COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

The Factor Analysis Model: Understanding of Causes

Factor analysis was invented nearly 100 years ago by psychologist Charles Spearman, who hypothesized that the enormous variety of tests of mental ability measures of mathematical skill, vocabulary, other verbal skills, artistic skills, logical reasoning ability, etc. could all be explained by one underlying "factor" of general intelligence that he called *g*. He hypothesized that if *g* could be measured and you could select a subpopulation of people with the same score on *g*, in that subpopulation you would find no correlations among any tests of mental ability. In other words, he hypothesized that *g* was the only factor common to all those measures.

Factor Analysis

- Factor analysis has a tremendous appeal for the behavior and social sciences.
- In these areas, it is natural to regard multivariate observations on human processes and behavior as manifestations of underlying unobservable “traits”.
- Factor analysis provides a way of explaining the observed variability in behavior in terms of these traits.

The Factor Analysis Model

- Data:

$\boldsymbol{X} = (x_1, \ldots, x_p)$ is an *observable* random vector which has a p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

- Factors:

$F_1, F_2, \ldots, F_m$ are *unobservable/latent* random variables called *the common factors*

- Errors:

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are random variables called *errors* or *specific* factors.

- Loadings:

Suppose that there exist constants λ_{ij} (*the loadings*) such that:

$$x_1 = \mu_1 + \lambda_{11}F_1 + \lambda_{12}F_2 + \ldots + \lambda_{1m}F_m + \varepsilon_1$$
$$x_2 = \mu_2 + \lambda_{21}F_1 + \lambda_{22}F_2 + \ldots + \lambda_{2m}F_m + \varepsilon_2$$
$$\ldots$$
$$x_p = \mu_p + \lambda_{p1}F_1 + \lambda_{p2}F_2 + \ldots + \lambda_{pm}F_m + \varepsilon_p$$

Factor Analysis Model in Matrix Notation

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{\varepsilon}$$

where

$\boldsymbol{X}$ is $p \times 1$, $\boldsymbol{L}$ is $p \times m$, $\boldsymbol{F}$ is $m \times 1$, and $\boldsymbol{\varepsilon}$ is $p \times 1$

Assume: $\text{cov}(\boldsymbol{F}) = \boldsymbol{I}_{m \times m}$, and $\text{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$,

where

$$\Psi = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

Note:

$$\boldsymbol{\Sigma} = \text{cov}(\boldsymbol{X}) = \boldsymbol{LL}' + \boldsymbol{\Psi}$$

Hence

$$\sigma_{ii} = \text{Var}(X_i) = \sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i$$

and

$$\sigma_{ik} = \text{cov}(X_i, X_k) = \sum_{j=1}^{m} \lambda_{ij}\lambda_{kj}$$

$h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$ is called the i^{th} *communality*

i.e. the component of variance of x_i that is due to the common factors $F_1, F_2, \ldots, F_m$

ψ_i is called the *specific* variance

i.e. the component of variance of x_i that is **specific** only to that variable

(sometimes $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ are called the **specific factors**)

Factor Rotation

If $\widehat{\boldsymbol{L}}$ is the $p \times m$ matrix of estimated factor loadings, then let

$$\widehat{\boldsymbol{L}}^* = \widehat{\boldsymbol{L}}\boldsymbol{T}$$

where $\boldsymbol{T}$ is any orthogonal matrix (that is, $\boldsymbol{T'T} = \boldsymbol{TT'} = \boldsymbol{I}$).

Then $\widehat{\boldsymbol{L}}^*$ is a $p \times m$ matrix of *rotated* loadings. Moreover,

$$\text{cov}(\boldsymbol{X}) = \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}' + \widehat{\boldsymbol{\Psi}} = \widehat{\boldsymbol{L}}\boldsymbol{T}\boldsymbol{T}'\widehat{\boldsymbol{L}}' + \widehat{\boldsymbol{\Psi}} = \widehat{\boldsymbol{L}}^*\widehat{\boldsymbol{L}}^{*\prime} + \widehat{\boldsymbol{\Psi}}$$

That is, the estimated covariance matrix remains unchanged! Thus, from a mathematical point of view it is the same whether $\widehat{\boldsymbol{L}}$ or $\widehat{\boldsymbol{L}}^*$ is used, because the factor model is overparametrized and has many solutions.

Notes:

- Since the original loadings may not be easily interpretable, it is a common practice to rotate them until a “simpler structure” is achieved.
- Ideally, we want to see a pattern of loadings such that each variable loads heavily on a single factor and has small loadings on the remaining factors.
- It is not always possible to obtain such simple structure ☹
- There are graphical and analytical methods for choosing the optimal rotation.

