# Statistics 135

# Chapter 9

# Factor Analysis

Chris Drake

*Department of Statistics*

*University of California, Davis*

# Purpose of Factor Analysis

Describe the covariance relationship among many variables in terms of a few underlying but unobservable quantities. Identify variables measuring common traits.

Examples:

1 Wechsler Adult Intelligence Scale Subtest Scores. The data set consists of 11 variables measuring aspects of intellectual performance; we can think of underlying, unobserved factors that "explain" performance on these tasks; the idea is to identify underlying factors such as verbal ability similar to principal components.

2 Fowl bone dimensions where measurements were taken on the skull (2), the wing (2) and the leg (2) for a total of 6 measurements; question: are there underlying commonalities that explain the observed measurements?

3 Example 9.3 on page 491: consumer preference data; there are 5 variables: taste, good buy for money, flavor, suitable for snack, provides

lots of energy; question: do these variables form groups

# The Model

Let $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$ be a p-variate normal vector and $\tilde{\mathbf{X}} = \mathbf{X} - \mu$ is the centered version;

Suppose $F_1, ..., F_m$ are uncorrelated variables $m < p$ with $E[\mathbf{F}] = \mathbf{0}$ and $Cov(\mathbf{F}) = E[\mathbf{F}\mathbf{F}'] = \mathbf{I}$.

Consider the model

$$\begin{aligned}
\tilde{X}_1 &= l_{11}F_1 + l_{12}F_2 + ... + l_{1m}F_m + \epsilon_1 \\
\tilde{X}_2 &= l_{21}F_1 + l_{22}F_2 + ... + l_{2m}F_m + \epsilon_2 \\
&: \qquad\qquad\quad : \\
\tilde{X}_p &= l_{p1}F_1 + l_{p2}F_2 + ... + l_{pm}F_m + \epsilon_p
\end{aligned}$$

where $E[\epsilon] = \mathbf{0}$ and $Cov(\epsilon) = E[\epsilon\epsilon'] = \mathbf{\Psi} = diag(\psi_1, ...\psi_p)$ and $Cov(\mathbf{F}, \epsilon) = \mathbf{0}_{p\times m}$

*Definitions:*

1 The coefficients $l_{ij}$ for $i = 1, .., p$ and $j = 1, .., m$ are called the loadings of the $i^{th}$ variable on the $j^{th}$ factor.

2 $F_1, ..., F_m$ are called the common factors.

3 $\epsilon_1, .., \epsilon_p$ are called the errors or *specific factors*.

4 The Orthogonal Factor Model with $m$ Common Factors is given by:

$$\mathbf{X} = \mu + \mathbf{L}\mathbf{F} + \epsilon$$

where $\mathbf{L}$ is the matrix of factor loadings and $\mathbf{F}$ is the vector of common factors.

*Note:*

$$\begin{aligned}(\mathbf{X} - \mu)(\mathbf{X} - \mu)' &= (\mathbf{LF} + \epsilon)(\mathbf{LF} + \epsilon)' \\ &= \mathbf{LF}(\mathbf{LF})' + \epsilon(\mathbf{LF})' + (\mathbf{LF})\epsilon' + \epsilon\epsilon'\end{aligned}$$

and therefore

$$\mathbf{\Sigma} = \mathbf{L}E[\mathbf{FF}'](\mathbf{L})' + E[\epsilon\mathbf{F}']\mathbf{L}' + \mathbf{L}E[\mathbf{F}\epsilon'] + \epsilon\epsilon' = \mathbf{LL}' + \mathbf{\Psi}$$

5 $Cov(\mathbf{X} - \mu) = Cov(\mathbf{X}) = \mathbf{LL}' + \mathbf{\Psi}$ which gives

$$Var(X_i) = l_{i1}^2 + l_{i2}^2 + .. + l_{im}^2 + \psi_i$$

$$Cov(X_i, X_k) = l_{i1}l_{k1} + l_{i2}l_{k2} + .. + l_{im}l_{km}$$

therefore

$$Cov(X_i, F_j) = l_{ij}$$

6 The variance of $X_i$ is the sum of the specific variance $\psi_i$ and the communality

$$h_i = l_{i1}^2 + l_{i2}^2 + .. + l_{im}^2$$

7 Most covariance matrices cannot be factored as $\mathbf{LL}' + \mathbf{\Psi}$ with $m << p$; for an illustration see example (9.2) on page 486. The solution that is obtained leads to values for correlations that are greater than 1.

