Assessment of Data after Dimensionality Reduction through PCA using SVD

ABSTRACT

PURPOSE: To explore the use of PCA and evaluate the clusterings generated by applying it on raw and standardized data using DB index scores

METHODS: Principal Component Analysis using Singular Value Decomposition and DB Index algorithms are explored in this report. We make three varying approaches to reducing the dimensionality of the data in wine.csv to explore the models produced by the principal features that define the data.

The first makes use of a simple nested iteration of the features in the dataset to find the specific pair of features with the lowest DB index which best model the data in wine.csv in a lower dimensional space.

The second makes use of Principal Component Analysis (PCA) using Singular Value Decomposition (SVD) to reduce the dimensionality of wine.csv to its two principal components.

The third approach is identical to the second except for one important diverging aspect wherein the data is standardized prior to being subjected to PCA using SVD.

RESULTS: The results clearly demonstrate the effectiveness of each algorithm in modelling the three wine cultivars in the form of a summarized table Table 1 and scatter graphs for the algorithms with the respective K-Means analyses in Figures 1, 2 and 3.

CONCLUSIONS: We concluded that the application of PCA on standardized data for the wine cultivars was highly effective, in stark contrast to its use with raw data, and that this algorithmic model is suitable for large datasets with extremely high variances and many outliers. Also, that DB index is a highly useful tool to evaluate the clustering produced by any of these supervised algorithms.

Word Count: 261

INTRODUCTION

The purpose of this exploration was to explore the use of PCA and evaluate the clusterings generated by applying it on raw and standardized data using DB index scores.

PCA is a dimensionality reduction technique used to explore and visualize high-dimensional data in a low dimensional space whilst retaining the most important information in the dataset by computing principal components that define most of the variance present in the data. This was done with the aid of SVD which provided the right singular vector V, which when multiplied by a raw or standardized independent variable matrix produces a ranking of the principal components involved in the dataset. DB index index is, in essence, a clustering evaluation metric which is widely used to assess the quality of a clusters produced by any specialized algorithm.

The question being explored is the effectiveness of a PCA algorithm applied to raw and standardized datasets in conjunction with a DB index evaluatory metric and its use in actually finding principal components in contrast to doing so with SVD.

METHODS

We make three varying approaches to reducing the dimensionality of the data in wine.csv to explore the models produced by the principal features that define the data. The assessment of these approaches using DB index scores is appropriately discussed in the discussions section of the report.

The first makes use of a simple nested iteration of the features in the dataset to find the specific pair of features with the lowest DB index which best model the data in wine.csv in a lower dimensional space. The second makes use of Principal Component Analysis (PCA) using Singular Value Decomposition (SVD) to reduce the dimensionality of wine.csv to its two principal components. The third approach is identical to the second except for one important diverging aspect wherein the data is standardized prior to being subjected to PCA using SVD.

The function a3_20292366 is the one that contains all the executory code for producing the DB index scores and plots of the dimension-reduction techniques used in this exploration.

It makes use of the pre-given function dbindex which takes in Xmat, which is the matrix of all independent variables, in this case, raw datapoints of the three cultivars of wine being analyzed, and lvec (defined as yvec in function a3_20292366), which is the label vector for the three cultivars of wine. dbindex calculates the mean DB score of the DB scores calculated of all neighboring clusters by iterating through all indices.

Xmat and yvec (defined as lvec in dbindex) are first extracted from the data extracted from wine.csv using csvread and skipping over the first column. They are then transposed to make them viable arguments for dbindex and more suitable for analysis in a3_20292366.

Part (a) of the function iterate through the Xmat, which contains all the raw cultivar statistics, finding the DB index of all possible pairs of columns, and then finding the pair with the lowest possible DB index.

min_score is initialized as the largest value possible as to keep the algorithm bug-free as all possible datapoints are going to be compared against it, while the best pair vector stores the best possible pair ie. The pair with the lowest DB index. This allows for reduction to the most important dimensions during further analysis, as it ensures that our clustering results will be meaningful.

We next calculate the K-Means clusters of the pair with the lowest DB index which we later use to compare our dimension-reduction results against those produced by raw PCA and standardized PCA, using the built-in kmeans function and storing the relevant results in idx.

We then plot the results for the dimension reduction and the K-Means analysis in the same figure using the nexttile and gscatter functions in MATLAB.

