Combing through the Higher Dimensions of Wine

Christoph Stich

April 7, 2016

1 Preamble

I have chosen to use the WINE dataset for this assignment. The source code and all derived data are attached to this assignment as a pdf of my Jupyter Notebook.

2 Data processing

The datasets consists of 13 continuous variables representing various chemical and physical properties of wine. While all wines in the dataset were grown in the same region, the datasets also includes one class variable differentiating cultivars (see figure 1 for a histogram of the class variable).

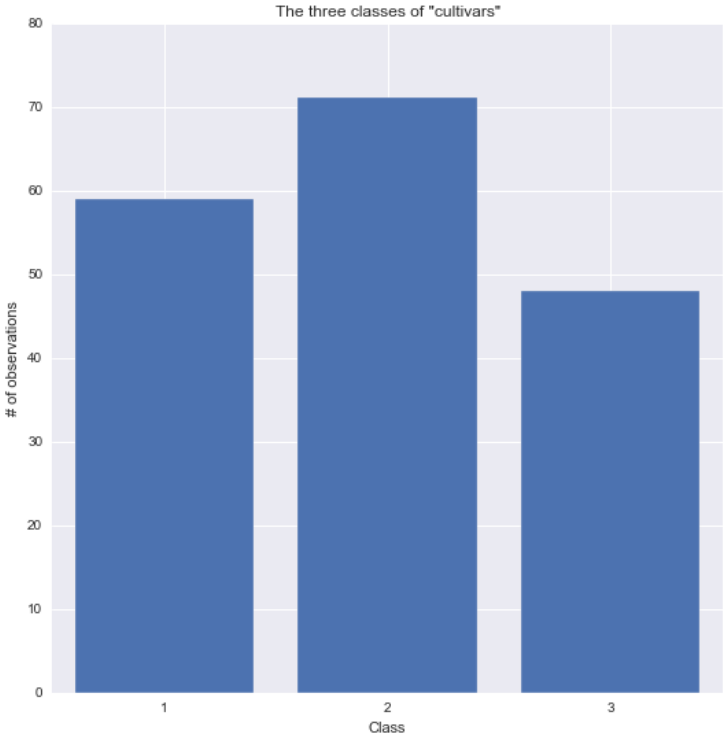
The data is read in and stored as the matrix X_{raw}, where rows represent individual samples and columns the 13 different attributes. As the 13 variables in the dataset were given in unknown units and furthermore differ wildly in their magnitude, I standardize each variable. For each attribute column vector $\vec{x_a}$ of X_{raw}, I calculate its mean $\overline{x}_a$ and standard deviation s_a, before applying$z = \frac{x-\overline{x}_a}{s_a}$ to all points $x \in \vec{x}_a$, yielding a new standardized feature vector $\vec{z}_a$. I then use all $\vec{z}_a$ as columns to construct the matrix X. I furthermore check whether X contains any missing data or duplicates, but cannot find either.

3 Research questions

For my assignment I tried to answer the following three research questions:

1. Does the chemical and physical composition vary between the three different cultivars and what are the most important dimensions?
2. Do wines with different levels of alcohol differ from one another? And is the level of alcohol related to the different cultivars?
3. What can we learn by looking at the color (i.e. hue and color intensity) of wine?

Figure 1: Histogram “cultivars”



4 Cultivars

The first question I was interested in was whether different cultivars also have a distinct composition of features and what are the most "important axis" for describing the data succinctly. As the dataset already contains a dedicated class variable representing the three different cultivars, I decided to use those labels for the experiment.

When trying to visualize high-dimensional data, one faces the problem that one cannot visualize 13 dimensions as in this dataset simultaneously nor can one arbitrarily project high dimensional data into a low dimensional space. Compare the two different projections of X into a 2 dimensional subspace in figure 2. Intuitively we could argue that the projection onto "matlic acid" and "hue" is a much better projection than the projection in figure 1 as we can observe the following pattern: Wines with an above average hue and above average level of maltic acid are produced from cultivar 1, wines with a below average hue are cultivar 2, and wines with below average hue and below average level of maltic acid belong to cultivar 3.

However, manually going through all possible plots of possible sub-spaces is a very biased approach nor feasible for an even higher dimensional dataset. One possible solution is to project the data in the subspace that preserves the greatest amount of variance and this is exactly what a projection based on a *principal component analysis* does.

Figure 3 projects the 13 dimensional data onto a sub-space spanned by the first and second eigenvector of the co-variance matrix of X (hereafter denoted as *Covar(X)*. While there does not exist a "perfect" boundary between the three different cultivars, nevertheless three distinct clusters emerge. This is a clear sign that we can separate the three cultivars based on the data available in X and that the three types of cultivars on average are different in their physical and chemical composition.

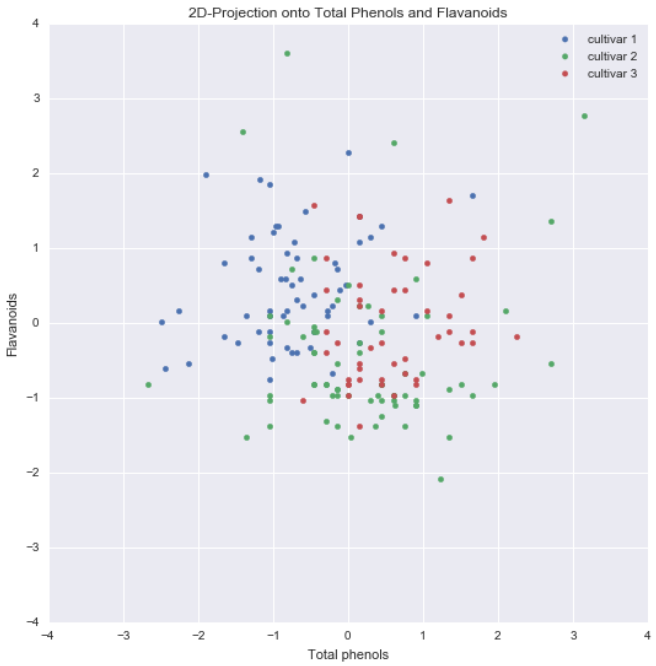
When looking at the eigenvectors and eigenvalues of *Covar(X)*, I can see that we need only the first six eigenvectors to explain 90% of the variance of the data (see figure 3.b). We can also observe that a few eigenvalues are relatively large and most are rather small (see figure 3.a), meaning that a projection onto the first few eigenvectors preserves a lot of variance. In fact the 2D eigenvector projection in figure 3 alone preserves 55% of all variance. Furthermore, the two top eigenvectors each preserve different dimensions of the data as I find a Pearson correlation of -0.39 for the absolute value of the elements of the two eigenvectors.

