Factor Analysis

- Exploratory data analysis method:
  - used to create new variables which summarize the information that might be available in the original variables.
  - Latent variables (Factors/ unobserved variables/ hypothetical variables)
  - Reduction in the number of variables

- Maximum number of factors = Number of variables

- Exploratory Factor Analysis (EFA):
  - Reducing number of variables, identifying latent relationship among set of variables in a dataset.
  - EFA seeks to model a large set of observed variables as linear combinations of some smaller set of unobserved latent factors.

- Confirmatory Factor Analysis (CFA): Each factor is associated with a particular set of observed variables.

- Principle:
  - whether the response variables exhibit patterns of relationships with each other, such that the variables can be partitioned into subsets of variables.
  - Study the correlational structure of the variables in a data set.

- Input Data:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Steps of FA:

1. Study the correlation structure matrix
2. Determine the number of underlying factors
3. Interpret these new factors
4. Use these factors in other statistical analyses of the data

Warning: If the original variables are already uncorrelated, then there is a little reason to consider carrying a FA.

Assumptions

- No outliers in data
- Sample size >> factors
- Should not have homoscedasticity (constant variance across sample)

Certain concepts

- Factor Loadings
- Non uniqueness in the factor loading matrix.
- Factor rotation
- Choosing appropriate number of factors
  - Eigenvalues > 1
  - Cumulative percentage variation > specified percentage (80% or 90%)Factor Rotation

Adequacy Test

- **Bartlett's Test**: *In Bartlett's test, if the p-value is < level of significance, the test is statistically significant, indicating that the observed correlation matrix is not an identity matrix.*
- **Kaiser-Meyer-Olkin (KMO) Test**: *The overall KMO for data > 0.8, is excellent. Value of KMO less than 0.6 is considered inadequate. This value indicates that you can proceed with your planned factor analysis.*

EXAMPLE 1

In a consumer-preference study, a random sample of customers were asked to rate several attributes of a new product. The response on a 7-point semantic differential scale were tabulated and the attribute correlation matrix is constructed. The correlation matrix is given as

Attributes (Variables)	**1**	**2**	**3**	**4**	**5**
1. Taste	1.00				
2. Good buy for money	0.02	1.00			
3. Flavour	0.96	0.13	1.00		
4. Suitable for snacks	0.42	0.71	0.50	1.00	
5. Provides lots of energy	0.01	0.85	0.11	0.79	1.00

Interpret the correlation matrix R, to show whether to go for Factor Analysis.

The factor loadings (obtained by Principal Component Analysis), the cumulative proportion of total sample variance is given below.

Variables			After factor rotation (Varimax)	
	Factor 1	Factor 2	Factor 1	Factor 2
1 2 3 4 5	0.56 0.78 0.65 0.94 0.80	0.82 - 0.53 0.75 - 0.11 - 0.54	0.02 0.94 0.13 0.84 0.97	0.99 - 0.01 0.98 0.33 - 0.02
Cumulative proportion of total sample variance	0.571	0.931	0.507	0.935

Interpret the factor loadings.

EXAMPLE 2

Stock prices on weekly basis is taken for 5 stocks. The Factor Analysis results are shown in the table below.

Variable	Factor 1	Factor 2
1. Allied Chemical 2. Du Pont 3. Union Carbide 4. Exxon 5. Texaco	0.783 0.773 0.743 0.713 0.712	- 0.417 - 0.458 - 0.434 0.472 0.524
Cumulative proportion of total sample variance	0.57	0.83

Interpret the factor loadings.

Example 3

- To explore the factors determining the decision of initiating remanufacturing business

Based on literature survey, the following issues are identified that make remanufacturing business infeasible

Issues	Notation	References
No specific market for remanufactured products	V_1	Ferrer et al. (2003); Thierry et al. (1995)
Relatively few customers in the market	V_2	
Disorganized business sector is already active	V_3	Majumder et al. (2001)
Secondhand market is thriving	V_4	
Customers think refurbished goods to be inferior	V_5	Gungor et al. (1999); Guide et al. (2000)
Mindset of people is not like the western world	V_6	
Economically not profitable	V_7	Mukherjee (2002); Guide et al. (1997); Dekker et al. (2000); Amezquita et al. (1995); Guide et al. (2003);
Technically infeasible	V_8	Ferrer et al. (2003); Giuntini et al. (2003); Hammond et at. (1998); Gungor et al. (1998)
Expertise not available in this area	V_9	Seitz et al. (2004); Giuntini et al. (2003)
No environmental compulsion for remanufacturing	V_{10}	Doppelt et al. (2001); Spicer et al. (2004); Guide et al. (2000); Guide et al. (2001)
Uncertainty in timing, quantity and quality of returns	V_{11}	Ferrer et al. (2003); Guide (2000); Brito et al. (2002); Spicer et al. (2004); Krumwiede et al. (2002); Thierry et al. (1995)
Acquisition of returns is difficult	V_{12}	
Reverse distribution of used products is difficult	V_{13}	
High cost is associated with logistics	V_{14}	

