

(Exploratory) Factor Analysis



Ying-Chao Hung
Department of Statistics
National Chengchi University
hungy@nccu.edu.tw

Introduction

- The origins of FA date back to 1905.
- Spearman hypothesized that students' performance in various courses are intercorrelated and can be explained by students' "general intelligence" levels.

Main Idea

- The essential purpose of FA is to describe the relationships between variables by a **small number** of underlying, but **unobservable (latent)** variables, called **factors**.
- In particular, the goal is to examine whether the observed variables are **linear functions of a small number of latent variables**.

Example

- ▣ Students' test scores (grades) are:

Math (M), Physics (P), Chemistry (C), English (E), History (H), French (F).

- ▣ Assume “grades” are a function of their **general intelligence level (I)** and their **aptitudes (A)**.

We then consider the following equations:

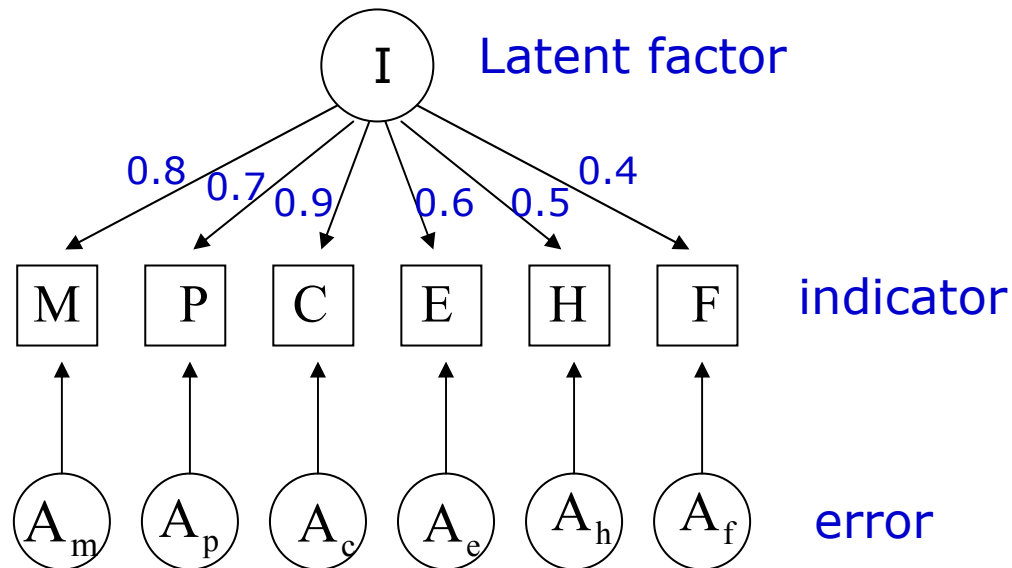
$$M = 0.8I + A_m, \quad P = 0.7I + A_p$$

$$C = 0.9I + A_c, \quad E = 0.6I + A_e$$

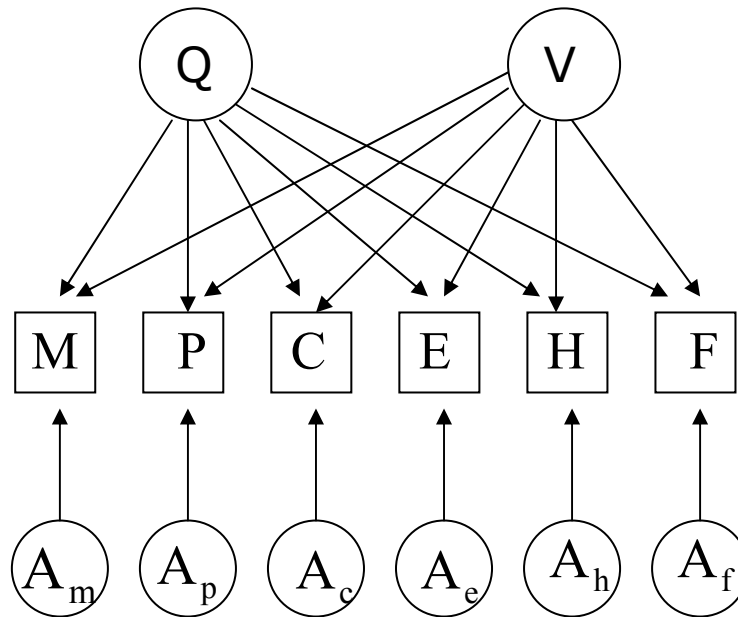
$$H = 0.5I + A_h, \quad F = 0.4I + A_f$$

