Canonical Correlation Analysis (CCA) Report for Quality of Red Wine

ST405 – Multivariate Methods II

S/18/837

1. Introduction

Canonical Correlation Analysis (CCA) is a multivariate statistical technique designed to identify and quantify the relationships between two sets of variables. Unlike other methods that focus on a single set of variables, CCA simultaneously explores multiple dependent and independent variables, providing a more comprehensive understanding of how these sets interact. By maximizing the correlations between linear combinations of the two variable sets, CCA uncovers the underlying patterns that link them together.

In this study, we apply CCA to a dataset measuring various chemical properties that influence the quality of wine. Our dataset includes variables related to acidity, sugar content, and other chemical properties such as chlorides, sulfur dioxide levels, density, pH, sulphates and alcohol. The primary question we aim to address is how do the variables related to acidity and sugar content correlate with those related to the chemical composition and stability of wine. The purpose of this study is to uncover the relationships between these two sets of variables, providing insights into which chemical properties most significantly influence wine quality.

We hypothesize that there will be significant canonical correlations between the sets, indicating strong interdependencies that can inform winemaking practices and quality control.

2. Methodology

In this analysis, I used “Red Wine Quality” data set. The dataset used in this study contains measurements of various chemical properties that influence the quality of wine. Specifically, the dataset includes the following variables:

- Fixed Acidity: The concentration of fixed acids present in the wine, which contribute to its tart taste.
- Volatile Acidity: The concentration of volatile acids, primarily acetic acid, which can impart a vinegar-like taste if present in high amounts.

- Citric Acid: An acid that can add freshness and tanginess to the wine.
- Residual Sugar: The amount of sugar remaining after fermentation, contributing to the wine's sweetness.
- Chlorides: The salt content in the wine, which can affect its taste.
- Free Sulfur Dioxide: Sulfur dioxide that is not bound to other molecules, playing a role in preventing microbial growth and oxidation.
- Total Sulfur Dioxide: The total amount of sulfur dioxide, including both free and bound forms.
- Density: Reflects the sugar content and can indicate the stage of fermentation.
- pH: A measure of the wine's acidity, influencing taste and microbial stability.
- Sulphates: Contribute to the bitterness and astringency of the wine.
- Alcohol: The alcohol content, affecting the wine's body, flavor, and preservation.

These variables were divided into two sets for Canonical Correlation Analysis (CCA):

Set A (Acidity and Fermentation):

- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- pH

Set B (Chemical Composition and Stability):

- Chlorides
- Free Sulfur Dioxide
- Total Sulfur Dioxide
- Density
- Sulphates
- Alcohol

Canonical Correlation Analysis (CCA) was performed to examine the relationships between the two sets of variables. This was performed to examine the relationships between the two sets of variables.

3. Results and Discussion

First split the data set into two sets. Set A is Acidity and Fermentation and set B is Chemical Composition and Stability.

	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	pH <dbl>
1	7.4	0.70	0.00	1.9	3.51
2	7.8	0.88	0.00	2.6	3.20
3	7.8	0.76	0.04	2.3	3.26
4	11.2	0.28	0.56	1.9	3.16
5	7.4	0.70	0.00	1.9	3.51
6	7.4	0.66	0.00	1.8	3.51

Set A

chlorides <dbl>	free.sulfur.dioxide <dbl>	total.sulfur.dioxide <dbl>	density <dbl>	sulphates <dbl>
0.076	11	34	0.9978	0.56
0.098	25	67	0.9968	0.68
0.092	15	54	0.9970	0.65
0.075	17	60	0.9980	0.58
0.076	11	34	0.9978	0.56
0.075	13	40	0.9978	0.56

Set B

Then we fitted Canonical Correlation Model and Canonical Correlations are obtained.

```r
ccModel <- cc(acidityNfermentation,chemicalCompNstability)
```

```r
ccModel$cor
```

```
[1] 0.78505514 0.50609008 0.33456777 0.24525275 0.02461554
```

There are five Canonical Correlations in this model. These values represent the strength of the relationships between the canonical variates of the two sets of variables. First Canonical Correlation (0.78505514) is the highest canonical correlation, indicating a strong relationship between the first canonical variates of the two sets. The second canonical correlation (0.50609008) is moderate. This indicates that there is a moderate relationship between the second canonical variates of the two sets. Third Canonical Correlation (0.33456777) value is a weaker, yet still notable but last two Canonical Correlation values are indicates a relatively weak relationship between the canonical variates of the two sets.

