Factor Analysis

An example

The mathematical ability of the graduate students in mathematics program are measured, based one the testscores in algebra, combinatorics, graph theory, real analysis, measure theory, probability, differential equations and data structure. It is observed that the testsores from algebra, combinatorics, graph theory and data structure are highly correlated; the testscores from real analysis, measure theory, probability and differential equations are also highly correlated. So, we believe there are two types of mathematical ability. One, which can be called algebraic ability, determines a student's performance in the first group of branches; the other one, which can be called analytic ability, determines a student's performance in the second group of mathematical branches. So, there are two factors, algebraic ability and analytic ability, underlying the testscores.

The essential purpose of factor analysis

describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors.

Orthogonal factor model

$$\begin{aligned}
X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \epsilon_1 \\
X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \epsilon_2 \\
&\vdots \\
X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \epsilon_p
\end{aligned}$$

or

$$X - \mu = LF + \epsilon$$

where

$$L = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{pmatrix}, \qquad F = \begin{pmatrix} F_1 \\ \vdots \\ F_m \end{pmatrix}, \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix}$$

$L = (l_{ij})$ is called the loading matrix (l_{ij} called loadings). $F_1, \ldots, F_m$ are called the common factors, $\epsilon_1, \ldots, \epsilon_m$ are called the specific factors.

Assumptions

$E(F) = 0, cov(F) = I($ the $m \times m$ identity matrix$), E(\epsilon) = 0$, and

$$cov(\epsilon) = \Psi = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}$$

and ϵ and F are independent, $cov(F, \epsilon) = 0$

Relation between Σ and L, Ψ

$$\Sigma_{p\times p} = E(X-\mu)(X-\mu)' = LL' + \Psi$$

Relation between X and F

$$cov(X, F) = E(X-\mu)F = L$$

So, $Var(X_i) = l_{i1}^2 + \cdots + l_{im}^2 + \psi_i$, where $i1^2 + \cdots + l_{im}^2$ is called the ith communality, and ψ_i is called the uniqueness, or specific variance. $cov(X_i, F_j) = l_{ij}$ is called the loading of the ith variable on the jth factor.

Some preliminary issues

1. Nonexistence of a proper solution

For example, suppose $p = 3$, $m = 1$ and $\Sigma = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix}$. And the model is

$$X_1 - \mu_1 = l_{11}F_1 + \epsilon_1 X_2 - \mu_2 = l_{21}F_1 + \epsilon_2 X_3 - \mu_3 = l_{31}F_1 + \epsilon_3$$

So,

$$\Sigma = LL' + \Psi = \begin{pmatrix} l_{11} \\ l_{21} \\ l_{31} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix} = \begin{pmatrix} l_{11}l_{11} & l_{11}l_{12} & l_{11}l_{31} \\ & l_{21}^2 & l_{21}l_{31} \\ & & l_{31}^2 \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{pmatrix}$$

So, we have,

$$0.7 = l_{11}l_{31}, \quad 0.4 = l_{21}l_{31} \quad 0.9 = l_{11}l_{21}$$

which implies $l_{21} = \frac{0.4}{0.7} l_{11}$ and $0.9 = l_{11} l_{21}$. Hence $l_{11}^2 = 1.575$, $l_{11} = 1.255$. Since $1 = l_{11}^2 + \psi_1$, ψ_1 is equal to -0.575, which is a contradiction because $\psi_1 = var(\epsilon_1) \geq 0$.

2. Inherent ambiguity (The solution is not unique).

Suppose $X - \mu = LF + \epsilon$ and $\Sigma = LL' + \Psi$. Let T be any $m \times m$ orthogonal matrix such that $TT' = T'T = I$. Let

$$F^* = TF, \qquad L^* = LT'$$

We have

$$X - \mu = LF + \epsilon = L^* F^* + \epsilon$$

and

$$\Sigma = LL' + \Psi = L^* L^{*\prime} + \Psi$$

. Hence, both (L, F, ϵ) and (L^*, F^*, ϵ) are the solutions to the orthogonal factor models.

