Evaluating effectiveness of differing methods of reducing dimensionality of data

Abstract

PURPOSE: Test the effectiveness of the PCA method for reducing the dimensionality of data on a data set with greater than 3 dimensions.

METHODS: Applying PSA, by using SVD and the components in order to create and score vectors that will define a projection of the dataset into a lower dimension, Then comparing the results to the best possible result of reduction of dimensionality using pre existing column vectors, and another application of PSA on standardized data.

RESULTS: The PCA dimension reduction had the highest DB index, the best fit 2 dimension had the second lowest DB index. The standardized PCA had the lowest DB index, and the most data retention. .

CONCLUSIONS:The PCA led to an accurate display of information, but a high margin of error due a lack of uniform units. The pair of columns led to a low error, but little information was sustained. The standardized PCA method allowed for low margin of error and a large amount of maintained variability. Thus standardized PCA should be used if the units are unimportant, PCA should be used when the units are important.

Introduction

Databases are large and with more than 3 dimensions of variability, it is impossible to visualize making it hard to interpret the data that has been collected. Reducing dimensionality of data is useful for comprehension and allows for clustering to be visualized in terms of what they are most similar to. Most importantly, it allows for the visualization of variability in data. This can be achieved in a few different ways, including principal component analysis (PCA). The idea of PCA is to project a zero mean observation onto a principal component that represents a lower dimension vector. This gives the best possible fit of lower dimensions to a higher dimension dataset allowing for a visualization for how the data interacts. PCA is a broad study of data, but commonly implemented via the use of the singular value decomposition (SVD). A method of factoring a real matrix into three distinct parts. The left singular vectors, the singular values, and the right singular values. These three parts allow for the use of the right singular vectors in combination with the zero mean dataset in order to create score vectors which will define the reduced dimension projection.

Methods

The data was loaded and prepared, removing the row headers and cultivar data (Which will be stored for later use) and transposing the data in order to get it into the form of column vectors being the headings. The data is then made into a zero mean data matrix, by subtracting the means multiplied by a vector containing all ones. In order to find the best pair of vectors to

represent the data in 2 dimensions, every possible pair of columns were iterated through and tested using the Davies-Bouldin (DB) index as a measure. The DB index was obtained using the given assignment function. In order to gather the part b values, the zero mean data was then used to do PCA. This was achieved by computing SVD of the zero mean data using matlab, gathering 2 vectors, the first and second singular vectors, these being the principal components. These vectors were evaluated using the Davies-Bouldin index to depict their effectiveness, inputting the two score vectors as a matrix and the cultivar matrix as a label. After evaluation, the vectors were multiplied by the zero mean matrix in order to get the projection of the dataset into two dimensions, now able to be plotted. After this was done, the data was standardized and the process ran again to gather the reduction of data using the standardized PCA. The score vectors were then analyzed in combination with the scatter plot to figure out the trends with each cultivar. The data was found by subtracting a cultivar's negative biases from the positive biases. then documenting which factors were above an arbitrary number for significance.

Results

Table 1: Displays DB values of the tested methods

	DB values	Variables	2D Representation Matrix
Lowest DB of existing pair of column vectors.	0.7875	[1,7]	Vector 1: [1;0;0;0;0;0;0;0;0;0;0;0;0;0] Vector 7: [0;0;0;0;0;0;1;0;0;0;0;0;0]
PCA	1.5148		PS1: [-0.0017; 0.0007; -0.0002; 0.0047; -0.0179; -0.0010; -0.0016; 0.0001; -0.0006; -0.0023; -0.0002; -0.0007; -0.9998] PS2: [-0.0012; -0.0022; -0.0046; -0.0265; -0.9993; -0.0009; 0.0001; 0.0014; -0.0050;-0.0151; 0.0008; 0.0035; 0.0178]
Standardized PCA	0.6392		PS1: [-0.1443; 0.2452; 0.0021; 0.2393; -0.1420; -0.3947; -0.4229; 0.2985; -0.3134; 0.0886; -0.2967; -0.3762; -0.2868] PS2: [0.4837; 0.2249; 0.3161; -0.0106; 0.2996; 0.0650; -0.0034; 0.0288; 0.0393; 0.5300; -0.2792; -0.1645; 0.3649]

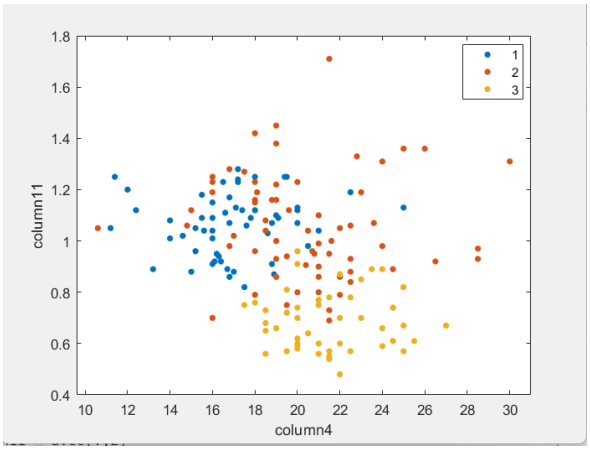

Figure 1: The scatterplot depicting the pair of 2 best columns to reduce dimensionality

