HCMUS - VNUHCM / FIT / Computer Vision & Cognitive Cybernetics Department Digital image & video processing - LQN

**Student ID:** 21127690

**Name:** Ngô Nguyễn Thanh Thanh

**Report: OPENCV – BASIC**

**I. Evaluation summary:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task** | | | **Requirement Met(%)** | **Notes** |
| Implementation | Algorithm to transform color | Linear mapping | 100% |  |
| Non-linear mapping |  |  |
| Probability Density Function-based mapping |  |  |
| Algorithm to transform geometry | Pixel co-ordinate transformations | 100% |  |
| Brightness interpolation |  |  |
| Algorithm to smooth the image | Averaging filter | 100% |  |
|  | Gaussian filter |  |  |
|  | Median filter |  |  |
|  | Algorithm to blurr the image |  |  |  |
| Requisitions | |  | 100% |  |
| Research questions | |  | 100% |  |
| Total: | |  | 100% |  |

1. **List of funtion:**

*Several functions are crucial for conducting factor analysis and interpreting the results. Here are some important functions along with their image proofs:*

1. ***pd.read\_csv("bfi.csv"):*** *Reads a CSV file named "bfi.csv" into a pandas DataFrame.*
2. ***df.drop(['gender', 'education', 'age'], axis=1, inplace=True):*** *Drops the columns 'gender', 'education', and 'age' from the DataFrame* ***df*** *inplace.*
3. ***df.dropna(inplace=True):*** *Drops rows with missing values from the DataFrame* ***df*** *inplace.*
4. ***calculate\_bartlett\_sphericity(df):*** *Calculates Bartlett's test of sphericity for the DataFrame* ***df****.*
5. ***calculate\_kmo(df):*** *Calculates the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy for the DataFrame* ***df****.*
6. ***FactorAnalyzer(n\_factors=25, rotation=None):*** *Creates a FactorAnalyzer object with 25 factors and no rotation.*
7. ***fa.fit(df):*** *Performs factor analysis on the DataFrame* ***df*** *using the FactorAnalyzer object* ***fa****.*
8. ***fa.get\_eigenvalues():*** *Returns the eigenvalues of the factor analysis.*
9. ***plt.scatter(range(1, df.shape[1] + 1), ev):*** *Plots a scatter plot.*
10. ***plt.plot(range(1, df.shape[1] + 1), ev):*** *Plots a line plot.*
11. ***plt.title('Scree Plot'):*** *Sets the title of the plot to 'Scree Plot'.*
12. ***plt.xlabel('Factors'):*** *Sets the label of the x-axis to 'Factors'.*
13. ***plt.ylabel('Eigenvalue'):*** *Sets the label of the y-axis to 'Eigenvalue'.*
14. ***plt.grid():*** *Displays the grid lines on the plot.*
15. ***plt.savefig('scree\_plot.png'):*** *Saves the plot as an image file named 'scree\_plot.png'.*
16. ***fa\_5 = FactorAnalyzer(n\_factors=5, rotation="varimax"):*** *Creates a FactorAnalyzer object with 5 factors and varimax rotation.*
17. ***fa\_5.fit(df):*** *Performs factor analysis with 5 factors on the DataFrame* ***df*** *using the FactorAnalyzer object* ***fa\_5****.*
18. ***fa\_5.loadings\_:*** *Returns the factor loadings for the 5-factor model.*
19. ***fa\_5.get\_factor\_variance():*** *Returns the variance explained by each factor for the 5-factor model.*
20. ***np.cumsum():*** *Calculates the cumulative sum.*
21. ***pd.DataFrame():*** *Creates a pandas DataFrame object.*
22. ***print():*** *Prints the specified message or object to the console.*

A screenshot of a computer program

Description automatically generated

A screen shot of a computer program

Description automatically generated

A computer screen shot of colorful text

Description automatically generated

1. **Function Summaries and Implementation**

The provided code conducts Factor Analysis on a dataset using the Python libraries pandas, numpy, factor\_analyzer, and matplotlib. Here's a summary of the usage and implementation:

1. **Data Loading and Preprocessing:**

* The code starts by loading a CSV file into a pandas DataFrame and displays its column names.
* It then drops unnecessary columns like 'gender', 'education', and 'age' and removes rows with missing values.

