Latent Variable Models and Factor Analysis

MSc Further Statistical Methods
Lectures 6 and 7
Hilary Term 2007

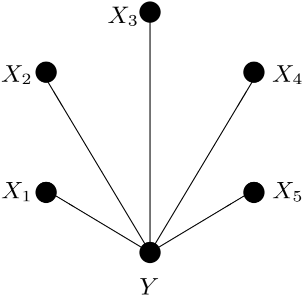
Steffen Lauritzen, University of Oxford; February 8, 2007

Basic idea

Latent variable models attempt to explain complex relations between several variables by simple relations between the variables and an underlying unobservable, i.e. *latent* structure.

Formally we have a collection $X = (X_1, \dots, X_p)$ of *manifest* variables which can be observed, and a collection $Y = (Y_1, \dots, Y_q)$ of *latent* variables which are unobservable and 'explain' the dependence relationships between the manifest variables.

Here 'explaining' means that the *manifest variables are assumed to be conditionally independent given the latent variables*, corresponding e.g. to the following graph:



Here Y is the latent variable(s) and there are 5 manifest variables $X_1, \ldots, X_5$.

For the model to be useful, *q must be much smaller than p*.

Data available will be repeated observations of the vector $X = (X_1, \ldots, X_p)$ of manifest variables.

Latent variable models are typically classified according to the following scheme:

	Manifest variable	
Latent variable	Metrical	Categorical
Metrical	Factor analysis	Latent trait analysis
Categorical	Latent profile analysis	Latent class analysis

Other terminologies are used, e.g. *discrete factor analysis* for latent trait analysis.

Categorical variables can either be ordinal or nominal, and metrical variables can either be discrete or continuous.

An example

A classical latent trait model is behind intelligence testing.

The intelligence of any individual is assumed to be a latent variable Y measured on a continuous scale.

An intelligence test is made using a battery of p tasks, and an individual scores $X_i = 1$ if the individual solves task i and 0 otherwise.

The test is now applied to a number of individuals to establish and estimate the parameters in the model.

Subsequently the test battery will be used to estimate the intelligence of a given individual by using

$$\mathbf{E}(Y \mid X_1 = x_1, \ldots, X_p = x_p)$$

as the estimate of intelligence for a given individual with score results $x = (x_1, \ldots, x_p)$.

Typical models will now have the intelligence distributed as

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

and the manifest variables as

$$\pi_i(y) = P(X_i = 1 \,|\, Y = y) = \frac{e^{\alpha_i + \beta_i y}}{1 + e^{\alpha_i + \beta_i y}}$$

corresponding to

$$\operatorname{logit}\{\pi_i(y)\} = \alpha_i + \beta_i y,$$

i.e. the response for each item being a logistic regression on the latent intelligence.

This model has too many parameters so we need to standardise and choose e.g. $\mu = 0$ and $\sigma^2 = 1$ to have a chance of estimating α_i and β_i.

We may increase the dimensionality of this model by assuming Y and β_i are q-dimensional and have

$$Y \sim \mathcal{N}_q(0, I), \quad \operatorname{logit}\{\pi_i(y)\} = \alpha_i + \beta_i^\top y.$$

This model is known as the *logit/normit model*.

Estimation is typically done by the EM-algorithm. The E-step involves numerical integration and the M-step needs in principle iterative methods as well.

See Bartholomew and Knott (1999), pp. 80–83 for details.

Estimation in latent variable models

Historically, algorithms for maximizing the likelihood function have been developed separately for each specific model.

Generally, estimation problems can be very difficult and there are problems with uniqueness of estimates.

The difficulties show in particular if sample sizes are small and p is not large relatively to q.

There are also severe problems with the asymptotic distribution of likelihood ratio tests.

Latent variable models are perfectly suitable for the EM algorithm as Y is MCAR.

However, the general ‘well-established’ knowledge is that the EM algorithm is too slow.

Typicallly, the EM algorithm quickly gets close to the MLE, but then slows down. This suggests a hybrid approach to be suitable, where the EM algorithm is applied initially to get good starting values, then special algorithms for the final convergence.

MIM implements a version of the EM-algorithm which is applicable for *latent class analysis, latent profile analysis,* and *factor analysis,* but *not* latent trait analysis.

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

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

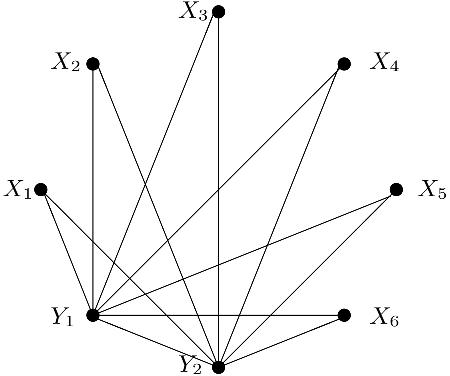
where Y and U are independent and follow multivariate normal distributions

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

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*.

Dependence graph of LNF model



Graph only displays conditional independences. In addition, $Y_1 \perp\!\!\!\perp Y_2$.

Linear factor analysis

The *idea* of the LNF model 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

$\tilde{Y} = OY$ and $\tilde{\Lambda} = \Lambda O^\top$ we have

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

and thus

$$X = \mu + \Lambda Y + U = X + \mu + \tilde{\Lambda}\tilde{Y} + U.$$

Since also $\tilde{Y} \sim \mathcal{N}_q(0, I)$ and

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

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

Hence Λ is only identifiable modulo orthogonal equivalence.

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

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

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 is a viable alternative which may not be sufficiently well exploited. See B & K(1999), §3.6 for details of this.

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, corresponding to possible dependence among the factors. Such rotations are *oblique*.

Example

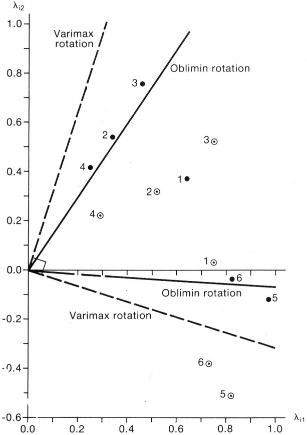
This example is taken from Bartholomew (1987) and is concerned with 6 different scores in intelligent tests. The $p = 6$ manifest variables are

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
3
2
4
1
3
2
4
1
6
5
Oblimin rotation
Varimax rotation
6
5
0.0
0.2
0.4
0.6
0.8
1.0
λi1

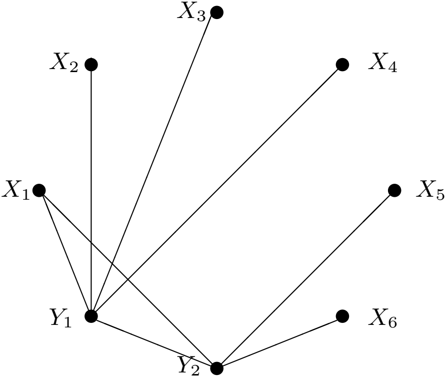
This diagram also shows axes corresponding to varimax and oblique rotations

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?

Dependence graph of simplified model



Y_1 and Y_2 are no longer independent.