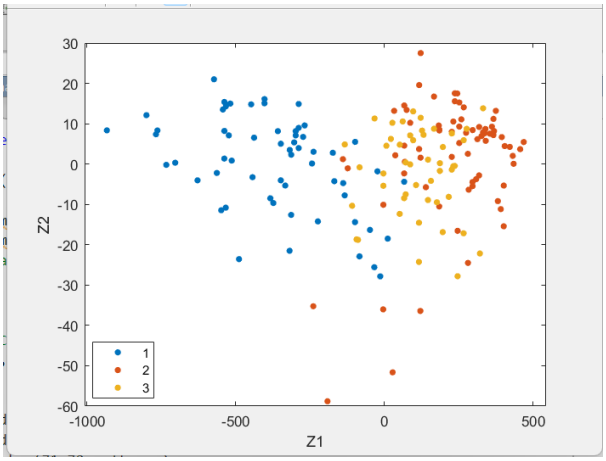

Figure 2: The scatterplot depicting the PCA of the data generated from the SVD

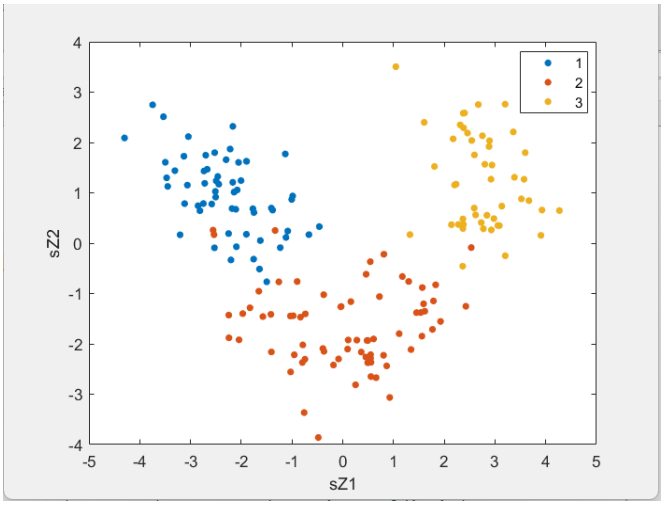

Figure 3: The scatterplot depicting the standardized PCA of the data generated from the SVD .

Discussion

Looking at part 1, and the corresponding DB value for that reduction of dimensionality. It is lower than the PCA value, but higher than the standardized PCA value. The score vectors being in vector one and seven corresponds to the best indication of the total wine is gotten from inspecting the Ethanol, and the Flavonoids. These are the most dependent variables out of the 13, these two columns give the best projection of the 13D data. The reduced data itself is not very useful, the scatter plot does not describe the data anymore than the 2 vectors, and thus very little correlation between the dimensional reduction and the actual dataset maintaining little of the variability within the dataset. There does not seem to be any reason to use this type of dimensional reduction for this data set, as the other options provide much more accuracy and retaining of variability.

Inspecting part 2, the 2D representation matrix is interesting, the values are similar to the vectors for part 1, where the values are being generated mostly from two vector columns, column 5 and column 13. These columns correspond to the data that has the highest number, ignoring the units. This indicates that a non standardized PCA is susceptible to numerical bias if there are units present. This shows that for this data, the PCA got skewed due to the large numbers of th columns, meaning that it should not be used on data such as the current data. Interestingly, the DB score is higher than the best combination of 2 columns, indicating a higher error. The PCA does a much better job than the column vectors of maintaining the variability of the data, an important factor of dimensionality reduction. The data is very skewed due to the nonstandardized nature of the dataset. Due to the fact that it takes in all the columns when creating the scores, much more of the variability is captured. Due to the nature of the units for the data, PCA should not be used unstandardized as it gets heavily skewed by the large numbers.

Inspecting part 3, we can see the DB index for the standardized PCA is the lowest out of the rest, leading to the conclusion that it has the lowest error out of the previous tested methods. The figure 3 helps to very clearly depict the relationship between the three cultivars, showing that the reduction also maintained the variance shown within higher dimensions, giving a much better sense of relationships between cultivars in the true data. We can also inspect the score vectors in order to see what the data means. Cultivar 1, because it's positively correlated with score vector 2, and negatively correlated with score vector 1, looking at the difference between scores 2 and 1, we can get a sense of what cultivar 1 wine is like. Having a positive value in rows 1,3,5,6,7,9,10,12,13. In this data set we can conclude a lot of information about cultivar 1’s wine. Having increased, Ethanol, Ash, Magnesium, Total Phenols, Flavonoids, Proanthocyanidins, Colour Intensity, OD280/OD315 Ratio, and Proline proportional to the average. This data can now be used to inspect the other cultivars to see how their wine is like. This type of analysis is easiest on the standardized PCA compared to the rest. This leads to a very effective dimensionality reduction, being able to depict most of the important information from the two dimensions. It also allows for the natural biases of the data to be displayed correctly. A standardized principal component analysis using singular value decomposition is the best way to reduce dimensionality of this dataset while sustaining the variability of the data.