# Confirmatory factor analysis

Dr. Wan Nor Arifin

Biostatistics and Research Methodology Unit

Universiti Sains Malaysia

wnarifin@usm.my / wnarifin.github.io



Updated January 18, 2024

# Outlines

- Introduction
- Confirmatory factor analysis
- Analysis steps in CFA
- Composite reliability
- Path diagram
- Results presentation

# Introduction

# Introduction

Types of factor analysis:

1. Exploratory factor analysis (EFA)
2. Confirmatory factor analysis (CFA)

# Introduction

**Confirmatory Factor Analysis (CFA)**

- Confirmatory analysis – confirm item-factor relationship, confirm theory
- Part of Structural Equation Modeling (SEM):
  - Measurement model (CFA)
  - Structural model (path analysis)
- Includes model fit assessment

# Introduction

- Needs strong theory, CFA model is specified ahead of analysis:
  - factors
  - items under each factor
  - patterns of relationship between them
- Analysis is usually on **variance-covariance** matrix
- To what extend the matrix expected by model **fits** the matrix of observed data $\rightarrow$ Model fit

# Introduction

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.

# EFA vs CFA revisited

| EFA | CFA |
| --- | --- |
| Exploratory | Confirmatory |
| No need theory | Theory |
| Explore & Generate theory | Confirm theory |
| Item not fixed to factor | Item fixed to factor |
| Rotation | No rotation |
| No Hx testing | Hx testing & model fit |

# Confirmatory factor analysis

# CFA - 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.

# CFA

Basic equation:

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

where $y_j$ is the $j$th of $p$ observed variables, $\lambda_{jm}$ is the $j$th factor loading corresponding to $m$ latent factor, $\eta_m$ is the latent factor and $\epsilon_j$ is the $j$th unique variance.

# CFA

Simplified equation:

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

where $y$ is the observed variables, $\Lambda_y$ is the factor loadings of $y$ variables, $\eta$ is the latent factors and $\epsilon$ is the unique variances.

# CFA

Matrix form:

$$\Sigma = \Lambda_y \Psi \Lambda_y^\intercal + \Theta_\epsilon$$

where $\Sigma$ is the $p \times p$ correlation matrix of $p$ items, $\Lambda_y$ is the $p \times m$ factor loading matrix, $\Psi$ is the $m \times m$ factor correlation matrix and $\Theta_\epsilon$ is the $p \times p$ diagonal matrix of unique variances.

# CFA

For example, our previous **Importance** factor from EFA consists of 3 items:

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

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

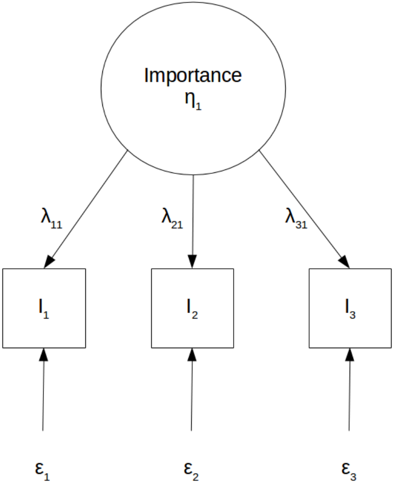
$$I_3 = \lambda_{31}\eta_1 + \epsilon_3$$

can be represented as:

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

# CFA

As path diagram:



# CFA - Scaling of 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.

# CFA - Degrees of freedom

To perform CFA, the model also needs statistical identification. Depending on the *df*:

- $df > 0$: Overidentified. Desired for performing CFA.
- $df = 0$: Just identified. Always gives perfect model fit, cannot apply the goodness-of-fit assessment. Not for analysis.
- $df < 0$: Underidentified. Cannot perform analysis.

