Investigating Properties of Wine from a 13-Dimensional Data Set
Intelligent Data Analysis

J.D. Preece

30th March 2016

School of Computer Science
University of Birmingham
Birmingham, B15 2TT

Abstract

The purpose of these experiments was to utilise dimension-reducing techniques such as principal component analysis (PCA) and coordinate-projection to explore relationships between varying attributes of wine. It was found that some of these attributes didn't correlate with the rest of the data (such as total phenols), though when reduced enough, certain attributes were found to hold relationships.

Acknowledgements

I would like to dedicate this report to a certain Amelia Elgey, whose passion and knowledge for fine wines has inspired me to explore what makes it such a delightful drink. May her supply of Echo Falls be endless unto the future.

I would also like to thank Peter Tino for providing many useful resources, as well as the online community for helping me through this project.

Contents

1 Introduction

Wine, an alcoholic beverage made by fermenting grapes (or other fruits on occasion), has been produced for thousands of years. [1] It is a beverage enjoyed by various demographics; from revellers to professional wine tasters, it is a go-to drink for many. However, to those drinking it, though the accumulation of the properties is apparent, the individual properties are rarely considered. It was the purpose of this project to investigate those individual properties, and to try to discover what gives different varieties of wine their colours, flavours, and alcohol content.

2 Results

A variety of MATLAB functions were designed to help with the analysis, as well as the MATLAB workspace. This meant importing large data sets was easy, and that it was easy to manipulate and use them.

2.1 Entire Data Set

Firstly, the entire data set was analysed, to establish which attributes would be of interest. This meant preprocessing the data for further use, and utilising methods of PCA to generate projections to analyse.

2.1.1 Preprocessing

It was important to judge the scale of the measurements in the data, to establish whether standardising the data was necessary. To do this, a box plot was made of the raw data, which can be seen in Figure 1. As is seen, each attribute contributes a different unit to the data - especially that of proline and magnesium. It was therefore necessary to standardise the data.

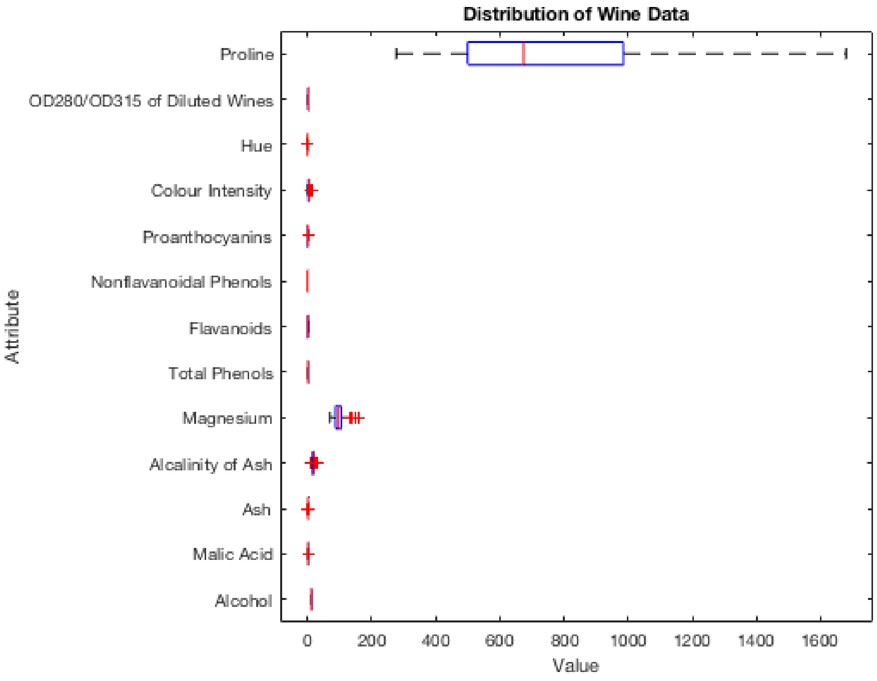

Figure 1: A box plot showing the range of values of the raw data

Standardisation can be approached in numerous ways. For this data set, MATLAB's built in `zscore` function was used, which centred the values to have mean 0 and scaled to have standard deviation 1. [2] After this standardisation was performed, a new box plot was generated, and can be seen in Figure 2. As can be seen,

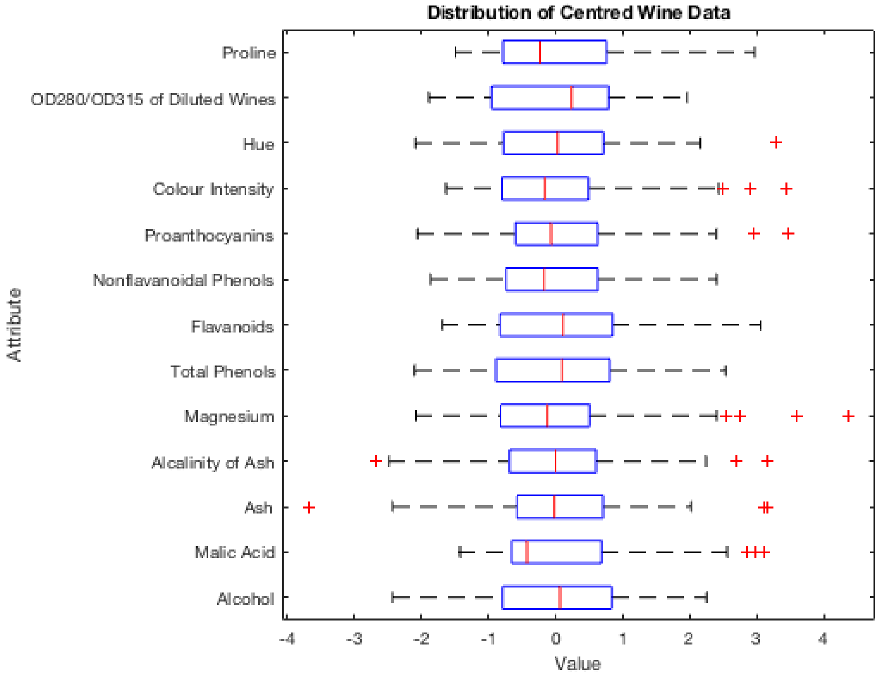

Figure 2: A box plot showing the range of values of the standardised data

the data shared a scale, meaning plots made after this wouldn't be skewed due to certain attributes.

