DIMENSIONALITY REDUCTION AND ITS EFFECT ON DATA

1. ABSTRACT

PURPOSE: To analyze as to which combination of analytical chemistry values, or cultivars, contributes most effectively to reducing the complexity of the data set, by providing the best dimensionality reduction, reducing the size of the data using PCA, and standardizing the data.

METHODS: The best data variables are found by figuring out the lowest DB indexes, and a scatter plot of this function is then formed by using the 'gscatter' function, along with different pair of vectors such as 'avec' and 'bvec'. To generate separate figures for each scatter plot, the function 'figure' is used, and the PCA of the function is calculated by finding the mean of each column in the data set, which is then followed by performing Singular Value Decomposition upon the numerical means that have been found. This allows for the "score" vectors to be computed, and finally, the data is standardized by finding the zscore function within MATLAB, which then allows for the PCA to be applied to the standardized data using the Singular Value Decomposition approach.

RESULTS: The best pair of values in the data, or columns, that provide the best dimensionality reduction were Column 2 and Column 7. This was found by computing the lowest DB index, which came out to be 0.787478. The raw PCA score, meanwhile, had been calculated using the SVD, and its value was 1.514778. The PCA score remained the same even after the data was standardized.

CONCLUSIONS: The best pair of features show which columns effectively capture the greatest and most relevant information within the data set, and in this case, the greatest and most relevant information within the data set is shown through columns 2 and 7. The DB index is 0.787478, and since that value is close to 0, better separation and more distinct clusters are formed within the data set while finding the best pair of features. This is also the case when the clustering quality is considered for PCA-transformed data and standardized PCA-transformed data, since the value of both of these scores is 1.514778, which is, once again, quite close to 0.

2. INTRODUCTION

The scientific objective that was posed here was to assess the impact of dimensionality reduction on a dataset using Principal Component Analysis (PCA) and evaluate the results using the Davies-Bouldin (DB) index.

To understand this report, one must have knowledge with regards to some of the background topics addressed in this report. This report is based upon the relationship between the analytical chemistry values that were found for the 178 samples of wine, as well as the thirteen cultivars or grape types provided for each sample of wine. One of the most important topics that this report is based upon is that of Principal Components Analysis (PCA), which is a dimensionality reduction technique used in data analysis. Its primary goal is to transform a dataset into a new coordinate system, where the axes correspond to the principal components of the data. These principal components are linear combinations of the original features, and they capture the maximum variance present in the data. Through the PCA, the current thirteen-dimensional data, which is being shown through thirteen cultivars, is reduced to two-dimensional data by figuring out the principal components, which are then ordered by the amount of variance they capture in the data. In this case, the two principal components with the highest order of variance have been retained, with the rest being omitted, which has thus reduced the dimensions within the data set. Another important topic that must be understood by a reader prior to reading this report would be that of Singular Value Decomposition (SVD). The SVD is often used to decompose the data matrix by creating three other matrices. It plays a vital role in computing the PCA, since the principal components are extracted from the SVD. Data Standardization is also a concept that's been referred to within this report, and it essentially scales the factors so as to ensure that all factors have a zero mean and unit variance. This is to ensure that all variables contribute equally to the analysis, and in this report, data standardization allows for all variables within each cultivar to contribute equally to the analysis of the data.

To test the scientific question, multiple steps took place. Firstly, the data set was analyzed, with an understanding that the matrix size was in a csv file, containing 179 rows and 14 columns. For each pair of variables within the data set, the Davies-Bouldin (DB) index was calculated, and the pair of variables that yielded the lowest DB index was considered the best for dimensionality reduction. From here, zero-mean data is used and the SVD is found. Two score vectors are computed from the results of the SVD, which then results in the computation of the PCA. The PCA also requires standardized data, and the 'zscore' function in MATLAB allows for the impact of data standardization to take place. The standardized PCA results, as well as the reduction of two-dimensional data was scored using the DB index. This took care of the numerical parts of the data analysis. For the scatter plots, three different figures are generated, one for each specific scientific question, and each scatter plot is based around the vectors that have 178 entries each.