Example 1: *Olympic decathlon Scores*

Data was collected for $n = 280$ starts from 1960 to 2004 for the ten decathlon events (*100-m run, Long Jump, Shot Put, High Jump, 400-m run, 110-m hurdles, Discus, Pole Vault, Javelin, 1500-m run*).

Correlated Factors

We now change the assumption that the factors are *uncorrelated* with the following:

$$\text{cov}(\boldsymbol{F}) = \boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1m} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{m1} & \phi_{m2} & \cdots & \phi_{mm} \end{bmatrix}$$

Then the covariance of the data becomes

$$\boldsymbol{\Sigma} = \text{cov}(\boldsymbol{X}) = \boldsymbol{L\Phi L'} + \boldsymbol{\Psi}$$

Structural Equation models:

Path Diagrams

The above equation is just one way to specify the model we want to fit to the data. Alternatively, we may use *path diagrams* to specify the model graphically.

In a path diagram, observed variables are denoted with rectangles and latent (unobservable) variables (factors) are denoted with ovals.

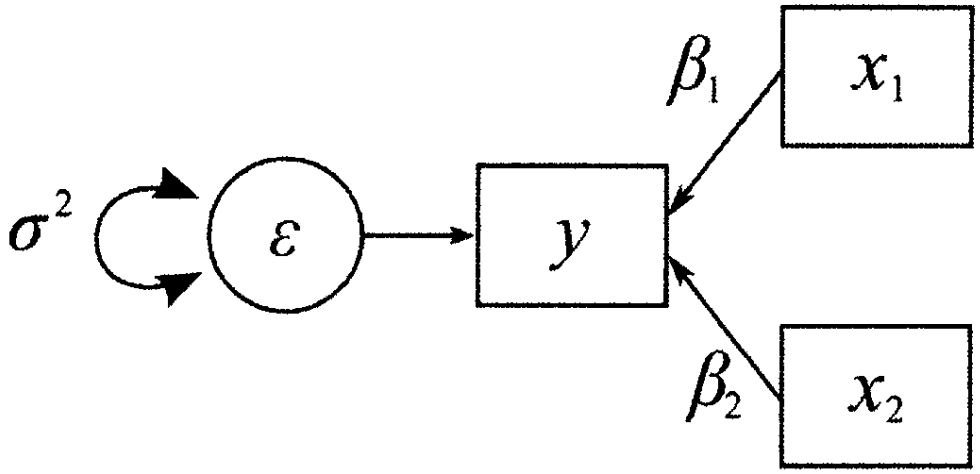
The ovals and rectangles are connected with arrows. Note that a one-headed arrow denotes a structural or unidirectional relationship, while a two-headed arrow denotes a covariance among variables. A two-headed arrow pointing from one variable and back to itself is used to specify the variance of the variable, which is usually an error or "disturbance" variable.

Example:

Consider the two-predictor regression model

$$y = \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \boldsymbol{\varepsilon}, \text{ Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma}^2$$

It can also be represented with a path diagram as follows:



Example 2: 14.2.2 from Rencher

Table 14.1, which was obtained from a university business statistics class. As shown in Table 14.1, each of the 94 students in the class received scores for laboratory assignments (Lab), homework assignments (HW), pop quizzes (PopQuiz), midterm exam #1 (Exam1), midterm exam #2 (Exam2), and the final exam (FinalExam).

The instructor might hypothesize that the 6-dimensional measure of performance in the class is being driven by an underlying 2-dimensional factor process, with the first factor associated with daily effort and the second factor associated with knowledge mastery.

A possible model is:

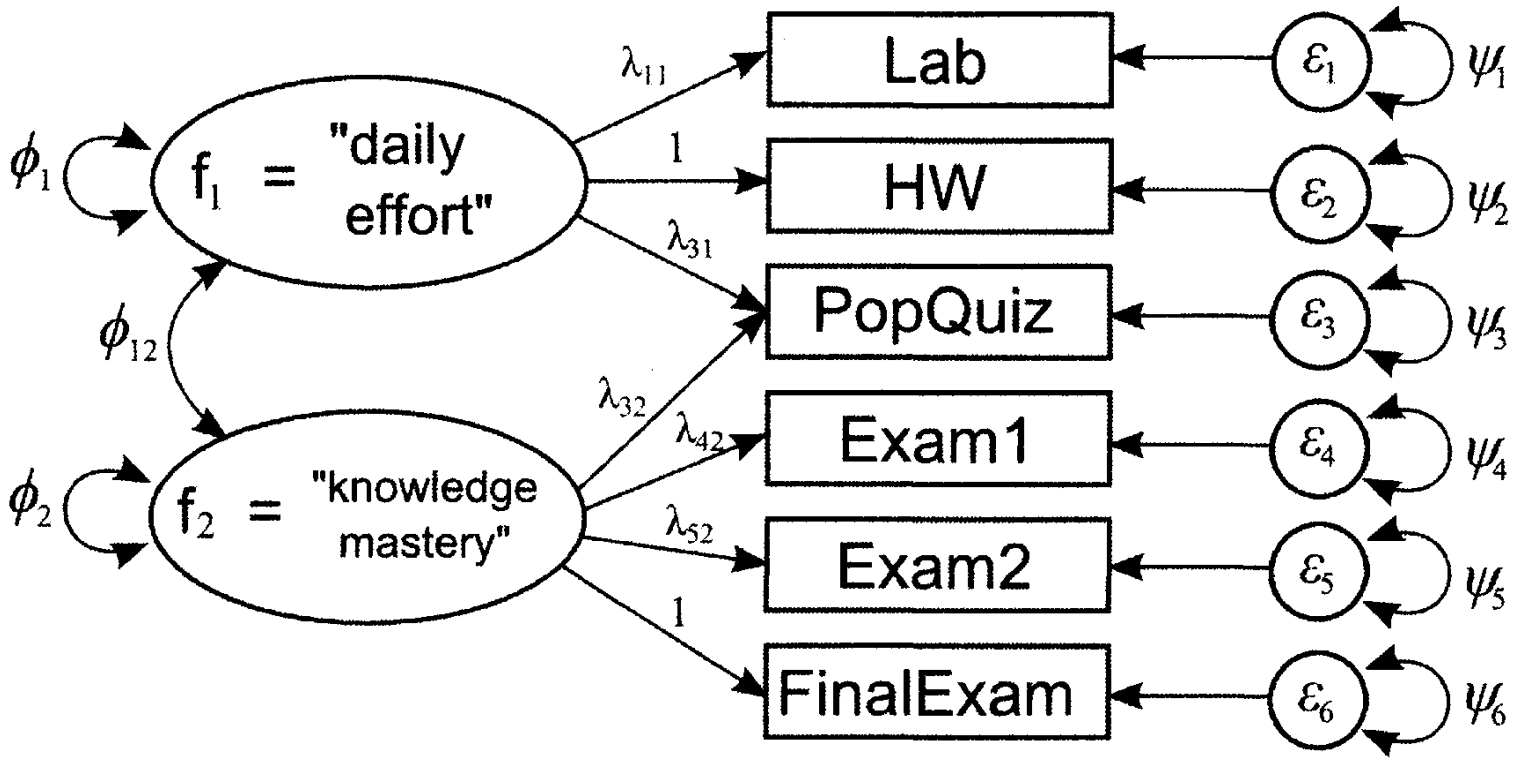
$$
\begin{aligned}
\text{Lab} &= \boldsymbol{\mu}_1 + \boldsymbol{\lambda}_{11} f_1 + \boldsymbol{\lambda}_{12} f_2 + \boldsymbol{\varepsilon}_1 \\
\text{HW} &= \boldsymbol{\mu}_2 + f_1 + \boldsymbol{\varepsilon}_2 \\
\text{PopQuiz} &= \boldsymbol{\mu}_3 + \boldsymbol{\lambda}_{31} f_1 + \boldsymbol{\lambda}_{32} f_2 + \boldsymbol{\varepsilon}_3 \\
\text{Exam1} &= \boldsymbol{\mu}_4 + \boldsymbol{\lambda}_{41} f_1 + \boldsymbol{\lambda}_{42} f_2 + \boldsymbol{\varepsilon}_4 \\
\text{Exam2} &= \boldsymbol{\mu}_5 + \boldsymbol{\lambda}_{51} f_1 + \boldsymbol{\lambda}_{52} f_2 + \boldsymbol{\varepsilon}_5 \\
\text{Final Exam} &= \boldsymbol{\mu}_6 + f_2 + \boldsymbol{\varepsilon}_6
\end{aligned}
$$

Note: Some of the loadings are assumed to be 1 because of model identifiability issues.

