Chapter 8: Factor Analysis

Jingyuan Liu
Department of Statistics, School of Economics
Wang Yanan Institute for Studies in Economics
Xiamen University

Outline

1. Introduction to Factor Analysis
   - Intuition and Definitions
   - Orthogonal Factor Model and Assumption
   - Distinguishing Factor Analysis and Other Methods
2. Methods of Estimation
   - Principal Component Methods
   - Other Methods
   - Estimating Factor Scores
3. Factor Rotation and Interpretation
   - Intuition and Introduction
   - Orthogonal Rotation and Oblique Rotation
4. Application Example with R Implementation
   - Test Score Example

Outline

1 Introduction to Factor Analysis
- Intuition and Definitions
- Orthogonal Factor Model and Assumption
- Distinguishing Factor Analysis and Other Methods

Intuition Example

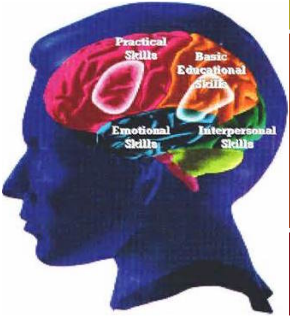
Consider a simple example of children's examination marks in three subjects, Classics (Y_1), French (Y_2) and English (Y_3), from which the following correlation matrix is calculated:

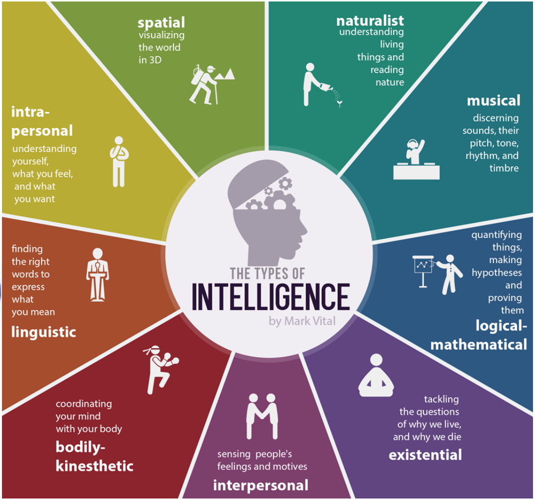
$$\mathbf{R} = \begin{array}{l} \text{Classics} \\ \text{French} \\ \text{English} \end{array} \begin{pmatrix} 1.00 & & \\ 0.83 & 1.00 & \\ 0.78 & 0.67 & 1.00 \end{pmatrix}.$$

The three variables are fairly correlated, thus can possibly be represented by a common "factor", such as " the general linguistic ability", although this factor can neither be observed, nor fully represent the original three variables.

Intuition of Factor Analysis

- Factor analysis is motivated by the following argument: "Variables can be grouped by their correlations." That is, for the original observable p variables $Y_1, \ldots, Y_p$, those variables that are highly correlated would follow a common single underlying constuct, or factor, that is responsible for the observed correlations.
- However, this "common factor" is probably unobservable, which is called **latent variables**. This is typical in psychology and other disciplines of behavioral science, such as "intelligence" and "social class".


Practical Skills
Basic Educational Skills
Emotional Skills
Interpersonal Skills


THE TYPES OF
INTELLIGENCE
by Mark Vital
spatial
visualizing the world in 3D
naturalist
understanding living things and reading nature
musical
discerning sounds, their pitch, tone, rhythm, and timbre
quantifying things, making hypotheses and proving them
logical-mathematical
tackling the questions of why we live, and why we die
existential
sensing people's feelings and motives
interpersonal
coordinating your mind with your body
bodily-kinesthetic
finding the right words to express what you mean
linguistic
intra-personal
understanding yourself, what you feel, and what you want

Introduction and Definitions

- The essential purpose of factor analysis is to describe, if possible, the covariance or correlation structure among many observable variables $Y_1, \ldots, Y_p$ in terms of a few underlying, but unobservable, random quantities called **factors**, denoted by $F_1, \ldots, F_m$, where $m \leq p$, hopefully $m < p$.
- In the intuition example, since Y_1, Y_2, Y_3 are all highly correlated, probably only one factor is needed, such as "linguistic capability".

Two Types of Factor Analysis

- **Exploratory factor analysis (EFA)** is an exploration of multivariate data to identify possible latent structure. In EFA, the structure of latent factors is not determined before the analysis.
- **Confirmatory factor analysis (CFA)** allows specifying the number of latent factors and the specific nature of the latent structure in the data, and then test the hypotheses.
- EFA is useful for simplifying complex multivariate data and formulating hypotheses for CFA.
- The details for CFA can be found in Chapter 14 of the book. Only EFA is covered in class. The factor analysis refers to EFA here by default.

Outline

1 Introduction to Factor Analysis
- Intuition and Definitions
- Orthogonal Factor Model and Assumption
- Distinguishing Factor Analysis and Other Methods

Orthogonal Factor Model

Suppose the observable random vector $\mathbf{y} = (Y_1, \dots, Y_p)'$ has mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The factor model postulates that $\mathbf{y}$ linearly depends upon a few unobservable random variables $\mathbf{f} = (F_1, \dots, F_m)'$, called **common factors**, and p additional sources of variation $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_p)'$, called **specific factors** or **errors**. Typically $m < p$. The model is

$$
\begin{aligned}
Y_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\
Y_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\
&\ldots \\
Y_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p
\end{aligned}
$$

The coefficient l_{jk} is called the **loading** of the jth variable on the kth factor.