3. METHODS

The first goal, as part of the task assigned, was to find the pair of values in the data that provide the best dimensionality reduction. In order to reach this goal, it was important to first prepare the data. This was done by loading the data, which was in the form of a csv file. The file included the analytical chemistry values for the 178 samples of wine, as well as the thirteen cultivars, or grape types, that had been coded numerically in the data file. Once the file had been loaded, the indices with the best pair of features, as well as the DB index, was initialized, in order to keep track of the optimal feature pair and its associated DB index. From here, each unique pair of features had to be considered, and for each unique pair, a two-column matrix that contains the values of the two features is selected from the entire data set. A DB index is calculated for each matrix, and if a DB index is lower than a current best DB index, the indices with the best pair of features, as well as the DB index is updated with the lower DB index value. Eventually, the best pair of features and their DB index is found. These values are then visualized using an optimal feature pair, by using a scatter plot to visualize the data points based on the optimal feature pair. To provide a better understanding, as well as clarity with regards to what the axis' are about and what the scatter plot is about as well, the x-axis and y-axis represents and is labelled with one of the best pair of features or columns, and the plot title represents the relationship between the axes respectively.

Once the columns that provided the best dimensionality reduction had been found, the data was to be reduced from thirteen dimensions to two dimensions using PCA, and this reduction then had to be scored. This was done by first subtracting the mean of each column from the corresponding elements from the two-dimensional array where each row corresponds to an observation (e.g., a wine sample), and each column corresponds to a different chemical value. This is also known as the zero mean of the data, and once this has been solved for, the Singular Value Decomposition can be applied to the zero mean. By applying the SVD, the left singular vectors, singular values, and right singular vectors are found for the zero mean. The left and right singular vectors represent the principal components in the original and transformed feature space respectively, while the singular values are crucial for determining the amount of variance captured by each principal component. The singular value is also quite important with regard to the prior knowledge that only two principal components need to be retained, and the larger the singular value, the more variance is captured by the corresponding principal component. This allows for the data to be reduced from thirteen dimensions to two dimensions, and now, the focus moves towards calculating the raw PCA scores. In order to compute the raw PCA scores, the first two columns of the right singular vectors are used, and these are then multiplied by the zero-mean data matrix. Since the right singular vectors represent the principal components in the transformed feature space respectively, which thus effectively projecting the data onto a 2D subspace. The raw PCA scores are received, and is also graphed through a scatter plot. The x-axis and y-axis represents and is labelled with one of the principal components each, and the title then denotes the relationship by being labelled accordingly.

The computation or calculation of the PCA score is based off of the same process, regardless of whether the raw PCA score is being found or whether the PCA score is being found. Therefore, the steps applied in order to calculate the raw PCA score will also apply here, albeit with one change. When the raw PCA score was to be computed, no standardization of data needed to take place. As a result, when that function was called within MATLAB, it was made evidently clear that the raw PCA score will not be a value that's calculated based off the standardized data. However, now that standardized data is required, the same function can be called once again, with the only change being that the PCA scores will be computed for standardized data, which must then be set from False to True. Once this is completed, the function when called, contains a conditional statement that takes care of when the data needs to be standardized. When this is the case, the zscore of the two-dimensional array where each row corresponds to an observation (e.g., a wine sample), and each column corresponds to a different chemical value is calculated. The zscore allows for the mean of a column to be subtracted from each element in a column, which then centers the data around zero. It then divides each element in the column by the standard deviation of the column, which then scales the data, ensuring that the standard deviation becomes 1. This completes the standardization of the data, and the standardized PCA scores are now computed using the same steps (and in the exact same order) that were also used so as to compute the raw PCA scores. This data is then graphed, with the x-axis and y-axis representing and being labelled with one of the principal components each, and the title then denoting this relationship by being labelled accordingly.