2.1.2 Labelling

The original data set came labelled, with class numbers between 1 and 3 being assigned to each wine. Therefore, it was not necessary to establish a new labelling system for the investigation into the entire data set.

2.1.3 Principal Component Analysis

After the data had been standardised, MATLAB's `pca` function was used to generate the eigenvalues, eigenvectors, and coefficients to the eigenvalues. Before the results from the PCA function were projected, a scree plot of the percent variability was plotted to see which principal components would be of use to the PCA projection. This scree plot can be seen in Figure 3. This shows that the first and second principal components make up approximately 55% of the total variance, which makes these two components ideal for the projection. It was also considered that a 3D plot with the third principal component would also have been informative, as that brings it up to around 66% of the total variance, which is a majority. However, for ease, just the first two were plotted first.

Once the main principal components were established, a 2D projection was generated, which can be seen in Figure 4. This is a good result, as clusters of similar classes are easily present. It may be improved further by taking into account the first three principal components, though not by a great deal. Subsequent to the PCA projection, a biplot was plotted to visualise both the coefficients of the eigenvalues for each attribute and the principal component scores for each data point. It can be seen in Figure 5. The direction and length of the vectors produced indicates the contribution to the principal components from each attribute. The attributes to the right of the y-axis had the largest contribution to the first principle component, whilst those above the x-axis had larger contributions towards the second principle component. Furthermore, vectors that are close to each other provide relationships. For example, it proves that total phenols is strongly related to proanthocyanins and flavinoids, which are what make up total phenols. The nonflavanoidal phenols seem to contribute to the alkalinity of ash, and vice versa. These were tested through some coordinate projections.

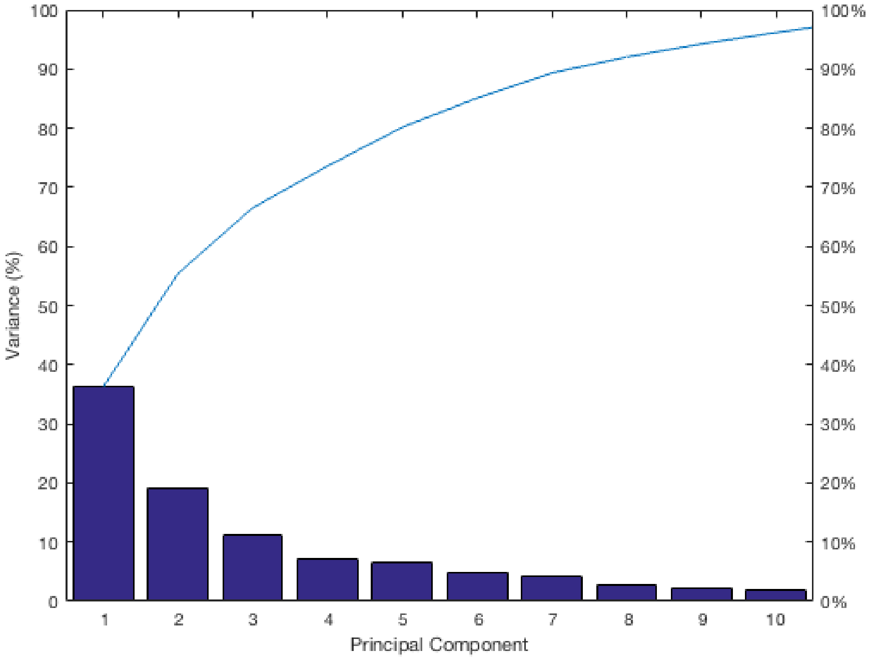

Figure 3: A scree plot showing the percentage variance of each principle component, and a running cumulative total

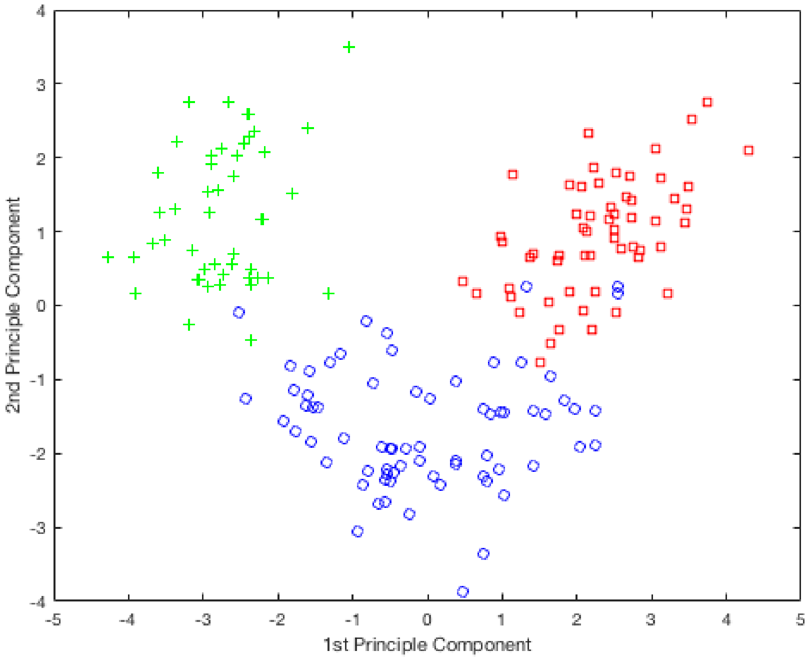

Figure 4: A 2D projection of the PCA results. Class 1 is red, class 2 is blue, and class 3 is green.

