

Deciding on the number of factors

PSY544 – Introduction to Factor Analysis

Week 9

Introduction

- The objective of exploratory (unrestricted) factor analysis is to determine the **number** and **nature** of the common factors.
- Deciding on the number of factors to extract is key, also because subsequent analyses (such as rotation) depend on this decision.
- There's a lot of literature out there on the how-many-factors problem. Unfortunately, most of this literature focuses on developing mechanistical rules that will indicate the “true” number of factors.

Introduction

- This is very misguided.
- First of all, there is no such thing as “the true number of factors”.
- Second, mechanistical rules – while sometimes useful – do not work every time.

Introduction

- We need to recognize that no parsimonious model will hold exactly.
- We must use informed judgement and need to take into account various sources of information.
- At times, the decision won't be easy and all the available information will not clearly point us in either direction.

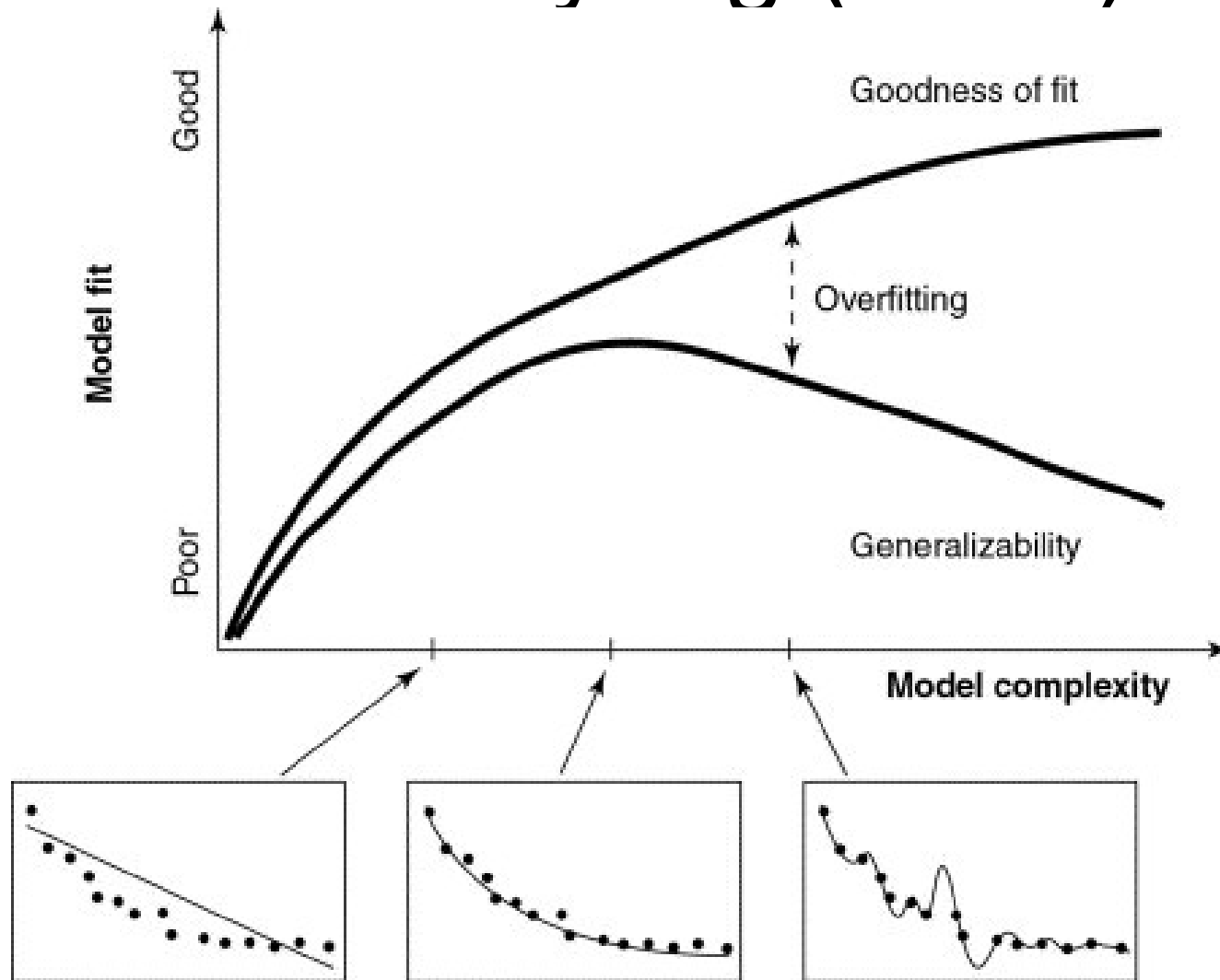
Introduction

- A lot of researchers use the mechanistical rules or rules of thumb without careful judgement or understanding.
- This is not effective and can result in misleading solutions or interpretations.
- The mechanical rules can (and often do) provide useful information, but none can be effectively used by itself, without considering other information.

Introduction

- There is a trade-off in what we do – we would like to find such a number of factors so that the model would fit reasonably well, but still provides a considerable degree of parsimony.
- **Ideal situation:** to find a model that fits reasonably well, while a model with fewer factors fits significantly worse and a model with more factors doesn't fit significantly better.
- Also, the identified factors have to be meaningful and interpretable.

Pitt & Myung (2002)



Introduction