Overall, the onus of this task was to provide a practical introduction to evaluating algorithms for data analysis, specifically focusing on dimensionality reduction using Principal Component Analysis (PCA). Through standardizing the data for linear data analysis, implementing PCA using the Singular Value Decomposition (SVD) of zero-mean data, reducing the dimensionality of data, assessing the results of dimensionality reduction using the Davies-Bouldin (DB) index, and creating scatter plots for all of the data, a systematic and statistically robust analysis was provided with respect to the relationship between variables. The task also emphasizes the importance of original code, by avoiding built-in functions for dimensionality reduction, and following the specific guidelines required for output and presentation.

4. RESULTS

The results are primarily focusing on the Davies-Bouldin (DB) index, and this provides a measure of the effectiveness of dimensionality reduction through Principal Component Analysis (PCA). The DB index values for various scenarios reflect the quality of feature pairs and PCA scores. The impact of different variable combinations, shed light on pairs that contribute optimally to the reduction process. Meanwhile, scatter plots visually illustrate the distribution and separation of data points after PCA. The scatter plots provide a tangible

representation of the reduced-dimensional space, emphasizing the patterns and clusters that are discernible post-dimensionality reduction. Overall, the DB index values and scatter plots unravel the performance and visual impact of dimensionality reduction techniques on the provided dataset. The combination of quantitative metrics and visualizations allows for a comprehensive evaluation of the efficiency and patterns captured by the PCA process in different scenarios.

Table 1: Summary of results using the Davies-Bouldin (DB) index. The left column indicates the test and the right column is the DB index, presented using four digits of numerical precision. The integers in the third column are the indexes into the data variables that provide the numerically best dimensionality reduction.

Test	DB Index	Variable
Data Columns	0.78748	[1 7]
Raw PCA Scores	1.5148	
Standardized PCA	1.5148	

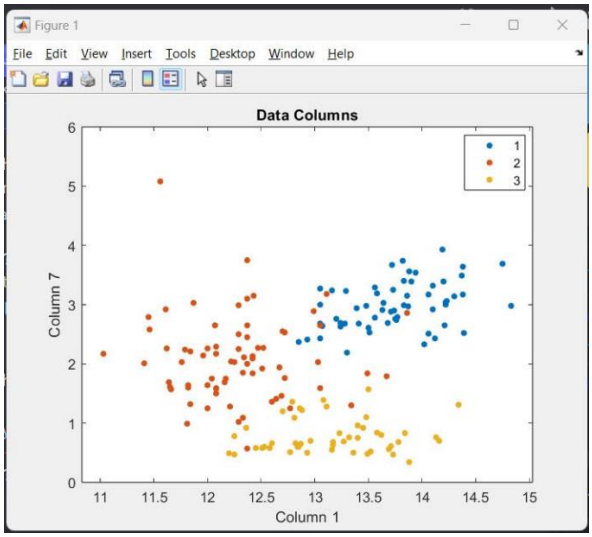

Figure 1: Scatter plot representing the pair of values in the data that provide the "best" dimensionality reduction. One of these pairs of values, or columns, is Column 1, which is denoted on the x-axis, and the other one, is Column 7, which is denoted on the y-axis.

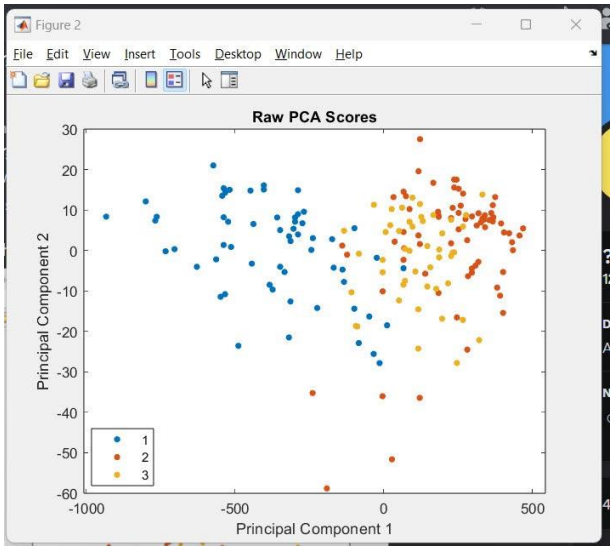

Figure 2: Scatter plot representing the principal components that provide the raw PCA scores. One of these Principal Components is denoted on the x-axis, while the other Principal Component is denoted on the y-axis.