8 Factor loadings $\mathbf{L}$ are not unique and can be determined only up to an orthogonal matrix $\mathbf{T}$. Recall, for orthogonal matrices $\mathbf{T}$ we have $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$; therefore

$$\mathbf{X} - \mu = \mathbf{L}\mathbf{F} + \epsilon = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{F} + \epsilon$$

for $\mathbf{L}^* = \mathbf{L}\mathbf{T}$ and $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$ and

$$E(\mathbf{F})^* = \mathbf{T}'E(\mathbf{F}) = \mathbf{0}$$

$$Cov(\mathbf{F}^*) = \mathbf{T}'Cov(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}$$

and therefore

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{L}' + \mathbf{\Psi} = (\mathbf{L}^*)(\mathbf{L}^*)' + \mathbf{\Psi}$$

# Samples and Estimation

1. Sample $\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_n$ of p-variate observations; it is assumed that the components of $\mathbf{X}$ are correlated, otherwise a factor analysis is not useful.
2. The sample covariance matrix $\mathbf{S}$ is an estimate of $\mathbf{\Sigma}$.

Given the sample and an assumption of non-zero correlations of the components of the observations, how do we estimate the factor loadings $l_{ij}$ and specific variances $\psi_i$; keep in mind that the factors $F_j$ for $j = 1, .., m$ are not observed, so a multivariate regression with the factors as a design matrix is not an option.

We will consider two methods:

1. The principal component method
2. Maximum likelihood

*Principal component method*

1 $Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + .. + e_{ip}X_p$ is the $i^{th}$ principal component where $\mathbf{e}_i$ is the $i^{th}$ eigenvector of $\boldsymbol{\Sigma}$; in matrix notation

$$\mathbf{Y} = \mathbf{P}'\mathbf{X} \qquad \textit{where} \qquad \boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$$

2 $Cov(Y_i, Y_j) = 0$

3 $\mathbf{PY} = \mathbf{PP}'\mathbf{X}$ and therefore $\mathbf{X} = \mathbf{PY}$

then taking $\mathbf{Y}$ as the common factors gets us almost what we want, except $Cov\mathbf{Y} = \boldsymbol{\Lambda}$. If we choose instead $\mathbf{L} = \mathbf{P}\boldsymbol{\Lambda}^{\frac{1}{2}}$ we get

$$\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi} = \mathbf{P}\boldsymbol{\Lambda}^{\frac{1}{2}}(\mathbf{P}\boldsymbol{\Lambda}^{\frac{1}{2}})' + \mathbf{0}$$

When $\lambda_{m+1}, .., \lambda_p$ are small, we can drop columns $l_{m=1}, .., l_p$ from the matrix $\mathbf{L}$.

Then

$$\mathbf{\Sigma} = [\sqrt{\lambda_1}\ \mathbf{e_1}|...|\sqrt{\lambda_m}\ \mathbf{e_m}] \begin{pmatrix} \sqrt{\lambda_1}\ \mathbf{e_1} \\ \sqrt{\lambda_2}\ \mathbf{e_2} \\ \vdots \\ \sqrt{\lambda_m}\ \mathbf{e_m} \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_m \end{pmatrix}$$

and

$$\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2$$

Note:

1 Principal components requires all components to account for the total variation and the purpose is to summarize the variability in the observed data.

2 Factor analysis accounts for the total variance with only a few common factors (the specific variance accounts for the residual variance not explained by the common factors).

