Factor Analysis

Alexandros Korolis

July 2025

Introduction

What is factor analysis?

Factor analysis is a statistical method that aims to create new latent variables (factors) from the existing ones in a dataset, based on a somewhat subjective interpretation of their underlying relationships. By using these factors, we can reduce the dimensionality of the dataset and better explain the correlations among the original variables. Factor analysis differs from principal component analysis in that it aims to explain the underlying structure of the data rather than merely accounting for its variability. In factor analysis, there is not always a unique solution. Different estimation methods, or even the same method applied in slightly different ways, can lead to different results. Moreover, the extracted factors can have varying interpretations, which may even contradict each other.

Orthogonal Factor Model

In the orthogonal factor model, we assume that there are some underlying common factors, initially unknown,that we aim to estimate and that are responsible for the correlations among the observed variables. In this manner, if there are pp observed variables, each can be expressed as a linear combination of kk underlying factors, such that:

$$X - \mu = LF + \epsilon \ (1)$$

where: X : a vector of observed variables, with dimension px1
μ: a vector of means of each variable, with dimension px1
L : a matrix with dimension pxk, where L_{ij} is the loading of factor F_j in variable X_i
F : a kx1 matrix consisting of factors
ϵ : an error. ϵ_i represents the portion of the i-th observed variable that cannot be explained by the factors.

Assume that each variable has a mean of zero, and that k<p then :

$$\begin{aligned} X_1 &= L_{11}F_1 + L_{12}F_2 + ... + L_{1k}F_k + \epsilon_1 \\ X_2 &= L_{21}F_1 + L_{22}F_2 + ... + L_{2k}F_k + \epsilon_2 \\ &..... \\ X_p &= L_{p1}F_1 + L_{p2}F_2 + ... + L_{pk}F_k + \epsilon_1 \end{aligned}$$

Note that the factors can also be written as a linear combination of the observed variables, which is particularly useful when creating new variables. In this context, the coefficients are called factor score coefficients, which are distinct from the factor loadings.

Assumptions of Orthogonal Factor Model

1. E(F)=0, each factor has a mean value of zero
2. Cov(F) = I, there is no correlation between each factor (Orthogonal Factors)
3. E(ϵ) = 0, each error term has a mean value of zero
4. Cov($\epsilon = \Psi$,

where $\Psi = \begin{pmatrix} \Psi_1 & 0 & .. & 0 \\ 0 & \Psi_2 & & \\ 0 & .. & .. & 0 \\ 0 & .. & .. & \Psi_p \end{pmatrix}$,

meaning that there is no correlation between the error terms.
5. $cov(\epsilon_i, F_j) = 0$ for i different than j, error terms and factor are unrelated.
6. If method of estimation is maximum likelihood, the we assume that the data follow a multivariate normal distribution.

Communality-Specificity

Based on the assumption of the orthogonal model, observe that: $\Sigma = cov(X) = cov(LF + \epsilon) = Lcov(F)L^T + cov(\epsilon) = LL^T + \Psi$

This means that the variance-covariance matrix consists of two components: the first corresponds to the common factors and is referred to as communality, while the second represents the portion that the model cannot explain and is known as specificity.

Loading refers to the correlation between each observed variable and each underlying factor.

Factor Analysis Steps

1. Examine the adequacy of correlations among variables in order to apply factor analysis.

2. Determine the appropriate number of factors and estimate the model parameters.

3. Examine the rotation method to enhance interpretability.

First Step - Examine Correlations among variables

We are interested in variables with absolute correlation values higher than 0.4. If this condition is not met, those variables should be excluded, as they are likely to form a standalone factor on their own.

- Bartlett's test of sphericity tests whether the population covariance matrix is diagonal. A significant result indicates that correlations exist among the

variables, justifying the use of factor analysis.

$$H_0 : \Sigma = \sigma^2 I_p$$
$$vs$$
$$H_1 : \Sigma \neq \sigma^2 I_p$$

Test statistic:

$$L = -[n - \frac{1}{6p}(2p^2 + p + 2)][ln|S| - ln(\prod_{i=1}^{p} s_i^2)]$$

Critical value:

$$c = x^2_{\frac{p(p-1)}{2}}$$

Where
S : sample variance-covariance matrix
s_i^2: sample variance of i-th variable

- Partial correlation measures the relationship between two variables while removing the effect of one or more other variables. It helps avoid misleading results that may occur when a third variable affects both variables of interest.

