Conceptual overview

PSY544 – Introduction to Factor Analysis

Week 1

Basic principles

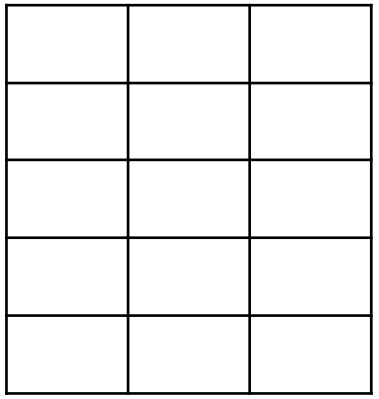
- Consider the type of data we typically apply factor analysis to:

Multivariate data – Data for a sample of individuals on a number of manifest (measured, observed) variables (like psychological tests)

Data matrix:

One column for each variable

One row for each person



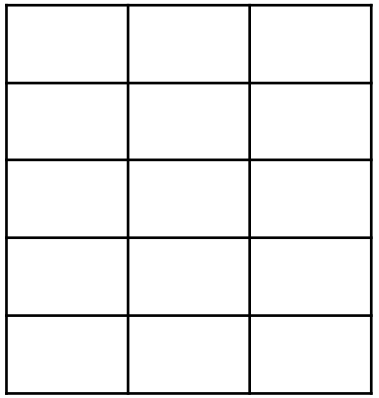
Basic principles

- Entries in the data matrix represent scores of each person on each manifest variable
- The underlying premise of factor analysis is that these data are not completely random, but have some systematic aspects that can be studied

One column for each variable

Data matrix:

One row for each person



Basic principles

- Factor analysis is **one way** to study the underlying structure
- i.e., we are trying to get a simpler understanding of data

- FA was originally developed to study the structure of human mental abilities. Factor models provide a formal basis for various theories of intelligence and the structure of mental abilities.
- We will mostly use examples from this domain in the course
- However, FA is used in other domains, like personality, attitudes, etc.

Basic principles

- Two aspects of FA:
    - **Theory** – statistical models specifying the underlying structure of data
    - **Methodology** – procedures that allow us to analyze these data and reveal the specified structure

Key terms and definitions

- **Manifest variable** – variable that can be directly measured (or observed)
- **Latent variable** – variable that cannot be directly measured (or observed) – a hypothetical construct. A latent variable is a **factor** in factor analysis. Thus, a factor is a variable and individuals have scores on those factors (hypothetically).
- **Population** – The entire set of individuals of interest
- **Sample** – A selected group of individuals from the population (N persons)

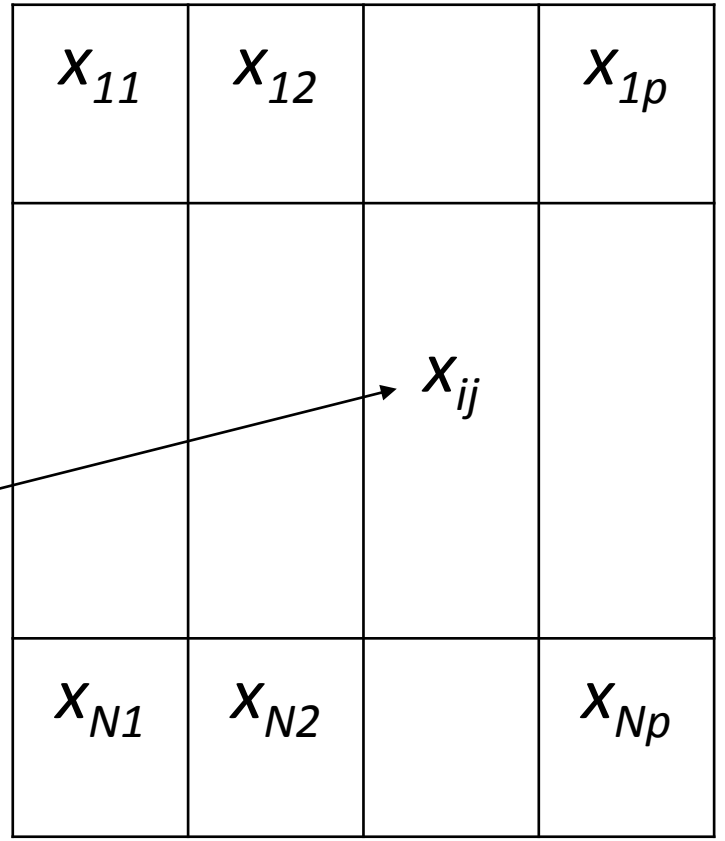
Key terms and definitions

Data matrix:

p columns (variables)

X **=** *N rows (individuals)*

Score of person *i* on variable *j*



Key terms and definitions

- **What can we observe in these data?**
  - Variation of each variable over individuals (measured by variance / SD)
  - Covariation in a pair of variables over individuals (measured by covariance / correlation)

x_{11}	x_{12}		x_{1p}
		x_{ij}	
x_{N1}	x_{N2}		x_{Np}

Key terms and definitions

Correlation matrix:

R:

p manifest variables

1	r_{12}	r_{13}			r_{1p}
r_{21}	1	r_{23}			r_{2p}
r_{32}	r_{32}	1			r_{3p}
				r_{kj}	
			r_{jk}		
r_{p1}	r_{p2}	r_{p3}			1

p manifest variables

R:

1	r_{12}	r_{13}		r_{1p}
r_{21}	1	r_{23}		r_{2p}
r_{32}	r_{32}	1		r_{3p}
			r_{kj}	
		r_{jk}		
r_{p1}	r_{p2}	r_{p3}		1

- To understand the pattern of relationships among the MVs, we could just try to describe it in terms of the entries in this matrix
- However, this gets increasingly difficult as *p* increases.

