STAT 550 Final Project

Multivariate Statistical Analysis for Chemical Compounds of Wine

by STUDENT NAMES

STAT 550: Multivariate Statistical Analysis

Prof. Sung Kim

Date

Table of Contents

1. Introduction

Wine is one of the most popular alcoholic beverages around the world. People who know how to select wine usually check the label consisting of information such as what company and brand made the wine, grape varietal, vintage, etc attached to the bottle. The label gives general information for customers to help choose a wine depending on their personal favorite because more information is not necessary to select a wine. However, the label does not include what chemical compounds are contained in wine, so people do not know how many complex chemical compounds are composed in wine. A question arises at this point. 'Different type of wine contains different type of chemical compounds in wine?', 'Different amount of each chemical compound decides the different type of wine?' and 'What chemical compounds determine a type of wine?'. For this final project, with a few major chemical compounds of the wine, we are going to statistically analyze which chemical compounds are major compounds to distinguish different types of wine. Also, since there is no specific information for each class, we are going to present characteristics or relationships for each class as much as we can. Our project will be mostly focused on multivariate statistical techniques and a little knowledge of chemistry and (plant) biochemistry may be needed to draw a statistical conclusion.

2. Data

The dataset is based on the results of a chemical analysis of wines grown in the same region in Italy, but derived from three different cultivars. There are 178 rows (observations) in the dataset. There are 13 columns with one class identifier (Class 1, 2, and 3), but this variable does not include a specific name of wine and there is no information about the name of each class of wine on the internet. There is no way out to find the specific name of each class since we have an old dataset. Also, since the raw dataset does not identify what each column is for, we have added an index for each variable in our dataset. Here is index of each column. (See Table 1)

Table 1: Index of each variable in the wine dataset.

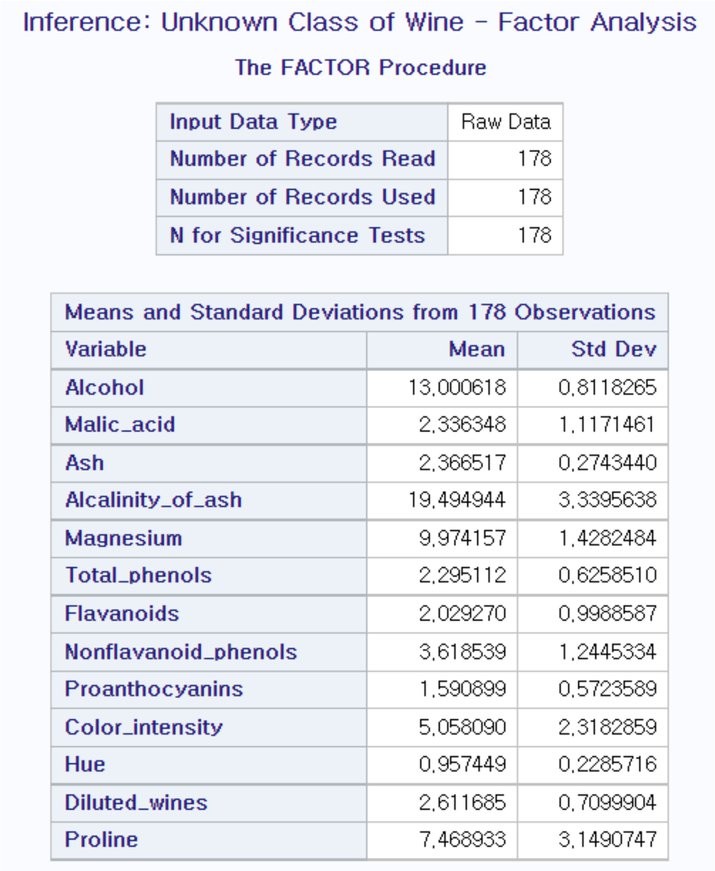
# of Column	Name of Variable	Type of Variable	Special Note
Column 1	Type (Class identifier)	Categorical	N/A
Column 2	Alcohol	Numeric	N/A
Column 3	Malic acid	Numeric	N/A
Column 4	Ash	Numeric	N/A
Column 5	**Alkalinity of ash**	Numeric	N/A
Column 6	Magnesium	Numeric	Re-scaled (Divided by 10)
Column 7	Total phenols	Numeric	N/A
Column 8	Flavanoids (Polyphenol)	Numeric	N/A
Column 9	Non-flavonoids (Polyphenol)	Numeric	Re-scaled (Multiplied by 10)
Column 10	Proanthocyanins	Numeric	N/A
Column 11	Color intensive	Numeric	N/A
Column 12	Hue	Numeric	N/A
Column 13	OD280/OD315 of diluted wines	Numeric	N/A
Column 14	**Proline**	Numeric	Re-scaled (Divided by 100)

Two variables 'Alkalinity of ash' and 'Proline' will be the most important variables for our project. A specific reason will be presented soon.

3. Multivariate Analysis (A. Factor Analysis)

Here is the short descriptive statistics of our dataset. (See Table 2)

Table 2: Descriptive statistics



Inference: Unknown Class of Wine - Factor Analysis

The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	178
Number of Records Used	178
N for Significance Tests	178

Means and Standard Deviations from 178 Observations		
Variable	Mean	Std Dev
Alcohol	13.000618	0.8118265
Malic_acid	2.336348	1.1171461
Ash	2.366517	0.2743440
Alcalinity_of_ash	19.494944	3.3395638
Magnesium	9.974157	1.4282484
Total_phenols	2.295112	0.6258510
Flavanoids	2.029270	0.9988587
Nonflavanoid_phenols	3.618539	1.2445334
Proanthocyanins	1.590899	0.5723589
Color_intensity	5.058090	2.3182859
Hue	0.957449	0.2285716
Diluted_wines	2.611685	0.7099904
Proline	7.468933	3.1490747

Since we already have fixed a scale of each variable, we have used a covariance matrix for our factor analysis. From the covariance matrix, we can compute all eigenvalues and proportion of each eigenvalue by the principal component method. (See Figure 1 and Table 3)

Figure 1: Scree plot and variance explained plot

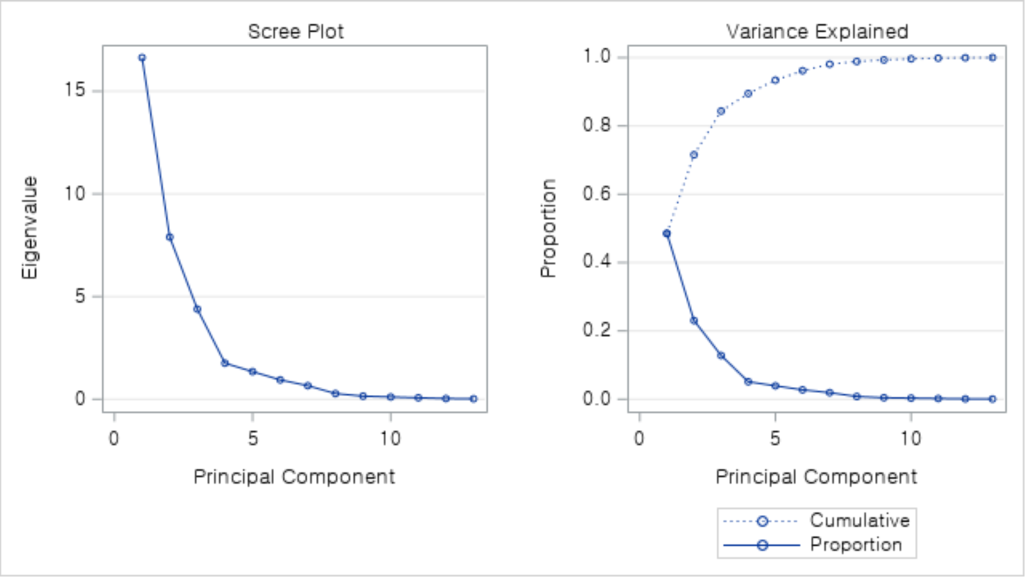
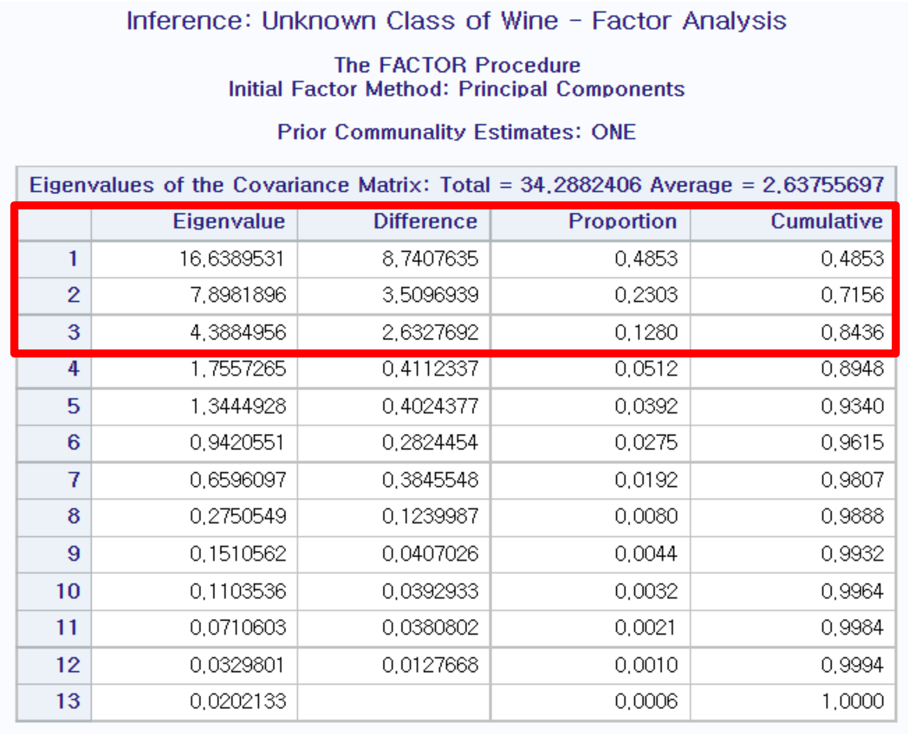