A more detailed analysis of the first eigenvector (see table 1) reveals that five out of the seven dimensions with the greatest magnitude are related to environmental stressors for plants. Plants produce flavanoids "growth and development, attraction of pollinator animals, nitrogen-fixation in leguminous plants, and for protection against damage by herbivores, microbes, UV and reactive oxygen species"[1]. Plants synthesize phenols as a response to ecological pressures such

[1] Miranda, C L; et al. (2012). "Flavonoids" eLS

Figure 2: 2D Projections

(a) Projection 1



(b) Projection 2

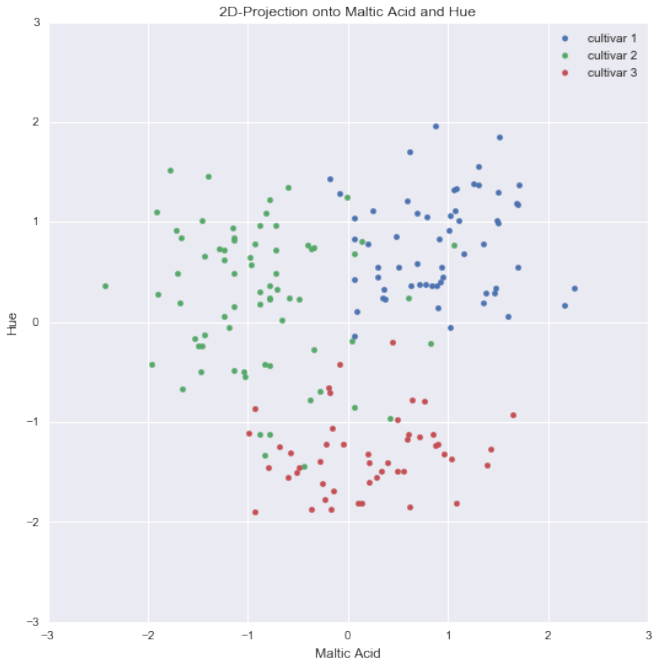


Figure 3: 2D PCA projection of X

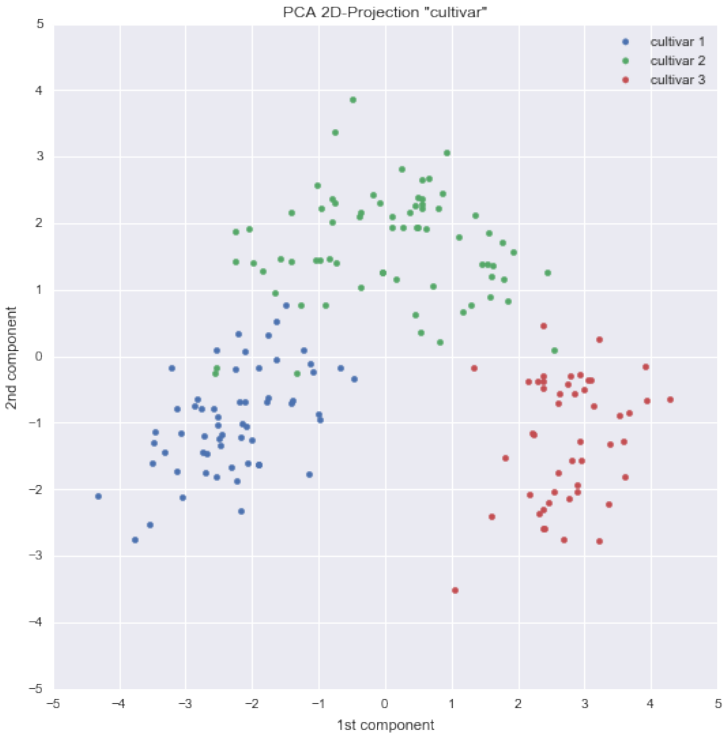
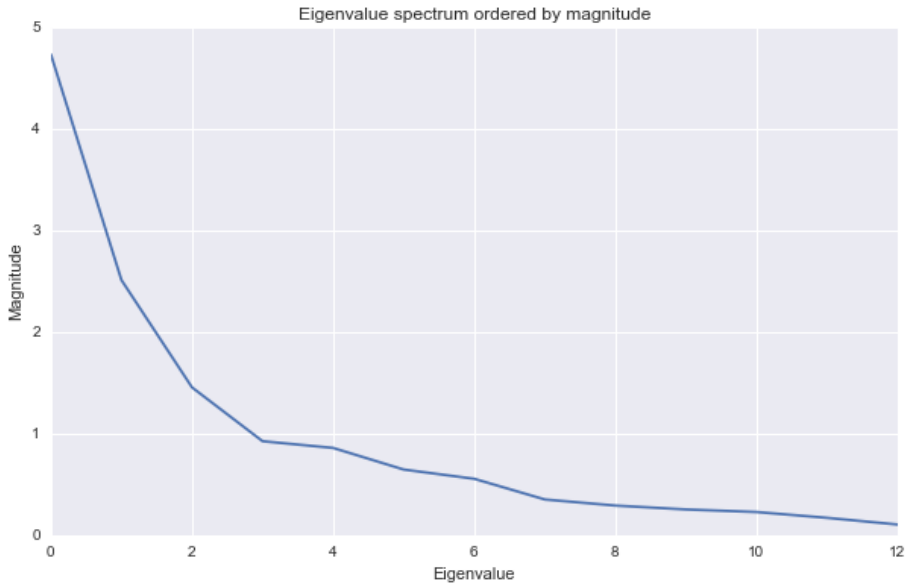


