UNIVERSITY OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY

MULTIVARIATE STATISTICAL ANALYSIS

Report Practice 04
Factor Analysis

Lecturers Lý Quốc Ngọc
Nguyễn Mạnh Hùng
Phạm Thanh Tùng

Student Võ Nguyễn Hoàng Kim
21127090

CONTENTS

I. SELF-ASSESSMENT FORM

Features		Level of completion
Source	Implement code	100%
Requests	Request 1	100%
	Request 2	100%
	Request 3	100%
Research questions	Question 1	100%
	Question 2	100%
	Question 3	100%
	Question 4	100%

II. REQUESTS

1. Data

- Download: https://vincentarelbundock.github.io/Rdatasets/csv/psych/bfi.csv
- The BFI dataset based on personality assessment project, which was collected using a 6-point response scale: 1 Very Inaccurate, 2 Moderately Inaccurate, 3 Slightly Inaccurate 4 Slightly Accurate, 5 Moderately Accurate, and 6 Very Accurate.

2. Request 1

Student explain the meaning of chi_square_value, p_value, KMO values

a. Chi_square_value and p_value

- In factor analysis, Bartlett's test of sphericity is used to test the hypothesis that the correlation matrix of variables is an identity matrix (all elements on the main diagonal are equal to 1, and all elements off the main diagonal are equal to 0). The chi-square value (chi_square_value) and the p-value (p_value) calculated from Bartlett's Sphericity test are two important statistical indices in factor analysis for assessing the sphericity of the correlation matrix.
- If the chi-square value is large and the p-value is small (usually < 0.05), we have enough evidence to reject the hypothesis that the correlation matrix is not spheric, meaning the data are structured and suitable for factor analysis.
- Conversely, if the chi-square value is small and the p-value is large (usually > 0.05), there is not enough evidence to reject the hypothesis that the correlation matrix is spheric, implying that the data may lack structure and may not be suitable for factor analysis.
- In this source code, with a chi-square value of 18184.306307820785 and a p-value of 0, it indicates that this is not a spherical matrix (which can be verified based on the correlation matrix represented from the dataset), meaning the data are structured.

b. KMO values

- The KMO test (Kaiser-Meyer-Olkin test) evaluates the adequacy of data for factor analysis by assessing the degree of coherence among variables. The test score ranges from 0 to 1, where values above 0.5 are generally considered suitable for factor analysis [1].
- Typically, KMO test values above 0.6 are considered acceptable for analysis, values above 0.7 indicate good suitability, values above 0.8 indicate very good suitability, and values above 0.9 indicate excellent suitability (Kaiser and Meyer, 1974).
- In this source code, the KMO test yields a result of 0.84, indicating that the data are well-suited for factor analysis.

3. Request 2

Students explain the eigenvalues and base on that eigenvalues choose the best number of factors to do the Factor Analysis. Explain why you chose this number.

a. Eigenvalues

- The eigenvalue represents the amount of variance each factor accounts for. Each extracted factor will have an eigenvalue (the integer multiple of the original vector). The first extracted factor is going to try to absorb as much of the variance as possible, so successive eigenvalues will be lower than the first [2].
- The significance of eigenvalues is as follows:
    - A larger eigenvalue indicates that the corresponding factor explains a large portion of the variance in the data.
    - A smaller eigenvalue indicates that the corresponding factor explains a small portion of the variance in the data.
    - Factors with similar eigenvalues may be considered equivalent in explaining the data.

b. Number of factors

- Eigenvalues greater than or equal to 1 are stable values, so we can use this criterion to choose the number of vectors that meet this condition.
- Alternatively, we can visualize the eigenvalues using a scree plot. These plots depict the amount of variance explained by each factor, and the "elbow point" is the number of factors just before the plot starts to level off.
- In this source code, we choose to count the number of factors with eigenvalues greater than or equal to 1 because it's noticed that finding the "elbow point" in the scree plot can be challenging and prone to error. Therefore, the number of factors obtained is 6.

4. Request 3

Students look at the loadings table explain the significant of each factor versus each property. If there are factor(s) that has no “high loading” value, you can remove these and perform Factor Analysis again with the remain factor. Otherwise, explain the Factor Variance Table.

a. Factor Loadings:

- Factor loadings are a matrix of how observed variables are related to the factors you’ve specified. In geometric terms, loadings are the numerical coefficients corresponding to the directional paths connecting common factors to observed variables. They provide the basis for interpreting the latent variables. Higher loadings mean that the observed variable is more strongly related to the factor. A rule of thumb is to consider loadings above 0.3.
- In the loading table represented in the source code, with a threshold value of 0.3 for high loadings, it can be observed that all factors have coefficients greater than this value. Therefore, there is no need to remove any factors from the table.

