Exploratory Factor Analysis

Tsung-Chi Cheng

Department of Statistics
National Chengchi University
Taipei 11605, Taiwan

E-mail: chengt@nccu.edu.tw

- Assume that we have a set of observed or manifest variables, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$, which are assumed to be linked to k unobserved latent variables or common factors $\boldsymbol{F} = (\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_k)$, where $k < p$, by a regression model of the form

$$
\begin{aligned}
\boldsymbol{x}_1 &= \lambda_{11}\boldsymbol{f}_1 + \lambda_{12}\boldsymbol{f}_2 + \; + \lambda_{1k}\boldsymbol{f}_k + u_1 \\
\boldsymbol{x}_2 &= \lambda_{21}\boldsymbol{f}_1 + \lambda_{22}\boldsymbol{f}_2 + \; + \lambda_{2k}\boldsymbol{f}_k + u_2 \\
&\vdots \\
\boldsymbol{x}_p &= \lambda_{p1}\boldsymbol{f}_1 + \lambda_{p2}\boldsymbol{f}_2 + \; + \lambda_{pk}\boldsymbol{f}_k + u_p
\end{aligned}
$$

- The λ_js are essentially the regression coefficients of the X-variables on the common factors,
  - in the context of factor analysis these regression coefficients are known as the factor loadings and show how each observed variable, $\boldsymbol{x}_i$, depends on the common factors.

- The factor loadings are used in the interpretation of the factors
- Larger values relate a factor to the corresponding observed variables and from these we can often, but not always, infer a meaningful description of each factor

- The regression equations above may be written more concisely as

$$\boldsymbol{X} = \boldsymbol{\Lambda f} + \boldsymbol{u}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \lambda_{21} & \cdots & \lambda_{2k} \\ \vdots & \vdots & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pk} \end{pmatrix} \quad \boldsymbol{F} = \begin{pmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \\ \vdots \\ \boldsymbol{f}_k \end{pmatrix} \quad \boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}$$

- We assume that the random disturbance terms $u_1, u_2, \cdots, u_p$ are uncorrelated with each other and with the factors $\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_k$.
- The elements of $\boldsymbol{u}$ are specific to each $\boldsymbol{x}_i$ and hence are generally better known in this context as *specific variates*.

- The two assumptions imply that, given the values of the common factors, the manifest variables are independent
  - that is, the correlations of the observed variables arise from their relationships with the common factors.
- Because the factors are unobserved, we can fix their locations and scales arbitrarily and we shall assume they occur in standardised form with mean zero and standard deviation one.
- We will also assume, initially at least, that the factors are uncorrelated with one another, in which case the factor loadings are the *correlations* of the manifest variables and the factors.

- With these additional assumptions about the factors, the factor analysis model implies that the variance of variable $\boldsymbol{x}_i$, σ_i^2, is given by

$$\sigma_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i,$$

where ψ_i is the variance of u_i.
- Consequently, we see that the factor analysis model implies that the variance of each observed variable can be split into two parts
    - $\sum_{j=1}^{k} \lambda_{ij}^2$ is known as the *communality* of the variable and represents the variance shared with the other variables via the common factors.
    - ψ_i is called the *specific* or *unique* variance and relates to the variability in X_i not shared with other variables.

- In addition, the factor model leads to the following expression for the covariance of variables X_i and X_j

$$\sigma_{ij} = \sum_{\ell=1}^{k} \lambda_{i\ell}\lambda_{j\ell}$$

- We see that the covariances are *not* dependent on the specific variates in any way.
- It is the common factors only that aim to account for the relationships between the manifest variables.

- The results above show that the k-factor analysis model implies that the population covariance matrix, $\mathbf{\Sigma}$, of the observed variables has the form

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$$

  where $\mathbf{\Psi} = diag(\Psi_i)$.
- In practise, $\mathbf{\Sigma}$ will be estimated by the sample covariance matrix $\boldsymbol{S}$.