Estimation method

Principal component method

Suppose $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$ are the eigenvalue-eigenvector pairs of Σ. According to the spectral decomposition theorem,

$$\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p' = (\sqrt{\lambda_1} e_1)(\sqrt{\lambda_1} e_1)' + (\sqrt{\lambda_2} e_2)(\sqrt{\lambda_2} e_2)' + \cdots + (\sqrt{\lambda_p} e_p)(\sqrt{\lambda_p} e_p)'$$

$$= (\sqrt{\lambda_1} e_1 \sqrt{\lambda_2} e_2 \cdots \sqrt{\lambda_p} e_p) \begin{pmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \vdots \\ \sqrt{\lambda_p} e_p' \end{pmatrix}$$

Let $L = (\sqrt{\lambda_1} e_1 \sqrt{\lambda_2} e_2 \cdots \sqrt{\lambda_p} e_p)$, then

$$\Sigma_{p \times p} = L_{p \times p} L'_{p \times p} + 0_{p \times p}.$$

But it is not a interesting solution (why?). A more interesting is

$$\tilde{L} = ((\sqrt{\lambda_1} e_1 \sqrt{\lambda_2} e_2 \cdots \sqrt{\lambda_m} e_m)$$

and

$$\tilde{\Psi} = diag(\tilde{\psi}_1, \cdots, \tilde{\psi}_p)$$

where $\tilde{\psi}_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2$. Hence

$$\Sigma \approx \tilde{L}\tilde{L}' + \tilde{\Psi}$$

The Choice of m

Suppose S is the sample covariance matrix. The residual matrix from the principal component solution is

$$S - (\tilde{L}\tilde{L}' + \tilde{\Psi})$$

The norm of a matrix A, denoted by $\|A\|$, is defined to be the sum of squared entries of A. It can be shown that

$$\|S - (\tilde{L}\tilde{L}' + \tilde{\Psi})\| \leq \lambda_{m+1}^2 + \cdots + \lambda_p^2$$

Contribution to s_{ii} from F_1 is l_{i1}^2.

Contribution to $tr(S) = s_{11} + \cdots + s_{pp}$ from F_1 is $l_{11}^2 + l_{21}^2 + \cdots + l_{p1}^2 = \lambda_1$.

The same can be stated for the jth factor F_j. Hence the proportion of the total sample variance due to the jth factor is

$$\frac{\lambda_j}{s_{11} + s_{22} + \cdots + s_{pp}}$$

Principal factor method

It is an iterative procedure to approximate S by $LL' + \Psi$.

$$\min_{L,\Psi} \|S - (LL' + \Psi)\|^2$$

1. Initialize Ψ as Ψ_1.
2. Decompose $S - \Psi_1$, and select the largest m eigenvectors to form L_1.
3. Set $\Psi_2 = diag(S - L_1L_1')$.
4. Iterate step 2 and step 3 until convergence.

(Question: when we expect a solution such that $S = LL'$ and $\Psi = 0$?)

Maximum likelihood method

Assume that F and ϵ are normally distributed. Let $L(\mu, \Sigma)$ be the likelihood function dependent on μ, L and Ψ. Then,

$$\max_{L'\Psi^{-1}L=\Delta} L(\mu, L, \Psi) \Rightarrow \hat{L}, \hat{\Psi}(\text{mles for } L \text{ and } \Psi)$$

MLE for the ith communality: $\hat{h}_i^2 = \hat{l}_{i1}^2 + \cdots + \hat{l}_{im}^2$.

Proportion of the total sample variance due to the jth factor is

$$\frac{\hat{l}_{1j}^2 + \cdots + \hat{l}_{pj}^2}{s_{11} + \cdots + s_{pp}}$$

Comparison between principal component method and maximum likelihood method

Suppose $x_1, x_2, \ldots, x_5$ denote the observed weekly rates of return Applied Chemical, Du Pont, Union Carbide, Exxon, and Texaco, respectively. The sample correlation matrix is as follows.

$$R = \begin{pmatrix} 1.000 & .577 & \cdots & 0.462 \\ .577 & 1.000 & \cdots & 0.322 \\ \vdots & \vdots & & \vdots \\ 0.462 & 0.322 & \cdots & 1.000 \end{pmatrix}$$

	Principal component			Maximum likelihood		
variable	F_1	F_2	$\tilde{\psi}_i$	F_1	F_2	$\hat{\psi}_i$
1.Applied Ch	.783	$-.217$	.34	.684	.189	.50
2.Du Pont	.772	$-.458$	.19	.694	.517	.25
3.Union Car	.794	$-.234$	.31	.681	.248	.47
4.Exxon	.713	.412	.27	.621	$-.073$	.61
5.Texaco	.712	.524	.22	.792	$-.442$	.18
Cum.Prop	.571	.733		.485	0.598	

