UNIVERSITY OF SCIENCE

Multivariate Statistical Analysis

Practice 04: Factor Analysis

Lecturer: Nguyễn Mạnh Hùng

Lý Quốc Ngọc

Phạm Thanh Tùng

Student name: Đoàn Việt Hưng

Student ID: 21127289

TABLE OF CONTENT

1. Self-assessment table:

No.	Work name	Work description	Percentage of complettion	Work not done
1	3 requests in the additional guidance	Based on the source code and its ouput, answer the request below each instruction	100% done in jupiter notebook file	None
2	3 requests based on the additional guidance	Follow these instruction to make a report	100%	None
3	Additional research questions	Answer and make a report	100%	None

2. Overview of Factor Analysis:

- Factor analysis is a statistical technique used to identify patterns among observed variables, understand the underlying structure of those variables, and reduce the complexity of data by summarizing it into a smaller set of factors. It's commonly employed in fields such as psychology, sociology, economics, and market research to uncover latent variables that explain the relationships among observed variables.

- Here's an overview of the key concepts and steps involved in factor analysis:

1. Data Collection: Factor analysis starts with a dataset containing multiple variables. These variables could be measurements, survey responses, test scores, or any other type of observable data.

2. Correlation Matrix: The first step in factor analysis is to examine the correlations between all pairs of variables in the dataset. This correlation matrix provides information about the relationships among variables.

3. Factor Extraction: The main goal of factor analysis is to identify a smaller number of underlying factors that explain the correlations among observed variables.

Factor extraction methods, such as principal component analysis (PCA) or maximum likelihood estimation, are used to identify these factors.

4. Factor Rotation: Once the initial factors are extracted, factor rotation techniques are applied to simplify the factor structure and make it easier to interpret. Rotation aims to maximize the variance of factor loadings, thereby clarifying the relationship between variables and factors.

5. Factor Loading: Each variable in the dataset is associated with one or more factors through factor loadings. Factor loadings represent the strength and direction of the relationship between variables and factors. High factor loadings indicate that a variable is strongly associated with a particular factor.

6. Interpretation: After factor extraction and rotation, researchers interpret the meaning of each factor based on the variables with high loadings on that factor. Factors are often given labels or names based on the variables that define them, helping to understand the underlying structure of the data.

7. Validity and Reliability: It's important to assess the validity and reliability of the factors identified through factor analysis. Validity refers to whether the factors accurately represent the underlying structure of the data, while reliability assesses the consistency of factor solutions across different samples or datasets.

8. Application: Finally, the results of factor analysis can be used for various purposes, such as developing psychometric tests, creating composite scores, reducing data dimensionality for further analysis, or gaining insights into complex relationships among variables.

Overall, factor analysis provides a powerful tool for exploring the underlying structure of data and uncovering hidden patterns that may not be immediately apparent from the observed variables alone.

3. Installation and implementation:

1) Import vital libraries:

```
# Import required libraries
import pandas as pd
from sklearn.datasets import load_iris
from factor_analyzer import FactorAnalyzer
import matplotlib.pyplot as plt
```

- **pandas:** This library is widely used for data manipulation and analysis. It offers data structures and functions to work with structured data, such as data frames. In the code, it's used to read the CSV file, manipulate the data frame, and perform various operations on it.
- **sklearn.datasets:** This module from scikit-learn provides utilities to load datasets, including sample datasets for practice or testing purposes. In this code, it's used to load the Iris dataset, although it seems it's not being used later in the code.
- **factor_analyzer:** This is a Python package specifically designed for factor analysis. Factor analysis is a statistical method used to uncover underlying factors (or latent variables) that explain patterns of correlations within a set of observed variables. The package provides functions and classes to perform factor analysis, such as calculating Bartlett's test of sphericity, Kaiser-Meyer-Olkin (KMO) statistic, factor extraction, and rotation methods.
- **matplotlib.pyplot:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. The `pyplot` module provides a MATLAB-like interface for creating plots and visualizations. In this code, it's used to create a scree plot, which helps visualize the eigenvalues obtained from the factor analysis.

These libraries collectively enable the user to perform factor analysis on a dataset, assess the suitability of the data for factor analysis using statistical tests like Bartlett's test and KMO statistic, and visualize the results through a scree plot.

