Chapter 9
Factor Analysis

Factor Analysis and Principal Component Analysis

The principal component analysis is used for data reduction. Similar to principal component analysis, factor analysis is a multivariate method used for data reduction. The idea is to represent a set of variables by a smaller number of variables, called factors. Factor analysis represents the relationships among multiple correlated/interconnected variables by identifying a few underlying factors.

Factor analysis differs from analysis of variance, multiple linear regression, and discriminant analysis in that, in all such techniques, one variable is considered dependent and the others independent, whereas factor analysis does not make such a distinction. Factor analysis helps explore the possible interrelationships among multiple variables and assess the underlying causes of these relationships.

In factor analysis, the observed variables are expressed as linear functions of a smaller number of underlying factors, analogous to a multiple linear regression model. The key difference is that the explanatory variables, called factors, are not directly observed but are latent variables that are inferred from the correlations among the observed variables.

The model in factor analysis is based on the observed vector, which is partitioned into an unobserved systematic part and an unobserved error part. The systematic part is taken as a linear combination of a relatively small number of unobserved factor variables. The components of error vectors are considered as uncorrelated or independent.

Factor analysis reveals the underlying factors that explain the correlations among variables. It also helps create a smaller set of uncorrelated variables, simplifying further multivariate analyses such as discriminant or regression analysis. It helps identify a smaller and more meaningful set of variables for further use in multivariate analysis.

The applications of factor analysis are exploratory in nature, so it is also called as exploratory factor analysis (EFA).

The essential purpose is to describe the covariance relationships among many variables in terms of a few underlying, but unobservable random quantities called factors. The motivation behind the factor model is that, suppose variables can be grouped by their correlations, such that all variables within a group are highly correlated among themselves but have relatively small correlations with variables in a different group. It can be inferred that each group of variables represents a single underlying construct (or factor) that accounts for the observed correlations. Factor analysis was originally developed for analysing scores on mental tests. For example, when a battery of tests is given to a group of individuals, it is observed that the score of an individual on a given test is more related to his scores on other tests than to the scores of other individuals on the other tests, i.e., usually the scores for any particular individual are interrelated to some degree. This interrelation is "explained" by condensing a test score of an individual as made up of a part which is peculiar to this particular test (called error) and a part which is a function of more fundamental quantities called factor scores. Since they enter several test scores, it is their effect that connects the various test scores. The idea behind this is that a person who is more intelligent in some respects will do better on many tests than someone who is less intelligent.

Both factor analysis and principal component analysis aim to reduce the dimensionality of the data, but they differ in their underlying approaches. Principal component analysis constructs new variables, called principal components, which are linear combinations of the observed variables. In contrast, factor analysis models the observed variables as linear functions of a smaller number of underlying latent factors.

The Model

The factor model can be considered in the term of a multiple linear regression model where the observable variable X_i is to be predicted from the values of unobservable common factors f_i as follows:

$$\begin{aligned}
X_1 &= \mu_1 + \lambda_{11} f_1 + \lambda_{12} f_2 + \ldots + \lambda_{1m} f_m + \epsilon_1 \\
X_2 &= \mu_2 + \lambda_{21} f_1 + \lambda_{22} f_2 + \ldots + \lambda_{2m} f_m + \epsilon_2 \\
&\vdots \\
X_p &= \mu_p + \lambda_{p1} f_1 + \lambda_{p2} f_2 + \ldots + \lambda_{pm} f_m + \epsilon_p
\end{aligned}$$

where $E(X_i) = \mu_i$ is the population mean of variable i and can be viewed as intercept terms in the multiple regression model. The i^{th} common factor is denoted by f_i and generally $m << p$. These variables can be expressed as

$$\underset{\sim}{X} = \underset{\sim}{\mu} + \Lambda \underset{\sim}{f} + \underset{\sim}{\epsilon}$$

where $\underset{\sim}{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)'$ is a $p \times 1$ vector of population mean, $\underset{\sim}{f} = (f_1, f_2, \cdots, f_m)'$ is an $m \times 1$ vector of common factors. The loading of the i^{th} variable on the j^{th} factor is denoted by l_{ij} and are represented in the following $p \times m$ matrix

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \ldots & \lambda_{pm} \end{pmatrix}$$

as the matrix of factor loadings, the errors ϵ_i is called the specific factor for the variable i and $\underset{\sim}{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_p)'$ is a $p \times 1$ vector of specific factors. Note that l_{ij}'s are like regression coefficients and each of the response variables X is predicted as a linear function of the unobserved common factors $f_1, f_2, \ldots, f_m$, which are m explanatory variables. Thus m unobserved factors control the variation in the data.

Assumptions

We assume $E(\underset{\sim}{\epsilon}) = 0$, $E(\underset{\sim}{\epsilon}\underset{\sim}{\epsilon}') = \Psi$ is a diagonal matrix and $\underset{\sim}{\epsilon}$ is distributed independently of $\underset{\sim}{f}$. The vector $\underset{\sim}{f}$ is treated as a random vector and as a vector of parameters that varies from observation to observation.

In terms of mental tests, each component of X is a score on a test and corresponding component of $\underset{\sim}{\mu}$ is the average score of this test in the population. The components of f are the scores of the mental factors, and the linear combinations of these enter into the test scores. The coefficients of these linear combinations are the factor loadings, which are the elements of Λ.