Example (continued)

- ❑ The coefficients are called **pattern loadings**.
- ❑ The relationships can be viewed as a set of regression equations.
- ❑ The relationship can be shown as



Example (two-factor model)



The FA Model

- Data matrix: $X_{N \times m}$
- Denote the observed variables by X_1, X_2, \dots, X_m
- Denote the latent variables by F_1, F_2, \dots, F_k
- The *k*-factor model can be presented as

$$X_1 = \sum_{j=1}^k \lambda_{1j} F_j + U_1 = \lambda_{11} F_1 + \dots + \lambda_{1k} F_k + U_1$$

$$X_2 = \sum_{j=1}^k \lambda_{2j} F_j + U_2 = \lambda_{21} F_1 + \dots + \lambda_{2k} F_k + U_2$$

\vdots

The FA Model (continued)

■ In a matrix notation: $X = \Lambda F + U$ where

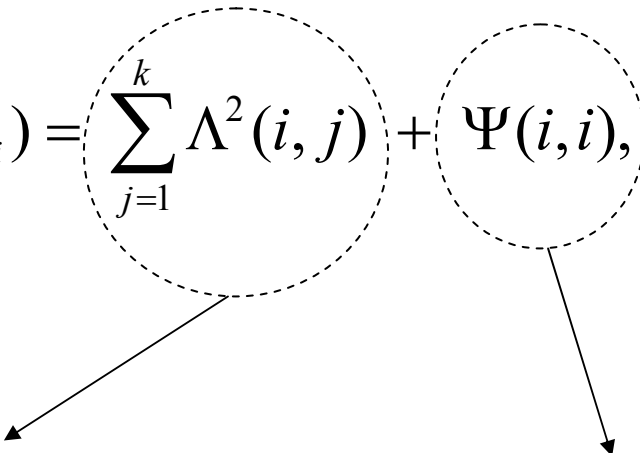
$$\begin{cases} \Lambda : m \times k \text{ matrix of factor loadings} \\ F : k \times 1 \text{ vector of common factors} \\ U : m \times 1 \text{ vector of error terms} \end{cases}$$

■ **Assumptions:**

- (i) F and U are independent ($\text{Cov}(F, U) = 0$)
- (ii) $E(F) = 0$ and $\text{Cov}(F) = I_k$ (i.e. $\text{Cov}(F_i, F_j) = 0, i \neq j$)
- (iii) $E(U) = 0$ and $\text{Cov}(U) = \Psi$, Ψ is an $m \times m$ diagonal matrix.
(i.e. $\text{Cov}(U_i, U_j) = 0, i \neq j$)
- (iv) X is standardized

Variance Decomposition

It is clear that

$$\text{Var}(X_i) = \sum_{j=1}^k \Lambda^2(i, j) + \Psi(i, i), \quad i = 1, \dots, m.$$


The diagram shows the equation $\text{Var}(X_i) = \sum_{j=1}^k \Lambda^2(i, j) + \Psi(i, i)$. The first term, $\sum_{j=1}^k \Lambda^2(i, j)$, is enclosed in a dashed circle. An arrow points from the bottom of this circle to the definition of Communality. The second term, $\Psi(i, i)$, is also enclosed in a dashed circle. An arrow points from the bottom of this circle to the definition of Uniqueness.

Communality, the variance
shared with other variables
via the common factors

Uniqueness, the variance
not shared with other
variables

The Matrix Form

□ Let Σ be the **covariance matrix** of X , then

$$\begin{aligned}\text{Cov}(X) &= \Sigma = E(XX') \\ &= E[(\Lambda F + U)(\Lambda F + U)'] \\ &= \Lambda E(FF')\Lambda' + 2\Lambda \underbrace{E(FU')}_0 + E(UU') \\ &= \Lambda\Lambda' + \Psi\end{aligned}$$

□ **Goal**: to determine Λ , Ψ , and F .

Rotational Indeterminacy of the Solution

- ▣ The solution to the previous problem is **NOT** unique !!