*The principal components estimates*

1. Sample data $\mathbf{x}_1, ..., \mathbf{x}_n$ are p-variate observation vectors with sample covariance matrix $\mathbf{S}$ and correlation matrix $\mathbf{R}$
2. $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ for $i = 1, .., p$ are the eigenvalues and eigenvectors of $\mathbf{S}$
3. For some $m < p$ we obtain the matrix $\tilde{\mathbf{L}}$ of factor loadings

$$\tilde{\mathbf{L}} = [\sqrt{\hat{\lambda}_1}\ \hat{\mathbf{e}}_1 \mid \sqrt{\hat{\lambda}_2}\ \hat{\mathbf{e}}_2 \mid ... \mid \sqrt{\hat{\lambda}_m}\ \hat{\mathbf{e}}_m]$$

and

$$\tilde{\mathbf{\Psi}} = diag(\tilde{\psi}_i) \qquad with \qquad \tilde{\psi} = s_{ii} - \sum_{j=1}^{m} \tilde{l}_{ij}^2$$

with communalities

$$\tilde{h}_i^2 = \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + ... + \tilde{l}_{im}^2$$

4. Note,that $\mathbf{S}$ is not reproduced with the principal components solution. That would require $p$ factors.

*Maximum likelihood*

1. The $\mathbf{F}$ and $\epsilon$ are assumed jointly normally distributed.
2. The observations $\mathbf{X}_j - \mu = \mathbf{L}\mathbf{F}'_j + \epsilon_j$ are then also normally distributed.
3. The likelihood $L(\mu, \mathbf{\Sigma})$ is expressed as a function of $\mathbf{L}$ and $\mathbf{\Psi}$ through $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}$. It is maximized subject to the condition $\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L} = \mathbf{\Delta}$ where $\mathbf{\Delta}$ is a diagonal matrix; the condition is required to make the MLE's well-defined.
4. Note, $m$ the number of factors needs to be chosen prior to obtaining the MLE's.
5. The maximum likelihood estimates of the loadings and specific variances are obtained iteratively.
6. The maximum likelihood estimates of the commonalities are given by

$$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + .... + \hat{l}_{im}^2 \qquad for \qquad i = 1, .., p$$

# Determining the number of factors

1 The test is carried out for the null hypothesis

$$H_0 : \quad \mathbf{\Sigma}_{p\times p} = \mathbf{L}_{p\times m}\mathbf{L}'_{p\times m} + \mathbf{\Psi}_{p\times p}$$

vs the alternative $H_1 : \quad \mathbf{\Sigma}$ is any other symmetric, positive, definite matrix.

2 The test is based on the likelihood ratio statistic

$$-2ln\Lambda = -2ln\left(\frac{\mid \hat{\mathbf{\Sigma}} \mid}{\mid \mathbf{S}_n \mid}\right)^{-n/2} + n[tr(\hat{\mathbf{\Sigma}}^{-\mathbf{1}}\mathbf{S}_n) - p] = nln\left(\frac{\mid \hat{\mathbf{\Sigma}} \mid}{\mid \mathbf{S}_n \mid}\right)$$

where $\hat{\mathbf{\Sigma}} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}}$; $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$ are the maximum likelihood estimates of the factor loadings and the specific factors. Note: $tr(\hat{\mathbf{\Sigma}}^{-\mathbf{1}}\mathbf{S}_n) - p = 0$ if $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$ are the MLE's.

3 The degrees of freedom are

$$\nu - \nu_0 = \frac{1}{2}p(p+1) - [p(m+1) - \frac{1}{2}m(m-1)] = \frac{1}{2}[(p-m)^2 - p - m]$$

4 An improved version of the test was suggested by Bartlett; this test rejects at level $\alpha$ if

$$n - 1 - \frac{2p+4m+5}{6} ln \frac{\mid \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}} \mid}{\mid \mathbf{S}_n \mid} \cdot \chi^2_{[(p-m)^2-p-m]/2}(\alpha)$$

5 The Bartlett approximation requires that $n$ and $n-p$ are large.

6 Since the degrees of freedom must be positive we need

$$m < \frac{1}{2}(2p + 1 - \sqrt{8p+1})$$

# Factor Rotation

Factor loadings are not unique. If $\hat{\mathbf{L}}$ is any estimate of the $p \times m$ matrix of factor loadings, then so is

$$\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T} \qquad \textit{where} \qquad \mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$$

so $\mathbf{T}$ is an orthogonal transformation.