Estimating the parameters in the k-factor analysis model

- The estimation problem in factor analysis is essentially that of finding $\mathbf{\Lambda}$ (the estimated factor loading matrix) and $\mathbf{\Psi}$ (the diagonal matrix containing the estimated specific variances)
- Assuming the factor model outlined in $\mathbf{\Sigma}$, reproduce as accurately as possible the sample covariance matrix, $\boldsymbol{S}$

$$\boldsymbol{S} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}}$$

- Given an estimate of the factor loading matrix, $\hat{\mathbf{\Lambda}}$, it is clearly sensible to estimate the specific variances as

$$\hat{\Psi}_i = s_i^2 - \sum_{j=1}^{k} \hat{\lambda}_{ij}^2 \qquad i = 1, 2, \cdots, p$$

- The diagonal terms in $\boldsymbol{S}$ are estimated exactly.
- There are two main methods of estimation leading to what are known as *principal factor analysis* and *maximum likelihood factor analysis*, both of which are now briefly described.

Principal factor analysis I

- Principal factor analysis is an eigenvalue and eigenvector technique similar in many respects to principal components analysis but operating not directly on $\boldsymbol{S}$ (or $\boldsymbol{R}$) but on what is known as the *reduced covariance matrix*, $\boldsymbol{S}^*$, defined as

$$\boldsymbol{S}^* = \boldsymbol{S} - \hat{\boldsymbol{\Psi}}$$

  where $\hat{\boldsymbol{\Psi}}$ is a diagonal matrix containing estimates of the $\boldsymbol{\Psi}$.
- The "ones" on the diagonal of $\boldsymbol{S}$ have in $\boldsymbol{S}^*$ been replaced by the estimated communalities, $\sum_{j=1}^{k} \hat{\lambda}_{ij}^2$, the parts of the variance of each observed variable that can be explained by the common factors.

Principal factor analysis II

- Unlike principal components analysis, factor analysis does not try to account for *all* the observed variance, only that shared through the common factors.
- Of more concern in factor analysis is accounting for the covariances or correlations between the manifest variables.
- To calculate $\boldsymbol{S}^*$ (or with $\boldsymbol{R}$ replacing $\boldsymbol{S}$, $\boldsymbol{R}^*$) we need values for the communalities.
- Clearly we cannot calculate them on the basis of factor loadings because these loadings still have to be estimated.
- To get around this seemingly "chicken and egg" situation, we need to find a sensible way of finding initial values for the communalities that does not depend on knowing the factor loadings.

Principal factor analysis III

- When the factor analysis is based on the correlation matrix of the manifest variables, two frequently used methods are:
  - Take the communality of a variable X_i as the square of the multiple correlation coefficient of X_i with the other observed variables.
  - Take the communality of X_i as the largest of the absolute values of the correlation coefficients between X_i and one of the other variables.
- Each of these possibilities will lead to higher values for the initial communality when X_i is highly correlated with at least some of the other manifest variables, which is essentially what is required.

Principal factor analysis IV

- Given the initial communality values, a principal components analysis is performed on $\boldsymbol{S}^*$ and the first k eigenvectors used to provide the estimates of the loadings in the k-factor model.
- The estimation process can stop here or the loadings obtained at this stage can provide revised communality estimates calculated as $\sum_{j=1}^{k} \hat{\lambda}_{ij}^2$, where the $\hat{\lambda}_{ij}^2$s are the loadings estimated in the previous step.
- The procedure is then repeated until some convergence criterion is satisfied.
- Difficulties can sometimes arise with this iterative approach if at any time a communality estimate exceeds the variance of the corresponding manifest variable, resulting in a negative estimate of the variable's specific variance.

Maximum likelihood factor analysis

- Maximum likelihood is regarded, by statisticians at least, as perhaps the most respectable method of estimating the parameters in the factor analysis.
- The essence of this approach is to assume that the data being analysed have a multivariate normal distribution. Under this assumption and assuming the factor analysis model holds, the likelihood function L can be shown to be $-\frac{1}{2}nF$ plus a function of the observations where F is given by

$$F = \ln|\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}| + trace(\boldsymbol{S}|\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}|^{-1}) - \ln|\boldsymbol{S}| - p$$

- The function F takes the value zero if $\mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$ is equal to $\boldsymbol{S}$ and values greater than zero otherwise.
- Estimates of the loadings and the specific variances are found by minimising F with respect to these parameters.
- A number of iterative numerical algorithms have been suggested in the literature.
- factanal() in R