The elements of $\underset{\sim}{f}$ are also called as common factors because they are common to several different tests. The components of $\underset{\sim}{\epsilon}$ are the part of the test score not explained by the common factors and is considered as error of measurement in the test, along with a specific factor having to do only with this particular test. Since in our model, with one set of observations on each individual, we cannot distinguish between these two components of $\underset{\sim}{\epsilon}$, so we call them as error of measurement.

Based on the nature of $\underset{\sim}{f}$, we can have two types of models.

Model 1: We consider $\underset{\sim}{f}$ to be nonrandom. The components of $\underset{\sim}{f}$ varies from one individual to another. It is more accurate to write

$$X_\alpha = \mu + \Lambda f_\alpha + \epsilon$$

where the nonrandom factor score may describe the systematic part better, but the likelihood function may not have a maximum, which may pose problems of inference. It is suitable when the specific individuals involved and not just the structure are of interest.

Model 2: We consider $\underset{\sim}{f}$ to be random, which is more appropriate when different samples consist of different individuals.

When $\underset{\sim}{f}$ is taken as random, we assume $E(\underset{\sim}{f}) = 0$.

When $\underset{\sim}{f}$ is taken as nonrandom, we assume $E(\underset{\sim}{X}) = \underset{\sim}{\mu} + \Lambda E(\underset{\sim}{f})$ and $\underset{\sim}{\mu}$ can be redefined

to absorb $\Lambda E(\underset{\sim}{f})$ and the model is usually expressed as

$$\underset{\sim}{X} - \underset{\sim}{\mu} = \Lambda \underset{\sim}{f} + \underset{\sim}{\epsilon}.$$

Let $E(\underset{\sim}{f}\underset{\sim}{f}') = \Phi$ when $\underset{\sim}{f}$ is random and usually, $\underset{\sim}{f}$ and $\underset{\sim}{\epsilon}$ are considered to have normal distributions.

If f is not random, then $\underset{\sim}{f} = \underset{\sim}{f}_\alpha$ for the α^{th} individual and then we assume

$$\frac{1}{N}\sum_{\alpha=1}^{N} \underset{\sim}{f}_\alpha = 0 \quad \text{and} \quad \frac{1}{N}\sum_{\alpha=1}^{N} \underset{\sim}{f}_\alpha \underset{\sim}{f}_\alpha' = \Phi.$$

Indeterminacy in the Model

Let C be an $m \times m$ nonsingular matrix. Then let $\underset{\sim}{f}^* = C^{-1}f$ and $\Lambda^* = \Lambda C$ and the model

$$\underset{\sim}{X} = \underset{\sim}{\mu} + \Lambda \underset{\sim}{f} + \underset{\sim}{\epsilon} \tag{1}$$

can be written as

$$\begin{aligned} \underset{\sim}{X} &= \underset{\sim}{\mu} + \Lambda C C^{-1} \underset{\sim}{f} + \underset{\sim}{\epsilon} \\ &= \underset{\sim}{\mu} + \Lambda^* \underset{\sim}{f}^* + \underset{\sim}{\epsilon} \end{aligned} \tag{2}$$

When $\underset{\sim}{f}$ is random, then

$$E(\underset{\sim}{f}^* \underset{\sim}{f}^{*'}) = C^{-1}E(\underset{\sim}{f}\underset{\sim}{f}')(C^{-1})' = C^{-1}\Phi(C^{-1})' = \Phi^*.$$

When $\underset{\sim}{f}$ is nonrandom, then

$$\frac{1}{N}\sum_{\alpha=1}^{N} \underset{\sim}{f}_\alpha^* \underset{\sim}{f}_\alpha^{*'} = \Phi^*.$$

The models (1) with Λ and $\underset{\sim}{f}$ is equivalent to model (2) with Λ^* and $\underset{\sim}{f}^*$ in the sense that by observing $\underset{\sim}{X}$, we cannot distinguish between model (1) and model (2).

Some of the indeterminacy can be eliminated by requiring

$$E(\underset{\sim}{f}\underset{\sim}{f}') = I \quad \text{if } \underset{\sim}{f} \text{ is random}$$

or

$$\sum_{\alpha=1}^{N} \underset{\sim}{f}_{\alpha} \underset{\sim}{f}'_{\alpha} = NI \quad \text{if } \underset{\sim}{f} \text{ is nonrandom.}$$

In this case, the factors are said to be **orthogonal**.

If Φ is not diagonal, the factors are said to be **oblique**.

When $\Phi = I$, then

$$E(\underset{\sim}{f}^{*} \underset{\sim}{f}^{*'}) = C^{-1}(C^{-1})' = I \qquad (I = CC').$$

The indeterminacy is equivalent to multiplication by an orthogonal matrix. This is called as **problem of rotation**.

Requiring Φ to be diagonal needs components of $\underset{\sim}{f}$ to be independently distributed when $\underset{\sim}{f}$ is normally distributed. This condition appeals to psychologists because, by definition, one idea about common mental factors is that they are independent or uncorrelated quantities.

The assumption that the components of $\underset{\sim}{\epsilon}$ are uncorrelated is crucial. One point of view is that the errors of observation and the specific factors are, by definition, uncorrelated; i.e., the interrelationships among the test scores are caused by common factors that we want to investigate. Another point of view is that the common factors are supposed to explain as much of the variance of the test scores as possible.