Under such an orthogonal transformation, the estimated covariance matrix remains unchanged, since

$$\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}\mathbf{T}\mathbf{T}'\mathbf{L}' + \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}^*\hat{\mathbf{L}}^{*\prime} + \hat{\mathbf{\Psi}}$$

Furthermore, the residual matrix remains unchanged, since

$$\mathbf{S}_n - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}^*\hat{\mathbf{L}}^{*\prime} - \hat{\mathbf{\Psi}}$$

Therefore, orthogonal rotations do not change the estimate of the covariance matrix.

Factor rotation refers to a rotation of the factor loadings to achieve a simpler structure that makes the factors easier to interpret. Ideally, a rotation leads to a structure where each variable loads highly on one factor and has small or moderate loadings on the other factors.

For 2 factors, this rotation can be done visually by inspecting a plot, where the original factors are plotted along perpendicular axis ($F_1$ is the x-axis, $F_2$ is the y-axis), the factor loadings $\hat{l}_{1i}, \hat{l}_{2i}$ are plotted for $i = 1, .., p$ and the axis rotated at an angle $\phi$ such the rotated factors align more closely with clusters of points.

For more than 2 factors, computational techniques need to be employed. Computer programs such as SAS will do it.

*Varimax criterion*

Suppose $\tilde{l}^*_{ij} = \hat{l}^*_{ij}/\hat{h}_i$ where $\hat{l}^*_{ij}$ is the rotated coefficient. The (normal) *varimax* procedure selects the orthogonal transformation that maximizes

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

After the transformation $\mathbf{T}$ is determined, the loadings $\tilde{l}^*_{ij}$ are multiplied by $\hat{h}_i$, so that the original communalities are preserved.

Note: V corresponds to "spreading out" the squares of the loadings on each factor as much as possible. The goal is to find large and negligible coefficients in each column of $\hat{\mathbf{L}}^*$.

One criterion for rotation is the achieve positive and/or negative loadings for each factor.

Thurston (1945, *Multiple Factor Analysis*), gave the following criteria:

1 Each row of $\hat{\mathbf{L}}^*$ should contain at least one zero. The respective factor $F$ is not associated with the component of $X$ represented by that row.

2 Each column of $\hat{\mathbf{L}}^*$ should contain at least m zeros. The factor $F$ is not associated with m components of $\mathbf{X}$.

3 Every pair of columns of $\hat{\mathbf{L}}^*$ should contain several components of $\mathbf{X}$ whose loadings vanish in one but not the other. The respective factors load on separate $X's$.

4 If the number of factors is 4 or more, every pair of columns of $\hat{\mathbf{L}}^*$ should contain a large number of components of $\mathbf{X}$ with zero loadings in both columns.

5 For every pair of columns of $\hat{\mathbf{L}}^*$ only a small number of components of $\mathbf{X}$ should have nonzero loadings in both columns.

These criteria require that the components of $\mathbf{X}$ fall into mutually exclusive groups with loadings high on a single factor and low to moderate on the other factors.

# Estimating Factor Scores

1 Factor scores are estimates of the unobserved random factors $\mathbf{F}_j$ for $j = 1, .., n$; note: $\mathbf{X}$ is the random variable and $\mathbf{x}_1, ..., \mathbf{x}_n$ is a sample from the population; the possible value $\mathbf{f}_j$ of $\mathbf{F}_j$, $\mathbf{F}_j$ is a subject specific random effect in the same way that $\epsilon_j$ is.

