Dimensionality Reduction Methods in Modern Multivariate Analysis

Saadhana Ganesa Narasimhan
201703255

Supervisor: Rukia Nuermaimaiti

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analytics
as part of the degree of

Master of Science in Data Science and Analytics

The University of Leeds, School of Mathematics

September 2024

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

UNIVERSITY OF LEEDS

School of Mathematics

Declaration of Academic Integrity for Individual Pieces of Work

I declare that I am aware that as a member of the University community at the University of Leeds I have committed to working with Academic Integrity and that this means that my work must be a true expression of my own understanding and ideas, giving credit to others where their work contributes to mine.

I declare that the attached submission is my own work.

Where the work of others has contributed to my work, I have given full acknowledgement using the appropriate referencing conventions for my programme of study.

I confirm that the attached submission has not been submitted for marks or credits in a different module or for a different qualification or completed prior to entry to the University.

I have read and understood the University’s rules on Academic Misconduct. I know that if I commit an academic misconduct offence there can be serious disciplinary consequences.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties to verify that this is my own work, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and I wish to have taken into account.

Student Signature:	Saadhana. G	**Student Number:**	201703255
Student Name:	Saadhana Ganesa Narasimhan	**Date:**	31/08/2024

Please note:

When you become a registered student of the University at first and any subsequent registration you sign the following authorisation and declaration:

"I confirm that the information I have given on this form is correct. I agree to observe the provisions of the University's Charter, Statutes, Ordinances, Regulations and Codes of Practice for the time being in force. I know that it is my responsibility to be aware of their contents and that I can read them on the University web site. I acknowledge my obligation under the Payment of Fees Section in the Handbook to pay all charges to the University on demand.

I agree to the University processing my personal data (including sensitive data) in accordance with its Code of Practice on Data Protection http://www.leeds.ac.uk/dpa . I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material. I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity.'"

Abstract

Dimensionality reduction methods are crucial in handling high-dimensional datasets in modern multivariate analysis, providing the benefits in data visualization, noise reduction, and computational efficiency. This study investigates the application of dimensionality reduction techniques like Principal Component Analysis (PCA), Principal Component Regression (PCR), Kernel-Principal Component Analysis (K-PCA). The dataset contains multiple features representing the video characteristics and other metadata. The high dimensionality and potential multicollinearity of these features pose challenges for data interpretation. A comparative analysis was conducted using the dataset, by applying each technique like PCA which transforms the original variables into a set of principal components and evaluating the maximum variances in the data, Principal Component Regression explores the relationship between these components and the target variable, providing insights into the most significant predictors., Kernel-Principal Component Analysis, a radial basis function (RBF) kernel. The resulted outcomes highlight the effectiveness of these dimensionality reduction techniques in simplifying the dataset while preserving the essential information.

Contents

List of Figures

Chapter 1

Introduction

In the era of big data, modern multivariate analysis has emerged as a important tool for extracting meaningful insights from complex, high-dimensional datasets. As researchers and organizations increasingly rely on a data-driven decision-making, the ability to analyze and interpret the multivariate data where multiple variables are measured simultaneously on each observation has becomes essential.

High-dimensional data, characterized by a large number of variables or features, can lead to issues such as over-fitting, increased computational burden, and difficulties in visualizing and interpreting the data. To address these challenges, dimensionality reduction techniques have become indispensable in the field of the machine learning. These techniques help in simplifying the data by reducing the number of variables while retaining the most relevant information.

1.1 Multivariate Analysis

Multivariate Analysis is a set of statistical methods designed for analyzing data by involving the multiple variables simultaneously. This allows researchers to explore the interconnections between variables, identify patterns in complex datasets. By considering the multiple dimensions of data simultaneously, multivariate analysis provide a deeper insights than traditional univariate or bivariate methods.

Everitt, B. & Hothorn (2011) said when researchers collect values for multiple random variables from a group of individuals or various other items of interest, they generate multivariate data. This results in vector-valued or multidimensional observations for each subject or item. Such data are prevalent across numerous fields, and it is reasonable to assert that the majority of real-world datasets encountered in practice are multivariate.

The need for a new approach to multivariate analysis has emerged due to three recent developments such as the traditional methods of multivariate analysis, nature of the inquires and the costs associated with the data storage and processing as proposed by (Izenman & A.Julian , 2008). These developments have greatly enhanced the feasibility of analysing complex data with greater efficiency.

Multivariate statistical analysis involves the simultaneous examination of a collection of random variables. It goes beyond the analysis of a single variable, enabling the calculation of measures of location and variation, checking distributional assumptions, and detecting outliers. Multivariate analysis enhances individual univariate analyses by incorporating information about the relationships between all variables into the statistical analysis, providing a more comprehensive understanding of the data.

1.2 Literature Review

Principal Component Analysis is one of the most widely used technique in the field of dimensionality reduction. Derived from the Pearson, K. (1901) and Hotelling, H. (1933) formulation in 1933, PCA transforms the data into a lower-dimensional space while retaining the maximum variance. It operates by identifying the principal components-orthogonal directions in the feature space that capture the most variance in the data. According to Jolliffe, I. T (2002), comprehensive text on PCA elucidates the mathematical foundations and various applications across domains, highlighting its effectiveness in reducing the dimensions of large data sets without significant loss of information. However, PCA assumes linear relationships among the variables, which limits its applicability to more complex, non-linear datasets. Jackson, J. E. (2005) book discusses the intricacies of PCA, including the determination of the number of components to retain, the interpretation of component scores, and the application of PCA to various types of data. Also emphasizing the importance of understanding the assumptions behind PCA, such as linearity and the importance of scale in the data. Jolliffe, I. T. & Cadima, J. (2016) reviews PCs application in modern data analysis, image compression, and environmental data. Their review highlighted on PCs versatility, though against the over-reliance PCA for the data interpretation without considering its limitation such as the potential loss of information.

Kernel Principal Component Analysis extends the capabilities of PCA by enabling the analysis of non-linear data structures. This technique, developed by Schölkopf, B. et al (1998), uses kernel functions to project the original data into a higher-dimensional feature space where the linear separability might be achieved. KPCA has been particularly useful in fields such as image and signal processing, where the data often exhibit non-linear patterns. According to the work of Y. Gu, Y. Liu, & Y. Zhang. (2008), KPCA has proven effective in anomaly detection in hyperspectral imagery, demonstrating the techniques utility in handling the complex data distributions that standard PCA cannot adequately address. The flexibility of KPCA, however, comes at the cost of increased computational complexity and the challenge of choosing an appropriate kernel function. The work of Mika, S. et al. (1998) further refined by exploring its application in pattern recognition and classification tasks, demonstrating its superiority over linear PCA in capturing nonlinear relationships in data. The research underscores the flexibility, particularly in the contact of image recognition, where the data often resides on non-linear manifolds.

Schölkopf & Smola (2002) book provides a detailed exposition of the kernel trick, which is central to KPCA, and its application to a variety of machine learning algorithms, including the support vector machines and kernel ridge regression. The work is crucial for understanding the mathematical foundations of KPCA.

Principal Component Regression combines PCA with regression analyse to address mutlicollineartiy issues in datasets where predictor variables are highly correlated. PCR first applies PCA to extract principal components from the predictors, which are then used in a regression model. Frank & Friedman (1993) provided a critical evaluation of PCR, comparing it with the Partial Least Square Regression (PLSR) and Ridge Regression. Their analysis reveals that while the PCR is effective in reducing dimensionality and mitigating multicollinearity, it may not always yield the best predictive performance compared to other technique. They argue that PCRs reliance on variance as the criterion for selecting components may overlook the components with low variance. Their work was essential for understanding the trade-offs involved in using PCR, particularly in comparison to the other regression techniques. Another significant contribution to the literature on PCR is the work by Jolliffe (1982) who explored the partial challenges of applying PCR, particularly in selecting the number of components to retain. This paper discusses the potential pitfalls of PCR, such as the arbitrary nature of component selection and the risk of discarding relevant information, They also suggests some of the alternative approaches such as cross-validation to determine the optimal number of components to include in the regression model.

Chapter 2

Data Description

Izenman & A.Julian (2008) suggests that most multivariate datasets can be represented in a rectangular format, familiar from spreadsheets, where each row corresponds to the variable values for a particular unit in the dataset, and each column corresponds to the values taken by a specific variable.

According to Izenman & A.Julian (2008) multivariate data are ubiquitous, as illustrated by the following examples such as Psychologists and behavioral scientists often collect data on multiple cognitive variables from a wide range of subjects. Similarly, educational researchers analyze student exam scores across the different subjects. Archaeologists may record various measurements related to artifacts of interest, while environmentalists evaluate pollution levels across different cities, alongside other relevant characteristics pertaining to climate and human ecology.

This chapter introduces the dataset used throughout the project, with a summary, pre-process and steps to be taken to prepare the data ready for analysis.

2.1 Data Explanation

This investigation of what makes a trend on YouTube is utilized to understand the characteristics and the factors influencing the trendiness of the YouTube videos. This dataset is taken from the Statso.io website, available at statso.io/what-makes-you-trend-on-youtube-case-study contains multiple variables that provide comprehensive insights into various aspects of the YouTube videos, including metadata, performance metrics, and engagement statistics.

The dataset includes several numerical variables that quantify the aspects of video performance, such as ‘view_count’ , ‘like_count’ , ‘dislike_count’ , ‘favorite_count’ , and ‘comment_count’. These variables are used to capture the engagement metrics of the videos and are critical for understanding the audiences interaction. There are six numerical variables in total, including ‘category_id’, which is represented numerically, categorizes the videos into different content types.

