Lecture 10
Factor Analysis

DSA 8070 Multivariate Analysis
October 18 - October 22, 2021

Whitney Huang
Clemson University

Agenda

1. Background
2. Factor Model Analysis
3. Stock Price Data Example

Overview

Factor Analysis (FA) assumes the covariance structure among a set of variables, $\boldsymbol{X} = (X_1, \cdots, X_p)^T$, can be described via a linear combination of unobservable (latent) variables $\boldsymbol{F} = (F_1, \cdots, F_m)^T$, called factors.

There are three typical objectives of FA:

1. Data reduction: explain explain covariance between p variables using $m < p$ latent factors
2. Data interpretation: find features (i.e., factors) that are important for explaining covariance ⇒ exploratory FA
3. Theory testing: determine if hypothesized factor strucuture fits observed data ⇒ confirmatory FA

FA and PCA

FA and PCA have similar themes, i.e., to explain covariance between variables via linear combinations of other variables

However, there are distinctions between the two approaches:

- FA assumes a statistical model that describes covariation in observed variables via linear combinations of latent variables
- PCA finds uncorrelated linear combinations of observed variables that explain maximal variance

> FA refers to a statistical model, whereas PCA refers to the eigenvalue decomposition of a covariance (or correlation) matrix

Factor Model

Let $\boldsymbol{X} = (X_1, \cdots, X_p)^T$ is a random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The factor model postulates that $\boldsymbol{X}$ can be written as a linear combination of a set of m common factors $F_1, F_2, \cdots, F_m$:

$$
\begin{aligned}
X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\
\vdots \quad & \quad\quad\quad\quad \vdots \quad\quad\quad\quad\quad\quad \vdots \\
X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p
\end{aligned}
$$

where

- $\{\ell_{jk}\}_{p\times m}$ denotes the matrix of factor loadings, that is, ℓ_{jk} is the loading of the j-th variable on the k-th factor
- $(F_1, \cdots, F_m)^T$ denotes the vector of the latent factor scores, that is, F_k is the score on the k-th factor
- $(\varepsilon_1, \cdots, \varepsilon_p)^T$ denotes the vector of latent error terms, ε_j is the j-th specific factor

Factor Model in Matrix Notation

The factor model can be written in a matrix form:

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon},$$

where

- $\boldsymbol{L} = \{\ell_{jk}\}_{p \times m}$ is the factor loading matrix
- $\boldsymbol{F} = (F_1, \cdots, F_m)^T$ is the factor score vector
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_p)^T$ is the (latent) error vector

Unlike in linear model, we do not observe $\boldsymbol{F}$, therefore we need to impose some assumptions to facilitate the model identification

Model Assumptions

First, we assume:

$$\mathbb{E}(\boldsymbol{F}) = \mathbf{0}, \qquad \mathbb{V}\mathrm{ar}(\boldsymbol{F}) = \mathbb{E}(\boldsymbol{F}\boldsymbol{F}^T) = \boldsymbol{I}$$
$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \mathbb{V}\mathrm{ar}(\boldsymbol{\varepsilon}) = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \boldsymbol{\Psi} = \mathrm{diag}(\psi_i), i = 1, \cdots, p$$

Moreover, we assume $\boldsymbol{F}$ and ε are independent, so that $\mathbb{C}\mathrm{ov}(\boldsymbol{F}, \boldsymbol{\varepsilon}) = \mathbf{0}$

- The factors have variance one (i.e., $\mathbb{V}\mathrm{ar}(F_i) = 1$) and uncorrelated with one another
- The error vector are uncorrelated with one another with the specific variance $\mathbb{V}\mathrm{ar}(\varepsilon_i) = \psi_i$
- Under the model assumptions, we have $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}$

Variances and Covariances of Factor Models

Under the factor model, we have

$$\mathbb{Var}(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2 + \psi_i$$

$$\mathbb{Cov}(X_i, X_j) = \ell_{i1}\ell_{j1} + \ell_{i2}\ell_{j2} + \cdots + \ell_{im}\ell_{jm}$$

The portion of the variance that is contributed by the m common factors is the communality:

$$\ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2,$$

and the portion that is not explained by the common factors is called the uniqueness (or the specific variance):

$$\mathbb{Var}(\varepsilon_i) = \psi_i$$

Estimation in Factor Models

In this course we consider two methods to estimate the parameters of a factor model:

- Principal Component Method

$$\text{PCA}: \qquad \boldsymbol{\Sigma} = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T$$

$$\text{Factor Model:} \qquad \boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}$$

  Main idea: Use the first m PCs to form the factor loading matrix, then use the relationship $\boldsymbol{\Psi} = \boldsymbol{\Sigma} - \boldsymbol{L}\boldsymbol{L}^T$ to estimate the specific variances $\hat{\psi}_i = s_i^2 - \sum_{j=1}^m \lambda_j \hat{e}_{ji}^2$

- Maximum Likelihood Estimation: assuming data $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi})$, maximizing the log-likelihood $\ell(\boldsymbol{\mu}, \boldsymbol{L}, \boldsymbol{\Psi}) \propto$ $-\frac{n}{2}\log|\boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^n (\boldsymbol{X}_i - \boldsymbol{\mu})^T (\boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi})^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu})$ to obtain the parameter estimates

Choosing the Number of Common Factors

- The factor model assumes that the $p(p+1)/2$ variances and covariances of X can be reproduced from the $pm+p$ factor loadings and the variances of the p unique factors

- Situations in which m, the number of common factors, is small relative to p is when factor analysis works best. For example, if $p = 12$ and $m = 2$, then the $(12 \times 13)/2 = 78$ elements of $\mathbf{\Sigma}$ can be reproduced from $2 \times 12 + 12 = 36$ parameters in the factor model