# CFA

Calculate *df*

$$df = b - a$$

$$b = \frac{p(p+1)}{2}$$

here $b$ is the number of elements in input matrix (i.e the variance-covariance matrix/correlation matrix), $p$ is the number of items, and $a$ is the freely estimated parameters that must be calculated manually per model. These parameters are:

1. Factor loadings (FL)
2. Error variances
3. Factor variances
4. Factor covariances

# CFA

Calculate *df* for **Importance**:

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

$$a = 2(FL) + 3(Error\ \ VAR) + 1(Factor\ \ VAR) + 0(Factor\ \ COV) = 6$$

One less FL because one of the FL is fixed to 1 as marker indicator.

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

Just identified! Not good for CFA model as we cannot get model fit.

# CFA

Calculate *df* for **Affinity**:

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

$$a = 4(FL) + 5(Error\ \ VAR) + 1(Factor\ \ VAR) + 0(Factor\ \ COV) = 10$$

$$df = b - a = 15 - 10 = 5$$

Overidentified and ready for CFA.

# CFA

Now can you calculate for two-factor model obtained from EFA? i.e. 2 factors: AFFINITY, IMPORTANCE; 8 items; 1 between factor correlation.

$$b = ?$$

$$a = ?(FL) + ?(Error\ VAR) + ?(Factor\ VAR) + ?(Factor\ COV) = ?$$

$$df = b - a = ?$$

# CFA - Maximum likelihood estimation*

The most commonly used estimation method in CFA, but it needs multivariate normal data.

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, $trace(I) - p = 0$.

# Analysis Steps in CFA

# Preliminary steps

1. Descriptive statistics
2. Multivariate normality

If the **data are normally distributed**, we may use **maximum likelihood (ML)** estimation method for the CFA.

If the **data are not normally distibuted**, two common alternatives are:

1. **MLR** (robust ML), suitable for complete and incomplete, non-normal data (Rosseel, Jorgensen, & Rockwood, 2023).
2. **WLSMV** (robust weighted least squares), suitable for categorical response options (e.g. dichotomous, polynomous, ordinal (Brown, 2015))

# Step 1: Specify the measurement model

Specify the measurement model according to `lavaan` syntax:

```
model = "
FACTOR1 =~ Q1 + Q2 + Q3
FACTOR2 =~ Q4 + Q5 + Q6
"
```

# Step 2: Fit the model

Fit the specified model.

By default, the *marker indicator variable* approach is used in `lavaan` to scale a factor (item coefficient set to 1).

May also set to scale a factor by *fixing the factor variance* to 1.

# Step 2: Results

To interpret the results, we must looks at

1. Overall model fit - by fit indices
2. Localized areas of misfit
   - Residuals
   - Modification indices
3. Parameter estimates
   - Factor loadings
   - Factor correlations

# 1. Fit indices.

The following are a number of selected fit indices and the recommended cut-off values:

| Category | Fit index | Cut-off |
|---|---|---|
| Absolute fit | $\chi^2$ | $P$>0.05 |
| | Standardized root mean square (SRMR) | $\leq 0.08$ |
| Parsimony correction | Root mean square error of approximation (RMSEA) | and its 90% CI $\leq 0.08$,<br>CFit $P > 0.05$ (H0 $\leq 0.05$) |
| Comparative fit | Comparative fit index (CFI)<br>Tucker-Lewis index (TLI) | $\geq 0.95$ |

# 2. Localized areas of misfit

- Residuals
  - Residuals are the difference between the values in the sample and model-implied variance-covariance matrices.
  - Standardized residuals (SRs) $> |2.58|$ indicate the standardized discrepancy between the matrices.

# 2. Localized areas of misfit

- Modification indices (MIs)
  - A modification index indicates the expected parameter change if we include a particular specification in the model (i.e. a constrained/fixed parameter is freely estimated, e.g. by correlating between errors of Q1 and Q2).
  - Specifications with MIs $> |3.84|$ should be investigated.

# 3. Parameter estimates

- Factor loadings (FLs) (`Std.all` column under `Latent Variables` table).
  - FLs $\geq 0.5$ are practically significant. In addition, the *P*-values of the FLs must be significant (at $\alpha = 0.05$).
  - Look for out-of-range values. FLs should be in range of 0 to 1 (absolute values), thus values $> 1$ are called *Heywood cases* or *offending estimates*.

