Applied Multivariate and Longitudinal Data Analysis

Factor Analysis

Ana-Maria Staicu
SAS Hall 5220; 919-515-0644; astaicu@ncsu.edu

Factor analysis is another data dimension reduction tool. Factor analysis can be viewed as an extension of PCA. Specifically, for vector **centered** variate $\mathbf{X}$ of dimension p recall that PCA identifies a set of (much) fewer mutually uncorrelated variables, $Z_1, \ldots Z_m$ such that $\mathbf{X}$ can be approximated parsimoniously by

$$\mathbf{X} \approx \mathbf{a}_1\sqrt{\lambda_1}Z_1 + \ldots + \mathbf{a}_m\sqrt{\lambda_m}Z_m,$$

wherer $\mathbf{a}_1, \ldots, \mathbf{a}_m$ are mutually orthogonal unit one vectors in R^p and Z_ℓ's are zero mean, unit-variance and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Factor analysis introduces a statistical model that uses a parsimonious representation of the inner structure plus error. In factor analysis, the terms analogous to Z_ℓ's are referred to as *factors*. The statistical model will be something like

$\mathbf{X} = \boldsymbol{l}_1 Z_1 + \ldots + \boldsymbol{l}_m Z_m + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$ is error and independent of factors Z_ℓ's.

PCA focuses on the total variation and tries to explain this using only few main features; factor analysis focuses on the covariance or correlation of multiple variables and looks at finding the common factors that account for the largest covariance. Moreover, different from PCA, the approximation used in factor analysis is based on a statistical model.
Intuitively, factor analysis relies on the assumption that the variables that compose $\mathbf{X}$ can be grouped according to their correlations using the reasoning:

1) variables in the same group are highly correlated among them;

2) variables in different groups have smaller correlations;

factor analysis assumes that for each group of variables there is a single underlying structure or a factor that is responsible for the observed high correlation.

Applications of the factor analysis:

- identification of underlying factors:
    1. cluster variables into homogeneous groups
    2. create new variables (e.g. factors)
    3. allows one to gain insight into categories;
- screening of variables
    1. by identifying groups it allows us to select few variables to represent the variables from larger set
    2. useful in regression in handling collinearity

Example: Recall the stock-price data. Weekly stock return data. 103 weekly rates of return on 5 stocks listed on the NY stock exchange (JPMorgan, Citibank, WellsFargo, Shell, Exxon) are recorded for 103 successive weeks. We define

$$\textit{weekly return} = \frac{\text{current closing price - previous week closing price}}{\text{previous week closing price}}$$

adjusted for stock splits and dividend. Rates of returns across stocks are expected to be correlated. Can we see what factors/components could possibly be driving the stock-prices?

One Factor Model (Spearman, 1904)

The One Factor Model:

$$\begin{aligned} X_{i1} - \mu_1 &= l_1 F_i + \epsilon_{i1}, \\ X_{i2} - \mu_2 &= l_2 F_i + \epsilon_{i2}, \\ \vdots \quad &\quad \vdots \\ X_{ip} - \mu_p &= l_p F_i + \epsilon_{ip}, \end{aligned}$$

Key concepts:

- F_i is the latent (unobservable) variable
- $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ is the p-dimensional outcome
- $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})^T$ is the vector of measurement error for X's
- $\boldsymbol{l} = (l_1, \ldots, l_p)^T$ is the vector of *loadings*; l_j is the loading for X_j.

Model assumptions:

- $F_i \sim (0, 1)$; this means that F has zero-mean and unit-variance
- F_i and ϵ_{ij}'s are uncorrelated; $\mathrm{cov}(F_i, \epsilon_{ij}) = 0$ for all j
- $\epsilon_{ij} \sim (0, \sigma_j^2)$ for all j. Errors are uncorrelated: $\mathrm{cov}(\epsilon_{ij}, \epsilon_{ik}) = 0$

This the X_{ij}'s are only related to each other through the factor F_i. Conditional on F_i the variables are independent of one another (this property is sometimes referred to by *conditional independence*).

In class: Calculate $\mathrm{var}(X_{ij})$.