2) <u>Prepare data:</u>

```
df= pd.read_csv("C:\\Users\\ACER\\OneDrive\\Desktop\\MSA\\PRACTICE\\Practice 4\\21127289\Data\\bfi.csv")
df.columns
# Index(['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)
```

- Download data at this site:
https://vincentarelbundock.github.io/Rdatasets/csv/psych/bfi.csv

3) Evaluate the "factorability" of our dataset:

- Factorability means "can we found the factors in the dataset?". There are two methods to check the factorability or sampling adequacy: **Bartlett's Test** and **Kaiser-Meyer-Olkin Test**

```
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value, p_value = calculate_bartlett_sphericity(df)
print(chi_square_value, p_value)
# the p-value is 0
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(df)
print(kmo_model)
# KMO is 0.84.
```

- These functions are used to evaluate the suitability of the dataset for factor analysis. Bartlett's test of sphericity assesses whether variables in the dataset are correlated enough for factor analysis, with a low p-value indicating suitability. The Kaiser-Meyer-Olkin (KMO) measure determines the adequacy of the dataset for factor analysis, with higher values suggesting better suitability. In the provided code, Bartlett's test yields a significant result (p-value=0), indicating correlation among variables, while the KMO statistic indicates a high level of adequacy (KMO=0.84), supporting the use of factor analysis on the dataset.

- **Chi-Square Value:** This is a statistic that measures the difference between the observed frequencies and the frequencies we would expect to obtain according to

a specific theoretical distribution. In the context of factor analysis, it's used in Bartlett's test of sphericity, which checks if the observed variables intercorrelate at all using the observed correlation matrix against the identity matrix. If the test found statistically significant, it indicates that the observed variables are correlated and hence suitable for structure detection.

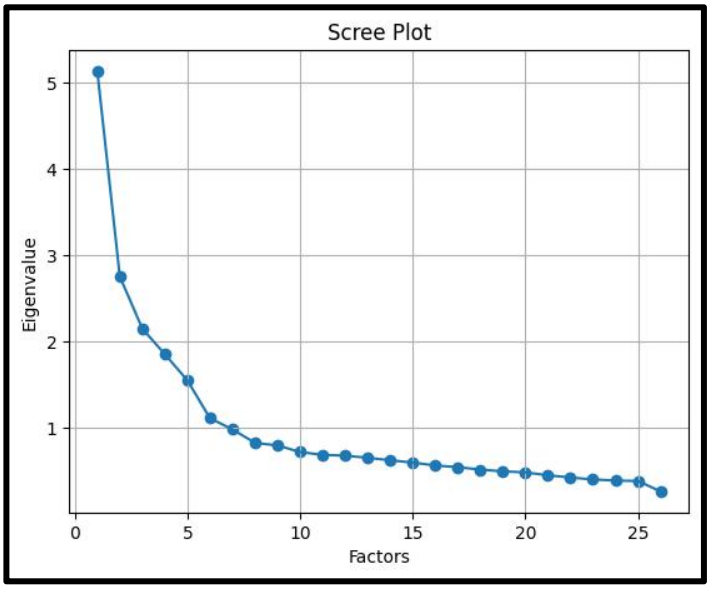
- **P-Value:** The p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis. In your case, a p-value of 0 indicates that the observed correlation matrix is not an identity matrix, meaning that there are some relationships among the variables.
- **Kaiser-Meyer-Olkin (KMO) Measure:** This is a statistic that indicates the proportion of variance in your variables that might be caused by underlying factors. High values (close to 1.0) generally indicate that a factor analysis may be useful with your data. If the value is less than 0.50, the results of the factor analysis probably won't be very useful. In your case, a KMO of 0.84 suggests that you're likely to have distinct and reliable factors.

4) <u>Choosing number of factors:</u>

```
# Create factor analysis object and perform factor analysis
fa = FactorAnalyzer(n_factors=25, rotation=None)
fa.fit(df)
# Check Eigenvalues
ev, v = fa.get_eigenvalues()

# You can create scree plot using matplotlib
plt.scatter(range(1,df.shape[1]+1),ev)
plt.plot(range(1,df.shape[1]+1),ev)
plt.title('Scree Plot')
plt.xlabel('Factors')
plt.ylabel('Eigenvalue')
plt.grid()
plt.show()
```

- This code segment initializes a FactorAnalyzer object with 25 factors and performs factor analysis on the provided dataset. It then retrieves the eigenvalues associated with each factor and creates a scree plot using matplotlib, displaying the eigenvalues against the number of factors to help determine the optimal number of factors to retain for the analysis.



- The scree plot helps to identify the "elbow" point, where the eigenvalues start to level off. This point indicates the number of factors that should be retained in the factor analysis.