Table 1: Top two eigenvectors of *Covar(X)*

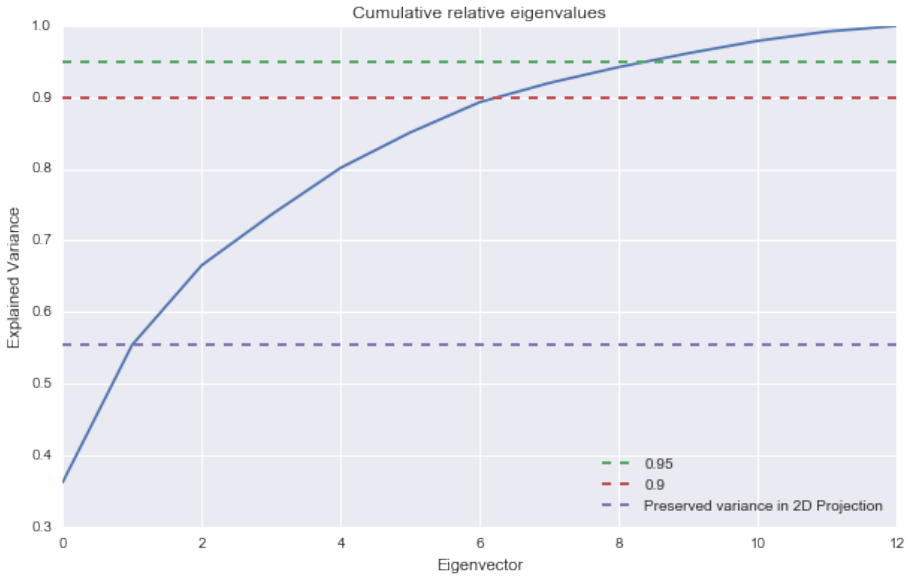
"Environmental Stressors"		"Quality"	
Eigenvalue	4.7324	Eigenvalue	2.5110
Dimensions		Dimensions	
Flavanoids	-0.4229	Color intensity	-0.5300
Total phenols	-0.3947	Alcohol	-0.4837
OD280/OD315	-0.3762	Proline C	-0.3649
Proanthocyanidins	-0.3134	Ash	-0.3160
Nonflavanoid phenols	0.2985	Magnesium	-0.3000
Hue	-0.2967	Hue	0.2792
Proline C	-0.2868	Malic acid	-0.2249
Malic acid	0.2452	OD280/OD315	0.1645
Alcalinity of ash	0.2393	Total phenols	-0.0650
Alcohol	-0.1443	Proanthocyanins	-0.0393
Magnesium	-0.1420	Nonflavoid phenols	-0.0288
Color intensity	0.0886	Alcalinity of ash	0.0105
Ash	0.0020	Flavanoids	0.0034

Figure 4: Eigenvalues and Eigenvectors of $Covar(X)$

(a) Eigenvalue spectrum



(b) Cumulative eigenvalues



as pathogens, insect attacks, UV radiation and wounding[2]. Similarily, plants defend themselves against pathogens and predators with the help of proanthocyanidins[3]. Resveratrol on the other hand is produced in response to injury or when the plant is threatened by pathogens[4]. Last but not least proline is positivly correlated with stress for a plant[5].

The first eigenvector is thus most likely related to how different cultivars deal with environmental stressors. Each wine made from a specific cultivar exhibits its own range of values for this "environmental stressor" eigenvector. Cultivar 1 wines have the lowest, cultivar 2 wines the medium, and cultivar 3 wines the highest value of the "environmental stressor" eigenvector (see figure 3).

For the second eigenvector the two biggest dimensions are color intensity and alcohol (table 2). Alcohol levels are directly related to how much sugar can be extracted from the grapes during mashing as yeast primarily ferments sugar into alcohol. Further, the color (represented by "hue" and "color intensity" in this dataset) of a wine is highly dependent on the amount of anthocyanins, which are created during the ripening process and are positively correlated with the amount of sugar[6]. Whereas, both ash and magnesium are used as important criteria for chemically evaluating the quality of wine[7]. The second eigenvector is thus clearly linked to the amount of sugar present in the wort and thus to the ripeness of the grapes, but also to chemical compounds that function as a measure of quality for wine. The second eigenvector seems to assess the characteristics—the "quality"—of the wines in the sample.

Observing the values for the second eigenvector I can spot a dichotomy between wines with a high "quality" value (cultivar 2 wines) and wines with an average or below average "quality" level (cultivar 1 and 3 wines) in figure 3. On a "quality" level this makes cultivar 1 and cultivar 2 wines indistinguishable. It would be interesting to see whether the "quality" eigenvector that I found, is related to taste at all, but more data is needed to answer that question.

5 Alcohol level

Another question I was interested in was how wines with different level of alcohol relate to each other and if the level of alcohol in a wine is cultivar specific. For the second experiment, I thus created labels based on the relative alcohol content. The distribution of values for alcohol more or less follows a bell curve and I thus

[2]Klepacka, J; et al. (2011). "Phenolic Compounds as Cultivar- and Variety-distinguishing Factors in Some Plant Products". Plant Foods Hum Nutr. 66 (1): 64–69

[3]Amil-Ruiz, F.; et al. (2011). "The Strawberry Plant Defense Mechanism: A Molecular Review". Plant and Cell Physiology 52 (11): 1873–903. doi:10.1093/pcp/pcr136. PMID 21984602.