Matrix Form and Assumption

In terms of matrix notation, the above model becomes

$$\mathbf{y}_{p\times 1} - \boldsymbol{\mu}_{p\times 1} = \mathbf{L}_{p\times m}\mathbf{f}_{m\times 1} + \boldsymbol{\varepsilon}_{p\times 1},$$

where $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ are assumed to satisfy

$$E(\mathbf{f}) = \mathbf{0}_{m\times 1}, \qquad COV(\mathbf{f}) = E(\mathbf{f}\mathbf{f}') = \mathbf{I}_{m\times m}$$

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}_{p\times 1}, \qquad COV(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Psi}_{p\times p} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

and that $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ are uncorrelated: $COV(\boldsymbol{\varepsilon}, \mathbf{f}) = \mathbf{O}_{p\times m}$.

Orthogonal Factor Model: Implication

- $\boldsymbol{\Sigma} = COV(\mathbf{y}) = \mathbf{LL}' + \boldsymbol{\Psi}$
  - $\sigma_{jj} = var(Y_j) = h_j^2 + \psi_j$, where $h_j^2 = l_{j1}^2 + \ldots + l_{jm}^2$. h_j^2, called the jth **communality**, refers to the portion of variance contributed by the m common factor. ψ_j is called the **uniqueness**, or **specific variance**, denoting the portion of variance due to the specific factor.
  - $\sigma_{jj'} = cov(Y_j, Y_j') = l_{j1}l_{j'1} + \ldots + l_{jm}l_{j'm}$
- $COV(\mathbf{y}, \mathbf{f}) = \mathbf{L}$
  - $cov(Y_j, F_k) = l_{jk}$, thus the coefficient/loading l_{jk} measures the association between the corresponding variable and common factor.

Nonuniqueness of Factor Loadings

From the model statement, the factor loadings $\mathbf{L}$ can be determined only up to any orthogonal matrix $\mathbf{T}$:

- The basic model is equivalent to $\mathbf{y} - \boldsymbol{\mu} = \mathbf{L}^*\mathbf{f}^* + \varepsilon$, where $\mathbf{L}^* = \mathbf{LT}$, $\mathbf{f}^* = \mathbf{T}'\mathbf{f}$, $\mathbf{T}$ is any orthogonal matrix s.t. $\mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$.
- The new loading matrix $\mathbf{L}^*$ reproduces $\boldsymbol{\Sigma}$ in the same fashion: $\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi} = \mathbf{L}^*\mathbf{L}^{*\prime} + \boldsymbol{\Psi}$.
- The new factors in $\mathbf{f}^*$ satisfy the assumptions as before: $E(\mathbf{f}^*) = \mathbf{0}$, $COV(\mathbf{f}^*) = \mathbf{I}$, and $COV(\mathbf{f}^*, \varepsilon) = \mathbf{O}$.
- This ambiguity provides the rationale for the “factor rotation”, which rotates the loadings to new ones with easier or more meaningful interpretations.

Outline

1 Introduction to Factor Analysis
- Intuition and Definitions
- Orthogonal Factor Model and Assumption
- Distinguishing Factor Analysis and Other Methods

Factor Analysis vs. PCA

- Factor analysis aims to explain covariances/correlations of the observed variables by means of common factors. However, principal component analysis (PCA) is primarily concerned with the variance of the observed variables.
- In factor analysis, the original variables are expressed as linear combinations of the factors. However, principal components are linear combinations of the original variables.
- The coefficients for principal components are unique, but the loadings of the common factors are only unique up to an orthogonal matrix.

Factor Analysis vs. Canonical Correlations

- Factor analysis deals with the covariance or correlation structure within one set of variables, while canonical correlation analysis focuses on the association between two sets of variables.
- In factor analysis, the original variables are expressed as linear combinations of the factors, while in the canonical correlation analysis, the canonical variates are linear combinations of the original variables.

Factor Analysis vs. Multivariate (Multiple) Regression

- In multivariate regression, the "factors"/predictors $\mathbf{x}$'s are observed, and often regarded as fixed. While in factor analysis, the factors F's are unobserved random variables.
- Factor analysis requires the specific factors $\varepsilon_1, \ldots, \varepsilon_p$ to be uncorrelated with different variances. In multivariate regression, the random error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$ has its own covariance structure $\mathbf{\Sigma}$ and typically the elements are correlated.

Outline

2 Methods of Estimation
- Principal Component Methods
- Other Methods
- Estimating Factor Scores

Principal Component Method

- Recall that the spectral decomposition of $\boldsymbol{\Sigma}$ provides

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' = \sum_{j=1}^{p} \lambda_j \mathbf{e}_j \mathbf{e}_j',$$

where $\boldsymbol{\Lambda} = (\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_p}\mathbf{e}_p)$, and $(\lambda_j, \mathbf{e}_j)$'s are eigen pairs of $\boldsymbol{\Sigma}$.

- Even if this factor representation of $\boldsymbol{\Sigma}$ is exact, it is not particularly useful: It employs as many common factors as there are variables and does not allow for any variation in the specific factor ε.
- The factor analysis tries to answer the question "Does the factor model with a small number $(m < p)$ of factors adequately represents the data?"

