STAT 24620=FINM 34700, STAT 32950
Multivariate Data Analysis
Lecture 4: Factor Analysis

Jingshu Wang

The University of Chicago

Outline

PCA $\rightarrow$ Factor analysis

- Lecture 1: multivariate normal setup and covariance structure.
- Lecture 2: PCA as orthogonal variance-maximizing projections.
- Lecture 3: PCA workflow + PPCA + practical interpretation.

Today: turn "directions" into a latent variable model.

- Separate common variation from variable-specific noise.
- Learn what is and is not identifiable in factor models.
- Estimate factors/loadings and rotate for interpretation.

Core factor analysis model

For observed vector $\mathbf{x}_i \in \mathbb{R}^p$ and latent factors $\mathbf{f}_i \in \mathbb{R}^m$ $(m < p)$:

$$\mathbf{x}_i = \boldsymbol{\mu} + L\mathbf{f}_i + \boldsymbol{\epsilon}_i,$$

with assumptions

$$\mathbb{E}[\mathbf{f}_i] = \mathbf{0}, \quad \mathrm{Cov}(\mathbf{f}_i) = I_m, \quad \mathbb{E}[\boldsymbol{\epsilon}_i] = 0, \quad \mathrm{Cov}(\boldsymbol{\epsilon}_i) = \Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p),$$

and $\mathrm{Cov}(\mathbf{f}_i, \boldsymbol{\epsilon}_i) = 0$.

- $L \in \mathbb{R}^{p \times m}$: loading matrix.
- Ψ: uniqueness (idiosyncratic variance) matrix.

Covariance decomposition

Under the model,

$$\Sigma = \mathrm{Cov}(\mathbf{x}_i) = LL^\top + \Psi.$$

Interpretation by variable j:

$$\mathrm{Var}(X_{ij}) = \underbrace{\sum_{k=1}^{m} l_{jk}^2}_{\text{communality } h_j^2} + \underbrace{\psi_j}_{\text{uniqueness}}.$$

- **Communality** h_j^2: variation explained by common factors.
- **Uniqueness** ψ_j: variable-specific noise/residual variance.

Conceptual goal of FA: explain covariance, not necessarily maximize total variance as PCA does.

Relation to PPCA

PPCA model:

$$\mathbf{x}_i = W\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \quad \mathbf{z}_i \sim \mathcal{N}(0, I_q), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2 I_p),$$

FA model generalizes PPCA by replacing isotropic noise with diagonal noise:

$$\sigma^2 I_p \longrightarrow \Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p).$$

- PPCA: equal uniqueness variance across variables.
- FA: variable-specific uniquenesses, often more realistic.

Takeaway: FA is more flexible allowing heterogeneous measurement noise.

Geometric vs probabilistic views of PCA and FA

PCA (geometric view)

- deterministic orthogonal directions;
- maximizes variance of projections;
- decomposes total variance;
- no explicit noise model.

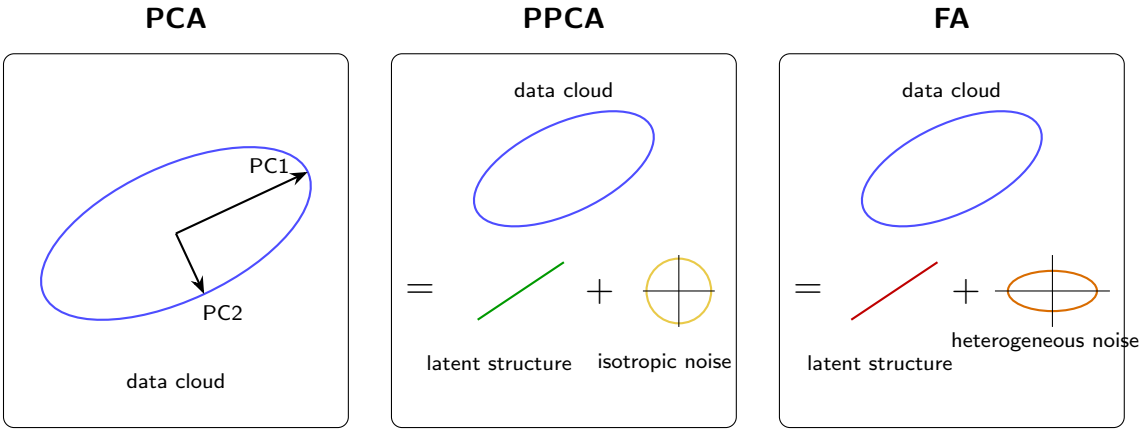
Factor Analysis (probabilistic view)