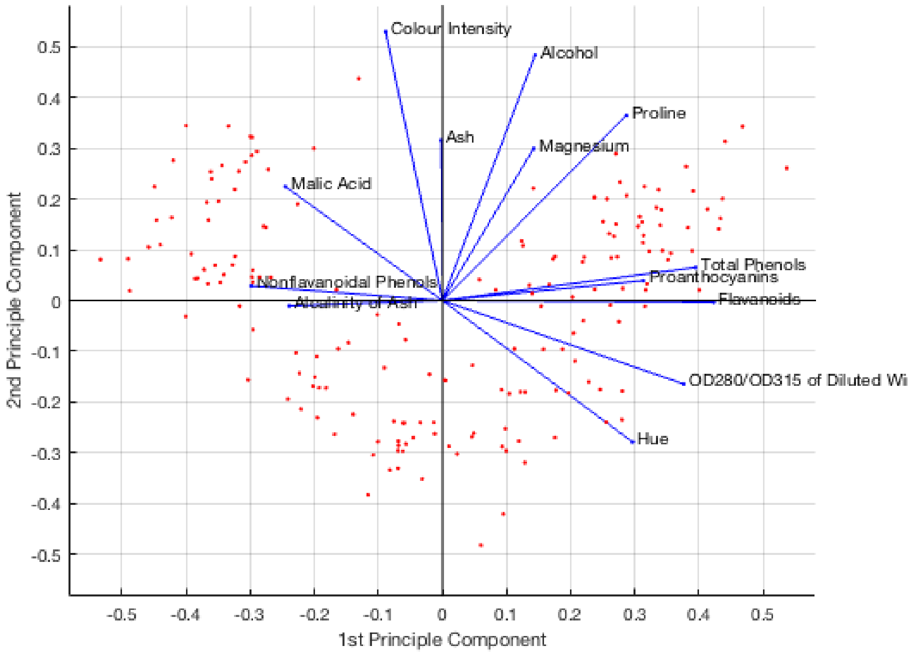

Figure 5: A biplot visualising the coefficients of the eigenvalues with the scores for each data point

2.1.4 Coordinate Projections

As mentioned, it appeared that total phenols was related to proanthocyanins and flavinoids. A coordinate projection of two attributes at a time was used to test this hypothesis. The first test was between total phenols and flavinoids, illustrated in Figure 6. This figure clearly shows that a class of wine can be distinguished by these attributes. This is also true for tests between flavinoids and proanthocyanins. However, when testing for the expected relationship between nonflavinoidal phenols and alkalinity of ash, the result is messy, and of no use, as seen in Figure 7. After further testing, such as that of alcohol and total phenols in Figure 8, it was determined that any attributes with a similar vector length in the PCA visualisation would lead to a good estimate of class when projected together.