If the factors are random, then the covariance matrix of observed X is

$$\begin{aligned} \Sigma &= E(\underset{\sim}{X} - \underset{\sim}{\mu})(\underset{\sim}{X} - \underset{\sim}{\mu})' \\ &= E(\Lambda \underset{\sim}{f} + \underset{\sim}{\epsilon})(\Lambda \underset{\sim}{f} + \underset{\sim}{\epsilon})' \\ &= \Lambda \Phi \Lambda' + \Psi. \end{aligned}$$

In case factors are orthogonal (i.e., $E(\underset{\sim}{f} \underset{\sim}{f}') = I$, then

$$\Sigma = \Lambda\Lambda' + \Psi.$$

If $\underset{\sim}{f}$ and $\underset{\sim}{\epsilon}$ are normal, all the information about the structure comes from Σ and $E(\underset{\sim}{X}) = \mu$.

Identification

Consider the following question: Given Σ $(= \Lambda\Phi\Lambda' + \Psi)$ and a number m of factors, can we find whether there exists a triplet

$$(\Lambda, \Phi > 0, \Psi > 0 \text{ and diagonal})$$

to satisfy $\Lambda\Phi\Lambda' + \Psi = \Sigma$, and if such a triplet is found, is it unique?

Since any triplet can be transformed into an equivalent structure ΛC, $C^{-1}\Phi C^{-1'}$ and Ψ, we can put m^2 independent conditions on Λ and Φ to rule out this indeterminancy.

The number of parameters in

- Λ is pm.
- Φ is $\frac{m(m+1)}{2}$ and
- Ψ is p.

So total number of parameters is $pm + \frac{m(m+1)}{2} + p = Q$, say.

The number of components in the observable Σ and the number of conditions (for uniqueness) is $\frac{p(p+1)}{2} + m^2 = R$, say.

The excess of observed quantities and conditions over the number of parameters is $R - Q = \frac{1}{2}[(p-m)^2 - p - m]$.

- If this excess is positive, one can expect a problem of existence, but can anticipate uniqueness if a set of parameters does exist.
- If this excess is negative, we can expect existence but possibly not uniqueness.
- If this excess is zero, we can hope for both existence and uniqueness (or at least a finite number of solutions).

The question of the existence of a solution is whether there exists a diagonal matrix Ψ with nonnegative diagonal entries such that $(\Sigma - \Psi)$ is positive semidefinite matrix of rank m.

The model is said to be **identified** if a solution exists and is unique.

Some m^2 conditions have to be put on Λ and Φ to eliminate a transformation $\Lambda^* = \Lambda C$ and $\Phi^* = C^{-1}\Phi C^{-1'}$. The condition $\Phi = I$ forces a transformation C to be orthogonal and there are $\frac{m(m+1)}{2}$ component equations in $\Phi = I$.

Sometimes, the restrictions $\Gamma = \Lambda'\Psi^{-1}\Lambda$ is diagonal is added. If the diagonal elements of Γ are ordered and different ($\gamma_{11} > \gamma_{22} > \ldots > \gamma_{mm}$), Λ is uniquely determined.

Units of Measurements

In many cases, the units of measurement of each component in $\underset{\sim}{X}$ are different. If the units of measurement are to be changed, it means multiplying each component of $\underset{\sim}{X}$ by a constant, and these constants may not necessarily be equal. When a given test score is multiplied by a constant, the factor loadings for the test are multiplied by the same constant and the error variance is multiplied by the square of the constant.

Let D be a diagonal matrix with positive diagonal elements. Consider the model

$$\underset{\sim}{X} = \underset{\sim}{\mu} + \Lambda \underset{\sim}{f} + \underset{\sim}{\epsilon}$$

and premultiply it with D as

$$D\underset{\sim}{X} = D\underset{\sim}{\mu} + D\Lambda\underset{\sim}{f} + D\underset{\sim}{\epsilon}$$
$$\text{or} \quad \underset{\sim}{X}^* = \underset{\sim}{\mu}^* + \Lambda^*\underset{\sim}{f} + \underset{\sim}{\epsilon}^*$$

where $\underset{\sim}{X}^* = D\underset{\sim}{X}$, $\underset{\sim}{\mu}^* = E(\underset{\sim}{X}^*) = D\underset{\sim}{\mu}$, $\Lambda^* = D\Lambda$, $\underset{\sim}{\epsilon}^* = D\underset{\sim}{\epsilon}$, $Cov(\underset{\sim}{\epsilon}^*) = D\Psi D$.

Then

$$E(\underset{\sim}{X}^* - \underset{\sim}{\mu}^*)(\underset{\sim}{X}^* - \underset{\sim}{\mu}^*)' = \Lambda^*\Phi\Lambda^{*'} + \Psi^* = \Sigma^*$$

where $\Sigma^* = D\Sigma D$. If the identification conditions are $\Phi = I$ and $\Lambda'\Psi^{-1}\Lambda$ is diagonal, then Λ^* satisfies these conditions.

Model with Orthogonal Factors

Now we restrict our attention to a model with orthogonal factors and develop the analysis further for this model.

For review, the model is now

$$\underset{\sim}{X} = \underset{\sim}{\mu} + \Lambda \underset{\sim}{f} + \underset{\sim}{\epsilon}$$

with $E(\underset{\sim}{\epsilon}) = 0$, $E(\underset{\sim}{f}) = 0$, $E(\underset{\sim}{X}) = \underset{\sim}{\mu}$, $Cov(\underset{\sim}{f}) = E(\underset{\sim}{f}\underset{\sim}{f}') = I$, $Cov(\underset{\sim}{\epsilon}) = E(\underset{\sim}{\epsilon}\underset{\sim}{\epsilon}') = \Psi$ and the specific factors are uncorrelated with the common factors, i.e., $Cov(\epsilon_i, f_j) = 0$, $i = 1, 2, \ldots, p; j = 1, 2, \ldots, m$.