Table 3: Eigenvalues and cumulative proportion of eigenvalues



Inference: Unknown Class of Wine - Factor Analysis

The FACTOR Procedure
Initial Factor Method: Principal Components

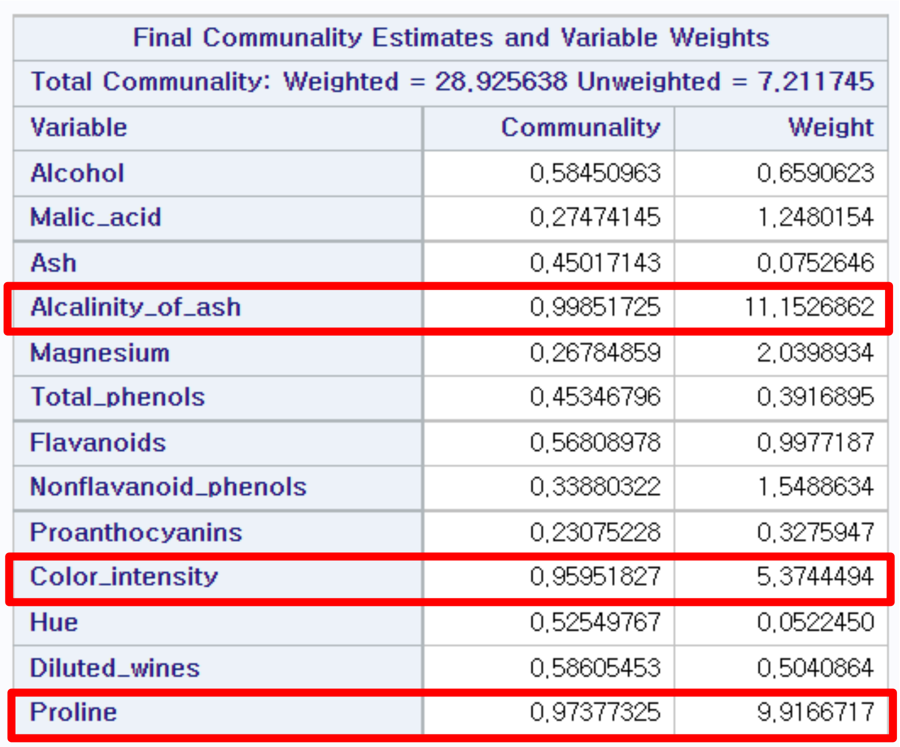
Prior Communality Estimates: ONE

Eigenvalues of the Covariance Matrix: Total = 34.2882406 Average = 2.63755697

	Eigenvalue	Difference	Proportion	Cumulative
1	16.6389531	8.7407635	0.4853	0.4853
2	7.8981896	3.5096939	0.2303	0.7156
3	4.3884956	2.6327692	0.1280	0.8436
4	1.7557265	0.4112337	0.0512	0.8948
5	1.3444928	0.4024377	0.0392	0.9340
6	0.9420551	0.2824454	0.0275	0.9615
7	0.6596097	0.3845548	0.0192	0.9807
8	0.2750549	0.1239987	0.0080	0.9888
9	0.1510562	0.0407026	0.0044	0.9932
10	0.1103536	0.0392933	0.0032	0.9964
11	0.0710603	0.0380802	0.0021	0.9984
12	0.0329801	0.0127668	0.0010	0.9994
13	0.0202133		0.0006	1.0000

From the table above, since total variance explained by top three eigenvalues is approximately 85% variance, we are going to choose three factor loadings for further analyses. Also, since we need a result of rotated factor loadings to compute the final factor scores for the next steps, additional analysis in this analysis using principal component method is not needed. Now, we are going to discuss the results of rotated factor loading by one of orthogonal rotation methods "Varimax". (See Tables and Figures)

Table 4: Communality

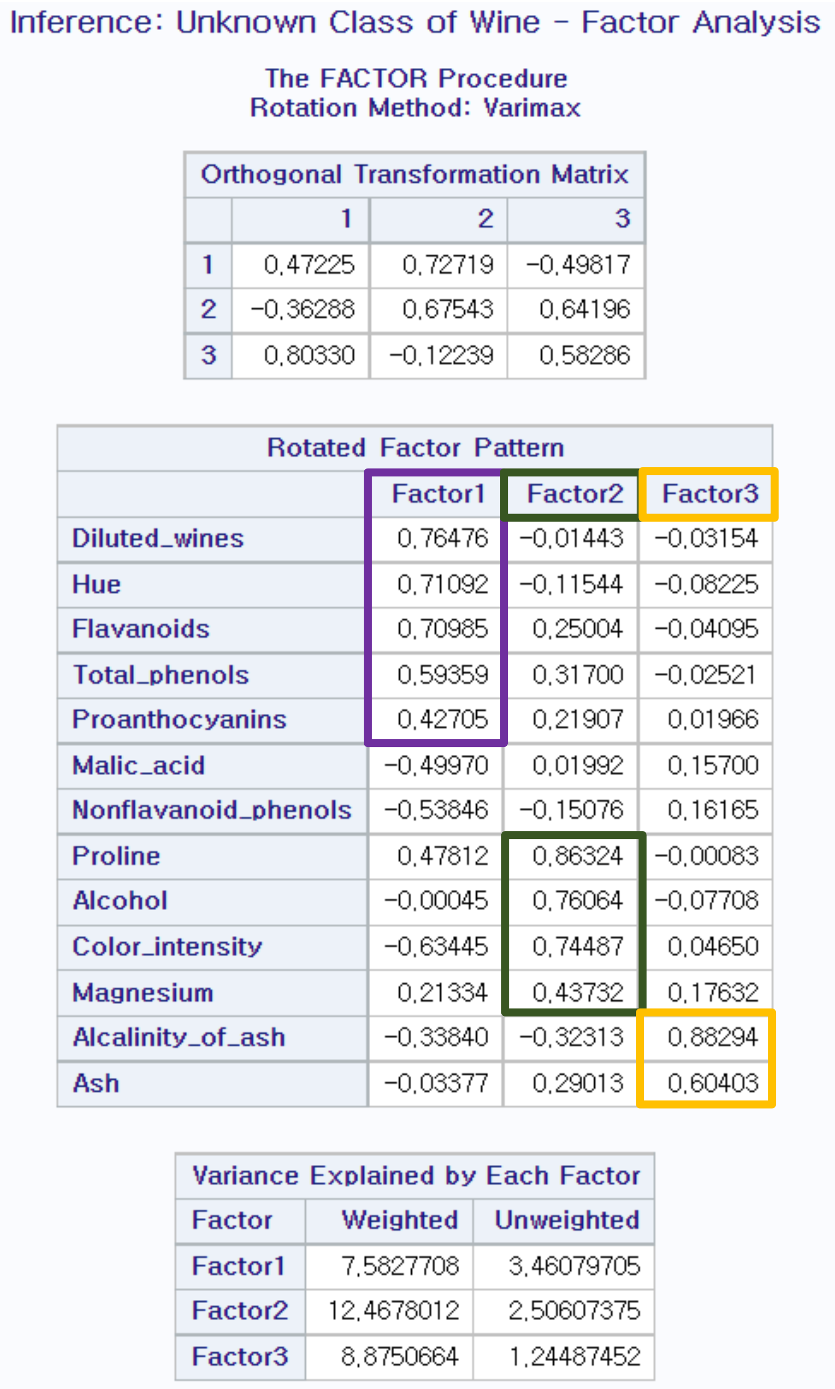


Final Communality Estimates and Variable Weights

Total Communality: Weighted = 28.925638 Unweighted = 7.211745

Variable	Communality	Weight
Alcohol	0.58450963	0.6590623
Malic_acid	0.27474145	1.2480154
Ash	0.45017143	0.0752646
Alcalinity_of_ash	0.99851725	11.1526862
Magnesium	0.26784859	2.0398934
Total_phenols	0.45346796	0.3916895
Flavanoids	0.56808978	0.9977187
Nonflavanoid_phenols	0.33880322	1.5488634
Proanthocyanins	0.23075228	0.3275947
Color_intensity	0.95951827	5.3744494
Hue	0.52549767	0.0522450
Diluted_wines	0.58605453	0.5040864
Proline	0.97377325	9.9166717

Table 5: Several outputs after rotated factor loadings



Inference: Unknown Class of Wine - Factor Analysis

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2	3
1	0.47225	0.72719	-0.49817
2	-0.36288	0.67543	0.64196
3	0.80330	-0.12239	0.58286

Rotated Factor Pattern

	Factor1	Factor2	Factor3
Diluted_wines	0.76476	-0.01443	-0.03154
Hue	0.71092	-0.11544	-0.08225
Flavanoids	0.70985	0.25004	-0.04095
Total_phenols	0.59359	0.31700	-0.02521
Proanthocyanins	0.42705	0.21907	0.01966
Malic_acid	-0.49970	0.01992	0.15700
Nonflavanoid_phenols	-0.53846	-0.15076	0.16165
Proline	0.47812	0.86324	-0.00083
Alcohol	-0.00045	0.76064	-0.07708
Color_intensity	-0.63445	0.74487	0.04650
Magnesium	0.21334	0.43732	0.17632
Alcalinity_of_ash	-0.33840	-0.32313	0.88294
Ash	-0.03377	0.29013	0.60403

Variance Explained by Each Factor

Factor	Weighted	Unweighted
Factor1	7.5827708	3.46079705
Factor2	12.4678012	2.50607375
Factor3	8.8750664	1.24487452

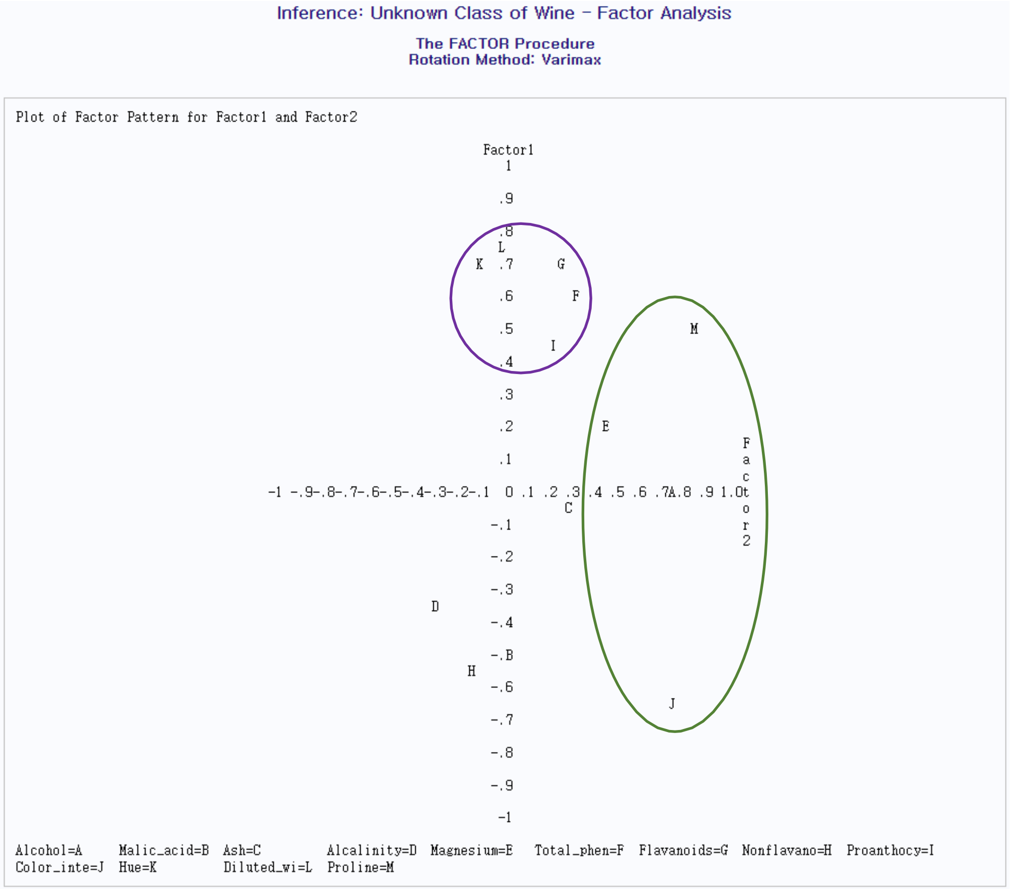
Now, we can add a name of each factor pattern. We will present a specific reason of the name for each factor pattern.