Consider X_{ij} is standardized to have unit variance. Then l_j^2 gives the percentage variance in X_{ij} that is explained by the latent factor F_i, when . The percentage of variance of X_{ij} explained by the factor/s, in this case l_j^2, is called in the literature *communality*.

In class: Calculate $\mathrm{cov}(X_{ij}, X_{ik})$. If X_{ij}'s are standardized then this is the same as $\mathrm{corr}(X_{ij}, X_{ik})$.

Note that the covariance between X_{ij} and X_{ik} for $k \neq j$ is solely determined by F_i!

Factor analysis is equally applicable to the original $\mathbf{X}_i$'s as well as to the standardized $\mathbf{X}_i$'s. In fact one could recover the factor/loadings for the original $\mathbf{X}_i$ from the ones corresponding to the standardized $\mathbf{X}_i$'s. The majority of factor analysis use correlation matrices; we will use the correlation matrix (i.e X_{ij}*'s have been standardized to have unit variance*) for the rest of this chapter.

The Orthogonal Factor Model

Next we discuss extensions of these ideas to the case where there are multiple factors. The model is called *orthogonal factor model*, where orthogonal is used with the sense of independent.

The Orthogonal Factor Model:

$$
\begin{aligned}
X_{i1} - \mu_1 &= l_{1,1}F_{i1} + l_{1,2}F_{i2} + \cdots + l_{1,m}F_{im} + \epsilon_{i1}, \\
X_{i2} - \mu_2 &= l_{2,1}F_{i1} + l_{2,2}F_{i2} + \cdots + l_{2,m}F_{im} + \epsilon_{i2}, \\
&\vdots \quad \vdots \\
X_{ip} - \mu_p &= l_{p,1}F_{i1} + l_{p,2}F_{i2} + \cdots + l_{p,m}F_{im} + \epsilon_{ip},
\end{aligned}
$$

where:

- μ_j is the $mean$ of the j-th variable
- $F_{i1}, F_{i2}, \ldots, F_{im}$ are the *common factors* (latent variables); $F_{ij} \sim (0, 1)$ and $cov(F_{ij}, F_{ik}) = 0$ for $j \neq k$.
- $\boldsymbol{l_k} = (l_{1,k}, \ldots, l_{p,k})^T$ is the vector of *loadings* for factor F_{ik}; $l_{j,k}$ is the *loading* of variable j, X_{ij}, on factor k, F_{ik};
- ϵ_{ij} is a measurement error, affecting only X_{ij}; sometimes is called *specific factor*; $cov(\epsilon_{ij}, F_{ik}) = 0$.

The loadings $l_{j,k}$ represent the degree to which (standardized) variable X_{ij} correlates to factor F_{ik}; thus $l_{j,k}$'s range between -1 and 1. Inspection of factor loadings provide an idea of the extent to which each variable contributes to the meaning of each factor.

In class exercise: Let $\mathbf{\Sigma}$ be the $p \times p$ covariance matrix of X's defined by $(\mathbf{\Sigma})_{jk} = \mathrm{cov}(X_{ij}, X_{ik})$. Determine $\mathbf{\Sigma}$ in terms of the loadings $l_{j,k}$'s and variances of the specific factors σ_j^2. Discuss what are the model parameters and what type of quantities they are (random/fixed).

Further terminology:

- The *communality* of X_{ij} (denoted by h_j^2) is the proportion of the variance of X_{ij} that is explained by the m common factors

$$h_j^2 = l_{j,1}^2 + l_{j,2}^2 + \cdots + l_{j,m}^2;$$

  Intuitively variables with high communality are 'informative'.

- $\mathrm{var}(X_{ij}) = h_j^2 + \sigma_j^2$. As X_{ij} have been standardized to have unit variance, it follows

$$\sigma_j^2 = 1 - h_j^2;$$

  this is the *part of variance that is unique* to X_{ij}.

