Factor Analysis

Further Statistical Methods, Lecture 8
Hilary Term 2004

Steffen Lauritzen, University of Oxford; February 15, 2005

The linear normal factor model

The p *manifest* variables $X^\top = (X_1, \ldots, X_p)$ are linearly related to the q *latent* variables $Y^\top = (Y_1, \ldots, Y_q)$ as

$$X = \mu + \Lambda Y + U, \tag{1}$$

where Y and U are independent and follow multivariate normal distributions

$$Y \sim \mathcal{N}_q(0, I), \quad U \sim \mathcal{N}_p(0, \Psi),$$

where Ψ is a *diagonal* matrix, i.e. the indidividual error terms U_i are assumed independent.

The latent variables Y_j are the *factors* and Λ the matrix of *factor loadings*.

The *idea* is to describe the variation in X by variation in a latent Y plus noise, where the number of factors q is considerably smaller than p.

The *problem* is now to determine the smallest q for which the model is adequate, estimate the factor loadings and the error variances.

The marginal distribution of the observed X is

$$X \sim \mathcal{N}_p(\mu, \Sigma), \quad \Sigma = \Lambda\Lambda^\top + \Psi.$$

The factor loadings Λ cannot be determined uniquely. For example, if O is an orthogonal $q \times q$-matrix and we let $\tilde{\Lambda} = \Lambda O$ we have

$$\tilde{\Lambda}\tilde{\Lambda}^\top = \Lambda O O^\top \Lambda^\top = \Lambda\Lambda^\top$$

so Λ and $\tilde{\Lambda}$ specify same distribution of the observable X.

Maximum likelihood estimation

Let

$$S = \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X})(X_n - \bar{X})^\top$$

be the empirical covariance matrix. The likelihood function after maximizing in μ to obtain $\hat{\mu} = \bar{X}$ is

$$\log L(\Sigma) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det(\Sigma) - \frac{n}{2}\operatorname{tr}(\Sigma^{-1}S).$$

Maximizing this under the constraint $\Sigma = \Lambda\Lambda^\top + \Psi$ can be quite tricky.

After some (complex) manipulation, the likelihood equations can be collected in two separate equations. One is the obvious equation

$$\Psi = \mathrm{diag}(S - \Lambda\Lambda^\top) \qquad (2)$$

which gives Ψ in terms of S and Λ.

To express Λ in terms of S and ψ is more complex. Introduce

$$S^* = \Psi^{-1/2} S \Psi^{-1/2}, \quad \Lambda^* = \Psi^{-1/2}\Lambda.$$

Then the MLE of Λ^* can be determined by the following two criteria:

1. The columns of $\Lambda^* = (\lambda_1^* : \cdots : \lambda_q^*)$ are eigenvectors of the q largest eigenvalues of S^*.

2. If Γ is a diagonal matrix with Γ_{ii} being the eigenvalue associated with λ_i^*, then

$$\Gamma_{ii} > 1, \quad S^* \Lambda^* = \Lambda^* \Gamma. \tag{3}$$

A classic algorithm begins with an initial value of Ψ, finds the eigenvectors e_i^* corresponding to the q largest eigenvalues of S^*, lets $\lambda_i^* = \theta_i e_i^*$ and solves for θ_i in (3). When Λ^* and thereby Λ has been determined in this way, a new value for Ψ is calculated using (2).

The algorithm can get severe problems if at some point the constraints $\psi_{ii} > 0$ and $\Gamma_{ii} > 1$ are violated.

The EM algorithm

This is straight-forward. Initialize with Λ and Ψ and $\mu = \bar{X}$. The E-step imputes the latent variables Y as $\hat{Y}_n$ by exploiting

$$\hat{Y}_n = \mathbf{E}(Y \mid X_n) = \Lambda^\top \Sigma^{-1}(X_n - \mu).$$

The M-step estimates μ, Λ, Ψ by standard linear least squares in the model

$$X_n = \mu + \Lambda \hat{Y}_n + U_n.$$

The algorithm is claimed to be slow, but it is conceptually simpler and each step is straight-forward so demands very little computation.

Choice of the number of factors

Under regularity conditions, the *deviance*

$$\begin{aligned} D &= -2\{\log L(H_0) - \log L(H_1)\} \\ &= n\{\mathrm{tr}(\hat{\Sigma}^{-1}S) - \log\det(\hat{\Sigma}^{-1}S) - p\} \end{aligned}$$

has an approximate χ^2-distribution with ν degrees of freedom where

$$\nu = \frac{1}{2}\{(p-q)^2 - (p+q)\}.$$

One can now either choose q as small as possible with the deviance being non-significant, or one can minimze AIC or BIC where

$$AIC = D + 2\nu, \quad BIC = D + \nu \log N.$$

Interpretation

To interpret the results of a factor analysis, it is customary to look at the *communality* c_i of the manifest variable X_i

$$c_i = \frac{\mathbf{V}(X_i) - \mathbf{V}(U_i)}{\mathbf{V}(X_i)} = 1 - \frac{\psi_{ii}}{\psi_{ii} + \sum_{j=1}^{q} \lambda_{ij}^2}$$

which is the proportion of the variation in X_i explained by the latent factors. Each factor Y_j contributes

$$\frac{\lambda_{ij}}{\psi_{ii} + \sum_{j=1}^{q} \lambda_{ij}^2}$$

to this explanation.

Typically the variables X are standardized so that they add to 1 and have unit variance, corresponding to considering just the empirical correlation matrix C instead of S.

Then

$$\psi_{ii} + \sum_{j=1}^{q} \lambda_{ij}^2 = 1$$

so that $c_i = 1 - \psi_{ii}$ and λ_{ij}^2 is the proportion of $\mathbf{V}(X_i)$ explained by Y_j.

Orthogonal rotation

Since Y is only defined up to an orthogonal rotation, we can choose a rotation ourselves which seems more readily interpretable, for example one that 'partitions' the latent variables into groups of variables that mostly depend on specific factors, known as a *varimax* rotation

A little more dubious rotation relaxes the demand of orthogonality and allows skew coordinate systems and other variances than 1 on the latent factors. Such rotations are *oblique*

Example

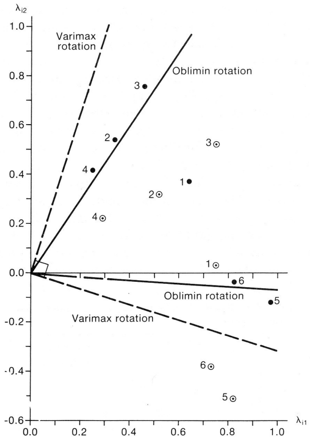
This example is taken from Bartholomew (1987) and is concerned with 6 different scores in intelligent tests. The $p = 6$ manifest variables are

1. Spearman's G-score
2. Picture completion test
3. Block Design
4. Mazes
5. Reading comprehension
6. Vocabulary

A 1-factor model gives a deviance of 75.56 with 9 degrees of freedom and is clearly inadequate.

A 2-factor model gives a deviance of 6.07 with 4 degrees of freedom and appears appropriate.

The loadings of each of the 6 variables can be displayed as black dots in the following diagram


λi2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
Varimax rotation
Oblimin rotation
Oblimin rotation
Varimax rotation
0.0
0.2
0.4
0.6
0.8
1.0
λi1

This diagram also shows axes corresponding to varimax and oblique rotations

It is tempting to conclude that 2, 3 and 4 seem to be measuring the same thing, whereas 5 and 6 are measuring something else. The G-score measures a combination of the two.

The axes of the oblique rotation represent the corresponding "dimensions of intelligence".

Or is it all imagination?