Study of Italian Grape types through Principal Component Analysis

Student Number: 10219553

0.1 Abstract:

PURPOSE: Investigate & evaluate the effectiveness of Principal Component Analysis (PCA) for dimensionality reduction, scored on the Davies-Bouldin index.

METHODS: The Singular Value Decomposition (SVD) was taken of a zero-mean data matrix. The PCA was used to reduce this dataset from 13D to 2D. The smallest DB was calculated from all column pairs and subsequently plotted.

RESULTS: PCA is effective for the reduction of the dimensionality of a zero-mean dataset. It can provide DB indices of under 1.5 with standardized zero-mean data.

CONCLUSIONS: The algorithm implemented over the given dataset met the metrics evaluation criteria specified in the scientific question. Using PCA on standardized zero-mean data can yield DB indices below 1.5, and further extension of the algorithm using Robust PCA may reduce anomalies and noise, resulting in better overall metrics.

Word Count: 127

0.2 Introduction:

The objective of this study is to investigate the effectiveness of Davies-Bouldin (DB) indices and Principal Component Analysis (PCA) for the reduction of the dimensionality of data.

The DB index takes the distances between values within a cluster, and divides it by the distances of all centroids of clusters in a matrix. A smaller DB index implies that the values within a cluster are close together while the centroids are further apart. A small DB index is a solid metric to measure the accuracy of the dimensionality reduction.

Singular Value Decomposition (SVD) takes an MxN zero-mean matrix and computes three new matrices, the product of which returns the original matrix. The three matrices are, in order, an orthogonal matrix U, MxM, which has the left singular values, a matrix S, which is diagonal, and finally V, NxN, which has the right singular values and is also orthogonal.

Principal Component Analysis (PCA) utilizes the SVD to reduce the dimensionality of a data set, while considering accuracy. Reducing the dimensionality of data comes at the cost of losing data, and PCA is a method that preserves much of that data. To perform PCA, the first two right singular vectors from the V matrix are taken from the SVD, and multiplied with the Zero-mean data matrix. In our case, this will reduce the dimensionality of the dataset from 13D to 2D.

The scientific question to be looked at is, given a 13D dataset, can PCA provide us with Davies-Bouldin scores under 1.5.

0.3 Methods:

The analytical framework of this study is structured into four distinctive phases, each meticulously designed to explore and analyze the dataset from initial calculations to the final visualization of clusters. This approach ensures a comprehensive understanding of the data characteristics and underlying patterns.

The study commences with the computation of the Davies-Bouldin index for each pair of columns within the data matrix. Utilizing the dbinfex function, this process involves the selection of two distinct columns along with the yvec parameter for the calculation. The indices generated from this operation are systematically recorded and stored in the Scoremat matrix, establishing a foundational assessment of the dataset's initial clustering tendencies.

Following the initial analysis, the data matrix undergoes a zero-mean transformation, a critical step for normalizing the dataset. This transformation facilitates the subsequent Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). By multiplying the zero-mean matrix with the first two right singular vectors from the SVD output, we perform PCA. The results of this calculation are encapsulated in the Zeromat1, which specifically stores the outcomes related to the zero-mean transformation. In this phase, the Davies-Bouldin index is recalculated to evaluate the clustering efficiency post-transformation, providing insights into the data's dimensional reduction and its impact on cluster delineation.

Building on the previous phase, the study progresses with the standardization of the Zero-mean matrix. This step is crucial for normalizing the dataset further, ensuring that the subsequent analysis is not biased by varying scales within the data. Following standardization, the study repeats the analytical process akin to Phase Two, emphasizing the robustness and consistency of the analysis. The resulting data from this phase is stored in the Smat matrix, which serves as a repository for the standardized computations.

The culmination of the study is marked by the visualization of the analyzed data, employing the gscatterplot function. This phase focuses on graphically representing the clusters and their labels, differentiated by distinct colors. The visualization not only facilitates an intuitive understanding of the cluster formations but also highlights the analytical journey from raw data to structured insights. This graphical representation serves as a critical component of the study, enabling the visual assessment of clustering effectiveness and patterns within the dataset.