- **Eigenvalue Significance:** Eigenvalues measure the variance in all variables which can be attributed to that factor. The higher the eigenvalue, the more variance that factor explains.
- **Choosing Factors:** The 'elbow' method is used to determine the optimal number of factors. This involves looking for a point where the slope of the eigenvalues levels off, indicating diminishing returns for additional factors.
- **Scree Plot Analysis:** Based on the scree plot description, there's an 'elbow' around the 4th or 5th factor. This suggests that subsequent factors contribute less to explaining the variance.
- **Recommendation:** I would recommend using 4 or 5 factors for your factor analysis. This choice is supported by the eigenvalues leveling off after the 5th factor, implying that additional factors would not significantly improve the model's explanatory power.

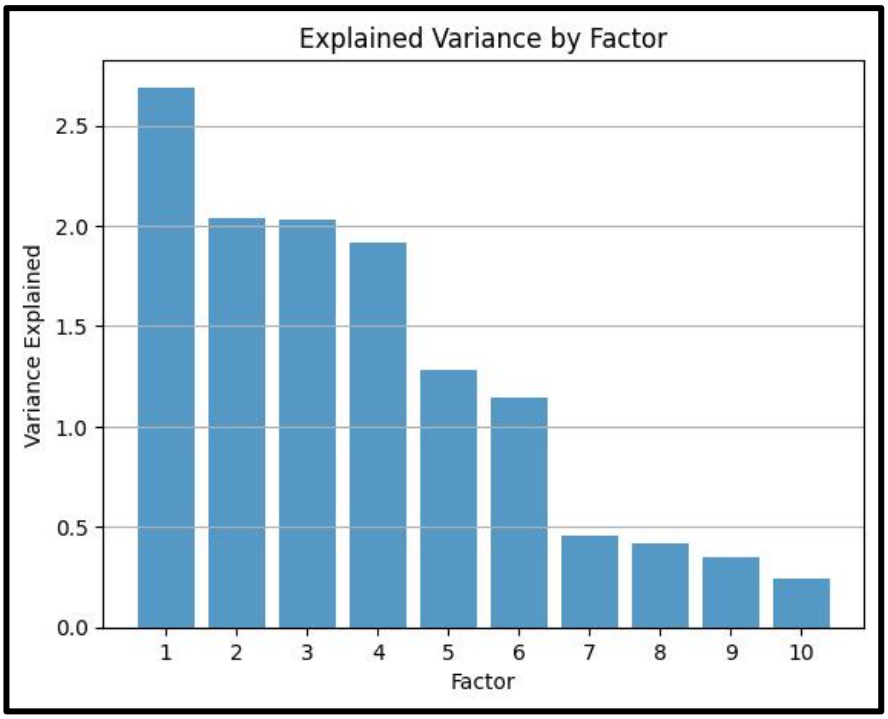
This approach helps to balance between capturing sufficient variance and maintaining model simplicity. Remember, the goal is to explain the data with the fewest factors possible without losing significant information.

5) <u>Performing Factor Analysis:</u>

```
# Create factor analysis object and perform factor analysis
fa = FactorAnalyzer()
num_factors = int(input('Input the magic number you want: '))
fa = FactorAnalyzer(num_factors, rotation="varimax")
fa.fit(df)
loadings = fa.loadings_
# Get variance of each factor
factor_variance = fa.get_factor_variance()
print(factor_variance)
```

- This code conducts factor analysis on a dataset after prompting the user to input the desired number of factors. It initializes a FactorAnalyzer object with the specified number of factors and performs factor analysis with the varimax rotation method. The factor loadings, representing correlations between variables and extracted factors, are then obtained. Additionally, the code calculates and prints the variance explained by each factor, aiding in understanding the contribution of each factor to the dataset's variability.

4. Output and analysis/evaluation:



1. Title and Purpose:

- The title "Explained Variance by Factor" indicates that the graph is showing how much variance each factor explains in the data. In the context of factor analysis, this means each bar represents a factor's ability to account for the variability among the observed variables.

2. Axis Labels:

- The x-axis is labeled "Factor," and it has discrete values ranging from 1 to 10, which correspond to the individual factors extracted from the analysis.

- The y-axis is labeled "Variance Explained," which represents the amount of total variance that each factor accounts for in the dataset. It is measured in the units on the y-axis which range from 0 to a little above 2.5.

3. Bar Chart Description:

- There are 10 bars, each corresponding to one of the 10 factors on the x-axis.

- The height of each bar indicates the level of explained variance for that factor.