Part (b) of the function conducts PCA using SVD on the raw data in wine.csv We do so using the right singular vectors 'V' from the Singular Value Decomposition analysis of the Xmat, which contains the raw data from all the features of the cultivars. The columns of V are ordered by decreasing magnitude of corresponding singular values in matrix 'S', which determines the order of importance based on the level of variance present inside the raw data for each of the features.

We start by computing the mean of Xmat, which contains the raw data from all the features of the cultivars, and subtracting the zero-means of all features from their raw data to convert Xmat to a mean-centered XmatM for effective use of PCA. We then find the left singular vectors 'V' through an SVD analysis on the mean-centered XmatM using the builtin function svd.

We then explicitly state the no. of principal components that we want to use in our PCA, which in our case is 2, by stating compNum = 2, which stays consistent for part (c) as well. An interesting manner to obtain the most efficient no. of principal components is to do r = rank(XmatM) and then plot(sum(S)/sum(sum(S))) which shows the optimal no. of principal components to be used at the inflection points on curve, which really is a product of the level of variance in each dimension or feature which contributes most to modelling the data as a whole.

Then, Z2 calculates PCA scores here and obtains the principal components by multiplying Xmat with V which, as outlined in the paragraph above, help extract the variables that 'carry' most of the weight in term of information and use them to produce a meaningful visualization. This PCA is then scored using the dbindex function and the results rounded to 4 decimal places for accurate interpretation and comparison.

The K-Means clusters are then calculated in the same manner as that in part (a), and the results for both the raw PCA and the K-Means clusters of the data in the raw PCA is plotted in the same manner as part (a) using the nexttile and gscatter functions in MATLAB.

Part (c) of the function conducts the standardized PCA using SVD for the data in wine.csv. The process for this is identical to the one outlined for part (b) except for the salient difference where we standardize the data in Xmat. The only change we make for this part is in the initialization of Xmat_std wherein we make use of the built-in function zscore on Xmat.

The final part of the function is involved in creating summarized output for the DB index scores for each dimensional-reduction technique using the table function and vectors of relevant output.

RESULTS

Table 1 contains the summary of results of using the Davies-Bouldin (DB) index on Raw Data, PCA conducted on raw data and PCA conducted on standardized data.

Figure 1 shows the plot resulting clusters of the best pair of features for the 3 cultivars and their respective K-Means clusters.

Figure 2 shows the plot of resulting clusters from PCA conducted on raw data and their respective K-Means clusters.

Figure 3 shows the plot of resulting clusters from PCA conducted on standardized data and their respective K-Means clusters.

Table 1: Summary of results using the Davies-Bouldin (DB) index on Raw Data, PCA conducted on raw data and PCA conducted on standardized data. The integers in the third column are the indices into the data variables that provide the numerically best dimensionality reduction.

Test	DB Index	Selected Variables
"Data Columns"	0.7875	"[1 7]"
"Raw PCA Scores"	1.5148	<missing>
"Standardized PCA"	0.6392	<missing>