These assumptions are necessary for the unique estimation of the parameters.

The **communality** for variable i is $\sum_{j=1}^{m} \lambda_{ij}^2$ and variance of trait i is

$$\sigma_i^2 = Var(X_i) = \sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i.$$

The covariance between trait i and factor j is $Cov(X_i, f_j) = \lambda_{ij}$. So we can express

$$\Sigma = \Lambda\Lambda' + \Psi$$

where Λ is the matrix of factor loadings and $\Psi = diag(\psi_1, \psi_2, \ldots, \psi_p)$ is the diagonal matrix of specific variances.

It is worth noting that the model assumes the data are linear functions of unobservable common factors, so the linearity cannot be checked.

The covariance matrix Σ, having $\frac{p(p+1)}{2}$ unique elements, is approximated by mp factor loadings in Λ and p specific variances. Ideally $(mp + p) << \frac{p(p+1)}{2}$. In case mp is too small, then $(mp + p)$ parameters may not adequately describe Σ. In such conditions, the model may not be the right one, but we cannot reduce the data to a linear combination of factors.

It is important to note that the calculations in the original model and the model obtained by an orthogonal matrix remain the same. This is shown as follows.

With an $m \times m$ orthogonal matrix D, the model

$$\underset{\sim}{X} = \underset{\sim}{\mu} + \Lambda \underset{\sim}{f} + \underset{\sim}{\epsilon}$$

is written as

$$\begin{aligned}\underset{\sim}{X} &= \underset{\sim}{\mu} + \Lambda DD' \underset{\sim}{f} + \underset{\sim}{\epsilon} \\ &= \underset{\sim}{\mu} + \Lambda^* \underset{\sim}{f}^* + \underset{\sim}{\epsilon}\end{aligned}$$

and does not change any calculations because $DD' = I_m$ and multiplication by I to any matrix does not alter the original matrix.

Also,

$$\begin{aligned}E(\underset{\sim}{f}^*) &= E(D'\underset{\sim}{f}) = D'E(\underset{\sim}{f}) = 0 \\ Var(\underset{\sim}{f}^*) &= D'Var(\underset{\sim}{f})D = D'ID = I \\ Cov(\underset{\sim}{f}^*, \underset{\sim}{\epsilon}) &= Cov(D'\underset{\sim}{f}, \underset{\sim}{\epsilon}) = D'Cov(\underset{\sim}{f}, \underset{\sim}{\epsilon}) = 0.\end{aligned}$$

Thus $\underset{\sim}{f}^*$ satisfies all the assumptions and is an equally valid collection of common factors.

It is worth noting that there are different choices of orthogonal matrices that give rise to different models, but all these models satisfy the model assumptions.

Next we consider the estimation of parameters of a factor model. The parameters can be estimated by the principal component method, the maximum likelihood method, the principal factor method, etc.

Principal Component Method of Estimation of Parameters

It may be recalled that the principal component method is used for dimension reduction and its methodology is based on the characteristic roots and characteristic vectors of covariance matrix. So, the principal component method should be understood in the context of the developments and details of principal component analysis.

Let the observations for the l^{th} subject on p variables is expressed in a $p \times 1$ vector $\underset{\sim}{X}_l = (X_{l1}, X_{l2}, \ldots, X_{lp})'$ and its sample covariance matrix is

$$S = \frac{1}{n-1} \sum_{l=1}^{n} (\underset{\sim}{X}_l - \bar{\underset{\sim}{X}})(\underset{\sim}{X}_l - \bar{\underset{\sim}{X}})'$$

which is based on $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n$. We find the characteristic roots of S as $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$ and the corresponding characteristic vectors as $\underset{\sim}{\hat{e}}_1, \underset{\sim}{\hat{e}}_2, \ldots, \underset{\sim}{\hat{e}}_p$, respectively.

Using the spectral decomposition theorem over Σ, we can express

$$\Sigma = \sum_{i=1}^{p} \lambda_i \underset{\sim}{e}_i \underset{\sim}{e}_i' \approx \sum_{i=1}^{m} \lambda_i \underset{\sim}{e}_i \underset{\sim}{e}_i'$$

where the last $(p-m)$ terms are ignored, similar to the case in principal component analysis, and Σ is approximated by first m terms. This can be further expressed in terms of Λ as

$$\begin{aligned}
\Sigma &= \sum_{i=1}^{m} \lambda_i \underset{\sim}{e}_i \underset{\sim}{e}_i' \\
&= (\sqrt{\lambda_1}\underset{\sim}{e}_1, \sqrt{\lambda_2}\underset{\sim}{e}_2, \ldots, \sqrt{\lambda_m}\underset{\sim}{e}_m) \begin{pmatrix} \sqrt{\lambda_1}\underset{\sim}{e}_1' \\ \sqrt{\lambda_2}\underset{\sim}{e}_2' \\ \vdots \\ \sqrt{\lambda_m}\underset{\sim}{e}_m' \end{pmatrix} \\
&= \Lambda\Lambda'.
\end{aligned}$$

It is important to note that the maximum likelihood estimates of the characteristic roots and characteristic vectors in principal component analysis were obtained as the sample-based characteristic roots and characteristic vectors from the sample covariance matrix. In case the observations are standardized, the sample covariance matrix will be replaced by the sample correlation matrix.