- Nowadays, simply doing an EFA is not enough (oh, the old days...). So it is also important to view EFA as a step in the entire factor analysis process – it is very likely that you will subsequently perform a CFA on similar data, so EFA can work as a tool for limiting the space of potential models and identifying a couple of candidate models (with a reasonable number and interpretability of the common factors)

EFA, CFA, UFA, RFA...

- I have been using the terms *exploratory*, *confirmatory*, *restricted*, *unrestricted*, ..., in a pretty chaotic way sometimes, so let me explain.
- Traditionally, people recognize EFA (exploratory) and CFA (confirmatory). But this sounds a bit more like it's about the purpose and not about the model itself.
- Better to use *restricted* (typically CFA) and *unrestricted* (typically EFA) factor analysis. This simply refers to whether there are any restrictions on Λ

Sources of error

- When performing EFA (or most statistical analyses, actually), we face two sources of error:
 1. Sampling error (our **R** is not our **P**)
 2. Model error (even if we had **P**, it would not fit perfectly – all models are wrong)

Number of factors

- OK, so, we established there is not such thing as the *true number of factors*
- Our objective, then, is to identify the plausible number of *major common factors*
- What kind of information is available?

Number of factors

- Types of information available:
 1. Rules of thumb (mostly based on eigenvalues)
 2. Statistical tests
 3. Common sense
 4. Informed judgement

Rules of thumb

- **Number of eigenvalues greater than 1**
- Sometimes known as the *Kaiser criterion* or the *Kaiser-Guttman criterion*.
- According to Guttman (1954), if the model holds exactly in the population, the number of eigenvalues greater than 1 provides a lower bound to the number of factors.

Rules of thumb

- **Number of eigenvalues greater than 1**
- Although there is theoretical justification for this rule at the population level and in the ideal case the model holds exactly, it is routinely used with sample correlation matrices.
- People use this VERY often. It has been repeatedly demonstrated to be misleading.