Additionally, the dataset features several categories variables, such as ‘video_id’ , ‘title’ , ‘description’, ‘published_at’, ‘channel_id’, ‘channel_title’, ‘tags’, ‘duration’, ‘definition’, and ‘caption’. These variables provide qualitative details about the each video, including the unique identifier ‘video_id’, textual content like title and description, and other metadata contents like channel id and title. There are ten categorical variables, reflecting different dimensions of each videos identify and presentation.

2.2 Data preprocessing

The Data preprocessing phase is the fundamental step to ensure the quality and suitability of the dataset for the dimensionality reduction techniques such as PCA, KPCA, and PCR.

The initial step involved in generating the summary statistics to understand the datasets overall structure. This analysis reveals substantial variability in key features such as ‘view_count’ and ‘like_count’, indicating a highly skewed distribution. Additionally, categorical features like ‘category_id’ showing eight distinct categories, highlighting the datasets diversity in content types.

Following the summary statistics, the data quality examination was conducted on a subset of the data. This includes in checking for unique values and potential anomalies. The analysis identified the columns with little variability, such as ‘definition’ and ‘favorite_count’, which does not contribute the meaningful information to the further analysis. The presence of outliers in key numerical features suggested that the careful consideration would be needed to manage effectively, preventing skewed results in the dimensionality reduction techniques.

2.2.1 Exploratory Data Analysis

Using predefined method in Python,

```
summary_stats = df.describe()
print(summary_stats)
```

The numerical features, such as ‘views’, ‘likes’, ‘dislikes’, and ‘comment count’, show significantly variability, exemplified by an average view count of approximately 3,000,000 and a standard deviation of around 5,000,000. The categorical features, including ‘category’, ‘comments disables’, ‘ratings disabled’, reveal that ‘Entertainment’ is the most frequent category among 15 unique genres, and most videos do not have comments or ratings disabled.

The descriptive analysis of the initial 20 rows of the dataset reveals significant variability in numerical features such as ‘views’, ‘likes’, ‘dislikes’, and ‘comment count’, indicating diverse video popularity and engagement levels using pre-defined method in Python,

```
df_subset = df.iloc[0:20]
print(df_subset.info())
```

Categorical features like ‘category’, ‘comments disabled’, and ‘ratings disabled’ provide insights into video content and interactivity settings, with ‘Entertainment’ being the most frequent category and most videos having comments and rating enabled.

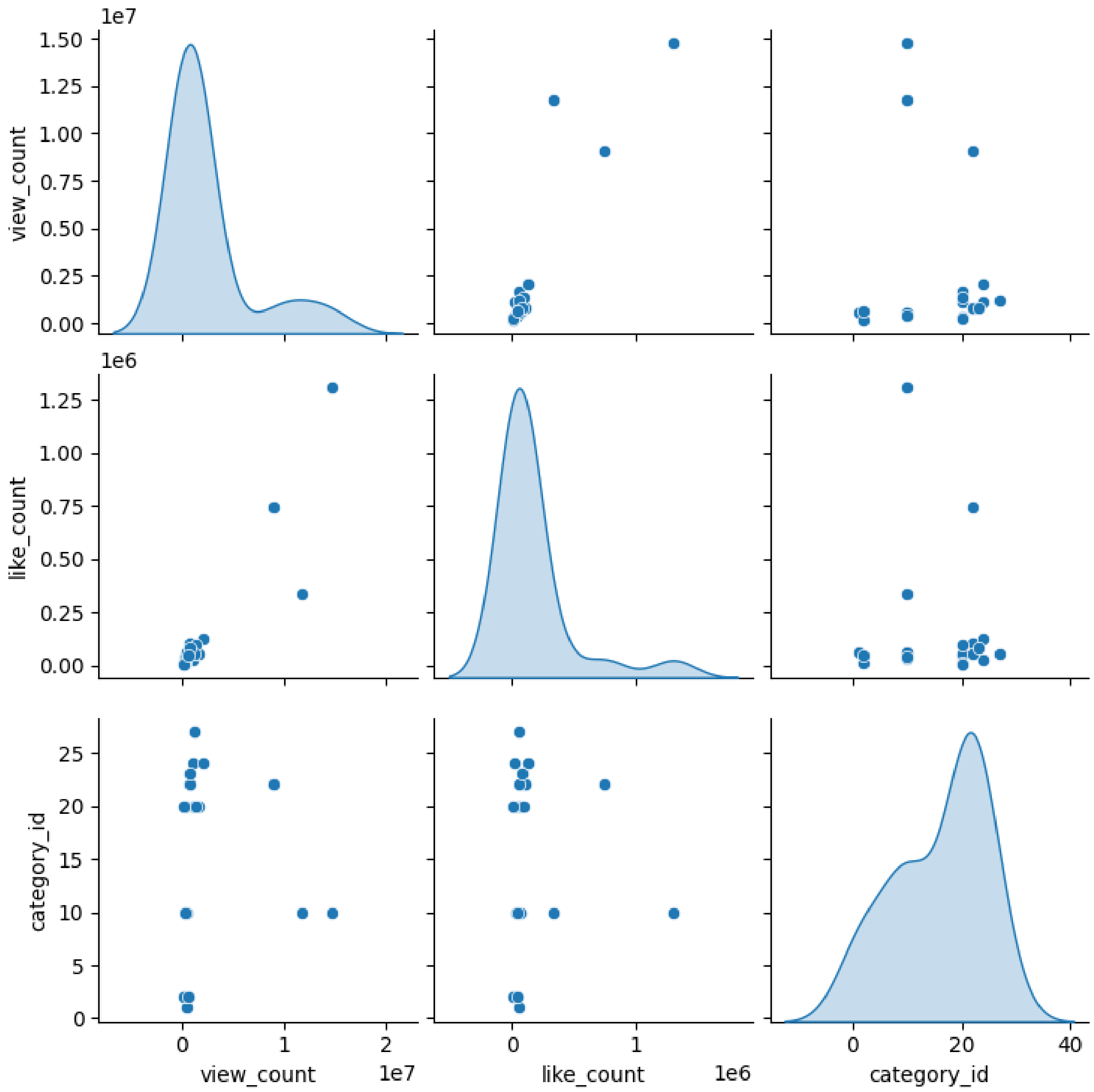

Figure 2.1: Diagnostic Analysis: A pair-plot visualization

The figure 2.1, reveals a pair-plot style visualization of variables ‘view_count’ and ‘like_count’ distributions are highly right-skewed, indicating most videos have low engagement with a few outliers achieving high counts. The density plots along the diagonal revels that both ‘view_count’ and ‘like_count’ are highly right-skewed, indicating that the majority of the videos receive relatively low engagement, while a small number of videos achieve significantly higher counts, manifesting as outliers in the dataset. This skewness is typically in the digital content platforms where only a few videos go viral, garnering millions of views and likes. ‘category_id’ variable shows a more uniform distribution with multiple peaks suggesting a diverse representation of different content categories. This indicates that the videos from various categories being dominated by a few specific ones.

Scatter plots indicate a strong positive correlation between ‘view_count’ and ‘like_count’, suggesting that the higher views typically result in more likes, while there is no clear patterns observed between ‘category_id’ and the other variables. The data ranges up to approximately 15 million views and 1.25 million likes, with outliers present in the high-performing videos.

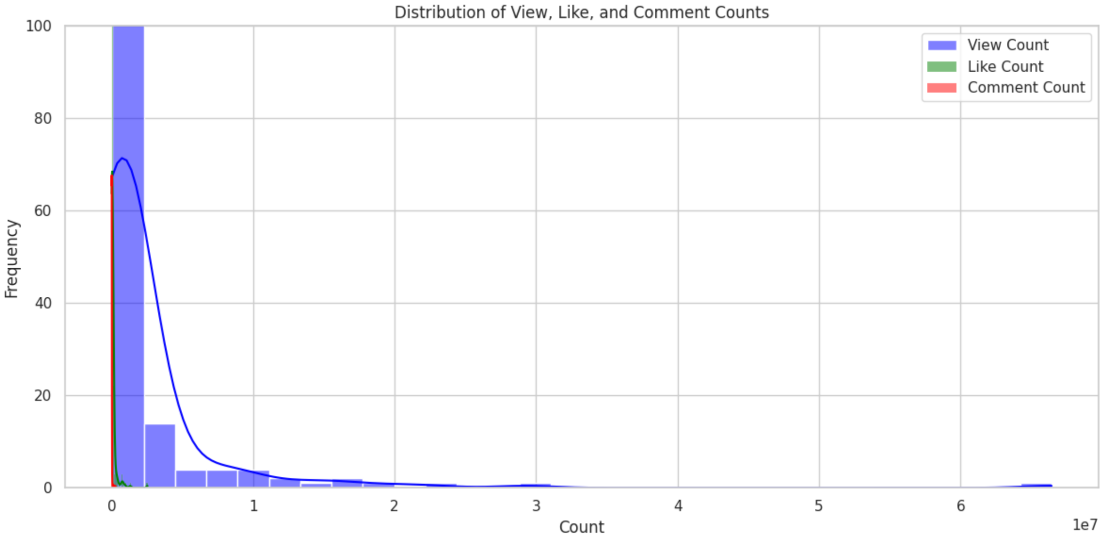

Figure 2.2: Histogram Visualization of Frequency and Density of Key engagement metrics

In figure 2.2 visually examines the distribution of three key engagement metrics: ‘view_count’, ‘like_count’, ‘comment_count’. Each metrics histogram is overlaid with a Kernel Density Estimate (KDE) curve, which provides a smooth estimate of the distribution, offering insights into the frequency and density of these engagement measures. The plot reveals that all three distribution are heavily right-skewed, indicating that the vast majority of videos have relatively low counts in terms of views, likes and comments.