The residual matrix from the principal component method

$$R - \tilde{L}\tilde{L}' - \tilde{\Psi} = \begin{pmatrix} 0 & -.127 & -.164 & -.069 & .017 \\ & 0 & -.122 & .055 & .012 \\ & & 0 & -.019 & -.017 \\ & & & 0 & -.232 \\ & & & & 0 \end{pmatrix}$$

The residual matrix from the maximum likelihood method:

$$R - \hat{L}\hat{L}' - \hat{\Psi} = \begin{pmatrix} 0 & .005 & -.004 & -.024 & -.004 \\ & 0 & -.003 & -.004 & .000 \\ & & 0 & .031 & -.004 \\ & & & 0 & -.000 \\ & & & & 0 \end{pmatrix}$$

Which method is better? and what is your conclusion

Test for the number of common factors (m)

$H_0 : \Sigma = L_{p\times m}L'_{p\times m} + \Psi$ $\qquad$ $H_1 : \Sigma$any other positive definite matrix

Likelihood ratio statistic:

$$-2ln\Gamma = nln(\frac{\mid \hat{\Sigma} \mid}{\mid S_n \mid}) \sim \chi^2_{df}$$

where

$$df = \frac{1}{2}p(p+1) - [p(m+1) - \frac{1}{2}m(m-1)] = \frac{1}{2}[(p-m)^2 - (p+m)]$$

(For any given p, $m < \frac{1}{2}(2p+1-\sqrt{8p+1})$ to guarantee that df is positive)

Bartlett correction:

we reject H_0 at the α level of significance if

$$(n-1-(2p+4m+5)/6)ln(\frac{\mid \hat{\Sigma} \mid}{\mid S_n \mid}) > \chi^2_{df}(\alpha)$$

Factor rotation

Let $\hat{L}^* = \hat{L}T$, where $TT' = T'T = I$, we have

$$\hat{L}\hat{L}' + \hat{\Psi} = \hat{L}^*\hat{L}^{*\prime} + \hat{\Psi}$$

Idea: Find T to give a simpler and more interpretable solution

1. Graphical method:

$$\hat{L}^*_{p\times 2} = \hat{L}_{p\times 2}T_{2\times 2}$$

where $T = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}$ (clockwise rotation), or $T = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$ (counterclockwise rotation).

For example,

	MLE		rotated		
variable	F_1	F_2	F_1^*	F_2^*	$\hat{\psi}$
Applied Ch	.684	.189	**.601**	.377	.50
Du Pont	.694	.517	**.850**	.164	.25
Union Car	.681	.248	**.643**	.335	.47
Exxon	.621	−.073	.365	**.507**	.61
Texaco	.792	−.442	.208	**.883**	.18
Cum.Prop.	.485	.598	**.335**	.598	

Questions: 1. how do you determine the rotation angles? 2. Do the communalities change?

3. Does the proportion of the total variance due to each factor change after rotation?

2. Analytic method (varimax criterion)

$$L = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{pmatrix}$$

1. Define $\tilde{l}_{ij}^* = \hat{l}_{ij}^*/\hat{h}_i$

2.

$$\max V = \max \frac{1}{p} \sum_{j=1}^{m} [\sum_{i=1}^{m} \tilde{l}_{ij}^{*4} - (\sum_{i=1}^{p} \tilde{l}_{ij}^{*2})^2/p]$$

3. Scale back the solution from step 2, $\tilde{l}_{ij}^* \hat{h}_i$

Oblique rotation

Orthogonal rotation sometime still does not give an easy interpretation. No-orthogonal rotation will be used. This allows for possible simplicity at the expense of losing the independence of the factors.

Factor scores

$\hat{f}_j$ = the estimates of the values f_j attained by F_j (the jth factors.