Factor pattern 1: OD280/OD315 of diluted wines, Hue, Flavanoids, Total phenols and Proantocyanins

→ **Factor 1: <u>Phytochemical group</u>**

Factor pattern 2: Proline, Alcohol, Color Intensity and Magnesium → **Factor 2: <u>Proline group</u>**

Factor pattern 3: Alkalinity of ash and Ash → **Factor 3: <u>Alkalinity of ash group</u>**



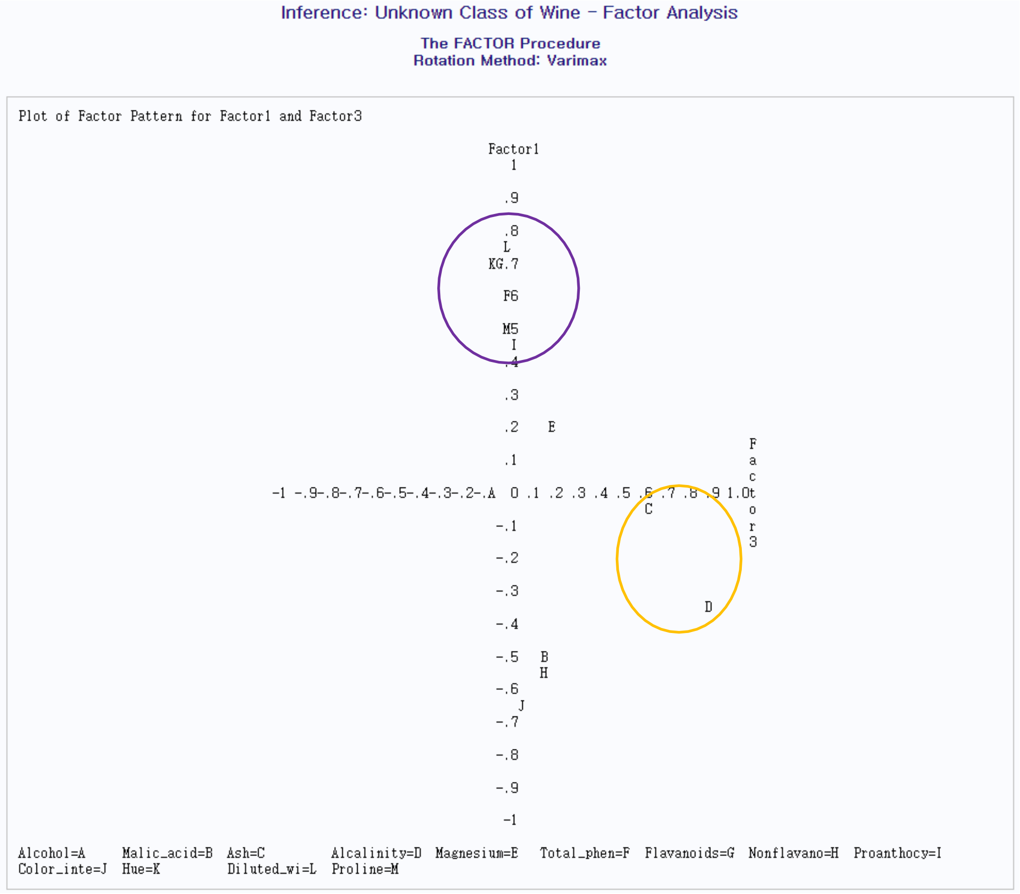
Inference: Unknown Class of Wine - Factor Analysis

The FACTOR Procedure
Rotation Method: Varimax

Plot of Factor Pattern for Factor1 and Factor2

```
Alcohol=A     Malic_acid=B  Ash=C         Alcalinity=D  Magnesium=E   Total_phen=F  Flavanoids=G  Nonflavano=H  Proanthocy=I
Color_inte=J  Hue=K         Diluted_wi=L  Proline=M
```

Figure 3: Plot for rotated factor loading (Factor 1 – Factor 3)

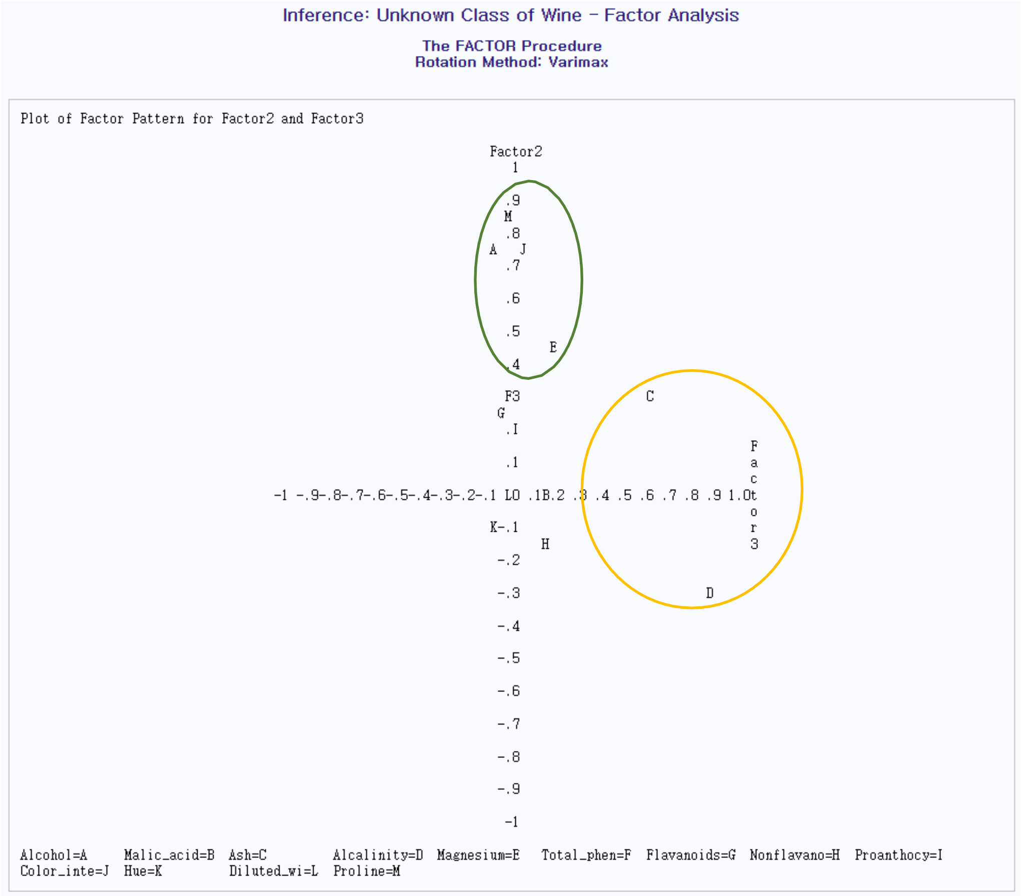


Inference: Unknown Class of Wine - Factor Analysis

The FACTOR Procedure
Rotation Method: Varimax

Plot of Factor Pattern for Factor1 and Factor3

```
Alcohol=A     Malic_acid=B  Ash=C         Alcalinity=D  Magnesium=E   Total_phen=F  Flavanoids=G  Nonflavano=H  Proanthocy=I
Color_inte=J  Hue=K         Diluted_wi=L  Proline=M
```



Inference: Unknown Class of Wine – Factor Analysis

The FACTOR Procedure
Rotation Method: Varimax

Plot of Factor Pattern for Factor2 and Factor3

```
Alcohol=A     Malic_acid=B  Ash=C         Alcalinity=D  Magnesium=E   Total_phen=F  Flavanoids=G  Nonflavano=H  Proanthocy=I
Color_inte=J  Hue=K         Diluted_wi=L  Proline=M
```

One thing what we want to present is that there is a problem in the communality table. Before we add a name of each factor pattern, we must discuss the communality problem. Here is the weighted communality table below. (See Table 6)

Table 6: Weighted communalities

Variable	Weighted communality	Proportion (Approx.)
Alcohol	0.3852282611	1%
Malic acid	0.3428815606	1%
Ash	0.0338819726	0.1%
Alkalinity of ash	11.13614955	38%
Magnesium	0.5463825709	2%
Total phenols	0.1776186385	0.7%
Flavanoids (Polyphenol)	0.5667937968	2%
Non-flavonoids (Polyphenol)	0.5247599073	2%
Proanthocyanins	0.0755932239	0.2%
Color intensive	5.15688239	18%
Hue	0.0274546258	0.1%
OD280/OD315 of diluted wines	0.2954221182	1%
Proline	9.656633742	33%
Total	**28.92568236**	**100%**

Let’s the greatest lower bound for the weighted communalities is generously 0.3 (30%). By the definition of communality in the Factor Analysis, communality means a proportion of variation for that variable explained by the factor loadings. When we check the weighted communality in the Table 6 above, we can see that all variables except ‘Alkalinity of ash’ and ‘Proline’ have a very low proportion of communality explained by all three factor loadings. In our perspective, all variables except ‘Alkalinity of ash’ and ‘Proline’ have very low proportion of communality, so we do not have to consider those variables. Based on the result of the weighted communality, when we check the table “Variance explained by each factor” in Table 5 shown above, we can also see that the factor loading 2 (Proline group) and the factor loading 3 (Alkalinity of ash group) have high variance in sequence. At this moment, for the further analyses, three options arise in this situation.

<Options for the next analyses>

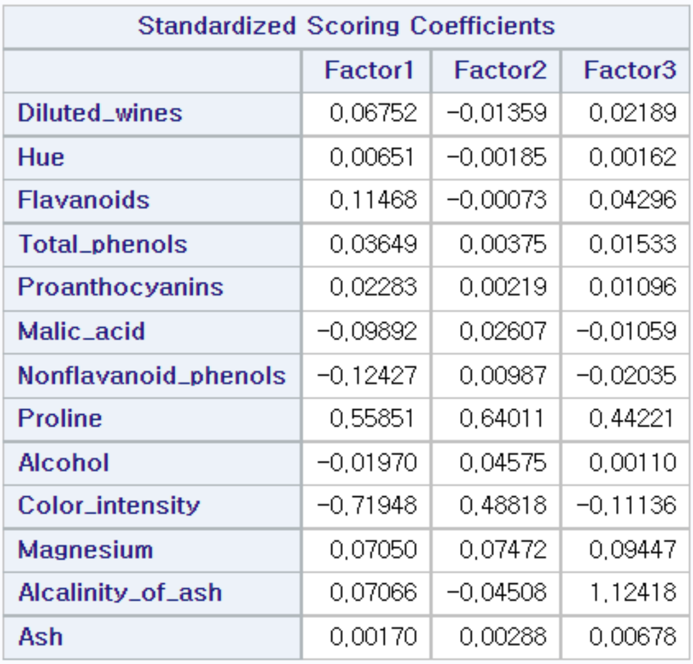
Option 1) Interpret three factor loadings and compute three factor scores and use scores for the next analysis as usual.

Option 2) Interpret two factor loadings (Factor 2 and Factor 3) with the variables which have a proportion of weighted communality greater than 0.3 (Consider ‘Alkalinity of ash’ and ‘Proline’ only) in each factor loading, compute factor scores respected to these two factor loadings and use scores as an input for the next analysis.

Option 3) Based on the weighted communality, all variables except ‘Alkalinity of ash’ and ‘Proline’ seem useless, so just use these two variables for the next analysis. It means that factor analysis with this dataset is meaningless.

We are going to take the **first** option and we are also going to explore the **third** option as a special case. For the first option, we need to compute the factor scores corresponded to each factor loading. (See Table 7) The scores in the table have been estimated by the regression method. We are going to use the first factor score (The phytochemical group), the second factor score (The Proline group) and the third factor score (The alkalinity of ash group) as an input of the Canonical discriminant analysis and the Logistic regression analysis soon.