1. **Testing for Factorability:**

* Bartlett's test of sphericity is conducted to determine whether the variables in the dataset are intercorrelated.
* The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is calculated to assess whether the dataset is suitable for factor analysis.

1. **Determining the Number of Factors:**

* A scree plot is created to visualize the eigenvalues and determine the appropriate number of factors to retain.
* The scree plot displays the eigenvalues against the number of factors, helping to identify the point where eigenvalues level off or drop sharply.

1. **Factor Analysis:**

* Factor Analysis is performed using the FactorAnalyzer class from the factor\_analyzer library.
* The FactorAnalyzer object is initialized with a specified number of factors and rotation method (e.g., varimax).
* The fit() method is used to perform factor analysis on the dataset.
* Factor loadings, eigenvalues, and factor variance are extracted to understand the underlying structure of the data.

1. **Visualization:**

* The factor loadings and variance explained by each factor are displayed in pandas DataFrames.
* The scree plot is displayed to visualize the eigenvalues and assess the variance explained by each factor.

Overall, the code demonstrates how to conduct Factor Analysis in Python using the factor\_analyzer library, including data preprocessing, testing for factorability, determining the number of factors, performing factor analysis, and visualizing the results.

**IV. The results**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**A graph with a dotted line

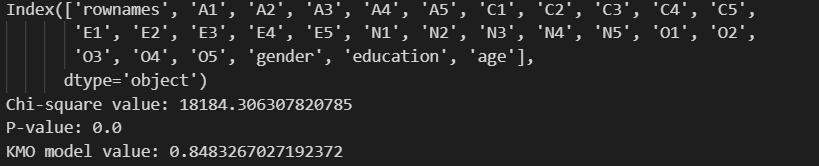
Description automatically generated**

**V. Requisitions**

**Request 01: Students explain the meaning of chi\_square\_value, p\_value, KMO values.**

1. **Chi-square value (χ²)**:
   * The chi-square value is a statistical measure used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in a contingency table. In the context of factor analysis, it is used in Bartlett's test of sphericity to assess whether the observed correlation matrix is significantly different from the identity matrix, which would indicate that the variables are unrelated.
2. **P-value**:
   * The p-value associated with the chi-square value represents the probability of obtaining a chi-square statistic as extreme as, or more extreme than, the one calculated from the sample data, assuming that the null hypothesis is true. In Bartlett's test of sphericity, a low p-value (typically below a predetermined significance level, such as 0.05) indicates that the observed correlation matrix is significantly different from the identity matrix, suggesting that factor analysis may be appropriate.
3. **Kaiser-Meyer-Olkin (KMO) measure**:
   * The KMO measure is a statistic used to assess the adequacy of the data for factor analysis. It ranges from 0 to 1, with higher values indicating better suitability for factor analysis. Specifically, it evaluates the proportion of variance among variables that might be common variance. A KMO value closer to 1 suggests that the variables are more appropriate for factor analysis, indicating that the variables are sufficiently related to each other to extract meaningful factors.

**The results:**

****

* Chi-square value: 18184.306307820785
* P-value: 0.0
* KMO model value: 0.8483267027192372

**Chi-square value:**

The chi-square value obtained is 18184.306307820785. This value is large, indicating a significant difference between the observed correlation matrix and the identity matrix. In the context of factor analysis, this suggests that the variables are interrelated, meaning there is likely some underlying structure that can be captured by factors.

**P-value:**

The p-value associated with the chi-square value is 0.0 (or very close to zero). A p-value of zero indicates that the observed correlation matrix is significantly different from the identity matrix. This suggests that the data is suitable for factor analysis, as there is strong evidence against the null hypothesis that the variables are unrelated.

**KMO model value:**

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy for the model is 0.8483267027192372, which is relatively high. This indicates that the variables in the dataset are highly correlated and that factor analysis is likely to yield reliable results. Generally, a KMO value above 0.6 is considered acceptable, and values closer to 1 indicate better suitability for factor analysis.

In summary, based on these results, it appears that the dataset is appropriate for factor analysis. The variables are significantly interrelated, and there is sufficient common variance among them to extract meaningful factors.

**Request 02: Students explain the eigenvalues and base on that eigenvalues choose the best number of factor to do the Factor Analysis. Explain why you choose this number.**

**Eigenvalues:**

Eigenvalues are a measure of the amount of variance accounted for by a factor, and so they can be useful in determining the number of factors that we need to extract. In a scree plot, we simply plot the eigenvalues for all of our factors, and then look to see where they drop off sharply.