Orthogonal factor model:

$$X - \mu = LF + \epsilon$$

Weighted least squares method:

$$\min\left(\sum_{i=1}^{p}\frac{\epsilon_i^2}{\psi_i}\right) = \epsilon'\Psi^{-1}\epsilon = (x-\mu-lf)'\Psi^{-1}(x-\mu-lf)$$

And

$$\hat{f}_j = (\hat{L}'\hat{\Psi}^{-1}\hat{L})^{-1}\hat{L}'\hat{\Psi}^{-1}(\vec{x}_j - \bar{x}).$$

Regression method:

Assume that F and ϵ are normally distributed. The joint distribution of $x-\mu$ and F is $N_{m+p}(0, \Sigma^*)$, and

$$\Sigma^* = \begin{pmatrix} \Sigma = LL' + \Psi & L \\ L' & I \end{pmatrix}$$

$$\text{mean} = E(F \mid x) = L'\Sigma^{-1}(x-\mu) = L'(LL'+\Psi)^{-1}(x-\mu)$$

$$\text{covariance} = Cov(F \mid x) = I - L'\Sigma^{-1}L = I - L'(LL'+\Psi)^{-1}L$$

The factor scores are

$$\hat{f}_j = \hat{L}'\hat{\Sigma}^{-1}(\vec{x}_j - \bar{x}) = \hat{L}'(\hat{L}\hat{L}' + \hat{\Psi})^{-1}(\vec{x}_j - \bar{x})$$

Miscellaneous issues in Factor Analysis

1. m, the number of common factors

(1) The proportion of the total variance explained

(2) Small residual matrix

(3) Likelihood ratio test under normal assumptions

(4) Subject-matter knowledge

(5) reasonableness of the results

2. Factor scores

Factor scores are used for diagnostic purposes, as well as subsequent analysis.

(1) outliers detection: plot the scores of F_i against those of F_j

(2) Compare results from different methods, identify insignificant factors

3. S or R Let

$$V = \begin{pmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_{pp} \end{pmatrix}$$

Then $R = V^{-1/2} S V^{-1/2}$.

(1) Principal component method

Suppose

$$S \approx L_s L_s' + \Psi_s, \qquad R \approx L_r L_r' + \Psi_r$$

$(V^{-1/2} L_s,\ V^{-1/2} \Psi_s V^{-1/2})$ and $(L_r,\ \Psi_r)$ are different, but usually they are close to each other.

(2)Principal factor method

Suppose $(\tilde{L}_s, \tilde{\Psi})$ is the solution to $\min \|S - LL' - \Psi\|$, and $(\tilde{L}_r, \tilde{\Psi}_r)$ is the solution to $\min \|R - LL' - \Psi\|$. Since for any given data, there usually exits a constant C such that

$$\|V^{-1/2} S V^{-1/2} - V^{-1/2} L (V^{-1/2} L)' - V^{-1/2} \Psi V^{-1/2}\| < C \|S - LL' - \Psi\|$$

i.e.,

$$\|R - V^{-1/2} L (V^{-1/2} L)' - V^{-1/2} \Psi V^{-1/2}\| < C \|S - LL' - \Psi\|$$

Hence, $(V^{-1/2} \tilde{L}_s, V^{-1/2} \tilde{\Psi}_s V^{-1/2}$ would not be very different from $(\tilde{L}_r, \tilde{\Psi}_r)$.

(3)Maximum likelihood method.

let $s_{ii} = \sum (x_{ki} - \bar{x}_i)/n$ for $i = 1, 2, \ldots, p$. Suppose $(\hat{L}_s, \hat{\Psi}_s)$ is the solution based on S; $(\hat{L}_r, \hat{\Psi}_r)$ is the solution based on R. They are equivalent under the transformation involving V.

In general, we don't expect significant different between the solution directly derived from R and the transformed solution from S, vise versa.

4. Relationship of FA to PCA

(1)PCA $\Leftarrow$ Original variables $\Leftarrow$ common factors

(2)

PCA: 1. eliminate correlation via linear transformation. 2. focus on explain the (sample) total variance $\sum s_{ii}$.

FA: model the covariance structure via a small number of factors

(3)

PCA: no assumptions

FA: make many assumptions. validations are difficult

(4)

PCA: interpretability is limited.

FA: provide flexibility in interpretation

(5)

PCA: results can be used directly for subsequent analysis

FA: need to be cautious when used for subsequent analysis

5. Factor analysis in practice.

(1) Try all possible methods and compare the results.

(2) For large datasets, split them in half and perform FA on each part, and compare results

(3) WOW criterion

Some concerns:

(1)unverifiable assumptions.

(2)existence of unobservable variables (latent variables)

(3)the number of factors is subjective

(4)solution is not unique