Example (continued):

We will further simplify the model to test the theory that variables mostly load on one of the factors:

$$
\begin{aligned}
\text{Lab} &= \boldsymbol{\mu}_1 + \boldsymbol{\lambda}_{11} f_1 + \boldsymbol{\varepsilon}_1 \\
\text{HW} &= \boldsymbol{\mu}_2 + f_1 + \boldsymbol{\varepsilon}_2 \\
\text{PopQuiz} &= \boldsymbol{\mu}_3 + \boldsymbol{\lambda}_{31} f_1 + \boldsymbol{\lambda}_{32} f_2 + \boldsymbol{\varepsilon}_3 \\
\text{Exam1} &= \boldsymbol{\mu}_4 + \boldsymbol{\lambda}_{42} f_2 + \boldsymbol{\varepsilon}_4 \\
\text{Exam2} &= \boldsymbol{\mu}_5 + \boldsymbol{\lambda}_{52} f_2 + \boldsymbol{\varepsilon}_5 \\
\text{Final Exam} &= \boldsymbol{\mu}_6 + f_2 + \boldsymbol{\varepsilon}_6
\end{aligned}
$$


ϕ_1
f_1 = "daily effort"
ϕ_{12}
ϕ_2
f_2 = "knowledge mastery"
λ_{11}
1
λ_{31}
λ_{32}
λ_{42}
λ_{52}
1
Lab
HW
PopQuiz
Exam1
Exam2
FinalExam
ε_1
ε_2
ε_3
ε_4
ε_5
ε_6
ψ_1
ψ_2
ψ_3
ψ_4
ψ_5
ψ_6

Introduction to multivariate latent

- The regression model from last time is somewhat limiting because it only describes the conditional distribution of one variable given the others.
- In general, we may be interested in the relationships among all the variables in a dataset.
- If the variables were approximately jointly normally distributed, then we could describe the relationships among the variables with the sample covariance matrix or a multivariate normal model.
- However, such a model is inappropriate for nonnumeric ordinal variables.
- To accommodate ordinal variables, we can extend the ordered probit model above to a latent, multivariate normal model

The Gaussian copula model

Let the data be p-dimensional vectors satisfying:

$$Y_{ij} = g_j(Z_{ij}), i = 1, \ldots, n$$
$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \sim N_p(0, \boldsymbol{\Psi})$$

We assume $g_1, \ldots, g_p$ are non-decreasing functions and $\boldsymbol{\Psi}$ is a correlation matrix with diagonal elements equal to 1. Under this model the cdf of j^{th} component of $\boldsymbol{Y}$ is $F_j(y) = \Phi\left(g_j^{-1}(y)\right)$. This is known as the *multivariate normal copula model.* The term "copula" refers to the method of "coupling" a model for multivariate dependence to a model for the marginal distributions of the data.

Copula estimation

The unknown parameters in the copula model are the matrix Ψ and the non- decreasing functions $g_1, \ldots, g_p$. Bayesian inference for all of these parameters would require that we specify a prior for Ψ as well as p prior distributions over the complicated space of arbitrary non-decreasing functions. There is a rank likelihood method which avoids this. Since each g_j is non-decreasing, observing the $n \times p$ data matrix $\boldsymbol{Y}$ tells us that the matrix of latent variables $\boldsymbol{Z}$ must lie in the set

$$R(Y) = \left\{Z : z_{i_1 j} < z_{i_2 j} \; if \; y_{i_1 j} < y_{i_2 j}\right\}$$

As a function of Ψ, $P(Z \in R(Y)|\Psi)$ is called the rank likelihood for the multivariate normal copula model. Computing the likelihood for a given value of Ψ is very difficult, but we can make an MCMC approximation to $f(\Psi, Z|Z \in R(Y))$ using Gibbs sampling, provided we use a prior for Ψ based on the inverse-Wishart distribution.

A parameter-expanded prior distribution for Ψ

Unfortunately, there is no simple conjugate class of prior distributions for our correlation matrix Ψ. Let's consider an alternative model:

$$Y_{ij} = g_j(Z_{ij}), i = 1, \ldots, n$$
$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \sim N_p(0, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is a covariance matrix. In this case a natural prior distribution for Σ would be an inverse-Wishart distribution, which would give an inverse-Wishart full conditional distribution and thus make posterior inference available via Gibbs sampling. Careful inspection of the rank likelihood indicates that it does not provide us with a complete estimate of Σ. For this reason, we say that the diagonal entries of Σ are non-identifiable in this model. The value of Ψ is identifiable from the rank likelihood, and so one estimation approach for the Gaussian copula model is to reparametrize the model in terms of a non-identifiable covariance matrix Σ, but focus our posterior inference on the identifiable correlation matrix Ψ = h(Σ).