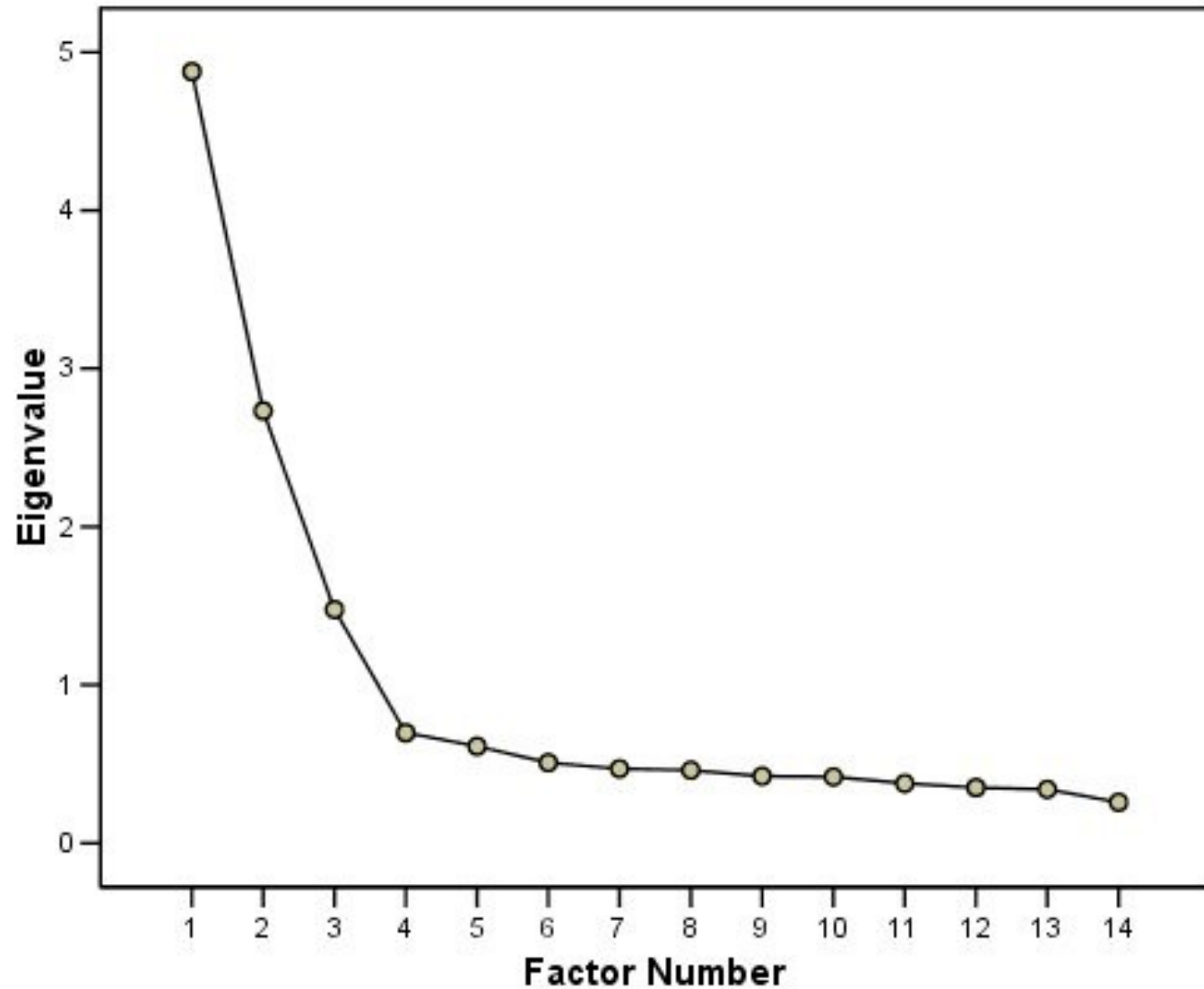
Rules of thumb

- **Number of eigenvalues greater than 1**
- Furthermore, the theoretical justification only applies to the population correlation matrix, but most EFA software gives you the eigenvalues for the sample *reduced* correlation matrix.
- It can serve you as a guide or a reference point, but dogmatic application will (probably) lead you nowhere.

Rules of thumb

- **Scree plot**
- Cattell (1966). Simply plot the eigenvalues of the sample correlation or reduced sample correlation matrix and look at the last large discontinuity.
- Let's look at an example.

Rules of thumb



Rules of thumb

- **Scree plot**
- In this case, we would choose $m = 3$.
- This procedure is very subjective and does not have any theoretical justification. Only an informal rationale is available – if there exists m factors, then there will be m relatively large eigenvalues. The rest of the eigenvalues (small) will only represent noise.

