Journal of Wine Economics, Volume 16, Number 1, 2021, Pages 56–67
doi:10.1017/jwe.2020.35

Classification of Wines Using Principal Component Analysis

Jackson Barth [a], Duwani Katumullage [b], Chenyu Yang [c] and Jing Cao [d]

Abstract

Classification of wines with a large number of correlated covariates may lead to classification results that are difficult to interpret. In this study, we use a publicly available dataset on wines from three known cultivars, where there are 13 highly correlated variables measuring chemical compounds of wines. The goal is to produce an efficient classifier with straightforward interpretation to shed light on the important features of wines in the classification. To achieve the goal, we incorporate principal component analysis (PCA) in the k-nearest neighbor (kNN) classification to deal with the serious multicollinearity among the explanatory variables. PCA can identify the underlying dominant features and provide a more succinct and straightforward summary over the correlated covariates. The study shows that kNN combined with PCA yields a much simpler and interpretable classifier that has comparable performance with kNN based on all the 13 variables. The appropriate number of principal components is chosen to strike a balance between predictive accuracy and simplicity of interpretation. Our final classifier is based on only two principal components, which can be interpreted as the strength of taste and level of alcohol and fermentation in wines, respectively. (JEL Classifications: C10, Cl4, D83)

Keywords: cross-validation, k-nearest neighbor classification, principal component analysis.

I. Background

When asked to describe the characteristics of a certain wine, most people would rely on sensory descriptions. Adjectives such as “smooth,” “robust,” “full-bodied,” and

The authors gratefully acknowledge helpful comments and advice from the editor, Karl Storchmann, and an anonymous reviewer.

[a] Department of Statistical Science, Southern Methodist University, 3225 Daniel Ave, Dallas, Texas, 75275; e-mail: jbarth@smu.edu.
[b] Department of Statistical Science, Southern Methodist University, 3225 Daniel Ave, Dallas, Texas, 75275; e-mail: dkatumullage@smu.edu.
[c] Department of Statistical Science, Southern Methodist University, 3225 Daniel Ave, Dallas, Texas, 75275; e-mail: chenyuy@smu.edu.
[d] Department of Statistical Science, Southern Methodist University, 3225 Daniel Ave, Dallas, Texas, 75275; e-mail: jcao@smu.edu (corresponding author).

"dry," to name a few, are common among both novice and expert wine enthusiasts to describe the perceived differences in wines. These descriptors are subjective, and they can be inconsistent and often unreliable because opinions and personal preferences can vary greatly from person-to-person. This inconsistency of wine ratings by experts has been well-documented in previous studies (Hodgson, 2008; Cao and Stokes, 2010; Cao, 2014). But the importance of subjective qualities should not be overlooked. Oczkowski (2016) found that subjective expert opinion had similar predictive power as objective qualities when predicting the prices of Australian wines. Luxen (2018) examined the wine ratings by prominent critics and discovered that wines with below-average ratings tend to cost less (i.e., less than 35 euro), wines with higher ratings tend to cost more (i.e., between 35 and 100 euro). McCannon (2020) constructed a regression model using wine ratings and wine texts review to predict the price of wines.

However, many of these subjective qualities can be broken down into objective measurements via chemical analysis. Use of chemical analysis to classify wines has been widely implemented with some success. Cabrita et al. (2012) implemented principle component analysis, variance partition methods, and neural networks to classify wines based on non-flavonoid phenolic compounds. The study found that the variance partitioning method has the best performance. Beltran et al. (2008) employed both dimension reduction and pattern recognition techniques in the aroma classification on three types of Chilean wines (Cabernet Sauvignon, Merlot, and Carmenere), where the most successful classification algorithm has a correct-classification rate above 85%. Santos et al. (2017) have conducted a classification study by geographical origin, applying partial least squares-discriminant analysis to vibrational spectroscopic data from Portuguese white wines and observed a misclassification rate of 12.3%. Using an ordered probit model, Corsi and Ashenfelter (2019) found that weather patterns (summer rainfall, in particular) were significant in predicting expert vintage ratings.

These classification studies generally rely on high-dimensional datasets and "black box" machine-learning techniques. While capable of producing satisfactory classification results, such studies have limited capacity for clear interpretation. There are two main reasons: (1) many of the original objective measurements of wines (i.e., chemical compounds) are technical and difficult to understand for those without expert knowledge, and (2) with a large number of variables contributing to the algorithm, it can be hard to visualize the classification rules and determine which aspects of the data have the most significant effects (hence the "black box"). The ability to interpret these classifiers with familiar and relatable language is of great interest.