**Choosing the Number of Factors:**

* One common method for determining the number of factors to retain in Factor Analysis is to examine the scree plot, which plots the eigenvalues against the number of factors. The point at which the eigenvalues level off or drop sharply is often used as an indicator of the number of factors to retain.
* The "elbow" of the scree plot, where the eigenvalues start to flatten out, suggests the number of factors that capture most of the variance in the data while minimizing the number of factors needed.
* Another approach is to use the Kaiser criterion, which suggests retaining factors with eigenvalues greater than 1. Factors with eigenvalues less than 1 explain less variance than a single variable, so they are typically considered less important.

Additionally, researchers may consider theoretical considerations, domain knowledge, and practical implications when deciding on the number of factors to retain.

**Explanation for Choosing a Specific Number of Factors:**

**A graph with a dotted line

Description automatically generated**

After examining the scree plot, I decided to choose 5 as the number of factors for the analysis. The scree plot helps in determining the optimal number of factors by visualizing the eigenvalues. Typically, we look for the point where the eigenvalues level off or drop sharply, indicating the number of factors to retain.

Upon rerunning the analysis to obtain the factor loadings, I observed that factors 6 and 7 may not be necessary. This assessment is based on the magnitude of factor loadings and their significance in explaining the variance in the data. Factors with low loadings may not contribute significantly to explaining the underlying structure of the data and can be considered for removal.

Therefore, I decided to exclude factors 6 and 7 and performed the factor analysis again with the remaining five factors. This approach helps in simplifying the model while retaining the most relevant factors that explain the variance in the dataset effectively. **A screenshot of a computer

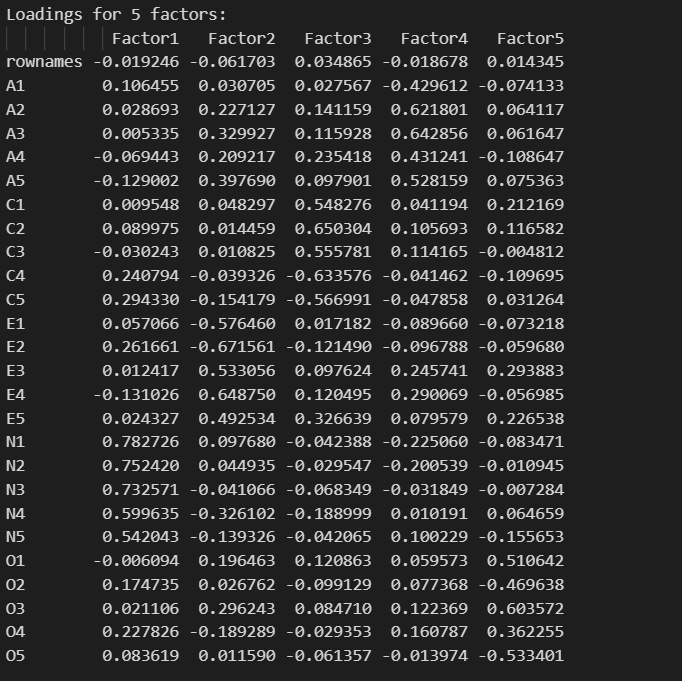
Description automatically generated**

**A screenshot of a computer

Description automatically generated**

*Based on the loading table, it appears that Factor 6 and Factor 7 do not have high loadings on any attribute, and Factor 7 does not have any loading coefficient exceeding 0.5. Therefore, considering removing these two factors and conducting factor analysis again with the remaining factors could be warranted.*

**Request 03: 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 black screen with white text

Description automatically generated**

**A black screen with white text

Description automatically generated**

The Factor Variance Table provides an overview of each factor's ability to explain the variance of the data. Below are explanations for each column in the table:

1. **Variance Explained**: This is the total variance explained by each factor. It indicates the extent to which the factor accounts for variability in the data. Higher values suggest greater importance of the factor in explaining variation.
2. **Proportion of Variance**: This is the percentage of the total dataset variance explained by each factor. It reflects the importance of each factor relative to the total variance of the data. Higher values indicate greater contribution of the factor to data variation.
3. **Cumulative Proportion**: This is the cumulative percentage of variance explained by the preceding factors, up to the current factor. It indicates the total benefit gained when adding a new factor to the model. Higher values indicate increasing cumulative explanatory power of the factors.