Correlation matrix

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}
V_1	1.00													
V_2	0.91	1.00												
V_3	-0.38	-0.34	1.00											
V_4	-0.59	-0.52	0.74	1.00										
V_5	0.71	0.74	-0.33	-0.47	1.00									
V_6	0.62	0.61	-0.46	-0.50	0.80	1.00								
V_7	0.48	0.55	-0.05	-0.49	0.42	0.39	1.00							
V_8	-0.01	0.04	-0.42	-0.42	0.07	0.24	0.11	1.00						
V_9	0.03	0.05	-0.32	-0.22	0.11	0.22	-0.06	0.69	1.00					
V_{10}	0.39	0.41	0.14	-0.04	0.50	0.37	0.44	-0.19	-0.17	1.00				
V_{11}	-0.27	-0.32	0.60	0.56	-0.27	-0.44	-0.22	-0.73	-0.57	0.29	1.00			
V_{12}	-0.18	-0.19	0.62	0.67	-0.21	-0.40	-0.24	-0.55	-0.44	0.22	0.76	1.00		
V_{13}	-0.20	-0.20	0.48	0.53	-0.14	-0.25	-0.19	-0.69	-0.60	0.05	0.74	0.67	1.00	
V_{14}	-0.26	-0.26	0.55	0.66	-0.12	-0.26	-0.22	-0.60	-0.45	0.09	0.61	0.77	0.84	1.00

This justifies possibility of selecting groups of variables, each representing a single underlying construct (factor)

Existence of high correlation among the following sets of variables,

Set I: V1, V2, V5 and V6,

Set II: V3 and V4,

Set III: V8 and V9 and

Set IV: V11, V12, V13 and V14

Variance explained by the six factors

Factor	Total	% of Variance	Cumulative %
1	3.191	22.792	22.792
2	2.922	20.874	43.666
3	2.571	18.362	62.028
4	1.624	11.600	73.627
5	1.147	8.196	81.824
6	1.133	8.096	89.920

Factor Analysis (PCA Technique & Varimax Rotation)

Variables	Factors					
	1	2	3	4	5	6
V_1	-0.214	0.021	0.914	0.221	0.117	0.107
V_2	-0.138	-0.052	0.903	0.235	0.215	0.101
V_3	0.770	0.157	-0.230	-0.330	0.232	0.184
V_4	0.774	0.125	-0.390	-0.198	-0.255	0.059
V_5	-0.093	-0.062	0.572	0.693	0.108	0.241
V_6	-0.238	-0.171	0.373	0.803	0.123	0.144
V_7	-0.141	-0.012	0.334	0.134	0.877	0.183
V_8	-0.311	-0.822	-0.102	0.110	0.132	-0.114
V_9	-0.077	-0.898	0.040	0.096	-0.176	-0.043
V_{10}	0.131	0.130	0.256	0.259	0.207	0.860
V_{11}	0.454	0.659	-0.170	-0.241	-0.175	0.373
V_{12}	0.763	0.401	0.037	-0.206	-0.196	0.177
V_{13}	0.577	0.716	-0.105	0.119	-0.073	-0.142
V_{14}	0.780	0.489	-0.130	0.165	-0.087	-0.126

Variables that are correlated (absolute correlation being greater than 0.4) with the six rotated factors are as follows.

Factor 1 (F1): V3 (0.770), V4 (0.774), V11 (0.454), V12 (0.763), V13 (0.577) and V14 (0.780)
Factor 2 (F2): V8 (-0.822), V9 (-0.898), V11 (0.659), V12 (0.401), V13 (0.716) and V14 (0.489)
Factor 3 (F3): V1 (0.914), V2 (0.903) and V5 (0.572)
Factor 4 (F4): V5 (0.693) and V6 (0.803)
Factor 5 (F5): V7 (0.877)
Factor 6 (F6): V10 (0.860)

Summary

The remanufacturing business in India is probably governed by the six underlying factors termed as

1. acquisition of returns
2. technology for remanufacturing
3. market
4. customers' attitude
5. profitability
6. legislation

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

# Load data
df= pd.read_csv(“……./Exp3bfi.csv")

#Preprocess Data
print(df.columns)

# Dropping unnecessary columns
df.drop(['gender', 'education', 'age'],axis=1,inplace=True)

# Dropping missing values rows
df.dropna(inplace=True)

print(df.head())
```

```
#Adequacy Test

#Bartlett’s 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)

#Kaiser-Meyer-Olkin (KMO) Test
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(df)
print(kmo_model)
```

```
#Choosing the Number of Factors

# Create factor analysis object and perform factor analysis
fa = FactorAnalyzer(n_factors = 6, rotation=None)
fa.fit(df)
fa.loadings_

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

# 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()
```

```
# Create factor analysis object and perform factor analysis
fa = FactorAnalyzer(n_factors = 6, rotation="varimax")
fa.fit(df)
print(fa.loadings_)

# Get variance of each factors
print(fa.get_factor_variance())
```