In this study, principal component analysis (PCA) is combined with a k-nearest neighbor (kNN) algorithm to produce a simpler and more interpretable classification rule. We used a publicly available dataset on wines with 13 highly correlated variables measuring their chemical compounds (Wine Data Set, 1991). The outcome variable for classification is cultivar, which represents three Italian varieties: Nebbiolo (the grape that produces Barolo and Barbaresco wines), Grignolino, and

Barbera. The dataset has been cited by several articles and is commonly used to demonstrate machine-learning techniques, specifically within the realm of classification. For example, it has been used to show how combinations of linear and quadratic programming can be leveraged to create stronger classification algorithms (Bredensteiner and Bennett, 1999). Kubica and Moore (2003) have employed the dataset to apply the learning explicit noise systems method to distinguish between corrupted and uncorrupted data. As recently as 2018, it has also been used to demonstrate the visualization of neural networks in classification problems (Duch, 2018).

PCA can be thought of as a dimension reduction and feature extraction tool, which has been applied in wine data analysis. It identifies patterns in a (large) group of correlated variables and transforms the patterns into a reduced number of uncorrelated features (i.e., principal components) without loss of important information. In the previously referenced study analyzing aroma chromatograms (Beltran et al., 2008), PCA was applied to reduce the number of classifiers from 600 original covariates to 20 principal components. The phenolic compounds study (Cabrita et al., 2012) also had success in reducing the dimensionality from 11 to 2 using PCA, while still maintaining roughly 85% of the original variability. There are also a number of studies applying PCA to the same dataset used in this article (Suthampan, 2017; Zhong and Fukushima, 2006) to reduce the effects of multicollinearity, an unavoidable issue when dealing with a large number of correlated covariates. In this article, PCA is applied not only to reduce multicollinearity but also to make the results of our analysis more interpretable—an arguably greater benefit to using PCA in many cases.

The goal of this study is to transform complex classification rules into terms that are easily understood by any wine connoisseur. In the following section, we will take a closer look at the dataset in question, categorizing each covariate into different categories and examining the correlation between covariates. In the methodology section, a brief overview is provided for kNN classification and PCA. In the analysis and results section, we implement the procedures combining kNN and PCA to the dataset. In the discussion section, we will interpret our findings and provide direction for further analysis.

II. Examination of the Dataset

The dataset we use consists of measurements on 178 wines obtained from a chemical analysis of three different cultivars (i.e., Nebbiolo, Grignolino, and Barbera), grown in a region called Piedmont in northwestern Italy. The data from the chemical analysis includes 13 numerical explanatory variables: Alcohol, Malic Acid, Ash Levels, Alkalinity of Ash, Magnesium Levels, Total Phenols, Flavonoids Phenols, Nonflavonoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280/OD315, and Proline.

Nebbiolo, Grignolino, and Barbera cultivars produce red wines. The Nebbiolo-based wines Barolo and Barbaresco wines are inarguably the most renowned

wines from Italy. They are aromatic with a floral character and have bright ruby color that fades over time. They have elevated acidity (makes the tongue feel wet) and tannins (makes the tongue feel astringent and dry), which are critical features of their success. They also have relatively high alcohol levels. Known to be the most widely grown grape in Piedmont, Barbera produces ruby-hued wines that have bright cherry flavors, floral and spicy aromas, and high alcohol levels. These grapes are high in acidity and have softer tannins and some roundness, which gives a sense of viscosity and thickness to the wine. Grignolino grapes produce wines that are less popular compared to the Nebbiolo-based Barolo and Barbaresco, or Barbera but are known for their extraordinary pale red color. Although the acidity level and tannin level of these grapes are high, Grignolino wines are known for the absence of warmth and intrigue. They tend to have fruity aromas, fruity and herbal flavors, and light alcohol levels (Wine-Searcher, 2020).