Orthogonal factor model in matrix form:

$$\underset{p\times 1}{\mathbf{X}_i} - \underset{p\times 1}{\boldsymbol{\mu}} = \underset{p\times m}{\mathbf{L}}\ \underset{m\times 1}{\mathbf{F}_i} + \underset{p\times 1}{\boldsymbol{\epsilon}_i}\ .$$

In terms of the observable variables $\mathbf{X}$, the model assumptions mean that

$$\begin{aligned} \mathsf{E}(\mathbf{X}_i) &= \boldsymbol{\mu}, \\ \mathsf{cov}(\mathbf{X}_i) &= \mathbf{\Sigma} = \underset{p\times m}{\mathbf{L}}\ \underset{m\times p}{\mathbf{L}^T} + \underset{p\times p}{\mathbf{\Psi}}; \end{aligned}$$

where $\mathsf{cov}(\boldsymbol{\epsilon}) = \mathbf{\Psi} = \mathsf{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2\right)$.

As $\mathbf{X}$ is standardized, we have $\mathbf{\Sigma} = \mathbf{R}$ (the correlation matrix).

The observable $\mathbf{X}$ and the latent (unobservable) $\mathbf{F}$ are related by

$$\mathsf{Cov}(\mathbf{X}_i, \mathbf{F}_i) = \mathbf{L}.$$

Note that if $\mathbf{T}$ is $m \times m$ orthogonal matrix ($\mathbf{T}^T\mathbf{T} = \mathbf{T}\mathbf{T}^T = I_m$), then $(\mathbf{LT})(\mathbf{LT})^T = \mathbf{L}\mathbf{L}^T$, so the loadings $\mathbf{LT}$ generate the same $\mathbf{\Sigma}$ as $\mathbf{L}$! Conclusion: the loadings are not uniquely defined.

Remark:

- Model parameters: $\mathbf{L}$ (fixed) and $\mathbf{F}_i$ (random).
- First step: estimate $\mathbf{L}$ (factor loadings). Recall that $\mathbf{L}$ is not unique, lets focus on a rotation of $\mathbf{L}$ that improves interpretation.
- Second step: estimate $\mathbf{F}_i$ (factor scores) given $\mathbf{L}$.

Estimation of factor loadings and scores

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample from a p-multivariate distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and covariance matrix $\boldsymbol{\Sigma} = (\Sigma)_{jk}$. Denote by

$$\begin{aligned} Z_{i1} &= \frac{X_{i1} - \mu_1}{\sqrt{\Sigma_{11}}} \\ Z_{i2} &= \frac{X_{i2} - \mu_2}{\sqrt{\Sigma_{22}}} \\ \cdots &\quad \cdots \\ Z_{ip} &= \frac{X_{ip} - \mu_p}{\sqrt{\Sigma_{pp}}} \end{aligned}$$

the standardized data and by $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$ the standardized $\mathbf{X}_i$. Notice that $\mathbf{Z}_i$'s have mean 0_p and covariance equal to $\boldsymbol{\Gamma}$.

We assume the orthogonal factor model for the standardized data $\mathbf{Z}$'s; that is

$$\underset{p\times 1}{\mathbf{Z}_i} = \underset{p\times m}{\mathbf{L}}\ \underset{m\times 1}{\mathbf{F}_i} + \underset{p\times 1}{\boldsymbol{\epsilon}_i} \qquad i = 1, 2, \ldots, n; \tag{1}$$

where $\mathbf{F}_i \sim (0_p, \mathbf{I}_p)$ and $\boldsymbol{\epsilon}_i \sim (0_p, \boldsymbol{\Psi})$ for diagonal matrix $\boldsymbol{\Psi}$, and the factors $\mathbf{F}_i$ are independent of errors $\boldsymbol{\epsilon}_i$. The loadings $\mathbf{L}$ and the latent scores $\mathbf{F}_i$ are unknown. In the remaining of this section we discuss estimation of both these quantities.

For observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ we estimate the mean and the covariance by the sample mean $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^T$ and sample covariance $\mathbf{S} = (S_{jk})_{jk}$. If the off-diagonal elements of $\mathbf{S}$ are very small, relative to the diagonal elements of $\mathbf{S}$, then the specific factors play a big role in how the variables vary/co-vary and thus a factor analysis would not be useful. (Essentially this means that the varaibles are nearly uncorrelated.) In the following we assume that $\mathbf{S}$ deviates considerably from a diagonal matrix. Thus the number of uncorrelated features is less than the number of variables; hence performing a factor analysis makes sense.