Thus we obtain

$$\hat{\lambda}_{ij} = \hat{e}_{ji}\sqrt{\hat{\lambda}_j}$$

Since $\Psi = \Sigma - \Lambda\Lambda'$, so the diagonal elements of Ψ, are estimated by

$$\hat{\psi}_i = s_i^2 - \sum_{j=1}^{m} \hat{\lambda}_j \hat{e}_{ji}^2$$

where s_i^2 is the sample variance for the i^{th} variable.

The factor loadings, which are the correlations between the factors and the variables, are obtained by

$$\hat{\underset{\sim}{e}}_i\sqrt{\hat{\lambda}_i}.$$

The interpretation of factor loadings is similar to that of the coefficients for principal components in principal component analysis. Some criterion for retaining the correlation coefficients is used, which may be arbitrary and can be developed within the framework of a test of hypothesis, as in the case of principal component analysis.

The communalities for the i^{th} variable are found from the factor loadings as

$$l_i^2 = \sum_{j=1}^{m} \hat{\lambda}_{ij}^2.$$

Performance Evaluation

The communality for a given variable is interpreted as the proportion of variation in that variable explained by the required variables of interest. It is equivalent to the coefficient of determination R^2 in the case of a multiple linear regression model in the sense that the model is fitted with these variables of interest.

The model performance can be assessed from the communalities. Values close to 1 are observed, indicating that the model explains most of the variation in these variables. In such a case, the model performs better for some variables than it does for others.

The sum of all communality values is the total communality value

$$\sum_{i=1}^{p} l_i^2 = \sum_{i=1}^{m} \lambda_i$$

The proportion of total variation explained by the selected factors is obtained as

$$\frac{\text{Total communality}}{\text{Number of variables}} = \frac{\sum_{i=1}^{p} l_i^2}{p}$$

which gives the percentage of variation explained by the model, and is considered as a measure of the overall assessment of the model performance. This proportion is the same as the proportion of variation explained by the m characteristic roots. The individual

communalities shed light on how the model performs for individual variables, and the total communality provides an overall assessment of performance.

As the data are standardized, so the variance of the standard data is equal to 1. The specific variances are computed by

$$\hat{\psi}_i = 1 - l_i^2$$

for all the variables.

Such a model provides an approximation to the correlation matrix, and the model's appropriateness can be assessed with the residuals obtained from the calculation

$$s_{ij} - \sum_{k=1}^{m} \lambda_{ik}\lambda_{jk}, \quad i \neq j = 1, 2, \ldots, p$$

which is essentially the difference between R and $\Lambda\Lambda'$. Such residuals are expected to be close to 0 as far as possible for a good model fit.

The principal component method does not provide a test of lack of fit, and the decisions are based on the numbers the experimenter feels are suitable. Such a test is available for the maximum likelihood method.

When the number of factors increases, the estimated factor loadings do not change in the principal component method, but they change in the maximum likelihood method.

As the number of factors increases, the communalities increase and tend towards 1 whereas the specific variances decrease towards zero.

The diagonal elements of the sample covariance matrix S are equal to the diagonal elements of $\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$, but the off-diagonal elements are not exactly reproduced due to variability in the data. Therefore, the number of factors is selected such that the off-diagonal elements of the residual matrix $S - (\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})$ are small.

When we want to choose m to be as small as possible, then the residual for that model can be large. On the other hand, when m is chosen to be large, the residuals will become smaller, but the model will become more complex and less interpretable.

The sum of squared elements of the residual matrix is equal to the sum of the squared values of the eigenvalues left out of the matrix, i.e., $\sum_{j=m+1}^{p} \hat{\lambda}_j^2$.

Maximum Likelihood Estimation of Parameters

We assume the observations are sampled from a multivariate normal distribution with a continuous random vector. Data in the social sciences are often on a Likert scale, which is discrete and bounded, which poses a limitation on the maximum likelihood estimates that can be used.

Assuming the $\underset{\sim}{f}_i$ and $\underset{\sim}{\epsilon}_i$ to follow jointly normal, then $\underset{\sim}{x}_i - \underset{\sim}{\mu} = \Lambda \underset{\sim}{f}_i + \underset{\sim}{\epsilon}_i$ are then normal and the likelihood function is

$$\begin{aligned}
L(\underset{\sim}{\mu}, \Sigma) &= \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}(\underset{\sim}{X}_i - \underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{X}_i - \underset{\sim}{\mu})\right] \\
&= \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left(\sum_{i=1}^{N}(\underset{\sim}{X}_i - \underset{\approx}{\bar{x}})(\underset{\sim}{X}_i - \underset{\approx}{\bar{x}})' + n(\underset{\approx}{\bar{x}} - \underset{\sim}{\mu})(\underset{\approx}{\bar{x}} - \underset{\sim}{\mu})'\right)\right\}\right] \\
&= \frac{1}{(2\pi)^{\frac{(N-1)p}{2}}|\Sigma|^{\frac{N-1}{2}}} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left(\sum_{i=1}^{N}(\underset{\sim}{X}_i - \underset{\approx}{\bar{x}})(\underset{\sim}{X}_i - \underset{\approx}{\bar{x}})'\right)\right\}\right] \\
&\quad \times \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{n}{2}(\underset{\approx}{\bar{x}} - \underset{\sim}{\mu})'\Sigma^{-1}(\underset{\approx}{\bar{x}} - \underset{\sim}{\mu})\right].
\end{aligned}$$

The likelihood function depends on Λ and Ψ through $\Sigma = \Lambda\Lambda' + \Psi$. This model is still not well-defined because of the multiplicity of choices for Λ enabled by orthogonal transformations.