To better understand the relationship among the 13 explanatory variables, we categorized them as described in Table 1. It is essential to take note that these were broad categorizations based on the similar traits observed in the variables and that they are not strictly defined categories. Resonating the similarities that we see in Table 1, the correlation matrix in Figure 1 shows that most of the variables are highly correlated (the higher the correlation between variables, the higher the intensity of the color in Figure 1). As we emphasized earlier, since these are not strictly defined categories, we also observe a high correlation between variables across the categories. For example, the variable OD280/OD315 in "Appearance" shows a high correlation (magnitude of correlation > 0.5) with the three variables that describe the Phenol content in "Taste." In addition, the variables Alcohol in "Alcohol/Fermentation" and Color Intensity in "Appearance" also have a high correlation.

Our primary goal is to categorize wines by cultivar with an interpretable classifier based on the information of this chemical analysis. The initial correlational examination demonstrates high levels of multicollinearity present in the data. The methods we used in the study are constructed to reach our goal while dealing with the significant multicollinearity among a large number of covariates in the dataset.

III. Methodology

In this section, we briefly introduce PCA and kNN classification with 10-fold cross-validation, which we use to decide the optimum kNN algorithm.

A. PCA

PCA is primarily a dimension reduction method. That is, if we have a dataset with a large number of correlated variables, we can use PCA to transform those to a smaller number of uncorrelated principal components that would still explain most of the

Table 1
Description and Categorization of the Explanatory Variables

Category	*Explanatory Variable*	*Description*
Alcohol/fermentation	Alcohol	Percentage of alcohol by volume
	Proline	Amino acid that affects yeast growth and contributes to fermentation
Taste	Total Phenols	Chemical compounds that affect the taste and mouthfeel
	Flavonoid Phenols (found in vine, skin, and seed of grape)	
	Non-flavonoid Phenols (found in the flesh of grape)	
	Malic Acid	Acid contributing to the tartness
	Proanthocyanins	Glycoside affecting astringency or dryness
Mineral content	Ash Levels	Amount of inorganic minerals (i.e., potassium, iron, calcium)
	Alkalinity of Ash	The pH level of the ash
	Magnesium Levels	Level of magnesium
Appearance	Hue	General color
	Color Intensity	Shade (light vs. dark color)
	OD280/OD315	The measure of chemical concentration, contributing to cloudiness or haziness

information in the dataset. The use of PCA paves the way for the interpretation of the data structure based on the underlying dominant features that are capable of providing a more accurate summary of the correlated covariates.

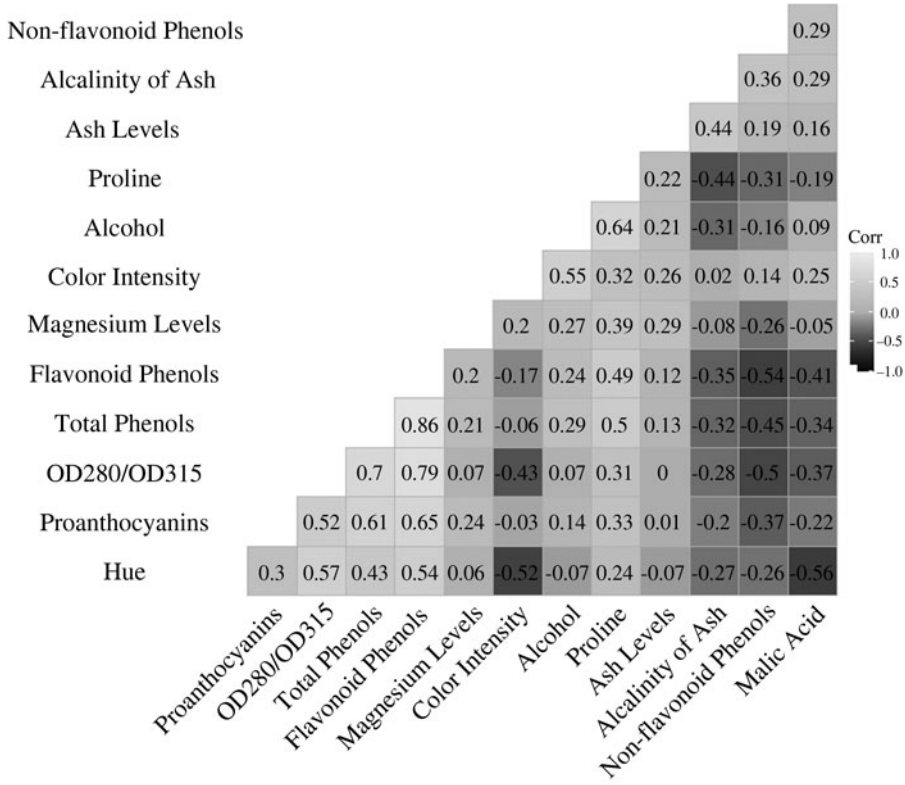
Principal components are linear combinations of the original variables. If there are g number of variables, we can produce g number of principal components. The first principal component (PC1) explains the highest proportion of variance within the data. The second principal component (PC2) is independent of PC1, and it explains the second-highest proportion of variance within the data. Likewise, the subsequent PCs will have the next highest proportion of variance, and they will be independent of the others. Theoretically, all g PCs are required to reproduce the total variability structure. However, most of the variability can often be captured by a much smaller number (h, $h \ll g$) of PCs (Johnson and Wichern, 2019) without loss of important information. Reducing the dimension in a dataset helps make data analysis more efficient and data visualization more straightforward. Usually, we standardize the variables before PCA is employed to avoid issues of variables with larger scales dominating variables with smaller scales.

