Multivariate Data Analysis

Huong Viet Ngo

1 Introduction

During my time in the Statistics and Probability Association's Directed Reading Program, I was introduced to various multivariate data analysis methods through a book titled *An Introduction to Applied Multivariate Analysis with R*. Although I have taken introductory statistics courses that introduced me to the notions of exploratory data analysis and statistical inference, the material only featured examples that focused on univariate analysis. Thus, I was quite excited to learn more complex methods that would support my work with more common datasets that would be utilized best when studying the interaction of many variables all at once. Following this introduction, I will introduce some notable methods I learned and discuss my experience with them.

2 Methods Learned

2.1 Principal Component Analysis

The first and most important multivariate analysis method I learned was **Principal Component Analysis** (**PCA**). This method is not only powerful but is the foundation for other techniques and visualizations that rely on working with data in two or three dimensions. PCA tackles the infamous **Curse of Dimensionality**, which refers to the problem of there being too many variables in a dataset and thus can be difficult to work with. The method's goal is to reduce the dimensionality of a multivariate dataset while trying to retain as much of the variation of the original dataset as possible. The result of this method is a small number, generally 2 or 3, of new variables that are called **principal components**.

This unit in the book was the most difficult to tackle, given the complex mathematics of the method's implementation. However, I was determined to gain full comprehension of it, before even realizing how critical this method is. With the help of my amazing mentor, Sarah Teichman, whom I bombarded with many questions, I was able to fully grasp how this amazing technique works.

I would like to mention another method that is similar to PCA but still worth mentioning: **Multidimensional Scaling**. While PCA is applied to the usual multivariate data matrices (dataset), Multidimensional Scaling is applied to distance matrices which is derived from the usual multivariate data matrix.

2.2 Cluster Analysis

As stated in the book, **Cluster Analysis** is a generic term for a wide range of numerical methods with the common goal of uncovering or discovering groups or clusters of observations that are homogeneous and separated from other groups. As humans, we tend to organize or categorize things based on how similar they are. Like data, which in its raw form, is simply randomness. Cluster Analysis is a tool for us to sort through that randomness and discover interesting findings about data that informs us a little bit about the world. Because of this, it was very appealing to use this technique to learn more about the airline operations data I was working with for my final project.

K-means clustering aims to partition n observations in a set of multivariate data into k clusters and k is given by minimizing a numerical criterion called the **within-group sum of squares**. The implementation starts of piggybacking on another cluster analysis method that uses **hierarchical clustering** to initialize a partition of the observations into k clusters. We then move the observations from its own cluster to another cluster and calculate the change in the numerical criterion value. We

make the change that results in the greatest improvement in the criterion value. The calculation and changes are repeated until there is no further improvement of the criterion value.

This method proved to be quite useful, along with principal components and parallel coordinates plots to better understand the behavior of airline carriers pre and post COVID.

3 Follow-up Work

There will be more work done on this project to discover more interesting conclusions by improving on the methods used in the current iteration and including a supplementary data set (traffic-capacity). Moreover, the next work will also involve other multivariate exploratory methods such as **Analysis of Repeated Measurements** to characterize changes in response variables and determine any explanatory variables most associated with any change.

4 Acknowledgements

I would like to thank the Statistics and Probability Association's Directed Reading Program for granting me this rare opportunity to work with seasoned mentors in my field of interest and learn something amazing. I would also like to thank my awesome and engaging mentor, Sarah Teichman, for her generous help and advice.