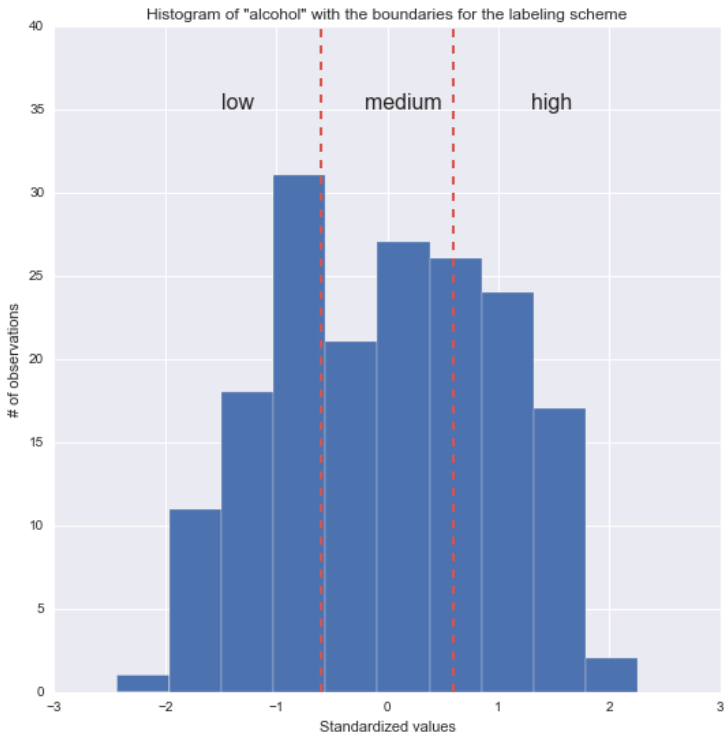
[4]Fremont, Lucie (2000). "Biological Effects of Resveratrol". Life Sciences 66: 663–673

[5]Hayat, S; et al. (2012) "A role of proline under changing environments". Plant Signaling & Behavior 7 (11): 1456-1466

[6]Robinson, J (2006) "The Oxford Companion to Wine" Third Edition, Oxford University Press, Oxford

[7]Pascal, R; et al (2006) "Handbook of Enology, The Chemistry of Wine: Stabilization and Treatments " Second Edition, John Wiley & Sons, Chichester

Figure 5: Histogram “alcohol”



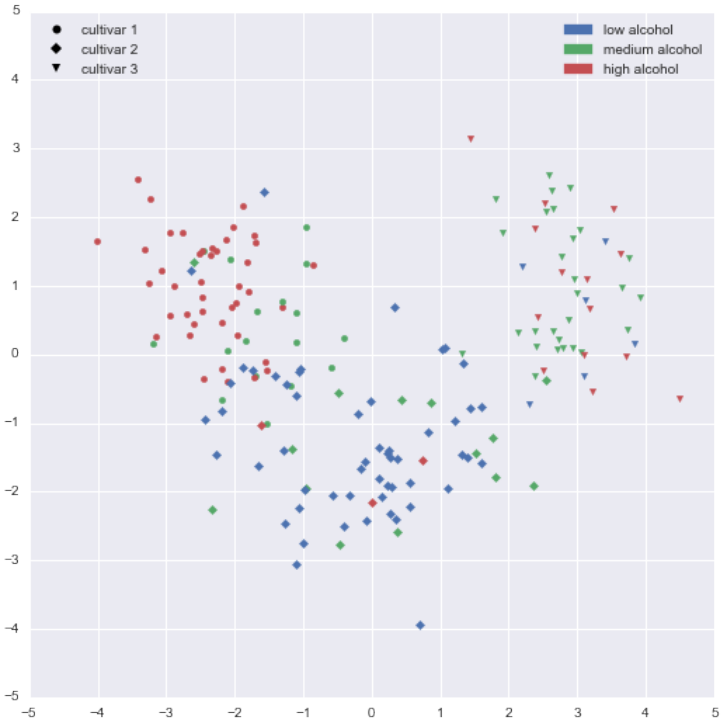
opted to split the data into three equally large groups (see figure 5). The one third of samples with the lowest value were assigned the label "low alcohol", the next third the label "medium alcohol" and the last third "high alcohol"[8].

Let $X_{-alcohol}$ be the matrix X without the column “alcohol”. Figure 6 shows the 2D-projection of $X_{-alcohol}$ onto the first and second eigenvector of $Covar(X_{-alcohol})$. In this case the clustering of our labels is much less pronounced than in figure 3. While the "low alcohol" and "high alcohol" points are relatively well separated and form relatively distinct clusters, points with the label "medium alcohol" are more spread out. In this instance two eigenvectors do not seem to be enough to separate the classes as clearly as before.

My assumption is that important independent variables for the alcohol content are missing from the dataset; neither the actual sugar level of the grapes used to make the wine nor the amount of anthocyanins (remember they are positively correlated with the level of sugar in the grapes) are reported in the data. And as yeast mainly utilizes sugar for fermentation, it is not surprising that the visualization based on an PCA does not reveal as highly distinct clusters as

[8] All bins that are based on percentiles are inclusive of the lower bound but exclusive of the upper bound; except for the last category which has an inclusive upper bound so that all values are assigned labels.

Figure 6: 2D PCA projection of $X_{-alcohol}$



before.

However, when examining the features “cultivar” and “alcohol” simultaneously, I can see that different cultivars usually correspond to a certain level of alcohol. Low alcohol wines are mostly comprised of cultivar 2 wines, medium alcohol wines of cultivar 3, and high alcohol wines of cultivar 1. And indeed it is quite plausibly that different types of grapes with different innate levels of sugar tend to produce wines with different levels of alcohol.