Test for independence between canonical variate pairs as follows.

H_0 : All the canonical correlations in the current row and all that follow are zero.

H_1 : At least one canonical correlation in the current row and all that are not equal to zero.

```
Wilks' Lambda, using F-approximation (Rao's F):
              stat      approx df1      df2   p.value
1 to 5:  0.2380771 111.3738718  25 5904.375 0.0000000
2 to 5:  0.6204960  51.3369426  16 4858.168 0.0000000
3 to 5:  0.8341426  33.2847794   9 3872.227 0.0000000
4 to 5:  0.9392816  25.3252179   4 3184.000 0.0000000
5 to 5:  0.9993941   0.9658232   1 1593.000 0.3258733
```

We tested this using Wilk's lamda test. The results suggest first four canonical correlations are significant, indicating strong relationships between the sets of variables. The fifth correlation is not significant, suggesting no additional meaningful relationship at that level.

These are the significant Canonical Correlations.

```
```{r}
ccModel$cor[1:4]
```
```

```
[1] 0.7850551 0.5060901 0.3345678 0.2452527
```

Hence we can compute squared Canonical Correlations.

```
```{r}
ccModel$cor[1:4]^2
```
```

```
[1] 0.61631157 0.25612717 0.11193559 0.06014891
```

These values are suggested that, 61.63% of variation in first canonical variable of set A (Acidity and Fermentation) is explained by the variation in first canonical variable in set B (Chemical Composition and Stability). 25.61% of variation in first canonical variable of set A is explained by the variation in second canonical variable in set B. Other values are sufficient lower values. Therefore first two canonical correlations are more important.

Canonical coefficients represent the weights applied to each original variable to obtain the canonical variates. The canonical variates are the linear combinations of the original variables that maximize the correlation between the two sets of variables.

These are the estimates canonical correlations for the set “Acidity and Fermentation”.

```{r}
ccModel$xcoef
```

	[,1]	[,2]	[,3]	[,4]	[,5]
fixed.acidity	1.29277730	-0.8350833	0.1227291	0.38339170	-0.08450542
volatile.acidity	0.08316702	0.7813110	0.5003307	0.56428640	0.56454614
citric.acid	-0.23740581	0.9349157	-0.5801105	-0.03715415	1.16403961
residual.sugar	0.31632309	0.1936772	0.4370638	-0.83446111	-0.07867608
pH	0.41694951	-0.8138926	0.1011651	-0.06929742	1.03360623

- First Canonical Variate: Dominated by fixed acidity, indicating its major role in the first dimension of the relationship between sets.
- Second Canonical Variate: Influenced by fixed acidity (negatively), volatile acidity, citric acid, and pH (negatively), highlighting these as key variables in the second dimension.
- Third Canonical Variate: Influenced by citric acid (negatively) and volatile acidity and residual sugar (positively).
- Fourth Canonical Variate: Strongly influenced by residual sugar with a significant negative contribution.
- Fifth Canonical Variate: Citric acid and pH are the primary positive influences.

These are the estimates canonical correlations for the set “Chemical Composition and Stability”.

```{r}
ccModel$ycoef
```

	[,1]	[,2]	[,3]	[,4]	[,5]
chlorides	-0.21641437	0.7513297	0.10167128	0.31173314	-0.68778273
free.sulfur.dioxide	0.06592875	-0.4673532	0.29386062	-0.80817166	-0.92952742
total.sulfur.dioxide	-0.24918075	0.8300691	0.08851987	-0.13101998	1.02751220
density	1.01429997	0.0878536	0.13341218	-0.08700913	0.00676233
sulphates	0.01392734	-0.1122272	-0.99425896	-0.38046601	0.15998760

- First Canonical Variate: Dominated by density, indicating its major role in the first dimension of the relationship between sets.
- Second Canonical Variate: Influenced by chlorides and total sulfur dioxide, highlighting these as key variables in the second dimension.
- Third Canonical Variate: Strongly influenced by sulphates, showing a significant negative contribution.
- Fourth Canonical Variate: Influenced by free sulfur dioxide, with a strong negative contribution.
- Fifth Canonical Variate: Influenced by total sulfur dioxide (positive) and free sulfur dioxide and chlorides (negative).