B. Classification: kNN Classification with Cross-Validation

One of the simplest classification learning algorithms is the kNN classification. It is a non-parametric learning algorithm that identifies the k nearest neighbors in the

Figure 1

Sample Correlation Matrix of the Explanatory Variables



dataset to a new instance and assigns the most common label of the k nearest neighbors to the new instance based on a similarity measure. The similarity measure that we used in our study is the following Euclidean distance,

$$d(x_p, x_q) = \sqrt{(x_p^{(1)} - x_q^{(1)}) + (x_p^{(2)} - x_q^{(2)}) + \ldots + (x_p^{(D)} - x_q^{(D)})} = \sqrt{\sum_{j=1}^{D} x_p^{(j)} - x_q^{(j)}},$$

where $d(x_p, x_q)$ is the square root of the sum of the squared differences between a new instance (x_p) and an existing instance (x_q) across all input attributes $j = (1, \ldots, D)$.

The value k is known as the tuning parameter, and it has a direct effect on the performance of kNN. A conventional approach to choose k is through cross-validation. In particular, we employ the 10-fold cross-validation, where we randomly split the dataset into ten folds. Then, we keep one of the folds as the test set and use the training data in the other nine folds to predict the labels for wines in the test set. We rotate this procedure ten times for all the folds and obtain the average of the

misclassification rates across all the rotations pertaining to a particular choice of k. By conducting this procedure over a range of values of k, we can identify the optimal value for the tuning parameter in kNN.

IV. Analysis and Results

In the analysis, we first applied PCA to the 13 variables on the chemical compounds to reduce the dimensionality and extract underlying features on the chemical property of the wines. Then we used cross-validation to determine the optimal number of PCs and the number of neighbors (k) used in kNN. Our analysis found that using only the first two PCs yielded the best classification performance.

Table 2 provides a summary of covariates that have large contributions in the formation of PC1 and PC2. Of the four important variables in PC1, three of them (Total Phenols, Flavonoid Phenols, and Proanthocyanins) are measurements related to "Taste," while OD280/OD315 measures the cloudiness or haze of wine under "Appearance." Although we have divided these variables into distinct categories, it is clear from the correlation matrix (Figure 1) that there is a certain amount of overlap across the categories. Given that OD280/OD315 has a strong positive correlation with these "Taste" variables and that it also measures the concentration of chemical compounds, it is reasonable to interpret PC1 as a measure on the strength of the taste of wines.

Table 2 also shows that Color Intensity, Alcohol Content, Proline, Ash Levels, and Magnesium Levels are the key variables in the construction of PC2. Alcohol Content and Proline are both related to fermentation (Table 1). Color Intensity, although in the appearance category, has a strong association with Alcohol and a moderate association with Proline, which is shown in the correlation matrix (Figure 1). Ash Levels and Magnesium Levels also have a moderate association with Alcohol and Proline. Based on this examination, it is reasonable to interpret PC2 mainly as a measure of alcohol and fermentation levels.

We incorporated principal component (PC) selection in the cross-validation process and treated the number of PCs and the number of neighbors in kNN as the tuning parameters. Under each configuration of the tuning parameters, we repeated the 10-fold cross-validation 10^5 times with randomized training/testing split and computed the averages to be the empirical misclassification rates.

