PCA & Exploratory Factor Analysis Quick Reference

PCA: Principal Component Analysis

Goal: Reduce dimensions, maximize variance

Steps:
1. Standardize data (mean=0, sd=1)
2. Compute correlation matrix $\mathbf{R}$
3. Eigendecomposition: $\mathbf{Rv} = \lambda\mathbf{v}$
4. Project data onto eigenvectors

Key formula:
$\text{PC}_j = v_{1j}X_1 + v_{2j}X_2 + \cdots + v_{pj}X_p$

Variance explained by PC_j:

$$\frac{\lambda_j}{\sum_i \lambda_i} \times 100\%$$

How many components?
Kaiser: keep if $\lambda > 1$
Scree: look for "elbow"
Cumulative variance $\geq 80\%$

Loadings vs. Scores:
Loading: correlation of variable with PC
Score: each observation's value on PC

EFA: Exploratory Factor Analysis

Goal: Find latent factors causing correlations

Model: $\mathbf{X} = \mathbf{\Lambda F} + \boldsymbol{\varepsilon}$
$\mathbf{X}$ = observed variables
$\mathbf{\Lambda}$ = loading matrix
$\mathbf{F}$ = latent factors
$\boldsymbol{\varepsilon}$ = unique variance

Pre-checks:
KMO > 0.6 (sampling adequacy)
Bartlett $p < 0.05$ (not identity matrix)

How many factors?
Parallel analysis (recommended)
Scree plot
Theory / content knowledge

Loading interpretation:
$|\lambda| > 0.4$: meaningful
$|\lambda| > 0.7$: strong

Communality:

$$h_i^2 = \sum_j \lambda_{ij}^2 \quad \text{(shared variance)}$$

PCA vs. EFA: When to Use

PCA:
Data compression / reduction
No model assumptions
Components = linear combos of *all* variables
Use when: preprocessing, visualization

EFA:
Discover latent structure
Model-based (factors *cause* correlations)
Factors explain correlations, not total variance
Use when: questionnaire validation, scale construction

Decision:
"Reduce variables?" → PCA
"Find hidden constructs?" → EFA
"Factors expected to correlate?" → promax
"Factors independent?" → varimax

Key distinction:
PCA: total variance (including unique)
EFA: shared variance (communality) only

Rotation Methods

Varimax (orthogonal):
Factors stay uncorrelated
Maximizes simple structure
Most common default

Promax (oblique):
Factors may correlate
Check $\mathbf{\Phi}$ matrix for correlations
More realistic for psychology / social science

Rule of thumb:
If $\mathbf{\Phi}$ correlations < 0.3 → varimax sufficient
If $\mathbf{\Phi}$ correlations ≥ 0.3 → use promax

Simple structure (Thurstone):
Each variable loads high on one factor
Each factor has some high, some near-zero loadings

R Commands

PCA:

```
pca <- prcomp(data, scale.=TRUE)
summary(pca)                          variance explained
pca$rotation                          loadings
pca$x                                 scores
biplot(pca)                           visualization
screeplot(pca)                        scree plot
```

EFA:

```
library(psych)
KMO(cor(data))                        sampling adequacy
fa.parallel(data)                     number of factors
fa(data, nfactors=k, rotate="varimax")
fa(..., rotate="promax")
$loadings                             factor loadings
$communalities                        shared variance
$Phi                                  factor correlations (promax)
```

Common Mistakes

PCA:
Forgetting `scale.=TRUE`
Interpreting PCA loadings as factor loadings (they differ)
Keeping too many components (use scree + variance)

EFA:
Choosing # factors by Kaiser alone
→ use parallel analysis instead
Using varimax when factors should correlate
Reporting unrotated loadings
Ignoring communalities < 0.3 (poorly explained items)
Not checking KMO / Bartlett before EFA

Both:
Using PCA when you need EFA (or vice versa)
Skipping standardization
Over-interpreting borderline loadings (0.3–0.4)

Always report: KMO/Bartlett, # factors + criterion, rotation method, rotated loadings, variance explained, communalities