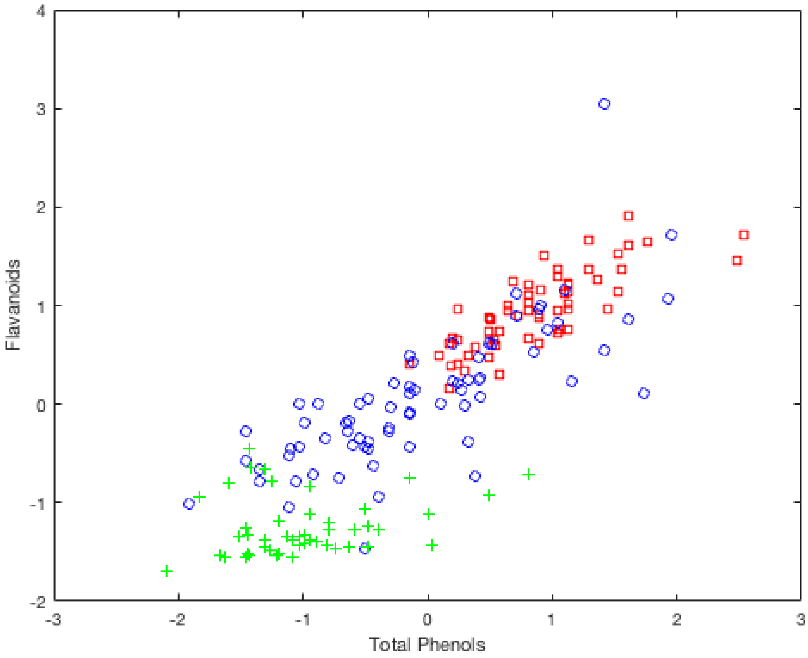

Figure 6: A coordinate projection between total phenols and flavinoids

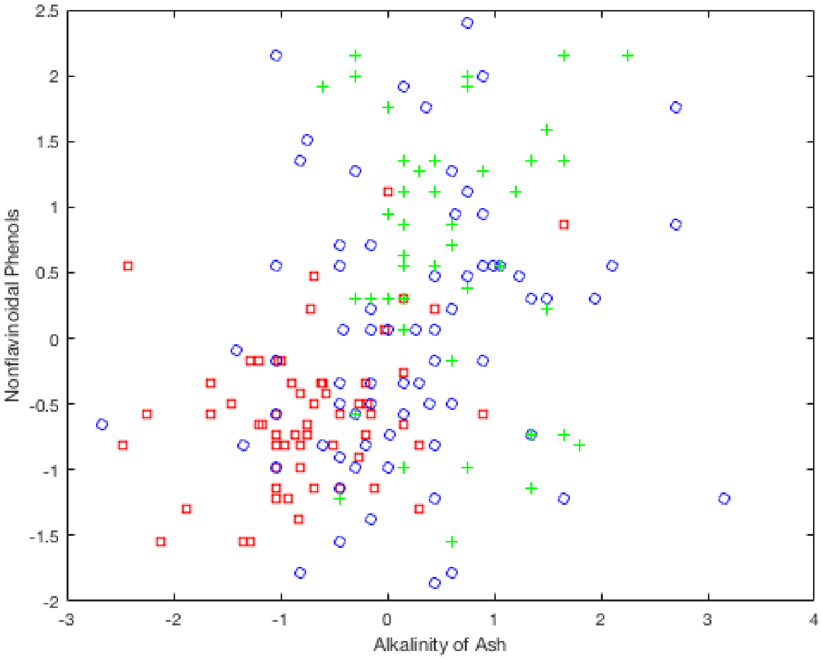

Figure 7: A coordinate projection between nonflavinoidal phenols and alkalinity of ash.

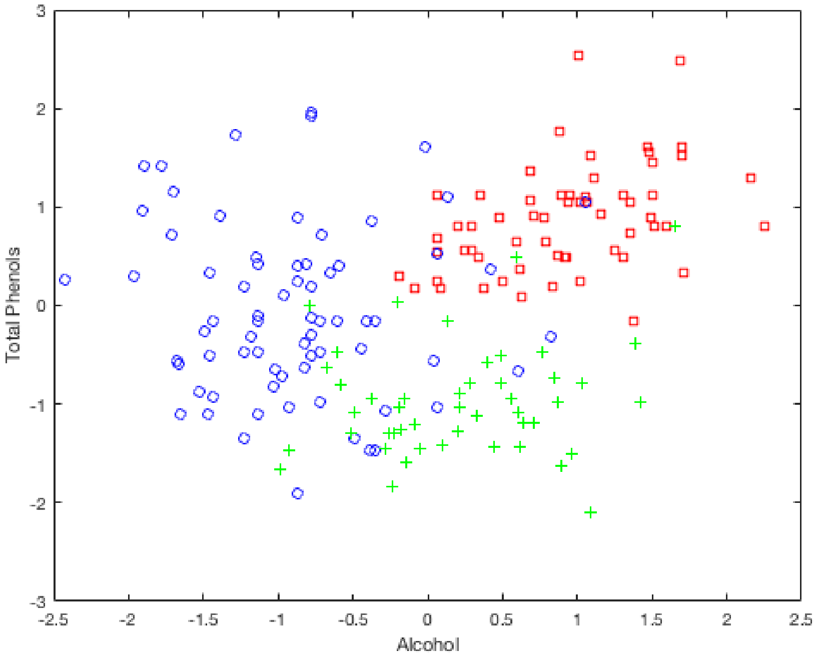

Figure 8: A coordinate projection between alcohol and total phenols.

2.2 Alcohol

Referring to Figure 5, it was hypothesised that any two attributes with similar vector magnitude, no matter what direction, would help determine a class for that wine. To test this, a new labelling scheme for the alcohol of each wine was made.

2.2.1 Preprocessing

The data was imported, and the alcohol attribute removed and placed as its own variable. As before, the main data set was standardised, to provide a uniform scale across all attributes.

2.2.2 Labelling

Various sources were used to establish an appropriate labelling system for the alcoholic content of wine, which can be seen in Table 1. [3][4] After these labels were established, the alcohol data was iterated over to determine the label of each data point, and assign it a value from 1 to 4, as shown by the Value column in Table 1. This can be seen visually scross the data in Figure 9.These values were stored as a local variable for later use.

Table 1: Labels for Alcohol Content of Wines

Label	Percentage p	Value
Low	$p < 12.5$	1
Average	$12.5 \leq p < 13.5$	2
High	$13.5 \leq p < 14.5$	3
Very High	$p > 14.5$	4

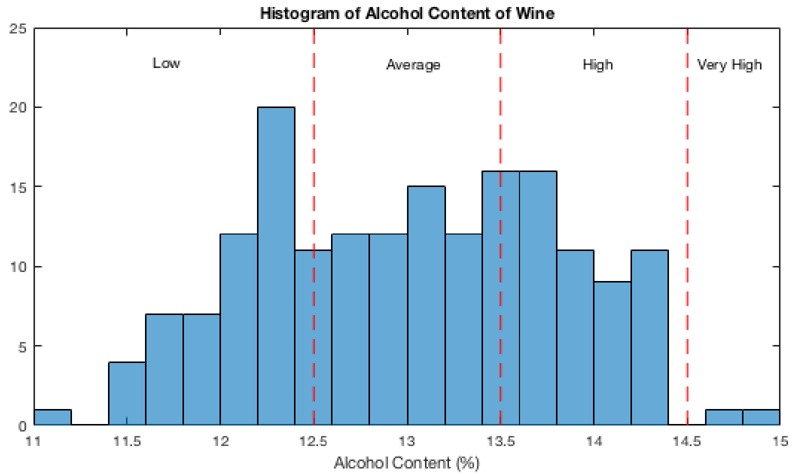

Figure 9: A histogram showing the labelling scheme of alcohol level

2.2.3 Principle Component Analysis

As previously done, a scree plot of the variance percentages of each principal component was plotted, to determine the key components. This can be seen in Figure 10. The majority of the contribution comes from