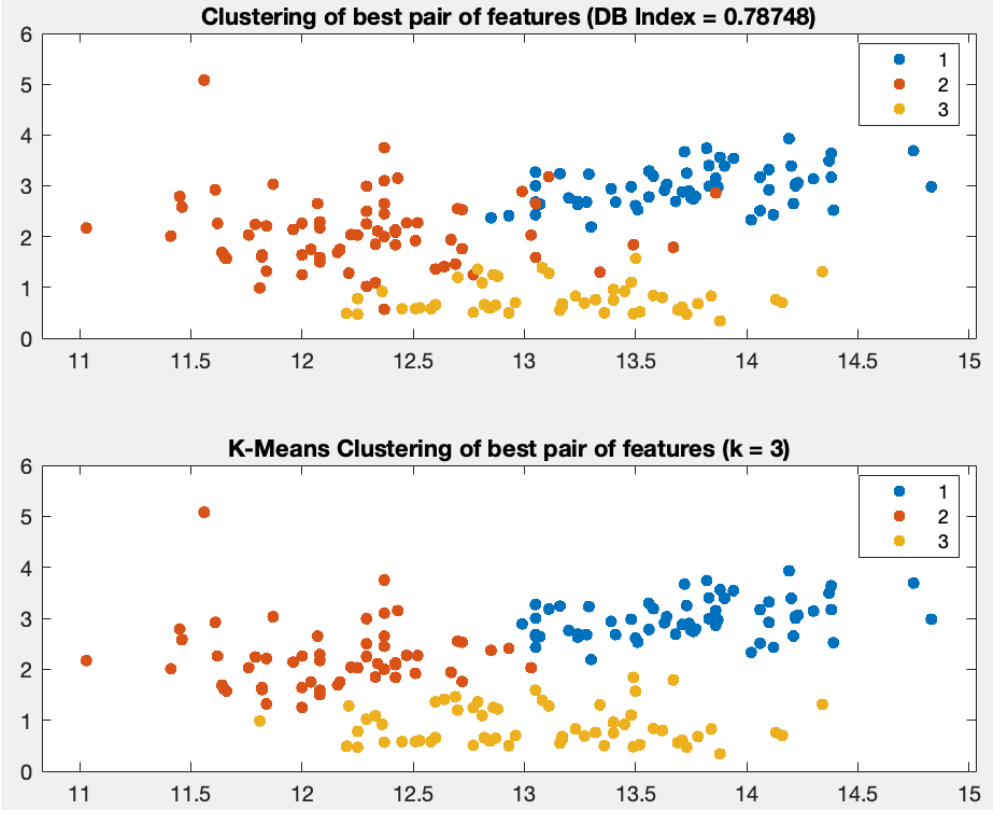

Figure 1: Plot of resulting clusters of the best pair of features for the 3 cultivars and their respective K-Means clusters.

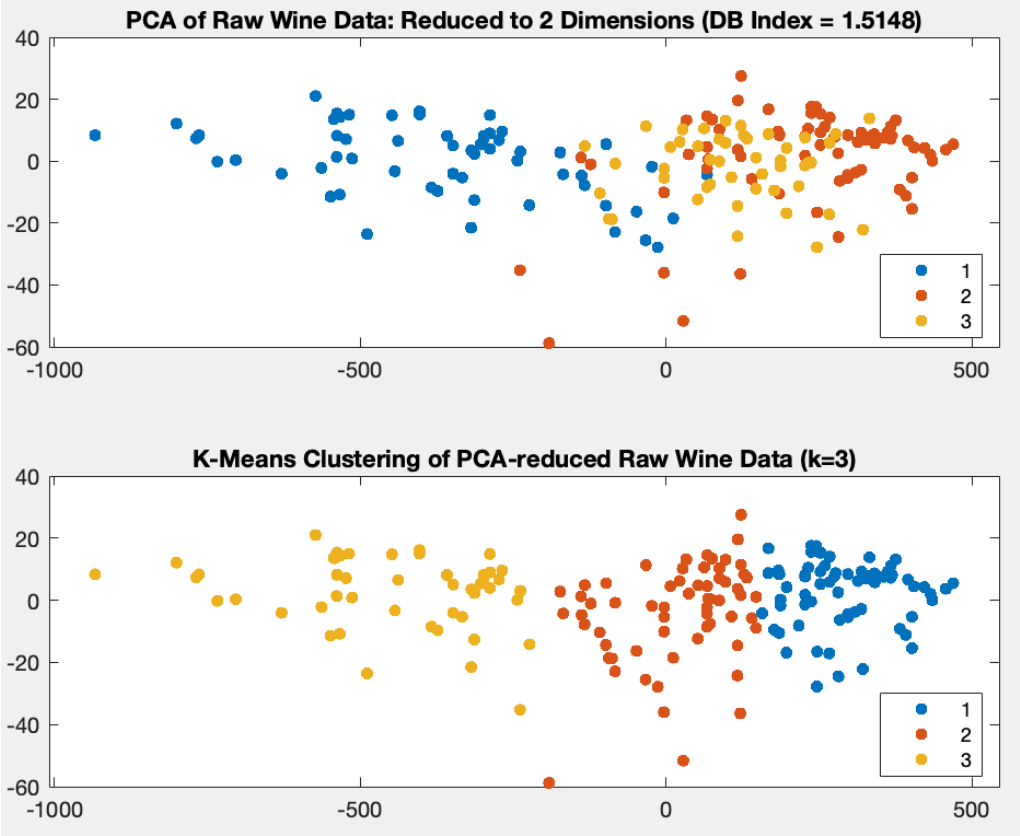

Figure 2: Plot of clusters from PCA on raw data and their respective K-Means clusters.

