Exploratory factor analysis

Dr Wan Nor Arifin
Unit of Biostatistics and Research Methodology, Universiti Sains Malaysia.
wnarifin@usm.my

Last update: 11 December, 2018

Outlines

Introduction

Factoring

- Group things that have common concept.
- Simplify long list of items/variables into smaller groups.
- **Factoring = Grouping**.
- **Factor = Construct = Concept.**

Intuitive factoring

List of items

Orange, motorcycle, bus, durian, banana, car

Do these items have anything in common?

Group the items

[Orange, durian, banana]

[Motorcycle, bus, car]

Name the groups

Fruit	***Motor vehicle***
Orange, durian, banana	**Motorcycle, bus, car**

- By finding something in common among the items, factoring the items and naming the factors are basically factor analysis!
- Factor out the common comcepts from the items.

Correlation matrix

- Let say the same items are rated on a Likert-type options from 1 (fruit) to 5 (motor vehicle) on their characteristics of being fruit or motor vehicle. Then the Pearson's correlation coefficients among the items are tabulated:

Items	**1**	**2**	**3**	**4**	**5**	**6**
1. Orange	1.00					
2. Durian	***.67***	1.00				
3. Banana	***.70***	***.81***	1.00			
4. Motorcycle	.11	.08	.05	1.00		
5. Bus	.08	.12	.09	***.75***	1.00	
6. Car	.18	.12	.22	***.89***	***.83***	1.00

- We then examine the patterns of correlation in the correlation matrix, then group highly correlated items into factors.

	Factors	
Items	***Fruit***	***Motor vehicle***
1. Orange	X	-
2. Durian	X	-
3. Banana	X	-
4. Motorcycle	-	X
5. Bus	-	X
6. Car	-	X

- However such approach is tedious for large number of items, for example for 100 items, we have to examine 100(100-1)/2 = 4950 correlations.

- Factor analysis enables objective assessment of these correlations and factor/group the items.

Factor analysis

Introduction

- A multivariate statistical analysis i.e. many outcomes.
- It refers to a mathematical method known as **multivariate linear factor model** (Gorsuch, 2014).
- A member of an analysis group known as latent variable model analysis (Bartholomew et al., 2008)
- The aim is to determine of number and nature of factors that are responsible for the **correlations** among the items (Brown, 2006).
- From a **number** of outcomes, factors are extracted and determined. These factors are **unobserved (latent)** independent factors.
- In contrast to multiple linear regression, the **one** outcome and **many** independent factors are measurable.
- By comparing the equations:

  Simple linear regression:

  $$y = a + bx$$

  Multiple linear regression:

  $$y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

  Factor analysis:

  Still: $y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$

  Written in different way:

  $$X_{i1} = w_{1A}A_i + w_{1B}B_i + \ldots + w_{1f}F_i + c$$

  In form of multivariate linear factor model:

  $$X_{i1} = w_{1A}A_i + w_{1B}B_i + \ldots + w_{1f}F_i$$
  $$X_{i2} = w_{2A}A_i + w_{2B}B_i + \ldots + w_{2f}F_i$$
  $$. . .$$
  $$X_{iv} = w_{vA}A_i + w_{vB}B_i + \ldots + w_{vf}F_i$$

  X_{iv} : Item ***v*** score for person ***i***
  W_{vf} : Factor weight/coefficient for item ***v***
  A_i to F_i : Factor score for person ***i***

* Constant, ***c*** is dropped as all scores are deviations from mean.

In a more human friendly form:

Item score = Factor Weight x Factor score

- The analysis can be (Brown, 2006):
  - Exploratory – Exploratory Factor Analysis (EFA).
  - Confirmatory – Confirmatory Factor Analysis (CFA).

- Analysis of latent variable such as factor analysis is important in fields like psychology and psychiatry, because we cannot observe directly psychological states, thus measured indirectly in form items, e.g. depression:
  - depression causes symptoms of depression.
  - depression (latent) is measured indirectly by items representing its symptoms.
  - prove the symptoms are correlated to each other, representing the concept of depression by factor analysis.