The log likelihood is maximized to obtain the estimates of $\underset{\sim}{\mu}$, Λ and Ψ, but the solutions for these factor models are not unique. The equivalent models are obtained by rotation. For example, if $\Lambda'\Psi^{-1}\Lambda$ is a diagonal matrix, then a unique solution can be obtained. In general, iterative computation methods are employed to obtain the estimates. So there are no closed-form maximum-likelihood estimates, and the process is computationally intensive and complex. Thus the maximum likelihood estimates of Λ and Ψ are by numerically maximizing the log likelihood and the maximum likelihood estimates of $\underset{\sim}{\mu}$ is $\underset{\approx}{\bar{x}}$ subject to $\Lambda'\Psi^{-1}\Lambda$ being diagonal.

Asymptotic Distribution of the Estimates

Theorem: If Λ and Ψ are identified by $\Lambda'\Psi^{-1}\Lambda$ being diagonal, if the diagonal elements are different and ordered and if $C \xrightarrow{Pr} \Psi + \Lambda\Lambda'$, then $\hat{\Psi} \xrightarrow{Pr} \Psi$ and $\hat{\Lambda} \xrightarrow{Pr} \Lambda$ where $C = \frac{A}{N}$ and $A = \sum_{i=1}^{N}(\underset{\sim}{X}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{X}_i - \underset{\sim}{\bar{x}})'$.

A sufficient condition for $C \xrightarrow{Pr} \Sigma$ is that $(\underset{\sim}{f}', \underset{\sim}{\epsilon}')'$ has a distribution with finite second order moments.

The estimators $\hat{\Lambda}$ and $\hat{\Psi}$ are the solutions of

$$\hat{\Lambda}(\hat{\Lambda}'\hat{\Psi}^{-1}\hat{\Lambda} + I) = C\hat{\Psi}^{-1}\hat{\Lambda}$$

and

$$diag(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}) = diag(C)$$

with $\hat{\Lambda}'\hat{\Psi}^{-1}\hat{\Lambda}$ being diagonal.

Goodness of Fit

The goodness of fit test is based on the comparison of the covariance matrix under the assumed model $\Sigma = \Lambda\Lambda' + \Psi$ with the covariance matrix that can take any values.

We have considered a specific structure of Λ and Ψ (i.e., the diagonal elements of Ψ are equal to some specified values). A more general framework is to allow these elements to take any value.

The Bartlett corrected likelihood ratio test statistic

$$\chi^2 = \left[n - 1 - \frac{2p + 4m - 5}{6}\right] \log \frac{|\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}|}{|\hat{\Sigma}|}.$$

The smaller values of χ^2 indicate that the factor model fits well and larger values of χ^2 indicate that the factor model does not fit well.

Under the null hypothesis

H_0 : Factor model adequately describes the relationships among the variables,

$$\chi^2 \sim \chi^2_{\frac{(p-m)^2-p-m}{2}}.$$

So reject H_0 if $\chi^2 > \chi^2_{\frac{(p-m)^2-p-m}{2}}$.

Factor Rotations

The motivation for factor rotations is due to the fact that the factor models are not unique. Recall that the models $\underset{\sim}{X} = \underset{\sim}{\mu} + \Lambda \underset{\sim}{f} + \underset{\sim}{\epsilon}$ and $\underset{\sim}{X} = \underset{\sim}{\mu} + \Lambda^* \underset{\sim}{f}^* + \underset{\sim}{\epsilon}$ where $\Lambda^* = \Lambda D$ and $\underset{\sim}{f}^* = D' \underset{\sim}{f}$ for some orthogonal matrix D such that $DD' = I$ are equivalent. The choice of D corresponds to a particular factor rotation, and there are an infinite number of possible orthogonal matrices. So the aim here is to find an appropriate rotation, defined by an orthogonal matrix D, that yields easily interpretable factors.

This can be done by considering a scatter plot of factor loadings. The orthogonal matrix rotates the axes of this scatter plot and finds a rotation such that each of the p variables has a high loading on only one factor. Each rotation corresponds to a choice of orthogonal matrix D. Our aim is to rotate the axes to obtain a cleaner interpretation of the data and to define a new coordinate system so that, after rotation, the points fall close to the vertices of the new axes. For example, considering two factors, we want to find each of the plotted points at the four tips of the rotated axes. So the rotation accounts for the factor pattern plot and rotates the axes in such a way that the points fall close to the axes.

There are two types of rotation methods - orthogonal and oblique. In **orthogonal rotation** the rotated factors will remain uncorrelated whereas in **oblique rotation**, the resulting factors will be correlated. There are a number of different methods of rotation of each type. The most common orthogonal method is varimax rotation.

Varimax Rotation

The varimax rotation is the most widely used rotation criterion. It seeks the rotated loadings that maximize the variance of the squared loadings for each factor with the objective to make some of the loadings as large as possible and the rest as small as possible in absolute value. The varimax rotation criterion encourages the detection of

factors each of which is related to few variables and discourages the detection of factors influencing all variables.