# 3. Parameter estimates

- Factor correlations
  - Factor correlation must be $< 0.85$, which indicates that the factors are distinct.
  - Correlation $> 0.85$ indicates multicollinearity problem.
  - Also look for out-of-range values. Factor correlations should be in range of 0 to 1 (absolute values).
- When a model has Heywood cases, the solution is *not acceptable*. The variance-covariance matrix (of our data) could be *non-positive definite* i.e. the matrix is not invertible for the analysis.

# Step 3: Model revision

Model does not fit well? Revise the model.

The causes of poor model fit in CFA could be:

1. Item – the item has low FL ($< 0.3$), is specified to load on wrong factor or has cross-loading issue.
2. Factor – the factors have multicollinearity problem (correlation $>$ 0.85), or the presence of redundant factors in a model. This can detected by residuals and MIs.

# Step 3: Model revision

3. Correlated error (method effect) – some items are similarly worded (e.g. "I like ...", "I believe...") or have almost similar meaning/content. This is usually detected by residuals and MIs.
4. Improper solution – the solution with Heywood cases. It could be because the specified model is not supported by the data and the misspecification could be a combination of all the first three causes listed above. A small sample may also lead to improper solution.

The problems might not surface if a proper EFA is done in the first place and the model is theoretically sound.

# Step 3: Model revision

Model-to-model comparison following revision is done based on:

1. $\chi^2$ difference
   - for nested[1] models only.
2. AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion)
   - for nested and unnested models.
   - an improvement in the model is shown as a reduction in AIC and BIC values. Better model = Smaller AIC/BIC.

[1] model with same number of items, but with different model specifications e.g. number of factors

# Composite reliability

# Composite reliability: Omega

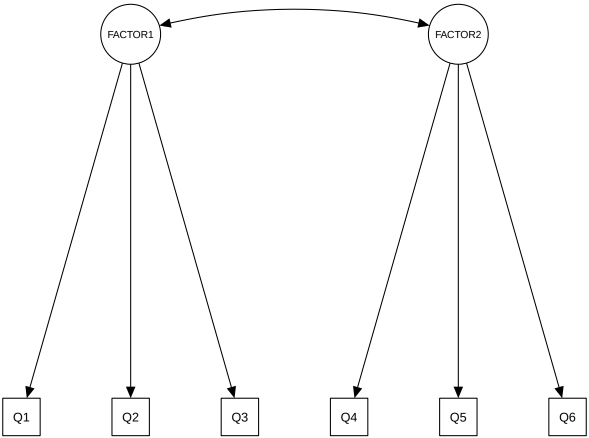
**Omega $\omega$ coefficient**

- One of the reliability indices applicable to CFA
- It takes into account correlated errors
- Construct reliability $\geq 0.7$ (Hair, Black, Babin, & Anderson, 2010) is acceptable

# Path diagram

# Path diagram

A CFA model can be nicely presented in the form of a path diagram



# Results presentation

# Results presentation

In the report, you must include a number of important statements and results pertaining to the CFA,

1. The **estimation method** e.g. ML, MLR, WLSMV etc.
2. The **model specification** and the **theoretical background** supporting the model.
3. **Details** about the selected fit indices, residuals, MIs, FLs and factor correlations and the accepted cut-off values.
4. **Detailed comments** on the fit and parameters of the tested models. This is usually done in reference to summary tables.

# Results presentation

5. Details about the **revision process**, i.e. item deletion, addition of correlated errors or any other modifications and the effects on the model fit. Also mention the reasons e.g. high SRs, low FIs etc.
6. **Summary tables**, which outlines the model fit indices, model comparison, FLs, reliability, and factor correlations.
7. The **path diagram** (most of the time, of the final model). This may be requested by some journals.

# References


Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. New Jersey: Prentice Hall.

Rosseel, Y., Jorgensen, T. D., & Rockwood, N. (2023). *Lavaan: Latent variable analysis*. Retrieved from https://lavaan.ugent.be