  The Kaiser-Meyer-Olkin (KMO) statistic measures the adequacy of sampling by comparing the magnitudes of observed correlation coefficients to the magnitudes of partial correlation coefficients. If partial correlations are small compared to regular correlations then variables share common factors and factor analysis is appropriate

$$KMO = \frac{\sum \sum_{i \neq j} r_{ij}^2}{\sum \sum_{i \neq j} r_{ij}^2 + \sum \sum_{i \neq j} \alpha_{ij}^2}$$

  In practice, KMO values higher than 0.8 suggest that the dataset is suitable for factor analysis.

- The measure of sampling adequency is calculated for each variable:

$$MSA_i = \frac{\sum_j r_{ij}^2}{\sum_j r_{ij}^2 + \sum_j \alpha_{ij}^2}$$

indicates to what extent a variable is suitable for a factor analysis. Values near 1 indicate that the i-th variable can be used for factor analysis.

Second Step - Number of factors, parameters estimation

- Number of factors: To determine the number of factors, one can use techniques similar to those used in Principal Component Analysis, such as the scree plot and the eigenvalues of the variance-covariance matrix. The number of factors must be decided before estimating the model's parameters. Moreover, one can increase the number of factors iteratively and determine the optimal number by comparing the factor covariance matrix (based on loadings) with the sample covariance matrix.

- Model's parameters estimation using Principal Components: We are interested in estimating $\hat{L}$ and $\hat{\Psi}$ such as that $\hat{L}\hat{L}^T + \hat{\Psi}$ closely approximating sample covariance matrix. From spectrum analysis of a covariance matrix, $\Sigma = AA^T$, where $A = \Pi\Lambda^{1/2}$, L : diagonal matrix with eigen values of Σ and Pi: containes eigen vectors of Σ. Therefore, if we choose $\hat{L} = \Pi\Lambda^{1/2}$ we can obtain a complete representation of Σ matrix. In practise, we use the sampling S covariance matrix.
  If the number of factors k is equal to the number of variables then we can completely represent the sample covariance matrix S and therefor specificities estimates $\hat{\psi}_i = 0$.
  If the number of factors k is less than the number of variables then we can estimate specificities:

$$\hat{\psi} = s_i^2 - \sum_{j=1}^{p} L_{ij}^2$$

where the sum is called communality and L_{ij} is the loading of the j-th factor to the i-th variable.

Third Step - Rotation

By using factor rotation, we aim to enhance interpretability.
Suppose that L: is a matrix that contains factor loadings and G: an orthogonal matrix $GG^T = I$ then $(LG)(LG)^T = (LG)(G^T L^T) = LL^T$. Therefor, LG can be considered as a factor loadings matrix.
One method of rotation is Varimax. Varimax rotation aims to reduce the number of variables with high loadings on each factor.

Confirmatory Factor Analysis

In confirmatory factor analysis, we are interested in testing specific assumptions based on established theories, for example in psychology or sociology.

Dataset

Background

The big five personality traits are the best accepted and most commonly used model of personality in academic psychology. If you take a college course in personality psychology, this is what you will learn about. The big five come from the statistical study of responses to personality items. Using a technique called factor analysis researchers can look at the responses of people to hundreds of personality items and ask the question "what is the best was to summarize an individual?". This has been done with many samples from all over the world and the general result is that, while there seem to be unlimited personality variables, five stand out from the pack in terms of explaining a lot of a persons answers to questions about their personality: extraversion, neuroticism, agreeableness, conscientiousness and openness to experience. The big-five are not associated with any particular test, a variety of measures have been developed to measure them. This test uses the Big-Five Factor Markers from the International Personality Item Pool, developed by Goldberg (Goldberg, Lewis R. "The development of markers for the Big-Five factor structure." Psychological assessment 4.1 (1992): 26).

Descriptive Statistics

The dataset consisted of 50 likert rated statements, race, age, native speaker, gender, hand (right handed or left handed), source (method of taking the test, online etc) and coountry. There were 19627 observations and 57 variables in total after removing non available observations and observations with age higher than 100 years old.