The varimax rotation scales the loadings by dividing them by the corresponding communality as

$$\tilde{\lambda}^*_{ij} = \frac{\hat{\lambda}^*_{ij}}{l_i}.$$

The varimax rotation finds the rotation that maximizes the $\tilde{\lambda}^*_{ij}$. The varimax procedure selects the rotation, or equivalently, the orthogonal matrix D that makes

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

as large as possible.

This is the sample variance of the standardized loadings for each factor summed over the m factors. After determining D, the loadings $\tilde{\lambda}^*_{ij}$ are multiplied by l_i so that the original communalities are preserved. The simple interpretation from V is

$$V \propto \sum_{j=1}^{m}\left[\begin{array}{c}\text{variance of the square of}\\ \text{(scaled) loadings for } j^{th} \text{ factor}\end{array}\right].$$

Maximizing V corresponds to the square of the loadings on each factor as much as possible. So we expect to find groups of large and negligible coefficients in any column of the rotated loadings matrix $\hat{\Lambda}^*$.

Oblique Rotation

Orthogonal rotations are appropriate for factor models where the common factors are assumed to be independent. Oblique (non-orthogonal) rotations are also useful in the social sciences, where the factors considered are correlated. Oblique rotations are more complex than orthogonal rotations, as they can involve one or two coordinate systems, a system of primary axes, or a system of reference axes. The oblique rotation produces a pattern matrix that contains the factor loadings and a factor correlation matrix that includes the correlations between the factors. The commonly used oblique rotation techniques are Direct Oblimin and Promax.

The **Direct Oblimin technique** simplifies the structure and the mathematics of the output of the problem under consideration. The **Promax technique** involves raising the loadings to the fourth power, resulting in greater correlations among the factors and achieving a simple structure. The Promax technique is used in larger datasets because of its speed.

Estimation of Factor Scores

A factor score can be considered to be a variable describing how much an individual would score on a factor.

Considering the model

$$\underset{\sim}{X}_i = \underset{\sim}{\mu} + \Lambda \underset{\sim}{f}_i + \underset{\sim}{\epsilon}_i, \quad i = 1, 2, \ldots, n.$$

We aim to estimate the vectors of factor scores $\underset{\sim}{f}_1, \underset{\sim}{f}_2, \ldots, \underset{\sim}{f}_n$ for each observation using the ordinary least squares method, the weighted least squares method and the regression method.

1. Ordinary Least Squares Method

The vector of common factors for subject i, i.e., $\hat{\underset{\sim}{f}}_i$ is obtained by minimizing the sum of squares due to $\underset{\sim}{\epsilon}$'s as

$$\sum_{j=1}^{p} \epsilon_{ij}^2 = (\underset{\sim}{X}_i - \underset{\sim}{\mu} - \Lambda \underset{\sim}{f}_i)'(\underset{\sim}{X}_i - \underset{\sim}{\mu} - \Lambda \underset{\sim}{f}_i).$$

This is similar to obtaining $\hat{\underset{\sim}{\beta}}$ from $(\underset{\sim}{y} - X\underset{\sim}{\beta})'(\underset{\sim}{y} - X\underset{\sim}{\beta})$ in the usual multiple linear regression model $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ and we get $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}$.

So we get the ordinary least squares estimates of $\underset{\sim}{f}_i$ as

$$\hat{\underset{\sim}{f}}_i = (\Lambda'\Lambda)^{-1}\Lambda'(\underset{\sim}{X}_i - \underset{\sim}{\mu}).$$

In practice, we substitute $\hat{\Lambda}$ and $\bar{\underset{\sim}{x}}$ in place of Λ and $\underset{\sim}{\mu}$, respectively which are obtained from the sampled data and obtain

$$\hat{\underset{\sim}{f}}_i = (\hat{\Lambda}'\hat{\Lambda})^{-1}\hat{\Lambda}'(\underset{\sim}{X}_i - \bar{\underset{\sim}{x}}).$$

Next, using the principal component method with the unrotated factor loadings, this gives

$$\underset{\sim}{\hat{f}}_i = \left(\frac{1}{\sqrt{\hat{\lambda}_1}} \underset{\sim}{\hat{e}}_1'(\underset{\sim}{X}_i - \underset{\sim}{\bar{x}}), \frac{1}{\sqrt{\hat{\lambda}_2}} \underset{\sim}{\hat{e}}_2'(\underset{\sim}{X}_i - \underset{\sim}{\bar{x}}), \ldots, \frac{1}{\sqrt{\hat{\lambda}_m}} \underset{\sim}{\hat{e}}_m'(\underset{\sim}{X}_i - \underset{\sim}{\bar{x}}) \right)'$$

where $\underset{\sim}{e}_1, \underset{\sim}{e}_2, \ldots, \underset{\sim}{e}_m$ are the first m characteristic vectors corresponding to the characteristic roots $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m$ obtained from the sample data.

We observe that $\underset{\sim}{\hat{f}}_i$'s are approximately the same as those obtained in the principal component method with the unrotated factor loadings. Here

$$\frac{1}{n} \sum_{j=1}^{n} \underset{\sim}{\hat{f}}_j = 0 \quad \text{(sample mean)}$$

$$\frac{1}{n-1} \sum_{j=1}^{n} \underset{\sim}{\hat{f}}_j \underset{\sim}{\hat{f}}_j' = 1 \quad \text{(sample variance.)}$$

2. Weighted Least Squares Estimation

The weighted least squares method gives more weight to variables in the estimation that have low specific variances. The factor model fits the data best for variables with low specific variances. Variables with low specific variances should provide more information about the true values of the specific factors.