The right skewness implies that most videos does not achieve widespread popularity, with only a few outliers reaching very high engagement levels. This pattern highlights the visibility and user interaction are concentrated among a few videos. Furthermore, the KDE curve suggests that most videos fall within a low range of engagement, there is a long tail where some videos outperform others, due to the factors of content quality, timing, and promotional strategies.

The figure 2.3, illustrates a correlation matrix for the engagement metrics. The matrix visually represents the Pearson Correlation coefficients between ‘view_count’, ‘like_count’, and ‘comment_count’. High positive correlations are observed among all three metrics, with the view count and like count having a correlation coefficient of 0.93, ‘view_count’ and ‘comment_count’ at 0.89, and ‘like_count’ and ‘comment_count’ at 0.89. These strong correlations imply that the videos with high views tend to receive more likes and comments, indicating that

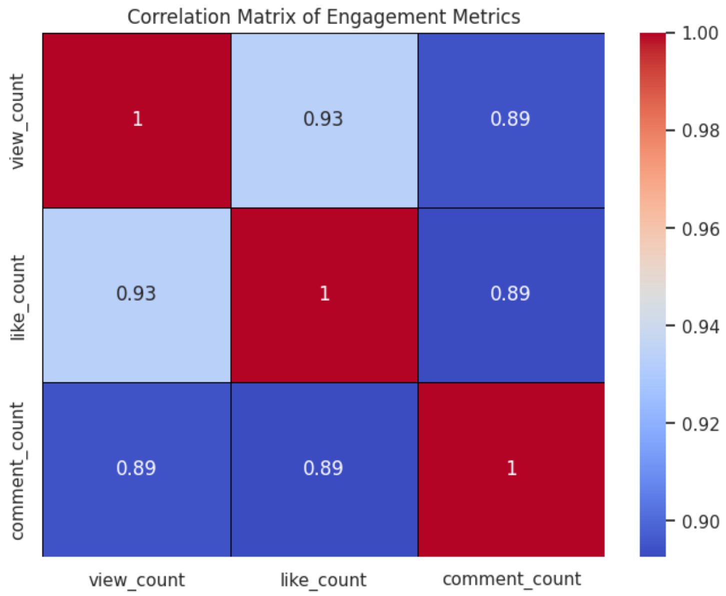

Figure 2.3: Correlation matrix between the Engagement metrics

these metrics are independent and collectively contribute to a video's popularity.

2.3 Data Preparation

The features in the dataset include the video and channel title, genre, publish time, tags, views, comment count, comments disabled, video error or removed, likes, dislikes and description provide a compressive overview of each video's performance and the context in which it became trending. For instance, the 'views' feature indicates the numbers of times a video has been watched, while 'likes' and 'dislikes' reflect viewer engagement.

This section prepares the data for dimensionality reduction and modeling using several pre-processing steps including numerical features to ensure that they are on a comparable scale and encoding categorical variables to convert them into a numerical format suitable for the analysis. This conversion ensures that the algorithms used for dimensionality reduction and subsequent modeling can effectively process the data.

2.3.1 Pre-processing Pipelines

The dataset includes categorical variables such as 'definition' and 'caption', which describes whether a video is in high definition and includes captions. These categorical variables are encoded using a `OneHotEncoder`, which transforms them into a numerical format that machine learning algorithms can process effectively. Additionally, missing values in these categorical variables are imputed with a constant value to ensure that no data is lost during analysis.

Numerical variables, which include metrics such as 'like_count', 'dislike_count', and 'comment_count' are first imputes using the mean values to fill in any missing data, ensuring that no gaps remain in the dataset. After imputation, the numerical data is standardized using the `StandardScaler`. Standardization rescales the data to have a mean of zero and a standard deviation of one. This step is crucial when the numerical features have different unit or scales, as it ensures that each features contributes equally to the analysis, preventing features with the larger numerical ranges from dominating the models learning process.

The preprocessing is applied prior to splitting the data into features 'X_preprocessed' and target variables 'y_preprocessed'. This ensures that the data fed into the machine learning models is clean, consistent, and appropriately scaled. For the target variable, 'view_count', standardization is also performed, allowing for a more balanced analysis, especially in models sensitive to the scale of the output variable.

Chapter 3

Methodology

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique widely used in machine learning for reduction of dimensions. PCA achieves by transforming the large set of variables into a smaller ones by identifying the principal components along which the variation in the data is maximal. These principal components are orthogonal to each other.

PCA allows for a more concise and interpretable representation of the data. This can help in uncovering hidden structures, reducing noise, and improving the efficiency for other algorithms.

Principal Component Analysis is a technique used to emphasise variation and capture strong patterns in the dataset. It carries out by transforming the data into a new coordinate system where the greatest variance by any projection of the data comes to lie on the first coordinate called the first principal component, the second variance on the second coordinate.

A. C. Rencher & W. F. Christensen (2002) suggests that the first principal component represents the linear combination of variables that accounts for the maximum variance in the data, essentially identifying the dimension along which the observations are most widely dispersed. The second principal component, orthogonal to the first, also represents a linear combination with maximal variance in a new direction. Generally, Principal Components (PCs) define distinct dimensions, differing from those established by discriminant functions or canonical variate.

Everitt, B. & Hothorn (2011) said that PCA transforms a set of correlated variables, $x^T = (x_1, \cdots, x_q)$, into a new set of uncorrelated variables, $y^T = (y_1, \cdots, y_q)$, where each new variable is a linear combination of the original variables. These new variables, called principal components, are ordered by their importance, with y_1 capturing the maximum possible variation in the data. Subsequent components, such as y_2, are selected to account for as much of the remaining variation as possible while being uncorrelated with the previous components.

The first principal component of the observations, y_1 is the linear combination

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q$$

whose sample variance is greatest among all such linear combinations. Because the variance of y_1 could be increased without the limit simply by increasing the coefficients $a_1^T = (a_{11}, a_{12}, \cdots, a_{1q})$. To derive the coefficients of the first principal component, maximizing the variance of the linear combination of the variables is constrained by the sum of the squares of the coefficients being one. Using the Lagrange multiplier method, the solution reveals that these coefficients are the eigenvector of the sample covariance matrix corresponding to its largest eigenvalue (Everitt, B. & Hothorn , 2011).

The second principal component, y_2, is defined to the linear combination

$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q$$

i.e., $y_2 = a_2^T x$, where $a_2^T = (a_{21}, a_{22}, \cdots, a_{2q})$ and $x^T = (x_1, x_2, \cdots, x_q)$ that has a greatest variance subject and so on (Everitt, B. & Hothorn , 2011).

Consider, for an example, a set of data consisting of examination scores for several different subjects for each of a number of students. One simple method that determines the informative index of overall examination performance is by computing the mean scores for each student. However, if the score ranges differ across subjects, it might be better to either weight the scores or standardize them before averaging. This could enhance by predicting the accurate ranking in the distribution of students scores. Similarly, applying PCA to the examination scores and using the scores on the first principal component can provide a measure of examination success that effectively distinguishes between the students.

$$X = \begin{pmatrix} 90 & 85 & 80 \\ 70 & 60 & 85 \\ 85 & 75 & 90 \end{pmatrix}$$

where X is the data matrix of examination scores for three students in three subjects.

3.1.1 Standardize the Data

Standardizing the data is crucial, especially when dealing with the examination scores from different subjects, as it ensures that each subject contributes equally to the analysis regardless of its original scale. Without standardization, subjects with higher scoring ranges would disproportionately influence the overall performance index or the principal components.

The first step in PCA is to standardize the dataset. This ensures that each feature contributes equally to the analysis,

$$X' = \frac{X - \mu}{\sigma}$$

where X is the data matrix, μ is the mean of the data, and σ is the standard deviation.

Computing the mean of each column (subject):

$$\text{mean} = \begin{pmatrix} 81.67 & 73.33 & 85 \end{pmatrix}$$

Centering the data by subtracting the column means:

$$X_{\text{centered}} = X - \text{mean} = \begin{pmatrix} 8.33 & 11.67 & -5 \\ -11.67 & -13.33 & 0 \\ 3.33 & 1.67 & 5 \end{pmatrix}$$

Computing the standard deviation of each column (subject):

$$\text{std} = \begin{pmatrix} 9.61 & 12.20 & 5 \end{pmatrix}$$

Standardizing the data by dividing each mean-centered value by the corresponding standard deviation:

$$X_{\text{standardized}} = \begin{pmatrix} \frac{8.33}{9.61} & \frac{11.67}{12.20} & \frac{-5}{5} \\ \frac{-11.67}{9.61} & \frac{-13.33}{12.20} & \frac{0}{5} \\ \frac{3.33}{9.61} & \frac{1.67}{12.20} & \frac{5}{5} \end{pmatrix} = \begin{pmatrix} 0.87 & 0.96 & -1 \\ -1.21 & -1.09 & 0 \\ 0.35 & 0.14 & 1 \end{pmatrix}$$

3.1.2 Calculation of Principal Components

Covariance Matrix

The Covariance matrix is crucial in PCA because it encapsulates the extent to which variables in the dataset change together, reflecting their pairwise relationships. The covariance matrix allows to understand how scores in one subject correlate with scores in another.

$$\text{Cov}(X_{\text{standardized}}) = \frac{1}{n-1} X_{\text{standardized}}^{\top} X_{\text{standardized}}$$

$$\text{Cov}(X_{\text{standardized}}) = \frac{1}{2} \begin{pmatrix} 0.87 & -1.21 & 0.35 \\ 0.96 & -1.09 & 0.14 \\ -1 & 0 & 1 \end{pmatrix}^{\top} \begin{pmatrix} 0.87 & 0.96 & -1 \\ -1.21 & -1.09 & 0 \\ 0.35 & 0.14 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1.03 & 1.05 & -0.26 \\ 1.05 & 1.08 & -0.41 \\ -0.26 & -0.41 & 1 \end{pmatrix}$$

Calculation of Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors helps in determining the principal components, which are the directions in which the data varies the most. The eigenvalues indicate the amount of variance captured by each principal component, while the eigenvectors define the direction of these components.

