# Factor analysis

## Introduction

- In social sciences (e.g., psychology), it is often not possible to measure the variables of interest directly. Examples:
  - Intelligence
  - Social class

  Such variables are called *latent variables* or *common factors*.
- Researchers examine such variables indirectly, by measuring variables that can be measured and that are believed to be indicators of the latent variables of interest. Examples:
  - Examination scores on various tests
  - Occupation, education, home ownership

  Such variables are called *manifest variables* or *observed variables*.
- Goal: study the relationship between the latent variables and the manifest variables



## Factor analysis model

- Multiple linear regression model:

$$\begin{aligned} x_1 &= \lambda_{11} f_1 + \cdots + \lambda_{1k} f_k + u_1 \\ x_2 &= \lambda_{21} f_1 + \cdots + \lambda_{2k} f_k + u_2 \\ \vdots &= \qquad \vdots \\ x_p &= \lambda_{p1} f_1 + \cdots + \lambda_{pk} f_k + u_p \end{aligned}$$

  where
  - $x = (x_1, \ldots, x_p)'$ are the observed variables (random)
  - $f = (f_1, \ldots, f_k)'$ are the common factors (random)
  - $u = (u_1, \ldots, u_p)'$ are called *specific factors* (random)
  - $\lambda_{ij}$ are called *factor loadings* (constants)

## Factor analysis model

- In short: $x = \Lambda f + u$, where $\Lambda$ is the $p \times k$ matrix containing the $\lambda_{ij}$'s.
- Difference with multiple regression: common factors $f_1, \dots, f_k$ are unobserved.
- Assumptions:
  - $E(x) = 0$ (if this is not the case, simply subtract the mean vector)
  - $E(f) = 0$, Cov$(f) = I$
  - $E(u) = 0$, Cov$(u_i, u_j) = 0$ for $i \neq j$
  - Cov$(f, u) = 0$

5 / 35

## Variance of $x_i$

- Notation:
  - Cov$(u) = \Psi =$ diag$(\psi_{11}, \dots, \psi_{kk})$
  - Cov$(x) = \Sigma$
- Then (see board):
  - $\sigma_{ii} =$ Var$(x_i) = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_{ii}$
  - Var$(x_i)$ consists of two parts:
    - $h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2$, called communality of $x_i$, represents variance of $x_i$ that is shared with the other variables via the common factors
    - $\psi_{ii}$, called the specific or unique variance, represents the variance of $x_i$ that is not shared with the other variables

6 / 35

## Covariance matrix of $x$

- Note that (see board):
  - $\sigma_{ij} =$ Cov$(x_i, x_j) = \sum_{\ell=1}^{k} \lambda_{i\ell}\lambda_{j\ell}$
- Hence, the factor models leads to: $\Sigma = \Lambda\Lambda' + \Psi$
- The reverse is also true: If one can decompose $\Sigma$ in this form, then the $k$-factor model holds for $x$

7 / 35

## Non-uniqueness of factor loadings

- Suppose that $k$-factor model holds for $x$: $x = \Lambda f + u$
- Let $G$ be a $k \times k$ orthogonal matrix.
- Then $x = \Lambda G G' f + u$.
- Note that $G' f$ satisfies assumptions that we made about the common factors (see board).
- Hence the $k$-factor model holds with factors $G' f$ and factor loadings $\Lambda G$.
- $\Sigma = (\Lambda G)(G' \Lambda') + \Psi = \Lambda \Lambda' + \Psi$
- Hence, factors $f$ with loadings $\Lambda$, or factors $G' f$ with loadings $\Lambda G$ are equivalent for explaining the covariance matrix of the observed variables.

8 / 35

## Non-uniqueness of factor loadings

- Non-uniqueness can be resolved by imposing an extra condition. For example:
  - $\Lambda' \Psi^{-1} \Lambda$ is diagonal with its elements in decreasing order (constraint 1)
  - $\Lambda' D^{-1} \Lambda$ is diagonal with its elements in decreasing order, where $D = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$ (constraint 2)

9 / 35

## Estimation

- $\Sigma$ is usually estimated by $S$ (or often: correlation matrix is estimated by $R$) .
- Given $S$ (or $R$), we need to find estimates $\hat{\Lambda}$ and $\hat{\Psi}$ that satisfy constraint 1 or 2, so that $S$ (or $R$) $\approx \hat{\Lambda} \hat{\Lambda}' + \hat{\Psi}$.
- Note that typically, the number of parameters in $\hat{\Lambda}$ and $\hat{\Psi}$ is smaller than the number of parameters in $S$. Hence, there is no exact solution in general.
- Two main methods to estimate $\hat{\Lambda}$ and $\hat{\Psi}$:
  - principal factor analysis
  - maximum likelihood estimation (requires normality assumption)
- In practice, we also need to determine the value of $k$, the number of factors.