The eigenvalues of the co-variance matrix of $X_{-alcohol}$ show a similar pattern to the eigenvalues of X, both in magnitude as well as in their overall distribution (see appendix A). Thus, the interpretation of eigenvectors and eigenvalues mostly stays the same and indeed the projection onto the first and second eigenvector of $Covar(X_{-alcohol})$ is very similar to the projection based on the eigenvectors of $Covar(X)$; it looks as if mainly the signs of the principal component are flipped.

6 Color

Last but not least I am trying to answer the question of how much we can learn about a wine by just “looking” at it, or in other words what is the relationship between color and the chemical composition of wine. As color in this dataset is comprised of “hue” and “color intensity” I chose to include both as dependent variables in the third experiment.

Consequently, I created labels based on the relative hue as well as the relative color intensity. As the distribution of hue values follows more or less a bell curve, I decided to split the data into three equal groups. Analog to the labels for "alcohol" I labeled the lowest third of hue samples "low hue", the next third "medium hue", and the last third "high hue" (see figure 7). Unfortunately the dataset does not include a description of what colors the "hue" values map to and thus cannot be used to plot the actual colors.

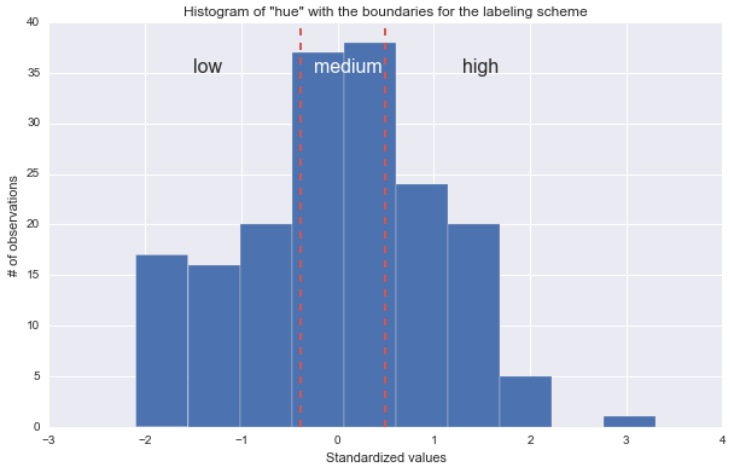
The distribution of values of color intensity however is skewed right with a relatively long tail. I thus decide to use four categories “low”, “medium”, "high", and "very high color intensity" to preserve more information about the tail. The category "low" comprises the first 30 percent and "medium" the next 30 percent. The remaining 40 percent are evenly split between the categories "high" and "very high" (see again figure 7).

Let X_{-hue} be the matrix X without the column “hue” and $X_{-color\ intensity}$ the matrix X without the column “color intensity”. Figure 8 depicts the projection of X_{-hue} onto the first and second eigenvector of the $Covar(X_{-hue})$. The most salient feature in the visualization is that wines with a low hue, form a distinct cluster. The other two categories are rather interspersed with each other, but together form their own cluster. There are clearly two groups of wine in this plot: wine with a low hue value and wine with a medium and high hue value.

What is more wines with a low hue content are dominantly cultivar 3, while wines with a medium or high hue are predominantly either cultivar 1 or 2.

Figure 7: Histograms “color”

(a) “Hue”



(b) “Color intensity”

