EXPERIMENT 4

CO2: Utilize a modeler to create, manage, and run data mining streams.

Aim: To identify underlying factors that group together related symptoms in patients, thereby aiding in the diagnosis and understanding of potential medical conditions.

Objective

To identify and interpret the underlying factors that group related symptoms in patients using Exploratory Factor Analysis. This will aid in understanding the relationships between symptoms and potentially improve diagnostic accuracy and treatment planning.

Theory

Exploratory Factor Analysis (EFA) is a statistical technique used to uncover the underlying structure of a relatively large set of variables. It is used primarily to identify the underlying relationships between measured variables and to detect possible constructs that are not directly observed but inferred from the observed variables.

Conceptual Framework

1. **Latent Variables**: EFA assumes that observed variables (e.g., symptoms) are influenced by underlying latent factors. These latent factors are not directly measured but are inferred from the patterns of correlations among the observed variables.

2. **Common and Unique Variance**: Each observed variable is considered to be influenced by one or more common factors (shared across multiple variables) and unique factors (specific to each variable, including measurement error). The total variance of each variable is partitioned into common variance and unique variance.

Mathematical Foundations of EFA

1. **Correlation Matrix (R)**

The first step in EFA is to compute the correlation matrix of the observed variables. The correlation matrix R is a symmetric matrix that shows the Pearson correlation coefficients between pairs of variables.

2. **Factor Extraction**

Factor extraction aims to identify the underlying factors that explain the correlations between observed variables. Common methods include:

- Principal Axis Factoring (PAF)
- Maximum Likelihood Estimation (MLE)

The goal is to express the correlation matrix R as the product of a factor loading matrix Λ and its transpose plus a unique variance matrix Ψ:

$$R = \Lambda\Lambda' + \Psi$$

3. **Eigenvalues and Eigenvectors**

Eigenvalues and eigenvectors of the correlation matrix are used to determine the number of factors to retain. The eigenvalue decomposition of R is given by:

$$R = VDV'$$

where V is the matrix of eigenvectors and D is the diagonal matrix of eigenvalues.

4. **Factor Loadings**

The factor loading matrix Λ represents the coefficients that map each variable to the underlying factors. If $\mathbf{X}$ is the vector of observed variables and $\mathbf{F}$ is the vector of common factors, then:

$$\mathbf{X} = \Lambda\mathbf{F} + \mathbf{U}$$

where $\mathbf{U}$ is the vector of unique factors (specific to each observed variable).

Sample Medical Data Table

Let's assume we have the following sample data for 10 patients and 5 variables (e.g., different symptoms):

Patient	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5
1	4	5	6	7	3
2	5	6	7	8	4
3	6	7	8	6	5
4	7	8	5	5	6
5	3	4	6	7	2
6	5	5	7	6	3
7	4	6	8	7	4
8	6	7	7	5	5
9	7	8	6	8	6
10	5	6	5	6	4

Step-wise Procedure for EFA in Origin Pro 2024b

Step 1: Input Data

1. Open Origin Pro 2024b.
2. Create a new workbook and enter the data from the sample medical data table.
3. Label the columns appropriately (e.g., Symptom 1, Symptom 2, etc.).

Step 2: Data Preprocessing

1. **Check for Missing Values**: Ensure that there are no missing values in the dataset.
2. **Standardize the Data**: If the variables are on different scales, it may be helpful to standardize them. Origin Pro offers tools for data normalization and standardization.

Step 3: Open EFA Tool

1. Go to the Statistics menu.
2. Navigate to Multivariate Analysis.
3. Select Exploratory Factor Analysis.

Step 4: Configure EFA

1. **Input Data**: Select the data range (all columns with symptoms).
2. **Extraction Method**: Choose an extraction method such as Principal Axis Factoring or Maximum Likelihood. For simplicity, we will use Principal Axis Factoring.
3. **Rotation Method**: Select a rotation method such as Varimax (orthogonal) or Promax (oblique). Varimax is commonly used and will be selected here.
4. **Number of Factors**: Choose the number of factors to extract. You can use criteria such as the Kaiser criterion (eigenvalues greater than 1) or a scree plot to determine this. For this example, we will extract 2 factors.