Table 7: Factor scores estimated by the regression method for the next steps (Option 1)



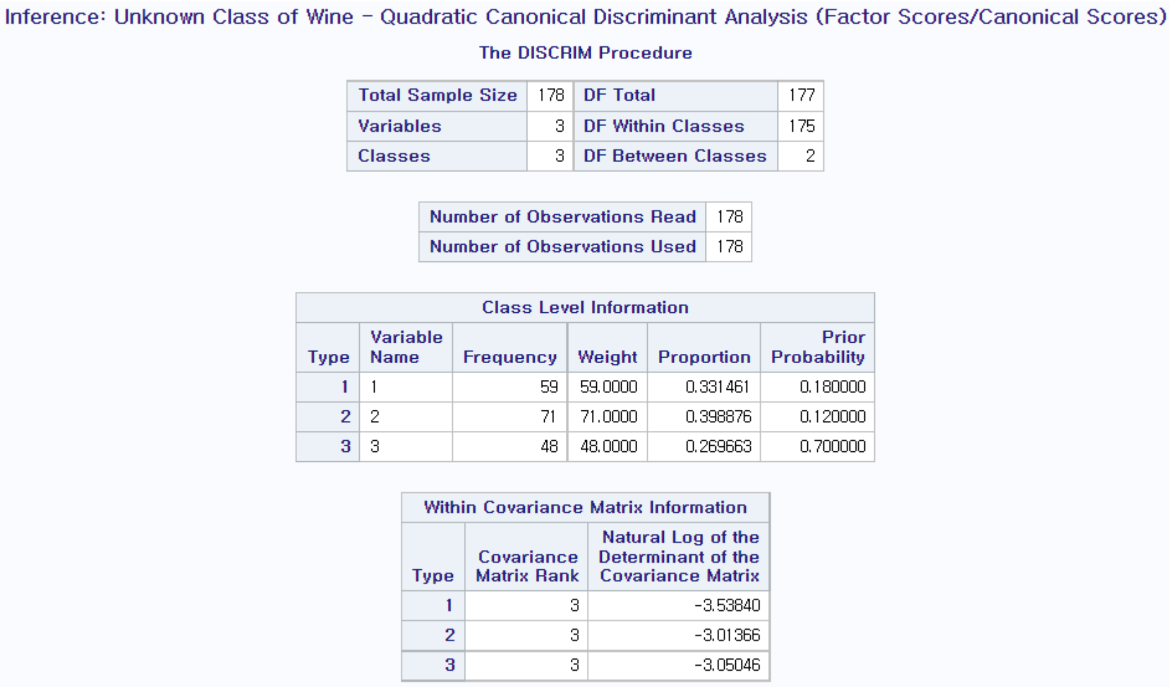
Standardized Scoring Coefficients

	Factor1	Factor2	Factor3
Diluted_wines	0,06752	-0,01359	0,02189
Hue	0,00651	-0,00185	0,00162
Flavanoids	0,11468	-0,00073	0,04296
Total_phenols	0,03649	0,00375	0,01533
Proanthocyanins	0,02283	0,00219	0,01096
Malic_acid	-0,09892	0,02607	-0,01059
Nonflavanoid_phenols	-0,12427	0,00987	-0,02035
Proline	0,55851	0,64011	0,44221
Alcohol	-0,01970	0,04575	0,00110
Color_intensity	-0,71948	0,48818	-0,11136
Magnesium	0,07050	0,07472	0,09447
Alcalinity_of_ash	0,07066	-0,04508	1,12418
Ash	0,00170	0,00288	0,00678

3. Multivariate Analysis (B. Canonical Discriminant Analysis)

Through this Canonical discriminant analysis, we are going to use the factor scores and then do classification of each class using new canonical scores for our final objective of this analysis. Here is our descriptive statistic table for the Canonical discriminant analysis using factor scores. (See Table 8)

Table 8: Descriptive statistics (Factor scores)



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Factor Scores/Canonical Scores)

The DISCRIM Procedure

Total Sample Size	178	DF Total	177
Variables	3	DF Within Classes	175
Classes	3	DF Between Classes	2

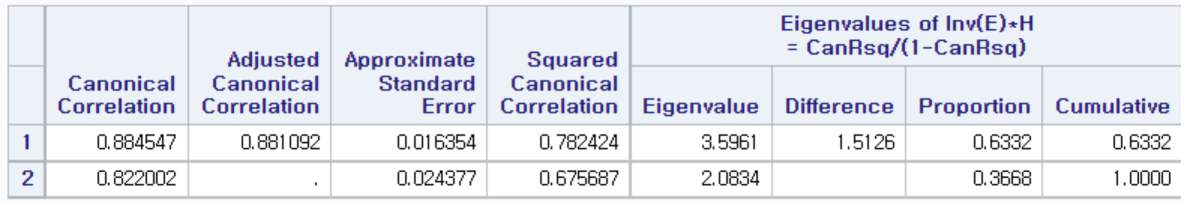
Number of Observations Read	178
Number of Observations Used	178

Class Level Information					
Type	**Variable Name**	**Frequency**	**Weight**	**Proportion**	**Prior Probability**
1	1	59	59.0000	0.331461	0.180000
2	2	71	71.0000	0.398876	0.120000
3	3	48	48.0000	0.269663	0.700000

Within Covariance Matrix Information		
Type	**Covariance Matrix Rank**	**Natural Log of the Determinant of the Covariance Matrix**
1	3	-3.53840
2	3	-3.01366
3	3	-3.05046

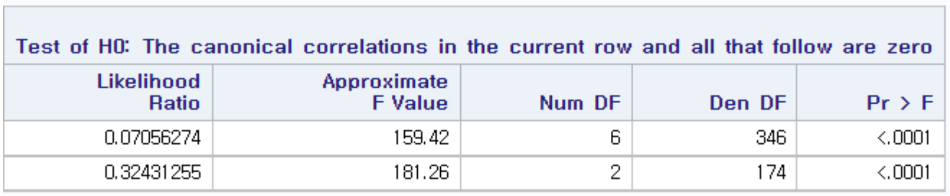
We have put the best prior ($P_1 = 0.18,\ P_2 = 0.12$ and $P_3 = 0.7$) to minimize the total error rate. Based on these statistics, we need to check canonical correlations for two canonical scores. (See Table 9 and 10)

Table 9: Canonical correlations for two canonical scores 1



					Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalue	Difference	Proportion	Cumulative
1	0.884547	0.881092	0.016354	0.782424	3.5961	1.5126	0.6332	0.6332
2	0.822002	.	0.024377	0.675687	2.0834		0.3668	1.0000

Table 10: Canonical correlations for two canonical scores 2

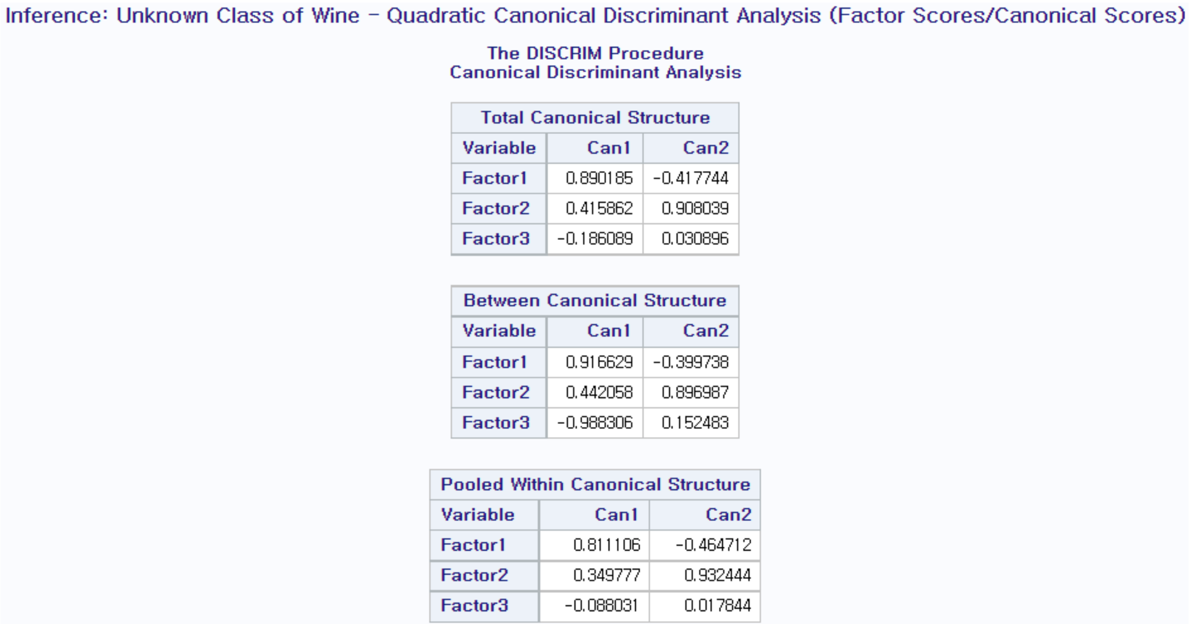


Test of H0: The canonical correlations in the current row and all that follow are zero				
Likelihood Ratio	**Approximate F Value**	**Num DF**	**Den DF**	**Pr > F**
0.07056274	159.42	6	346	<.0001
0.32431255	181.26	2	174	<.0001

We can check both canonical scores have high correlation (80%~90%) with each class of wine. These tables are also presenting that the first and second canonical variable (Can1 and Can2) explained by the first and second eigenvalue accounts for 63% and 37% of the variation using three factor scores in each class of wine.

Since square of canonical correlation for each canonical score is equivalent to the variation explains difference in each class of wine, we can say the first and second canonical score gives 78% and 68% variation respectively for the difference in each class of wine. At the end column of the table, we can see that both canonical scores are significant. For now, we should compute the canonical structures like below. (See Table 11)

Table 11: Canonical structure loadings (Factor scores)



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Factor Scores/Canonical Scores)

The DISCRIM Procedure
Canonical Discriminant Analysis

Total Canonical Structure

Variable	Can1	Can2
Factor1	0.890185	-0.417744
Factor2	0.415862	0.908039
Factor3	-0.186089	0.030896

