MACT 4233: Applied Multivariate Analysis - Dr. Ali Hadi

Project 1

The Application of the BACON Approach and the Hotelling's T^2 Test on the Wine Dataset

Farida Simaika - 900201753

Joyce Wassef - 900191248

Katia Gabriel - 900202272

Table Of Contents

Introduction

The wine industry contains various types of wines and relies solely on the quality of the drinks. It is one of the most important characteristics that differentiates a wine industry from the other. Therefore, the measure and the assurance of the quality of wines is significant. The aim of this project is to measure the quality of both white and red wines based on various features that will be thoroughly discussed in the course of this report. Moreover, important features and their levels that ensure the best quality of wines can be determined. Our aim is to help winemakers maximize the appeal of their wines through modifying certain aspects of their wines such as sweetness, acidity and shelf life.

Data Description

The data set encompassed 600 observations and 13 variables

Variable	Meaning	Unit	Type
Fixed Acidity	most acids involved with wine or fixed or nonvolatile (do not evaporate readily)	g / dm^3	Quantitative
Volatile acidity	the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste	g / dm^3	Quantitative
Citric acid	found in small quantities, citric acid can add 'freshness' and flavor to wines	g / dm^3	Quantitative
Residual sugar	the amount of sugar remaining after fermentation	g / dm^3	Quantitative

	stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet		
Chlorides	The amount of salt in the wine	g / dm^3	Quantitative
Free sulfur dioxide	The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine	mg / dm^3	Quantitative
Total sulfur dioxide	Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine	mg / dm^3	Quantitative
Density	The density of water is close to that of water depending on the percent alcohol and sugar content	g / cm^3	Quantitative
PH	Describes how acidic or basic a wine is on a scale from 0 (very acidic)	N/A	Quantitative

	to 14 (very basic); most wines are between 3-4 on the pH scale		
Sulfates	a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant	g / dm^3	Quantitative
Alcohol	The percent alcohol content of the wine	% by volume	Quantitative
Quality	Score between 0 and 10	N/A	Quantitative
Style	Type of wine (red or white)	N/A	Categorical

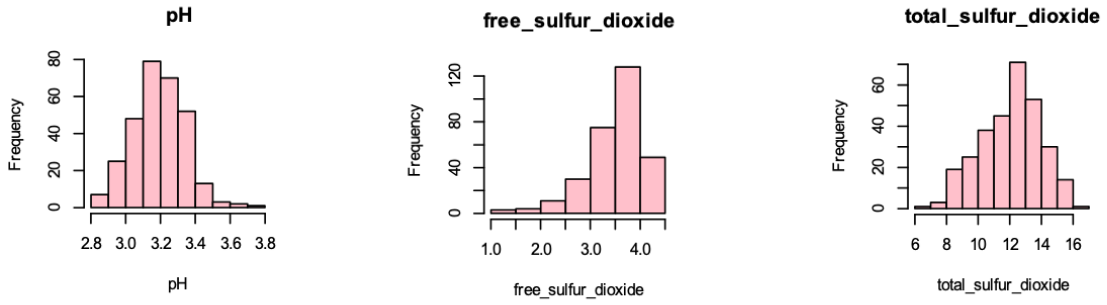
Source of the data:
https://www.kaggle.com/datasets/numberswithkartik/red-white-wine-dataset

Data Transformation

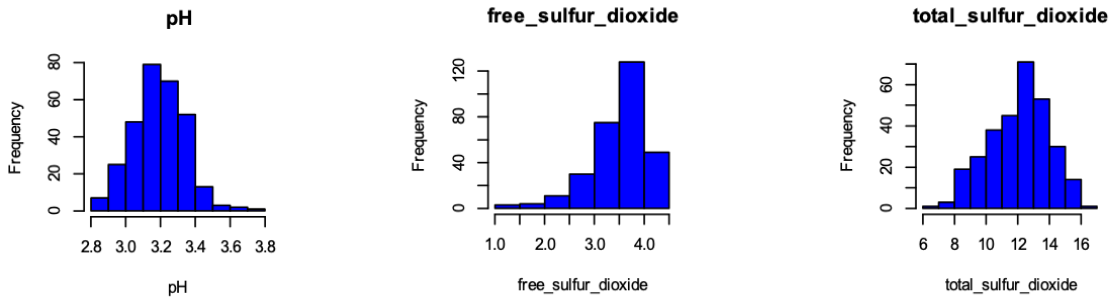
Upon plotting some of the variables in the data set, it was found that some of them do not follow a multivariate normal distribution and thus the appropriate transformations (such as log transformation or the square root) were applied. Transforming the variables is essential, it will allow the BACON approach to successfully identify the outlying data points.