...if *p* is large, the number of correlations ($\frac{p(p-1)}{2}$) is too big to understand fully

- **The general rationale of factor analysis is that these correlations are structured and can be explained in a relatively simple way.**

- The **objective** of factor analysis, then, is to **uncover** and **understand** the structure that produces the correlations in the data

- Essential to this objective is the notion of **factors**

- **Factors** are latent, unobservable variables – hypothetical constructs

- The basic principle of FA is that there exists a small number of factors (within a particular domain) which influence the MVs and thus produce the correlations (covariances) between manifest variables.

- A correlation between two MVs is due to these two MVs being dependent on one or more of the same factor(s)

- So, again, what we want is to identify the number and nature of the factors that produce the observed correlations between the MVs.

- Interrelationships between all possible MVs in a given domain can be explained by a limited number of factors. The number of factors is considered to be (much) smaller than the number of MVs (if this were not the case, we would gain very little by doing factor analysis)

  → e.g., it is assumed that a **limited** number of mental abilities will explain relationships between **all** ability tests

  …no MV single-handedly represents a distinct ability or trait.

- Again, the factors influence the MVs. One of the objectives in FA is to estimate the degree of these influences. We measure these by the means of **factor loadings**.

- The numerical values of factor loadings indicate the strength of the factor's influence on the MV (a zero indicates no influence). Factor loadings are equivalent to regression coefficients, standing for the influence of a factor (independent variable) on a MV (dependent variable)

- The pattern of factor loadings helps us determine the nature of a factor …in other words, a factor is defined by the subset of MVs that it substantially influences

To recap:

- Correlations between manifest variables exist because the manifest variables are influenced by one or more of the same factors.
  (e.g., text comprehension and verbal fluency are correlated because both are influenced by a common underlying factor of verbal ability)

- Our aim is to determine the number and nature of the underlying factors and their pattern of influence on the manifest variables.

- We want to obtain a simple explanation of relationships in the data using a small number of factors.

Example

- Suppose we have scores from a sample of individuals on 4 performance measures: paragraph comprehension, vocabulary, arithmetic skills, and mathematical problem solving. We get the following correlation matrix:

	PC	VO	AR	MPS
PC	1			
VO	.49	1		
AR	.14	.07	1	
MPS	.48	.42	.48	1

Example

- We would like to identify the underlying factors to explain the correlations. Thus, we employ factor analysis methods and obtain a factor loading matrix:

	Factor 1	Factor 2
PC	.70	.10
VO	.70	.00
AR	.10	.70
MPS	.60	.60

Example

	Factor 1	Factor 2
PC	.70	.10
VO	.70	.00
AR	.10	.70
MPS	.60	.60

- Elements in the matrix represent the linear influence of each factor on each measure.

- In this course, we will study methods that will allow us to obtain such interpretable factor loading matrices.

- Keep in mind that we are using a model – a one which represents some hypothesized structure of observed data. Any mathematical model is – at least to some extent – wrong and does not perfectly correspond to reality.
- A model that makes sense conceptually but does not fit reasonably well is useless.
- A model that fits great but does not make sense is useless as well.

- A factor analysis is not applicable to just any data.

- In the world of factor analysis, situations differ regarding the existence of prior hypotheses / knowledge about the number and nature of the factors:

Exploratory (unrestricted) FA:

We have little prior idea of how many and what kind of factors there are.

Confirmatory (restricted) FA:

We do have a hypothesis (or hypotheses) about the number and nature of factors.

...the underlying theoretical model is **the same**!