- The bars are presented in descending order, suggesting that Factor 1 explains the most variance, and each subsequent factor explains less.

4. Interpretation:

- Factor 1 is by far the most significant, with a variance explained of over 2.5 units. It dominates the chart, indicating it is the most important factor in explaining variability in the dataset.

- Factors 2 through 4 also contribute significantly, but less so than Factor 1, with their explained variances between approximately 1.5 and 2.5 units.

- From Factor 5 onwards, the variance explained decreases more gradually. These factors are increasingly less important in explaining the dataset's variance.

5. Evaluation:

- The fact that the bars decrease in height from left to right suggests that the factors have been ordered by the amount of variance they explain. This is typical in analyses like principal component analysis where factors are ranked by their eigenvalues.

- A graph like this is useful for determining the 'cut-off' point – where the factors cease to be meaningful. This is sometimes decided by looking for an 'elbow' in the graph, which is the point where the explained variance stops decreasing rapidly and starts to flatten out. In this chart, this point appears to occur at Factor 4 or 5.

- It is also apparent that by Factor 10, the variance explained is quite low, indicating that this factor might be considered negligible or not worth considering in further analysis or data reduction strategies.

In summary, this graph is a clear visual representation that would typically be used to help make decisions about the number of factors to retain in a factor analysis or principal component analysis by assessing the explained variance of each factor. Factors with higher explained variance are generally considered more important. The graph shows that in this particular analysis, the first few factors explain most of the variance, with a sharp drop-off after the first factor and a more gradual decline from the fifth factor onwards.

5. Conclusion and application:

- Factor analysis is a powerful statistical technique used to identify underlying relationships between observed variables, allowing researchers to reduce the complexity of their data and uncover latent factors that explain the patterns in the

data. The analysis involves extracting factors that capture the common variance among variables, thereby providing insights into the underlying structure of the data.

- From the graph depicting the explained variance by factor, it is evident that the factors extracted from the dataset vary in their ability to explain the total variability. Factors with higher explained variance are considered more influential in shaping the data patterns, while those with lower variance explained may be less significant. The graph aids in determining the optimal number of factors to retain, striking a balance between capturing enough variance and avoiding overfitting.

Applications of Factor Analysis:

1. Dimensionality Reduction: Factor analysis is widely used in fields such as psychology, sociology, marketing, and finance to reduce the dimensionality of data. By identifying latent factors that summarize the common variance among variables, researchers can simplify complex datasets and focus on the most critical dimensions.

2. Construct Validity: Factor analysis helps assess the construct validity of measurement scales by examining the relationships between observed variables and latent factors. It allows researchers to confirm whether the variables are indeed measuring the underlying constructs they are intended to represent.

3. Market Segmentation: In marketing research, factor analysis can be used to identify consumer segments based on shared characteristics or preferences. By clustering individuals into groups with similar factor loadings, marketers can tailor products and services to specific market segments effectively.

4. Psychological Assessment: Factor analysis is instrumental in developing psychometric tests and assessing personality traits. By identifying underlying factors that explain patterns of responses, psychologists can create more reliable and valid assessment tools.

5. Predictive Modeling: Extracted factors can serve as input variables in predictive modeling tasks, improving the accuracy and interpretability of predictive models. Factor scores derived from factor analysis can be used as features in machine learning algorithms.

6. Data Visualization: Visualizations, such as scree plots and bar charts of explained variance by factor, provide valuable insights into the structure of the data and aid in decision-making regarding the number of factors to retain.

In conclusion, factor analysis is a versatile tool with diverse applications across various disciplines. By uncovering latent factors and simplifying complex datasets, factor analysis enables researchers to gain deeper insights, improve measurement accuracy, and make informed decisions based on the underlying structure of the data.

6. Research questions:

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

- In Factor Analysis, factors are latent variables that represent underlying dimensions or constructs that explain the patterns of correlations among observed variables. These factors are not directly observed but are inferred from the relationships between the observed variables in the dataset. Factors capture the common variance shared by a group of variables, allowing for a more concise representation of the data.

- Importance of Factors in Factor Analysis:

1. Dimension Reduction: Factors help in reducing the dimensionality of the data by summarizing the variation in a set of variables into a smaller number of latent factors. This simplification makes it easier to interpret and analyze complex datasets.

2. Identifying Patterns: Factors reveal the underlying structure of the data by uncovering relationships and patterns that may not be apparent when looking at individual variables. They provide insights into the interrelationships among variables and help in understanding the data more comprehensively.