- Since $\mathbf{\Sigma} = \sum_{j=1}^{p} \lambda_j \mathbf{e}_j \mathbf{e}_j'$, so when the last $p - m$ eigenvalues are small, we can neglect the contribution of $\sum_{j=m+1}^{p} \lambda_j \mathbf{e}_j \mathbf{e}_j'$ to $\mathbf{\Sigma}$, resulting the following factoring:

$$\mathbf{\Sigma} \approx \mathbf{L}\mathbf{L}' + \mathbf{\Psi},$$

  where $\mathbf{L} = (\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_m}\mathbf{e}_m)$, $\mathbf{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_p)$, and $\psi_j = \sigma_{jj} - \sum_{k=1}^{m} l_{jk}^2$, $j = 1, \ldots, p$.
- In practice, when this representation is applied to the sample covariane matrix $\mathbf{S}$ (or the sample correlation matrix $\mathbf{R}$), it is known as the **principal component solution**.

Theorem (Principal component solution)

The principal component factor analysis of $\mathbf{S}$ *is specified in terms of its eigen pairs* $(\hat{\lambda}_j, \hat{\mathbf{e}}_j)$, $j = 1, \ldots, p$, *where* $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$. *Let* $m < p$ *be the number of common factors. Then the estimated factor loading matrix is given by*

$$\hat{\mathbf{L}} = (\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1, \ldots, \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m).$$

The estimated specific variances are the diagonal elements of $\mathbf{S} - \hat{\mathbf{L}}\hat{\mathbf{L}}'$, *i.e.* $\hat{\boldsymbol{\Psi}} = diag(\hat{\psi}_1, \ldots, \hat{\psi}_p)$ *with* $\hat{\psi}_j = s_{jj} - \sum_{k=1}^{m} \hat{l}_{jk}^2$. *The communalities are estimated by* $\hat{h}_j^2 = \hat{l}_{j1}^2 + \ldots + \hat{l}_{jm}^2$.

Remarks:

- The diagonal elements of $\mathbf{S}$ are equal to that of $\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}$. However, the off-diagonal elements of $\mathbf{S}$ are not usually reproduced by $\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}$.
- The loadings on the kth factor are proportional to the coefficients in the kth principal component.
- The contribution of the kth factor to the total sample variance $s_{11} + s_{22} + \ldots + s_{pp} = tr(\mathbf{S})$ is $\sum_{j=1}^{p} \hat{l}_{jk}^2 = \hat{\lambda}_k$, and hence the corresponding proportion of contribution by the kth factor to the total variance is

$$\frac{\hat{\lambda}_k}{tr(\mathbf{S})}, \;\; k = 1, \ldots, m.$$

Principal Component Method based on Standardized Variables

- As in the PCA, when the units of the variables are not commensurate, it is usually desirable to work with the standardized variables whose sample covariance matrix is the sample correlation $\mathbf{R}$ of the original observations.
- In this case, the principal component solution is applied by replacing $\mathbf{S}$ with $\mathbf{R}$.
- In practice, $\mathbf{R}$ is used more often than $\mathbf{S}$ and is default in most software packages.

Determining m

A natural question for the factor analysis is how many factors should we retain to capture the original variables.

- Researchers in psychology, sociology, or other behavioral sciences might specify the number of factors based on the theory in those fields or previous work.
- If no priori knowledge is available, we can choose m based on the estimated eigenvalues in much the same manner as with the principal components.

Determining m: Rationale

- Consider the residual matrix $\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}})$. The diagonal elements are 0, and if other elements are also small, we may subjectively consider the m-factor model to be appropriate.
- Analytically, we have

$$\text{Sum of squared entries of } (\mathbf{S} - (\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}})) \leq \sum_{k=m+1}^{p} \hat{\lambda}_k^2.$$

- Consequently, we may evaluate the contributions of the neglected eigenvalues, as in the PCA. E.g. percentage cutoff, average cutoff, and scree graph.

Principal Component Method: Example

Example: A 12-year-old girl made five ratings of perceptions on a 9-point scale for each of seven of her acquaintances. The ratings were based on the five adjectives "kind", "intelligent", "happy", "likeable", and "just". Her ratings are given in the following table.

Perception Data: Ratings on Five Adjectives for Seven People

People	Kind	Intelligent	Happy	Likeable	Just
FSM1[a]	1	5	5	1	1
SISTER	8	9	7	9	8
FSM2	9	8	9	9	8
FATHER	9	9	9	9	9
TEACHER	1	9	1	1	9
MSM[b]	9	7	7	9	9
FSM3	9	7	9	9	7

[a]Female schoolmate 1.
[b]Male schoolmate.

The correlation matrix for the five variables (adjectives) is as follows, with the larger values bolded:

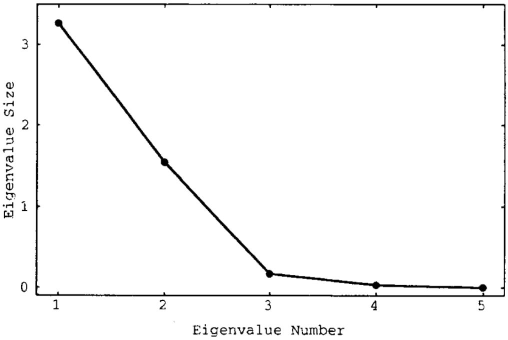
$$\mathbf{R} = \begin{pmatrix} 1.000 & .296 & \mathbf{.881} & \mathbf{.995} & .545 \\ .296 & 1.000 & -.022 & .326 & \mathbf{.837} \\ \mathbf{.881} & -.022 & 1.000 & \mathbf{.867} & .130 \\ \mathbf{.995} & .326 & \mathbf{.867} & 1.000 & .544 \\ .545 & \mathbf{.837} & .130 & .544 & 1.000 \end{pmatrix}.$$

The boldface values indicate two groups of variables: $\{1, 3, 4\}$ and $\{2, 5\}$. We would therefore expect that the correlations among the variables can be explained fairly well by two factors.