Form the standardized data using the sample mean and sample covariance, $z_{ij} = (x_{ij} - \bar{x}_j)/\sqrt{S_{jj}}$. Pretend $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^T$ arises from model (1). Denote by $\mathbf{R}$ the sample covariance of $\mathbf{z}_i$'s; $\mathbf{R}$ is the sample correlation of the original data $\mathbf{x}_i$'s. We will estimate the loadings and scores based on the standardized data $\mathbf{z}_i$'s.

Model estimation consists of two steps:

1) estimate the loadings $\mathbf{L}$;

2) estimate the scores $\mathbf{F}_i$ given the estimated loadings.

Step 1: Estimation of the factor loadings

Three approaches are common for the estimation of the loadings matrix $\mathbf{L}$: i) principal component analysis (PCA)-based approach; ii) a modified version of PCA; and iii) maximum likelihood estimation.

(i) **Principal components solution**

Consider the spectral decomposition of $\mathbf{R}$,

$$\mathbf{R} = \widehat{\lambda}_1\widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^{\mathsf{T}} + \cdots + \widehat{\lambda}_m\widehat{\mathbf{a}}_m\widehat{\mathbf{a}}_m^{\mathsf{T}} + \widehat{\lambda}_{m+1}\widehat{\mathbf{a}}_{m+1}\widehat{\mathbf{a}}_{m+1}^{\mathsf{T}} + \cdots + \widehat{\lambda}_p\widehat{\mathbf{a}}_p\widehat{\mathbf{a}}_p^{\mathsf{T}};$$

in matrix form this is $\mathbf{R} = \widehat{\mathbf{C}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{C}}^{\mathsf{T}} = \left(\widehat{\mathbf{C}}\widehat{\mathbf{\Lambda}}^{1/2}\right)\left(\widehat{\mathbf{C}}\widehat{\mathbf{\Lambda}}^{1/2}\right)^{\mathsf{T}}$.

Choose m such that $\widehat{\lambda}_1 + \widehat{\lambda}_2 + \cdots + \widehat{\lambda}_m$ is much much larger that $\widehat{\lambda}_{m+1} + \cdots + \widehat{\lambda}_p$. Then the first m terms give the best rank-m approximation to $\mathbf{R}$. The m rank approximation matrix is $\mathbf{R}^{(m)} = \widehat{\lambda}_1\widehat{\mathbf{a}}_1\widehat{\mathbf{a}}_1^{\mathsf{T}} + \cdots + \widehat{\lambda}_m\widehat{\mathbf{a}}_m\widehat{\mathbf{a}}_m^{\mathsf{T}}$. Here

- Estimate $\mathbf{L}$ by $\widehat{\mathbf{L}} = \mathbf{L}^{(m)}$, where $\mathbf{L}^{(m)}$ is the $p \times m$ matrix with the columns $\sqrt{\widehat{\lambda}_1}\widehat{\mathbf{a}}_1, \ldots, \sqrt{\widehat{\lambda}_m}\widehat{\mathbf{a}}_m$. $\mathbf{R}^{(m)} = \mathbf{L}^{(m)}\mathbf{L}^{(m)^T}$

- Estimate $\mathbf{\Psi}$ by $\widehat{\mathbf{\Psi}} = \text{diag}\left(\mathbf{R} - \mathbf{L}^{(m)}\mathbf{L}^{(m)^T}\right)$

  Note: The remainder term $\mathbf{R} - \mathbf{L}^{(m)}\mathbf{L}^{(m)^T}$ is non-negative definite, so its diagonal entries are non-negative. Hence we can get a closer approximation $\mathbf{R} \approx \mathbf{L}^{(m)}\mathbf{L}^{(m)^T} + \widehat{\mathbf{\Psi}}$.

(ii) **Principal factor solution**

This approach estimates the loadings iteratively.

1 Step 1: Denote by ψ_{jj}^* an initial value for the jth diagonal element of $\mathbf{\Psi}$. Then we obtain initial value for communalities by $h_j^{*2} = 1 - \psi_{jj}^*$.