The output below shows some of the variables where transformations were applied. As we can see below, they approximately follow a multivariate normal distribution.

The variables pH, free sulfur dioxide, and total sulfur dioxide before transformation:



The variables pH, free sulfur dioxide, and total sulfur dioxide after transformation:



Data Analysis

Outlier Detection using BACON on the Whole Data Set

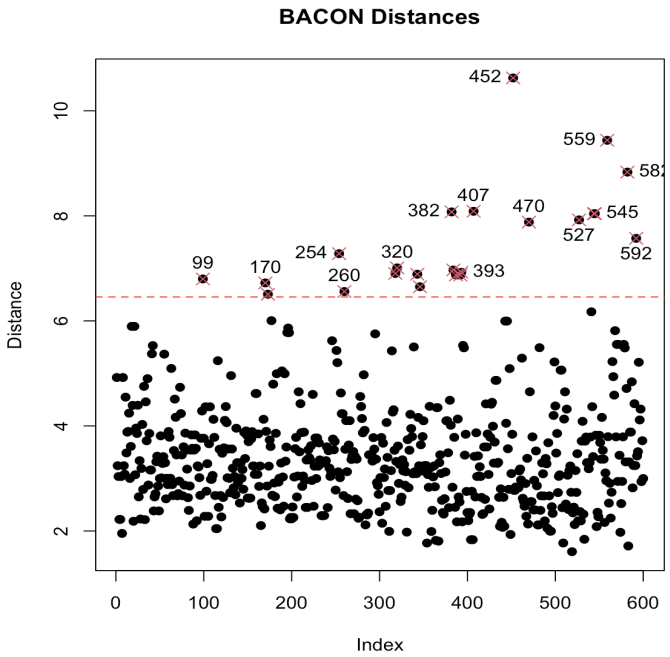
Outliers are commonly found in data sets. These outliers can drastically affect the results of multivariate techniques. Therefore, it is crucial to identify outliers when they exist in the data. Identification of outliers improves data quality and reliability, hence it improves the quality of the decisions drawn from the data and analysis.

Billor, Hadi, and Velleman's (2000) approach, BACON(Blocked, Adaptive,Computationally-Efficient Outlier Nominator,") is going to be implemented on this multivariate data for outlier detection.

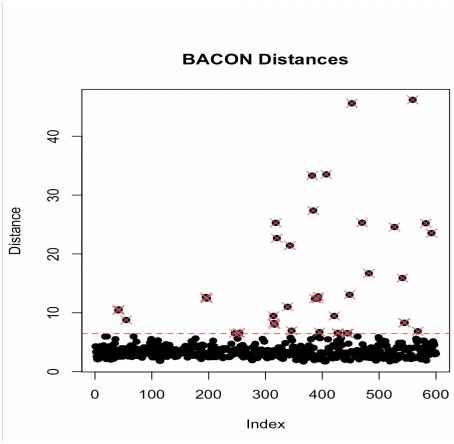
```
> output= mvBACON(d);
rank(x.ord[1:m,] >= p  ==> chosen m =  48
MV-BACON (subset no. 1): 48 of 600 (8 %)
MV-BACON (subset no. 2): 490 of 600 (81.67 %)
MV-BACON (subset no. 3): 555 of 600 (92.5 %)
MV-BACON (subset no. 4): 566 of 600 (94.33 %)
MV-BACON (subset no. 5): 572 of 600 (95.33 %)
MV-BACON (subset no. 6): 573 of 600 (95.5 %)
MV-BACON (subset no. 7): 575 of 600 (95.83 %)
MV-BACON (subset no. 8): 577 of 600 (96.17 %)
MV-BACON (subset no. 9): 577 of 600 (96.17 %)
```

The BACON approach was applied to the data set. Nine iterations were implemented in order to detect the outliers in the data set. The BACON algorithm was able to successfully detect the presence of 23 outliers in the data set. Approximately 3.83% of the data points were classified as outliers.

```
> output$limit
[1] 6.455448
> y = cbind(1:nrow(d),output$dis)
> colnames(y) <- c("Index","Distance");
> plot(y, pch=19, main = "BACON Distances")
> abline(h=output$limit,  col= "red",  lty=2)
> points(y[ ! output$subset, ], pch = 4, col = 2, cex = 1.5)
```