In the remainder of the section, we compare different classifiers with and without PCA. The misclassification rates for the different classifiers are reported in Table 3. They are also plotted against each other for a more straightforward visual comparison in Figure 2. In the 1-PC case, we applied the kNN algorithm to one PC at a time (i.e., only on PC1 and only on PC2, respectively). The lowest misclassification rate achieved by PC1 is around 15% at $k = 5$, whereas the lowest rate on PC2 is about 22% at $k = 15$. Note that kNN with PC1 enjoys a substantially lower misclassification rate, by a margin of about 10% on every configuration of k than that with PC2.

Table 2
Structure for the First 2-PCs with the Key Variables

Category	*Explanatory Variable*	*PC1 Loadings*	*PC2 Loadings*
Alcohol/fermentation	Alcohol		X
	Proline		X
Taste	Total Phenols	X	
	Flavonoid Phenols	X	
	Non-flavonoid Phenols		
	Malic Acid		
	Proanthocyanins	X	
Mineral content	Ash Levels		X
	Alkalinity of Ash		
	Magnesium Levels		X
Appearance	Hue		
	Color Intensity		X
	OD280/OD315	X	

In the 2-PC case, the kNN algorithm is applied to the first two PCs (i.e., on PC1 and PC2 together). The misclassification rate in the 2-PC case demonstrated a substantial improvement over the 1-PC case. As shown in Figure 2, the lowest misclassification rate is achieved at 3.3% at k around 10. We also examined the performance of classification with more than two PCs and found that adding more PCs in addition to the first two PCs will not reduce the misclassification rate. Furthermore, the performance of the 2-PC classifier is almost the same as that of the kNN classifier based on the 13 original variables. Taken together, the 2-PC classifier with $k = 10$ seems to provide an optimal balance between classification accuracy and interpretability.

Now we take a more in-depth examination of the classification results comparing the 1-PC classifier and the 2-PC classifier. In Figure 3, all 178 observations in the dataset are denoted with an asterisk, filled triangle, and circle corresponding to their respective cultivar labels (i.e., Nebbiolo, Grignolino, and Barbera). The coordinates of the observations are their respective PC1 and PC2 scores. Notice that there are three distinctive clusters belonging to the three cultivars. On the PC1 dimension (i.e., the strength of taste), Nebbiolo wines have the weakest strength, Grignolino wines have medium strength, and Nebbiolo wines have the strongest. On the PC2 dimension (i.e., level of alcohol and fermentation), Nebbiolo and Barbera wines have mostly positive values, whereas the Grignolino cluster has mostly negative values. Figure 3 unveils the reason for the classifier to perform relatively poorly in the 1-PC case. The data in the 1-PC case correspond to the projections of observations in the scatter plot onto the PC1 axis and to the PC2 axis, respectively. In the PC2-only case, it is evident from the graph that many observations would be mingled together on the PC2 coordinates, that is, Nebbiolo and Barbera are indistinguishable by PC2. In the PC1-only case, Nebbiolo and Barbera overlap with Grignolino, yet the overlapping is not quite as severe as that

Table 3
Misclassification Rates Among Different Classifiers

Number ofNeighbors	*1-PC*		*2-PC*	*13 OriginalCovariates*
	PC1	*PC2*	*PC1 and PC2*	
1	0.209	0.313	0.051	0.046
2	0.203	0.309	0.054	0.052
3	0.207	0.294	0.051	0.044
4	0.193	0.300	0.050	0.042
5	0.149	0.255	0.039	0.034
6	0.148	0.249	0.038	0.041
7	0.152	0.221	0.036	0.034
8	0.158	0.229	0.034	0.036
9	0.163	0.223	0.033	0.032
10	0.163	0.230	0.033	0.034
11	0.164	0.230	0.033	0.034
12	0.156	0.230	0.033	0.039
13	0.151	0.222	0.033	0.038
14	0.152	0.221	0.032	0.039
15	0.153	0.221	0.032	0.035

