An Overview of Factor Analysis

Jonathan Gillard

Cardiff University

20th April 2018

Contents

Factor analysis

The purpose of *common factor analysis* is to explain the correlations or covariances among a set of variables in terms of a limited number of unobservable latent variables.

The latent variables are not generally computable as linear combinations of the original variables.

In common factor analysis, it is assumed that the variables are linearly related if not for uncorrelated random error or *unique variation* in each variable; both the linear relations and the amount of unique variation can be estimated.

Introduction

- Factor analysis aims to represent the original variables as a linear combination of unmeasurable variables (hidden factors) and a specific error term
- Examples of variables which cannot be directly measured (*latent variables*)
  - intelligence
  - social class
- We can measure these concepts indirectly (*manifest variables*)
  - IQ test, performance in school
  - Occupation, salary, value of home

The factor model

- Full model

$$\begin{array}{lll} X_1 & = l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m & + \varepsilon_1 \\ X_2 & = l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m & + \varepsilon_2 \\ \vdots & \vdots & \vdots \\ \underbrace{X_p}_{\text{Observed variables}} & = \underbrace{l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m}_{\text{Linear dependence on latent variables}} & \underbrace{+ \varepsilon_p}_{\text{Specific factors}} \end{array}$$

- Matrix notation

$$\boldsymbol{X} = \boldsymbol{L}\boldsymbol{F} + \varepsilon.$$

- $\boldsymbol{L}$ is the matrix of factor loadings and
- $\boldsymbol{F}$ is the matrix of factors (latent).

Factors, loadings, errors

- Factors
  - m factors $(F_1, F_2, \ldots, F_m)$
  - all p variables are regressed on all factors
  - each factor must have a non zero loading for at least two of the variables (i.e. they must be common to variables)
  - since the factors are unobserved we can fix their scale and location arbitrarily
- Loadings
  - l_{ij} is the loading of the i-th variable on the j-th factor
  - loadings assumed to be fixed
- Errors or specific factors
  - $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m)$ are errors
  - each variable has a unique specific factor

Example

Imagine that 200 primary school children were psychologically tested. Five tests were administered, where the children were tested on the following:

- paragraph comprehension (PARA)
- sentence completion (SENT)
- word meaning (WORD)
- addition (ADD)
- counting dots (DOTS)

The factor model assumes that we can write these five variables as a linear combination of m latent variables.

$$
\begin{aligned}
\text{PARA} &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\
\text{SENT} &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\
\text{WORD} &= l_{31}F_1 + l_{32}F_2 + \ldots + l_{3m}F_m + \varepsilon_3 \\
\text{ADD} &= l_{41}F_1 + l_{42}F_2 + \ldots + l_{4m}F_m + \varepsilon_4 \\
\text{DOTS} &= l_{51}F_1 + l_{52}F_2 + \ldots + l_{5m}F_m + \varepsilon_5
\end{aligned}
$$

Example

We extract two factors (decision usually based on a PCA, more in a moment). We have obtained the following factor loadings:

	Factor 1	Factor 2
PARA	0.81	0.06
SENT	0.72	0.08
WORD	0.91	0.01
ADD	0.02	0.69
DOTS	0.11	0.92

This implies that

$$\begin{aligned}
\text{PARA} &= 0.81F_1 + 0.06F_2 + \varepsilon_1 \\
\text{SENT} &= 0.72F_1 + 0.08F_2 + \varepsilon_2 \\
\text{WORD} &= 0.91F_1 + 0.01F_2 + \varepsilon_3 \\
\text{ADD} &= 0.02F_1 + 0.69F_2 + \varepsilon_4 \\
\text{DOTS} &= 0.11F_1 + 0.92F_2 + \varepsilon_5
\end{aligned}$$

Example

- Variance of the variable PARA
  - $\text{PARA} = 0.81F_1 + 0.06F_2 + \varepsilon_1$
  - Since we used the correlation matrix in the extraction of factors, the variable PARA has been standardised, and thus has unit variance
  - F_1 and F_2 are uncorrelated, with zero mean and unit variance
  - The factors are uncorrelated with the specific factor $\mathbb{VAR}[\text{PARA}] = \mathbb{VAR}[0.81F_1 + 0.06F_2 + \varepsilon_1] =$

    $0.81^2 + 0.06^2 + \mathbb{VAR}[\varepsilon_1] = 1$

Example

- Communality of the variable PARA:
  - $\mathbb{VAR}[\text{PARA}] = 0.81^2 + 0.06^2 + \mathbb{VAR}[\varepsilon_1]$
  - The communality is $0.81^2 + 0.06^2 = 0.6597$
- The first two components make the communality
- The nearer the communality is to 1, the more variation can be explained by the common factors

Example

- Specific variance of the variable PARA:
  - $\mathbb{VAR}[\text{PARA}] = 0.81^2 + 0.06^2 + \mathbb{VAR}[\varepsilon_1] = 1$
  - The specific variance is given by $\mathbb{VAR}[\varepsilon_1] = \phi_1 = 1 - (0.81^2 + 0.06^2) = 0.3403$
- For the total variation in the variable PARA, 66% of the variation in the variable is accounted for by common factors in the model and 34% of the variation is accounted for by a unique variable only associated to PARA.