If we use $m = 2$ factors, the principal component method based on the correlation matrix $\mathbf{R}$ yields the following results:

	Loadings			
Variables	$\hat{l}_{j1}$	$\hat{l}_{j2}$	Communalities, $\hat{h}_j^2$	Specific Variances, $\hat{\psi}_j$
Kind	.969	−.231	.993	.007
Intelligent	.519	.807	.921	.079
Happy	.785	−.587	.960	.040
Likeable	.971	−.210	.987	.013
Just	.704	.667	.940	.060
Variance accounted for	3.263	1.538	4.802	
Proportion of total variance	.653	.308	.960	
Cumulative proportion	.653	.960	.960	

The the first two factors account for 96% of the total sample variance. And the scree graph also shows $m = 2$ is sufficient.



Scree graph for the perception data.

To see how well the two-factor model reproduces the correlation matrix, we examine

$$\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}} = \begin{pmatrix} .969 & -.231 \\ .519 & .807 \\ .785 & -.587 \\ .971 & -.210 \\ .704 & .667 \end{pmatrix} \begin{pmatrix} .969 & .519 & .785 & .971 & .704 \\ -.231 & .807 & -.587 & -.210 & .667 \end{pmatrix}$$

$$+ \begin{pmatrix} .007 & 0 & 0 & 0 & 0 \\ 0 & .079 & 0 & 0 & 0 \\ 0 & 0 & .040 & 0 & 0 \\ 0 & 0 & 0 & .013 & 0 \\ 0 & 0 & 0 & 0 & .060 \end{pmatrix}$$

$$= \begin{pmatrix} 1.000 & .317 & .896 & .990 & .528 \\ .317 & 1.000 & -.066 & .335 & .904 \\ .896 & -.066 & 1.000 & .885 & .161 \\ .990 & .335 & .885 & 1.000 & .543 \\ .528 & .904 & .161 & .543 & 1.000 \end{pmatrix},$$

which is very close to the original **R**.

Outline

2 Methods of Estimation

- Principal Component Methods
- Other Methods
- Estimating Factor Scores

Principal Factor Method

- In the principal component approach, we neglected $\mathbf{\Psi}$ and factored $\mathbf{S}$ or $\mathbf{R}$.
- The **principal factor/axis method**, however, uses an initial estimate $\hat{\mathbf{\Psi}}^{(0)}$ and factors $\mathbf{S} - \hat{\mathbf{\Psi}}^{(0)}$ or $\mathbf{R} - \hat{\mathbf{\Psi}}^{(0)}$, by the same routine as the principal component approach.
- A popular choice of $\hat{\mathbf{\Psi}}^{(0)}$ is $1/\text{diag}(\text{diag}(\mathbf{S}^{-1}))$ when $\mathbf{S}$ is used, and is $1/\text{diag}(\text{diag}(\mathbf{R}^{-1}))$ when $\mathbf{R}$ is used.
- We can also iteratively update the estimate $\hat{\mathbf{\Psi}}^{(0)}$ and the decomposition of $\mathbf{S} - \hat{\mathbf{\Psi}}^{(0)}$ or $\mathbf{R} - \hat{\mathbf{\Psi}}^{(0)}$ until convergence.
- This method can cause difficulty in interpretation with negative eigenvalues, as $\mathbf{S} - \hat{\mathbf{\Psi}}^{(0)}$ and $\mathbf{R} - \hat{\mathbf{\Psi}}^{(0)}$ may not be positive definite.

Principal Factor Method: Example

The following table compares the loadings obtained by the principal component method and the principal factor method, and they are faily similar. The communalities in the table are for the principal factor method.

	Principal Component Loadings		Principal Factor Loadings		
Variables	F1	F2	F1	F2	Communalities
Kind	.969	−.231	.981	−.210	.995
Intelligent	.519	.807	.487	.774	.837
Happy	.785	−.587	.771	−.544	.881
Likeable	.971	−.210	.982	−.188	.995
Just	.704	.667	.667	.648	.837
Variance accounted for	3.263	1.538	3.202	1.395	
Proportion of total variance	.653	.308	.704	.307	
Cumulative proportion	.653	.960	.704	1.01	

Maximum Likelihood Method

When the samples $\mathbf{f}_1, \ldots, \mathbf{f}_n$ and $\varepsilon_1, \ldots, \varepsilon_n$ are jointly normal, and hence $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are i.i.d. $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$, the likelihood function of $\boldsymbol{\Sigma}$ can be presented as follows with $\hat{\boldsymbol{\mu}}_{MLE} = \bar{\mathbf{y}}$ already plugged in:

$$L(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left[-\frac{1}{2} tr \left\{\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'\right)\right\}\right]$$

Then the maximum likelihood estimators of $\mathbf{L}$ and $\boldsymbol{\Psi}$ can be computed iteratively subject to the uniqueness condition that $\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L}$ is diagonal.