10 / 35

### Procedure - initialization

- Estimate correlation matrix by $R$
- Make preliminary estimates $\hat{h}_i^2$ of the communalities $h_i^2$, using:
  - The square of the multiple correlation coefficient of the $i$th variable with all the other variables, or
  - The largest correlation coefficient between the $i$th variable and one of the other variables



### Idea

- Given $R$ $(p \times p)$, we want to find $\hat{\Psi}$ $(p \times p)$ and $\hat{\Lambda}$ $(p \times k)$ that satisfy constraint 2, so that $R - \hat{\Psi} \approx \hat{\Lambda}\hat{\Lambda}'$
- We look at $R - \hat{\Psi}$, because we are interested in explaining the (co)variances that are shared through the common factors.
- $R - \hat{\Psi}$ is symmetric. Hence there is a spectral decomposition $R - \hat{\Psi} = GAG' = \sum_{i=1}^{p} a_i g_{(i)} g'_{(i)}$
- If the first $k$ eigenvalues are positive, and the remaining ones are close to zero, then $R - \hat{\Psi} \approx \sum_{i=1}^{k} a_i g_{(i)} g'_{(i)} = \sum_{i=1}^{k} (a_i^{1/2} g_{(i)})(a_i^{1/2} g_{(i)})'$.
- $\hat{\Lambda}\hat{\Lambda}' = \sum_{i=1}^{k} \hat{\lambda}_{(i)} \hat{\lambda}'_{(i)}$. Hence, a natural estimate for $\lambda_{(i)}$ is $\hat{\lambda}_{(i)} = a_i^{1/2} g_{(i)}$.
- In matrix form: $\hat{\Lambda} = G_1 A_1^{1/2}$.



### Procedure

- Determine the spectral decomposition of the *reduced correlation matrix* $R - \hat{\Psi}$, where the ones on the diagonal are replaced by $\hat{h}_i^2 = 1 - \hat{\psi}_{ii}$. Thus, $R - \hat{\Psi} = GAG'$, where $A = \text{diag}(a_1, \ldots, a_p)$ contains the eigenvalues of $R - \hat{\Psi}$, $a_1 \geq \cdots \geq a_p$, and $G$ contains the corresponding orthonormal eigenvectors.
- Estimate $\Lambda$ by $\hat{\Lambda} = G_1 A_1^{1/2}$, where $G_1 = (g_{(1)}, \ldots, g_{(k)})$ and $A_1 = \text{diag}(a_1, \ldots, a_k)$.
- Estimate the specific variances $\psi_{ii}$ by $\hat{\psi}_{ii} = 1 - \sum_{j=1}^{k} \hat{\lambda}_{ij}^2$, $i = 1, \ldots, p$.
- Stop, or repeat the above steps until some convergence criterion has been reached.

### Constraint 2

- $D = \text{diag}(\sigma_{11}, \dots, \sigma_{pp}) = I$ because working with the correlation matrix is equivalent to working with standardized variables.
- Hence, $\hat{\Lambda}$ satisfies constraint 2:

$$\hat{\Lambda}' D^{-1} \hat{\Lambda} = \hat{\Lambda}' \hat{\Lambda} = (A_1^{1/2} G_1')(G_1 A_1^{1/2}) = A_1$$

  is diagonal with decreasing elements.

15 / 35

### Heywood cases

- It can happen that $\hat{\psi}_{ii} < 0$ or $\hat{\psi}_{ii} > 1$.
- This makes no sense:
  - $\psi_{ii}$ is a variance, so must be positive.
  - Working with the correlation matrix means we are working with standardized variables. So $Var(x_i) = 1$, and $Var(\psi_i)$ cannot exceed 1.
- Such cases are called Heywood cases.

16 / 35

### Example

- See R-code.

17 / 35

## Maximum likelihood estimation

18 / 35

### MLE

- Assume that $X$ has a multivariate normal distribution
- Then log likelihood function (plugging in $\bar{x}$ for $\mu$) is (see board):

$$l(\Sigma) = -\frac{1}{2} n \log |2\pi\Sigma| - \frac{1}{2} n \cdot tr(\Sigma^{-1} S)$$

- Regard $\Sigma = \Lambda\Lambda' + \Psi$ as a function of $\Lambda$ and $\Psi$, and maximize the log likelihood function over $\Lambda$ and $\Psi$.
- Optimization is done iteratively:
  - For fixed $\Psi$, one can maximize analytically over $\Lambda$
  - For fixed $\Lambda$, one can maximize numerically over $\Psi$
- This method is used by the R-function `factanal()`.
- This method can also have problems with Heywood cases.