Considering the model

$$\underset{\sim}{X}_i = \underset{\sim}{\mu} + \Lambda \underset{\sim}{f}_i + \underset{\sim}{\epsilon}_i,$$

the weighted least squares method provides $\underset{\sim}{\hat{f}}_i$ that are obtained by minimizing

$$\sum_{j=1}^{p} \frac{\epsilon_{ij}^2}{\Psi_j} = (\underset{\sim}{X}_i - \underset{\sim}{\mu} - \Lambda \underset{\sim}{f}_i)' \Psi^{-1} (\underset{\sim}{X}_i - \underset{\sim}{\mu} - \Lambda \underset{\sim}{f}_i).$$

This is again similar to obtaining the generalized or weighted least squares estimator $\underset{\sim}{\hat{\beta}}$ in the usual multiple linear regression model $\underset{\sim}{y} = X \underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ with $Cov(\underset{\sim}{\epsilon}) = \Psi$ and we get $\underset{\sim}{\hat{\beta}} = (X' \Psi^{-1} X)^{-1} X' \Psi^{-1} \underset{\sim}{y}$.

So we get the generalized least squares estimator of $\underset{\sim}{f}_i$ as

$$\underset{\sim}{\hat{f}}_i = (\Lambda' \Psi^{-1} \Lambda)^{-1} \Lambda' \Psi^{-1} (\underset{\sim}{X}_i - \underset{\sim}{\mu}).$$

In practice, we substitute $\hat{\Lambda}$ and $\bar{\underset{\sim}{x}}$ in place of Λ and $\underset{\sim}{\mu}$, respectively and obtain

$$\hat{\underset{\sim}{f}}_i = (\hat{\Lambda}'\hat{\Psi}^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Psi}^{-1}(\underset{\sim}{X}_i - \bar{\underset{\sim}{x}}).$$

3. Regression Method of Estimation

This method is used for maximum likelihood estimates of factor loadings considering the joint distribution of $\underset{\sim}{X}_i$ and $\underset{\sim}{f}_i$ as

$$\begin{pmatrix} \underset{\sim}{X}_i \\ \underset{\sim}{f}_i \end{pmatrix} \sim N\left[\begin{pmatrix} \underset{\sim}{\mu} \\ \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \Lambda\Lambda' + \Psi & \Lambda \\ \Lambda' & I \end{pmatrix} \right].$$

Next obtain the conditional expectation of the common factor score $\underset{\sim}{f}_i$ given $\underset{\sim}{X}_i$ as

$$E(\underset{\sim}{f}_i | \underset{\sim}{X}_i) = \Lambda'(\Lambda\Lambda' + \Psi)^{-1}(\underset{\sim}{X}_i - \underset{\sim}{\mu}).$$

In practice, we replace Λ and $\underset{\sim}{\mu}$ by $\hat{\Lambda}$ and $\bar{\underset{\sim}{x}}$ and obtain

$$\hat{\underset{\sim}{f}}_i = \hat{\Lambda}'(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})^{-1}(\underset{\sim}{X}_i - \bar{\underset{\sim}{x}}).$$

Estimation for Fixed Models

Let $\underset{\sim}{x}_\alpha = (x_{1\alpha}, x_{2\alpha}, \ldots, x_{p\alpha})'$ be an observation on $\underset{\sim}{X}_\alpha$, $\alpha = 1, 2, \ldots, N$, given by

$$\underset{\sim}{X}_\alpha = \underset{\sim}{\mu} + \Lambda \underset{\sim}{f}_\alpha + \underset{\sim}{\epsilon}_\alpha$$

with $\underset{\sim}{f}_\alpha$ being nonstochastic vector satisfying $\sum_{\alpha=1}^{N} \underset{\sim}{f}_\alpha = 0$.

The likelihood function is

$$L = \frac{1}{[(2\pi)^p \prod_{i=1}^{p} \psi_{ii}]^{\frac{N}{2}}} \prod_{i=1}^{p} \exp\left[-\frac{1}{2} \sum_{\alpha=1}^{N} \frac{(x_{i\alpha} - \mu_i - \sum_{j=1}^{m} \lambda_{ij} \underset{\sim}{f}_{j\alpha})^2}{\psi_{ii}} \right].$$

The likelihood function does not have a maximum. To show this fact, let $\mu_i = 0$, $\lambda_{11} = 1$, $\lambda_{1j} = 0$ $(j \neq 1)$, $f_{1\alpha} = x_{1\alpha}$. Then $x_{1\alpha} - \mu_1 - \sum_{j=1}^{m} \lambda_{1j} f_{j\alpha} = 0$ and ψ_{11} does not appear in the exponent but appears only in the constant. As $\psi_{11} \to 0$, $L \to \infty$. Thus, the likelihood function does not have a maximum, and the maximum-likelihood estimators do not exist.

Under such a condition, one possibility is to use the maximum likelihood method appropriate for random factors. The sample covariance matrix under normality has a noncentral Wishart distribution, depending on Ψ, $\Lambda\Phi\Lambda'$, and $N-1$.