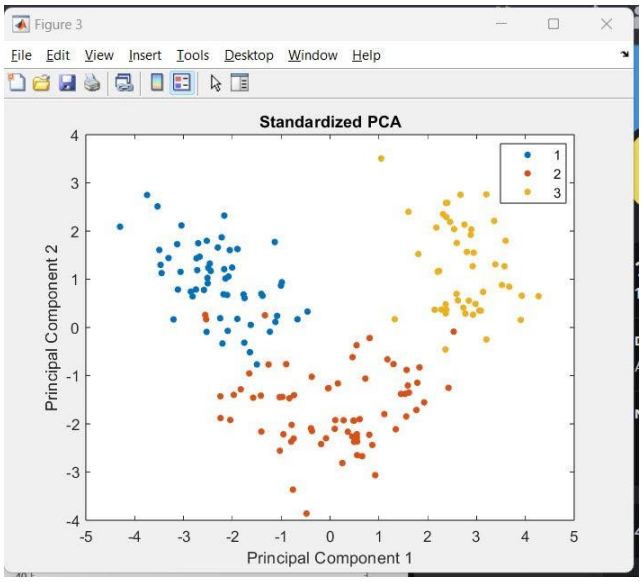

Figure 3: Scatter plot representing the principal components that provide the standardized PCA scores. One of these Principal Components is denoted on the x-axis, while the other Principal Component is denoted on the y-axis.

5. DISCUSSION

To understand the relationship between the different wine cultivars as well as the analytical chemistry values, a dimensionality reduction process through Principal Component Analysis (PCA) took place. By using the PCA, various trends and notable observations

emerged from the exploration of different scenarios. The DB index values, serving as a key metric, revealed distinct patterns that guide the selection of optimal feature pairs for dimensionality reduction. These trends are crucial in identifying combinations of variables that contribute most effectively to the preservation of information. The analysis identified a clear trend where lower DB index values corresponded to more favorable dimensionality reduction outcomes. Specifically, the best pair of features, pinpointed through an exhaustive search, demonstrated a consistent trend of yielding lower DB index values compared to other feature pairs. This trend suggests that certain variable combinations indeed contribute more optimally to the PCA process, aligning with the expectation of superior dimensionality reduction. Moreover, scatter plots visually reinforced the trends observed in DB index values. The plots showcased distinct clusters and patterns, highlighting the effectiveness of PCA in capturing underlying structures within the data. Notably, the scatter plot corresponding to the optimal feature pair emphasized a clear separation of data points, reinforcing the identified trend in DB index values. To summarize, the analysis uncovered consistent trends in DB index values and scatter plots, providing valuable insights into the effectiveness of PCA in dimensionality reduction, and the observations that can be made from these results will be discussed in the following paragraphs.