The eigenvectors tells how to combine the scores in different subjects to form the principal components, and the corresponding eigenvalues indicate the importance of each component.

Let $\mathrm{Cov}(X_{\text{standardized}})$ be A:

$$A = \begin{pmatrix} 1.03 & 1.05 & -0.26 \\ 1.05 & 1.08 & -0.41 \\ -0.26 & -0.41 & 1 \end{pmatrix}$$

Solve for λ and v in $Av = \lambda v$.

$$\text{Eigenvalues:} \quad \lambda_1, \lambda_2, \lambda_3$$

$$\text{Eigenvectors:} \quad v_1, v_2, v_3$$

Using a numerical method or software:

$$\text{Eigenvalues:} \quad \lambda_1 \approx 2.62, \lambda_2 \approx 0.52, \lambda_3 \approx 0.01$$

$$\text{Eigenvectors:} \quad v_1 \approx \begin{pmatrix} 0.58 \\ 0.61 \\ -0.53 \end{pmatrix}, \quad v_2 \approx \begin{pmatrix} -0.35 \\ -0.54 \\ -0.76 \end{pmatrix}, \quad v_3 \approx \begin{pmatrix} 0.73 \\ -0.57 \\ 0.38 \end{pmatrix}$$

Transforming standardized data into Principal Components

Transforming the data into principal components reduces the datasets dimensionality while retaining the most significant information. Transforming the examination scores into principal components simplifies the data, facilitating easier analysis and visualizations. The new set of variables (principal components) are uncorrelated and ranked according to the amount of variance captured, with the first component explaining the most variance.

$$X_{\text{pca}} = X_{\text{standardized}} V$$

where V is the matrix of eigenvectors:

$$V = \begin{pmatrix} 0.58 & -0.35 & 0.73 \\ 0.61 & -0.54 & -0.57 \\ -0.53 & -0.76 & 0.38 \end{pmatrix}$$

$$X_{\text{pca}} = \begin{pmatrix} 0.87 & 0.96 & -1 \\ -1.21 & -1.09 & 0 \\ 0.35 & 0.14 & 1 \end{pmatrix} \begin{pmatrix} 0.58 & -0.35 & 0.73 \\ 0.61 & -0.54 & -0.57 \\ -0.53 & -0.76 & 0.38 \end{pmatrix}$$

So the transformed data into the principal component space (principal components) is:

$$X_{\text{pca}} = \begin{pmatrix} 1.27 & -1.50 & 0.03 \\ -1.31 & 1.36 & -0.16 \\ 0.16 & 0.14 & -0.12 \end{pmatrix}$$

Variance by Each Principal Component

Variance by each principal component is vital to understand the significance of each component. The eigenvalues are directly related to the variance explained by each principal component. The total variance is the sum of the eigenvalues. The proportion of variance explained by each principal component is given by the ratio of its eigenvalue to the total variance.

$$\text{Total Variance} = \lambda_1 + \lambda_2 + \lambda_3 = 2.62 + 0.52 + 0.01 = 3.15$$

Variance Explained by Each Principal Component:

$$\text{Variance of PC1} = \frac{\lambda_1}{\text{Total Variance}} = \frac{2.62}{3.15} \approx 0.83 \quad (83\%)$$

$$\text{Variance of PC2} = \frac{\lambda_2}{\text{Total Variance}} = \frac{0.52}{3.15} \approx 0.17 \quad (17\%)$$

$$\text{Variance of PC3} = \frac{\lambda_3}{\text{Total Variance}} = \frac{0.01}{3.15} \approx 0.003 \quad (0.3\%)$$

3.2 Principal Component Regression

Principal Component Regression is a statistical technique that combines Principal Component Analysis with linear regression. It is primarily used to address multicollinearity issues in regression models. PCA is first applied to transform the original set of correlated predictors into a smaller set of uncorrelated PCs and then used as predictors in a linear regression model.

PCA is a widely used technique for dimension reduction that converts the original variables into a new set of variables called Principal Components (PCs). Typically, researchers (Hotelling (1957); Jackson, J. E. (2005); Jolliffe, I. T (2002); Kendall (1957)) select a subset of these PCs to replace the original variables, significantly reducing the datas dimensionality for further analysis. A common application of PCA is Principal Component Regression , which uses a subset of selected PCs to predict an outcome variables.

From an article by Jolliffe (1982) the concept of using the principal components in regression analysis is well-established. Kendall (1957) and Hotelling (1957) both discussed this approach in their respective works, with Jeffers (1967) providing a notable example. The principal components method involves in substituting the original predictor variables with their principal components, thereby orthogonalizing the regression problem and enhancing computational stability and simplicity.

Considering the same example from section 3.1, a set of data values represented in the data matrix X of examination scores for three students (rows) in three subjects (columns).

$$X = \begin{pmatrix} 90 & 85 & 80 \\ 70 & 60 & 85 \\ 85 & 75 & 90 \end{pmatrix}$$

Let the dependent variable y be the total scores of the students:

$$y = \begin{pmatrix} 255 \\ 215 \\ 250 \end{pmatrix}$$

3.2.1 Calculation of Principal Components

1. Computing the mean of each column (subject):

$$\text{mean} = \begin{pmatrix} 81.67 & 73.33 & 85 \end{pmatrix}$$

Center the data by subtracting the column means:

$$X_{\text{centered}} = \begin{pmatrix} 8.33 & 11.67 & -5 \\ -11.67 & -13.33 & 0 \\ 3.33 & 1.67 & 5 \end{pmatrix}$$

2. Calculating the covariance matrix of the centered data:

$$S = \frac{1}{n-1} X_{\text{centered}}^T X_{\text{centered}}$$

$$= \frac{1}{2} \begin{pmatrix} 8.33 & -11.67 & 3.33 \\ 11.67 & -13.33 & 1.67 \\ -5 & 0 & 5 \end{pmatrix}^T \begin{pmatrix} 8.33 & 11.67 & -5 \\ -11.67 & -13.33 & 0 \\ 3.33 & 1.67 & 5 \end{pmatrix}$$

$$S = \frac{1}{2} \begin{pmatrix} 223.39 & 261.11 & -74.95 \\ 261.11 & 323.11 & -85.00 \\ -74.95 & -85.00 & 50.00 \end{pmatrix}$$

3. Finding the eigenvalues and eigenvectors of the covariance matrix S:
Solve the characteristic equation $\det(S - \lambda I) = 0$:

$$\begin{vmatrix} 223.39 - \lambda & 261.11 & -74.95 \\ 261.11 & 323.11 - \lambda & -85.00 \\ -74.95 & -85.00 & 50.00 - \lambda \end{vmatrix} = 0$$

Solving for λ, we get the eigenvalues $\lambda_1, \lambda_2, \lambda_3$.
Assume the eigenvalues are $\lambda_1 = 512, \lambda_2 = 80, \lambda_3 = 4$.
Next, finding the corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$.
Assume eigenvectors are:

$$\mathbf{v}_1 = \begin{pmatrix} 0.577 \\ 0.577 \\ -0.577 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 0.707 \\ -0.707 \\ 0 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 0.408 \\ 0.408 \\ 0.816 \end{pmatrix}$$

4. Transforming the centered data into principal components:

$$Z = X_{\text{centered}} \cdot V$$

where V is the matrix of eigenvectors.

$$V = \begin{pmatrix} 0.577 & 0.707 & 0.408 \\ 0.577 & -0.707 & 0.408 \\ -0.577 & 0 & 0.816 \end{pmatrix}$$

$$Z = \begin{pmatrix} 8.33 & 11.67 & -5 \\ -11.67 & -13.33 & 0 \\ 3.33 & 1.67 & 5 \end{pmatrix} \cdot \begin{pmatrix} 0.577 & 0.707 & 0.408 \\ 0.577 & -0.707 & 0.408 \\ -0.577 & 0 & 0.816 \end{pmatrix}$$

Performing the multiplication to get the PCs:

$$Z = \begin{pmatrix} 11.18 & 2.38 & -5.91 \\ -18.31 & -3.33 & 0.91 \\ 7.13 & 0.95 & 4.99 \end{pmatrix}$$

5. Finding the Variance for each PC:
The variance for each PCs is given by the ratio of its eigenvalue to the sum of all eigenvalues.
For example, for the first principal component:

$$\text{Variance explained by } PC_1 = \frac{\lambda_1}{\sum \lambda} = \frac{512}{512 + 80 + 4} \approx 0.857$$

This means that the first PC explains approximately 85.7% of the total variance in the data.

3.2.2 Performing the Linear Regression on PCs

Performing linear regression on Principal Components involves using the principal components (PCs) derived from PCA as the predictor variables in a regression model instead of the original variables. Least Squares Estimation (LSE) is a method used in regression analysis to find the line of best fit by minimizing the sum of the squares of the difference between observed and predicted values.