Rules of thumb

- **Scree plot**
- As before, this method can be informative to some extent, but should not be used exclusively.

Goodness of fit tests

- **Test of perfect fit**
- Using the likelihood-ratio test statistic to test a hypothesis that the model fits perfectly in the population.
- We've covered this one already.
- For a model with m factors, rejecting the null hypothesis means we need more factors. When we fail to reject the null, we should stop adding more factors.

Goodness of fit tests

- **Test of perfect fit**
- Problems!
- The null hypothesis is not true, we know that, so what's the point in testing it?
- Sequential tests are not independent.
- Functionally a test of sample size.
- Mechanical.

Fit indices

- **RMSEA**
- Selecting the number of factors (or selecting a couple of plausible numbers of factors) is fundamentally a problem of model fit.
- We can sequentially estimate models with increasing numbers of factors and look at changes in model fit.
- Encourages mechanistical use. Don't rely too much on the suggested values for RMSEA.

Fit indices

- **Tucker-Lewis Index (TLI)**
- The TLI is a so-called *incremental* fit index. What does that mean?
- The index is based on comparing the fit of a model with m factors to the fit of two reference models.
- The first one is the „world’s worst model”, the so-called “null model”, with $m = 0$. This model implies zero correlation between the MVs in the population
- The second one is the “world’s best model”, the so-called “ideal model”, which fits perfectly in the population.

Fit indices

- **Tucker-Lewis Index (TLI)**
- The model is based on the properties of the χ^2 -distributed likelihood ratio test statistic. When a particular model is “correct”, the expected value of the test statistic is the number of degrees of freedom of the given χ^2 distribution, $E(\chi^2) = df$
- Therefore, we would expect the ratio $\frac{\chi^2}{df}$ to be roughly 1 for a “correct model”. As the model fits worse, the ratio becomes larger than 1.

Fit indices

- **Tucker-Lewis Index (TLI)**
- If we have a model with m common factors, we can compute the ratio as:

$$Q_m = \frac{\chi_m^2}{df_m}$$

- ...and we can compute the same ratio for the null model:

$$Q_0 = \frac{\chi_0^2}{df_0}$$

Fit indices

- **Tucker-Lewis Index (TLI)**
- As the null model will probably fit the data very badly, we would expect Q_0 to be very large. For an ideal model, we would expect Q to be roughly 1.
- The TLI is then defined as:

$$TLI_m = \frac{Q_0 - Q_m}{Q_0 - 1}$$

Fit indices

- **Tucker-Lewis Index (TLI)**

$$TLI_m = \frac{Q_0 - Q_m}{Q_0 - 1}$$

- The numerator represents the improvement in fit of the model with m factors over the null model. The denominator represents the improvement in fit of the ideal model over the null model.
- The ratio represents where does the m -factor model stand on the continuum between the “worst fit” and the “best fit”.

Fit indices

- **Tucker-Lewis Index (TLI)**

$$TLI_m = \frac{Q_0 - Q_m}{Q_0 - 1}$$

- A value over .90 is desirable, preferably higher than that.

Selecting the number of factors

1. Consider prior hypotheses about the number of factors
2. There is no need for a single “right answer” at the exploratory phase of an analysis
3. Consider some rules-of-thumb as a source of guidance, but don't rely on them blindly (scree plot, eigenvalues...)
4. Consider the results of goodness of fit tests and measures of fit, including RMSEA and TLI
5. Based on what you have observed, decide about the optimal number of factors, or a couple of alternatives

Selecting the number of factors

6. Go the extra mile and extract one or more additional factors, and see what happens – it's better to “over-factor” than to “under-factor”
7. Look for converging evidence
8. Use your judgment and knowledge about the data. If someone says you are not allowed to, they are wrong.

What not to do

- Use one piece of evidence or one mechanistical approach by itself
- Not consider multiple sources of information