QTM 3610: Exploratory Factor Analysis

Eric Chan

Fall 2020

learning objectives

- What is Factor Analysis?
- How does it differ from PCA?

What is factor analysis?

- The goal of factor analysis is dimension reduction
- where a set of variables are combined into one more more "factors"
- it attempts to explain the SHARED variance.
- Under many conditions, FA and PCA can yield similar results.

What is the difference between FA AND PCA?

Difference Number 1:

- In FA, the observed variables are expressed as a linear combination of the factors
- In PCA, the components are linear combinations of the observed variables. (Remember, PC scores)

Difference Number 2:

- FA seeks to account for the covariances (or correlations) among observed variables
- PCA seeks to explain the total variance in the observed variables through the number of components

Difference Number 3:

- FA scores are computed using factor loadings
- PC scores are computed using observed variable weights.

Difference Number 4:

- It is generally easier to interpret FA than PCA, but still very subjective!

What is the PRACTICAL difference between FA AND PCA?

In terms of a simple rule of thumb, I'd suggest that you:

- Run factor analysis if you assume or wish to test a theoretical model of latent factors causing observed variables.
  - A latent factor is some underlying variable that is not directly observed.
- Run principal component analysis If you want to simply reduce your correlated observed variables to a smaller set of important independent composite variables.

The underlying philosophy of factor analysis is that the observed variables are themselves related to unobserved factors (or latent variables) that would better explain the data. Examples of this approach:

- We can classify the teaching evaluation questions into categories that make sense to you.
- A supermarket chain might combing observed purchases from categories of variables into attributes of shoppers. (Young families, empty-nesters, single adult, etc.)
- A factor analysis of places rated might aim to combine variables like arts, education and recreation into a culture factor.

Assumptions of FA

To conduct a Factor Analysis, we need to test for:

- Sphericity. Does sufficient correlation exists in the correlation matrix to continue?
  - This is done using the Bartlett test. If the test is statistically significnat, we can proceed with FA. We cannot do FA without significance here. This is a REQUIRED assumption.
- Sampling Adequacy
  - We can use the KMO Test, but one rule of thumb is that we want 20 observations per variable as a minimum. We want the KMO test, which is from 0 to 1, to be as close to 1 as possible.
- Missing Data
  - If more than 10% of data is missing from a variable, best to either impute or not use factor analysis.
- Determinant must be positive

Interpretation of FA

Your output will contain factor LOADINGS. For example:

```
Factor Analysis using method =  minres
Call: fa(r = carpurchase, nfactors = 4, rotate = "oblimin", fm = "minres")
Standardized loadings (pattern matrix) based upon correlation matrix
                     MR1   MR2   MR4   MR3   h2   u2 com
Price                0.54  0.12 -0.08 -0.05 0.30 0.70 1.2
Safety              -0.33  0.36  0.11 -0.25 0.23 0.77 3.0
Exterior_Looks       0.12  0.07 -0.55  0.28 0.34 0.66 1.6
Space_comfort       -0.04  0.78 -0.13  0.08 0.67 0.33 1.1
Technology           0.01  0.36  0.06  0.02 0.13 0.87 1.1
After_Sales_Service  0.10  0.54  0.18 -0.05 0.33 0.67 1.3
Resale_Value         0.73 -0.14 -0.02 -0.16 0.57 0.43 1.2
Fuel_Type            0.01  0.57 -0.01 -0.12 0.32 0.68 1.1
Fuel_Efficiency      0.43  0.23  0.31  0.16 0.47 0.53 2.7
Color                0.07 -0.06  0.73  0.13 0.61 0.39 1.1
Maintenance          0.56  0.08  0.21  0.01 0.44 0.56 1.3
Test_drive           0.11  0.16  0.01  0.36 0.20 0.80 1.6
Product_reviews      0.34  0.15  0.05  0.36 0.32 0.68 2.4
Testimonials        -0.19 -0.01  0.06  0.68 0.51 0.49 1.2
```

These loadings can be interpreted as the correlation between the original variables and components/factors. Given you know the meaning of the variables you have used , you will look at the variables highly loaded (as absolute value) on the components/factors (in this case the loading can only vary between -1 and 1 so the interpretation is easy) and give a name to the components depending on which variables are more weighted on the components.