Using the first principal component for regression. Suppose the first principal component is Z_1:

$$Z_1 = \begin{pmatrix} 11.18 \\ -18.31 \\ 7.13 \end{pmatrix}$$

Performing the linear regression with Z_1 as the predictor and y as the response:

Fitting the model $y = \beta_0 + \beta_1 Z_1$.

Using LSE:

$$\hat{\beta}_1 = \frac{\sum(Z_1 - \bar{Z_1})(y - \bar{y})}{\sum(Z_1 - \bar{Z_1})^2}$$

$$\hat{\beta}_1 = \frac{(11.18 - 0)(255 - 240) + (-18.31 - 0)(215 - 240) + (7.13 - 0)(250 - 240)}{11.18^2 + (-18.31)^2 + 7.13^2}$$

$$\hat{\beta}_1 = 1.37$$

Calculating the intercept β_0:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{Z_1} = 240 - 1.37 \cdot 0 = 240$$

Thus, the regression model is:

$$\hat{y} = 240 + 1.37 Z_1$$

To find the performance of the regression model, Mean Squared Error (MSE) which measures the average of the squared difference between the estimated values $\hat{y}_i$ and the actual value y_i.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

and the R-squared (R^2) values measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values y_i.

$$\hat{y} = 240 + 1.37 Z_1$$

Given data:

$$Z_1 = \begin{pmatrix} 11.18 \\ -18.31 \\ 7.13 \end{pmatrix}$$

Actual y values:

$$y = \begin{pmatrix} 255 \\ 215 \\ 250 \end{pmatrix}$$

1. Calculating the Predicted Values ($\hat{y}$):

$$\hat{y} = 240 + 1.37 Z_1 = 240 + 1.37 \begin{pmatrix} 11.18 \\ -18.31 \\ 7.13 \end{pmatrix} = \begin{pmatrix} 255.066 \\ 215.021 \\ 249.761 \end{pmatrix}$$

2. Calculating MSE value:

$$\text{MSE} = \frac{1}{3} \sum_{i=1}^{3} (y_i - \hat{y}_i)^2$$

$$= \frac{1}{3} \left[(255 - 255.066)^2 + (215 - 215.021)^2 + (250 - 249.761)^2 \right]$$

$$\text{MSE} = 0.020639 \quad \text{(approx)}$$

3. Calculating the $\bar{y}$ (mean of actual y values):

$$\bar{y} = \frac{255 + 215 + 250}{3} = 240$$

4. Calculating R^2 value:

$$R^2 = 1 - \frac{\sum_{i=1}^{3} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{3} (y_i - \bar{y})^2}$$

$$= 1 - \frac{0.061918}{(255 - 240)^2 + (215 - 240)^2 + (250 - 240)^2}$$

$$R^2 = 0.999935 \quad \text{(approx)}$$

Final Estimation:

- Mean Squared Error (MSE): 0.020639 (approx)
- R-squared (R²): 0.999935 (approx)

3.3 Kernel-Principal Component Analysis

Kernel Principal Component Analysis is an extension of the Principal Component Analysis that captures the non-linear relationships within the data by mapping it into a higher-dimensional features space. PCA is limited to linear transformations, making it less effective for the data where the strictures is inherently nonlinear.

K-PCA addresses this limitation by applying a kernel function to the data, which implicitly transforms the input space into a higher-dimensional space where the linear PCA can be effectively applied.

Considering a data matrix X with two data points and two dimensions:

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

3.3.1 Computing the Kernel Matrix K

In Kernel-PCA, the kernel matrix K is a fundamental component which encapsulates the relationships between the data points in the transformed feature space.

Common choices of kernel function includes linear kernel, polynomial kernel, and radial basis function (RBF) kernel are considered: The linear kernel is the simplest, computing the inner product between two vectors, which is effective when the data is linearly separable in the input space. The polynomial kernel extends the linear kernel by allowing the interactions between the features, making it q suitable for the problems between the features that are not merely addictive but involve higher-order interactions. The Radial Basis Function kernel, also known as the Gaussian kernel, which is used to map the data into an infinite-dimensional space, capturing the complex, nonlinear patterns.

Considering a polynomial kernel to measure the similarity between pairs of the data points x_i and x_j for this example :

$$K(x_i, x_j) = (x_i \cdot x_j^\top + 1)^2$$

Here, $x_i \cdot x_j^\top$ represents the dot product between the vectors x_i and x_j.

First, we compute the dot products:

$$\begin{pmatrix} x_1 \cdot x_1^\top & x_1 \cdot x_2^\top \\ x_2 \cdot x_1^\top & x_2 \cdot x_2^\top \end{pmatrix} = \begin{pmatrix} 1 \cdot 1 + 2 \cdot 2 & 1 \cdot 3 + 2 \cdot 4 \\ 3 \cdot 1 + 4 \cdot 2 & 3 \cdot 3 + 4 \cdot 4 \end{pmatrix} = \begin{pmatrix} 5 & 11 \\ 11 & 25 \end{pmatrix}$$

Now, apply the polynomial kernel formula:

$$K = \begin{pmatrix} (5+1)^2 & (11+1)^2 \\ (11+1)^2 & (25+1)^2 \end{pmatrix} = \begin{pmatrix} 36 & 144 \\ 144 & 676 \end{pmatrix}$$

3.3.2 Center the Kernel Matrix K_c

Centering the Kernel Matrix is a preprocessing step in KPCA and other kernel methods which ensures that the data in the transformed feature space has zero mean. The mean of the each variable is subtracted to the focus on the variability and the structure of the data rather than the mean value.

To center the kernel matrix, we use the formula:

$$K_c = K - \mathbf{1}K - K\mathbf{1} + \mathbf{1}K\mathbf{1}$$

Kernel matrix K, which represents the pairwise similarities between the data points, adjusts the matrix and centers so that the transformed features are centered around the origin in the high-dimensional space.

Where $\mathbf{1}$ is a matrix of ones divided by the number of data points (here, 2):

$$\mathbf{1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

Let's calculate each part:

$$\mathbf{1}K = \begin{pmatrix} 90 & 410 \\ 90 & 410 \end{pmatrix}$$

Similarly:

$$K\mathbf{1} = \begin{pmatrix} 90 & 90 \\ 410 & 410 \end{pmatrix}$$

$$\mathbf{1}K\mathbf{1} = \begin{pmatrix} 250 & 250 \\ 250 & 250 \end{pmatrix}$$

Finally, the Centered kernel matrix is:

$$K_c = K - \mathbf{1}K - K\mathbf{1} + \mathbf{1}K\mathbf{1}$$

$$K_c = \begin{pmatrix} 106 & -106 \\ -106 & 106 \end{pmatrix}$$

3.3.3 Calculation of Principal Components

Computing the Eigenvalues and Eigenvectors

For the considered square matrix X, and eigenvalue λ and an eigenvector V satisfy the equation $Av = \lambda V$. The covariance matrix of the data is decomposed to identify the eigenvectors along which the data variance is maximized, with the corresponding eigenvalues indicating the amount of variance captured by each principal component.

The eigenvalues (λ) and eigenvectors (v) of the centered kernel matrix K_c:

$$\text{Eigenvalues}: \lambda_1 = 212, \lambda_2 = 0$$

$$\text{Eigenvectors}: v_1 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}, v_2 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

Transforming Data into Principal Components

In KPCA, after computing the eigenvalues and eigenvectors from the centered kernel matrix, the original data is transformed into PCs by projecting the centered kernel matrix onto eigenvectors.

K' is the centered kernel matrix, and V are the eigenvectors corresponding to the non-zero eigenvalues, then the transformed data is obtained by multiplying K' by v.

To transform the data into principal components, we project the centered kernel matrix onto the eigenvectors:

$$Z = K_c \cdot v$$

$$Z_1 = \begin{pmatrix} 106 & -106 \\ -106 & 106 \end{pmatrix} \cdot \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} \sqrt{2} \cdot 106 \\ -\sqrt{2} \cdot 106 \end{pmatrix}$$

$$Z_2 = \begin{pmatrix} 106 & -106 \\ -106 & 106 \end{pmatrix} \cdot \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Chapter 4

Dimensionality-Reduced Modelling

This chapter involves in applying various dimensionality reduction methods like Principal Component Analysis, Kernel-PCA and Principal Component Regression to the preprocessed dataset.

4.1 Principal Component Analysis

Importing the PCA module from scikit-learn allows the application of Principal Component Analysis, a technique used to reduce the dimensionality of the dataset as seen in the section 3.1, while preserving its variance ratio. PCA is instantiated 'n_components=2', by indicating the reduction of the dataset to two principal components. PCA transforms the data into a lower-dimensional space by fitting 'X_preprocessed', where each sample is represented by its projection onto the principal components. Additionally, 'explained_variance_pca' variance ratio provides the variance explained by each principal components,crucial for assessing the information which is retained after the dimension reductions.