Outline

2 Methods of Estimation

- Principal Component Methods
- Other Methods
- Estimating Factor Scores

Introduction to Factor Scores

- In many applications of factor analysis, the researcher wishes only to ascertain whether a factor model fits the data, and focuses on the parameters of the factor model, rather than the random factor values.
- In other applications, the researcher wishes to obtain factor scores, $\hat{\mathbf{f}}_i = (\hat{F}_{i1}, \ldots, \hat{F}_{im})$, $i = 1, \ldots, n$, i.e. the estimated factor values for each observation.
- There are two potential uses for such scores:
  1. The behavior of the observations in terms of the factors may be of interest;
  2. We may wish to use the factor scores as input of other analyses, such as classification.

Weighted Least Squares Method

- Recall that the factor model is

$$\mathbf{y} - \boldsymbol{\mu} = \mathbf{L}\mathbf{f} + \boldsymbol{\varepsilon}.$$

- Similar to the estimation of regression parameters, we might obtain the estimates of $\mathbf{f}$ by weighted least squares method since the variances of $\boldsymbol{\varepsilon}$ need not be equal. Then we obtain the estimated factor score

$$(\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{\mu}).$$

- In deriving the factor scores, we usually treat the estimated parameters $\hat{\mathbf{L}}$, $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$ as true values. Then the estimated factor scores for the ith case is

$$\hat{\mathbf{f}}_i = (\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}}).$$

Regression Method

- Another method, the regression method, is based on the normality assumption. Assuming multivariate normality of $\mathbf{f}$ and ε, we have

$$\begin{pmatrix} \mathbf{y}-\boldsymbol{\mu} \\ \mathbf{f} \end{pmatrix} \sim N_{p+m}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}, & \mathbf{L} \\ \mathbf{L}', & \mathbf{I} \end{bmatrix}\right).$$

- The conditional distribution of $\mathbf{f}$ given $\mathbf{y}$ is normal with

$$E(\mathbf{f}|\mathbf{y}) = \mathbf{L}'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}).$$

- Then we can obtain the following estimated factor scores for the ith case:

$$\hat{\mathbf{f}}_i = \hat{\mathbf{L}}'(\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}})^{-1}(\mathbf{y}_i - \bar{\mathbf{y}}), \text{ or } \hat{\mathbf{f}}_i = \hat{\mathbf{L}}'\mathbf{S}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}}).$$

Outline

3 Factor Rotation and Interpretation

- Intuition and Introduction
- Orthogonal Rotation and Oblique Rotation

Intuition to Factor Rotation

- As in PCA, we often need provide rational interpretation to the calculated factors. However, the original factors may not have sensible meanings directly.
- Recall that the factor loading matrix $\hat{\mathbf{L}}$ is unique only up to multiplication by an orthogonal matrix $\mathbf{T}$. The rotated loading matrix $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}$ are equivalent to $\hat{\mathbf{L}}$ in the sense that $\mathbf{\Sigma} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}^*\hat{\mathbf{L}}^{*\prime} + \hat{\mathbf{\Psi}}$.
- Thus we could find some $\mathbf{T}$ to rotate the loadings (or say, factors), if possible, for a simpler structure and a clearer interpretation.

Introduction to Factor Rotation

- Ideally, we should 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.
- Geometrically, the loadings in the jth row of $\hat{\mathbf{L}}$ constitute the coordinates of a point Y_j in the factor/loading space. The multiplication by an orthogonal matrix corresponds to a rigid rotation/reflection of the coordinate axes.
- To this end, the goal of rotation is to place the axes close to as many points as possible. This may be achieved if there are clusters of points (groupings of Y_j's).

Two Types of Rotation

We here consider two basic types of rotation:

1. **Orthogonal rotation:** The original perpendicular axes are rotated rigidly and remain perpendicular. Angles and distances are preserved, communalities are unchanged, and the basic configuration of the points remains the same. Only the reference axes differ.
2. **Oblique rotation:** The axes are not required to remain perpendicular and are thus free to pass closer to clusters of points.

Outline

3 Factor Rotation and Interpretation
- Intuition and Introduction
- Orthogonal Rotation and Oblique Rotation

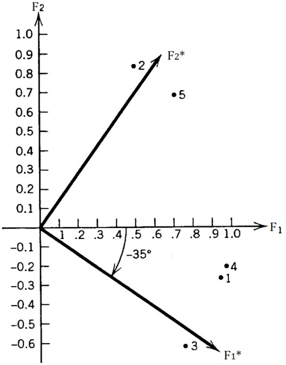
Orthogonal Rotation: Graphical Approach

- If $m = 2$, we can determine the rotation graphically based on a visual inspection of a plot of factor loadings.
- The rows of $\hat{\mathbf{L}}$ are pairs of loadings $(\hat{l}_{j1}, \hat{l}_{j2})$. A plot of these pairs yields p points, corresponding to $Y_1, \ldots, Y_p$.
- We choose an angle ϕ through which the axes can be rotated to move them closer to groupings of points. The new rotated loadings $(\hat{l}^*_{j1}, \hat{l}^*_{j2})$ can be measured on the graph as coordinates, or calculated from $\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T}$ using

$$\mathbf{T} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$$

Graphical Approach: Example

Back to the perception data example. The five pairs of loadings corresponding to the five variables are plotted below. An orthogonal rotation through $\phi = -35^\circ$ would bring the axes (factors) closer to the two clusters of points (variables).



