style=mystyle

HO CHI MINH UNIVERSITY OF SCIENCE

MULTIVARIATE STATISTICS ANALYSIS - LAB04

April 13, 2024

Author

Nguyen Viet Kim - 21127333

Contents

1 Introduction **4**

2 Algorithm Description **4**

2.1 Imported Libraries 4

2.2 Import Data 4

2.3 Evaluate The "Factorability" of The Dataset 4

2.3.1 Bartlett's Test 4

2.3.2 Kaiser-Meyer-Olkin Test 6

2.4 Choosing the Number of Factors 6

2.5 Performing Factor Analysis 7

1 Introduction

The used dataset is BFI dataset. 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 Algorithm Description

2.1 Imported Libraries

The required libraries are Numpy, Pandas, Scikitlearn, Factor-Analyzer and Matplotlib:

```
import pandas as pd
from sklearn.datasets import load_iris
from sklearn.decomposition import PCA, FactorAnalysis
from factor_analyzer import FactorAnalyzer
import matplotlib.pyplot as plt
import numpy as np
```

The requirement for the versions of the libraries are specified in the requirements.txt which is included in the source code folder.

2.2 Import Data

The data can be installed via https://vincentarelbundock.github.io/Rdatasets/csv/psych/bfi.csv link. The program then read the data from the csv file with the index specified in the Index variable.

```
df= pd.read_csv("../data/bfi.csv")
df.columns
Index = np.array(['A1', 'A2', 'A3', 'A4', 'A5', 'C1', 'C2', 'C3', 'C4', 'C5', 'E1',
 ↪  'E2',
        'E3', 'E4', 'E5', 'N1', 'N2', 'N3', 'N4', 'N5', 'O1', 'O2', 'O3', 'O4',
        'O5', 'gender', 'education', 'age'], dtype ='object')
# Dropping unnecessary columns
df.drop(['gender', 'education', 'age'],axis=1,inplace=True)
# Dropping missing values rows
df.dropna(inplace=True)
```

The data is processed by dropping unnecessary columns and missing values rows.

2.3 Evaluate The "Factorability" of The Dataset

There are two method for evaluating the data: Bartlett's Test and Kaiser-Meyer-AOlkin Test.

2.3.1 Bartlett's Test

Bartlett's Test of Sphericity assesses the suitability of a dataset for factor analysis by examining whether the observed correlations among variables are significantly different from those expected under the assumption of no correlation. It involves formulating null and alternative hypotheses regarding the correlation structure of the variables and computing a chi-square statistic to quantify the degree of dissimilarity between the

observed correlation matrix and the identity matrix. A high chi-square value, along with a low associated p-value, indicates that the observed correlations are unlikely to have occurred by chance alone, suggesting that the dataset may be suitable for factor analysis. Conversely, a non-significant result suggests that the variables are uncorrelated and that factor analysis may not be appropriate for exploring underlying factors in the dataset.

```
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(df)
chi_square_value, p_value
# the p-value is 0
```

In the program, the chi-Square Value (χ^2) is statistic measures the discrepancy between the observed correlation matrix and the identity matrix. A higher chi-square value indicates a greater discrepancy, suggesting that the observed correlations are significantly different from what would be expected under the assumption of no correlation.

$$\chi^2 = -\left(N - 1 - \frac{2p+5}{6}\right)\ln\left(\frac{\det(R)}{\det(T)}\right)$$

where:

- χ^2 is the chi-square value,
- N is the sample size,
- p is the number of variables or factors,
- $\det(R)$ is the determinant of the correlation matrix,
- $\det(T)$ is the determinant of the anti-image covariance matrix.

The p-value associated with the chi-square statistic represents the probability of obtaining a chi-square value as extreme as or more extreme than the one observed, assuming that the null hypothesis (no correlation among variables) is true. A low p-value indicates that the observed discrepancies are unlikely to have occurred by chance alone, leading to the rejection of the null hypothesis. In the context of Bartlett's Test, a low p-value suggests that the dataset is suitable for factor analysis because the observed correlations are significantly different from zero. Conversely, a high p-value indicates that the observed correlations are not significantly different from zero, indicating that factor analysis may not be appropriate for the dataset.

$$p = 1 - F_{\chi^2}(x; df), \tag{1}$$

where:

- p is the p-value,
- F_{χ^2} is the chi-square cumulative distribution function,
- x is the chi-square value,
- df is the degrees of freedom.

2.3.2 Kaiser-Meyer-Olkin Test

The Kaiser-Meyer-Olkin (KMO) test is a statistical measure used to assess the sampling adequacy for factor analysis and related techniques. It evaluates whether the data is suitable for factor analysis by examining the correlation structure among variables. The test examine the sampling adequacy, correlation structure, interpretation and variable selection of the variables.

```
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(df)
kmo_model
# KMO is 0.84
```

The KMO can be demonstrated by the following equations:

Individual KMO for each variable (KMOi):

$$KMO_i = \frac{\sum(\text{Corr}_{ij})^2}{\sum(\text{Corr}_{ij})^2 + \sum(\text{PartialCorr}_{ij})^2}$$

where Corr_{ij} represents the correlation between variable i and variable j, and PartialCorr_{ij} represents the partial correlation between variable i and variable j, controlling for all other variables.

Overall KMO:

$$KMO = \frac{\sum KMO_i}{\text{Number of Variables}}$$

The range of KMO is between 0 and 1. In this case, the KMO achieved 0.84, meaning that this data is suitable for factor analysis since the correlation between the variables are high.