These values represent the correlation coefficients between the variables in Set A and the canonical variables obtained from Canonical Correlation Analysis (CCA). Here are the correlations.

```{r}
loadingswine$corr.X.xscores
```

	[,1]	[,2]	[,3]	[,4]	[,5]
fixed.acidity	0.86354847	0.1708854	-0.4140121	0.1654559	-0.162174423
volatile.acidity	-0.01822384	0.2878225	0.8140102	0.4687345	0.185745484
citric.acid	0.40447783	0.4111758	-0.7661729	-0.1736514	0.223956136
residual.sugar	0.39506455	0.3032720	0.3601541	-0.7887734	-0.008693976
pH	-0.34489281	-0.5832134	0.4118189	-0.1069661	0.599895440

- Canonical Variable 1: Strongly correlated with fixed acidity.
- Canonical Variable 2: Moderate positively correlated with volatile acidity and Strong negatively correlated with pH.
- Canonical Variable 3: Strongly positive correlation with volatile acidity and Strongly negative correlation citric acid.
- Canonical Variable 4: Strongly negative correlation with residual sugar.
- Canonical Variable 5: Strongly correlated with pH.

These values represent the correlation coefficients between the variables in Set B and the canonical variables obtained from Canonical Correlation Analysis (CCA). Here are the correlations.

```{r}
loadingsWine$corr.Y.yscores
```

	[,1]	[,2]	[,3]	[,4]	[,5]
chlorides	-0.01918692	0.76403658	-0.234860599	0.14231872	-0.58349452
free.sulfur.dioxide	-0.12318482	0.08330968	0.299239026	-0.91165986	-0.23920137
total.sulfur.dioxide	-0.14253372	0.55508790	0.256347946	-0.67837369	0.38164962
density	0.95374271	0.29134320	0.006016647	-0.07256875	-0.01384043
sulphates	0.07691548	0.19126511	-0.917718091	-0.32502831	-0.09824343

- Canonical Variable 1: Strongly correlated with density.
- Canonical Variable 2: Strongly correlated with chlorides and total sulfur dioxide.
- Canonical Variable 3: Strong negative correlation with sulphates.
- Canonical Variable 4: Strongly negatively correlated with free sulfur dioxide and total sulfur dioxide.
- Canonical Variable 5: Moderately negatively correlated with chlorides.

These values represent the correlation coefficients between the variables in Set A and the canonical variables obtained from Canonical Correlation Analysis (CCA) for Set B. Here are the correlations.

```{r}
loadingsWine$corr.X.yscores
```

	[,1]	[,2]	[,3]	[,4]	[,5]
fixed.acidity	0.67793316	0.08648342	-0.1385151	0.04057851	-0.0039920106
volatile.acidity	-0.01430672	0.14566412	0.2723416	0.11495842	0.0045722249
citric.acid	0.31753740	0.20809200	-0.2563368	-0.04258849	0.0055128006
residual.sugar	0.31014745	0.15348295	0.1204960	-0.19344885	-0.0002140069
pH	-0.27075988	-0.29515852	0.1377813	-0.02623372	0.0147667485

- Canonical Variable 1: Moderately positively correlated with fixed acidity, citric acid, and residual sugar.
- Canonical Variable 2: Weakly negatively correlated with pH.
- Canonical Variable 3: Moderately correlated with volatile acidity and citric acid, but in opposite directions.
- Canonical Variable 4: Weakly negatively correlated with residual sugar.
- Canonical Variable 5: Very weakly positively correlated with volatile acidity, citric acid, and pH.

These values represent the correlation coefficients between the variables in Set B and the canonical variables obtained from Canonical Correlation Analysis (CCA) for Set A. Here are the correlations.\

```{r}
loadingsWine$corr.Y.xscores
```

	[,1]	[,2]	[,3]	[,4]	[,5]
chlorides	-0.01506279	0.38667134	-0.078576788	0.03490406	-0.0143630311
free.sulfur.dioxide	-0.09670687	0.04216220	0.100115735	-0.22358708	-0.0058880703
total.sulfur.dioxide	-0.11189683	0.28092448	0.085765761	-0.16637301	0.0093945103
density	0.74874062	0.14744590	0.002012976	-0.01779768	-0.0003406895
sulphates	0.06038289	0.09679738	-0.307038897	-0.07971408	-0.0024183148

- Canonical Variable 1: Strongly correlated with density.
- Canonical Variable 2: Moderately correlated with chlorides and total sulfur dioxide.
- Canonical Variable 3: Moderately negatively correlated with sulphates.
- Canonical Variable 4: Weakly negatively correlated with free sulfur dioxide and total sulfur dioxide.
- Canonical Variable 5: Very weakly negatively correlated with chlorides and free sulfur dioxide.