We obtain the following rotated loadings with $\phi = -35^\circ$:

$$\hat{\mathbf{L}}^* = \hat{\mathbf{L}}\mathbf{T} = \begin{pmatrix} .969 & -.231 \\ .519 & .807 \\ .785 & -.587 \\ .971 & -.210 \\ .704 & .667 \end{pmatrix} \begin{pmatrix} .819 & .574 \\ -.574 & .819 \end{pmatrix}$$

$$= \begin{pmatrix} \boxed{.927} & .367 \\ -.037 & \boxed{.959} \\ \boxed{.980} & -.031 \\ \boxed{.916} & .385 \\ .194 & \boxed{.950} \end{pmatrix}.$$

Then the groupings according to the correlations are clear, and are consistent with those given by the pattern of **R**.

The comparison between the rotated loadings and the original are reported as follows.

Variables	Principal Component Loadings		Graphically Rotated Loadings		Communalities,
	F_1	F_2	F_1^*	F_2^*	$\hat{h}_i^2$
Kind	.969	−.231	**.927**	.367	.993
Intelligent	.519	.807	−.037	**.959**	.921
Happy	.785	−.587	**.980**	−.031	.960
Likeable	.971	−.210	**.916**	.385	.987
Just	.704	.667	.194	**.950**	.940
Variance accounted for	3.263	1.538	2.696	2.106	4.802
Proportion of total variance	.653	.308	.539	.421	.960
Cumulative proportion	.653	.960	.539	.960	.960

- The interpretation of the rotated loadings is clear.
  - The first factor is associated with variables Y_1, Y_3, and Y_4: kind, happy, and likeable, which might be described as a person's perceived humanity or amiability.
  - The second factor consists of Y_2 and Y_5: intelligent and just, which involve more logical or rational practices.
- If the angle between the rotated axes is allowed to be less than 90° (an oblique rotation), the lower axis representing F_1^* could come closer to the points corresponding to variables 1 and 4.

Orthogonal Rotation: Varimax Approach

- The graphical approach to rotation is generally limited to $m = 2$. For $m > 2$, various analytical methods have been proposed. The most popular is the varimax technique, which seeks rotated loadings to maximize the "variance" of the squared loadings in each column of $\hat{\mathbf{L}}^*$.
- We hope to find groups of large and negligible coefficients in any column of the rotated loadings matrix.
- The varimax rotation is available in virtually all factor analysis software programs.

Varimax Approach: Example

The varimax rotated factor loadings of the perception data example are very close to the graphical rotation, and the interpretations are exactly the same.

	Principal Component Loadings		Graphically Rotated Loadings		Varimax Rotated Loadings		Communalities
Variables	F_1	F_2	F_1	F_2	F_1	F_2	$\hat{h}_i^2$
Kind	.969	−.231	**.927**	.367	**.951**	.298	.993
Intelligent	.519	.807	−.037	**.959**	.033	**.959**	.921
Happy	.785	−.587	**.980**	−.031	**.975**	−.103	.960
Likeable	.971	−.210	**.916**	.385	**.941**	.317	.987
Just	.704	.667	.194	**.950**	.263	**.933**	.940
Variance accounted for	3.263	1.538	2.696	2.106	2.811	1.991	4.802
Proportion of total variance	.653	.308	.539	.421	.562	.398	.960
Cumulative proportion	.653	.960	.539	.960	.562	.960	.960

Oblique Rotation/Transformation

- Instead of the orthogonal matrix $\mathbf{T}$ used in the orthogonal rotation of factors, an oblique rotation/transformation uses a general nonsingular transformation matrix $\mathbf{Q}$ to obtain $\mathbf{f}^* = \mathbf{Q}'\mathbf{f}$, then

$$COV(\mathbf{f}^*) = \mathbf{Q}'\mathbf{I}\mathbf{Q} = \mathbf{Q}'\mathbf{Q} \neq \mathbf{I}.$$

  Thus the new factors are correlated.
- Because distances and angles are not preserved, the communalities for $\mathbf{f}^*$ are different from those for $\mathbf{f}$.
- When the axes are not required to be perpendicular, they can more easily pass through the major clusters of points in the loading space, assuming there are such clusters.

Oblique Rotation: Example

Example: Recall the measurements of adult sons from 25 families in Chapter 7:

First Son		Second Son	
Head Length y_1	Head Breadth y_2	Head Length x_1	Head Breadth x_2
191	155	179	145
195	149	201	152
181	148	185	149
183	153	188	149
176	144	171	142
208	157	192	152
189	150	190	149
197	159	189	152
188	152	197	159
192	150	187	151
179	158	186	148
183	147	174	147
174	150	185	152
190	159	195	157

The correlation matrix is

$$\mathbf{R} = \begin{pmatrix} 1.000 & .735 & .711 & .704 \\ .735 & 1.000 & .693 & .709 \\ .711 & .693 & 1.000 & .839 \\ .704 & .709 & .839 & 1.000 \end{pmatrix}.$$

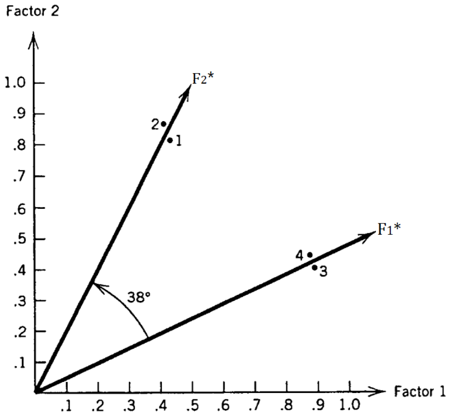
The groupings are not discernible directly from **R**, since all the variables are highly correlated.