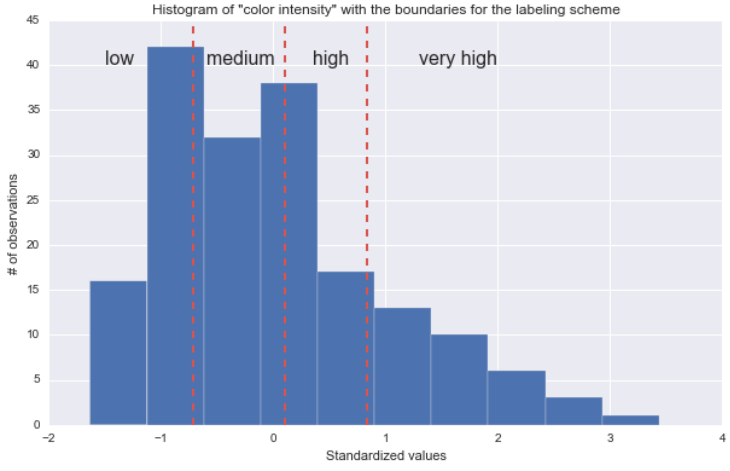


Figure 8: 2D PCA projection of X_{-hue}

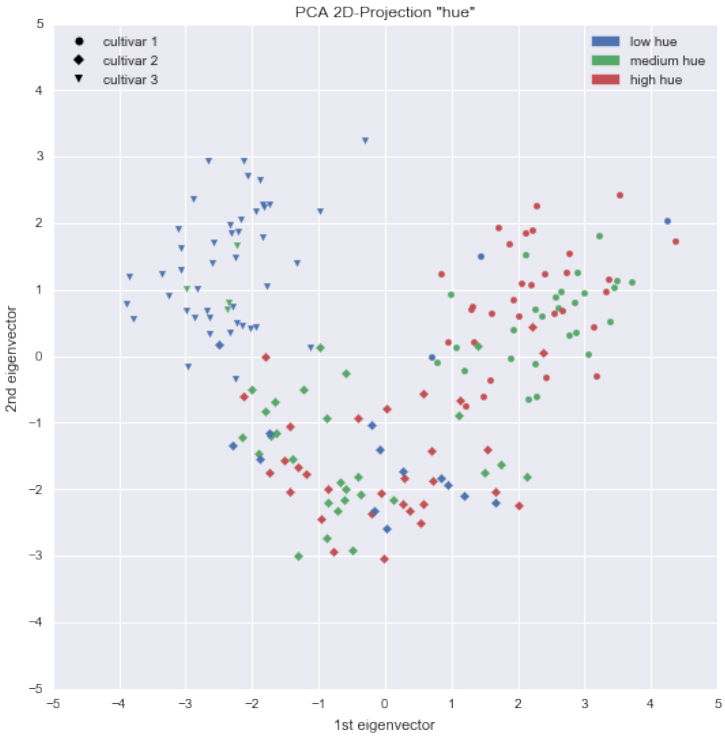
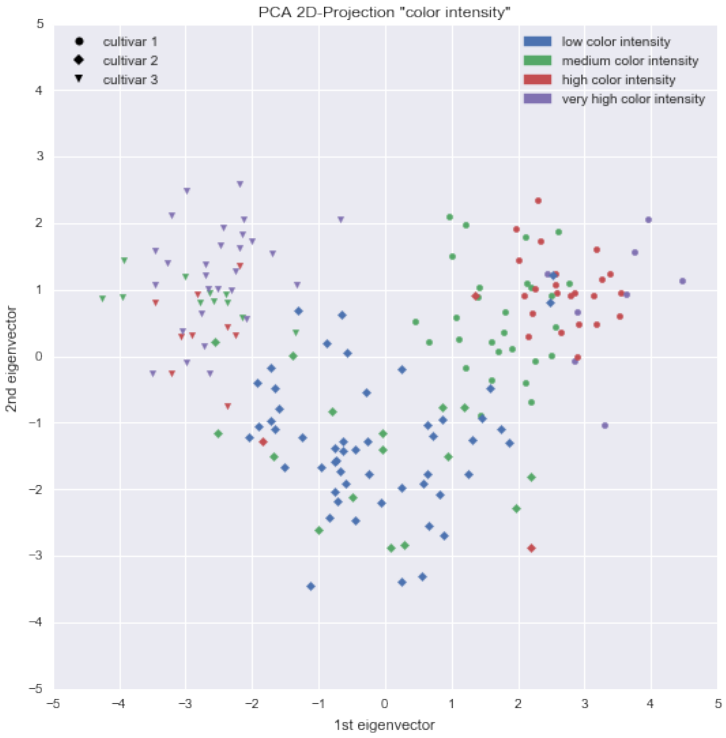


Figure 9: 2D PCA projection of $X_{-color\ intensity}$



(a) Including all cultivars

Potentially the two different clusters of hue values are related to the common wine dichotomy of red and white wines. However, without knowing the initial unit of measurement for hue, this is hard to know for sure.

The eigenvectors and the eigenvalues of $Covar(X_{-hue})$ are again very similar to the eigenvectors and eigenvalues of $Covar(X)$ (see appendix B) and thus reveal no new insights.

When projecting $X_{-color\ intensity}$ onto a 2D subspace spanned by the first and second eigenvector of $Covar(X_{-color\ intensity})$, I can identify a more or less distinct area for each level of color intensity (figure 9) First, points with a low color intensity have a considerable lower value for their second eigenvector than the other points. Second, points with a very high color intensity have a consequentially lower value for their first eigenvector. Third, points with a medium and high color intensity both have above average values for both their first and second eigenvector, although the points with a high color intensity tend to have a higher combined value for their eigenvectors.

Furthermore, wines with a low color intensity are mostly cultivar 2, wines with a medium and high color intensity mostly cultivar 1, and wines with a very high color intensity predominantly cultivar 3. Not surprisingly certain

color intensities are linked to certain cultivars.

It is however noteworthy that not all wines with similar color intensity are situated next to each other in the new subspace. The cluster of wines with a very high color intensity is much closer to the group of wines with a low color intensity than to wines with a high color intensity. For the other wines, color intensity progresses in a more linear fashion from low to high color intensity samples. What's more very high color intensity wines are mostly made up of cultivar 3 wines. This is further evidence that cultivar 3 wines are with respect to their color behaving quite differently than the other wines.

Looking at the eigenvectors and eigenvalues of $Covar(X_{-color\ intensity})$, I can see that for the most part they are once more very similar to the eigenvectors and eigenvalues of $Covar(X)$. However, as the most important dimension of the second eigenvector of $Covar(X)$ has now become a label and has thus been excluded from the decomposition, the values of the second eigenvector of $Covar(X_{-color\ intensity})$ change. The overall interpretation however that this eigenvector represents the characteristics, or quality, of a wine, still holds, as the top three dimensions are still related to the concept.

7 Conclusion

Overall different types of cultivar exhibit a distinct chemical and physical composition in at least three categories: First, the three cultivars are very well separated in a projection onto the first and second eigenvector of $Covar(X)$. Second, each type of cultivar also has a dominant type of alcohol level. Third, wines from cultivar 3 show a fundamentally distinct behavior with respect to hue and color intensity.

Furthermore, I find that the first two eigenvectors alone are able to explain 0.55 of the variance of X. Where the first eigenvector largely corresponds to how the wine plant chemically deals with "environmental stressors" and the second eigenvector describes the innate "qualities" of the fermented wine.

