Dimensionality Reduction Methods in Modern Multivariate Analysis

Saadhana Ganesa Narasimhan, 201703255

University of Leeds

Introduction

Background: The complexity and volume of data have increased significantly, prompting the development of new multivariate techniques for understanding patterns and relationships. High-dimensional data often require dimensionality reduction methods which map data into lower-dimensional spaces to interpret primary contributors to variability.

Aim: The primary objective of this research is to apply linear and non-linear dimensionality reduction methods such as Principal Component Analysis (PCA), Kernel PCA (K-PCA), and Principal Component Regression (PCR) to reduce the dimensionality and capturing the variance of the dataset.

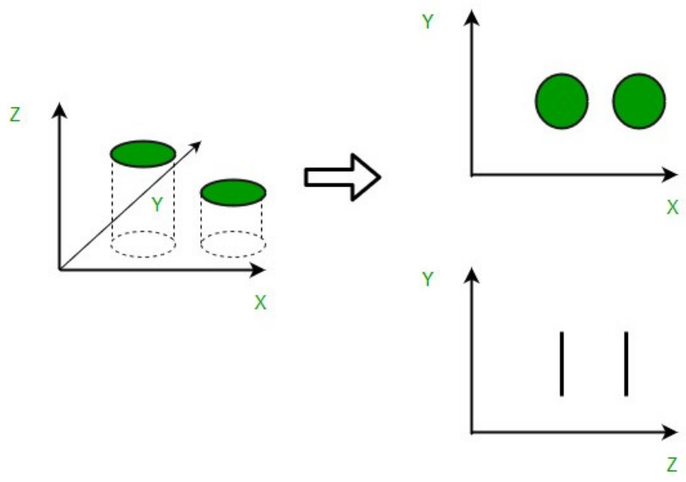

Figure 1. Dimensionality Reduction

Materials & Methods

Dataset: The dataset consists of 200 samples with 16 variables related to trending YouTube videos, including features such as views, likes, dislikes, and comment count. Missing values were imputed, and outliers were removed using the z-score method.

Dimensionality Reduction:

- PCA is applied to reduce dataset dimensionality, capturing linear relationships, particularly between 'view_count', 'like_count', 'comment_count'.
- K-PCA with a Radial Basis Function (RBF) kernel is utilized to capture non-linear relationships, particularly effective for modeling the complex, non-linear dependency.

Software and Tools: The analysis was conducted using Python and its libraries, including: `Pandas, Numpy, Scikit-Learn, Seaborn, and Matplotlib.`

Exploratory Data Analysis (EDA)

- Initial exploratory data analysis focused on visualizing distributions and relationships of key engagement metrics.

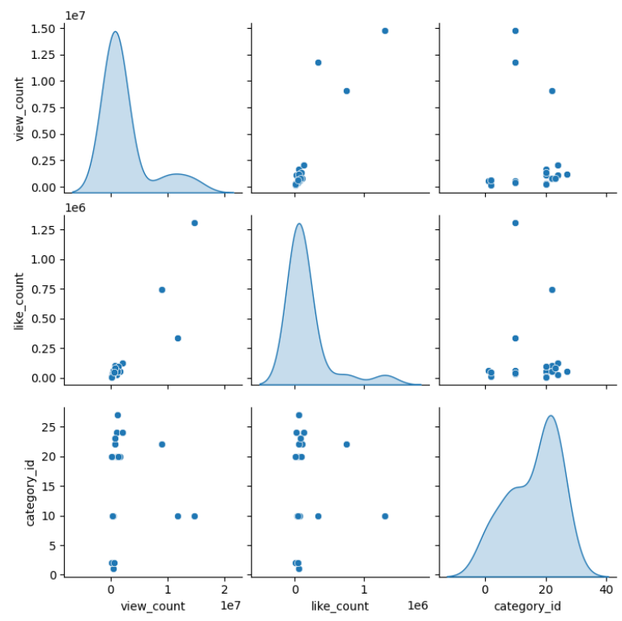

Figure 2. Diagnostic Analysis: Exploring Relationships

Figure 3. Distributions of Key Engagement Metrics

Future Work

- Principal Component Regression (PCR) integrates PCA with linear regression to address multicollinearity and reduce the dimensionality of predictor variables, aiming to enhance model accuracy and interpretability.
- Subsequently, predictive models can be developed and evaluated.

Further Exploratory Data Analysis

- Strong positive correlations among view count, like count, and comment count were observed.
- Analysis of tags against view count showed a weak relationship, indicating tags' role in video discoverability.

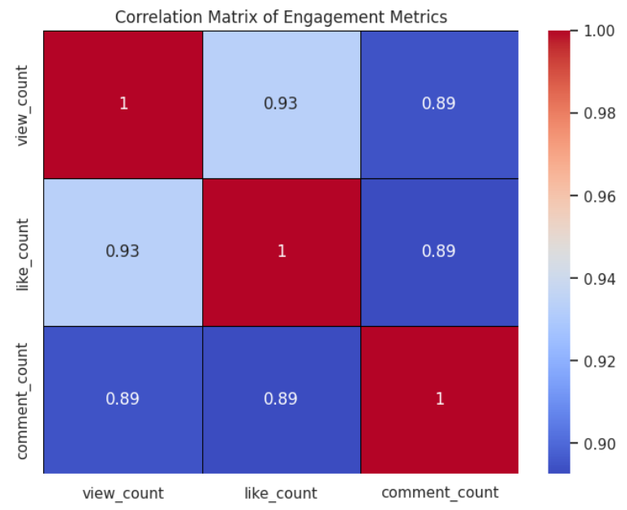

Figure 4. Correlation Matrix

- The number of tags for each video was calculated and analyzed against the view count, with a scatter plot.
- This plot indicated a weak relationship, suggesting that tags contribute to a videos discoverability.

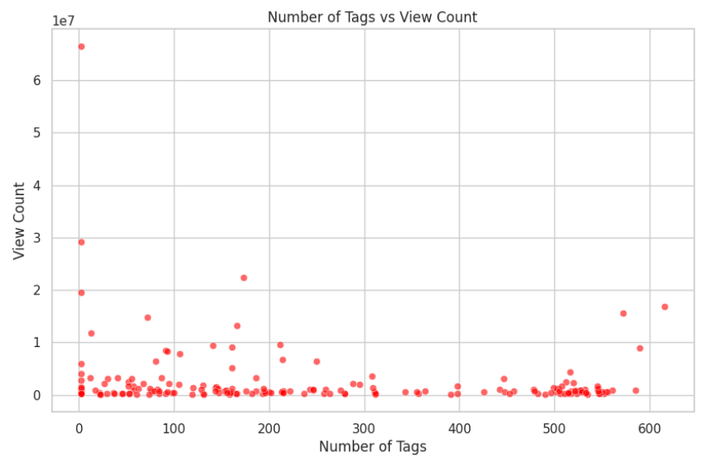

Figure 5. Number of Tags vs. View Count

References

Everitt, B. S. and T. Hothorn (2011). *An Introduction to Applied Multivariate Analysis with R*. 1st. New York, NY: Springer.

Izenman, Alan J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. 1st. New York: Springer.

Lu, Jiaxi (2023). "Trending Videos on YouTube". In: *Journal of Education Humanities and Social Sciences*. Available online: `https://www.researchgate.net/publication/367540346_Trending_Videos_on_YouTube`.