3. Construct Interpretation: Factors represent meaningful constructs or dimensions that explain the observed correlations among variables. By identifying and labeling these factors, researchers can better understand the theoretical concepts or relationships captured in the data.

4. Data Reduction: Factors help in summarizing the information contained in a large number of variables into a smaller set of factors. This reduction in data complexity facilitates further analysis and interpretation, making it easier to work with the data.

5. Modeling Relationships: Factors serve as building blocks for statistical models and can be used to predict or explain outcomes based on the relationships between variables. Factor scores derived from factor analysis can be used in regression analysis, clustering, or other modeling techniques.

6. Validity Assessment: Factors play a crucial role in assessing the validity of measurement scales and instruments. By examining the factor structure of variables, researchers can evaluate whether the variables are measuring the intended constructs accurately.

7. Decision Making: Factors help in making informed decisions about data analysis, such as determining the number of factors to retain, interpreting the results, and drawing conclusions based on the underlying factors that drive the data patterns.

In summary, factors in Factor Analysis are essential because they provide a concise representation of the data, reveal underlying patterns and relationships, aid in dimension reduction, facilitate construct interpretation, and support decision-making in data analysis and modeling. By extracting and examining factors, researchers can gain valuable insights into the structure of the data and enhance their understanding of complex datasets.

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

- Eigenvalues and eigenvectors play a crucial role in Factor Analysis as they help in extracting the underlying factors from the observed variables. Understanding the significance of eigenvalues and eigenvectors is essential for interpreting the results of Factor Analysis accurately. Here's why they are important:

1. Eigenvalues:

- Determining the number of factors: Eigenvalues indicate the amount of variance explained by each factor. In Factor Analysis, eigenvalues are used to determine the number of factors to retain. Factors with eigenvalues greater than 1 (Kaiser's criterion) or a scree plot showing a clear break in the eigenvalues are typically considered significant and retained.

- Explained variance: Higher eigenvalues represent factors that explain a larger proportion of the total variance in the data. Researchers can prioritize factors with higher eigenvalues as they capture more variability in the dataset.

- Eigenvalue ratio: The ratio of eigenvalues between successive factors can provide insights into the relative importance of each factor. A large difference in eigenvalues suggests a clear distinction between factors in terms of explained variance.

2. Eigenvectors:

- Factor loadings: Eigenvectors represent the factor loadings, which indicate the strength and direction of the relationship between variables and factors. High absolute values in eigenvectors suggest that the variables are highly correlated with the corresponding factor.

- Interpretation of factors: Eigenvectors help in interpreting the meaning of factors by identifying which variables are most strongly associated with each factor. By examining the pattern of loadings in the eigenvectors, researchers can label and interpret the factors in a meaningful way.

- Orthogonality: Eigenvectors corresponding to different factors are orthogonal to each other, meaning they are independent and capture unique variance in the data. This orthogonality property of eigenvectors ensures that factors are uncorrelated and distinct from each other.

3. Factor Rotation:

- Improving interpretability: Eigenvalues and eigenvectors are used in factor rotation techniques to enhance the interpretability of factors. Orthogonal or oblique rotations aim to simplify the factor structure by maximizing the variance explained by each factor and improving the clarity of factor loadings.

4. Model Fit Assessment:

- Eigenvalues as model fit indices: Eigenvalues can also be used as indicators of model fit in Factor Analysis. Higher eigenvalues suggest a better fit between the observed data and the extracted factors, indicating that the factors are effectively capturing the underlying structure of the data.

In conclusion, eigenvalues and eigenvectors are fundamental concepts in Factor Analysis that help in determining the number of factors to retain, assessing the explained variance, interpreting factor loadings, improving model fit, and enhancing the overall understanding of the underlying structure of the data. By leveraging eigenvalues and eigenvectors effectively, researchers can extract meaningful factors and derive valuable insights from their data.

3) Compare the Factor Analysis vs Principle Component Analysis ?