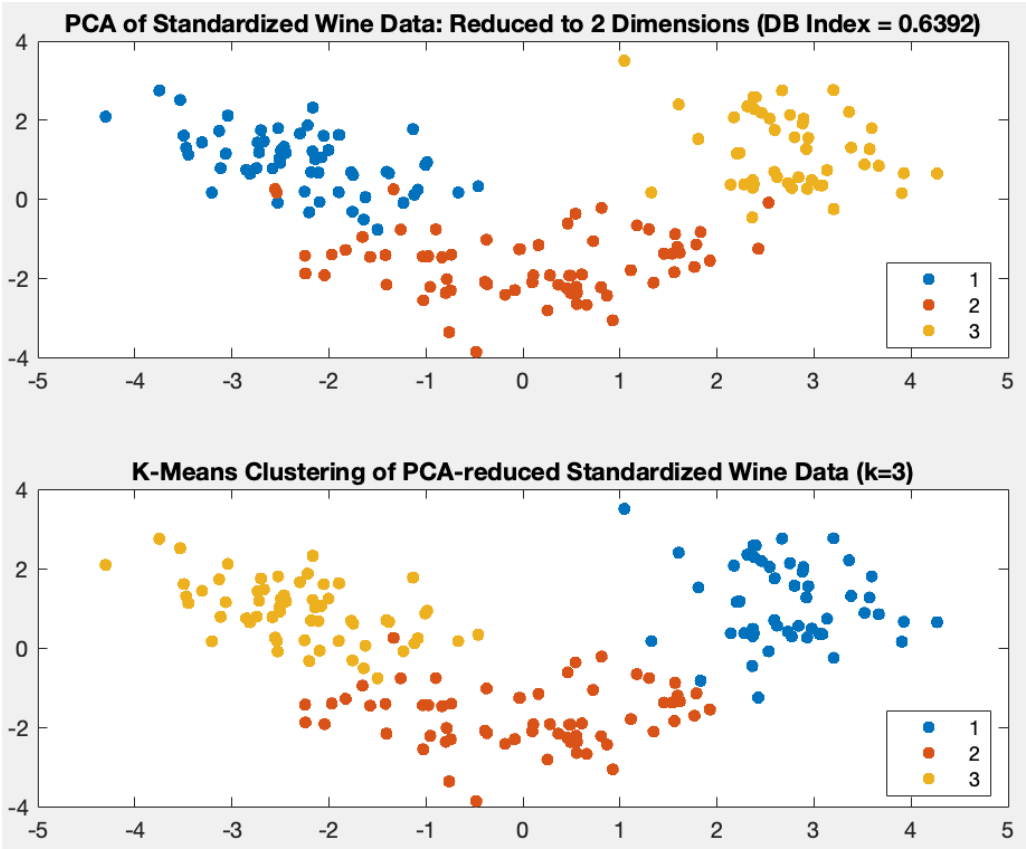

Figure 3: Plot of clusters from PCA on standardized data and their K-Means cluster

DISCUSSION

The results section shows the Davies-Bouldin (DB) index scores for three varying approaches to reducing and visualizing high-dimensional data into a lower-dimensional space while retaining the most salient information in the data.

The DB index is, in essence, a clustering evaluation metric which is widely used to assess the quality of a clusters produced by any specialized algorithm. It measures the average distance between each cluster center and the center of its nearest neighbor relative to the average distance between each cluster center and the objects within its own cluster (Davies & Bouldin, 1979).

To interpret the results, one must first understand the relationship between PCA and SVD. We know that the principal components of a dataset can be calculated by multiplying the data matrix with the matrix of the right singular vectors 'V' calculated using SVD. The importance of the V lies in its ordering. The columns of V are ordered by decreasing magnitude of their corresponding singular values in the eigenvalue matrix 'S'. An observation that must be made in this relationship is that the first column of V corresponds to the direction of maximum variation in a dataset, and the second column to the direction of the second maximum variation, and so on.

These represent the ranking of the principal features in any given dataset, which are the source of most of the variance present in the data. Thus, the features with the most weight that lend the most to the data's statistical architecture are captured by conducting PCA through SVD. This is especially relevant to our dataset as the level of variance between the different features is extremely high and the average variance is around 7645.5 arbitrary units.

The unsupervised algorithm in part (a) that relies on the DB index scores of the two best features that model the rest of the data produces a relatively good clustering of the three cultivars being modelled which is apparent from a brief glance at its K-Means counterpart and its low DB index score of 0.7875 shown in Table 1.

