Applied Multivariate Analysis

Project #4 - Dimension Reduction

Omar Moustafa

900222400

Nour Kahky

900221042

MACT 4233

Spring 2025

Introduction

Complete Source of the Data:

The dataset to be used and analyzed is sourced from Kaggle and serves the purpose of diabetes prediction based on biomedical data. The data was collected and displayed to reflect a range of metabolic health conditions, and at the end, includes a categorical variable which identifies which category of diabetes-level each patient falls under, with those categories being diabetic, prediabetic, and non-diabetic. Therefore, this dataset provides a strong foundation for applying various statistical procedures to determine patterns in large multivariate data and then comparing and contrasting such learning methods.

What Are the Main Objectives?

The primary objective of this study is to apply two different dimension-reduction techniques to a real-world dataset with a minimum of ten quantitative variables and then compare and contrast their levels of success. These two methods, **Principal Component Analysis (PCA)** and **Multidimensional Scaling (MDS)**, will be implemented and then compared and contrasted to determine which of these essential methodologies in multivariate statistics generates more success than the other.

From the chosen dataset, PCA will be applied to the centered data matrix, whereas MDS will be applied to the distance matrix. The success of each method will be evaluated through visualizations and quantitative criteria, such as explained variability and goodness of fit. By the end, it will be determined which technique better reduces the dimensionality with a genuine understanding of how each method operates and what aspects of the structure of the dataset each method is based on.

Description of the Data

What Are the Observations?

The dataset includes 1000 observations, each representing a patient's biomedical profile. Every observation corresponds to a unique patient and contains ten quantitative variables regarding age, kidney function, blood sugar levels, lipid profiles, and more. These measures are commonly utilized when needing to assess one's risk of diabetes and current medical conditions.

Full Definition of the Variables:

Variable Name:	**Variable Description:**	**Variable Type:**	**Unit of Measurement:**
ID	ID Number	Unique Identifier	N/A
No_Pation	Another identifier for the patient	Unique Identifier	N/A
Gender	Gender of the patient	Categorical (Binary)	F = Female, M = Male
AGE	Age of the patient	Quantitative	Years
Urea	Urea level in the blood	Quantitative	mg/dL or mmol/L
Cr	Creatinine level in the blood	Quantitative	mg/dL or μmol/L
HbA1c	Avg. blood sugar levels over the past 2-3 months	Quantitative	Percentage (%)
Chol	Cholesterol level in the blood	Quantitative	mg/dL or mmol/L

TG	Triglyceride levels in the blood	Quantitative	mg/dL or mmol/L
HDL	High-density lipoprotein cholesterol level	Quantitative	mg/dL or mmol/L
LDL	Low-density lipoprotein cholesterol level	Quantitative	mg/dL or mmol/L
VLDL	Very low-density lipoprotein cholesterol level	Quantitative	mg/dL or mmol/L
BMI	Body fat based on height and weight	Quantitative	kg/m^2
CLASS	Class label indicating the diabetes status of the patient	Categorical	N = Non-diabetic P = Prediabetic Y = Diabetic

Table 1

Data Analysis

Principal Component Analysis (PCA):

Principal Component Analysis (PCA) is a powerful technique for dimensional reduction of multivariate data, where it does so while still maintaining as much variability as possible. By transforming the original ten quantitative variables into a set of uncorrelated principal components, the PCA methodology allows for the identification of the key axes of variability within the dataset.

First and foremost, the eigenvalues of the PCA must be checked for any negative values, because if so, they must be set to zero before proceeding. The reason behind negative values arising in such scenarios is due to rounding errors, and changing them to zero would be the following step, as negative values do not contribute any meaningful variance. The output to this particular check was as follows:

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6
Standard deviation     1.4124932 1.2749371 1.2173538 1.0457334 1.0196633 0.95144498
Proportion of Variance 0.1995137 0.1625465 0.1481950 0.1093558 0.1039713 0.09052475
Cumulative Proportion  0.1995137 0.3620601 0.5102552 0.6196110 0.7235823 0.81410708
                           Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.75980216 0.70836251 0.65303005 0.59447804
Proportion of Variance 0.05772993 0.05017774 0.04264483 0.03534041
Cumulative Proportion  0.87183702 0.92201476 0.96465959 1.00000000
```

Table 2

Analyzing the output above, it can be seen that there are no negative eigenvalues, as all of the standard deviation values are positive. Thus, no action is required. This confirms that the correlation matrix used for PCA is **positive definite (p.d.)** and that the solution is indeed valid, coming from a mathematical perspective.

Next comes the thorough analysis of the effectiveness of these components at capturing the underlying structure of the dataset. This is achieved by visualizing the **Scree Plot,** as it is vital for observing the amount of variance explained by each component. This plot is a crucial tool for helping determine the ideal or optimal number of components to retain while balancing simplicity with valuable information.

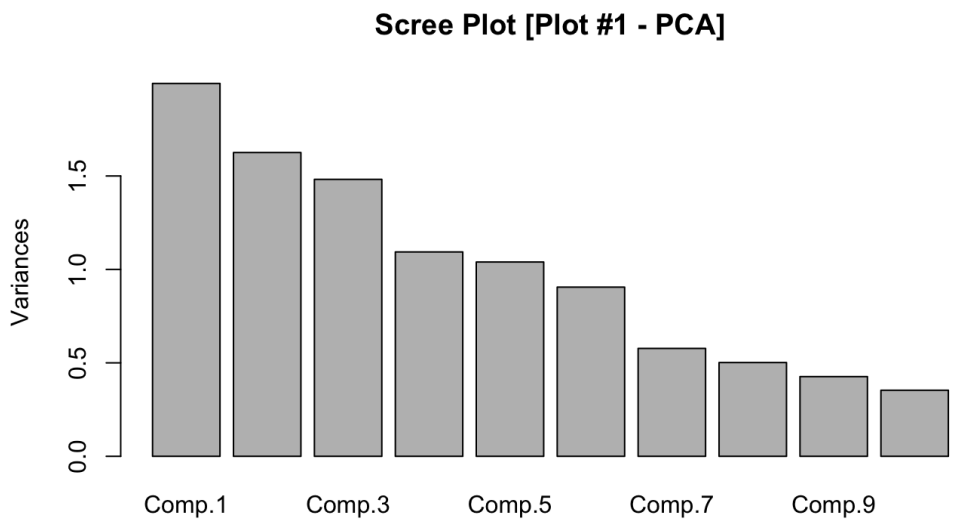

Figure 1

Analyzing the PCA scree plot above, it can be seen that the variances of the principal components are ordered in decreasing order. The first of the ten components, **Comp.1**, captures a great portion of the variance within the data, with the amount of variance explained by the rest of the components continuously decreasing until the end.

The steep decline after the third component, **Comp.3,** suggests that most of the structure of the data can be captured in approximately three components; no more than that is necessary. Specifically, if the first two to three components can explain a large portion of the variability, then those first couple of components are enough to summarize the original dataset with as little loss of information as possible. Another similarly steep decline happens after **Comp.6**, which further emphasizes that the number of components needed to explain the variability within this dataset is only a few and not excessive.

Overall, this plot assists in determining the optimal or ideal number of components to retain without jeopardizing the amount of lost information or becoming too complex.

Multidimensional Scaling (MDS):

In contrast to PCA, which aims to maximize variance, Multidimensional Scaling (MDS) projects data into a lower-dimensional space while preserving the pairwise distances between observations. This process is highly beneficial when the aim is to visualize existing similarities or differences in high-dimensional datasets.

To evaluate the effectiveness of this second methodology, the two-dimensional MDS projection, which plots the transformed coordinates of each observation, is examined. The idea behind arranging these transformed points is to help identify existing outliers, plausible clusters, and other patterns that reflect the structure of the original data. Therefore, MDS offers more perceptive and visual insights that complement the variance-based methodology of PCA.

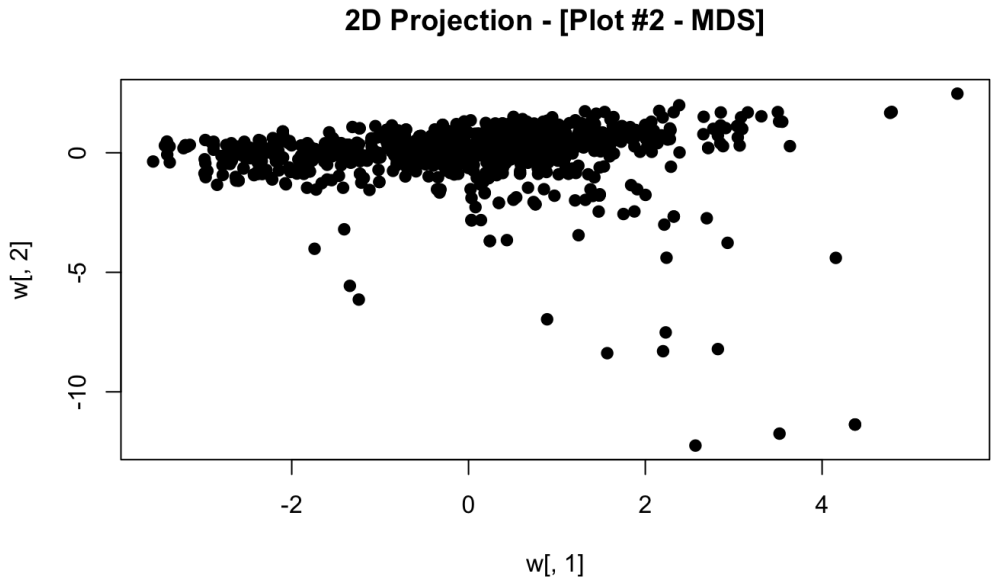

Figure 2

Analyzing Figure 2, the MDS two-dimensional projection visualizes the data in a reduced space while retaining the original pairwise distribution and distances. Patients with similar biomedical profiles or histories are represented by observations that are closer together on the MDS plot. On the other hand, patients with opposing biomedical profiles or histories are

represented by points that are further apart. Unlike PCA, which focuses on variability, MDS places more emphasis on maintaining the structure of relationships between observations. Because of this, MDS is especially helpful when maintaining relational structures between data points is more significant than the axes' interpretability.

Building on this, the shape and spread of the MDS configuration also assist in evaluating the appropriateness of reducing the data to being two-dimensional. From the figure above, even though some loss of information is expected and almost inevitable, the clear compact clustering of the majority of the points and separation of just a few points. This is highly indicative of MDS successfully retaining most of the information of the full-dimensional data. This layout reveals that MDS is a complementary and visually informative technique for dimensional reduction and analysis.

PCA vs. MDS:

PCA and MDS serve the same purpose and ultimate goal of dimensionality reduction, yet they accomplish this through fundamentally different principles. Starting with PCA, it heavily emphasizes the importance of maximizing variability captured in consecutively orthogonal components, which are derived from the original quantitative variables. Therefore, this eases the understanding of which combinations of features lead to most of the variability within the data. In this analysis, the resulting Scree Plot from the PCA implementation showed that only three components capture about half of the variability, revealing that most of the structure can be efficiently explained with an almost minimal number of linear combinations.

On the contrary, the MDS technique works to preserve pairwise distances between observations, regardless of the original variable axes. As a result, it offers a mapping that

preserves distance, which is perfect for displaying similarities and contrasts between samples. In contrast to PCA, the two-dimensional MDS projection produced a much more compressed plot with overlapping clusters, but it did maintain most of the relational structure.

Overall, of the two implemented dimension-reduction techniques, it is concluded that PCA was the one that generated more success in this specific study. Not only did it explain a larger proportion of the variance in fewer dimensions, but it also provided more precise guidance for dimensionality reduction based on the structure of its eigenvalues. The clearly steep drop-off in Figure 1 led to a straightforward confirmation of the number of required components for variability explanations. While MDS did indeed offer valuable visuals that lean into patient similarities and differences, it did not offer as many quantitative insights, and instead displayed the existing overlap between the observations in a reduced two-dimensional figure.

Conclusion

In conclusion, this project implemented two renowned dimension-reduction techniques, which were **Principal Component Analysis (PCA)** and **Multidimensional Scaling (MDS)**, to a real-world biomedical dataset that works to predict the level of diabetes that its patients have. Diving deeper, this study examined ten quantitative health predictors across a total of 1000 patients to assess the success and structure of each method when it comes to simplifying the data without leading to an excessive loss of information or important patterns.

The resulting plots revealed that PCA offered a more interpretable dimension-reduction figure. Only a very few number of components were required to account for a considerable portion of the variance, as indicated by the scree plot in Figure 1. Opposing this, MDS, although it is a useful and sufficient method when needing to visualize proximity relationships among the

observations, which in this case are medical patients, it ended up illustrating greater overlap between said observations in its two-dimensional display, and did not provide a quantitative threshold for dimensional adequacy. Therefore, it can be confidently confirmed that PCA was the most appropriate and successful of the dimensional-reduction methods that were implemented throughout this study. All in all, this strategy heavily emphasizes how crucial it is to choose analytical tools according to the objectives of the study as well as the characteristics and structure of the data.

References

Patel, Marshall. *Diabetes Prediction Dataset*. May 9, 2025.

https://www.kaggle.com/datasets/marshalpatel3558/diabetes-prediction-dataset-legit-dataset