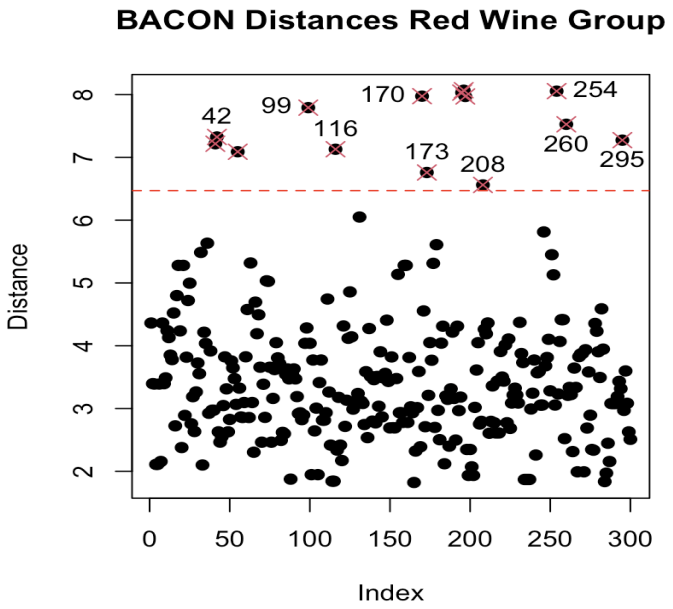


The adjusted χ^2(output limit) is 6.45448. This output limit separates between the basic subset (containing the non-outliers data points) and the non-basic subset (containing the outliers). Any data points with BACON distances greater than the output limit are considered outliers. As shown above, there are 23 outliers in this dataset. Transformation of the variables has helped us to narrow down the real outliers in the data set. As shown below, before transforming our data, the bacon algorithm detected 36 outliers in the data set some of which were very close to the basic subset (they were in fact not real outliers).

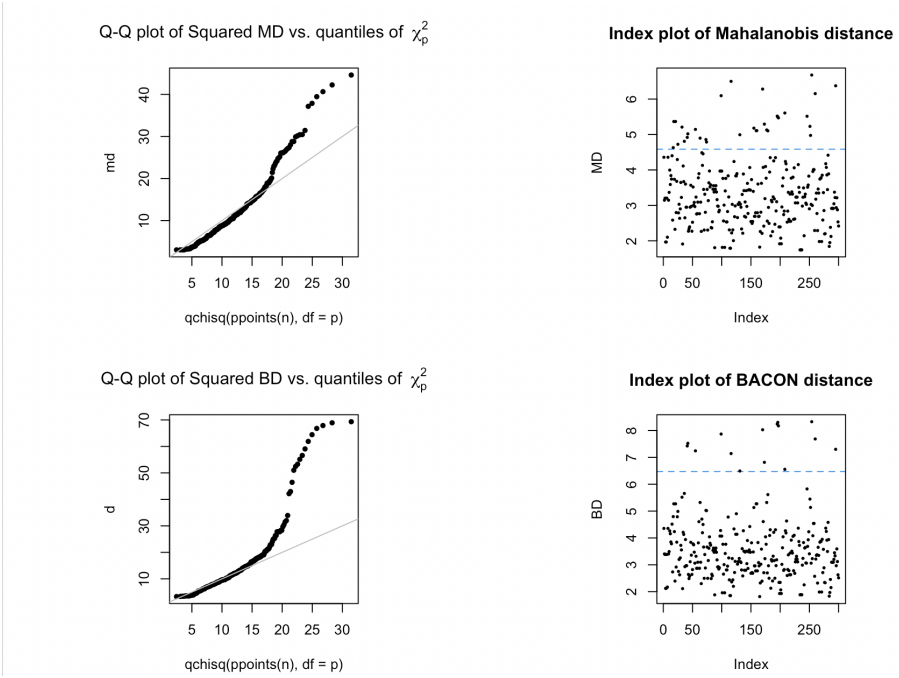


Outlier detection using BACON on Red Wine Group

```
> output= mvBACON(df);
rank(x.ord[1:m,] >= p  ==> chosen m =  48
MV-BACON (subset no. 1): 48 of 300 (16 %)
MV-BACON (subset no. 2): 254 of 300 (84.67 %)
MV-BACON (subset no. 3): 280 of 300 (93.33 %)
MV-BACON (subset no. 4): 286 of 300 (95.33 %)
MV-BACON (subset no. 5): 286 of 300 (95.33 %)
```



The adjusted χ^2(output limit) is still at 6.45448. As shown above, there are 14 outliers in the red wine group. The BACON algorithm was able to successfully detect the presence of 14 outliers in the data set after 5 iterations. Approximately 4.77% of the data points in the red wine group were classified as outliers.