4. Conclusion and Recommendation

In this study, we conducted Canonical Correlation Analysis (CCA) to explore the relationships between two sets of variables: Set A, comprising variables related to acidity and fermentation in wine, and Set B, containing variables related to chemical composition and stability.

From squared canonical correlations we can conclude, 61.63% of variation in first canonical variable of set A (Acidity and Fermentation) is explained by the variation in first canonical variable in set B (Chemical Composition and Stability) while 25.61% of variation in first canonical variable of set A is explained by the variation in second canonical variable in set B.

Our analysis revealed significant correlations between the sets, with Canonical Correlation coefficients indicating the strength and direction of these relationships. For Set A, variables such as fixed acidity, volatile acidity, and citric acid were found to be strongly correlated with certain canonical variables, suggesting their importance in explaining variation within the dataset. Similarly, for Set B, variables like density and sulphates showed notable correlations with canonical variables, highlighting their potential impact on chemical composition and stability.

5. References

- Härdle, W. K., & Simar, L. (2019). Applied Multivariate Statistical Analysis (5th ed.). Springer. https://doi.org/10.1007/978-3-662-45171-7

- Oliveira, L. S., Franca, A. S., & Ferreira, M. E. (2006). Canonical discriminant analysis for coffee classification. Journal of Food Science, 71(7), S519-S523. https://doi.org/10.1111/j.1750-3841.2006.00156.x

- Institute for Digital Research and Education, S. C. (n.d.). Canonical Correlation Analysis

6. Appendices

Dataset :

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.880	0.00	2.60	0.098	25	67	0.9968	3.20	0.68	9.8	5
7.8	0.760	0.04	2.30	0.092	15	54	0.9970	3.26	0.65	9.8	5
11.2	0.280	0.56	1.90	0.075	17	60	0.9980	3.16	0.58	9.8	6
7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.660	0.00	1.80	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.600	0.06	1.60	0.069	15	59	0.9964	3.30	0.46	9.4	5
7.3	0.650	0.00	1.20	0.065	15	21	0.9946	3.39	0.47	10.0	7
7.8	0.580	0.02	2.00	0.073	9	18	0.9968	3.36	0.57	9.5	7

Codes :

#Load packages

```{r}
library(ggplot2)
library(skimr)
library(tidyverse)
library(CCA)
library(CCP)
```

#Load dataset

```{r}
wine_data <- read.csv("../data/winequality-red.csv",sep = ";")
```

```{r}
view(wine_data)
```

#Summarize dataset

```{r}
skim(wine_data)
```

#Split dataset

```{r}
acidityNfermentation <- wine_data[,c(1,2,3,4,9)]
chemicalCompNstability <- wine_data[,c(5,6,7,8,10)]
```

#Standarize dataset

```{r}
acidityNfermentation <- scale(acidityNfermentation)
chemicalCompNstability <- scale(chemicalCompNstability)
```

#Check correlation

```{r}
matcor(acidityNfermentation,chemicalCompNstability)
```

#Canonical correlation model

```{r}
ccModel <- cc(acidityNfermentation,chemicalCompNstability)
```

#correlations

```{r}
ccModel$cor
```

#Wilk's lamda test

```{r}
rho <- ccModel$cor
n <- dim(acidityNfermentation)[1]
p <- dim(chemicalCompNstability)[2]
q <- dim(acidityNfermentation)[2]
```

```{r}
p.asym(rho,n,p,q,tstat = "Wilks")
```

```{r}
p.asym(rho,n,p,q,tstat = "Hotelling")
```

```{r}
p.asym(rho,n,p,q,tstat = "Pillai")
```

```{r}
p.asym(rho,n,p,q,tstat = "Roy")
```

#Significant correlations

```{r}
ccModel$cor[1:4]
```

#squared correlations

```{r}
ccModel$cor[1:4]^2
```

#estimate canonical correlations

```{r}
ccModel$xcoef
```

```{r}
ccModel$ycoef
```

#correlation coefficients

```{r}
loadingsWine <- comput(acidityNfermentation,chemicalCompNstability,ccModel)
```

```{r}
loadingsWine$corr.X.xscores
```

```{r}
loadingsWine$corr.Y.yscores
```

```{r}
loadingsWine$corr.X.yscores
```

```{r}
loadingsWine$corr.Y.xscores
```