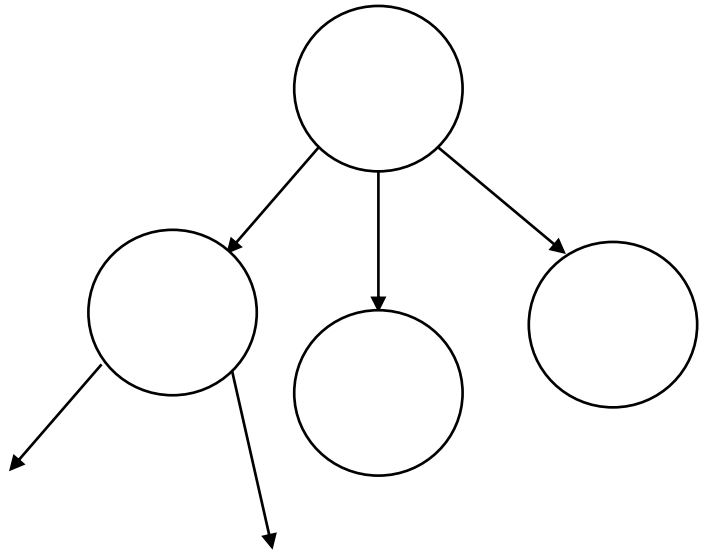
A bit of history

- Factor analysis began with the study of mental abilities

- Charles Spearman proposed the first factor model in 1904:
  - Performance on any test is a function of two factors – a **general** ability factor (Spearman's **g**) common to all ability tests, and a **specific** ability factor relevant only to the specific test in question.
  - → The two-factor theory of intelligence
  - Ability tests correlate because they all depend on the general factor.

A bit of history

- Burt and Vernon, on the other hand, proposed a **hierarchical model** of human abilities:
  - The human mind is organized in a hierarchy of abilities.
  - The general ability sits atop this hierarchy
  - More specific abilities are located lower in the hierarchy



A bit of history

- The **Common Factor Model** of L. L. Thurstone became the most prominent approach to FA since the 1940s. Thurstone disagreed with both the notion of **g** and a hierarchy of abilities.
- According to Thurstone, MVs depend on two kinds of underlying factors:
  - **Common factors** that are *common* to more than one MV
  - **Unique factors** that influence only one MV. Unique factors do not explain correlations between MVs.
- The p manifest variables depend on m common factors and p unique factors, where $m < p$

The Common Factor Model

- Therefore, for a given set of *p* MVs, there are *m+p* factors
- Each unique factor has two components:
  - Specific factor
  - Error of measurement

…the specific factor represents systematic factors affecting only a particular MV. The error component represents random error.

The Common Factor Model

We can break down the variance of a given MV in the following way:

Unique variance = Specific variance + Error variance

Observed variance = Common variance + Unique variance

= Common variance + Specific variance + Error variance

$$\text{Communality} = \frac{\text{Common variance}}{\text{Observed variance}} = 1 - \frac{\text{Unique variance}}{\text{Observed variance}}$$

= the proportion of observed variance due to common factors

The Common Factor Model

The mathematical expression of the Common Factor Model:

$$x_{ij} = \mu_j + \lambda_{j1}z_{i1} + \lambda_{j2}z_{i2} + \cdots + \lambda_{jm}z_{im} + 1u_{ij}$$

Mean + Common factor part + Unique factor part

x_{ij} **is the score of person *i* on manifest variable *j***

μ_j **is the mean of manifest variable *j***

The Common Factor Model

The mathematical expression of the Common Factor Model:

$$x_{ij} = \mu_j + \lambda_{j1}z_{i1} + \lambda_{j2}z_{i2} + \cdots + \lambda_{jm}z_{im} + 1u_{ij}$$

Mean + Common factor part + Unique factor part

z_{ik} **is the common factor score of person *i* on factor *k* (latent variable score)**

λ_{jk} **is the factor loading (regression weight) of MV *j* on factor *k***

u_{ij} **is the unique factor score of person *i* on unique factor *j* (also latent)**

...the unique factor score consists of a specific part and an error part:

$$u_{ij} = s_{ij} + e_{ij}$$

The Common Factor Model

We mentioned variances before – do not confuse the scores (x_{ij}, z_{ik}...) with the variances of those scores [var(x_j), var(z_k)], which is how these scores vary across persons.

The model can be re-written by subtracting the mean from both sides:

$$x_{ij} - \mu_j = \lambda_{j1}z_{i1} + \lambda_{j2}z_{i2} + \cdots + \lambda_{jm}z_{im} + 1u_{ij}$$

...thus, we can see that the model specifies the deviation from the mean as a function of the common and unique factors.

The Common Factor Model

Important assumption: In the model, the unique factor scores for different MVs are assumed to be uncorrelated over all persons. Therefore, all partial correlations between MVs, controlling for the effect of the common factors, are assumed to be zero.

- In other words, correlations between MVs are only due to the common factors (that's why they're called *common*)
- This assumption refers to the population

The Common Factor Model

- What factors are common and what factors are specific depends on the manifest variables in the dataset.
- If we change the set of MVs by introducing new MVs or deleting MVs, we can potentially change specific factors into common factors, and so on.

The Common Factor Model

- The model is will always be wrong to some degree (it's a *model* after all). What are some of the ways the model could be wrong?
  - 1) The assumption of linearity – the MVs are specified as linear functions of factors. Nobody really thinks the real world is perfectly linear.
  - 2) The number of common factors is generally assumed to be small ($m << p$). In reality, there are probably many, many influences on a score. However, we hope to identify the non-negligible ones.
- We should recognize the common factors will not perfectly explain the variation and covariation of the manifest variables.

The Common Factor Model

- The model equation looks like a multiple regression equation.
  - The manifest variables are dependent variables
  - The factors are independent variables
  - The factor loadings are regression weights / coefficients

- The factor analysis model is like a set of multiple linear regressions where the independent variables are unobservable.