Confirmatory factor analysis – A (very) brief introduction

Dr Wan Nor Arifin

Unit of Biostatistics and Research Methodology,
Universiti Sains Malaysia.
E-mail: wnarifin@usm.my

Contents

1 Objectives

1. Extend the knowledge of exploratory factor analysis to confirmatory factor analysis.
2. Understand and apply the basic knowledge about the analysis.
3. Specify the measurement model, fit the model, revise the model if required in lavaan, and interpret the results.

2 Introduction

- In CFA model is specified; the factors, the items under each factor, and the pattern of relationships between them.
- Usually analysis is done on covariance matrix.
- How the variance-covariance matrix produced from the model fits the variance-covariance matrix of the observed data $\rightarrow$ Goodness of fit of model to the data.
- Needs strong theoretical specification of the model ahead of the analysis.
- CFA is actually part of Structural Equation Modeling (SEM), which basically consists of two components:
    1. measurement model (CFA): dealing with latent variables (factors) and the relationships between the items and the factors, which is our main focus here.
    2. structural model (path analysis): dealing with how latent variables are related to each other.

3 Common factor model

- Recall back our common factor model, the variance consists of 2 parts:
    1. Common variance, which is the variance accounted by the latent factor, i.e. the variance shared between the related items.
    2. Unique variance, which is the variance specific to the item. It can be further partitioned into systematic error and random error variances.
- Basic equation revisited:

$$y_j = \lambda_{j1}\eta_1 + \lambda_{j2}\eta_2 + \ldots + \lambda_{jm}\eta_m + \epsilon_j$$

where y_j is the jth of p observed variables, λ_{jm} is the jth factor loading corresponding to m latent factor, η_m is the latent factor and ϵ_j is the jth unique variance. Or further simplified in form of

$$y = \Lambda_y \eta + \epsilon$$

where y is the observed variables, Λ_y is the factor loadings of y variables, η is the latent factors and ϵ is the unique variances. Or sometimes in its expanded matrix form as

$$\Sigma = \Lambda_y \Psi \Lambda_y^{'} + \Theta_\epsilon$$

where Σ is the $p \times p$ correlation matrix of p items, Λ_y is the $p \times m$ factor loading matrix, Ψ is the $m \times m$ factor correlation matrix and Θ_ϵ is the $p \times p$ diagonal matrix of unique variances.

- For example, our previous STATISTICS IMPORTANCE factor consists of 3 items:

$$I_1 = \lambda_{11}\eta_1 + \epsilon_1$$

$$I_2 = \lambda_{21}\eta_1 + \epsilon_2$$

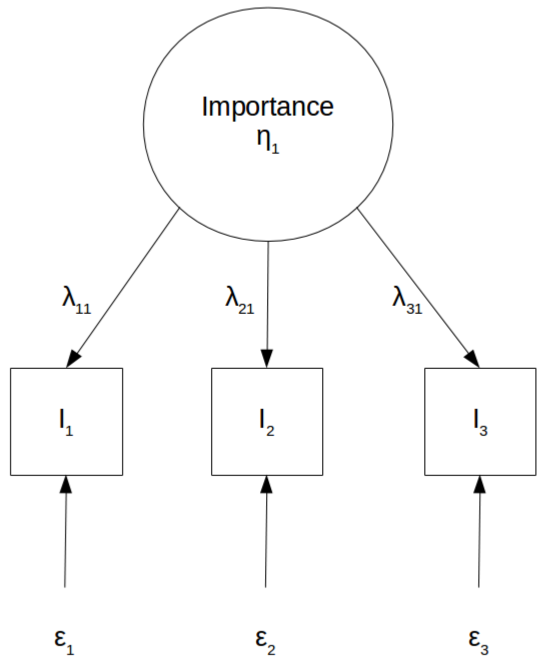
$$I_3 = \lambda_{31}\eta_1 + \epsilon_3$$

can be represented as

$$I = \Lambda_I \eta + \epsilon$$

or as a path diagram (Figure 1)

Figure 1: Path diagram for STATISTICS IMPORTANCE factor



4 Scaling the factor

- Latent variable is an unobserved variable, it has to be scaled by a method to define its metrics/unit of measurement. The approaches are:
  - Marker/reference indicator variable approach. By setting the metric of latent variable to one of its item. The most common approach.
  - Variance of latent variable is set to 1.

5 Degrees of freedom

- To perform CFA, the model also needs statistical identification. Depending on the df
  - $df > 0$: Overidentified, which is what we want to perform CFA. Number of known parameters, $b >$ unknown parameters, a (freely estimated parameters).

- df=0: Just identified, b=a. Always perfect fit, cannot apply the goodness-of-fit assessment. Also not for the analysis.
- df<0: Underidentified. b<a. Cannot perform the analysis.

- Calculating the df

$$df = b - a$$

$$b = p(p + 1)/2$$

where b is the number of elements in input matrix (i.e the variance-covariance matrix/correlation matrix) and p is the number of items. While for the a (the freely estimated parameters, have to calculate manually the number of model parameters to be estimated, which are:

1. Factor loadings
2. Error variances
3. Factor variances
4. Factor covariances

- For Figure 1 example the df

$$b = 3(3 + 1)/2 = 6$$

thus a, using marker indicator approach

$$a = 2(factor\,loadings)+3(error\,variances)+1(factor\,variance)+0(factor\,covariances) = 6$$

$$df = b - a = 6 - 6 = 0$$

which means our model is just identified! Which is not a good thing. If we calculate df for our AFFINITY OF STATISTICS factor (again from our previous lecture), consisting of 5 items

$$b = 5(5 + 1)/2 = 15$$

$$a = 4(FLs) + 5(error\,VARs) + 1(factor\,VAR) + 0(factor\,COVAR) = 10$$

$$df = 15 - 10 = 5$$

thus our model is overidentified and ready for CFA!

6 Maximum likelihood estimation

- The most commonly used estimation method in CFA, but it needs multivariate normal data as we will check later in hands-on.

- The fitting function that is minimized for the ML estimation is

$$F_{ML} = ln|S| - ln|\Sigma| + trace[(S)(\Sigma^{-1})] - p$$

where $|S|$ is the determinant of the input (i.e. observed) variance-covariance matrix that is compared to $|\Sigma|$ which is the determinant of variance-covariance matrix as predicted by the measurement model. If $(S) = (\Sigma)$, thus $(S)(\Sigma^{-1}) = SS^{-1} = I$, i.e the identity matrix. *trace* is the sum of the diagonal of the matrix, thus in this case, $trace(I) - p = 0$.

References

Kline, R. (2011). *Principles and Practice of Structural Equation Modeling*. Methodology in the social sciences. Guilford Press.

Brown, T. (2006). *Confirmatory Factor Analysis for Applied Research*. Methodology in the social sciences. Guilford Press.

Bartholomew, D. J., Steele, F., Moustaki, I., and Galbraith, J. I. (2008). *Analysis of multivariate social science data*. USA: CRC Press.