Example

The communalities for all the variables are:

PARA	SENT	WORD	ADD	DOTS
0.6597	0.5248	0.8282	0.4765	0.8585

- A key issue on interpreting communality is the interpretability of the factors.
  - A communality of, say, 0.75 for a variable seems high, but is meaningless unless the factor is interpretable.
  - A communality of, say, 0.25 seems low but may be meaningful if the item is contributing to a well-defined factor.

Example

Considering the factor loading pattern:

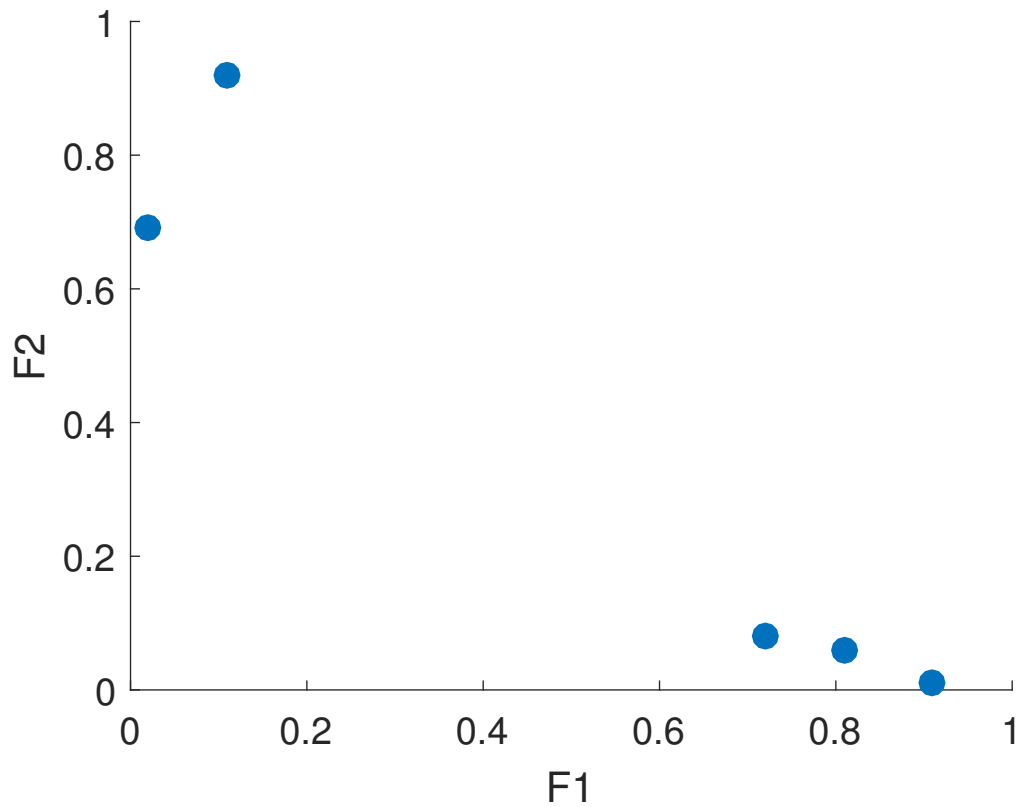
$$\begin{aligned} \text{PARA} &= 0.81F_1 + 0.06F_2 + \varepsilon_1 \\ \text{SENT} &= 0.72F_1 + 0.08F_2 + \varepsilon_2 \\ \text{WORD} &= 0.91F_1 + 0.01F_2 + \varepsilon_3 \\ \text{ADD} &= 0.02F_1 + 0.69F_2 + \varepsilon_4 \\ \text{DOTS} &= 0.11F_1 + 0.92F_2 + \varepsilon_5 \end{aligned}$$

- Noting that each loading also represents the covariance between each variable and each factor then
  - PARA, SENT AND WORD are highly loaded on F_1 and low on F_2 and may represent written ability.
  - ADD and DOTS are highly loaded on F_2 and low on F_1 and may represent numerical ability.