Exploratory factor analysis (EFA)

Introduction

- An exploratory method.
- Aims to explore the items, factor common concepts and generate theory.
- Generally two models (Gorsuch, 2014):
  - **Full Component Model.**
  - **Common Factor Model.**

- The choice of models determines the extraction methods.

Full Component Model

Item = (Weight 1 x Factor 1) + (Weight 2 x Factor 2) + … + (Weight n x Factor n)

- Extraction method: Principal component analysis (PCA)
- Takes into account for all variances, suitable for data reduction, e.g. items are condensed into smaller number of unrelated components, then used as variables in other statistical analysis.
- Do not account for **error** in measurement.
- Not the 'real' factor analysis (Gorsuch, 2014; Brown, 2006).
- Advantage: No problem with inability to come up with factor solution (indeterminate factor solution).
- Basically a descriptive method.

Common factor model

Item = (Weight 1 x Factor 1) + (Weight 2 x Factor 2) + … + (Weight n x Factor n) + ***Error***

- Extraction methods:

    - Classical: **Principal axis analysis**.
    - Other variants: Image analysis, alpha analysis, maximum likelihood.

- Attempts to account for **common** variances and also **error** variances.
- 'Real' factor analysis.
- Maximum likelihood variant allows assessment of factor model fit (chi-square).
- Problem – Indeterminate factor solution.

Rotation

- Rotation of factors is used to allow simpler analysis solution.
- Types of factor rotation:

    - **Orthogonal method – uncorrelated factors.**

        - **Varimax**, Quartimax, Equamax.

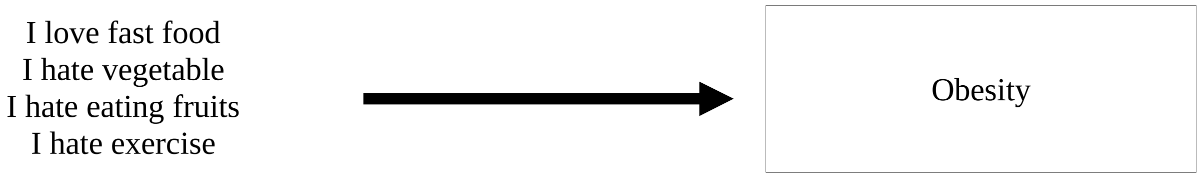
    - **Oblique method – correlated factors.**

        - **Oblimin,** Promax.

Confirmatory factor analysis (CFA)

Introduction

- A confirmatory method.
- It is also based on common factor model.
- A type of Structural Equation Modeling (SEM) analysis that deals with **measurement model**.
- Maximum likelihood estimation is commonly used for estimation.
- Allows assessment of measurement model fit.
- The main difference between EFA and CFA is that by using CFA, the researcher has already established the construct and which items belong to it. CFA is no longer exploratory.
- For example, CFA items:

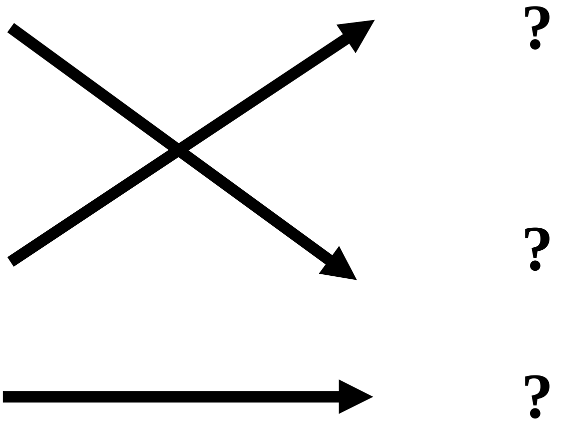


I love fast food
I hate vegetable
I hate eating fruits
I hate exercise

Obesity

- The items are probably based on his exploratory method, literature reviews, theories, or experience – strong theoretical basis for the items and factors.
- For example, EFA items:

I love cat
I hate snake
I love statistics
I love snorkelling
I support Harimau Malaya team
I love driving car
I love computer game
I like to have everything normally distributed
I think of independent t-test everyday
My favourite food is nasi ayam
I used to eat a lot of pisang goreng
I spend most of my time in front of computer
I love SPSS



- Can you explain easily the correlations between the items? No idea → EFA.