Between Canonical Structure

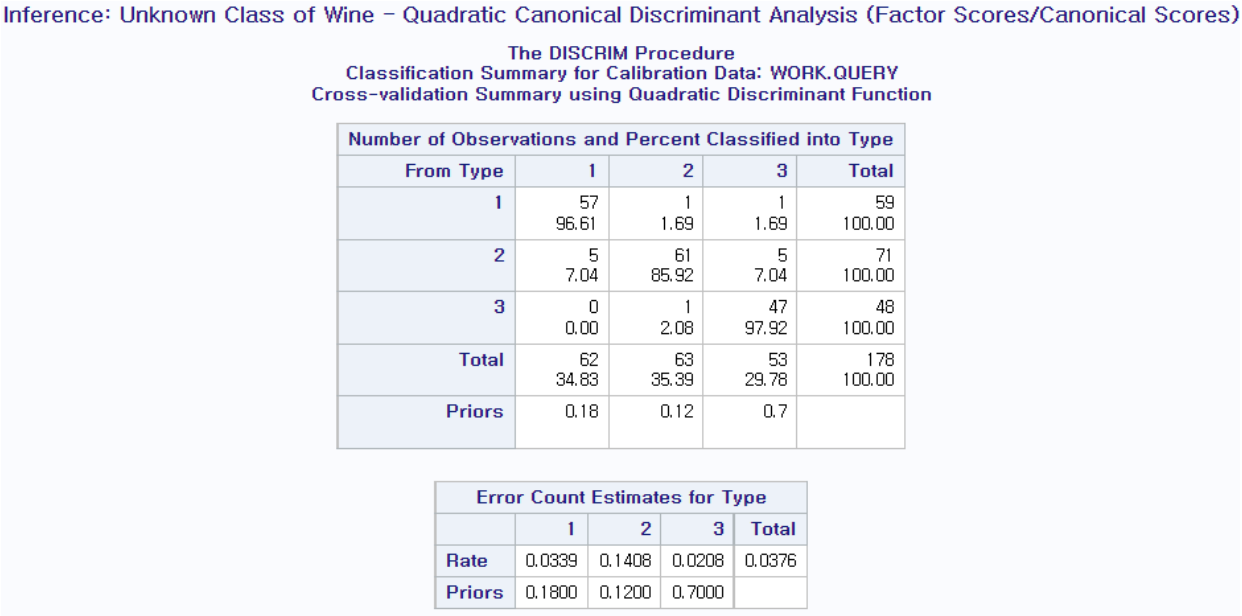
Variable	Can1	Can2
Factor1	0.916629	-0.399738
Factor2	0.442058	0.896987
Factor3	-0.988306	0.152483

Pooled Within Canonical Structure

Variable	Can1	Can2
Factor1	0.811106	-0.464712
Factor2	0.349777	0.932444
Factor3	-0.088031	0.017844

We can check that the first canonical score looks heavily weighted on the first factor score (Phytochemical group) and the second canonical score looks heavily weighted on the second factor score (Proline). We are ready to check the result of classification. There are two classification results and the scatter plot for each class of wine using factor scores and the original wine class first. (See Table 12, 13 and Figure 5)

Table 12: Classification result (Factor score – Original wine class)



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Factor Scores/Canonical Scores)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.QUERY
Cross-validation Summary using Quadratic Discriminant Function

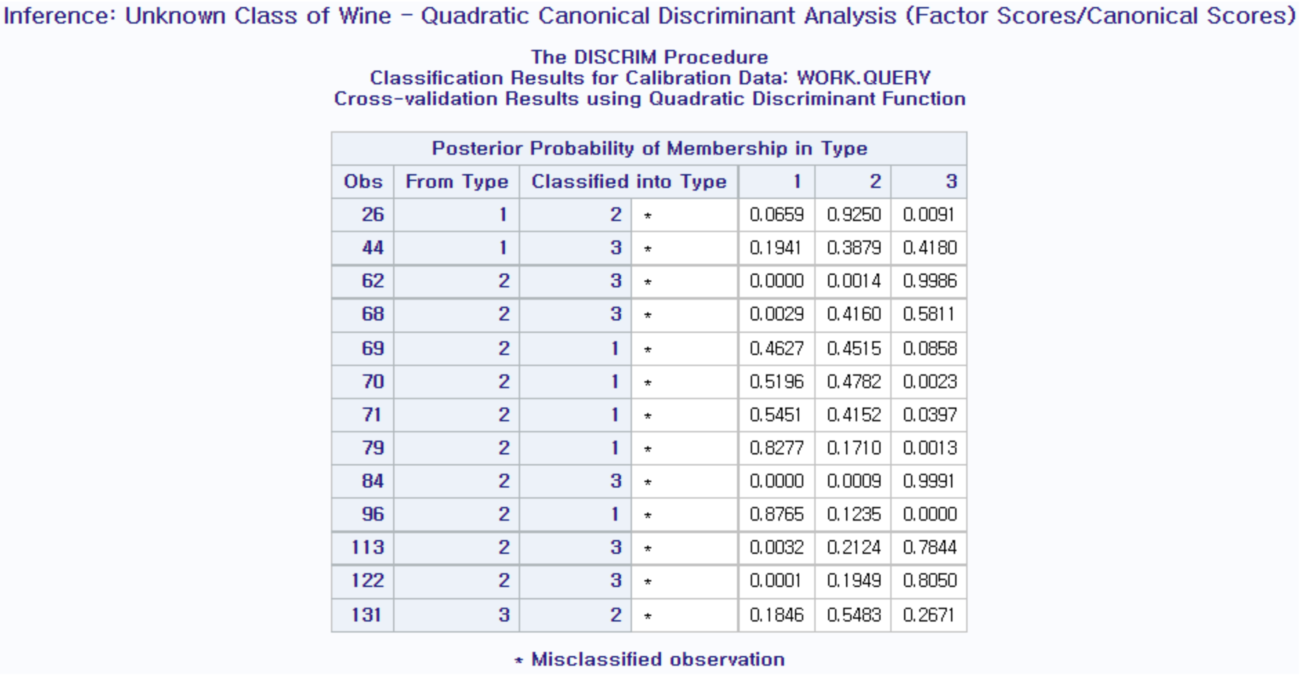
Number of Observations and Percent Classified into Type

From Type	1	2	3	Total
1	57 96.61	1 1.69	1 1.69	59 100.00
2	5 7.04	61 85.92	5 7.04	71 100.00
3	0 0.00	1 2.08	47 97.92	48 100.00
Total	62 34.83	63 35.39	53 29.78	178 100.00
Priors	0.18	0.12	0.7	

Error Count Estimates for Type

	1	2	3	Total
Rate	0.0339	0.1408	0.0208	0.0376
Priors	0.1800	0.1200	0.7000	

Table 13: Misclassification result (Factor score – Original wine class)



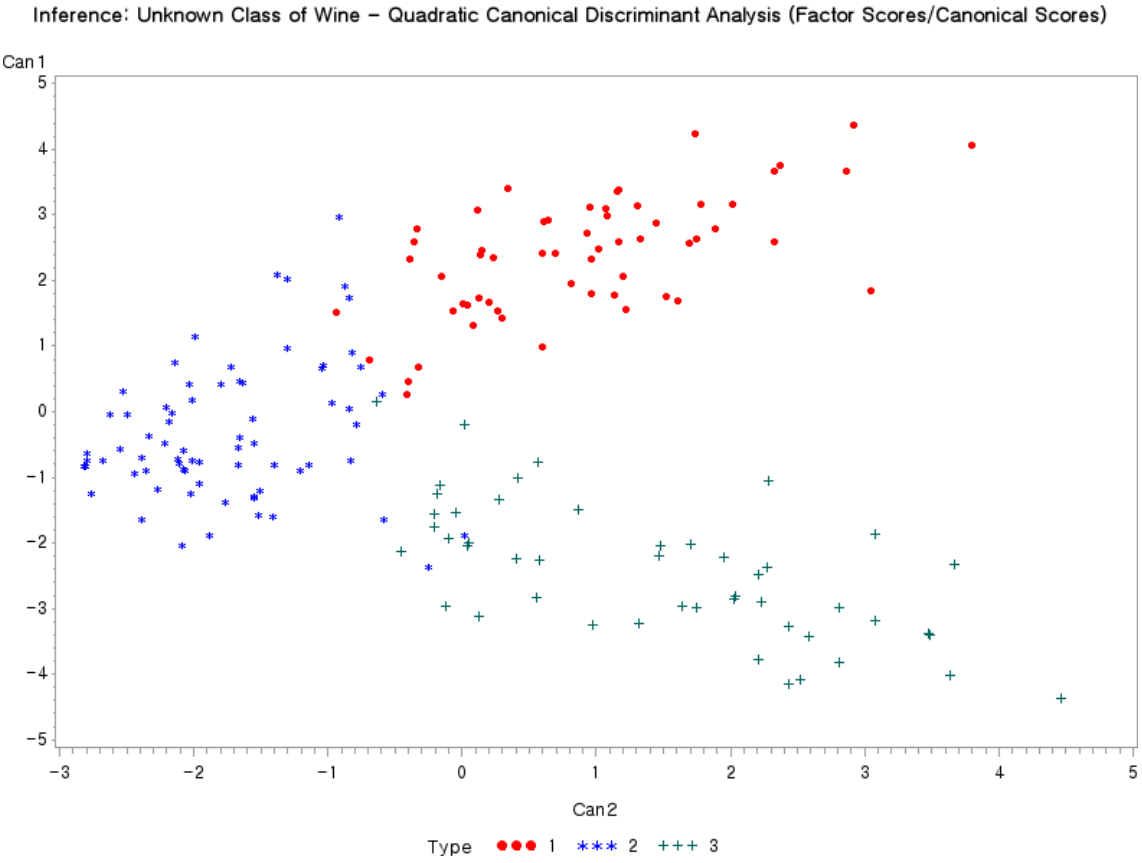
Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Factor Scores/Canonical Scores)

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.QUERY
Cross-validation Results using Quadratic Discriminant Function

Posterior Probability of Membership in Type						
Obs	From Type	Classified into Type		1	2	3
26	1	2	*	0.0659	0.9250	0.0091
44	1	3	*	0.1941	0.3879	0.4180
62	2	3	*	0.0000	0.0014	0.9986
68	2	3	*	0.0029	0.4160	0.5811
69	2	1	*	0.4627	0.4515	0.0858
70	2	1	*	0.5196	0.4782	0.0023
71	2	1	*	0.5451	0.4152	0.0397
79	2	1	*	0.8277	0.1710	0.0013
84	2	3	*	0.0000	0.0009	0.9991
96	2	1	*	0.8765	0.1235	0.0000
113	2	3	*	0.0032	0.2124	0.7844
122	2	3	*	0.0001	0.1949	0.8050
131	3	2	*	0.1846	0.5483	0.2671

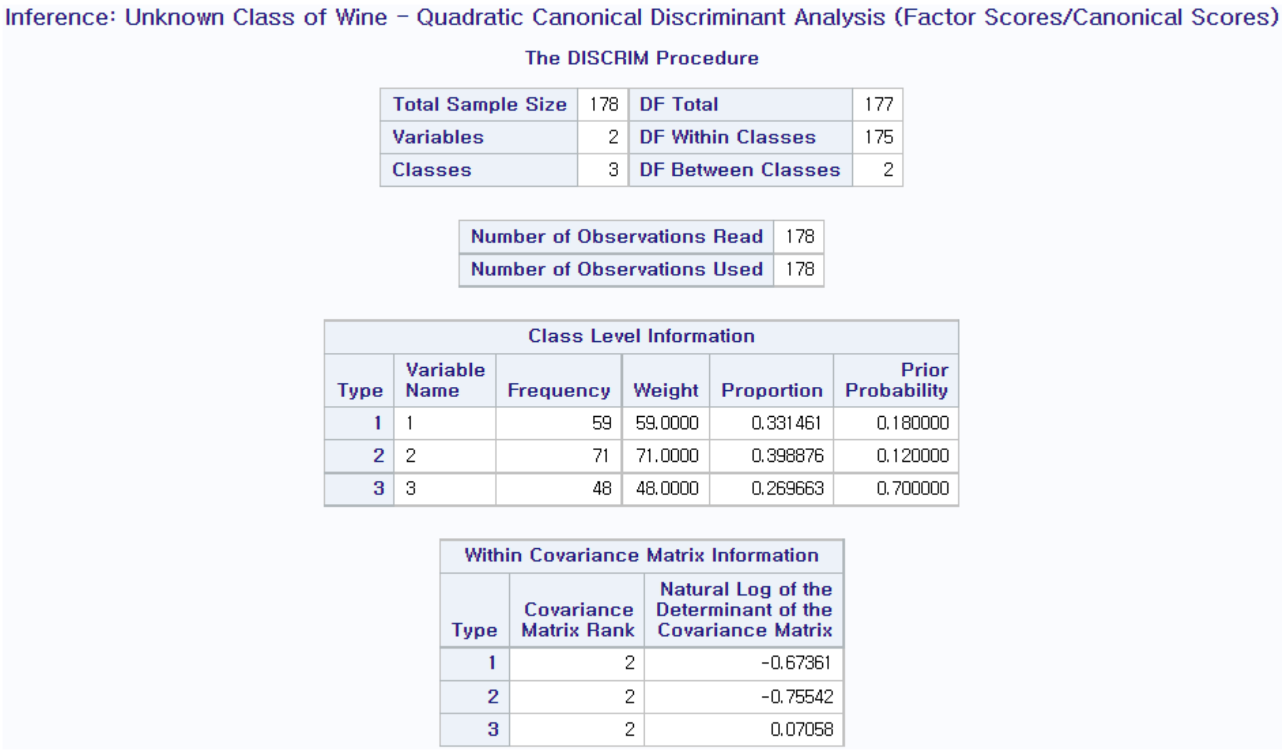
* Misclassified observation

Figure 5: Classification scatter plot (Factor score – Original wine class)