0.4 Results:

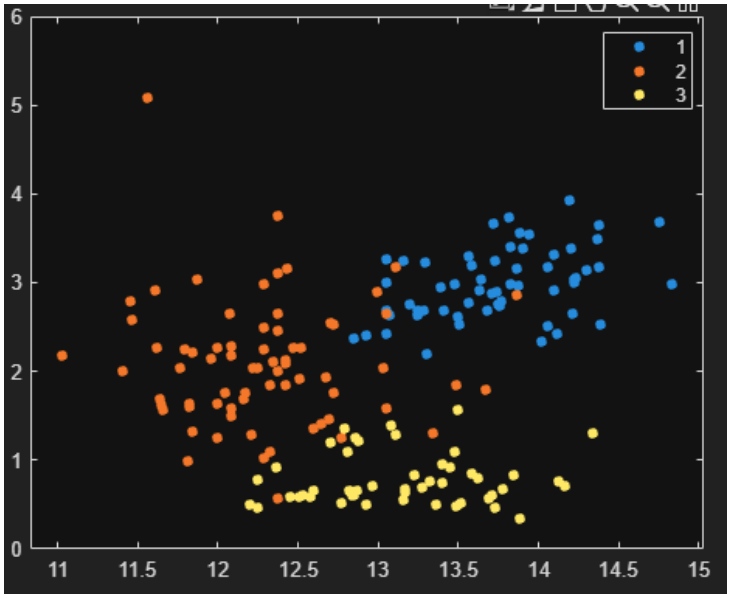

Figure 1: The above figure is a scatter plot of the first and seventh columns of the data matrix. The clusters are somewhat well-differentiated, but they lack some spacing.

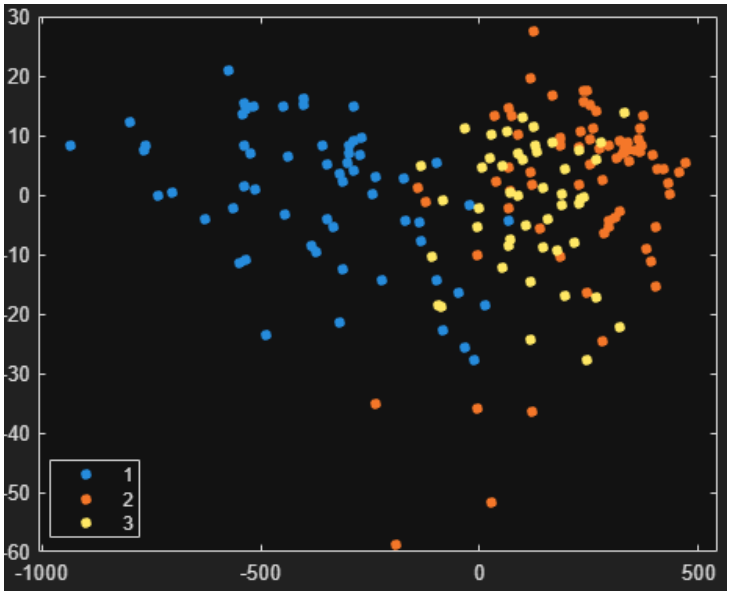

Figure 2: The above figure is a scatter plot of the PCA matrix when applied on the zero-mean data matrix. The data is poorly clustered vertically, with decent segmentation along the horizontal axis.

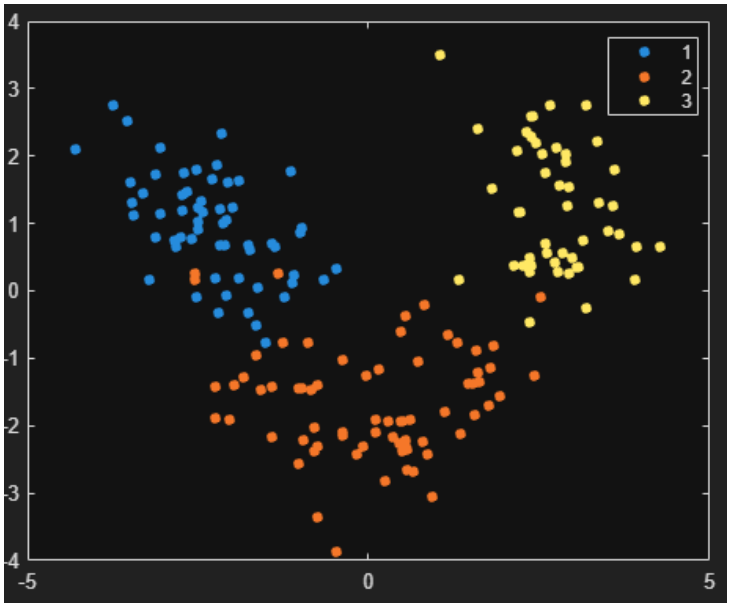

Figure 3: The above figure is a scatter plot of the PCA matrix when applied on the standardized data matrix. There is significant visible segmentation along the vertical and horizontal axes.