Figure 2
Misclassification Rates Among Different Classifiers

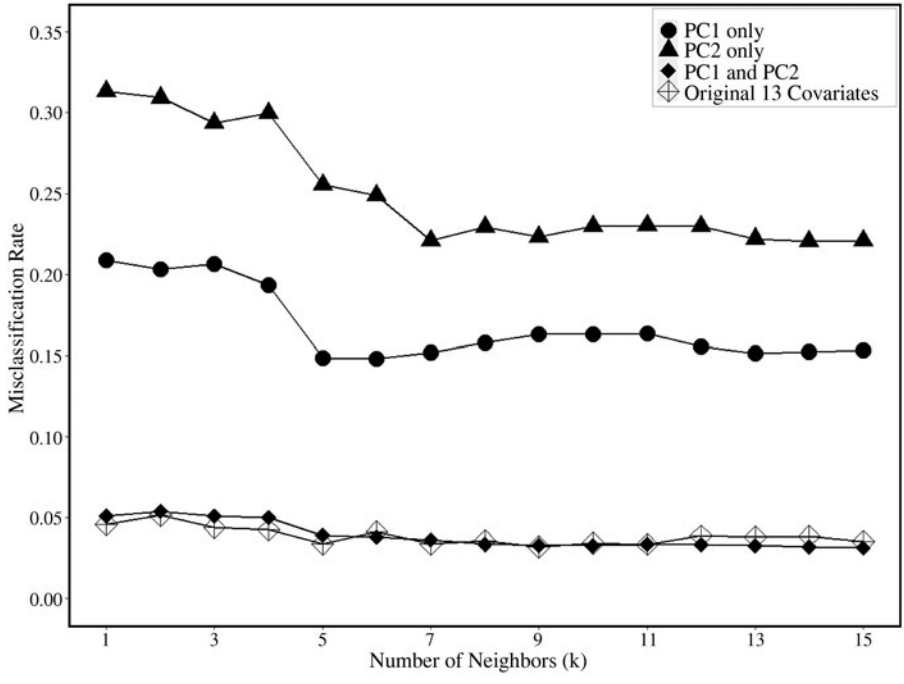
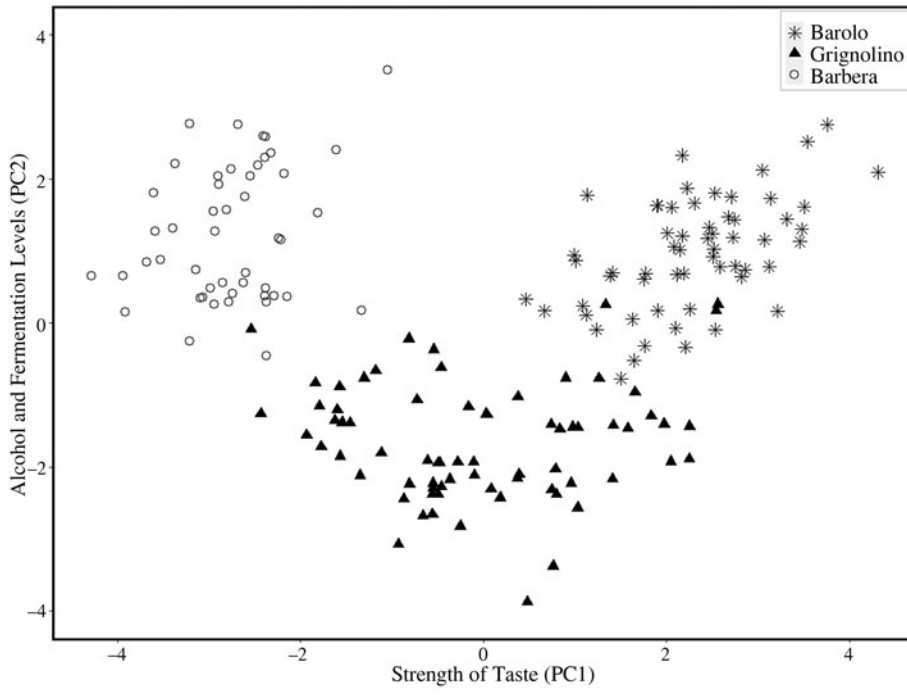


Figure 3

Scatter Plot of Wines on the PC1 and PC2 Coordinates



observed in the PC2-only case. This explains the improvement in the lowest misclassification rate by the PC1-only classifier (15%) over that by the PC2-only classifier (22%). By comparison, the 2-PC classifier results in a much better separation of the three cultivar groups, where they can be almost perfectly classified. Lastly, although the kNN classifier based on the 13 original variables has very similar performance compared to the 2-PC classifier, it cannot be interpreted based only on two features, nor can it be visualized in such a simple and informative plot.