- However, if m is too small, the $mp+p$ parameters may not be adequate to describe $\mathbf{\Sigma}$

A Goodness-of-Fit Test for Factor Model

We wish to test whether the factor model appropriately describes the covariances among the p variables

- Specifically, we test

$$H_0 : \mathbf{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \mathbf{\Psi}$$

versus

$$H_1 : \mathbf{\Sigma} \text{ is a positive definite matrix}$$

- Bartlett-Corrected Likelihood Ratio Test Statistic

$$-2\log\Lambda = (n - 1 - (2p + 4m + 5)/6)\log\frac{|\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\mathbf{\Psi}}|}{|\hat{\mathbf{\Sigma}}|}$$

- Reject H_0 at level α if

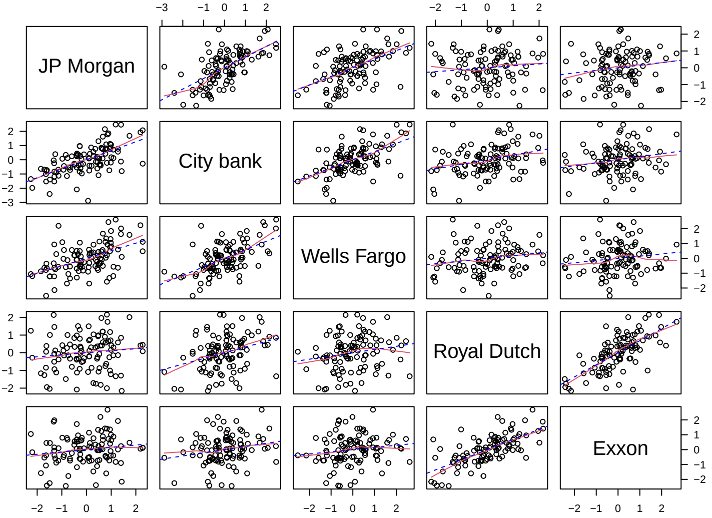
$$-2\log\Lambda > \chi^2_{df=\frac{1}{2}[(p-m)^2-p-m]}$$

Example: Stock Price Data

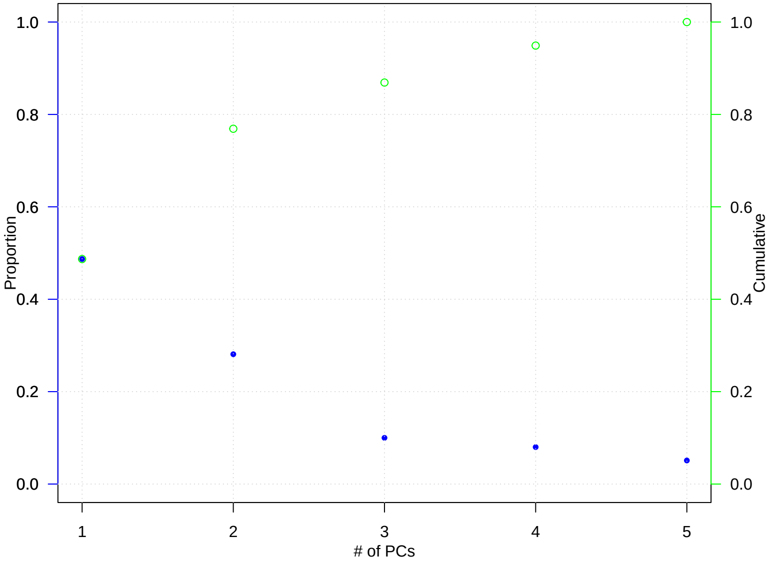
Data are weekly returns in stock prices for 103 consecutive weeks for five companies: JP Morgan, City bank, Wells Fargo, Royal Dutch (Shell), and Exxon

- The first three are banks and the last two are oil companies
- The data are first standardized and the sample correlation matrix is used for the analysis
- We will fit an $m = 2$ factor model

Scatter Plot Matrix of the Standardized Data



Screen Plot



Factor Loadings, Specific Variances, and Residual Matrix

Variable	Loadings 1	Loadings 2	Specific variances
JP Morgan	0.732	0.437	0.273
City bank	0.831	0.280	0.230
Wells Fargo	0.726	0.374	0.333
Royal Dutch	0.605	-0.694	0.153
Exxon	0.563	-0.719	0.166

The residual matrix is $\boldsymbol{\Sigma} - (\tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^T + \tilde{\boldsymbol{\Psi}})$:

$$\begin{bmatrix} 0 & -0.10 & -0.18 & -0.03 & 0.06 \\ & 0 & -0.13 & 0.01 & -0.05 \\ & & 0 & 0.00 & 0.01 \\ & & & 0 & -0.16 \\ & & & & 0 \end{bmatrix}$$

Question: Are these off-diagonal elements small enough?

Maximum Likelihood Estimation

```
> (stock.fac <- factanal(stock, factors = 2,
+ method = "mle", scale = T, center = T))

Call:
factanal(x = stock, factors = 2, method = "mle", scale = T, center = T)

Uniquenesses:
  JP Morgan   City bank Wells Fargo Royal Dutch
      0.417       0.275       0.542       0.005
      Exxon
      0.530

Loadings:
            Factor1 Factor2
JP Morgan   0.763
City bank   0.819   0.232
Wells Fargo 0.668   0.108
Royal Dutch 0.113   0.991
Exxon       0.108   0.677

               Factor1 Factor2
SS loadings      1.725   1.507
Proportion Var   0.345   0.301
Cumulative Var   0.345   0.646

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 1.97 on 1 degree of freedom.
The p-value is 0.16
```

Factor Loading Plot