Homework 2

IE 7275 Data Mining in Engineering

Readings and Practice:

1. Chapter 4: Dimension Reduction
2. Read the book chapter "Principal components and factor analysis.pdf" posted on Blackboard (also attached to the assignment). Practice example problems given in the book chapter.

Problem 1: Perform principal component analysis on NHL.xlsx, which contains statistics of 30 teams in the National Hockey League. The description of the variables is provided in the 'Description' sheet of the file. Focus only on the variables 12 through 25, and create a new data frame.

- Input the new data frame to fa.parallel() function to determine the number of components to extract
- Input the new data frame to principal() function to extract the components. If raw data is input, the correlation matrix is automatically calculated by principal() function.
- Rotate the components
- Compute component scores
- Graph an orthogonal solution using factor.plot()
- Interpret the results

Problem 2: Perform principal component analysis on `Glass Identification Data.xlsx`

- Input the raw data matrix to fa.parallel() function to determine the number of components to extract
- Input the raw data matrix to principal() function to extract the components. If raw data is input, the correlation matrix is automatically calculated by principal() function.
- Rotate the components
- Compute component scores
- Graph an orthogonal solution using factor.plot()
- Interpret the results

Problem 3: Perform factor analysis on `Herman74.cor`, which is a data structure available in the base installation (A correlation matrix of 24 psychological tests given to 145 seventh and eight-grade children in a Chicago suburb by Holzinger and Swineford).

- Input the correlation matrix to fa.parallel() function to determine the number of components to extract
- Input the correlation matrix to fa() function to extract the components. If raw data is input, the correlation matrix is automatically calculated by fa() function.

- Rotate the factors
- Compute factor scores
- Graph an orthogonal solution using factor.plot()
- Graph an oblique solutions using fa.diagram()
- Interpret the results

Problem 4: Perform factor analysis on `breast-cancer-wisconsin.xlsx`, is a multivariate dataset that is used to predict whether a cancer is malignant or benign from biopsy details of 699 patients with 11 attributes. Create a new data frame by removing the variable "BN".

- Calculate the correlation matrix from the new data frame. Visualize the correlation matrix using `pairs.panels` function of the "psych" package. How would you interpret the result in terms of correlation among the variables?
- Input the correlation matrix to fa.parallel() function to determine the number of components to extract
- Input the correlation matrix to fa() function to extract the components. If raw data is input, the correlation matrix is automatically calculated by fa() function.
- Rotate the factors
- Compute factor scores
- Graph an orthogonal solution using factor.plot()
- Graph an oblique solutions using fa.diagram()
- Interpret the results

Problem 5. Perform multidimensional scaling on Vertebral Column Data.xlsx

- Input the raw data matrix to fa.parallel() function to determine the number of components to extract
- Input the raw data matrix to cmdscale() function to perform multidimensional scaling. cmdscale() function which is available in the base installation performs a classical multidimensional scaling.
- Graph an orthogonal solution using factor.plot()
- Interpret the results

Files Included in the Folder:

Homework 2.pdf

PCA and FA Tutorial.pdf

NFL.xlsx

Glass Identification Data.xlsx

Glass Identification Data Description.pdf

Herman74.cor

breast-cancer-wisconsin.xlsx

breast-cancer-wisconsin-description.pdf

Vertebral Column Data.xlsx

Vertebral Column Description.pdf