EFA vs CFA

- The differences between EFA and CFA can be summarized in the table below:

EFA	**CFA**
Exploratory procedure.	Confirmatory procedure.
No pre-requisite to specify theoretical factors for a collections of items.	Pre-specified theoretical factors.
Aims to explore the items and extract common ideas. Theory generating based on empirical findings.	Strong theory. Just want to confirm.
Items free loading and not fixed to factors.	Items are fixed to pre-specified factors.
Rotation of factors is used to allow simpler solution.	Rotation not used.
Explicit hypothesis is not tested.	Explicit hypothesis testing. Allows assessment of model fit (X^2 GOF, Fit indices).

Analysis steps in EFA

The following steps allow systematic approach to EFA.

Preliminary step

1. Clean up the data for wrong entry, missing values. Replace missing values with appropriate imputation method of choice.
2. Descriptive statistics:

    - Check minimum-maximum values per item.
    - n(%) of response to options per item.

3. Normality of data:

    - Univariate normality

        - Maximum-Likelihood extraction requires multivariate normality.
        - Univariate normality → Multivariate normality.
        - If not normal, may use Principal Axis extraction.

    - Multivariate normality

        - Normality of the data at multivariate level.

Step 1

- Check suitability of data for analysis

    - Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy.
    - Bartlet's test of sphericity.

- Determine the number of factors by

    - Eigenvalues.
    - Scree plot.
    - Parallel analysis.

Assessment of results for Step 1

Result	**Cut-off points**	**Comments**
Suitability of data for analysis		
Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy	> 0.7	MSA is a relative measure of amount of correlation (Kaiser, 1970). It indicates whether it is worthwhile to analyze a correlation matrix or not. KMO is an overall measure of MSA for a set of items, given as: $\text{KMO}=\frac{\sum_{i\neq j}^{n}\sum_{i\neq j}^{n} r_{ij}^2}{\sum_{i\neq j}^{n}\sum_{i\neq j}^{n} r_{ij}^2+\sum_{i\neq j}^{n}\sum_{i\neq j}^{n} a_{ij}^2}$ where r_{ij} is the correlation between items *i* and *j* a_{ij} is the partial correlation coefficient (or anti-image correlation coefficient) between items *i* and *j*

		From the formula, we can imply that KMO → 1: Correlation → 1 and partial correlation → 0. KMO → 1: Correlation → 0 and partial correlation → 1. The following is the guideline on interpreting KMO values (Kaiser & Rice, 1974): < 0.5 – Unacceptable 0.5 – 0.59 – Miserable 0.6 – 0.69 – Mediocre 0.7 – 0.79 – Middling 0.8 – 0.89 – Meritorious 0.9 – 1.00 – Marvelous
Bartlet's test of sphericity	*P*-value < 0.05	Basically it tests whether the correlation matrix is an identity matrix (Bartlett, 1950; Gorsuch, 2014; Revelle, 2015). The determinant of the matrix, R_{vv} is converted to a chi-square statistic and tested for significance: $$\chi^2 = -\left(n-1-\frac{2v+5}{6}\right)\ln\lvert R_{vv}\rvert$$ where *n* is the sample size *v* is the number of items while the *df* for the χ^2 is $$df = v\frac{v-1}{2}$$ A significant test indicates that there are worthwhile correlations among the items based on correlation matrix. A non-significant test indicates that the items are not correlated to each other based on the correlation matrix.
Determination of the number of factors		
Eigenvalues	> 1	Look at number of factors at eigenvalues > 1 (Kaiser-Guttman rule). Eigenvalues can be interpreted as how worthwhile a factor in term of item. For an Eigenvalues of 4.5, the extracted factor is worth 4.5 times as much as a single variable. The cut-off value is 1 because if extracted factor is worth less than what a single variable can explain, the factor is not