2 Each subject has 2 random effects associated with it: $\mathbf{F}_j$ and $\epsilon_j$;

3 The relationship is given by

$$\mathbf{X}_{p\times 1} - \mu_{p\times 1} = \mathbf{L}_{p\times m}\mathbf{F}_{m\times 1} + \epsilon_{p\times 1}$$

and both $\mathbf{F}$ and $\epsilon$ are unknown.

4 To estimate $\mathbf{f}_j$ , the following assumptions are made, $\hat{l}_{ij}$ and $\hat{\psi}_i$ are the true values.

*The weighted least squares method*

1 Factor scores obtained by weighted least squares from the maximum likelihood estimates

$$\hat{\mathbf{f}}_j = \left(\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}}\right)^{-1}\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\left(\mathbf{x}_j - \hat{\mu}\right)$$

where $\hat{\mathbf{f}}$ is chosen to minimize

$$\epsilon'\boldsymbol{\Psi}^{-1}\epsilon = \left(\mathbf{x} - \mu - \mathbf{Lf}\right)'\boldsymbol{\Psi}^{-1}\left(\mathbf{x} - \mu - \mathbf{Lf}\right)$$

2 If factor loadings are estimated using the principal components approach, an unweighted least squares method is used to give

$$\hat{\mathbf{f}}_j = \left(\tilde{\mathbf{L}}'\tilde{\mathbf{L}}\right)^{-\mathbf{1}}\tilde{\mathbf{L}}'\left(\mathbf{x_j} - \hat{\mu}\right)$$

where $\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\ \hat{\mathbf{e}}_1 \mid \sqrt{\hat{\lambda}_2}\ \hat{\mathbf{e}}_2 \mid \ldots \mid \sqrt{\hat{\lambda}_m}\ \hat{\mathbf{e}}_m\right]$. The estimates are for the covariance matrix; scaled data use $\hat{\mathbf{L}}$ from the correlation matrix.

*The regression method*

The estimated factor scores are given by

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}'\mathbf{S}^{-1}\big(\mathbf{x_j} - \bar{\mathbf{x}}\big) \qquad\qquad j = 1, .., n$$

Here, $\mathbf{L}$ and $\boldsymbol{\Psi}$ are treated as known, $\mathbf{F}$ and $\epsilon$ are jointly normal and the joint distribution of $\mathbf{X} - \mu$ and $\mathbf{F}$ is $N_{p+m}(\mathbf{0}, \boldsymbol{\Sigma}^*)$ with

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} \boldsymbol{\Sigma}_{p\times p} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi} & \mathbf{L}_{p\times m} \\ \mathbf{L}'_{m\times p} & \mathbf{I}_{m\times m} \end{bmatrix}$$

then

$$E\big(\mathbf{F} \mid \mathbf{x}\big) = \mathbf{L}'\boldsymbol{\Sigma}^{-1}\big(\mathbf{x} - \mu\big) = \mathbf{L}'\big(\mathbf{F}\mathbf{F}' + \boldsymbol{\Psi}\big)^{-1}\big(\mathbf{x} - \mu\big)$$

and we substitute the maximum likelihood estimates of $\hat{\mathbf{L}}$, $\hat{\boldsymbol{\Psi}}$ and $\mu$; finally, $\mathbf{S}$ is substituted for $\hat{\mathbf{F}}\hat{\mathbf{F}}' + \hat{\boldsymbol{\Psi}}$

# Strategy for Factor Analysis

1. Perform a principal component factor analysis, ie use principal component approach to obtain factors.
   a. look for suspicious observations using methods for multivariate normal data
   b. try varimax rotation
2. Perform maximum likelihood factor analysis (use a computer program for that), including varimax rotation
3. Compare the solutions
4. Repeat the steps with different $m$.
5. If data set is large split into two data sets and perform factor analysis on each set; compare results
6. Consult with an expert in the field the data was collected for. If they think the results are non-informative, try again.