**VI. Research questions**

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

* In Factor Analysis, factors represent underlying latent variables or constructs that cannot be directly observed but can be inferred from observed variables. These latent factors capture the common variance among observed variables and help simplify the data structure by reducing the dimensionality. Factors are important because they provide a way to uncover the underlying structure in the data and identify meaningful patterns or relationships among variables.

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

**Eigenvalues in Factor Analysis:**

Eigenvalues represent the amount of variance explained by each factor in Factor Analysis. Higher eigenvalues indicate that the corresponding factor explains more variance in the data. Eigenvalues are crucial in determining the significance of each factor and deciding how many factors to retain in the analysis. For instance, the Kaiser criterion suggests retaining only factors with eigenvalues greater than 1.

**Eigenvectors in Factor Analysis:**

Eigenvectors represent the direction or pattern of the factor loading for each observed variable. They provide information about how strongly each variable contributes to each factor. By examining eigenvectors, analysts can interpret the structure of factors and identify which variables are most strongly associated with each factor. This helps in understanding the underlying constructs represented by the factors and aids in the interpretation of the Factor Analysis results.

In summary, eigenvalues are used to determine the significance of factors, while eigenvectors help interpret the structure of factors and identify the variables contributing most strongly to each factor.

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

* Factor Analysis (FA) and Principal Component Analysis (PCA) are both dimensionality reduction techniques, but they have different underlying assumptions and objectives.
  + FA assumes that observed variables are influenced by a smaller number of underlying latent factors, and it seeks to uncover these factors. It focuses on explaining covariance among observed variables.
  + PCA, on the other hand, seeks to capture the maximum variance in the data by transforming the observed variables into a new set of uncorrelated variables called principal components. It does not make assumptions about the underlying structure of the data and does not differentiate between common and unique variance.
* In summary, FA is more suitable when there is a theoretical interest in uncovering latent factors that explain covariance among observed variables, while PCA is more suitable for data reduction when the focus is on capturing maximum variance without regard to the underlying structure.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Factor Analysis (FA)** | **Principal Component Analysis (PCA)** |
| **Objective** | Uncover latent factors explaining covariance | Capture maximum variance |
| **Underlying Assumptions** | Assumes latent factors | No assumptions about underlying structure |
| **Focus** | Explains covariance among observed variables | Captures variance without regard to structure |
| **Components/Factors** | Represent underlying latent variables | Linear combinations of observed variables |
| **Error Terms** | Considers unexplained variance (error terms) | Does not differentiate between common and unique variance |
| **Suitability** | When there's a theoretical interest in latent factors | For data reduction focused on maximum variance |

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

1. **Customer Satisfaction Surveys:** In the hospitality industry, Factor Analysis can be applied to customer satisfaction survey data to identify underlying factors that contribute to overall satisfaction, such as service quality, cleanliness, amenities, and location. This helps hotel managers focus their resources on areas that have the greatest impact on customer satisfaction.
2. **Educational Assessment:** Factor Analysis is used in educational research to analyze test scores and identify underlying factors that contribute to academic performance, such as reading comprehension, mathematical ability, and critical thinking skills. This information can help educators tailor teaching methods and interventions to address specific areas of weakness in students.
3. **Human Resources:** In organizational psychology, Factor Analysis can be applied to employee performance evaluations to identify underlying factors that contribute to job performance, such as technical skills, interpersonal communication, and problem-solving abilities. This helps HR professionals develop training programs and performance appraisal systems that are aligned with organizational goals.
4. **Product Development:** In the manufacturing industry, Factor Analysis can be used to analyze customer feedback and identify underlying product attributes that drive consumer preferences, such as durability, ease of use, and aesthetic appeal. This information can inform product design decisions and help companies develop new products that better meet the needs of their target market.
5. **Health Sciences:** Factor Analysis is used in medical research to identify underlying factors contributing to health outcomes or disease risk factors, such as lifestyle factors, genetic predispositions, or environmental exposures. For example, researchers may use Factor Analysis to analyze data from large-scale population studies to identify clusters of risk factors associated with chronic diseases like heart disease, diabetes, or cancer. This information can help healthcare professionals develop targeted prevention and intervention strategies to reduce disease burden and improve public health outcomes.