We can see that the total error count ratio is approximately 0.4 (4%) which means our dataset is already well-classified and we can also check that the class 1 and 3 look independent each other. Now, we are going to do one more same test using canonical scores to figure out whether using canonical scores gives a good performance classifying class of wine rather than using factor scores. Here is our descriptive statistic table for the Canonical discriminant analysis using canonical scores. (See Table 14)

Table 14: Descriptive statistics (Canonical scores)



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Factor Scores/Canonical Scores)

The DISCRIM Procedure

Total Sample Size	178	DF Total	177
Variables	2	DF Within Classes	175
Classes	3	DF Between Classes	2

Number of Observations Read	178
Number of Observations Used	178

Class Level Information

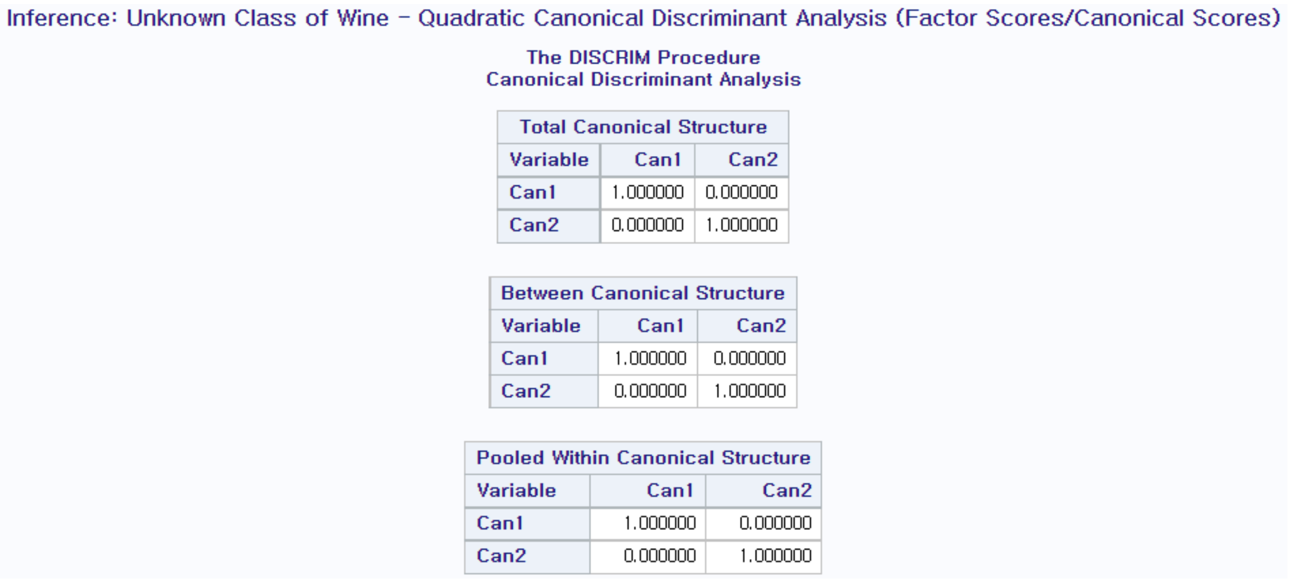
Type	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	1	59	59.0000	0.331461	0.180000
2	2	71	71.0000	0.398876	0.120000
3	3	48	48.0000	0.269663	0.700000

Within Covariance Matrix Information

Type	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	2	-0.67361
2	2	-0.75542
3	2	0.07058

We need to check the canonical structure loadings just in case. (See Table 15 and 16)

Table 15: Canonical structure loadings (Canonical variables)



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Factor Scores/Canonical Scores)

The DISCRIM Procedure
Canonical Discriminant Analysis

Total Canonical Structure

Variable	Can1	Can2
Can1	1.000000	0.000000
Can2	0.000000	1.000000

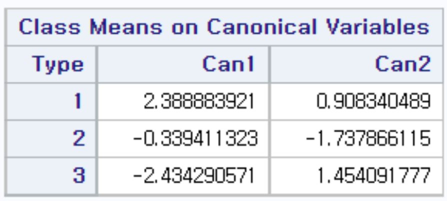
Between Canonical Structure

Variable	Can1	Can2
Can1	1.000000	0.000000
Can2	0.000000	1.000000

Pooled Within Canonical Structure

Variable	Can1	Can2
Can1	1.000000	0.000000
Can2	0.000000	1.000000

Table 16: Class means on canonical variables (Canonical variables)

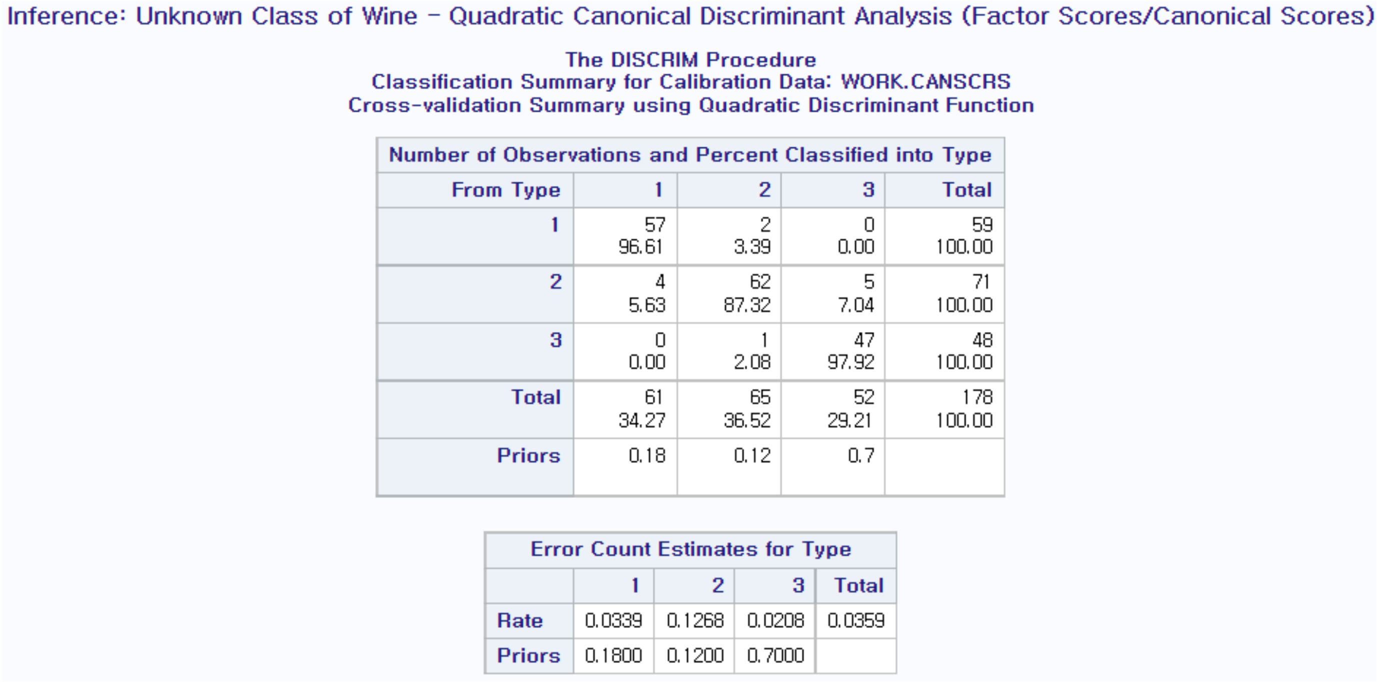


Class Means on Canonical Variables

Type	Can1	Can2
1	2.388883921	0.908340489
2	-0.339411323	-1.737866115
3	-2.434290571	1.454091777

From the Table 16, we can see how much each class of wine is weighted in each canonical variable. There are two classification results and the scatter plot for each class of wine using canonical scores and the wine class which is classified by factor scores. (See Table 17, 18, 19 and Figure 6)

Table 17: Classification result (Canonical score – Classified wine class by the factor score)



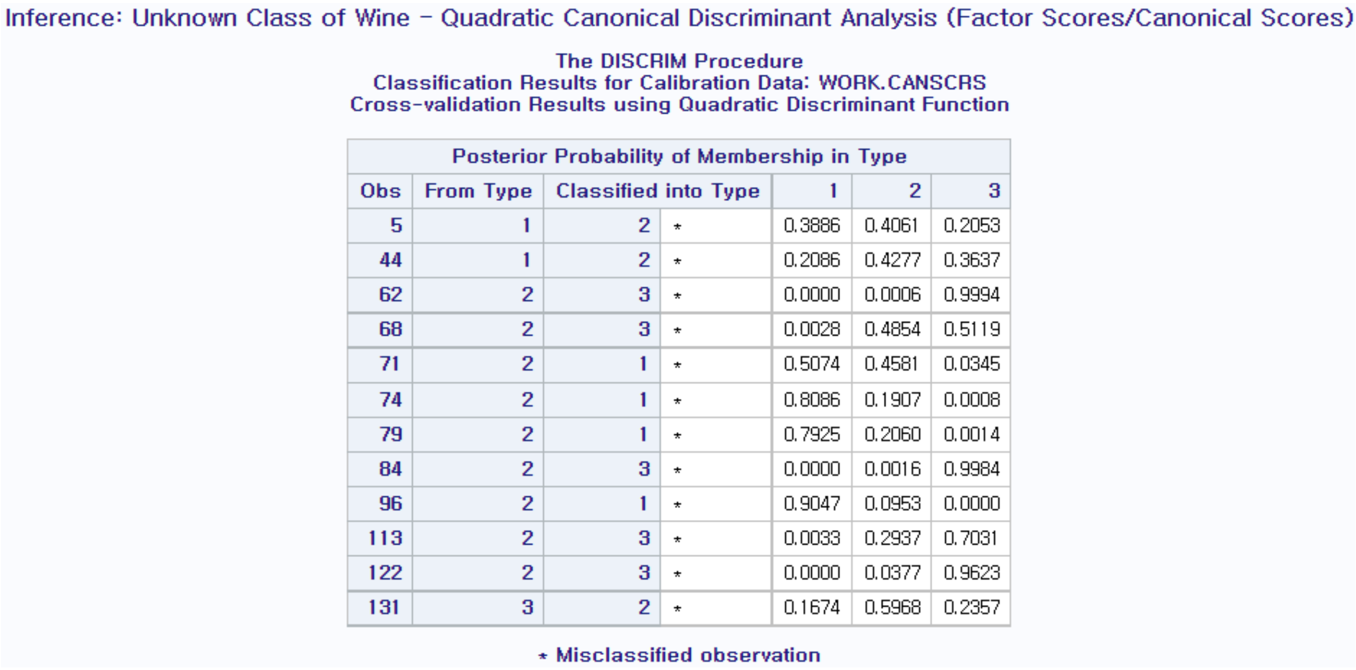
Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Factor Scores/Canonical Scores)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CANSCRS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Type				
From Type	1	2	3	Total
1	57 96.61	2 3.39	0 0.00	59 100.00
2	4 5.63	62 87.32	5 7.04	71 100.00
3	0 0.00	1 2.08	47 97.92	48 100.00
Total	61 34.27	65 36.52	52 29.21	178 100.00
Priors	0.18	0.12	0.7	