Step 5: Run EFA

1. Click the OK button to run the analysis.
2. Origin Pro will generate the output, including the factor loadings, communalities, eigenvalues, and a scree plot.

Step 6: Interpret Results

1. **Factor Loadings**: Examine the factor loadings to understand which symptoms load onto which factors. A loading greater than 0.4 is typically considered significant.
2. **Communalities**: Check the communalities to see how much of each variable's variance is explained by the extracted factors.
3. **Eigenvalues and Scree Plot**: Look at the eigenvalues and the scree plot to confirm the number of factors chosen.

Step 7: Report Findings

1. **Factor Structure**: Describe the factor structure, indicating which symptoms are associated with each factor.
2. **Implications**: Discuss the implications of the findings in the medical context. For example, certain symptoms may cluster together, suggesting they are influenced by the same underlying condition.

Example Interpretation

Suppose the results show the following factor loadings for two factors:

Symptom	**Factor 1**	**Factor 2**
Symptom 1	0.85	0.10
Symptom 2	0.80	0.20
Symptom 3	0.30	0.75
Symptom 4	0.25	0.80
Symptom 5	0.15	0.85

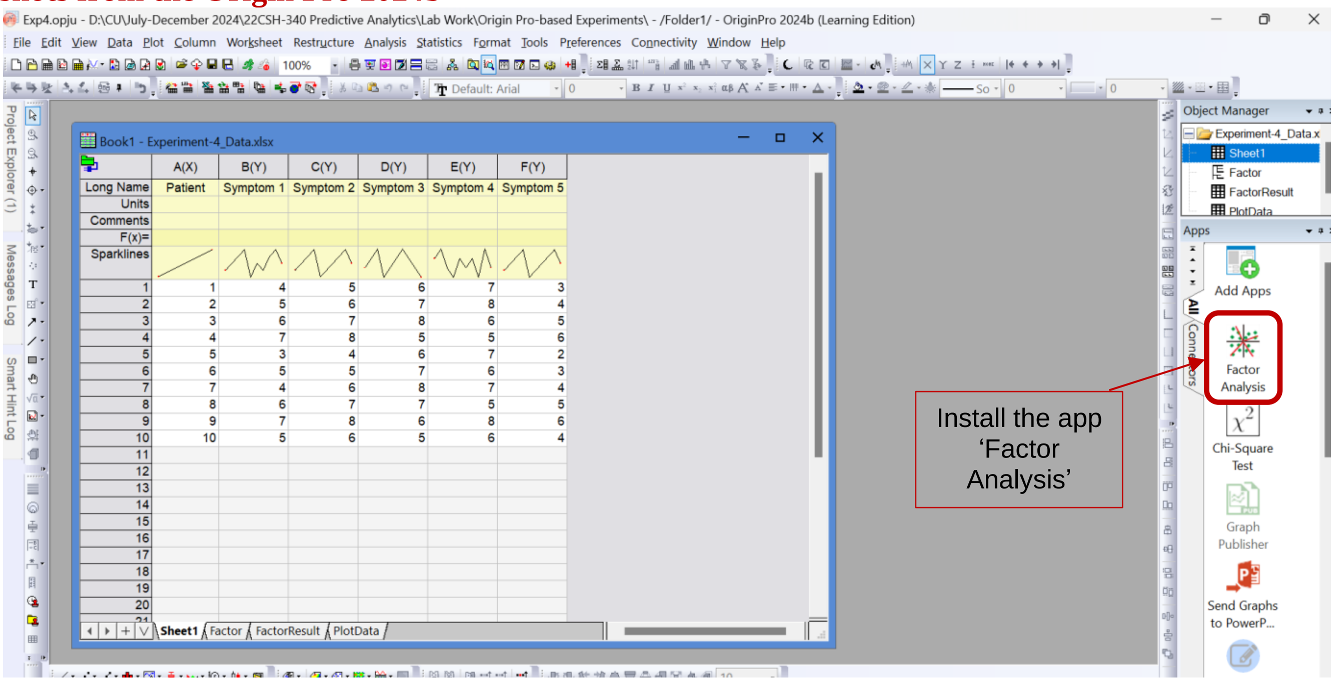
Interpretation:

- **Factor 1**: High loadings on Symptom 1 and Symptom 2 suggest this factor represents a cluster of symptoms likely related to one underlying condition.
- **Factor 2**: High loadings on Symptom 3, Symptom 4, and Symptom 5 suggest this factor represents another cluster of symptoms potentially related to a different underlying condition.