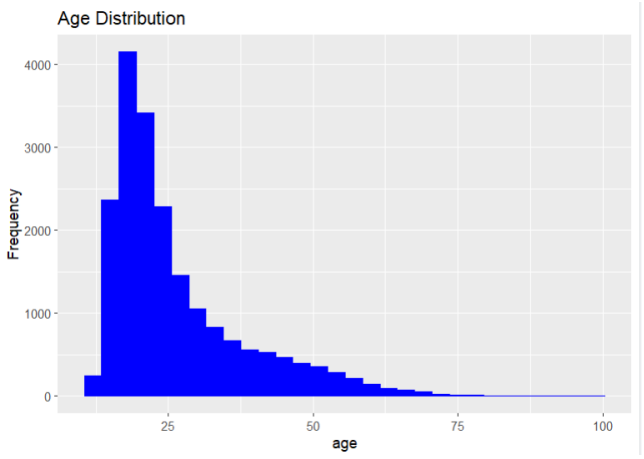

Figure 1: Age Distribution

In Figure 1, we observe the distribution of age. The mean age in the dataset is 26.26 years, and 50% of the observations are below 22 years old. The distribution is positively skewed, with a skewness coefficient of 1.49.

The average age of women is 26 years, of men is 26.7 years, and of other genders is 23.4 years.

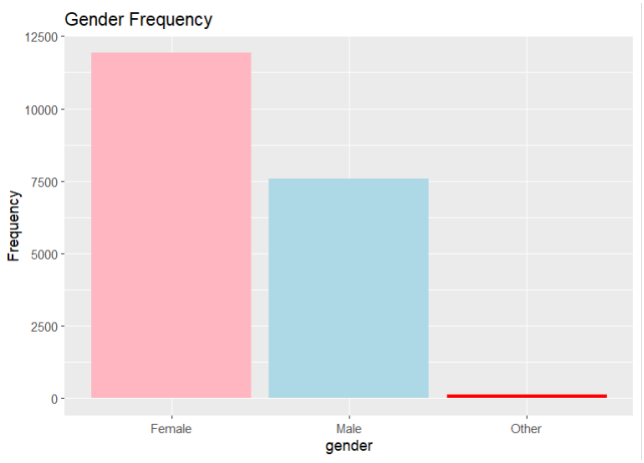

Figure 2: Genders

In Figure 2, we observe the total number for each gender. The total number of woman is 11922, of men is 7583 and of other is 122.

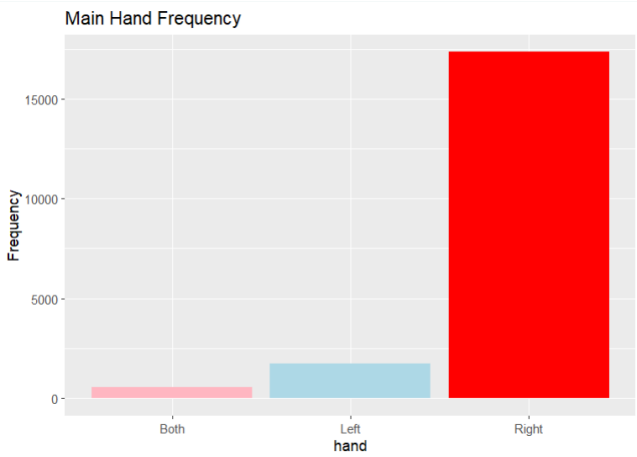

Figure 3: Main Hand

In Figure 3, we observe the frequency of individuals who are right-handed, left-handed, or ambidextrous. There are 17,343 right-handed individuals, 1,721 left-handed, and 563 who use both hands.

The dataset includes individuals from various ethnic backgrounds. The majority identify as Caucasian (European), with 10,527 observations. This is followed by individuals categorized as 'Other' (2,683), South East Asian (1,835), Caucasian (Indian) with 1,505 entries, and Mixed Race with 1,429 individuals. This distribution highlights a predominantly European representation, with notable diversity across other groups.

The dataset represents individuals from various countries, with the United States having the largest representation at 8,732 entries. The United Kingdom follows with 1,529, India with 1,456, Australia with 973, and Canada with 922. This shows a predominance of US-based individuals, with other countries contributing a smaller proportion to the dataset.

Step 1: Correlations

In order to implement confirmatory factor analysis, we split the data into two sets: demographics and questionnaire responses. We used only the questionnaire responses, which consisted of 50 variables and 19,627 observations.

We observed that the determinant of the covariance matrix is greater than 0.00001 (0.0683), which suggests that multicollinearity is not a serious concern and that the data may be suitable for factor analysis.

Based on Bartlett's Test of Sphericity, we reject the null hypothesis that the population covariance matrix is diagonal at the significance level $\alpha = 0.05$ (p-value 0.000).

MSA for each item is bigger than 0.6 and overall is MSA = 0.91 which suggest that all variables are suitable for factor analysis.

Step 2: Number of factors, parameters estimation