- The Factor loadings table in factor analysis displays the correlation coefficients between the observed variables (input variables) and the factors discovered during the factor analysis process. The significance of the factor loadings table is to describe the extent to which each observed variable is influenced by each factor. Specifically:
    - Each row in the factor loadings table corresponds to an observed variable. Each column in the table corresponds to a latent factor.
    - The value at each cell in the table represents the degree of correlation between the observed variable and the corresponding factor. This value typically ranges from -1 to 1.
    - Values close to -1 or 1: The observed variable has a strong relationship with the corresponding factor. When the value is negative, the relationship is negative; when the value is positive, the relationship is positive.
    - Values close to 0: The observed variable has no significant relationship with the corresponding factor.

b. Factor variance table

- The Factor Variance table provides detailed information about the variance explained by each latent factor in factor analysis. This table typically includes the following columns:
    - Factor: The factors analyzed in the model.
    - Variance: The variance value of each factor, indicating the extent to which that factor explains the variance of the original data.
    - Proportion Variance: The percentage ratio of the variance of each factor to the total variance of the original data. This is the percentage of variance that each factor contributes to the overall variance of the data.
    - Cumulative Variance: The cumulative sum of the proportion variance of the factors. This is the cumulative sum of the percentage of variance explained by the factors from the first factor to the current factor.
- In this program, we could conclude that:
    - The explanation level of each factor: The factors with higher variance values are Factor 0, Factor 1, and Factor 2.
    - Percentage variance ratio: The factors with higher percentage variance ratios are Factor 0, Factor 1, and Factor 2.
    - Cumulative variance: The cumulative variance achieved is 43.3333%, with 6 factors considered.

III. REASEARCHING QUESTIONS

1. What are factors in Factor Analysis, and why are they important?

- Factor analysis, or correlational analysis, is a statistical technique that reduces a large number of variables into a few data sets that are more manageable and understandable. This makes it easier to work with research data.
- Factors in factor analysis refer to latent variables that represent dimensions or underlying structures within a dataset. These factors are not directly observed but are inferred from the relationships among the observed variables.
- Factors play an important role because:
    - **Dimension Reduction**: Factor analysis helps reduce the dimensionality of a dataset by identifying latent factors that explain the correlations among the observed variables. Instead of dealing with a large number of observed variables, we can work with a smaller number of factors that capture the necessary information in the data.

- **Pattern Identification:** By identifying factors, we can discover patterns and structures in the data that may not be evident from individual observed variables alone. Factors help understand the latent relationships and dependencies among variables.
- **Data Interpretation:** Factors provide a meaningful way to interpret the relationships among the observed variables. They represent abstract concepts or theoretical structures and can help understand the underlying phenomena being studied.
- **Variable Selection:** Factor analysis can assist in selecting the most important variables or features in a dataset by identifying the factors that contribute most to the variance in the data.

2. Explain the significance of eigenvalues and eigenvectors in Factor Analysis?

- Eigenvalues and eigenvectors are fundamental concepts, play a significant role in factor analysis. In factor analysis, they are used to identify the underlying factors and understand the structure of the data. Eigenvalues measure the amount of variance explained by each factor, helping determine the number of factors to retain, while eigenvectors provide the direction and weights of the factors, aiding in explaining the underlying structure of the data in factor analysis.
- **Eigenvalues**: represent the amount of variance explained by each factor in the dataset. Each eigenvalue corresponds to a factor, and a higher eigenvalue indicates that the corresponding factor explains more variance in the data. They help determine the number of factors to retain in factor analysis. Factors associated with eigenvalues greater than 1 (the Kaiser criterion) or values above the "elbow" point in the scree plot are considered important and retained because they explain more variance than individual variables.
- Eigenvectors represent the direction of the latent factors in the original variable space. Each eigenvector corresponds to a factor and provides weights (loadings) for each observed variable on that factor. Eigenvectors aid in explaining the structure of the factors. By examining the loadings of variables on each factor, we can understand which variables have the strongest relationship with each factor and infer the significance or underlying structure represented by that factor.

3. Compare the Factor Analysis Vs. Principle Component Analysis.

Based on the understanding of these two methods, below is a table of data comparing PCA and FA based on some basic criteria:

Criteria	Factor Analysis	Principal Component Analysis
Objective	Identifying factors	Reducing dimension of data
Data transformation	Use the correlation matrix	Use the covariance matrix
Goal	Determine underlying structure	Retaining the largest variance
Factor evaluation	Eigenvalues and eigenvectors	Eigenvalues and eigenvectors
Number of vectors to select	Based on the eigenvalues (greater than 1)	No specific criteria
Variance explained	Degree of correlation between variables	Variance of variables
Differentiation	Examines relationship between variables	Identifies most important variables
Application	Exploratory analysis Data compression Visualization	Confirmatory analysis Data interpretation Hypothesis testing

4. Provide examples of real-world applications where Factor Analysis can be useful.

Market Research: Factor analysis can be used to identify latent factors influencing consumer preferences and behavior. For example, in a survey containing questions about various product features, factor analysis can help identify which features are most important to consumers and group them into different factors such as price sensitivity, product quality, or brand loyalty. This helps businesses formulate appropriate strategies.

Personality Traits: FA is used to analyze the structure of personality trait questionnaires such as conscientiousness, extraversion, confidence, and other aspects of personality. In this way, FA can help psychologists explain and classify complex personality traits into more intuitive and understandable factors.

Social and Behavioral Research: In this field, FA can be used to analyze relationships between variables such as life satisfaction, romantic relationships, or consumer behavior. In this way, FA helps psychologists understand the structure and factors influencing human behavior and cognition.

IV. REFFERENCES

[1] Statistischedatenanalyse, "Exploratory Factor Analysis (EFA), How to interpret KMO and Bartlett´s test," Statistischedatenanalyse, [Online]. Available: https://www.statistischedatenanalyse.de/images/Exploratory_Factor_Analysis-EFA-How_to_interpret_KMO_and_Bartletts_test.pdf.

[2] Columbia University Irving Medical Center, "Exploratory Factor Analysis," Columbia University Irving Medical Center, [Online]. Available: https://www.publichealth.columbia.edu/research/population-health-methods/exploratory-factor-analysis.