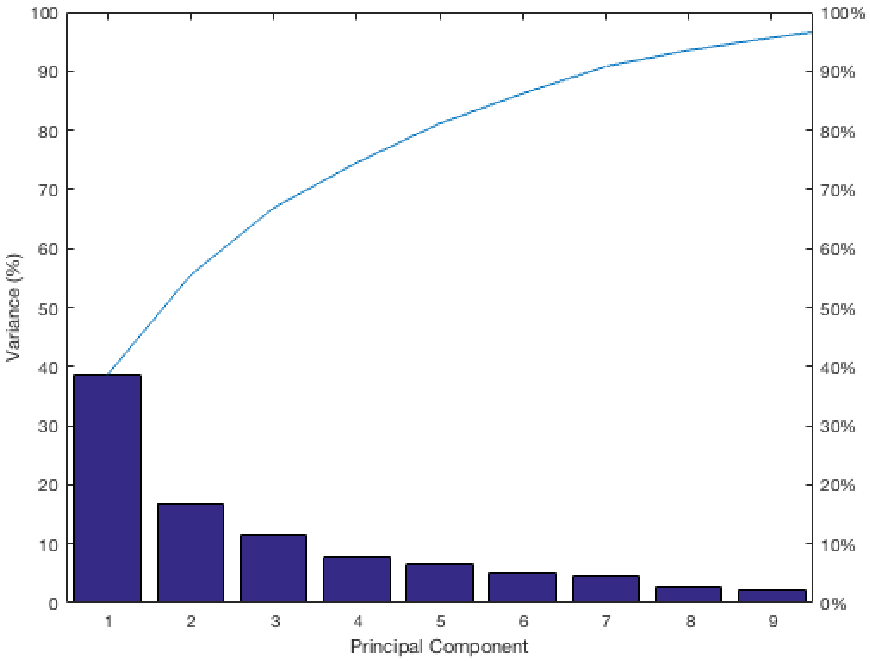

Figure 10: A scree plot showing the percentage variance of each principle component, and a running cumulative total.

the first component, though components two and three are also important, due to their percentage values. A projection was plotted with the first two principle components, illustrated in Figure 11. The plot is not very informative, as the majority of classes are spread, and each class overlaps. This being said, the high alcohol content class (green) provides the largest cluster. A biplot was plotted, as seen in Figure 12. From this, it can be seen that many of the attributes share a similar length. This indicates that many attributes contribute to

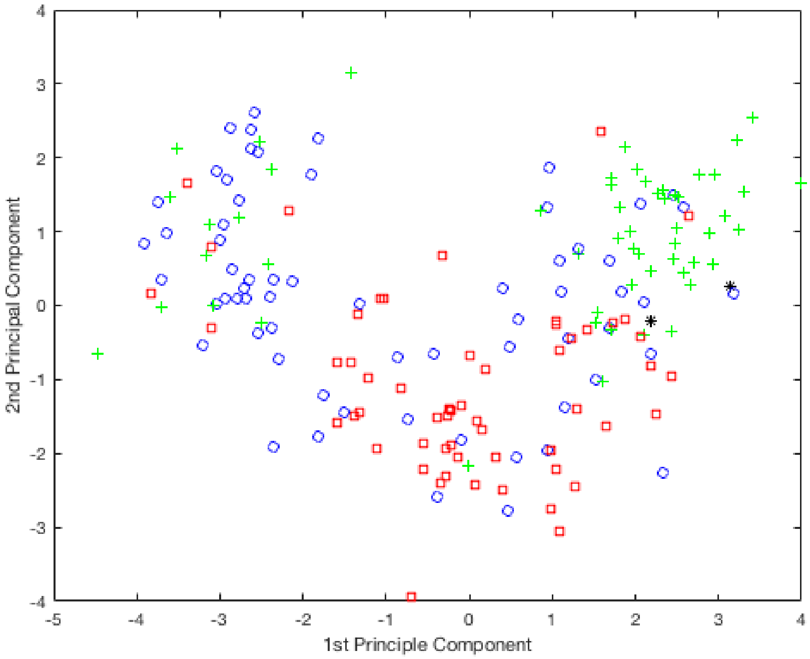

Figure 11: A 2D projection of the PCA results. Low is red, average is blue, high is green, and very high is black.

the alcohol of a wine.