2.4 Choosing the Number of Factors

Eigenvalues are the special set of scalar values that is associated with the set of linear equations most probably in the matrix equations. The eigenvectors are also termed as characteristic roots. It is a non-zero vector that can be changed at most by its scalar factor after the application of linear transformations.

```
# Create factor analysis object and perform factor analysis
fa = FactorAnalyzer()
fa.fit(df, 25)

# Check Eigenvalues
ev, v = fa.get_eigenvalues()
```

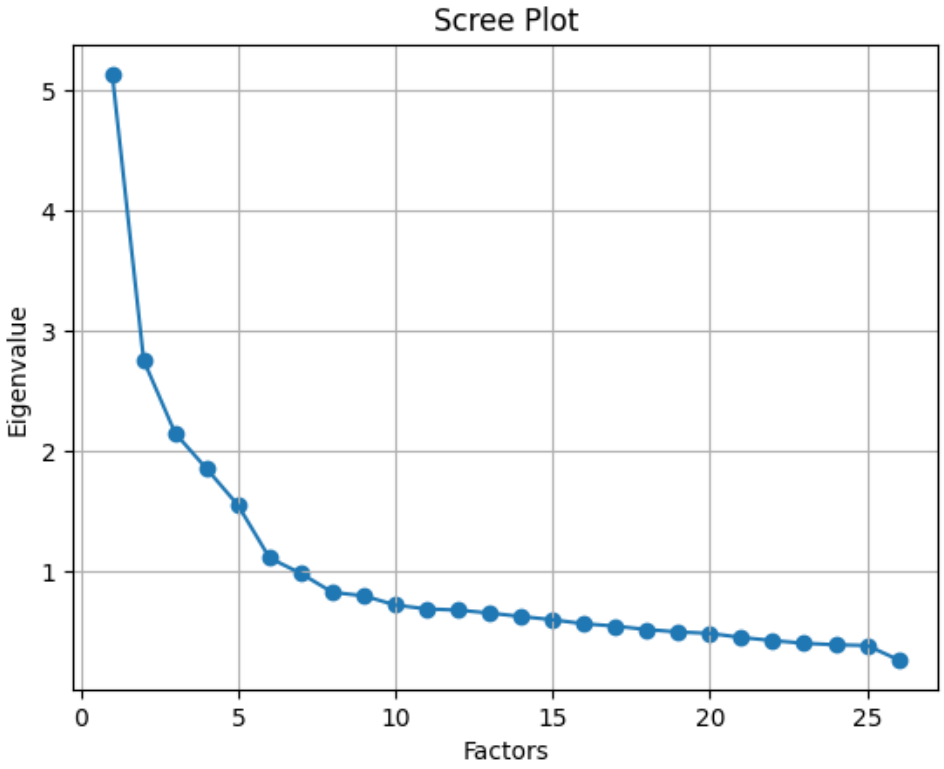

Figure 1: Scree plot

To choose the best number of factor for the factor analysis, the number should be chosen at the elbow of the Scree plot, which is at the factor 5. Therefore the chosen number is 5.

2.5 Performing Factor Analysis

The program use the varimax rotation for factor analysis with 5 factors:

```
    # Create factor analysis object and perform factor analysis with rotation
    fa = FactorAnalyzer(rotation="varimax")
    fa.fit(df, 5)

    # Retrieve factor loadings
    loadings = fa.loadings_

    # Print factor loadings
    print("Factor Loadings:")
    print(loadings)

    # Get variance of each factor
    variance = fa.get_factor_variance()
```

The result is shown as the below figure:

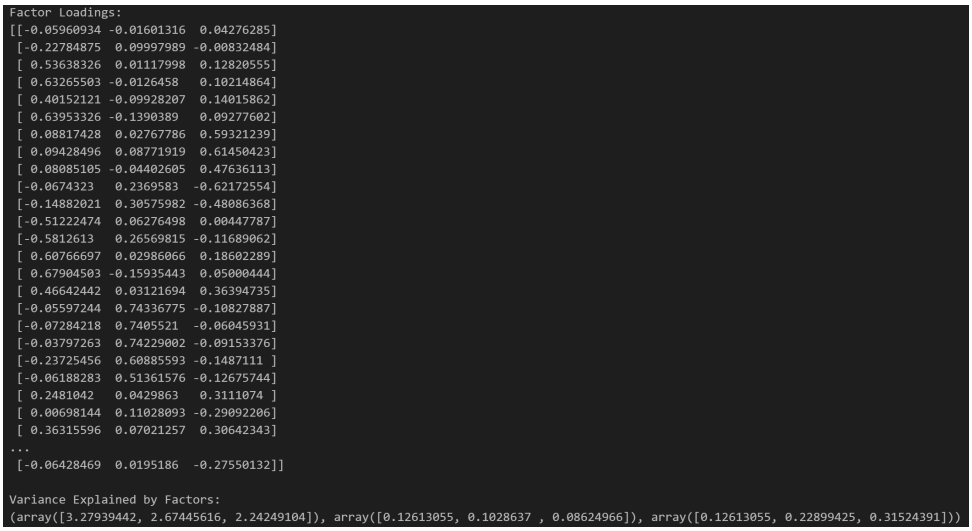

Figure 2: Factor loadings table

Tthe three factors explain 12.6%, 10.3%, and 8.6% of the total variance, respectively. The cumulative variance explained by the three factors together is 31.5%. This indicates that the extracted factors collectively capture a significant portion of the variability in the data.

References