Consider **any orthonormal matrix** T , i.e.,

$$T'T = TT' = I_k$$

Let $X = (\Lambda T)(T'F) + U$, then

$$\begin{aligned}\text{Cov}(X) = \Sigma &= (\Lambda T)(\Lambda T)' + \Psi \\ &= \Lambda\Lambda' + \Psi\end{aligned}$$

- ➔ The solution Λ is not unique, since ΛT does not alter the decomposition of Σ .

Estimation of the FA Model

- First, use the sample covariance matrix S as the estimate of the covariance matrix Σ .
- Two approaches:
 - Principal Factor Analysis
 - Maximum Likelihood Factor Analysis

Principal Factor Analysis (PFA)

- In practice, it is hard to find $\hat{\Lambda}$ and $\hat{\Psi}$ such that

$$S = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi} \quad \text{holds exactly.}$$

- Let $\hat{S} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$, the goal now becomes to find $\hat{\Lambda}$ and $\hat{\Psi}$ such that

$$\text{trace}(S - \hat{S})'(S - \hat{S}) \quad \text{is minimized.}$$

Note: This is equivalent to minimizing the sum of squared errors between two matrices.

Algorithm of PFA

(I) Guess $\hat{\Psi}$

(II) Write $(S - \hat{\Psi}) = U\Sigma U'$, let $\hat{\Lambda} = U_k \Sigma_k^{1/2}$,

where assume the first k eigenvalues are positive.

(III) Set $\hat{\Psi} = \text{diag}(S - \hat{\Lambda}\hat{\Lambda}')$

Then repeat steps (II) and (III) until convergence.

■ Remark: In case $\hat{\Psi} = 0$, this corresponds to PCA. The factor loadings then corresponds to rescaled versions of the PCs.

Maximum Likelihood Factor Analysis

▣ Assume that $F \sim N(0, I_k)$ and $U \sim N(0, \Psi_{m \times m})$

The **log-likelihood function** is then given by

$$l(\Lambda, \Psi) = -\frac{1}{2} Nm \log(2\pi) - \frac{N}{2} \log |\Lambda \Lambda' + \Psi| \\ - \frac{N}{2} \text{trace}((\Lambda \Lambda' + \Psi)^{-1} S)$$

➔ Can find $\hat{\Lambda}$ and $\hat{\Psi}$ which maximize $l(\Lambda, \Psi)$.

Remarks

- What does the maximum likelihood estimators (MLE) mean ?
- In some cases where $l(\Lambda, \Psi)$ is maximized with $\hat{\Psi}(i, i) < 0$.
In such cases we have to force $\hat{\Psi}(i, i) = 0$.
(so-called the Heywood cases, also happens in PFA)

Number of Factors

□ Confirmatory FA:

Assume that the factor structure is known, that is, **the number of factors is hypothesized**.

□ Exploratory FA:

Assume that we have no knowledge about the underlying structure of the factor models, that is, we don't know the number of factors.

The Maximum Number of Factors

- Examine the model's degrees of freedom, remember

$$S = \Lambda\Lambda' + \Psi$$

$df = m$ \rightarrow $\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix}$

$df = 1 + 2 + \cdots + m = \frac{m(m+1)}{2}$

For Λ , $df = m \times k$.
 But the rotational freedom for Λ is $\frac{k(k-1)}{2}$.
 (Since there are C_2^k equations $\langle F_i, F_j \rangle = 0, i \neq j$.)

$$df = mk - \frac{k(k-1)}{2}$$

$df = m$

The Maximum # of Factors (continued)

- Note that to estimate the unknown parameters Λ and Ψ , the model needs $df \geq 0$ (for estimating the errors).

Since the degrees of freedom for this model is

$$df = \frac{m(m+1)}{2} - \left[mk - \frac{k(k-1)}{2} + m \right]$$

➔ The maximum k is the one making the $df \geq 0$.

How to choose the best k ?

▣ An Informal Method:

Check out the percentage of the total variance accounted, as we did in PCA.

▣ A More Formal Method:

Consider testing the hypotheses

$$\begin{cases} H_0 : \Sigma = \Lambda\Lambda' + \Psi \text{ with } k \text{ common factors} \\ H_a : \Sigma \text{ is unconstrained} \end{cases}$$

➔ Can construct a likelihood ratio test by assuming normality and using the MLE.

