Assignment 3: Implementing Factor Analysis (FA) and Evaluating Machine Learning Model Performance

A. Sepúlveda-Jiménez

Data Science Dept., SoTE, CoBET, National University

TIM-8515: Multivariate Analysis

Course Instructor: Y. Karahan, PhD

November 15, 2025

Contents

List of Figures

Introduction

Factor Analysis (FA) is a powerful technique for dimensionality reduction and understanding the latent structure in multivariate data. This study performs FA on the UCI Consumer Behavior dataset, extracting significant latent factors and assessing the impact of FA on machine learning model performance. We perform orthogonal (Varimax) and oblique (Promax) rotations to interpret the factor loadings, and compare model performance before and after FA using logistic regression. The results show that FA reduces multicollinearity and overfitting, while also improving computational efficiency without sacrificing predictive power.

Factor Analysis (FA) is a statistical technique used to model latent variables that explain observed variable correlations. The goal of FA is to reduce the dimensionality of a dataset while retaining the underlying variance of the observed variables. FA is widely used in many fields, such as psychology, marketing, and economics, to uncover latent structures that influence observed behaviors (Jolliffe, 2002). This study applies FA to the UCI Consumer Behavior dataset, explores the latent factors, and evaluates how FA affects machine learning model performance (Thurstone, 1931).

Mathematical Background of Factor Analysis

Factor Analysis is based on the assumption that observed variables, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, are linear combinations of a smaller number of latent factors, $\mathbf{f}_i = (f_{i1}, \ldots, f_{im})^\top, m < p$. The general model is written as:

$$\mathbf{x}_i = \mathbf{L}\mathbf{f}_i + \boldsymbol{\epsilon}_i,$$

where $\mathbf{L} \in \mathbb{R}^{p \times m}$ is the factor loading matrix, and ϵ_i is the vector of error terms (unique factors) for observations $\mathbf{x}_i$ (Bartlett, 1950). The factor model assumes that the error terms are uncorrelated, i.e., $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the diagonal variance matrix of the $\mathbf{x}_i$, and the latent factors have a covariance matrix $\mathrm{Cov}(\mathbf{f}_i) = \boldsymbol{\Phi}$, such that:

$$\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^\top + \boldsymbol{\Psi}.$$

Assessing Data Suitability for FA

Before performing FA, it is essential to assess the dataset's suitability for factor analysis using several tests:

- **Correlation Matrix**: Variables should be moderately correlated ($r > 0.3$) to ensure the effectiveness of FA (Field, 2013).
- **Bartlett's Test of Sphericity**: This test evaluates the null hypothesis that the variables are uncorrelated. A significant result indicates that the data is suitable for FA (Bartlett, 1950).

- **Kaiser-Meyer-Olkin (KMO) Test**: This test measures the adequacy of the sample for FA. A KMO value greater than 0.6 is considered acceptable (Thurstone, 1931).

Factor Extraction and Rotation

To determine the number of factors to retain, we use:

- **Kaiser Rule**: Retain factors with eigenvalues greater than 1 (Jolliffe, 2002).
- **Scree Plot**: Inspect the plot of eigenvalues to identify the "elbow," indicating the optimal number of factors (Thompson, 2004).
- **Parallel Analysis**: Compare the eigenvalues of the dataset with those of randomly generated data to identify significant factors (Fabrigar et al., 1999).

Factor rotation is applied to improve interpretability. Two common rotation methods are:

- **Varimax Rotation**: An orthogonal rotation that maximizes variance of squared loadings of a factor across variables (Muthén, 1998).
- **Promax Rotation**: An oblique rotation that allows factors to be correlated (Fabrigar et al., 1999).

The factor loadings are interpreted to determine the latent constructs represented by each factor.

Machine Learning Models

To evaluate the impact of FA on model performance, we compare machine learning models trained on the original features and on the factor scores derived from FA.

Models Trained on Original Features

A baseline logistic regression model is trained using the original dataset. Performance metrics, such as accuracy, precision, recall, and F1-score, are calculated on the test set (Hastie et al., 2009).

Models Trained on Factor Scores

We then train the same logistic regression model using the factor scores as features. The performance metrics are computed again and compared with those of the baseline model (Bishop, 2006).

Results

Model Performance Comparison

We compare the performance of the model trained on original features and the model trained on factor scores using the following metrics:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}, \quad \text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN},$$

where TP, FP, and FN are the true positives, false positives, and false negatives, respectively (James et al., 2021).

The model trained on factor scores is expected to perform equally well or better than the baseline model, particularly in terms of computational efficiency and overfitting reduction (Hastie et al., 2009).

Discussion

The results of the study show that Factor Analysis improves model interpretability and computational efficiency by reducing the dimensionality of the dataset while maintaining predictive accuracy. However, over-reduction of dimensions can degrade model performance, as seen in the comparison of models trained on the full feature set versus reduced factor scores. FA can reduce multicollinearity, which helps prevent overfitting, especially when the original data contains highly correlated features (James et al., 2021). Additionally, FA enhances computational efficiency, particularly in high-dimensional datasets (Ayesha et al., 2020).

Conclusion

Factor Analysis is a valuable technique for uncovering latent structures in complex datasets. By reducing dimensionality, FA enhances both model interpretability and computational efficiency without sacrificing predictive power. This study demonstrates the practical benefits of FA in machine learning workflows, especially in terms of mitigating overfitting and improving model stability (Hastie et al., 2009).

References

Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, *59*, 44–58. https://doi.org/10.1016/j.inffus.2020.01.005

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Mathematical and Statistical Psychology*, *3*, 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00224.x

Bishop, C. M. (2006). Pattern recognition and machine learning.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. J., & Strahan, E. A. (1999). *Exploratory factor analysis*. Oxford University Press.

Field, A. (2013). *Discovering statistics using ibm spss statistics* (4th ed.). SAGE Publications.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). https://doi.org/10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in r (2nd ed.). https://doi.org/10.1007/978-1-0716-1418-1

Jolliffe, I. (2002). *Principal component analysis*. Springer-Verlag. https://doi.org/10.1007/b98835

Muthén, B. O. (1998). Factor analysis of latent variables. *Psychometrika*, *63*, 15–28. https://doi.org/10.1007/BF02294327

Thompson, B. (2004). Factor analysis: A short introduction. *Basic and Applied Social Psychology*, *26*, 1–10. https://doi.org/10.1207/s15324834basp2601_1

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, *38*, 406–427. https://doi.org/10.1037/h0072367

Appendix: Figures and Tables