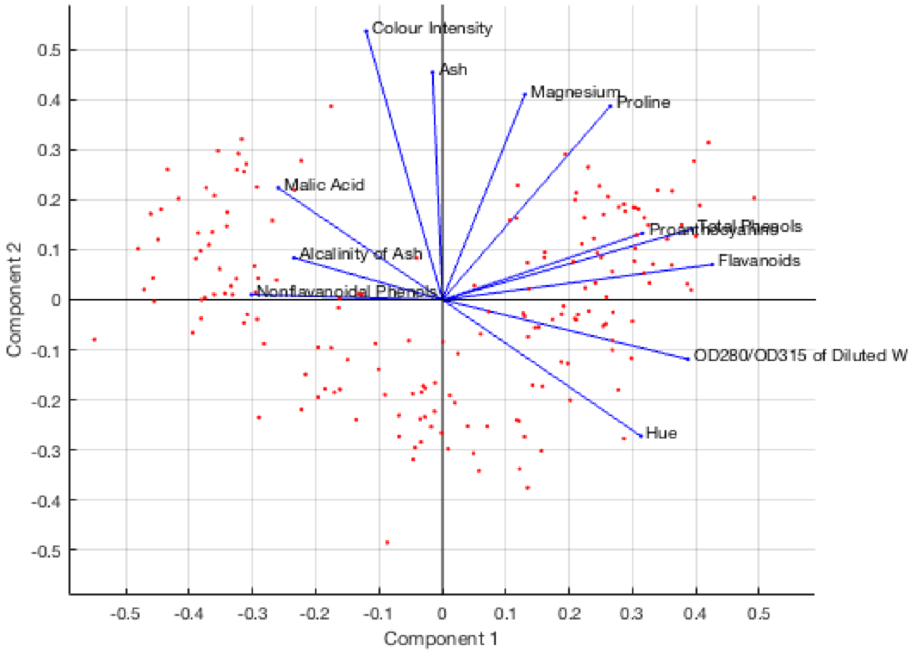

Figure 12: A biplot visualising the coefficients of the eigenvalues with the scores for each data point

2.2.4 Coordinate Projections

Again, the coordinates between total phenols and flavinoids were tested, due to their close proximity. The results, as seen in Figure 13, show a certain level of order, though due to the nature of alcohol being dependant on so many attributes, it still is fairly clustered. Because of the clustering, no more coordinate projections were made.

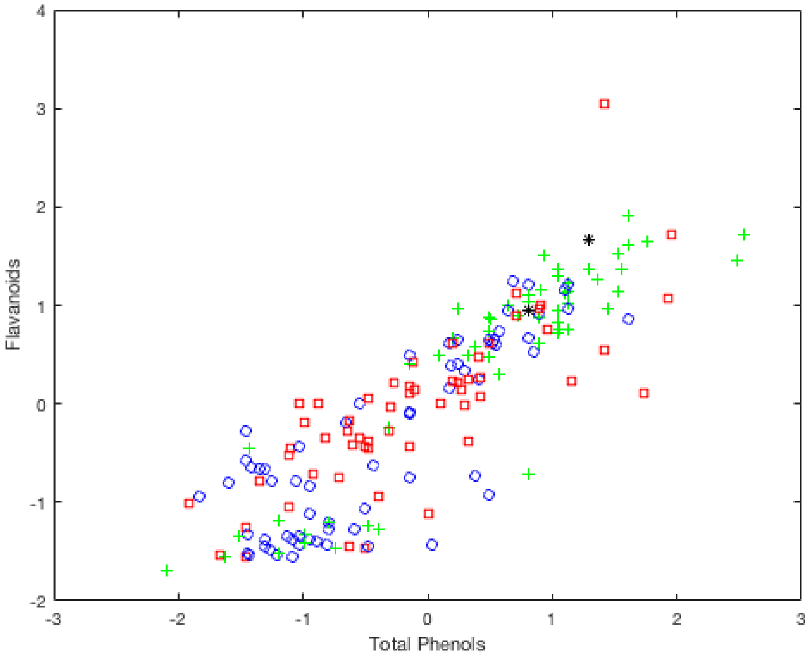

Figure 13: A coordinate projection between total phenols and flavinoids.

2.3 Total Phenols

As total phenols had been used for coordinate projections in both previous examples, it was decided it would be determined if there was anything causing these levels specifically.

2.3.1 Preprocessing

Preprocessing was identical to that of alcohol. The data was imported, and the total phenols attribute removed and placed as its own variable. As before, the main data set was standardised, to provide a uniform scale across all attributes.

2.3.2 Labelling

An arbitrary labelling scheme was set up, which was mathematically calculated. The data's minimum and maximum values were found, and four regions were created in between those values. Data points were then labelled as to which category they fell into. The histogram of this labelling can be seen in 14.

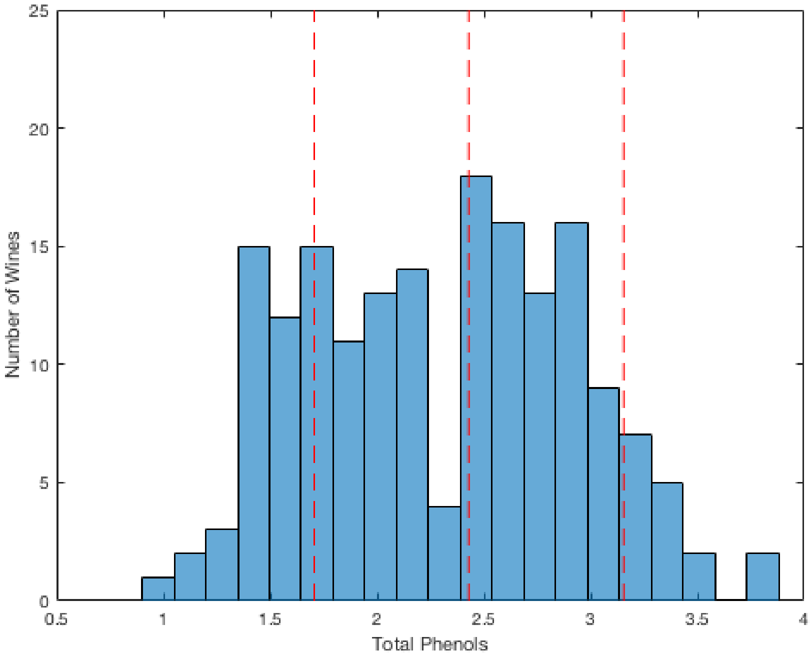

Figure 14: A histogram showing the labelling scheme of total phenols.

2.3.3 Principal Component Analysis

A scree plot was plotted to determine the adequate number of principle components, as seen in Figure 15. Once again, the first two principle componants were used, though it is worth mentioning that results would improve with a three-dimensional plot again. The results were then projected, as seen in Figure 16. These results are very scattered, and even the biplot of Figure 17 does not help determine much about the plot, other than the total phenols are not strongly correlated with the total of the remaining data.

