser rule: keep components with above-average importance



Multiple criteria (elbow, Kaiser, parallel analysis) should agree before you decide.

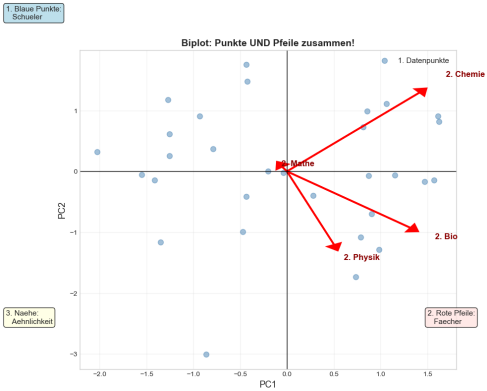
What Is a Biplot Telling You?

- **Arrows** = original variables – direction shows which PC they load on
- **Dots** = observations – their position shows how they score on PCs
- **Arrow proximity** = arrows pointing the same way mean correlated variables

A biplot shows variables and observations together in reduced PC space.

Can You Read This Biplot?

- Follow the step-by-step guide: arrows first, dots second, angles last
- Close arrows = correlated variables; opposite arrows = negatively correlated



Practice reading biplots – they are the main PCA output you will interpret.

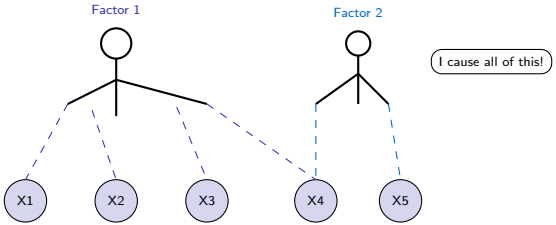
What Can PCA NOT Tell You?

- PCA **compresses** but does not explain *why* variables are correlated
- It finds directions of maximum spread, not hidden causes
- It cannot tell you what the components *mean* in theory

To move from "what" to "why", we need a different tool...

PCA is descriptive – it summarizes but does not explain.

What If Something Hidden Is Pulling the Strings?

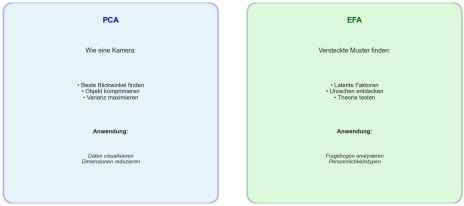


EFA looks for the puppet master behind your data.

EFA assumes hidden factors cause the correlations you observe.

PCA vs EFA – What Is the Real Difference?

- PCA = **compress** – uses all variance to reduce dimensions
- EFA = **explain** – uses only shared variance to find hidden causes



PCA asks “how can I simplify?” – EFA asks “what is causing this?”

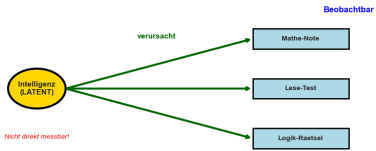
What Is a Latent Factor?

- A **hidden cause** you cannot directly measure but whose effects you can observe
- Example: "intelligence" is latent – you measure its effects via math, reading, and logic scores
- Factors **explain correlations**: if math and reading scores move together, a common factor may drive both

Latent factors are the unobservable drivers behind observable patterns.

Can You See the Hidden Factors?

- Arrows from factor to variables = causal influence
- Thicker arrows = stronger loadings (variable depends more on that factor)



Factor diagrams show the hidden structure behind your observed variables.

How Many Factors Should You Keep?

- **Parallel analysis** = gold standard method for deciding
- Compare your eigenvalues to eigenvalues from random data of the same size
- Keep only the factors that **beat the random benchmark** – those carry real signal

Parallel analysis protects you from extracting noise factors.

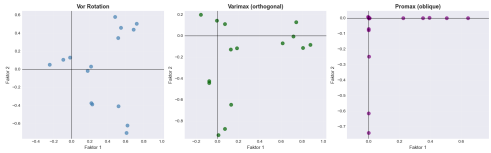
Why Does Rotation Matter?

- The initial factor solution is mathematically valid but **hard to interpret**
- Rotation redistributes loadings so each variable loads strongly on **one** factor
- Goal: achieve "simple structure" where the pattern is clear and clean

Rotation does not change the fit – it changes how the solution looks.

See How Rotation Cleans Up the Picture!

- Before rotation: messy loadings spread across multiple factors
- After rotation: clear simple structure – each variable "belongs" to one factor



Rotation is not optional – always rotate for interpretable results.

Varimax or Promax – Which Rotation?

- **Varimax** = factors stay uncorrelated (orthogonal) – simpler to interpret
- **Promax** = factors allowed to correlate (oblique) – more realistic in social sciences
- Default recommendation: **start with Varimax**; switch to Promax if factors should logically correlate

When in doubt, try both and compare the interpretability of the results.

How Do You Run This in R?

Three lines are all you need:

1. **PCA:** `prcomp(data, scale = TRUE)`
2. **EFA (base R):** `factanal(data, factors = 3)`
3. **EFA (psych):** `fa(data, nfactors = 3, rotate = "varimax")`

The psych package gives you parallel analysis, rotation options, and publication-ready output.

Start with these one-liners, then explore the output objects for details.

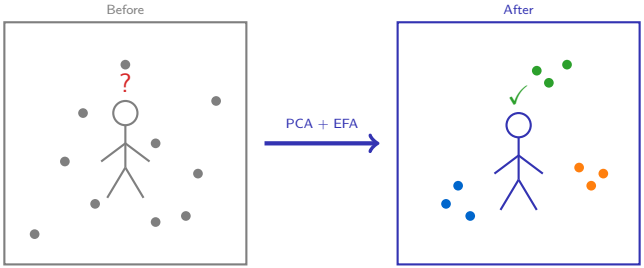
Remember the Three Big Takeaways!

1. **PCA compresses** – use it when you want fewer variables with minimal information loss
2. **EFA explains** – use it when you suspect hidden causes driving your observed correlations
3. **Always check** scree plots and parallel analysis before deciding how many components or factors to keep

---

PCA and EFA are complementary – learn both and choose based on your research question.

Can You See the Patterns Now?



From drowning in variables to seeing the structure.

PCA and EFA transform confusion into clarity – that is the power of dimensionality reduction.

Where Do You Go From Here?

- **Try PCA** on a real dataset – one line in R: `prcomp(data, scale = TRUE)`
- **Compare PCA and EFA** results side by side to see how they differ
- **Read the full technical lecture** for the mathematical foundations behind these methods

---

The best way to learn PCA and EFA is to run them on data you care about.