The varimax rotated loadings and the oblique rotated loadings (by Harris-Kaiser method in SAS) for two factors obtained by the principal component method are given below.

	Varimax Loadings		Oblique Pattern matrix	
Variable	F_1	F_2	F_1	F_2
1	.42	**.82**	.03	**.90**
2	.40	**.85**	−.03	**.96**
3	**.87**	.41	**.97**	−.01
4	**.86**	.43	**.95**	.01

The oblique loadings give a much cleaner and simpler structure than the varimax loadings, but the interpretation is essentially the same.

The oblique loadings are ploted below.



Remarks:

- The angle of the oblique axes is 38°, and the correlation between the two rotated factors is 0.79 obtained from the off-diagonal element of $\mathbf{Q}'\mathbf{Q}$, or from $\cos 38^\circ = 0.79$.
- Since the angle between axes is less than 45°, a single factor would be adequate. The suggestion to let $m = 1$ is also supported by the three criteria to choose m, as well as by the correlation matrix $\mathbf{R}$.

Outline

4 Application Example with R Implementation
- Test Score Example

R Implementation: Test Score Example

Example: Recall the test score example from Chapter 7. The test scores of 52 students on 6 subjects: Y_1 =math, Y_2 =physics, Y_3 =chemistry, Y_4 =Chinese, Y_5 =history, Y_6 =English, were recorded. Part of data are as follows.

1	Y1	Y2	Y3	Y4	Y5	Y6
2	65	61	72	84	81	79
3	77	77	76	64	70	55
4	67	63	49	65	67	57
5	78	84	75	62	71	64
6	66	71	67	52	65	57
7	83	100	79	41	67	50
8	86	94	97	51	63	55
9	67	84	53	58	66	56
10	69	56	67	75	94	80
11	77	90	80	68	66	60
12	84	67	75	60	70	63
13	62	67	83	71	85	77
14	91	74	97	62	71	66
15	82	70	83	68	77	85
16	66	61	77	62	73	64
17	90	78	78	59	72	66

From the following correlation matrix **R**, we suspect two groups (Y_1, Y_2, Y_3) and (Y_4, Y_5, Y_6).

```
> test<-read.table("/Users/jingyuan/Documents/Teaching/Multivariate Analysis/R
code/Chap8/test_score.csv", sep=",", header=T)
> (R<-round(cor(test), 3))    # sample correlation matrix
       Y1     Y2     Y3     Y4     Y5     Y6
Y1  1.000  0.647  0.696 -0.561 -0.456 -0.439
Y2  0.647  1.000  0.573 -0.503 -0.351 -0.458
Y3  0.696  0.573  1.000 -0.380 -0.274 -0.244
Y4 -0.561 -0.503 -0.380  1.000  0.813  0.835
Y5 -0.456 -0.351 -0.274  0.813  1.000  0.819
Y6 -0.439 -0.458 -0.244  0.835  0.819  1.000
```

So the factor model to be estimated is

$$Y_j - \mu_j = l_{j1}F_1 + l_{j2}F_2 + \varepsilon_j,\ j = 1, \ldots, 6.$$

Example: Maximum Likelihood Method

```
> ## maximum likelihood method to estimate the loadings
> factanal(test,factors=2,rotation="none")

Call:
factanal(x = test, factors = 2, rotation = "none")

Uniquenesses:
   Y1    Y2    Y3    Y4    Y5    Y6
0.228 0.459 0.333 0.148 0.210 0.150

Loadings:
   Factor1 Factor2
Y1 -0.676   0.562
Y2 -0.599   0.427
Y3 -0.487   0.656
Y4  0.917   0.104
Y5  0.856   0.239
Y6  0.883   0.266

               Factor1 Factor2
SS loadings      3.404   1.068
Proportion Var   0.567   0.178
Cumulative Var   0.567   0.745

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 3.64 on 4 degrees of freedom.
The p-value is 0.457
```

Example: Principal Component Method

```
> library(psych)
> (fac<-principal(test,nfactors=2,rotate="none",covar=F))
Principal Components Analysis
Call: principal(r = test, nfactors = 2, rotate = "none", covar = F)
Standardized loadings (pattern matrix) based upon correlation matrix
     PC1  PC2   h2   u2 com
Y1 -0.79 0.42 0.81 0.19 1.5
Y2 -0.73 0.40 0.70 0.30 1.5
Y3 -0.64 0.63 0.81 0.19 2.0
Y4  0.89 0.31 0.89 0.11 1.2
Y5  0.81 0.47 0.87 0.13 1.6
Y6  0.83 0.46 0.90 0.10 1.6

                       PC1  PC2
SS loadings           3.71 1.26
Proportion Var        0.62 0.21
Cumulative Var        0.62 0.83
Proportion Explained  0.75 0.25
Cumulative Proportion 0.75 1.00

Mean item complexity =  1.6
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.06
 with the empirical chi square  5.96  with prob <  0.2

Fit based upon off diagonal values = 0.99
```

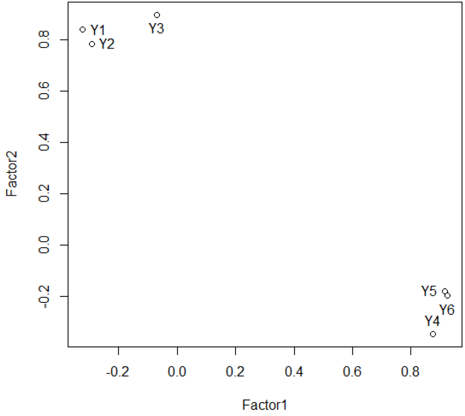
The cumulative contribution of the total variance by two common factors is 74.5% for the maximum likelihood method, yet 83% for the principal component method. Thus we focus on the latter for further discussion.