2.3.4 Coordinate Projections

After going through various coordinates, not many provided clear results. However, two examples that came close were that of magnesium and flavinoids, and (once again) flavinoids and proanthocyanins. These can be

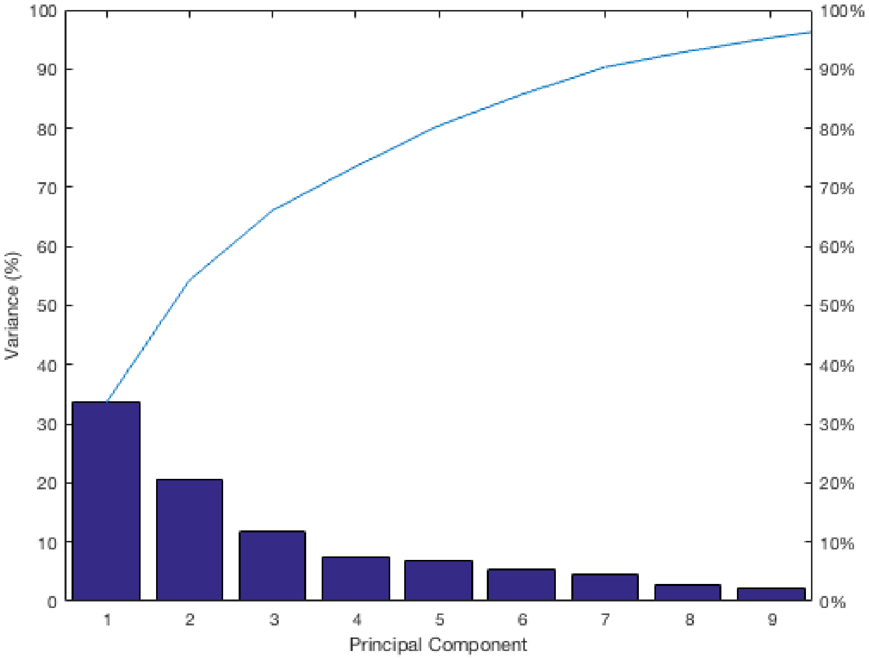

Figure 15: A scree plot showing the percentage variance of each principle component, and a running cumulative total.

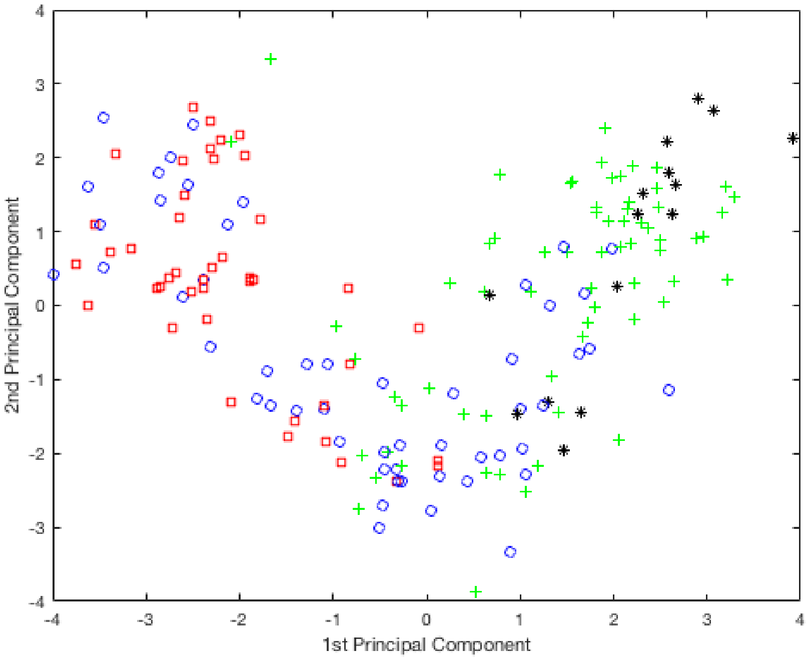

Figure 16: A 2D projection of the PCA results. Low is red, average is blue, high is green, and very high is black.

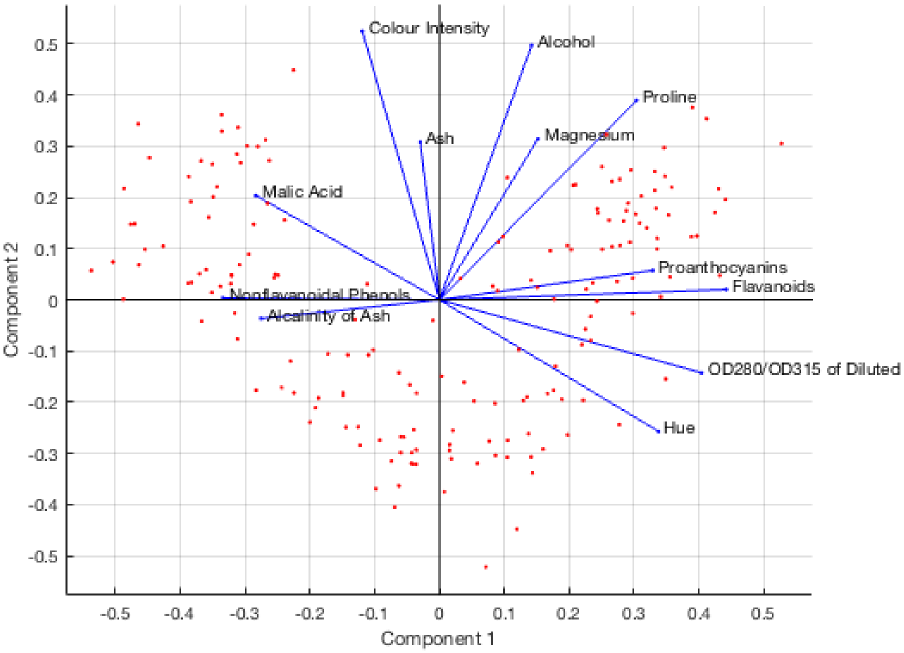

Figure 17: A biplot visualising the coefficients of the eigenvalues with the scores for each data point

seen in Figures 18 and 19 respectively. This suggests that these attributes may help when classing a wine's total phenols, though it may be uncertain due to classes overlapping.