2 Step (k): Update the "working" correlation matrix using current off-diagonal elements of $R_{\text{k-1}}$ and communalities

$$\mathbf{R}_{\text{k}} = \begin{bmatrix} h_1^{*2} & r_{1,2} & \dots & r_{1,p} \\ r_{2,1} & h_2^{*2} & \dots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p,1} & r_{p,2} & \ddots & h_p^{*2} \end{bmatrix}.$$

Use the spectral decomposition of $\mathbf{R}_{\text{r}}$ to find its best rank-m approximation

$$\mathbf{R}_{\text{k}} \approx \mathbf{L}^* \mathbf{L}_{\text{k}}^{*T}.$$

3 Step $(k+1)$: updated value of communalities is

$$h_j^{(k)*2} = \sum_{k=1}^{m} l_{j,k}^{*2}.$$

Updated value for $\mathbf{\Psi}$: diagonal matrix with diagonal elements: $\psi_j^{(k)*} = 1 - h_j^{(k)*2}$; equivalently:

$$\mathbf{\Psi}_k^* = \mathbf{I} - \text{diag}\left(\mathbf{L}_{\text{k}}^* \mathbf{L}_{\text{k}}^{*T}\right).$$

The method implies iteration until convergence; some approaches use a single step iteration.

There are several choices for selecting initial values for the unique variances ψ_{jj}^*; a common one is to use the mean squared error of the regression of X_{ij} onto the rest of the variables, without j'th one, $\{X_{ij'} : j' \neq j\}$.

Nevertheless this approach is far less popular than the principal components solution.

(iii) **Maximum likelihood method**

Maximum likelihood implies knowledge of the generating model. Assume that $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The log-likelihood function for the original data is:

$$\begin{aligned} \ell(\boldsymbol{\mu}, \Sigma) &= -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \quad (2) \\ &= -\frac{n}{2}\log|2\pi\Sigma| - \frac{n}{2}Tr(\Sigma^{-1}S) - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})^T\Sigma^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \quad (3) \end{aligned}$$

It follows that the induced log-likelihood for the standardized data $\mathbf{Z}_i$ with mean zero and covariance $\boldsymbol{\Gamma} = \mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi}$ is

$$\ell(\mathbf{L}, \boldsymbol{\Psi}) = -\frac{n}{2}\log|2\pi(\mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi})| - \frac{n}{2}Tr\{(\mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi})^{-1})S\}.$$

The model is not identified because if $\mathbf{L}$ is a solution, then any rotation of $\mathbf{L}$ is also a solution. To make it identifiable, we impose an additional constraint (uniqueness condition):

$$\mathbf{L}^T\boldsymbol{\Psi}^{-1}\mathbf{L} = \text{diagonal matrix.}$$

The likelihood function is optimized numerically:

$$\widehat{\mathbf{L}}, \widehat{\boldsymbol{\Psi}} = \arg max_{\mathbf{L},\boldsymbol{\Psi}}\ell(\mathbf{L}, \boldsymbol{\Psi})$$

under the constraint specified above. There is no closed form equation for $\hat{\mathbf{L}}$.

- The maximum likelihood estimates (MLE) of communalities are:

$$\widehat{h}_j^2 = \widehat{\ell}_{j1}^2 + \widehat{\ell}_{j2}^2 + \ldots + \widehat{\ell}_{jm}^2 \qquad j = 1, \ldots, p.$$

To asses the importance of the factors one can calculate the proportion of sample variance explaiend by each common factor. The proportion of total sample variance due to the kth common factor is

$$\frac{\widehat{\ell}_{1,k}^2 + \widehat{\ell}_{2,k}^2 + \ldots + \widehat{\ell}_{p,k}^2}{p};$$

where we use p as the denominator because p is the trace of the correlation matrix $\mathbf{R}$.

The differences between the maximum likelihood estimates and the"principal factors" approach can be substantial. If the data appear to be normally distributed (as shown by the usual tests), then the additional efficiency of maximum likelihood estimation is highly worthwhile.

How many factors to choose?