Example: Factor Rotation with “varimax”

```
> ## principal component method with varimax rotation
> (fac1<-principal(test,nfactors=2,rotate="varimax",covar=F))
Principal Components Analysis
Call: principal(r = test, nfactors = 2, rotate = "varimax", covar = F)
Standardized loadings (pattern matrix) based upon correlation matrix
     RC1   RC2   h2   u2 com
Y1 -0.32  0.84 0.81 0.19 1.3
Y2 -0.29  0.78 0.70 0.30 1.3
Y3 -0.07  0.90 0.81 0.19 1.0
Y4  0.88 -0.35 0.89 0.11 1.3
Y5  0.92 -0.18 0.87 0.13 1.1
Y6  0.93 -0.20 0.90 0.10 1.1

                       RC1  RC2
SS loadings           2.66 2.31
Proportion Var        0.44 0.39
Cumulative Var        0.44 0.83
Proportion Explained  0.54 0.46
Cumulative Proportion 0.54 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.06
 with the empirical chi square  5.96  with prob <  0.2

Fit based upon off diagonal values = 0.99
```

And the corresponding rotated factor plot is as follows.

```
> ## plot the rotated factors
> plot(fac1$loadings[,1], fac1$loadings[,2], xlab="Factor1",ylab="Factor2")
> identify(fac1$loadings[,1], fac1$loadings[,2], labels=c("Y1","Y2","Y3","Y4","Y5","Y6"))
```



Thus the first factor depends strongly on (Y_4, Y_5, Y_6), hence can be referred to as "liberal art factor". The second factor is much more correlated with (Y_1, Y_2, Y_3), which might be called "science factor". Apparently, the meanings of factors are clearer after rotation.

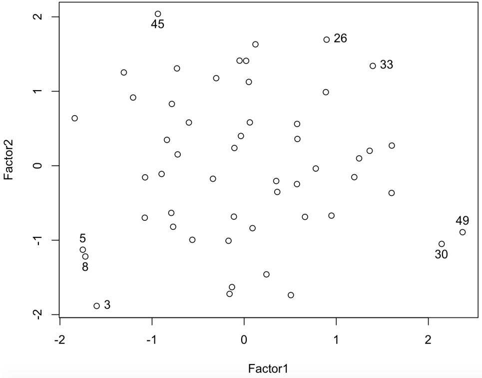
Example: Factor Scores

We could calculate the sample factor scores if necessary: (only show the first 20 students)

```
> fac1$scores[1:20,]
              RC1         RC2
 [1,]  0.66036380 -0.68718464
 [2,] -1.07567973 -0.15571837
 [3,] -1.60122590 -1.88323054
 [4,] -0.72215827  0.15234085
 [5,] -1.75198210 -1.12791008
 [6,] -1.84006051  0.63813576
 [7,] -1.30641265  1.25308371
 [8,] -1.72435369 -1.21908241
 [9,]  0.94800759 -0.66986805
[10,] -0.83830982  0.34763562
[11,] -0.89636993 -0.11048745
[12,]  0.57341472 -0.24629083
[13,] -0.30355745  1.17630966
[14,]  0.57503687  0.56148305
[15,] -0.77109905 -0.81988047
[16,] -0.59997498  0.58064466
[17,]  0.06168574  0.58222335
[18,]  1.24824003  0.09893997
[19,]  1.60089485 -0.36560636
[20,]  0.88689495  0.98857294
```

And the rotated factor scores plot can also be obtained.

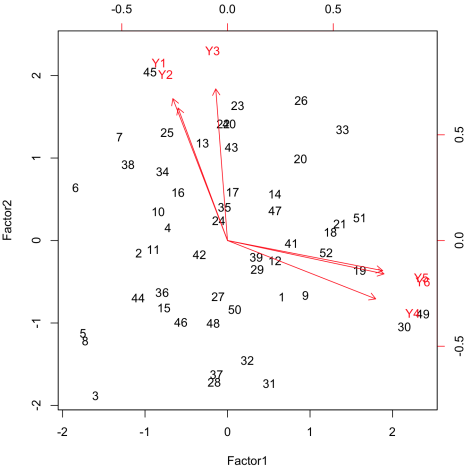
```
> ## plot factor scores in the factor space and identify the typical students
> plot(fac1$scores[,1], fac1$scores[,2], xlab="Factor1",ylab="Factor2")
> identify(fac1$scores[,1], fac1$scores[,2], labels=1:52)
[1]  3  5  8 26 30 33 45 49
```



The typical students were identified in the plot, and are consistent with the PCA ressults in Chapter 7. Specifically, students in the first (third) quadrant perform well (poorly) in both factors - science and liberal arts. While students in the second (fourth) quadrant perform worse (better) in factor 1 - liberal arts than factor 2 - science.

The original variables can be depicted in the factor plot.

```
> biplot(fac1$scores,fac1$loadings,xlab="Factor1",ylab="Factor2")
```



Summary and Take-home Messages

- What is factor analysis for?
- What are the differences between factor analysis and the principal component analysis?
- What are the factor model and the assumptions?
- How to estimate the factor loadings?
- Why do we need to rotate the factors and how to?
- How to interpret the factor analysis results?