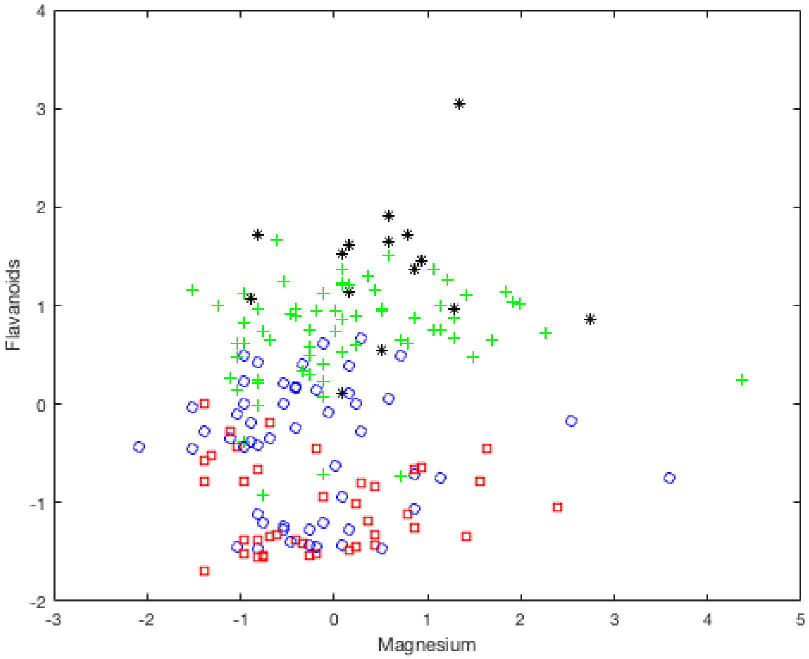

Figure 18: A coordinate projection between magnesium and flavinoids.

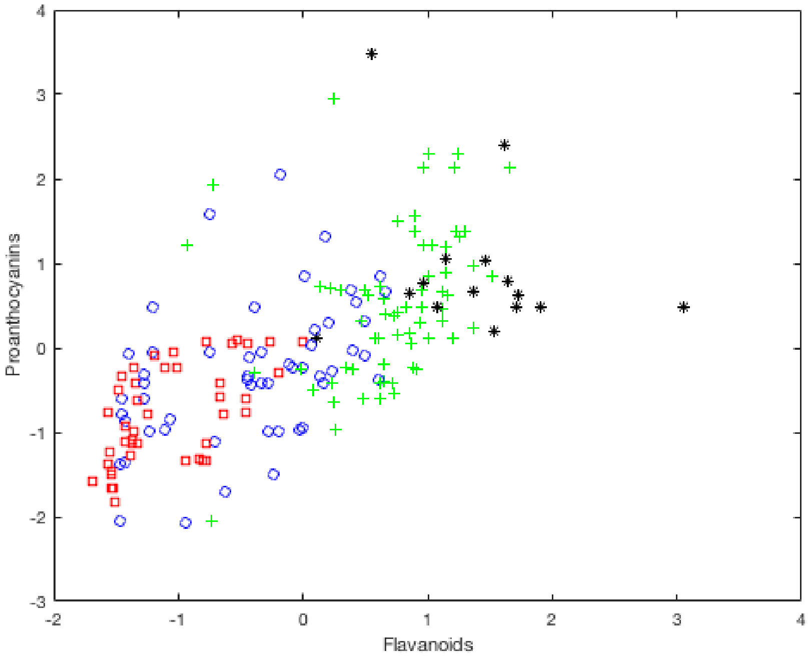

Figure 19: A coordinate projection between proanthocyanins and flavinoids.

3 Conclusion

It was necessary to standardise the data set being used, so to scale everything for use in PCA and coordinate-projection. The resulting linear predictive models are adequate at predicting labels, though the results of these experiments could be improved by introducing a third principle component during PCA, or even using different techniques completely. A coordinate projection of total phenols and flavinoids gave the clearest visualisation, and would make it easiest to predict which class a wine was from. Some coordinate projections made valid suggestions at how a wine could be classed for both alcohol and total phenols, though it mostly seemed that it was down to the flavinoids and proanthocyanins. However, with further research, and a deeper look at all the possible projects against all attributes, there may be more to find.

References

[1] H. Johnson. *Vintage: The Story of Wine*. Simon & Schuster, 1992.

[2] MathWorks. *zscore*. URL: `http://uk.mathworks.com/help/stats/zscore.html`.

[3] Wine Companion. *Alcohol Content in Red Wines*. URL: `http://www.winecompanion.com.au/wine-essentials/wine-education/alcohol-content-in-red-wines`.

[4] Real Simple. *A Guide to the Alcohol Content in Wine*. URL: `http://www.realsimple.com/holidays-entertaining/entertaining/food-drink/alcohol-content-wine`.