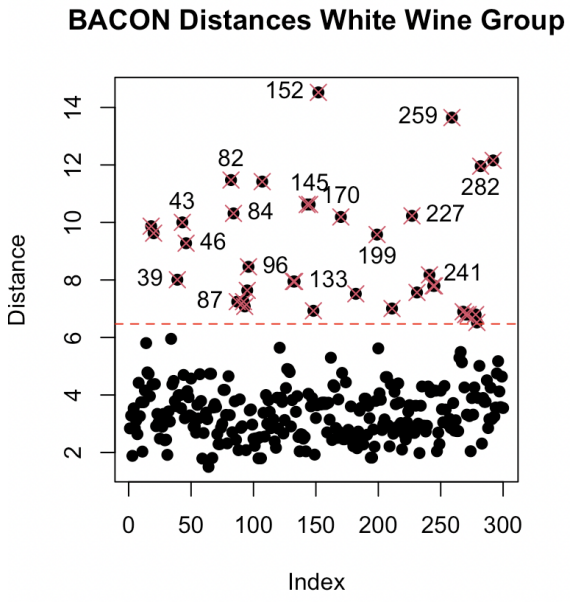


As shown above in the index plot of Mahalanobis distance, any point with a Mahlanobis distance greater than 4.5 is considered an outlier. The classic Mahlanobis distance is non-robust to outliers, hence why it detected more outliers than the BACON approach.

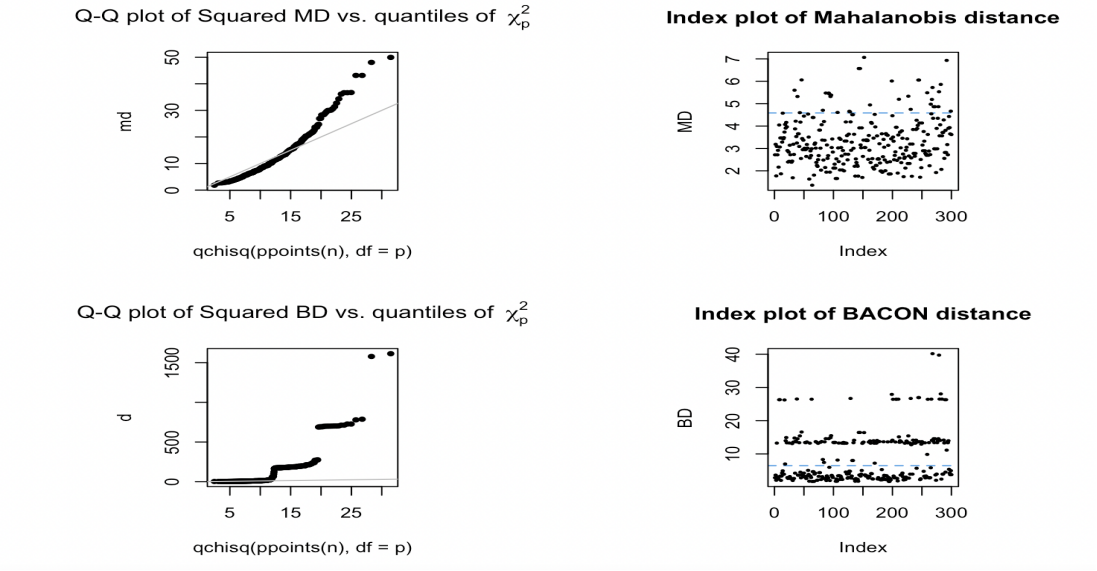
The QQ-plot of the Squared BACON distances shows that the data was following a Chi-Squared distribution before deviating from it as the quantiles increased.

Outlier detection using BACON on White Wine Group

```
> output= mvBACON(df);
rank(x.ord[1:m,] >= p  ==> chosen m =  48
MV-BACON (subset no. 1): 48 of 300
 (16 %)
MV-BACON (subset no. 2): 208 of 300
 (69.33 %)
MV-BACON (subset no. 3): 250 of 300
 (83.33 %)
MV-BACON (subset no. 4): 262 of 300
 (87.33 %)
MV-BACON (subset no. 5): 264 of 300
 (88 %)
MV-BACON (subset no. 6): 264 of 300
 (88 %)
```



The adjusted χ^2 (output limit) is also at 6.45448. As shown above, there are 36 outliers in the white wine group. The BACON algorithm was able to successfully detect the presence of 36 outliers in the data set after 6 iterations. Approximately 12% of the data points in the white wine group were classified as outliers.



As shown above in the index plot of Mahalanobis distance, any point with a Mahlanobis distance greater than 4.5 is considered an outlier. The classic Mahlanobis distance is non-robust to outliers, hence why it detected more outliers than the BACON approach.