- stochastic latent factors;
- covariance: $\Sigma = LL^\top + \Psi$;
- separates common and idiosyncratic variation;
- includes variable-specific noise (Ψ).

Bridge: PPCA

- special case of FA with $\Psi = \sigma^2 I_p$;
- maximum likelihood solution recovers PCA directions.

PCA vs PPCA vs Factor Analysis



Another visual comparison between PCA and FA

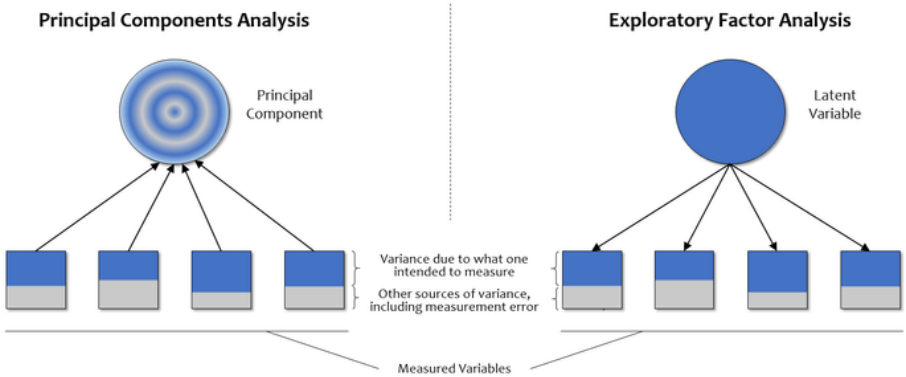


Figure: Figure from `https://community.jmp.com/t5/JMP-Blog/Principal-components-or-factor-analysis/ba-p/38347`

FA vs PCA: orthogonality and variance normalization

PCA

$$\mathbf{X}_c \approx Z_m V_m^\top$$

- Loading vectors are **orthonormal**: $V_m^\top V_m = I$.
- Score coordinates are uncorrelated with different variances:

$$\mathrm{Cov}(\mathbf{z}_i) = \mathrm{diag}(\lambda_1, \ldots, \lambda_m).$$

- Scale is absorbed into scores

Factor Analysis

$$\mathbf{x}_i = \boldsymbol{\mu} + L\mathbf{f}_i + \boldsymbol{\epsilon}_i$$

- Columns of the loading matrix L are **not required to be orthogonal**.
- Latent factors are standardized and uncorrelated:

$$\mathrm{Cov}(\mathbf{f}_i) = I_m.$$

- Scale is absorbed into loadings.

Why diagonal Ψ is strong but useful

Assuming diagonal Ψ means uniqueness components are uncorrelated across variables.

- Good approximation when common dependence is captured by factors.
- Keeps parameter count manageable.

Parameter counting intuition:

- Unrestricted covariance: $p(p+1)/2$ parameters.
- FA model: $pm + p$ parameters before identifiability adjustments.

If $m \ll p$, FA can be a strong dimensionality reduction for covariance modeling.

Identifiability and rotational invariance

If $Q \in \mathbb{R}^{m \times m}$ is orthogonal ($Q^\top Q = I_m$), then

$$L\mathbf{f}_i = (LQ)(Q^\top \mathbf{f}_i),$$

and since $Q^\top \mathbf{f}_i$ still has covariance I_m, the covariance of $\mathbf{x}_i$ is unchanged:

$$LL^\top = (LQ)(LQ)^\top.$$

Consequence: factors/loadings are not unique without additional conventions.

- "Raw" factor directions are only identified up to rotation/sign/permutation.
- Interpretation needs a rotation criterion.