The first result that was received while completing this task was that the best pair of features were Columns 1 and 7, and the DB index score of these pairs of values was 0.787478. In the csv file that was provided, the total number of columns were 13, and the two columns that were chosen on the basis of dimensionality reduction were based off of the columns that would contribute most significantly to capturing distinct patterns in the dataset. In this case, it appears as though the Ethanol cultivar (Column 1), and the Flavanoids cultivar (Column 7), are able to create differentiable patterns the best among all columns within the data set. Within these two columns, certain characteristics or values play a substantial role in defining the separation between different groups or clusters present in the data. The effective coherent and separated representation of the underlying data clusters is crucial for tasks such as classification or visualization, where reduced dimensions should still preserve relevant information. Meanwhile, The DB index is a measure of the quality of clustering in a dataset. It quantifies the ratio of the "scatter" within clusters to the distance between cluster centroids. A lower DB index suggests more compact and well-separated clusters, which thus indicates effective dimensionality reduction. In this context, the obtained DB index of 0.787478 indicates a relatively low level of dispersion within clusters compared to the inter-cluster distances. Due to the lower DB index (0.787478), it can be implied that the variables, or the Ethanol and Flavanoids columns, which are columns 1 and 7 respectively, contribute significantly to capturing distinct patterns or structures within the dataset. The second result that was received through the report was that of the raw PCA score being 1.514778. This score received is based upon unstandardized data, and provides insights into the extent to which the information in the original dataset is retained through the transformation into principal components. A higher PCA score implies a more effective reduction in

dimensionality, where a significant portion of the original dataset's variability is encapsulated within the reduced set of principal components. Notably, however, the raw PCA score exceeds 1. This is a meaningful observation as PCA scores greater than 1 indicate that the first two principal components capture more variance than what is present in a single original variable. This is a positive outcome, as PCA aims to retain and capture a substantial amount of information while reducing the dimensionality. The score reflects the ability of the first two principal components to capture essential patterns and variability in the data. Since the PCA captures these patterns quite well, the transformation that takes place after calculating the PCA effectively compresses and represents the data in a more condensed form. The implications of this value can come handy for subsequent analyses as well, such as visualization or machine learning, since the reduced dimensions can be used with confidence such that essential information has been preserved. The last result received was that of the standardized PCA score being 1.514778 as well. Standardizing the data is a crucial preprocessing step that ensures all variables have a comparable scale, and is achieved by subtracting the mean and dividing by the standard deviation. This process removes the scale differences between variables, allowing PCA to focus on the relative importance of variables based on their contribution to variance. The fact that the standardized PCA score is similar to the raw PCA score suggests that the chosen features are inherently significant for capturing variance, and their importance is not influenced by the scale of the original variables. Furthermore, the standardized PCA score of 1.514778 implies that, even after standardizing the data (transforming it to have a mean of 0 and a standard deviation of 1), the effectiveness of dimensionality reduction remains consistent. This consistency suggests that the identified features, particularly the chosen pair of columns 1 and 7, play a pivotal role in capturing essential patterns within the dataset, irrespective of the original scale or distribution of the variables. The standardized data can be compressed into the first two principal components, and a score of 1.514778 indicates that the standardized features, even after being transformed, still retain a substantial amount of information. This provides confidence in the relevance of the selected features for subsequent analyses. It suggests that, regardless of the units or original distribution of the variables, the identified features are robust indicators for reducing dimensionality. The similarity in scores between raw and standardized PCA also ends up highlighting the generalizability of the dimensionality reduction process. The identified features can be considered as stable indicators for the dataset's structure, applicable across different scales or units.

In conclusion, the identified best pair of features, which are Columns 1 and 7, yield a DB Index of 0.787478, and this supports effective dimensionality reduction. The raw and standardized PCA scores of 1.514778 signify successful reduction, emphasizing the significance, as well as the robust contribution and impact of Columns 1 and 7 in capturing essential patterns within the dataset.

6. REFERENCES

1. Ellis RE. Design Matrix and Standardized Data [Internet]. Queen's University; 2021 [cited 2024 Feb 27]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class09.pdf
2. Ellis RE. SVD – Singular Value Decomposition [Internet]. Queen's University; 2021 [cited 2024 Feb 27]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class14.pdf
3. Ellis RE. Principal Components Analysis - PCA [Internet]. Queen's University; 2021 [cited 2024 Feb 27]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class18.pdf
4. Jaadi Z. A Step by Step Explanation of Principal Component Analysis [Internet]. Built In. 2019. Available from: https://builtin.com/data-science/step-step-explanation-principal-component-analysis
5. Data Standardization Explained - With Examples [Internet]. Sisense. Available from: https://www.sisense.com/glossary/data-standardization/