Error Count Estimates for Type				
	1	2	3	Total
Rate	0.0339	0.1268	0.0208	0.0359
Priors	0.1800	0.1200	0.7000	

Table 18: Misclassification result (Canonical variable – Classified wine class by the factor score)



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Factor Scores/Canonical Scores)

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.CANSCRS
Cross-validation Results using Quadratic Discriminant Function

Posterior Probability of Membership in Type						
Obs	From Type	Classified into Type		1	2	3
5	1	2	*	0.3886	0.4061	0.2053
44	1	2	*	0.2086	0.4277	0.3637
62	2	3	*	0.0000	0.0006	0.9994
68	2	3	*	0.0028	0.4854	0.5119
71	2	1	*	0.5074	0.4581	0.0345
74	2	1	*	0.8086	0.1907	0.0008
79	2	1	*	0.7925	0.2060	0.0014
84	2	3	*	0.0000	0.0016	0.9984
96	2	1	*	0.9047	0.0953	0.0000
113	2	3	*	0.0033	0.2937	0.7031
122	2	3	*	0.0000	0.0377	0.9623
131	3	2	*	0.1674	0.5968	0.2357

* Misclassified observation

Figure 6: Classification scatter plot (Canonical variable – Classified wine class by the factor score)

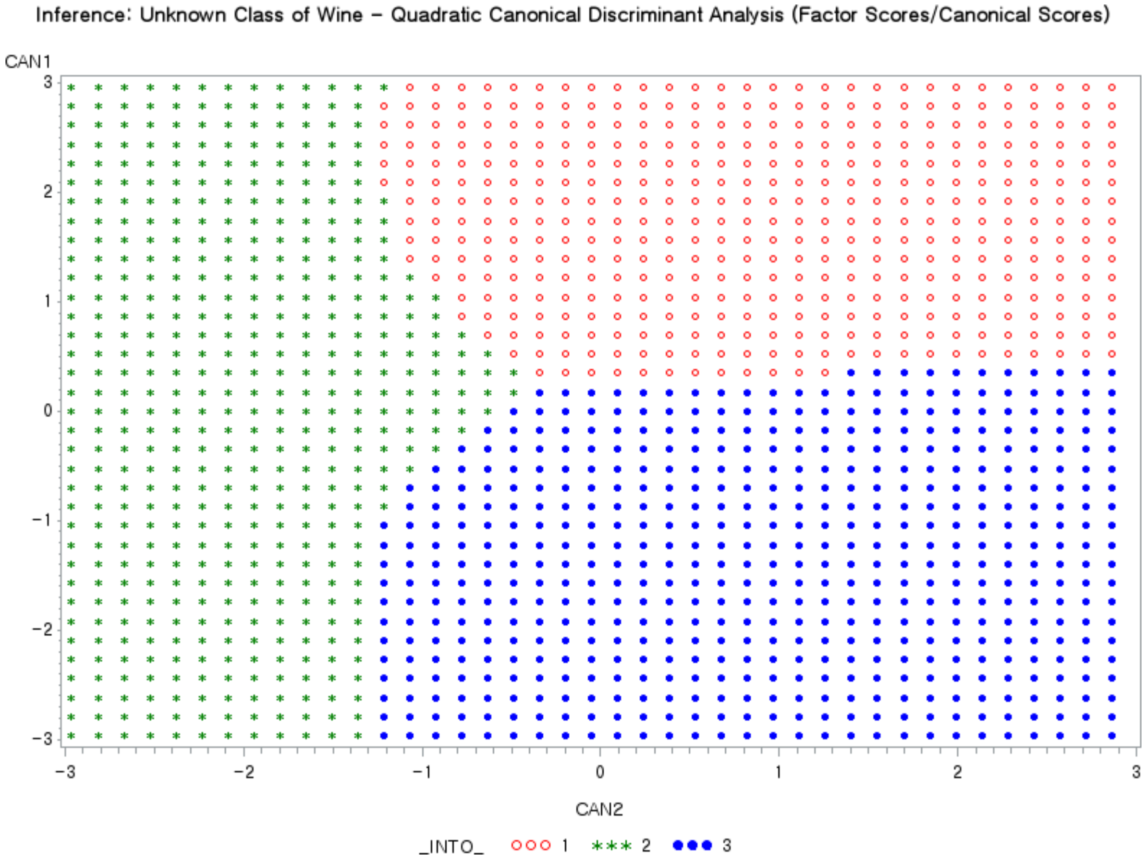
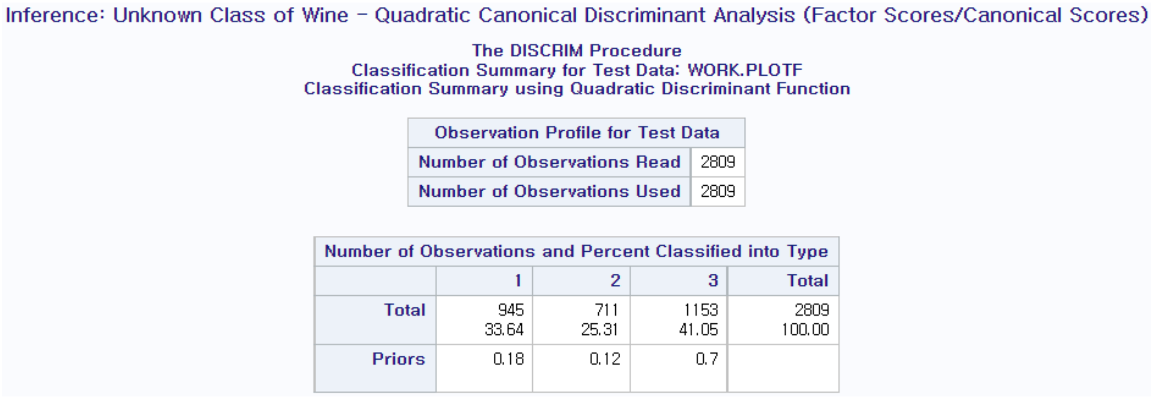


Table 19: Classification Summary



Inference: Unknown Class of Wine – Quadratic Canonical Discriminant Analysis (Factor Scores/Canonical Scores)

The DISCRIM Procedure
Classification Summary for Test Data: WORK.PLOTF
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	2809
Number of Observations Used	2809

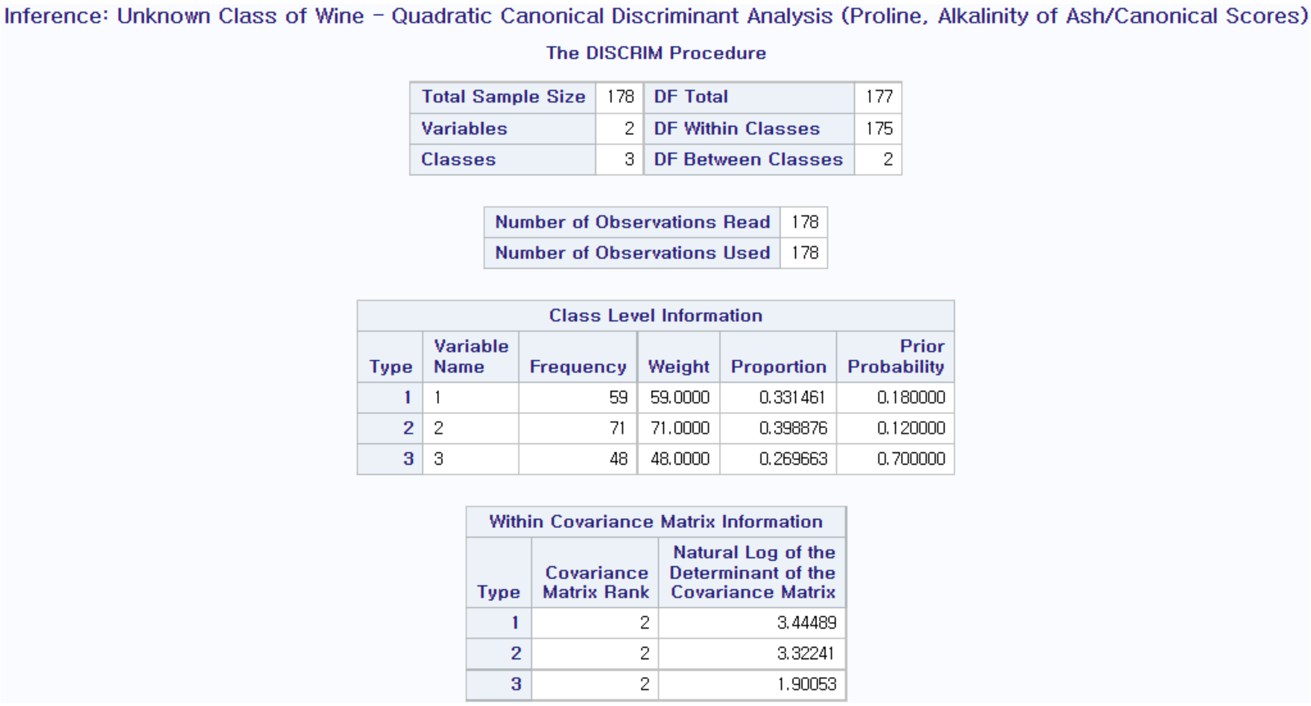
Number of Observations and Percent Classified into Type				
	1	2	3	Total
Total	945 33.64	711 25.31	1153 41.05	2809 100.00
Priors	0.18	0.12	0.7	

The result shows that the total error count ratio has been reduced approximately 0.002 (0.2%) which means there is not much difference between the factor score and the canonical variable. When we check the Table 16 and Figure 6, Type 1 and Type 3 are classified by the first canonical variable. In other words, by the phytochemical compounds, the first and third type of wine can be distinguished. Type 1 seems like more containing a relatively large amount of phytochemical compounds than Type 3. There is no way to distinguish Type 2 by phytochemical compounds. However, when we consider the second canonical loading which is heavily weighted on one of the amino acid (Proline), the first type of wine and the third type of wine seem containing large amount of Proline than Type 2.