Learning Outcomes:

1. Students will be able to effectively perform Exploratory Factor Analysis using statistical software, including data preparation, computation of correlation matrices, factor extraction, and rotation techniques.
2. Students will be able to interpret the results of EFA, including understanding factor loadings, identifying significant factors, and explaining the relationships between observed variables and underlying latent constructs.
3. Students will be able to apply EFA to practical scenarios, such as analyzing medical symptom data, to identify underlying patterns and inform evidence-based decision-making in research and clinical settings.

Screenshots from the Origin Pro 2024b



Book1 - Experiment-4_Data.xlsx

Factor Analysis (07-08-2024 00:38:37)

Notes

Input Data

Basic Statistics

Descriptive Statistics

	Mean	Standard Deviation
Symptom 1	5.2	1.31656
Symptom 2	6.2	1.31656
Symptom 3	6.5	1.08012
Symptom 4	6.5	1.08012
Symptom 5	4.2	1.31656

Correlation Matrix

	Symptom 1	Symptom 2	Symptom 3	Symptom 4	Symptom 5
Symptom 1	1	0.9359	-0.15627	-0.31254	0.9359
Symptom 2	0.9359	1	-0.07813	-0.2344	1
Symptom 3	-0.15627	-0.07813	1	0.14286	-0.07813
Symptom 4	-0.31254	-0.2344	0.14286	1	-0.2344
Symptom 5	0.9359	1	-0.07813	-0.2344	1

Variance Explained

Factor	Variance	Percentage of Variance (%)	Cumulative (%)
1	3.03667	60.73339	60.73339
2	1.05499	21.09973	81.83312