- Factor Analysis and Principal Component Analysis (PCA) are both dimensionality reduction techniques used in data analysis, but they have distinct differences in terms of their underlying assumptions, objectives, and applications. Here is a comparison between Factor Analysis and PCA:

Feartures	Factor Analysis	PCA
Objective	The primary objective of Factor Analysis is to identify latent factors that explain the correlations among observed variables. It aims to uncover the underlying structure of the data by extracting factors that represent meaningful constructs or dimensions.	PCA focuses on capturing the maximum amount of variance in the data using a smaller set of orthogonal components (principal components). It does not explicitly aim to uncover latent factors but rather emphasizes variance maximization.
Assumptions	Factor Analysis is based on the assumption that the observed variables are influenced by a smaller number of unobserved latent factors. It assumes that there is a linear relationship between the observed variables and the latent factors, and that the error terms are uncorrelated.	PCA assumes that the principal components are linear combinations of the observed variables and are orthogonal to each other. It does not require the presence of latent factors and treats all observed variables as equally important in capturing variance.
Interpretation	Factor Analysis aims to provide a meaningful interpretation of the underlying factors by identifying the variables that are most strongly related to each factor. Factors are often labeled based on the variables with high loadings on them, allowing for a clear interpretation of the data.	PCA focuses on capturing variance in the data without explicitly providing a meaningful interpretation of the components. Principal components are derived based on variance maximization and may not have direct interpretability in terms of underlying constructs or dimensions.
Rotation	Factor Analysis often involves factor rotation techniques (e.g., varimax, oblimin) to simplify the factor structure and improve interpretability. Rotation helps in aligning variables with factors to enhance the clarity of factor	PCA does not involve rotation of components since the principal components are already orthogonal to each other. The components are ordered based on the amount of variance they explain in the data.

	loadings.	
Model Assumptions	Factor Analysis assumes that the observed variables are a linear combination of latent factors and error terms, with the error terms being uncorrelated. It allows for the possibility of measurement error and unique variance in the data.	PCA assumes that the observed variables are directly related to the principal components without considering latent factors. It treats all variables as equally important in capturing variance and does not explicitly model measurement error.

- In summary, Factor Analysis and PCA differ in their objectives, assumptions, interpretability, emphasis on variance explained, rotation techniques, and model assumptions. Factor Analysis aims to uncover latent factors that explain the correlations among variables and provide a meaningful interpretation of the underlying structure, while PCA focuses on capturing variance in the data using orthogonal components without explicit consideration of latent factors. Researchers should choose between Factor Analysis and PCA based on their specific research goals, the nature of the data, and the level of interpretability required for their analysis.

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

1. Market Research:

- Brand Perception: Factor Analysis can be used to identify underlying factors that influence consumers' perceptions of brands. By analyzing survey data on brand attributes, Factor Analysis can reveal the key factors driving brand perception, such as quality, price, and customer service.

2. Psychology and Behavioral Sciences:

- Personality Traits: Factor Analysis is commonly used in psychology to study personality traits. It can help identify clusters of related traits (e.g., extraversion, agreeableness) and understand how these traits are interrelated, leading to the development of personality assessment tools like the Big Five personality model.

3. Healthcare and Medicine:

- Health-related Quality of Life: Factor Analysis can be applied to health-related quality of life surveys to identify dimensions that impact patients' well-being. By analyzing responses to questions about physical, emotional, and social aspects of health, Factor Analysis can reveal underlying factors influencing overall quality of life.

4. Education and Learning:

- Student Performance: Factor Analysis can be used in educational research to explore the factors influencing student performance. By analyzing academic assessment data, Factor Analysis can identify underlying factors such as study habits, motivation, and learning styles that contribute to student success.

5. Finance and Economics:

- Risk Assessment: Factor Analysis can be applied in finance to analyze the risk factors affecting investment portfolios. By examining the correlations among various financial indicators, Factor Analysis can identify common risk factors (e.g., market risk, interest rate risk) that influence portfolio performance.

6. Marketing and Customer Segmentation:

- Market Segmentation: Factor Analysis can help in market segmentation by identifying key factors that differentiate customer segments. By analyzing demographic, behavioral, and attitudinal data, Factor Analysis can reveal underlying factors that drive customer preferences and behaviors, leading to more targeted marketing strategies.

7. Environmental Studies:

- Environmental Impact Assessment: Factor Analysis can be used in environmental studies to assess the impact of various factors on environmental quality. By analyzing data on pollutants, land use, and ecological indicators, Factor Analysis can identify key factors influencing environmental degradation and inform policy decisions.

8. Human Resources and Organizational Behavior:

- Employee Engagement: Factor Analysis can be employed in human resource management to understand factors influencing employee engagement. By analyzing survey responses on job satisfaction, work environment, and career development, Factor Analysis can reveal underlying factors that impact employee engagement levels.

These real-world applications demonstrate the versatility and utility of Factor Analysis across diverse fields, highlighting its ability to uncover underlying structures, identify key factors, and provide valuable insights for decision-making and research in various domains.