19 / 35

### Testing for number of factors

- Advantage of the MLE method is that it allows to test if the number of factors is sufficient:
  - Null hypothesis: $k$ factors is sufficient
  - Alternative hypothesis: $k$ factors is not sufficient
  - p-value $< 0.05$ means ...
- Often sequential testing procedure is used: start with 1 factor and then increase the number of factors one at a time until test doesn't reject the null hypothesis.
- It can occur that the test always rejects the null hypothesis. This is an indication that the model does not fit well (or that the sample size is very large).

20 / 35

### Example

- See R-code

21 / 35

## Factor rotation

22 / 35

### Some general comments

- In factor rotation, we look for an orthogonal matrix $G$ such that the factor loadings $\Lambda^* = \Lambda G$ can be more easily interpreted than the original factor loadings $\Lambda$.
- Is it a good idea to look for such rotations?
  - Cons: One can keep rotating the factors until one finds an interpretation that one likes.
  - Pros: Factor rotation does not change the overall structure of a solution. It only changes how the solution is described, and finds the simplest description.

23 / 35

### What do we look for?

- Factor loadings can often be easily interpreted if:
  - Each variable is highly loaded on at most one factor.
  - All factor loadings are either large and positive, or close to zero.

24 / 35

### Two types of rotations

- Orthogonal rotation: the factors are restricted to be uncorrelated.
- Oblique rotation: the factors may be correlated.
- Advantage of orthogonal rotation: For orthogonal rotation (based on standardized variables), the factor loadings represent correlations between factors and observed variables (see board). This is not the case for oblique rotations.
- Advantage of oblique rotation: May be unrealistic to assume that factors are uncorrelated. One may obtain a better fit by dropping this assumption.

25 / 35

### Types of rotations

- Orthogonal:
  - Varimax: default in `factanal()`. Aims at factors with a few large loadings, and many near-zero loadings.
  - Quartimax: not implemented in base R.
- Oblique:
  - Promax: use option `rotation="promax"` in `factanal()`. Aims at simple structure with low correlation between factors.
  - Oblimin: not implemented in base R

26 / 35

### Example

- See R-code

27 / 35

## Estimating/predicting factor scores

28 / 35

### Random vs. deterministic factor scores

- So far, we considered the factor scores to be random. This is appropriate when we think of different samples consisting of different individuals, and we are interested in the general structure.
- One can also consider the factor scores to be deterministic. That is appropriate when we are interested in a specific group of individuals.

29 / 35

## Deterministic factor scores: Bartlett's method

- Assume normality, and suppose that $\Lambda$ and $\Psi$ are known.
- Denote the factor scores for the $i$th individual by $f_i$.
- Then $x_i$ given $f_i$ is normally distributed with mean $\Lambda f_i$ and covariance matrix $\Psi$.
- Hence, the log likelihood for one observation $x_i$ is given by

$$-\frac{1}{2}\log|2\pi\Psi| - \frac{1}{2}(x_i - \Lambda f_i)'\Psi^{-1}(x_i - \Lambda f_i).$$

- Setting the derivative with respect to $f_i$ equal to zero gives (see board):

$$\hat{f}_i = (\Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}x_i.$$

30 / 35

## Random factor scores: Thompson's method

- Consider $f$ to be random, i.e., $f$ has a normal distribution with mean 0 and covariance matrix $I$.
- Then

$$\begin{pmatrix} f \\ x \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I & \Lambda' \\ \Lambda & \Sigma \end{pmatrix}\right)$$

- Then $f|x$ has distribution $N(\Lambda'\Sigma^{-1}x, I - \Lambda'\Sigma^{-1}\Lambda)$ (see board).
- Hence, natural estimator for $f_i$ is $\Lambda'\Sigma^{-1}x_i$.

31 / 35

## Examples

- Both methods have advantages and disadvantages, no clear favorite.
- See examples in R-code.

32 / 35

## Factor analysis vs. PCA



### Common properties

- Both methods are mostly used in exploratory data analysis.
- Both methods try to obtain dimension reduction: explain a data set in a smaller number of variables.
- Both methods don't work if the observed variables are almost uncorrelated:
  - Then PCA returns components that are similar to the original variables.
  - Then factor analysis has nothing to explain, i.e. $\psi_{ii}$ close to 1 for all $i$.
- Both methods give similar results if the specific variances are small.
- If specific variances are assumed to be zero in principle factor analysis, then PCA and factor analysis are the same.



### Differences

- PCA required virtually no assumptions.
  Factor analysis assumes that data come from a specific model.
- In PCA emphasis is on transforming observed variables to principle components.
  In factor analysis, emphasis is on the transformation from factors to observed variables.
- PCA is not scale invariant.
  Factor analysis (with MLE) is scale invariant.
- In PCA, considering $k+1$ instead of $k$ components does not change the first $k$ components.
  In factor analysis, considering $k+1$ instead of $k$ factors may change the first $k$ factors (when using MLE method).
- Calculation of PCA scores is straightforward.
  Calculation of factor scores is more complex.