V. Discussion

In this article, we show that kNN classification combined with PCA can be an efficient classifier with straightforward interpretation, which helps to shed light on the important features of wines in the classification. Due to the advantage of only using a small number of principal components in kNN, we can easily evaluate the importance of each feature, represented by individual principal components. In addition, data visualization of the classifier can be constructed with less effort while delivering insightful information in the classification.

The misclassification rate achieved by this study is relatively standard for this dataset. However, the interpretation of the classifier in this analysis deserves more attention. As discussed in the introduction, intelligible interpretation is something many studies of this type lack, despite its importance to wine research. Based on a review of the current literature, this study is the first analysis of this particular dataset to describe the classification algorithm in terms easily understood by all.

Although the misclassification rate is good by many standards, this study cannot make broader inferences to other types of wine outside of Nebbiolo, Grignolino, and Barbera. However, it does provide a template for future analysis of wine classification that can be easily interpreted. Applying these methods to data with a greater number of variables and a broader variety of cultivars will be the subject of further analysis.

References

Beltran, N. H., Duarte-Mermoud, M. A., Soto Vicencio, V. A., Salah, S. A., and Bustos, M. A. (2008). Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer. *IEEE Transactions on Instrumentation and Measurement*, 57(11), 2421–2436.

Bredensteiner, E. J., and Bennett, K. P. (1999). Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12, 53–79.

Cabrita, M., Aires-De-Sousa, J., Gomes Da Silva, M., Rei, F., and Costa Freitas, A. (2012). Multivariate statistical approaches for wine classification based on low molecular weight phenolic compounds. *Australian Journal of Grape and Wine Research*, 18, 138–146.

Cao, J. (2014). Quantifying randomness versus consensus in wine quality ratings. *Journal of Wine Economics*, 9(2), 202–213.

Cao, J., and Stokes, L. (2010). Evaluation of wine judge performance through three characteristics: Bias, discrimination, and variation. *Journal of Wine Economics*, 5(1), 132–142.

Corsi, A., and Ashenfelter, O. (2019). Predicting Italian wine quality from weather data and expert ratings. *Journal of Wine Economics*, 14(3), 234–251.

Duch, W. (2018). Coloring black boxes: Visualization of neural network decisions. *ArXiv.Org; Ithaca*. Available at https://arxiv.org/pdf/1802.08478.pdf.

Hodgson, R. T. (2008). An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, 3(2), 105–113.

Johnson, R. A., and Wichern, D. W. (2019). *Applied Multivariate Statistical Analysis* (6th ed.). Upper Saddle River, NJ: Pearson.

Kubica, J., and Moore, A. (2003). Probabilistic noise identification and data cleaning. Paper presented at the Third IEEE International Conference on Data Mining, Melbourne, FL. In *2013 Third IEEE International Conference on Data Mining*, 131–138. Available at https://www.computer.org/csdl/proceedings-article/icdm/2003/19780131/12OmNzcPAqS.

Luxen, M. F. (2018). Consensus between ratings of red Bordeaux wines by prominent critics and correlations with Prices 2004–2010 and 2011–2016: Ashton revisited and expanded. *Journal of Wine Economics*, 13(1), 83–91.

McCannon, B. C. (2020). Wine descriptions provide information: A text analysis. *Journal of Wine Economics*, 15(1), 71–94.

Oczkowski, E. (2016). Identifying the effects of objective and subjective quality on wine prices. *Journal of Wine Economics*, 11(2), 249–260.

Santos, C. A. T., Páscoa, R. N. M. J., Sarraguça, M. C., Porto, P. A. L. S., Cerdeira, A. L., González-Sáiz, J. M., Pizarro, C., and Lopes, J. A. (2017). Merging vibrational spectroscopic data for wine classification according to the geographic origin. *Food Research International*, 102, 504–510.

Suthampan, E. (2017). Principle component analysis (PCA) for wine dataset. Available at https://rstudio-pubs-static.s3.amazonaws.com/253795_29cb3d89b03e476a99ee2d32a7886243.html#/

Wine Data Set (1991). University of California at Irvine. UCI Machine Learning Repository. Available at http://archive.ics.uci.edu/ml/datasets/wine (accessed May 5, 2020).

Wine-Searcher (2020). Piedmont [Piemonte] wine. Available at https://www.wine-searcher.com/regions-piedmont+%5Bpiemonte%5D (accessed May 5, 2020).

Zhong, P., and Fukushima, M. (2006). Second-order cone programming formulations for robust multiclass classification. *Neural Computation*, 19(1), 258–282.