From covariance model to likelihood

From the factor model,

$$\mathrm{Cov}(\mathbf{x}_i) = \Sigma = LL^\top + \Psi.$$

Goal: estimate (L, Ψ) so that Σ matches the sample covariance $\mathbf{S}$.

Under Gaussian assumptions:

$$\mathbf{f}_i \sim \mathcal{N}(0, I_m), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \Psi) \;\Rightarrow\; \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma).$$

This allows likelihood-based estimation.

Maximum likelihood estimation for FA

Under multivariate normality, the log-likelihood is

$$\ell(L, \Psi) \propto -\log|\Sigma| - \mathrm{tr}(\mathbf{S}\Sigma^{-1}), \quad \Sigma = LL^\top + \Psi.$$

Interpretation:

- choose Σ to approximate the sample covariance $\mathbf{S}$;
- decompose variability into
  - common structure: $LL^\top$,
  - variable-specific noise: Ψ.

Challenge:

- nonlinear parameterization $\Rightarrow$ no closed-form solution.

Estimation via EM: key idea

Latent factors $\mathbf{f}_i$ are unobserved.

Idea: treat latent factors as missing data.

E-step:

- estimate the latent factors given the observed data and current parameters.

M-step:

- update (L, Ψ) by fitting the model to these estimated factors.

Repeat until convergence.

Interpretation:

- alternates between estimating hidden structure and updating model parameters

This procedure is known as the EM algorithm.

Alternative estimation methods

Besides MLE, one can use:

- **The principal component / factor method:** approximate $S \approx LL^\top + \Psi$ using PCA as solutions or initializations;
- **MINRES:** minimize residual off-diagonal covariance.

Remarks:

- do not require full distributional assumptions;
- often simpler but less statistically efficient than MLE.

How many factors m?

Focus on two complementary diagnostics:

- **Scree plot / PA:** start from the eigenvalues of the sample *correlation* matrix.
  - Exactly the same steps as for PCA
- **Residual correlations after FA fit:** fit FA with a candidate m and check whether the off-diagonal entries of

$$R - \hat{R}$$

  are small. Large residual blocks mean some common dependence is still unexplained.

Interpretation:

- too small $m \Rightarrow$ scree / parallel analysis suggests more factors and residual correlations remain structured;
- increase m until residual structure is mostly negligible.

Choose the smallest m that removes systematic residual correlation and still gives interpretable factors.

Rotation for interpretation

Because of rotational invariance, we can choose a rotation to make factors interpretable

Orthogonal rotation (eg., varimax):

$$L \to LQ, \quad \mathbf{f}_i \to Q^\top \mathbf{f}_i, \quad Q^\top Q = I_m$$

- factors remain uncorrelated: $\mathrm{Cov}(\mathbf{f}_i) = I_p$;

Oblique rotation (eg., oblimin / promax):

$$L \to LQ, \quad \mathbf{f}_i \to Q^{-1} \mathbf{f}_i$$

- factors may become correlated: $\mathrm{Cov}(\mathbf{f}_i) \neq I_p$;

Typical outcome:

- sparse loadings (each variable loads on few factors);
- easier interpretation ("market", "sector", etc.).

Illustrating factor rotation

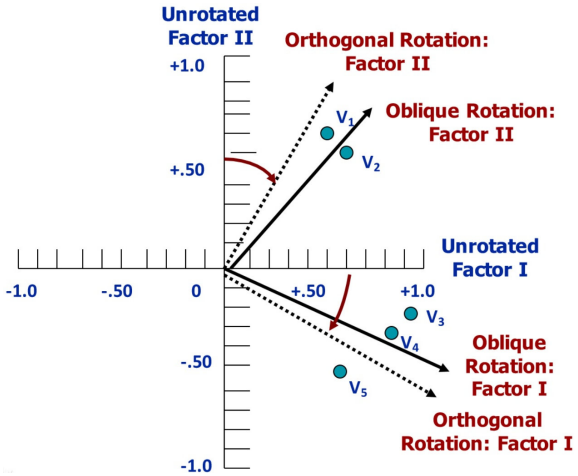


Figure: Source: `https://devopedia.org/factor-analysis`

Varimax criterion

For the loading matrix L and columns $l_{\cdot k}$, Varimax seeks a rotation maximizing spread of squared loadings:

$$\max_{Q} \sum_{k=1}^{m} \left[\frac{1}{p} \sum_{j=1}^{p} \tilde{l}_{jk}^{4} - \left(\frac{1}{p} \sum_{j=1}^{p} \tilde{l}_{jk}^{2} \right)^{2} \right], \quad \tilde{L} = LQ.$$

- Pushes loadings toward "large or near-zero" patterns.
- Helps create simple structure for communication.