A Appendix for $X_{-alcohol}$

Table 2: Eigenvectors of $Covar(X_{-alcohol})$

Eigenvector 1		Eigenvector 2	
Eigenvalue	4.6597	Eigenvalue	2.0372
Dimension		Dimension	
Flavanoids	-0.42622	Color intensity	0.5364
Total phenols	-0.3940	Ash	0.4545
OD280/OD315	-0.3874	Magnesium	0.4100
Proanthocyanins	-0.31560	Proline C	0.3871
Hue	-0.3131	Hue	-0.2721
Nonflavanoid phenols	0.3016	Malic acid	0.2236
Proline C	-0.2650	Total phenols	0.1411
Malic acid	0.2593	Proanthocyanins	0.1331
Alcalinity of ash	0.2346	OD280/OD315	-0.1185
Magnesium	-0.1300	Alcalinity of ash	0.0839
Color intensity	0.1205	Flavanoids	0.0704
Ash	0.0155	Nonflavanoid phenols	0.0103

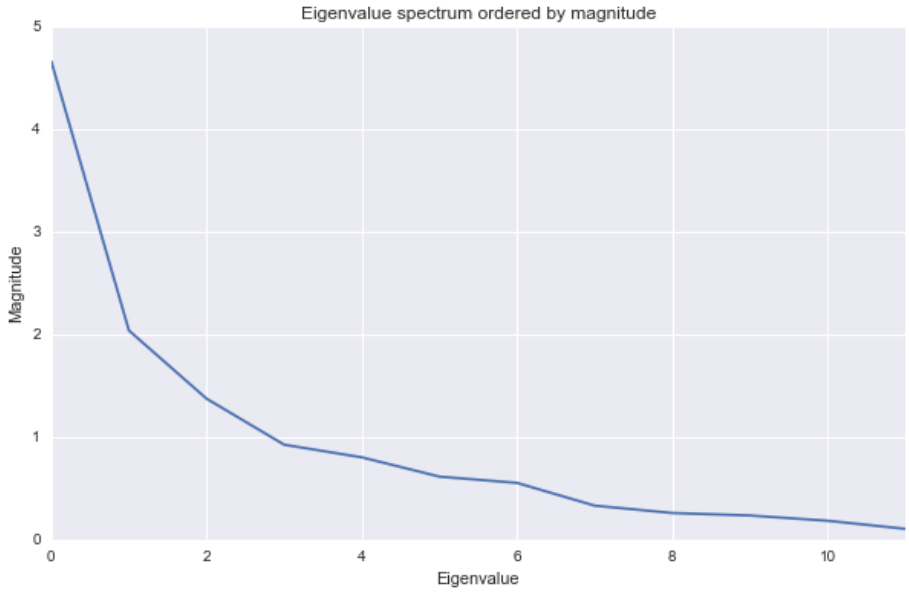
B Appendix for X_{-hue}

Table 3: Top eigenvectors of $Covar(X_{-hue})$

Eigenvector 1		Eigenvector 2	
Eigenvalue	4.3883	Eigenvalue	2.3284
Dimension		Dimension	
Flavanoids	0.4369	Color intensity	0.5442
Total phenols	0.4159	Alcohol	0.4772
OD280/OD315	0.3753	Ash	0.3610
Proanthocyanins	0.3339	Proline C	0.3435
Proline C	0.3220	Magnesium	0.2981
Nonflavanoid phenols	-0.3168	Malic acid	0.2501
Alcalinity of ash	-0.2512	0D280/OD315	-0.2353
Malic acid	-0.2249	Nonflavanoid phenols	0.1115
Alcohol	0.1896	Flavanoids	-0.0752
Magnesium	0.1687	Alcalinity of ash	0.0439
Color intensity	-0.0384	Proanthocyanins	-0.0304
Ash	0.0154	Total phenols	-0.0052