Our algorithm indicates that the first and seventh features of the data, namely, Ethanol and Total Phenols, are the defining components. A cursory look at the V matrix from the SVD analyses done for both parts (a) and (b) clearly show the eigenvectors of the first and seventh features having the largest values which relates to the weights of variance that they carry in the data. The largest value from the largest eigenvectors from part (a) were -0.9998 and 0.9233 for columns one and seven respectively. The same for part (b) were -0.4229 and 0.5954 for columns one and seven respectively. This supports our axiomatic reasoning about the relationship between PCA and SVD.

There are quite a few datapoints that are not don't match up in the K-Means clusters shown in Figure 1 which is expected because a dataset of such high variance is likely to have lost of noise present in the form of outliers. This is also the case in the graph for the raw PCA which, when compared to the clustering provided by part (a), is relatively less effective in producing clusters which have maximally concentrated centroids and maximized space between clusters. The clusters very clearly overlap to the point where two of them are inextricable from each other, specifically for cultivars 2 and 3 shown in Figure 2. This is also supported by the relatively high DB index score of 1.5148 shown in Table 1.

The scatter graph of the Raw PCA clustering is also obviously not suitably effective at modelling the clusters for the cultivars in comparison to its K-Means counterpart as which clearly has better clusters. This, though visually appealing, is likely not a good clustering for the dataset as it is entirely unsupervised and thus ignores the biases that the different weights the features embody in the data. This is true for all K-Means graphs in this exploration which is why they are not a good exploratory tool to interpret in isolation and must be backed up by a supervised algorithm which takes labels into account.

The graphs for part (c) are seemingly the best representation that this exploration produced, a fact which is apparent from its very low DB index score of 0.6392 shown in Table 1. The graph for PCA on standardized data shows clearly distinct clusters with a minimal number of outliers which speaks to the algorithm's efficacy at modelling the cultivars. When compared to its K-Means counterpart, we see an extremely close relationship as there is only three obvious outliers present which also speaks to how well the cultivars are modelled by a standardized dataset modelled by K-Means clustering. This symbiotic relationship present between an unsupervised and a supervised algorithm indicates a good clustering. Since the clustering results for the K-Means graph in Figure 3 show that the data points within each cluster are similar to each other, it is a good indication that the PCA has effectively captured the underlying structure of the data.

The above argues a great case for standardizing data even when features present in the data might be measured on differing scales, so that none of them may dominate the model, and equally contribute to whatever analysis is being done on them. As can be seen when the raw PCA and standardized PCA graphs are compared, standardization is effective at minimizing any noise that may skew the model due to outliers. Standardization in the case of this dataset does not diminish our ability to interpret the model, as is the case in many datasets where scales of features are vital in the process of interpretation or ones which have uniform scales, but rather enhance it a lot as is obvious by the great clustering seen in the Figure 3 for standardized PCA.

Variance and its relationship with PCA is a prime aspect of this standardized dimensional-reduction algorithm. Standardization can affect the variance of our dataset by shifting the mean of each feature to zero which affects covariance between features, while scaling each feature to have unit variance. The scaling effect of standardization, computed by comparing the average variance for all features in the raw dataset, which had an approximate average variance of 7645.5, with that of the standardized dataset, which had an approximate average variance of 0.1642, was 7645.3362. This ensured equally weighted analysis and helped avoid biases arising from units of measurement or differences in scale, which is the case in most real-world datasets.

In conclusion, the application of PCA on standardized data for the wine cultivars was highly effective and this algorithmic model is suitable for large datasets with extremely high variances and many outliers. Application of PCA on raw data, however, may not be suitable in those cases but may be necessary for when the scales in the data are similar and the raw weights of the features are heavily involved in interpretation. DB index is a highly useful tool to evaluate the clustering produced by any of these supervised algorithms.

REFERENCES

Davies, D. L., & Bouldin, D. W. (1979, April). *A cluster separation measure | IEEE Journals & Magazine | IEEE Xplore*. A Cluster Separation Measure. Retrieved March 2, 2023, from https://ieeexplore.ieee.org/document/4766909

MIT. (n.d.). *Singular Value Decomposition (SVD) tutorial*. Singular value decomposition (SVD) tutorial. Retrieved March 2, 2023, from https://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm