Principal Component Analysis (PCA)

Introduction

Principal Component Analysis (PCA) is a statistical technique used to reduce the complexity of datasets. By projecting the data onto a new coordinate system, PCA identifies the directions (principal components) that capture the most variance in the data. This is particularly useful when dealing with high-dimensional datasets, allowing them to be represented in fewer dimensions without significant information loss.

Key Concepts

PCA seeks to identify principal components—orthogonal directions in the data that maximise variance. These components are linear combinations of the original variables and are uncorrelated with each other.

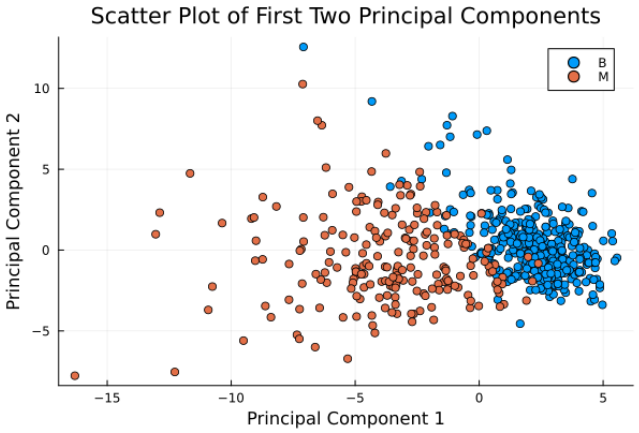

Figure 1: Example PCA

How Does PCA Work?

Step 1: *Standardise the Data*. Subtract the mean of each variable from the dataset to centre the data around the origin. This results in variables with a mean of zero, which is essential for calculating the covariance matrix correctly. We also need to divide each observation by the standard deviation, so that the standard deviation becomes 1.

Step 2: *Compute the Covariance Matrix*. Calculate the covariance matrix to understand how variables in the data vary together. The covariance matrix summarises the correlations between variables.

Step 3: *Perform Eigenvalue Decomposition*. Decompose the covariance matrix into eigenvalues and eigenvectors. The eigenvectors represent the direction of the principal components, and the eigenvalues indicate the magnitude of variance captured by each component.

Step 4: *Project Onto Principal Components*. Select the top k principal components that capture the most variance. Project the original data onto these components to obtain a reduced-dimensional representation.

Applications of PCA

- **Data Compression**: Reduces the number of dimensions in large datasets while retaining the most important information. This saves storage space and computational resources.
- **Visualisation**: Simplifies high-dimensional data to 2 or 3 dimensions for easy visualisation. This helps in identifying patterns and relationships within the data.
- **Noise Reduction**: Eliminates less significant components that may represent noise, which enhances the signal quality of the data.
- **Feature Extraction**: Identifies the most influential variables by capturing the variance, which can help with the development and implementation of machine learning algorithms.

Conclusion

PCA is helpful for reducing dimensionality and simplifying complex datasets. By focusing on the directions with the greatest variance, it can facilitate efficient data analysis and visualisation, as well as identify key features for further analysis.