Factor Rotations

- Remember ΛT for any orthonormal matrix T does not alter the decomposition of Σ .
- ➔ A rotation by T may succeed in revealing a “simpler structure” of factor models.

However, in many cases the solution **may be difficult to interpret** (e.g. factors have similar large loadings, or both positive and negative loadings, etc)

Orthogonal/Non-orthogonal Rotation

□ Orthogonal Rotations:

Factors are still linearly independent, but maximizing some other criterion.

- **Varimax Rotation** (e.g. keeping large/small loadings for easier interpretation)
- **Quartimax Rotation**

□ Oblique Rotations: (non-orthogonal rotation)

Might provide **easier interpretation** of the solution, but **losing the orthogonality of the factors** (i.e. factors are correlated)

- **Promax** (in R)

Factor Scores

- How do we calculate an object's location (called **factor score**) in the obtained factor space?

- **Deterministic Factor Scores: Bartlett's Method**

Denote the factor score of the i -th observation x_i by f_i .

Since $X = \Lambda F + U$, we can reasonably assume: $X_i | f_i \sim N(\Lambda f_i, \Psi)$.

Hence, for one observation x_i , the log likelihood is given by

$$-\frac{1}{2} \log |2\pi \Psi| - \frac{1}{2} (x_i - \Lambda f_i)' \Psi^{-1} (x_i - \Lambda f_i).$$

The MLE of f_i is then given by (setting the derivative to zero):

$$\hat{f}_i = (\Lambda' \Psi^{-1} \Lambda)^{-1} \Lambda' \Psi^{-1} x_i.$$

Factor Scores

□ Random Factor Scores: Thomson's Method

Consider F to be random, i.e., assume $\begin{pmatrix} F \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I & \Lambda' \\ \Lambda & \Sigma \end{pmatrix}\right)$,
then we have:

$$F | x \sim N(\Lambda' \Sigma^{-1} x, I - \Lambda' \Sigma^{-1} \Lambda).$$

Hence, a natural estimate is the associated mean:

$$\hat{f}_i = \Lambda' \Sigma^{-1} x_i.$$

Notes:

- (1) The Bartlett's method can be viewed as a **weighted least square estimate** by treating X as responses and Λ as input data.
- (2) The Thomson's method is like a **regression approach** by treating F as responses and X as inputs.
- (3) For calculating the factor scores, we use the estimate of $\hat{\Lambda}$, $\hat{\Psi}$, and S .

Summary for FA

- Estimate S , Λ and Ψ .
- Identify the smallest number of common factors (k) that best explain or account for the correlations among the indicators (variables).
- Identify, via factor rotations, the most plausible factor solution that can be meaningfully interpreted.
- Provide interpretation for the common factors, usually using a **cutoff point 0.4** for the factor loadings.

Comparison of FA and PCA

- PCA: find directions of maximum variation in data
FA : attempt to explain associations (correlations) between variables by a set of common factor.
- In PCA, the PCs are a unique set of orthogonal variables.
In FA, factors are not unique, and not necessarily orthogonal.

FA vs PCA

Common properties

- Both methods are mostly used in exploratory data analysis.
- Both methods try to obtain dimension reduction: explain a data set in a smaller number of variables.
- Both methods don't work if the observed variables are almost uncorrelated:
 - ◆ Then PCA returns components that are similar to the original variables.
 - ◆ Then factor analysis has nothing to explain, i.e. ψ_{ii} close to 1 for all i .
- Both methods give similar results if the specific variances are small.
- If specific variances are assumed to be zero in principle factor analysis, then PCA and factor analysis are the same.

FA vs PCA

Differences

- PCA required virtually no assumptions.
Factor analysis assumes that data come from a specific model.
- In PCA emphasis is on transforming observed variables to principle components.
In factor analysis, emphasis is on the transformation from factors to observed variables.
- PCA is not scale invariant.
Factor analysis (with MLE) is scale invariant.
- In PCA, considering $k + 1$ instead of k components does not change the first k components.
In factor analysis, considering $k + 1$ instead of k factors may change the first k factors (when using MLE method).
- Calculation of PCA scores is straightforward.
Calculation of factor scores is more complex.