The QQ-plot of the Squared BACON distances shows that the data was following a Chi-Squared distribution before deviating from it as the quantiles increased.

Outliers Interpretation

After careful analysis of the data, the reasons behind the presence of outliers in both groups were identified. In the red wine group, most of the data points classified as outliers had significantly high fixed acidity which increased their overall quality grades. The acidity level in wine is related to the overall lifespan of the drink and its quality. Acidity provides some of the backbone needed for long-term aging, so high acid wines are more likely to improve with time than those with lesser amounts. In the wine industry, high acid wines report better qualities and usually taste crisper on the palate. The high acidity levels are one of the reasons behind the presence of outliers in the red wine group.

In the white wine group, most of the data points classified as outliers had low sulfate quantities which degraded the overall quality of these wines. Sulfates are important additives in wine because they prevent its oxidation and are essential to preserving the wine. The outliers in the white win group had very low quality grades because of the low sulfate quantities.

In both the white wine and red wine groups, some of the observations classified as outliers had low sulfur dioxide quantities. This additive is an antioxidant and prevents microbial spoilage in the wine. Low quantities of sulfur dioxide degrade the overall quantity of wines. The outliers in the white wine group had very low quality grades because of the low sulfur dioxide quantities.

In both groups, the reasons behind some of the data points being classified as outliers are unknown and thus requiring more domain knowledge.

Hotelling's T^2 Test

An important question that can be discussed while analyzing the wine dataset is whether there is a difference between the means of red and white wines or not. The Hotelling's T2 Test is accordingly used in an attempt to verify the equality of the means between both wine types.
As a first step it is necessary to state both the null and the alternative Hypothesis.

The null hypothesis states that the means of the two groups are equal

$$H_0: \mu_r = \mu_w$$

The alternative hypothesis states that the means of the two groups are not equal

$$H_1: \mu_r \neq \mu_w$$

The Non-Robust Hotelling's T^2 Test

The non-robust Hotelling's T^2 Test was performed first with a significance level of $\alpha = 0.05$ and it yielded the following results:

Results

- T^2 = 4715.637
- F-statistic = 385.7412
- Degrees of Freedom = 12 & 587
- P-value = 0

After looking at the results the null hypothesis is to be rejected as the p-value is less than the significance level α. Accordingly, there is a difference between the means of the two wine types

and they are not equal. However, the above results have been affected by outliers and might not be reliable. Since it has been proven that the data contains outlier as detected by the bacon approach above, it is essential to perform the robust Hotelling's T^2 Test.

The Robust Hotelling's T^2 Test

The robust Hotelling's T^2 Test was performed as a second step with a significance level of $\alpha = 0.05$ and it yielded that following results:

Results:

Two-sample Hotelling test

data: x and y

- T2 = 4715.64
- F = 385.74
- df1 = 12
- df2 = 587
- p-value < 2.2e-16

alternative hypothesis: true difference in mean vectors is not equal to (0,0,0,0,0,0,0,0,0,0,0,0)

sample estimates:

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides
mean x-vector	6.875333	-1.2977099	0.5770903	1.5948103	-3.054284
mean y-vector	7.995333	-0.6473761	0.4271207	0.8188557	-2.433767

	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulfates	alcohol
mean x-vector	3.515331	12.066212	-0.005236610	3.194800	-0.7670916	2.297299
mean y-vector	2.537727	7.137096	-0.003063967	3.326733	-0.4145998	2.287228

quality

mean x-vector 5.723333

mean y-vector 5.406667

After looking at the results the null hypothesis is also to be rejected as the p-value is less than the significance level α. Accordingly, there is a difference between the means of the two wine types and they are not equal.

Conclusion

The wine industry relies heavily on the taste and the quality of wine. A lot of ingredients and ratios are needed to come up with the best tasting wine. Just one single addition of an acid or dioxides can improve or deteriorate the quality of the end product. The wine dataset represents a sample of wines produced in the wine industry, specifically both red and white wines. To ensure the quality of the dataset the BACON approach was performed to detect the existence of any outliers that can affect the results of statistical analysis. Approximately 12% of the data points in the white wine group were classified as outliers. Approximately 4.77% of the data points in the red wine group were classified as outliers. In attempt to understand the underlying reason for them being outliers, it was observed that in the red wine group, most of the outliers had significantly high fixed acidity which increased their overall quality grades and in the white wine group, most of the outliers had low sulfate quantities which degraded the overall quality of these wines. Moreover, the Hotelling's T^2 Tests performed showed that the means of the red wine group and the white wine group differed, hence they do not share the same extent of characteristics.