Figure 10: Eigenvalues and Eigenvectors of $Covar(X_{-alcohol})$

(a) Eigenvalue spectrum



(b) Cumulative eigenvalues

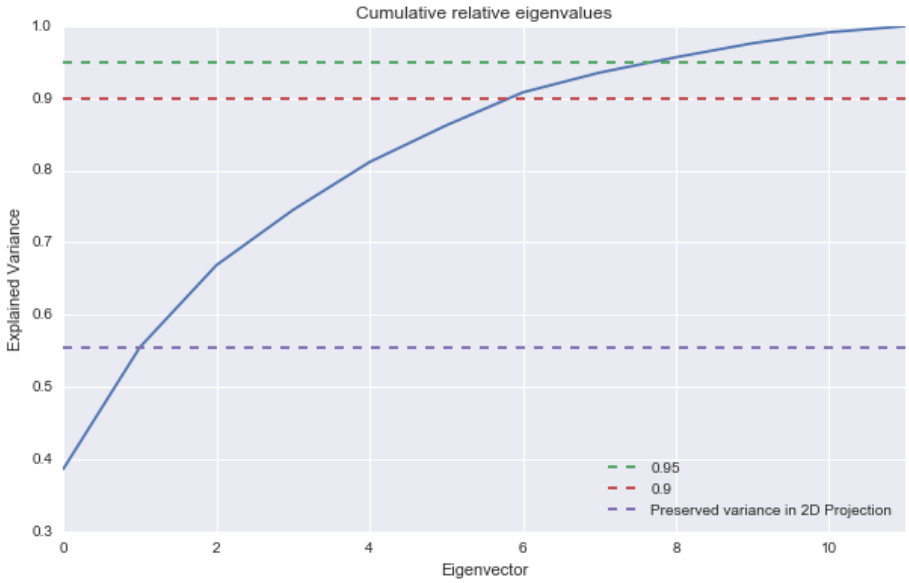
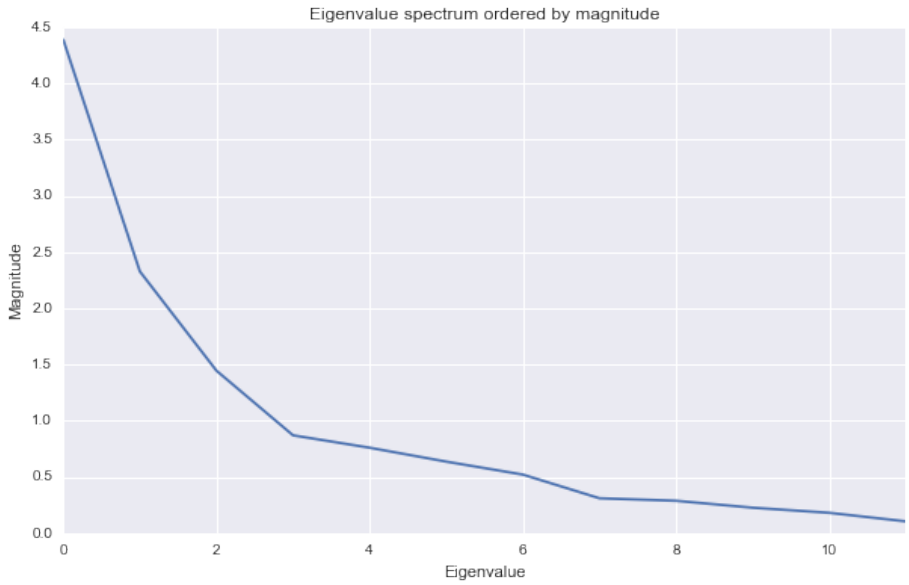
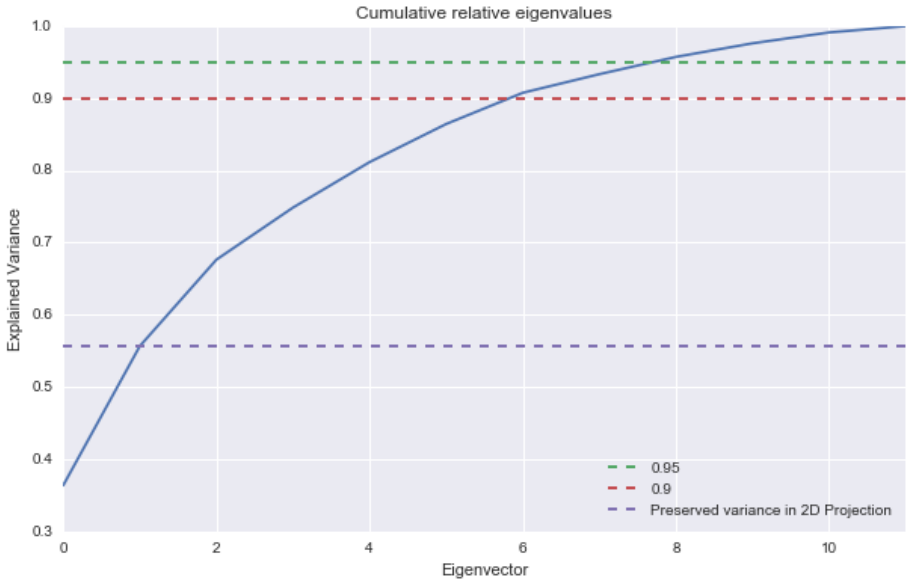


Figure 11: Eigenvalues and Eigenvectors of $Covar(X_{-huel})$

(a) Eigenvalue spectrum



(b) Cumulative eigenvalues



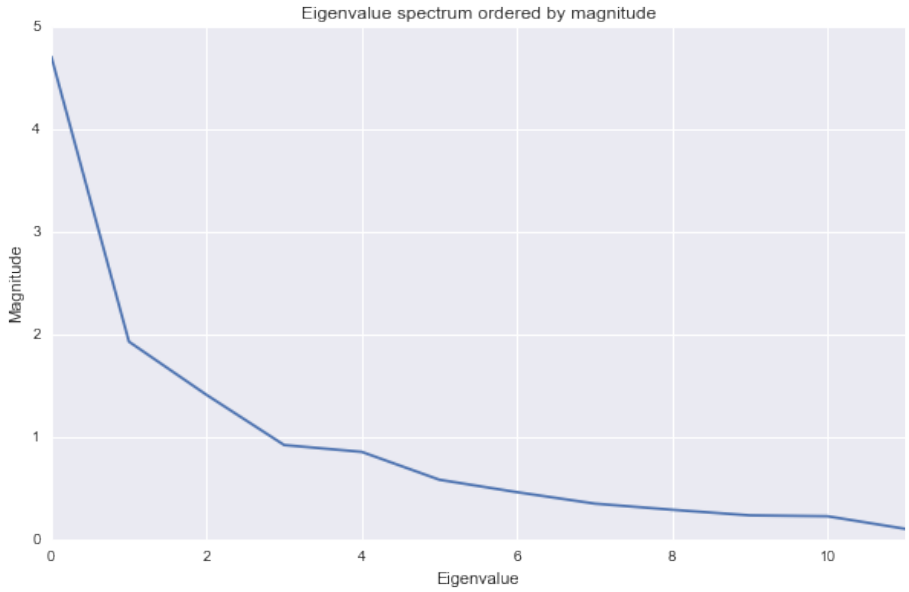
C Appendix for $X_{-color\ intensity}$

Table 4: Top eigenvectors of $Covar(X_{-color\ intensity})$

Eigenvector 1		Eigenvector 2	
Eigenvalue	4.7056	Eigenvalue	1.9286
Dimension		Dimension	
Flavanoids	0.4234	Alcohol	0.4931
Total phenols	0.3980	Ash	0.4790
OD280/OD315	0.3692	Magnesium	0.3858
Proanthocyanins	0.3161	Proline C	0.3731
Proline C	0.3007	Malic acid	0.3256
Nonflavanoid phenols	-0.2983	Hue	-0.2942
Hue	0.2856	OD285/OD315	-0.1615
Alcalinity of ash	-0.2416	Alcalinity of ash	0.1071
Malic acid	-0.2392	Nonflavanoid phenols	0.0788
Alcohol	0.1630	Total phenols	0.0499
Magnesium	0.1514	Flavanoids	-0.0183
Ash	0.0072	Proanthocyanins	0.0137

Figure 12: Eigenvalues and Eigenvectors of $Covar(X_{-color\ intensity})$

(a) Eigenvalue spectrum



(b) Cumulative eigenvalues