Method: Principal Components
Type: Correlation Matrix

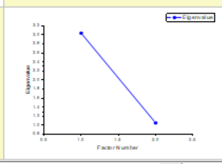
Scree Plot

Variance Plot

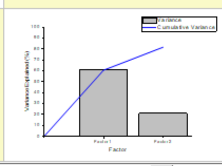
Sheet1 Factor FactorResult PlotData

Book1 - Experiment-4_Data.xlsx

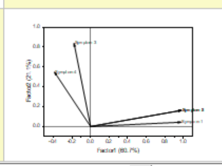
Scree Plot



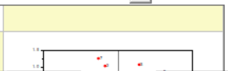
Variance Plot



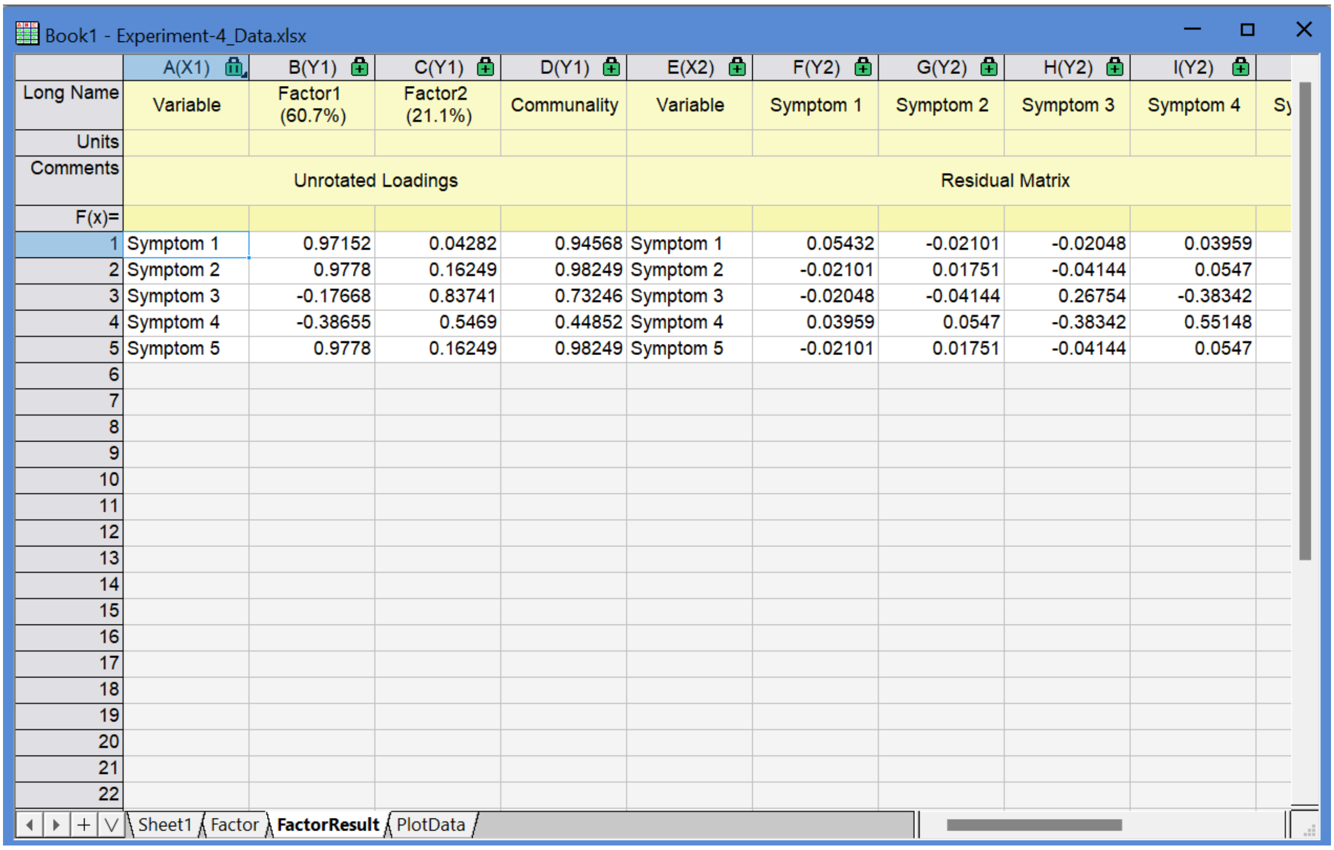
Loading Plot



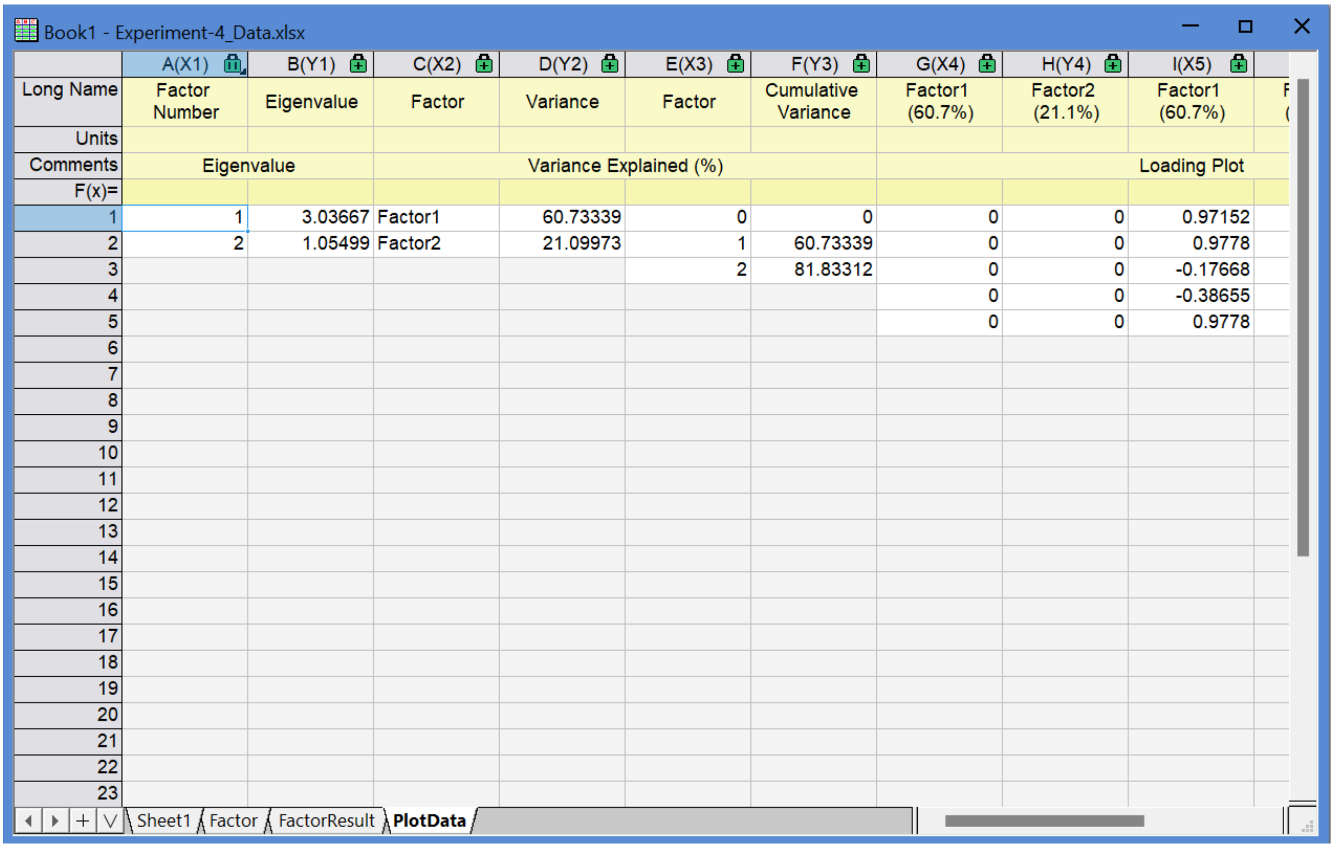
Score Plot



Sheet1 Factor FactorResult PlotData



	A(X1)	B(Y1)	C(Y1)	D(Y1)	E(X2)	F(Y2)	G(Y2)	H(Y2)	I(Y2)
Long Name	Variable	Factor1 (60.7%)	Factor2 (21.1%)	Communality	Variable	Symptom 1	Symptom 2	Symptom 3	Symptom 4
Units									
Comments	Unrotated Loadings				Residual Matrix				
F(x)=									
1	Symptom 1	0.97152	0.04282	0.94568	Symptom 1	0.05432	-0.02101	-0.02048	0.03959
2	Symptom 2	0.9778	0.16249	0.98249	Symptom 2	-0.02101	0.01751	-0.04144	0.0547
3	Symptom 3	-0.17668	0.83741	0.73246	Symptom 3	-0.02048	-0.04144	0.26754	-0.38342
4	Symptom 4	-0.38655	0.5469	0.44852	Symptom 4	0.03959	0.0547	-0.38342	0.55148
5	Symptom 5	0.9778	0.16249	0.98249	Symptom 5	-0.02101	0.01751	-0.04144	0.0547



	A(X1)	B(Y1)	C(X2)	D(Y2)	E(X3)	F(Y3)	G(X4)	H(Y4)	I(X5)
Long Name	Factor Number	Eigenvalue	Factor	Variance	Factor	Cumulative Variance	Factor1 (60.7%)	Factor2 (21.1%)	Factor1 (60.7%)
Units									
Comments	Eigenvalue		Variance Explained (%)				Loading Plot		
F(x)=									
1	1	3.03667	Factor1	60.73339	0	0	0	0	0.97152
2	2	1.05499	Factor2	21.09973	1	60.73339	0	0	0.9778
3					2	81.83312	0	0	-0.17668
4							0	0	-0.38655
5							0	0	0.9778