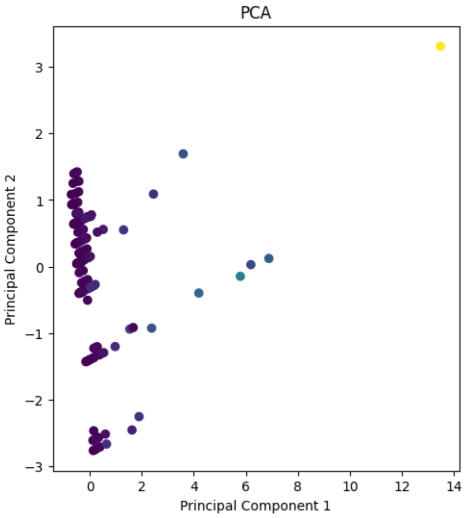

Figure 4.1: Visualization of data points using the PCA method

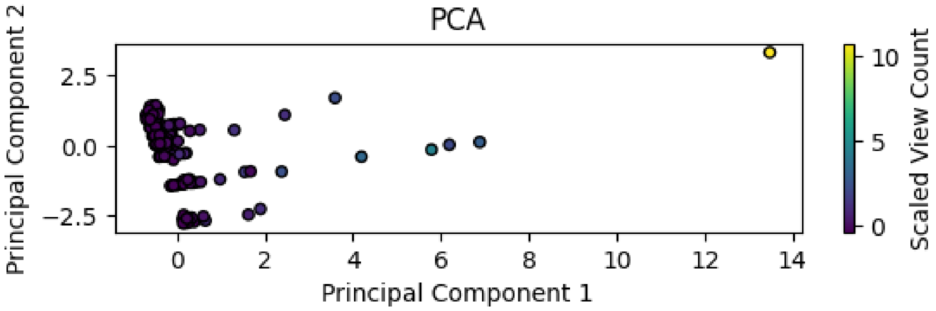

Figure 4.2: Visualization of data points in Scaled View Count

In figure 4.1 and 4.2, the visualizations of the PCA in a two-dimensional scatter plot reveals the greater variance explained by PC1 compared to PC2 indicates that the first principal component captures a significant amount of variability in the data. This suggests that the PCA effectively reduces the dimension while preserving essential information. The visualization hints at the potential clusters or subgroups within the data, which could be explored further using clustering algorithms. Additionally, PCA highlights outliers, such as yellow data points, prompting further investigation in the original feature space.

4.1.1 Outliers Reduction

The z-score method is a statistical technique used to identify the outliers in a dataset by measuring the standard deviation in a data points away from the mean. The z-score method assesses each principal component's values to detect extreme observations that deviate significantly from the rest of the dataset.

To identify outliers using the z-score method, computing the z-scores for each data plot in the dataset. This involves in subtracting the mean of the data from each point, dividing by the standard deviation, and taking the absolute value. A threshold, typically set to 3, is used to determine which z-scores are considered outliers, as they are more than three standard deviations away from the mean, suggesting that they significantly different from the rest of the data.

The outliers are returned as the indices of data points that meet this criterion. By applying this method in the PCA-transformed space, the analysis can reveal extreme deviations that might correspond to significant anomalies in the original dataset. Visualizing these outliers involves plotting the PCA-transformed data and highlighting the outlier points, providing a clear and intuitive representation of unusual data points relative to the principal components.

In figure 4.3, the scatter plot illustrates data points projected onto the first two principal components, which captures the highest variance directions within the dataset. The majority of data points are concentrated near the origin, indicating similar values in these principal components. Notably, several points, highlighted in red and labeled as ”Outliers”, are significantly distant from the main cluster, indicating substantial deviations. The density of the data points are denoted as a third variable in a color scale from dark purple to yellow.

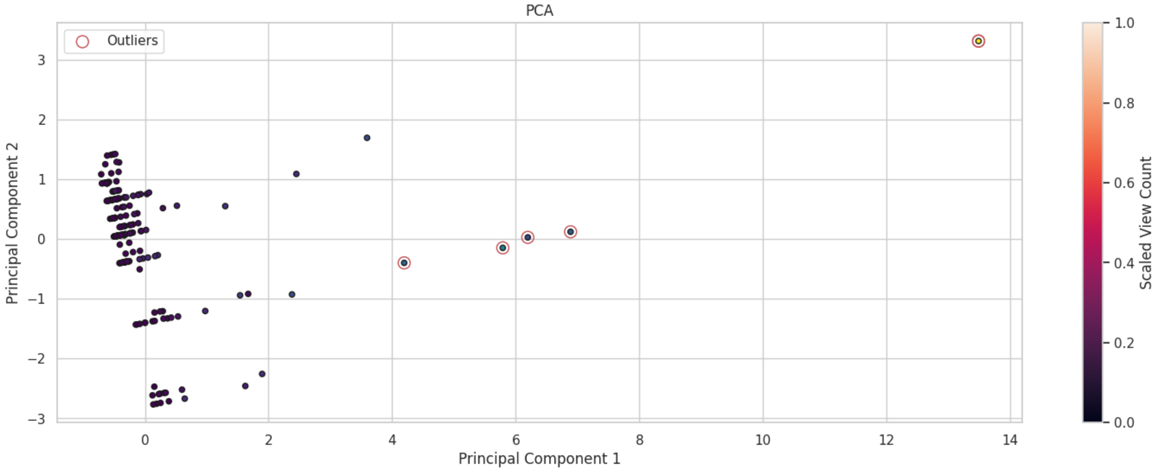

Figure 4.3: Visualization of data points with Outliers in PCA method

These identified outliers are then removed from both the feature set ('X_preprocessed')and the target variable ('y_preprocessed'), resulting in a cleaned dataset ('X_cleaned' and 'y_cleaned'). PCA is re-applied to this cleaned dataset to ensure that the principal components accurately represent the underlying data structure without the distortion caused by outliers.

4.1.2 Applying the PCA method

After the initial round of outlier removal and PCA application, further scrutiny reveals additional outliers at specific indices (18 and 103). These outliers are manually identified and removed from the already cleaned dataset, resulting in a further refined dataset ('X_further_cleaned' and 'y_further_cleaned'). PCA is applied once more to this further cleaned data to generate a new set of principal components approximately 67.66% and 23.46% of the variance. The explained variance ratio of these principal components is calculated to assess how much variance each principal component captures, ensuring that the significant dimensions of the dataset are preserved.

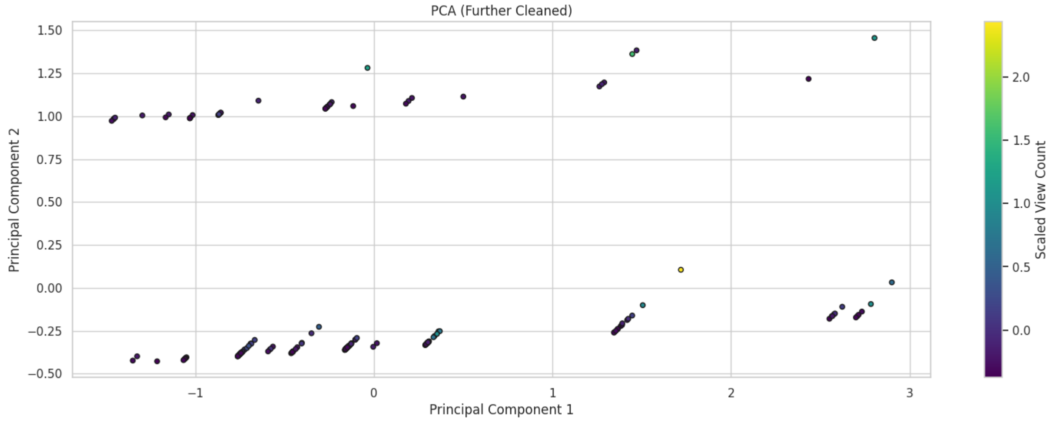

Figure 4.4: Data points that are further cleaned are visualized using PCA method

In figure 4.4, the scatter plots of the PCA-transformed data illustrate the distribution of the data points along with the principal components.

4.1.3 Data Visualization

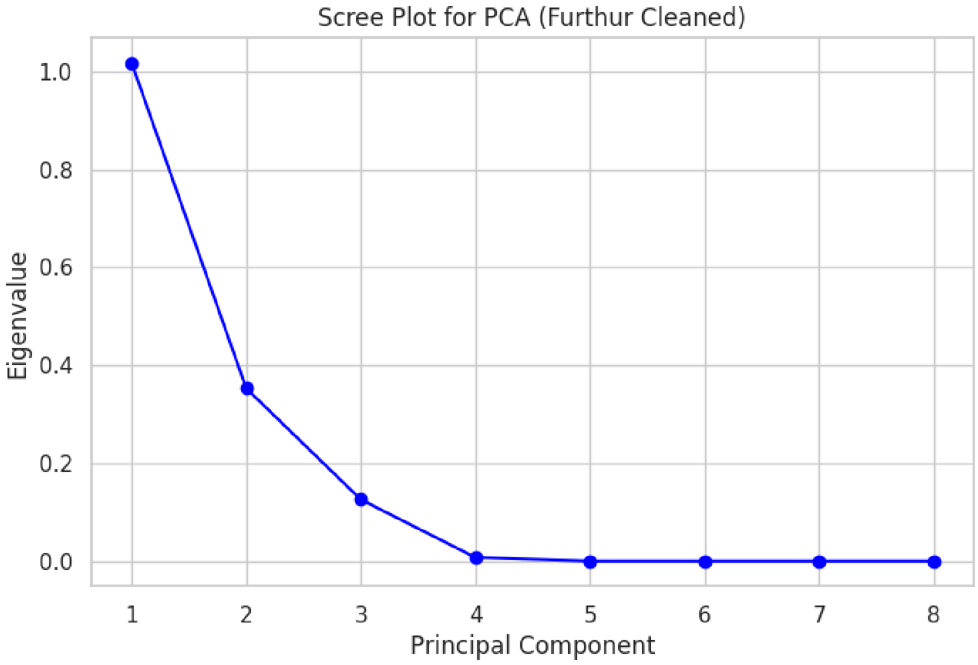

Figure 4.5: Scree plot visualization for Further cleaned data points using PCA

A Scree plot is a diagnostic tool used in PCA to evaluate the relative importance of each principal component in explaining the variance. The figure 4.5, visualizes the eigenvalues of the covariance matrix of the data, which quantify the amount of variance captured by each principal component.