Intuitively the number of factors is the number of uncorrelated constructs that are measured by $\mathbf{Z}_i$'s. Keep in mind that this is a dimension reduction if this number is less than the dimension of $\mathbf{Z}_i$. How to choose the number of factors ?

- Using the principal components. Select the number of factors equal to the reduced rank approximation of the covariance of $\mathbf{Z}_i$'s. Use the scree plot and the percentage of variance explained criterion in this regard.

- Large sample test for the number of common factors, if MLE is used to estimate the factor loadings.

  Using the normality assumption, a large sample test for the number of common factors m has been developed. Specifically, consider the hypothesis testing

  $$H_0 : m = m_0;\ H_A : m > m_0;$$

  Use the likelihood ratio test (LRT) statistic: $LRT = -2 \times$ log likelihood ratio. The null distribution of the LRT is approximately

  $$LRT \sim \chi^2_{\left\{(p-m_0)^2-p-m_0\right\}/2}.$$

  Note that Degrees of freedom > 0 if and only if $m_0 < \frac{1}{2}\left(2p+1-\sqrt{8p+1}\right)$.

Factor rotation

In PCA, the first factor describes most of variability. After choosing the number of factors to retain, we want to spread variability more evenly among the factors. To do this we "rotate" the factors;

- redefine factors such that loadings on various factors tend to be very high (-1 or 1) or very low (0)
- intuitively, it makes sharper distinctions in the meanings of the factors.

  " Ideally, we would like to see a pattern of loadings such that each variable loads highly on a single factor and has small to moderate loadings on the remaining factors." (Johnson & Wichern, page 504):

[That is, ideally, each *row* of $\mathbf{L}$ should have a single large entry.]

Recall from the corresponding equation

$$\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$$

that $\mathbf{L}$ and $\mathbf{LT}$ give the same $\mathbf{\Gamma}$ for any orthogonal matrix $\mathbf{T}$. We can choose $\mathbf{T}$ to make the *rotated* loadings $\mathbf{L}^* = \mathbf{LT}$ more readily interpreted.
Note: the rotation changes neither $\mathbf{R}$ nor $\mathbf{\Psi}$, and hence the communalities are also unchanged. Furthermore the rotation DOES NOT improve the fit. It only improves *interpretability* !

Varimax criterion searches for a rotation (i.e. linear combination) of the original factors such that the variance of the loadings is maximized. Intuitively, if $q_{j,k}$ is the loading of the jth variable on the kth factor then $\mathcal{V} = \sum\sum(q_{j,k}^2 - \bar{q}_{j,\cdot}^2)^2$, where $\bar{q}_{j,\cdot}^2$ is the mean of the squared loadings. [For computational stability, each of the loadings matrix is generally scaled to length one prior to the optimization procedure.]

Making this variance large tends to produce two clusters of scaled loadings, one of small values and one of large values. Each *column* of the rotated loading matrix tends to contain:

- a group of large loadings, which identify the variables associated with the factor;
- the remaining loadings are small.

Example. Let $\mathbf{L}$ be as below. Then the rotated loadings $\mathbf{L}^*$ are

$$\mathbf{L} = \begin{bmatrix} 0.5 & 0.5 \\ 0.8 & 0.8 \\ -0.7 & 0.7 \\ -0.5 & -0.5 \end{bmatrix}; \qquad \mathbf{L}^* = \begin{bmatrix} 0 & 0.6 \\ 0 & 0.9 \\ -0.9 & 0 \\ 0 & -0.9 \end{bmatrix}$$

Other orthogonal rotations are: QUARTIMAX (minimizes the number of factors used to explain each variable) and EQUIMAX which is a compromise between VARIMAX and QUARTIMAX.

Step 2: Estimation of the factor scores

Let $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{\Psi}}$ be the estimated factor loadings matrix and the estimated specific factor variance. The estimation of the factor scores relies on the assumption that these terms are known, and so the model used to estimate $\mathbf{F}_i$ is

$$\mathbf{Z}_i = \widehat{\mathbf{L}}\ \mathbf{F}_i + \boldsymbol{\epsilon}_i, \qquad \mathbf{F}_i \sim (0_p, \mathbf{I}_p), \quad \boldsymbol{\epsilon}_i \sim (0_p, \widehat{\mathbf{\Psi}});$$

where $\mathbf{Z}_i$ is the 'observed data'.