Table 1: Results of the study using the DB Index. The first column is the type of test, the second is the score achieved, and the third is the set of significant columns that provide the best reduction of dimensionality.

Test	DB Index	Variables
Data Columns	0.7875	[1 7]
Raw PCA	1.5148	
Standardized PCA	0.6392	

0.5 Discussion:

The findings of this study are wholly contained within the figures and table in the results section. The table compares the different tests and their respective Davies-Bouldin indices. The scatter plots visualize the distribution of the clusters found through the analysis, with each having its own color (Blue, Yellow, Orange).

The computational analysis found that the two columns of the data matrix that resulted in the lowest DB score were columns 1 and 7. These columns scored a 0.7875, which meets the evaluation critera specified in the scientific question. Figure 1 has clusters with centroids that are far apart, and values within each cluster that are relatively close. Applying PCA and SVD on the zero-mean data matrix using the matrix V results in a higher DB index of 1.5148, which is above the evaluation critera. The centroids of figure 2 were also close together but the values were much farther apart. Figure 3 was for the standardized data, and it had the lowest DB index of 0.6392, with centroids that were furthest apart and values within the clusters that were very close together.

Preliminary analysis shows that larger DB indices indicate smaller distances between the centroids of each cluster, and simultaneously imply larger distances between each point within a cluster. The third DB is the smallest, and it easy to tell visually that the centroids are much farther apart than the other two graphs and there is very little ”intermingling” between the points of different labellings. Additionally, when comparing Figures 1 and 2, it is clear that the results of PCA on zero-mean unstandardized data are poor, as the values within each cluster often overlap. Looking at the X-axis, we can also note that the steps are in the 500s, showing that the centroids are quite far apart, but the data within them is also very spaced out.

Through this analysis, it is clear that PCA is an effective for reducing the dimensionality of large datasets. However, the main pitfall of this is that, in order to meet high evaluation standards, the data must be standardized. This is shown in the vast differences between figures 2 and 3.

One extension of Principal Component Analysis is Robust PCA [1]. Robust PCA is designed to separate a data matrix into a low-rank matrix and a sparse matrix, and it aims to identify and handle outliers, noise, and other anomalies more effectively than standard PCA. This method is particularly useful in applications where the data is corrupted by substantial sparse errors or outliers, allowing for the extraction of the true underlying low-dimensional structure [2]. The data in this study comes from the composition of grapes from a particular region in Italy. The growth of these plants is, of course, subject to weather patterns and the composition of soil. Especially in smaller, rural areas, the surrounding shrubbery and composition of soil can differ greatly depending on location. As a result, there may be significant anomalies in the grapes based on a variety of variables not captured in this data, which Robust PCA could help mitigate.

In conclusion, Principal Component Analysis is an effective method for the dimensionality reduction of large datasets, especially when the data is standardized. It meets the evaluation critera specified in the scientific question, achieving a DB score of 0.6392 on standardized data, meeting the goal of 1.5. Further algorithmic exploration using Robust PCA may result in better metrics through de-noising and removal of anomalies.

0.6 References:

1. Wikipedia contributors. Robust principal component analysis [Internet]. Wikipedia, The Free Encyclopedia; 2024 [cited 2024 February 28]. Available from: https://en.wikipedia.org/wiki/Robu
2. Candes EJ, Li X, Ma Y, Wright J. Robust Principal Component Analysis? [Internet]. arXiv; 2009 [cited 2024 February 28]. Available from: https://arxiv.org/abs/0912.3599
3. Fazel M, Pong TK, Sun D, Tseng P. Hankel matrix rank minimization with applications to system identification and realization [Internet]. CORE; [cited 2024 February 28]. Available from: https://core.ac.uk/download/pdf/4727975.pdf
4. Ellis RE. CISC 271 Class 14: SVD – Singular Value Decomposition. 2021 [cited 2024 February 28]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class14.pdf
5. Ellis RE. CISC 271 Class 15: Orthonormal Basis Vectors and the SVD. 2021 [cited 2024 February 28]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class15.pdf
6. Ellis RE. CISC 271 Class 18: Principal Components Analysis – PCA. 2021 [cited 2024 February 28]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class18.pdf