Eigenvalues are plotted in descending order against the corresponding principal components. The plot identifies the ”elbow” point where the curve begins to flatten, indicating the number of principal components and the components to the left of the inflection point are generally considered significant, while those to the right contribute minimally to the variance, suggesting that they can be discarded.

4.2 Principal Component Regression

Principal Component Regression is one of the dimensionality reduction technique that combines PCA with a linear regression. PCA is initially transformed the original high dimensional dataset into a smaller set of uncorrelated principal components, capturing the majority of the variance in the data. Secondly, these principal components are used as a predictors in a linear regression model, facilitating the prediction of target engagement metrics.

A pipeline is created using scikit-learn library, which is then applied to the preprocessed datasets, transforming the original features into principal components and subsequently fitting the linear regression model.

4.2.1 Handling Outliers

The figure 4.6, visualizes a scatter plot of the transformed features. The transformed dataset, ‘X_pcr’, contains two principal components derived from PCA. Each point in the scatter plot represents a data sample, with the x-axis corresponding to the first principal component and y-axis to the second principal component.

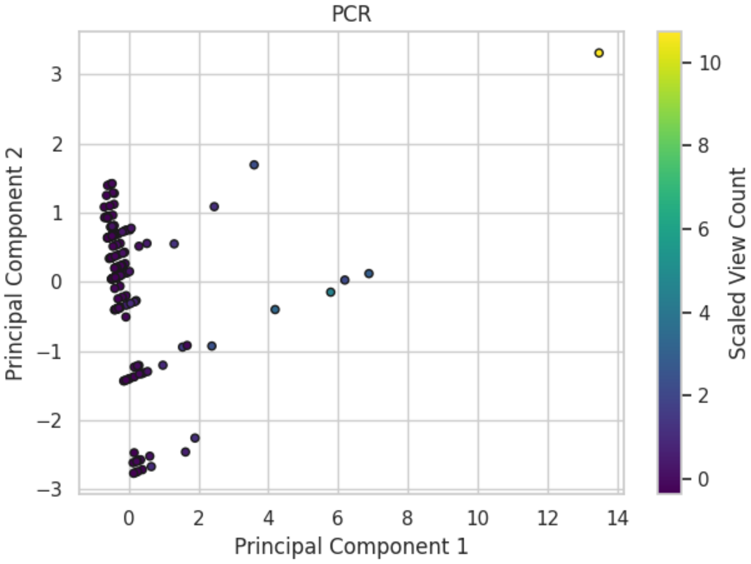

Figure 4.6: Visualization of data points using PCR method

From the figure 4.7 the ‘identify_outliers_zscore’ function calculates the z-scores for each data point in ‘x_pcr’, which represents the number of standard deviations a data point is from the mean of the dataset.

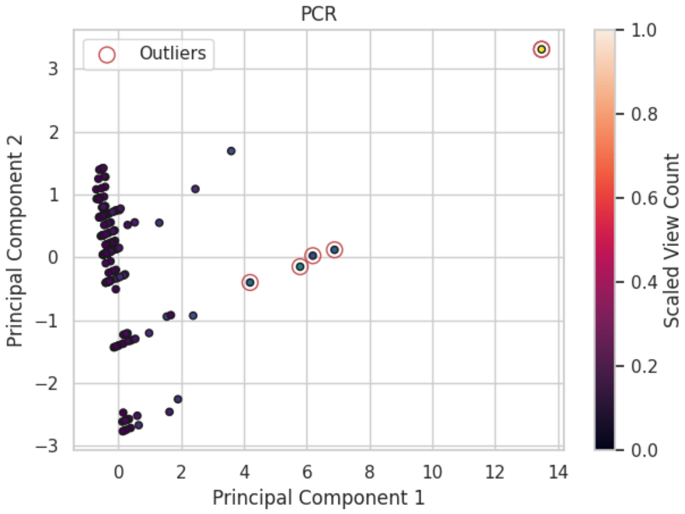

Figure 4.7: Visualizing the data points with outliers using PCR method

4.2.2 Data Visualization

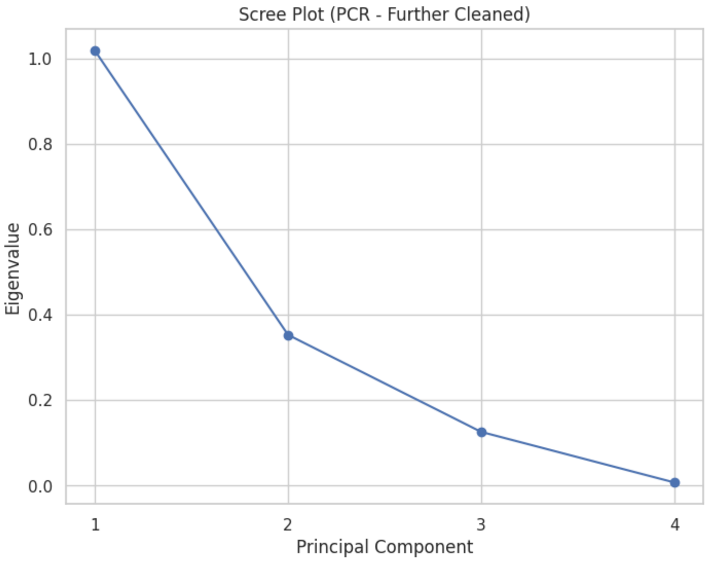

Figure 4.8: Scree Plot visualization of cleaned data points using PCR method

In PCR, a scree plot serves as a valuable tool for evaluating the significance of principal components in the context of regression modelling. The figure 4.8, illustrates the eigenvalues associated with each principal component derived from the covariance matrix of the predictors, with the eigenvalues reelecting the variance each component explains.

The plot highlights the components relative importance and facilities the selection of the optimal number of components to retain for PCR. The ”elbow” of the plot indicated the point where the additional components contribute the negligible to the explained variance. Non-positive eigenvalues are excluded as they do not contribute a meaningfully to the variance.

4.2.3 Linear Regression Model

Initially, the cleaned dataset is spilt into training and testing sets with 80% for training and 20% for testing. The ‘train_test_split’ function ensures a random split, preserving the representatives of the sample.

The PCR model is trained on the training data (‘X_train’, ‘y_train’), and predictions are made on the test data (‘X_test’). The models performance is evaluated using two metrics: Mean Squared Error (MSE) and R-squared (R^2).

MSE is evaluated and calculated by the average of the squares of the errors, indicating the average squared difference between the observed actual outcomes and the predicted values. A lower MSE signifies a better model performance.

R-squared indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from the 0 to 1, with a higher values indicating a better fit of the model.

The MSE value is **0.25**, which indicates that on a average the squared difference between the predicted and actual values are relatively small. The R-squared value is **0.06**, which suggests that only **6%** of the variance in the dependent variable is explained by the model. This relatively low R-squared value implies the model does not fit the data well and there might be other factors influencing the dependent variables.

4.3 Kernel-Principal Component Analysis

Kernel PCA is an extension of the PCA that uses technique from the Radial Basis Function kernel method to transform the preprocessed dataset into a new feature space defined by two principal components. The RBF kernel (’kernel=’rbf’) is a function for Kernel PCA for it’s ability to capture complex non-linear relationships in the data. ’x_kpca’ contains the transformed data points in the new principal component space derived from Kernel PCA.

In figure 4.9, the scatter plot visualizes the transformed data points (’X_kpca’) in the two-dimensional principal component space.

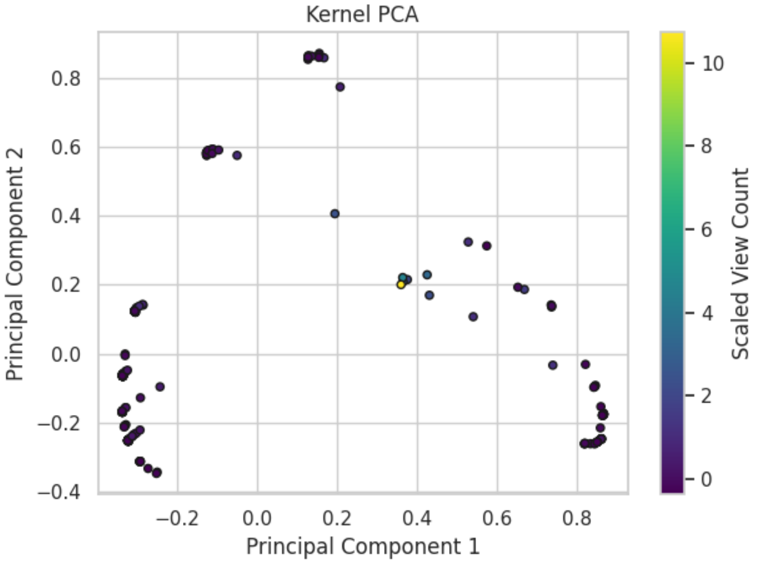

Figure 4.9: Visualization of data points using KPCA method

4.3.1 Outliers identification and Visualization

Outliers 'outliers_kpca' are identified through z-score analysis which calculate the z-scores of the transformed data points. Data points with z-scores exceeding the threshold of 3 are flagged as outliers, as this method ensures that the significant deviations from the mean, highlighting the data points from the rest of the dataset.