		worthwhile to be extracted.
Scree plot	–	Cattel’s scree test. "Scree" is a collection of loose stones at the base of a hill. This test is based on eye-ball judgement of an eigenvalues vs number of factors plot. Look for the number of eigenvalue points/factors before we reach the "scree". Look for last substantial decline or abrupt changes in the plot (elbow). Number of factors is the number of dots (eigenvalues) up to the 'elbow' of the plot. It is also suggested to to fix +/- 1 factor from the decided number of factor.
Parallel analysis	–	Comparison of the scree plot obtained from the data to the scree plot obtained from randomly generated data (Brown, 2006). Number of factors is the number of dots above the intersection between the plots.
Very simple structure (VSS) criterion	–	VSS compares the original correlation matrix to a simplified correlation matrix [Revelle, 2015]. Look for the highest VSS value at complexity 1 i.e. an item loads only on one factor.
Velicer's minimum average partial (MAP) criterion.	–	MAP criterion indicates the optimum number of factors that minimizes the MAP value. The procedure extracts the correlations explained by the factors, leaving only minimum correlations unrelated to the factors.

Step 2

- Run the analysis by fixing number of factors as decided from previous step.
- Choose an appropriate extraction method. We use **principal axis factoring** (PAF).
- Decide on rotation method. Choose an oblique rotation, **Oblimin**.

Assessment of results for Step 2

Result	**Cut-off points**	**Comments**
Judge quality of items by looking at the following results. Remove poor quality items.		
Factor loadings (pattern coefficients)	Ideally > 0.5	Only available when oblique rotation is used. Usually the pattern coefficients in the matrix are interpreted similarly to factor loadings. The coefficients are partials correlation coefficients of factors to the item. Factor loadings can be interpreted as follows (Hair Jr. et al., 2009):

		0.3 to 0.4 – Minimally acceptable ≥ 0.5 – Practically significant ≥ 0.7 – Well-defined structure The factor loadings are interpreted based on absolute values, ignoring the +/- signs. We may need to remove items based on this assessment. Usually we may remove items with FLs < 0.3 (or < 0.4, or < 0.5). But the decision depends on whether we want to set a strict or lenient cut-off value. Also check for cross-loading problem of an item across factors. This problem is indicated by having almost comparable factor loadings in two or more factors. It indicates that the item is not specific for a construct and to general, thus should be removed.
Communalities (Extraction)	Ideally > 0.5 Practically > 0.25	It is the % of item variance explained by the extracted factors. A cut-off of 0.5 is practical (Hair Jr. et. at., 2009), which means that 50% of item variance is explained by all extracted factors. The cut-off value depends on researcher as to what amount of explained variance is acceptable to him/her. However, for practical purpose I consider 0.25 cut-off point, considering factor loading > 0.5 is accepted, thus variance = square of factor loading = 0.5^2 = 0.25
Factor correlations	< 0.85	Only available when oblique rotation is used. If > 0.85, the is a multicollinearity between the factors, thus the factors are not distinct from each other, thus can be combined (change number of fixed factors) (Brown, 2006).

Step 3

- Repeat the analysis similar to **Step 2** every time an item is removed. Make judgment based on the results.
- The analysis is finished once we have:
  - satisfactory number of factors.
  - satisfactory item quality.

References

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2009). *Analysis of multivariate social science data (2nd eds.)*. Boca Raton, Florida: Chapman & Hall/CRC.

Bartlett, M. S., (1951), The effect of standardization on a chi square approximation in factor analysis. *Biometrika*, 38, 337-344.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford

Press.

Gorsuch, R. L. (2014). *Factor analysis: Classic edition*. New York: Rouledge.

Hair Jr., J. F., Black, W. C., Babin, B. J. & Anderson, R. E. (2009). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson Prentice-Hall.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, 35(4), 401-415.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, Mark IV. *Educational and Psychological Measurement*, 34, 111-117.

Revelle, W. (2015). *psych: Procedures for Personality and Psychological Research*. Evanston, Illinois, USA: Northwestern University. http://CRAN.R-project.org/package=psych Version = 1.5.8.