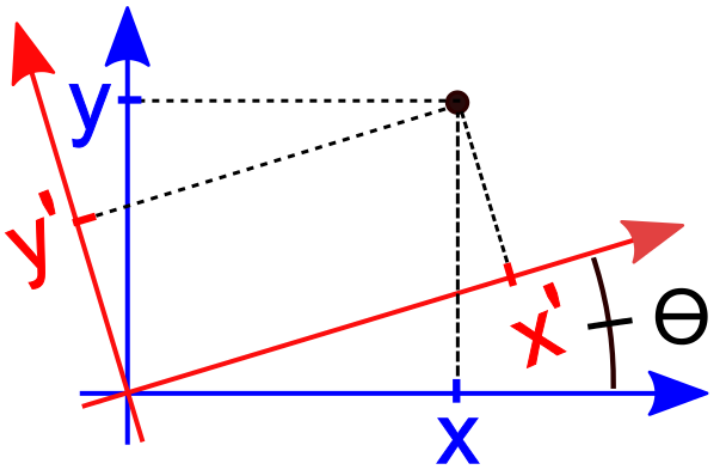
Non-uniqueness of solution: rotations

The factor loading matrix is not uniquely defined.

- The non-uniqueness of the factor loading matrix leads to the idea of rotations.
- Rotations sometimes allow the loading matrix to be more readily interpreted.



Rotations



Lots of named rotations available:

- Quartimax
- Varimax
- Equamax

which try to squeeze as many loadings to zero, in different ways.

Main idea: try lots of rotations, see which one gives you the most interpretable factors (if this is something you care about)

Selecting the number of factors

The decision is based on a principal component analysis

- *Scree plot* - Plot the variance (eigenvalue) of each principal component against the corresponding component number. We must look at the curve and locate an 'elbow'; a point in which the variances decrease in an approximately linear fashion.
- *Kaiser's rule* - Include components in the analysis that have an eigenvalue greater than 1. This dictates that a principal component must account for at least as much variation as one of the original variables used in the analysis. This ensures that no components are retained which are of less value than the original variables.
- *Proportion of variance* - the number of components to be retained in the analysis can be decided by choosing the number that account for a pre-specified amount of variation.

Example on SPSS (AthleticsData.sav)

Athletes were assessed on the following attributes:

- Score after a game of pinball
- Score after a game of snooker
- Score after a game of golf
- Time to run 1500m
- Time to row 2km
- Distance obtained after running 12 mins
- Number of bench presses of a certain weight
- Number of curls of a certain weight
- Number of push-ups until exhaustion

Perform a factor analysis to determine underlying factors behind the athlete's performance on the battery of tests.
[Analyse $\rightarrow$ Dimension Reduction $\rightarrow$ Factor...]

Example on R: Swiss Fertility and Economic (1888) Data

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

Data frame with 47 observations on 6 variables:

- Fertility
- Agriculture (% of males working in agriculture)
- Examination (% of draftees receiving highest mark on army examination)
- Education (% education beyond primary school for draftees)
- Catholic (% identifying as Catholic, as opposed to Protestant)
- Infant Mortality (live births who live less than one year)

Example on R: Swiss Fertility and Economic (1888) Data

```
data("swiss")
head(swiss)

fa0 <- factanal(swiss, factors = 1)
loadings(fa0)

fa1 <- factanal(swiss, factors = 2, rotation = "none")
loadings(fa1)

fa2 <- factanal(swiss, factors = 2, rotation = "promax")
loadings(fa2)
```

Exercise

The data sets SAQ.sav (if you'd like to use SPSS) and SAQ.R (if you'd like to use R) contain results from a questionnaire designed to predict how anxious an individual would be about learning how to use SPSS.

The designer of the questionnaire wished to know whether anxiety regarding SPSS could be broken down into specific forms of anxiety. In other words, are there other traits that might contribute to anxiety about SPSS?

Each question was a statement (see later slide) marked by a five point Likert scale ranging from 'strongly disagree' (5) through 'neither agree or disagree' (3) to 'strongly agree' (1). Analyse the data using factor analysis.

Exercise

Using SPSS

Open SAQ.sav and proceed.

Using R

First import the data using

```
SAQ <- read.csv(url("http://tiny.cc/SAQ"))
```

then start with (for example)

```
fa <- factanal(SAQ, factors = 1)
loadings(fa)
```

SAQ statements

1. Statistics makes me cry
2. My friends will think I'm stupid for not being able to cope with SPSS
3. Standard deviations excite me
4. I dream that Pearson is attacking me with correlation coefficients
5. I don't understand statistics
6. I have little experience of computers
7. All computers hate me
8. I have never been good at mathematics
9. My friends are better at statistics than me
10. Computers are useful only for playing games
11. I did badly at mathematics at school
12. People try to tell you that SPSS makes statistics easier to understand but it doesn't
13. I worry that I will cause irreparable damage because of my incompetence with computers
14. Computers have minds of their own and deliberately go wrong whenever I use them
15. Computers are out to get me
16. I weep openly at the mention of central tendency
17. I feel ill whenever I see an equation
18. SPSS always crashes when I try to use it
19. Everybody looks at me when I use SPSS
20. I can't sleep for thoughts of eigenvectors
21. I wake up under my duvet thinking that I am trapped under a normal distribution
22. My friends are better at SPSS than I am
23. If I'm good at statistics my friends will think I'm a nerd