There are two commonly used approaches to estimate the factor scores $\mathbf{F}_i$: i) one approach treats $\mathbf{F}_i$ as fixed parameters; ii) the second one treats $\mathbf{F}_i$ as random quantities.

(i) **Least squares estimation** (weighted least squares)

One method to estimate $\mathbf{F}_i$ is based on minimizing the weighted sum of squared errors:

$$WSS(\mathbf{F}_i) = \left(\mathbf{Z}_i - \widehat{\mathbf{L}}\mathbf{F}_i\right)^T\ \widehat{\mathbf{\Psi}}^{-1}\left(\mathbf{Z}_i - \widehat{\mathbf{L}}\mathbf{F}_i\right).$$

Specifically $\widehat{\mathbf{F}}_i = \arg min_{\mathbf{F}_i} WSS(\mathbf{F}_i)$. The minimizer has the following analytical solution:

$$\widehat{\mathbf{F}}_i = \left(\widehat{\mathbf{L}}^T\widehat{\mathbf{\Psi}}^{-1}\widehat{\mathbf{L}}\right)^{-1}\widehat{\mathbf{L}}^T\widehat{\mathbf{\Psi}}^{-1}\mathbf{Z}_i.$$

Remarks

- When MLE is used to estimate the factor loadings matrix then the quantity in brackets can be simplified due to the constraint used by MLE which is that $\widehat{\mathbf{L}}^T\widehat{\mathbf{\Psi}}^{-1}\widehat{\mathbf{L}}$ has to be diagonal;
- When PCA is used to estimate the loadings matrix, it is customary to use the LS method instead - that is $\widehat{\mathbf{F}}_i = \arg min_{\mathbf{F}_i}\left(\mathbf{Z}_i - \widehat{\mathbf{L}}\mathbf{F}_i\right)^T\ \left(\mathbf{Z}_i - \widehat{\mathbf{L}}\mathbf{F}_i\right)$.

(ii) **Conditional expectation** (regression method)

Another approach is to treat $\mathbf{F}_i$ as random and estimate them (or more appropriately said *predict* them) using conditional expectation. That is

$$\widehat{\mathbf{F}}_i = E[\mathbf{F}_i|\mathbf{Z}_i],$$

using the above model. If in addition we assume that all the random terms, $\mathbf{F}_i$ and $\boldsymbol{\epsilon}_i$ are normally distributed, then $\widehat{\mathbf{F}}_i$ are the best linear unbiased estimators. Furthermore the conditional expectation yields an analytical solution:

$$\widehat{\mathbf{F}}_i = \widehat{\mathbf{L}}^T \left(\widehat{\mathbf{L}}\widehat{\mathbf{L}}^T + \widehat{\boldsymbol{\Psi}}\right)^{-1} \mathbf{Z}_i$$

Remark

- One can establish the relationship between the score estimators using (weighted)LS $\widehat{\mathbf{F}}_i^{LS}$ and the score estimators using conditional expectation $\widehat{\mathbf{F}}_i^{CE}$

$$\widehat{\mathbf{F}}_i^{\mathsf{LS}} = \left\{\mathbf{I}_p + \left(\widehat{\mathbf{L}}^T \widehat{\boldsymbol{\Psi}}^{-1} \widehat{\mathbf{L}}\right)^{-1}\right\} \widehat{\mathbf{F}}_i^{\mathsf{CE}}.$$

- Studies to compare the two methods (LS and CE) have been carried and the two perform relatively similarly. None is recommended as uniformly superior. In general it is recommended to consider both approaches and compare the results.

Beyond factor analysis

Latent variable models are a broad class of models, which postulate some relationship between the statistical properties of observable variables and latent variables.

Category	Latent variable	Observed variable
Factor analysis	Continuous	Continuous
Latent profile analysis	Categorical	Continuous
Latent trait analysis	Continuous	Categorical
Latent class analysis	Categorical	Categorical