Special topics I

PSYn5440 – Introduction to Factor Analysis

Week 10

Factor scores

- Recall the common factor model: $\boldsymbol{x} = \boldsymbol{\Lambda z} + \boldsymbol{u}$

- The vector $\boldsymbol{z}$ contains individuals' scores on the common factors.
- Sometimes, researchers wish to obtain these scores so they can work with them in further analyses.

- For instance, they might wish to use them as independent or dependent variables in regression models or in t-tests for group comparisons.

Factor scores

- The problem is – these scores, as you already know, cannot be determined exactly. They are latent, unmeasurable, unknowable. They are indeterminate.

- Mathematically, the reason is as follows. In the model:

$$\boldsymbol{x} = \boldsymbol{\Lambda z} + \boldsymbol{u}$$

…the *p* MVs are defined as linear functions of *m* common factors and *p* unique factors. In effect, the model has *p* MVs and *p+m* LVs.

Factor scores

- It is impossible to determine $p+m$ scores from only p variables.
- This has long been criticized in the literature as a reason for shunning factor analysis altogether.
- Proponents of factor analysis argued that this issue is really only an issue when the factor scores are involved and that it does not affect the covariance / correlation structures.
- So, the problem only arises when one wishes to obtain the scores.

Factor scores

- In some textbooks and computer packages, you might encounter the procedure of "estimating" the factor scores.

- These procedures are problematic. Even more so if you consider that they assume you know $\boldsymbol{\Lambda}$ and $\boldsymbol{D}_{\psi}$, which you don't – all you have is $\widehat{\boldsymbol{\Lambda}}$ and $\widehat{\boldsymbol{D}}_{\psi}$. But let's assume you know the population parameters for the sake of argument.

Factor scores

- There are two common methods for estimating $\boldsymbol{z}$.

- The regression method: $\hat{\boldsymbol{z}}_R = \boldsymbol{\Lambda}'\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{D}_\psi\right)^{-1}\boldsymbol{x}$
  - The factors are considered to be dependent variables and the MVs are considered to be independent variables (opposite from what the common factor model implies).
  - Then, most accurate "predictions" are obtained for the factor scores.

Factor scores

- Bartlett's method: $\hat{\boldsymbol{z}}_B = (\boldsymbol{\Lambda}'\boldsymbol{D}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\Lambda})^{-1}\,\boldsymbol{\Lambda}'\boldsymbol{D}_{\boldsymbol{\psi}}^{-1}\boldsymbol{x}$

- This method yields factor scores estimates which, when plugged into the factor analysis data model, provide the most accurate reconstruction of the MV scores (in a least squares sense)

Factor scores

- Keep in mind that:
  - The factor scores obtained with either method are NOT the “true” factor scores
  - The factor scores obtained with either method are different
  - The correlations between the factor scores obtained with either method are not equal to the model-implied correlations between the common factors.
- In other words, they should not be treated as factor scores.

Factor scores

- So, what should you do if you wish to investigate the factor scores?
- You should use **structural equation modeling** (SEM).
- SEM allows you to use the factor scores as independent variables, dependent variables, mediators, etc. All without the need to obtain the actual scores.
- This is the only correct way of working with factor scores.

Factor scores

- Sometimes, researchers also calculate **composite scores**.
- That is, they obtain standard scores for each individual on MVs that have a high loading on some particular factor, sum the standard scores up (or take their average) and use this composite variable as a substitute for factor score.

- This is heavily used in practice.

Factor scores

- Researchers conduct a factor analysis, identify the MVs that load highly on each factor, and work with these manifest variables as if they were the actual factors.

- I’m not saying these composite scores have no meaning, but they are certainly not factor scores, or estimates of factor scores. Working with them is no longer factor analysis.

Sample size in factor analysis

- When doing factor analysis, how large of sample do we need?
- What *N* is “high enough” so that we know our model is accurate?
- This is a classic question, and you can find many different answers. There are multiple “rules of thumb”, some are claiming a minimum N, some focus on the minimum ratio of *N* to the number of MVs.

Sample size in factor analysis

- Gorsuch (1983) and Kline (1979): $N \geq 100$; $N / p \geq 5$
- Guilford (1954): $N \geq 200$
- Cattell (1978): $N \geq 250$; $N / p \geq 3$, better if $N / p \geq 6$
- Everitt (1975): $N / p \geq 10$
- Comrey & Lee (1992): *$N = 100$ … poor*

  $N = 200$ … fair

  $N = 300$ … good

  $N = 500$ … very good

  $N = 1000$ … excellent

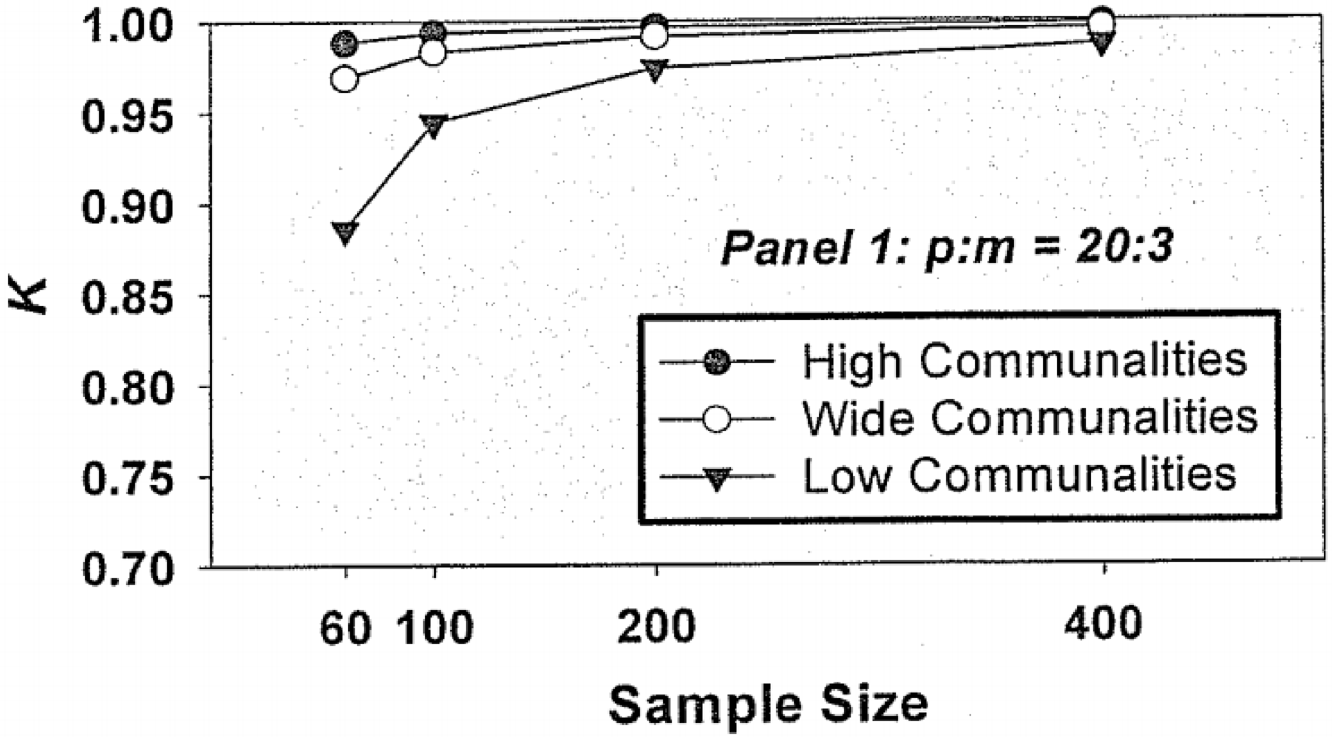
Sample size in factor analysis

- Consistent? Not quite.
- MacCallum et al. (1999) argue that these guidelines are not useful because they are based on a misconception that the minimum *N* required to achieve the model is accurate does not change across different situations / different data / different studies.