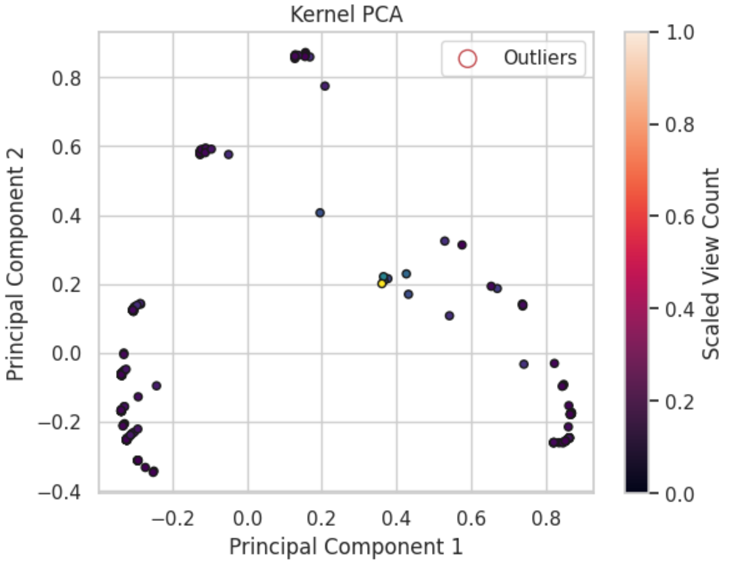

Figure 4.10: Visualization of data points with none outliers using KPCA method

In figure 4.10, the plot denotes distinct clusters of the data points predominately located at the extremes of the x-axis, with additional scattered distribution across the origin. The scatter plot denotes points as ’Outliers’, marked by red circles, though none are visible in this particular view.

4.3.2 Visualization of Cumulative Explained Variance

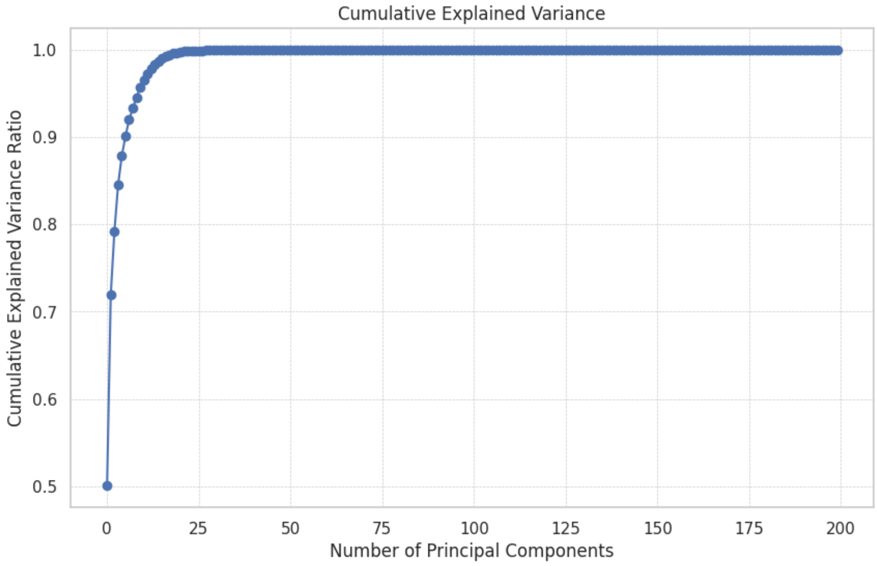

Figure 4.11: Visualization of Cumulative Explained Variance for KPCA method

The cumulative explained variance plot is a crucial tool in KPCA for understanding the proportion of total variance captured by successive principal components. This is particularly useful in KPCA, where non-linear mappings can introduce components with diminishing significance.

The figure 4.11, illustrates the cumulative sum of the sorted eigenvalues from KPCA, which reflects the variance explained by each component. The plot features an descending curve that starts steep and gradually flattens, reflecting the diminishing returns of the additional components. This visualization helps in determining the number of principal components to retain, balancing between capturing a substantial amount of variance and avoiding over-fitting by including too many components.

Chapter 5

Discussion

The focus was on applying modern dimensionality reduction techniques- PCA, Kernel PCA, and PCR to the dataset to undercover the data patterns and improve the model efficiency. Initially the datasets preprocessed to address the issues such as multicollinearity. PCA was employed to transform the data into a set of uncorrelated principal components, which facilitated a clear interpretation of the underlying structure. KPCA was subsequently applied to capture non-linear relationships within the data, and PCR was utilized to enhance the predictive accuracy by incorporating the principal components into the regression model.

During the process, outliers were identified and reduced to prevent them from skewing the results, particularly in the linear methods. The methods successfully reduced the dimensions of the dataset while retaining the critical information, thereby improving the performance and the interpreability of the model. Currently, these techniques continue to be pivotal in handling the high-dimensional data, the outcomes underscores the dimensionality reduction in simplifying data complexity and the quality of insights from the multivariate data analysis.

Chapter 6

Conclusion

In conclusion, the application of the dimensioanlity reduction techniques such as PCA, KPCA, and PCR on the YouTube dataset has been systemically carried out to enhance the understanding the underlying patterns within the data. The reduction of outliers was carefully undertaken to prevent skewed results and to ensure the accuracy of the analysis. Through PCA, the primary components contributing to data variance were identified, while KPCA provided a non-linear perspective, capturing more complex relationships. PCR was utilized to address the multicollinearity and improve the robustness of the predictive model. The results have demonstrated that these techniques, when appropriately applied, can significantly streamline the data, highlighting the most critical features without compromising the integrity of the information. This approach has not only facilitated a deeper insight into the dataset but also improved the effectiveness of subsequent analyses and predictors.

Chapter 7

References

Pearson, K. (1901). The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. *On lines and planes of closest fit to systems of points in space.* **2**(11), 559–572. Taylor & Francis.

Hotelling, H. (1933). *Analysis of a Complex of Statistical Variables into Principal Components.* Journal of Educational Psychology, 24(6), 417–441.

Hotelling, H. (1957). *The relations of the newer multivariate statistical methods to factor analysis*. *British Journal of Statistical Psychology, 10*(2), 69–79. Wiley Online Library.

Jolliffe, I. T. (2002). Springer. *Principal component analysis for special types of data.*

James, G. (2013). *An introduction to statistical learning*. Springer.

Jackson, J. E. (2005). John Wiley & Sons. *A user's guide to principal components.*

Jolliffe, I. T. & Cadima, J. (2016). The Royal Society Publishing. *Principal component analysis: a review and recent developments.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, **374**(2065), 20150202.

Jolliffe, I. T. (1982). *A note on the use of principal components in regression. Journal of the Royal Statistical Society. Series C (Applied Statistics), 31*(3), 300–303. Royal Statistical Society, Oxford University Press. Available at `http://www.jstor.org/stable/2348005`.

Kendall, M. G. (1957). *A course in multivariate analysis* (2nd ed.). Hafner Publishing Company.

Jeffers, J. N. R. (1967). *Two case studies in the application of principal component analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics), 16*(3), 225–236. Wiley Online Library.

Schölkopf, B., Smola, A., & Müller, K.-R. (1998). MIT Press. *Nonlinear component analysis as a kernel eigenvalue problem.* Neural Computation, **10**(5), 1299–1319.

Everitt, B. & Hothorn (2011). New York; Springer Science+Business Media. *An introduction to applied multivariate analysis with R.* **89**, 2011.

Izenman, A.Julian. (2001). New York, NY: Springer *Modern multivariate statistical techniques: regression, classification, and manifold learning* **17**, 2008.

J. Lu (2023) Trending Videos on YouTube *Trending Videos on YouTube, Journal of Education, Humanities and Social Sciences*, vol. 7, pp. 84-91, Jan. 2023. doi: 10.54097/ehss.v7i.4016.

G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker (2020), Analysis of Dimensionality Reduction Techniques on Big Data *Analysis of Dimensionality Reduction Techniques on Big Data, IEEE Access* vol. 8, pp. 54776-54788, 2020. doi: 10.1109/ACCESS.2020.2980942.

Y. Gu, Y. Liu, & Y. Zhang (2008) *A Selective KPCA Algorithm Based on High-Order Statistics for Anomaly Detection in Hyperspectral Imagery, IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 1, pp. 43-47, 2008. doi: 10.1109/LGRS.2007.907304.

Mika, S., Schölkopf, B., Smola, A., Müller, K.-R., Scholz, M., and Rätsch, G. (1998). *Kernel PCA and de-noising in feature spaces.* Advances in Neural Information Processing Systems, **11**.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT Press.

Frank, L. E., & Friedman, J. H. (1993). *A statistical view of some chemometrics regression tools*. *Technometrics, 35*(2), 109–135. Taylor & Francis.

Jolliffe, I. T. (1982). *A note on the use of principal components in regression. Journal of the Royal Statistical Society Series C: Applied Statistics, 31*(3), 300–303. Oxford University Press.

L. J. Cao, K. S. Chua, W. K. Chong, H. P. Lee, & Q. M. Gu (2003) *A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine, Neurocomputing* vol. 55, no. 1, pp. 321-336, 2003. doi: https://doi.org/10.1016/S0925-2312(03)00433-8. Available: https://www.sciencedirect.com/science/article/pii/S0925231203004338.

A. C. Rencher & W. F. Christensen *Methods of Multivariate Analysis.* John Wiley & Sons, 2002, vol. 727, pp. 2218–0230.