A rotation example

Before

$$L = \begin{pmatrix} 0.60 & 0.67 \\ 0.53 & 0.67 \\ 0.42 & 0.71 \\ -0.57 & 0.71 \\ -0.46 & 0.74 \\ -0.35 & 0.78 \end{pmatrix}$$

After (varimax)

$$\tilde{L} = \begin{pmatrix} 0.90 & 0.10 \\ 0.84 & 0.14 \\ 0.79 & 0.24 \\ 0.00 & 0.90 \\ 0.16 & 0.86 \\ 0.26 & 0.81 \end{pmatrix}$$

After (promax)

$$\tilde{L} = \begin{pmatrix} 0.93 & 0.00 \\ 0.86 & 0.00 \\ 0.79 & 0.00 \\ -0.12 & 0.94 \\ 0.00 & 0.87 \\ 0.12 & 0.81 \end{pmatrix}$$

- Before: variables load on both factors.
- After: each variable loads mainly on one factor.

Rotation reveals a simple structure.

How to use factor analysis (in practice)

1. choose number of factors m;
2. fit FA model;
3. rotate loadings for interpretation;
4. check and present residual correlations for model fit.

$$R_{\text{res}} = R - \hat{R}.$$

Inspect the largest off-diagonal residuals; persistent blocks suggest missing factors.

- **RMSR (root-mean-square-error) summary:**

$$\text{RMSR} = \sqrt{\frac{2}{p(p-1)} \sum_{i<j} r_{ij,\text{res}}^2}$$

computed from residual *correlations* gives one-number fit quality.

Goal: simple, interpretable structure with small residual correlations.

How do we get estimated factor scores?

After estimating Λ and Ψ, we often want a score estimate $\hat{\mathbf{f}}_i$ for each observation $\mathbf{x}_i$.

Regression scores:

$$\hat{\mathbf{f}}_i = \hat{\mathbb{E}}[\mathbf{f}_i \mid \mathbf{x}_i] = \hat{L}(\hat{L}\hat{L}^\top + \hat{\Psi})^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) = (I_m + \hat{L}^\top \hat{\Psi}^{-1}\hat{L})^{-1}\hat{L}^\top \hat{\Psi}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}).$$

- This is the conditional mean of the latent factor under the Gaussian FA model.
- It gives the best linear predictor of $\mathbf{f}_i$ from the observed variables.
- In R, this is what `factanal(..., scores = "regression")` returns.

Interpretation:

- scores summarize where an observation lies along the latent factors;
- unlike PCA scores, these are model-based and depend on both L and Ψ;

Live R notebook

Notebook file: `Lecture4_demo.nb.html`

Common pitfalls in factor analysis

- Rotated factors are not uniquely defined (rotation invariance).
- Rotation does not necessarily improve interpretation.
- More factors improve fit, but may reduce interpretability.
- Factors capture statistical patterns, not necessarily causal mechanisms.

Interpret factors cautiously.

Lecture 4 summary

1. Factor analysis models covariance through a small number of latent factors plus idiosyncratic noise.
2. Communalities capture shared variation; uniquenesses capture variable-specific noise.
3. PCA is a variance decomposition; factor analysis is a generative model for covariance.
4. Estimation balances fit and structure.
5. Rotation improves interpretability without changing model fit.
6. In practice, factor models provide an interpretable structure for multivariate data.

Suggested reading

- Johnson & Wichern (6th Edition), Chapter 9.
- Detailed derivations on EM for factor analysis:
  `https://www.math.hkbu.edu.hk/~hpeng/Math3806/EM-factor.html`