- The necessary *N* depends heavily on a couple of aspects of the data / study.
- Sometimes, a small *N* is enough. Sometimes, you need much more.

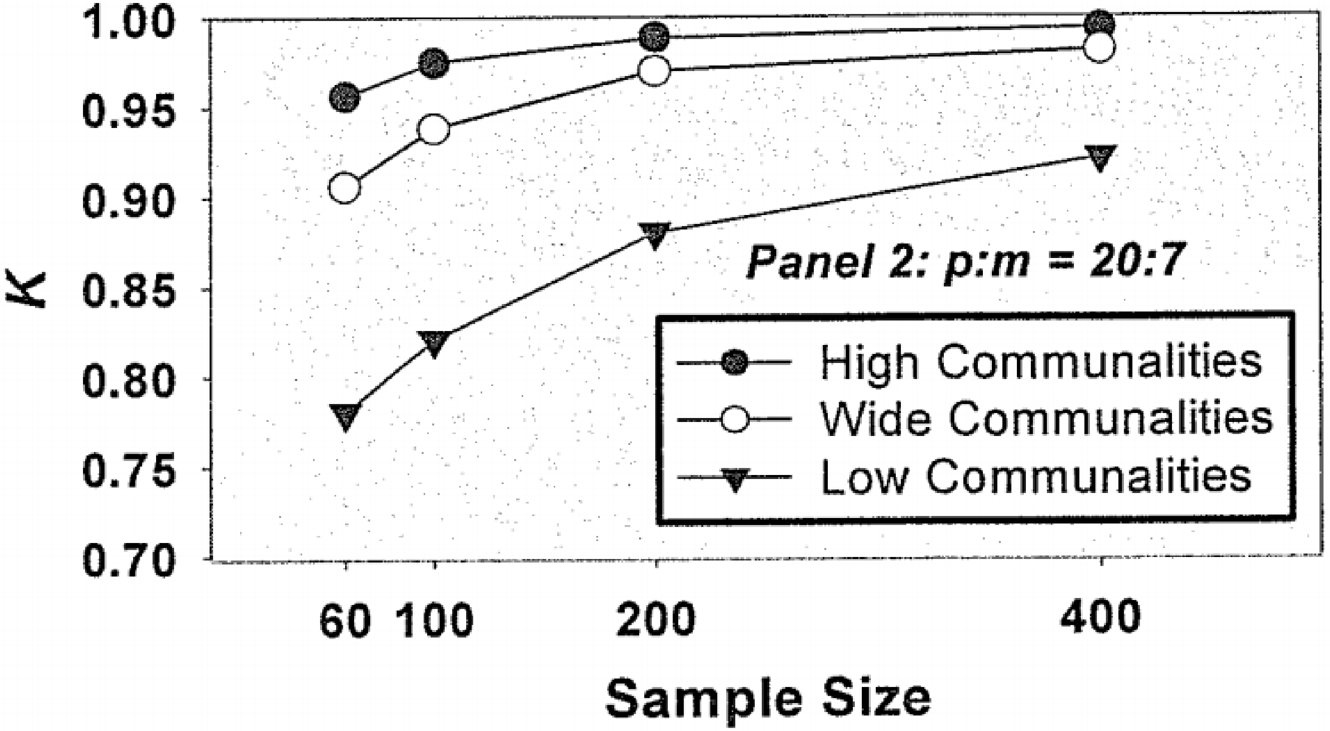
Sample size in factor analysis

- Simulation study done by MacCallum and Tucker (1991) shows two critical criteria - the **communalities**, and the number of MVs (p) per common factor (m). The y-axis (*K*) shows how well the model recovered known parameters



Sample size in factor analysis

- Simulation study done by MacCallum and Tucker (1991) shows two critical criteria - the **communalities**, and the number of MVs (p) per common factor (m). The y-axis (*K*) shows how well the model recovered known parameters



Sample size in factor analysis

- Conclusion?
- Don’t rely on the rules-of-thumb. What a surprise ☺