<**Special Case**: Proline & Alkalinity of ash>

Based on the result of the factor analysis, we have checked all variables except Proline and Alkalinity of ash have a very low communality. It means that Proline and Alkalinity of ash are the most important variables to explain difference among three types of wine and that also means the other variables are useless. Through this special analysis, we are going to use two main variables (Proline and Alkalinity of ash) only to compare the result of canonical discriminant analysis with factor scores to the result of same analysis with two important variables. We are going to do same processes as we have done before and present several essential results. (See Table 21, 22 and 23)

Table 21: Descriptive statistics (Proline & Alkalinity of ash)



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Proline, Alkalinity of Ash/Canonical Scores)

The DISCRIM Procedure

Total Sample Size	178	DF Total	177
Variables	2	DF Within Classes	175
Classes	3	DF Between Classes	2

Number of Observations Read	178
Number of Observations Used	178

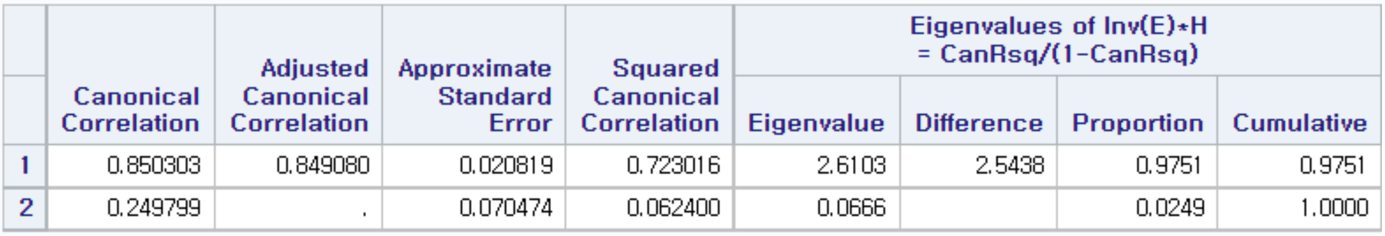
Class Level Information

Type	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	1	59	59.0000	0.331461	0.180000
2	2	71	71.0000	0.398876	0.120000
3	3	48	48.0000	0.269663	0.700000

Within Covariance Matrix Information

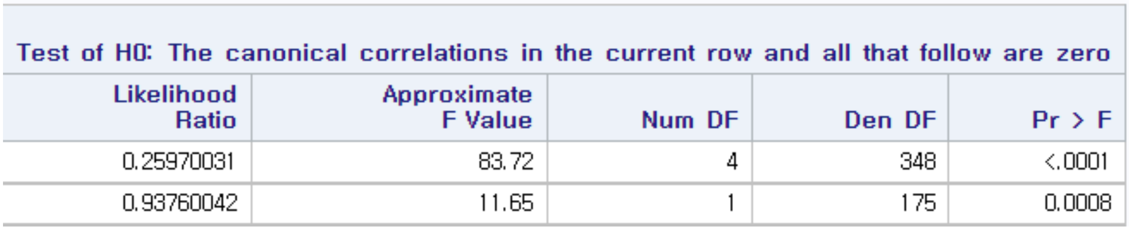
Type	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	2	3.44489
2	2	3.32241
3	2	1.90053

Table 22: Canonical correlations for two canonical scores 1



					Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalue	Difference	Proportion	Cumulative
1	0.850303	0.849080	0.020819	0.723016	2.6103	2.5438	0.9751	0.9751
2	0.249799	.	0.070474	0.062400	0.0666		0.0249	1.0000

Table 23: Canonical correlations for two canonical scores 2

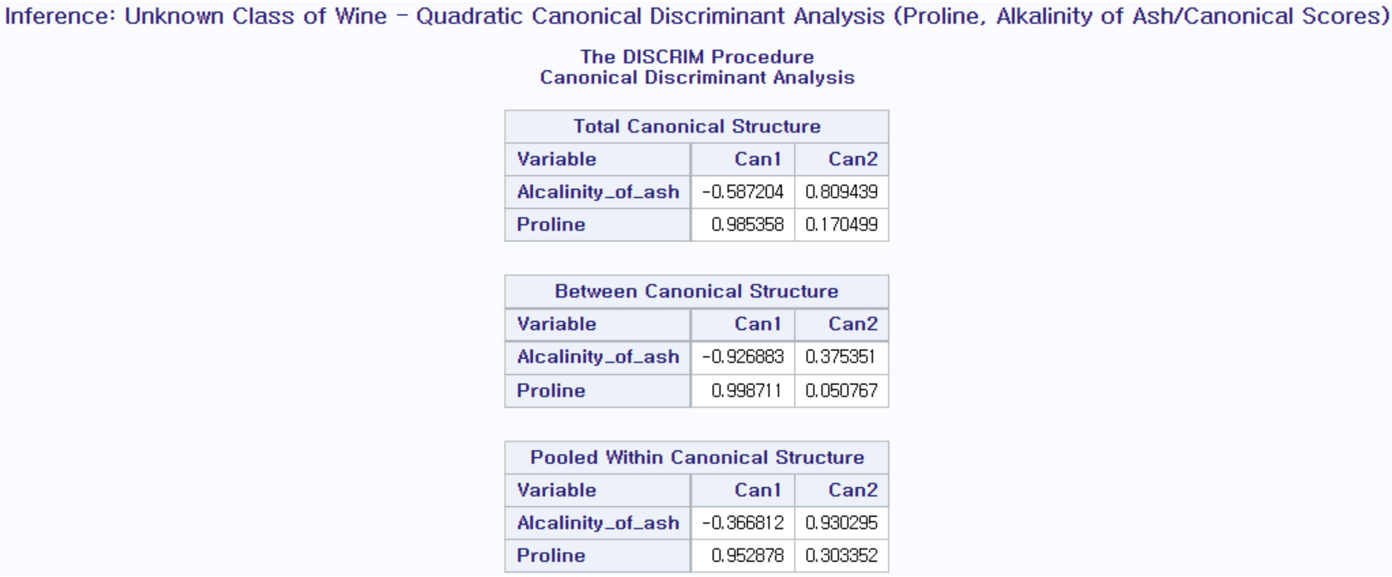


Test of H0: The canonical correlations in the current row and all that follow are zero

Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.25970031	83.72	4	348	<.0001
0.93760042	11.65	1	175	0.0008

We can check the first canonical score is significant and has high correlation (85%) with each class of wine. Also, the first canonical variable (Can1) explained by the first eigenvalue accounts for 98% of the variation using two important variables (Proline and Alkalinity of Ash) in each class of wine. We can check the first canonical variable gives 72% of the variation for the difference in each class of wine. On the contrary, although the second canonical score is significant, we can see that the second score looks containing not enough explanatory power. Let's check what the first canonical variable is. (See Table 24)

Table 24: Canonical structure loadings (Proline & Alkalinity of ash)



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Proline, Alkalinity of Ash/Canonical Scores)

The DISCRIM Procedure
Canonical Discriminant Analysis

Total Canonical Structure

Variable	Can1	Can2
Alcalinity_of_ash	-0.587204	0.809439
Proline	0.985358	0.170499

Between Canonical Structure

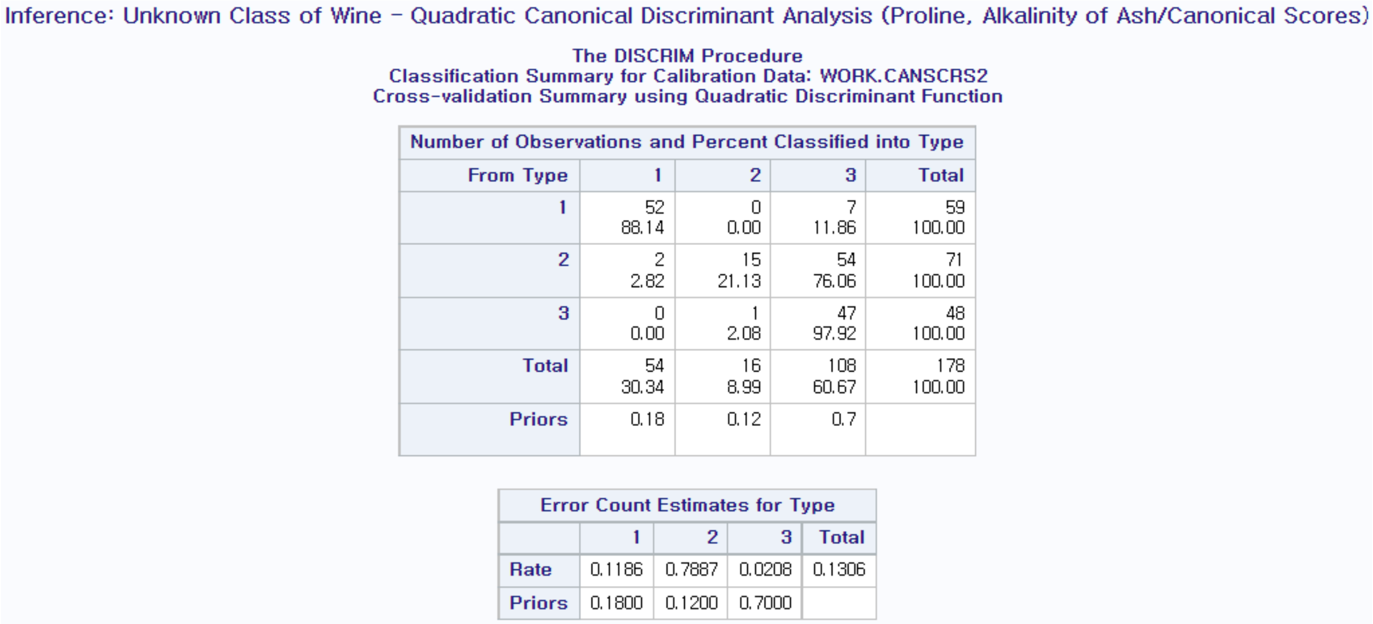
Variable	Can1	Can2
Alcalinity_of_ash	-0.926883	0.375351
Proline	0.998711	0.050767

Pooled Within Canonical Structure

Variable	Can1	Can2
Alcalinity_of_ash	-0.366812	0.930295
Proline	0.952878	0.303352

We can check that the first canonical score looks heavily weighted on proline and the second canonical score looks heavily weighted on alkalinity of ash. We are ready to check the result of classification using canonical variables and the classified result by proline and alkalinity of ash (See Table 25, 26, 27 and Figure 7)

Table 25: Classification result (Canonical variable – Classified class by Proline & Alkalinity of ash)



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Proline, Alkalinity of Ash/Canonical Scores)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CANSCRS2
Cross-validation Summary using Quadratic Discriminant Function

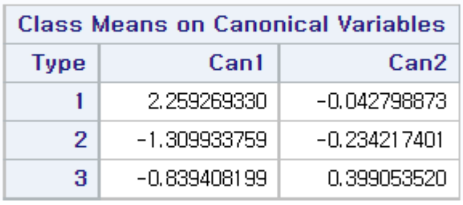
Number of Observations and Percent Classified into Type

From Type	1	2	3	Total
1	52 88.14	0 0.00	7 11.86	59 100.00
2	2 2.82	15 21.13	54 76.06	71 100.00
3	0 0.00	1 2.08	47 97.92	48 100.00
Total	54 30.34	16 8.99	108 60.67	178 100.00
Priors	0.18	0.12	0.7	

Error Count Estimates for Type

	1	2	3	Total
Rate	0.1186	0.7887	0.0208	0.1306
Priors	0.1800	0.1200	0.7000	

Table 26: Class means on canonical variables (Canonical variables)



Class Means on Canonical Variables		
Type	Can1	Can2
1	2.259269330	-0.042798873
2	-1.309933759	-0.234217401
3	-0.839408199	0.399053520

Figure 7: Plot of each type of wine (Canonical variable – Classified class by Proline & Alk. of ash)

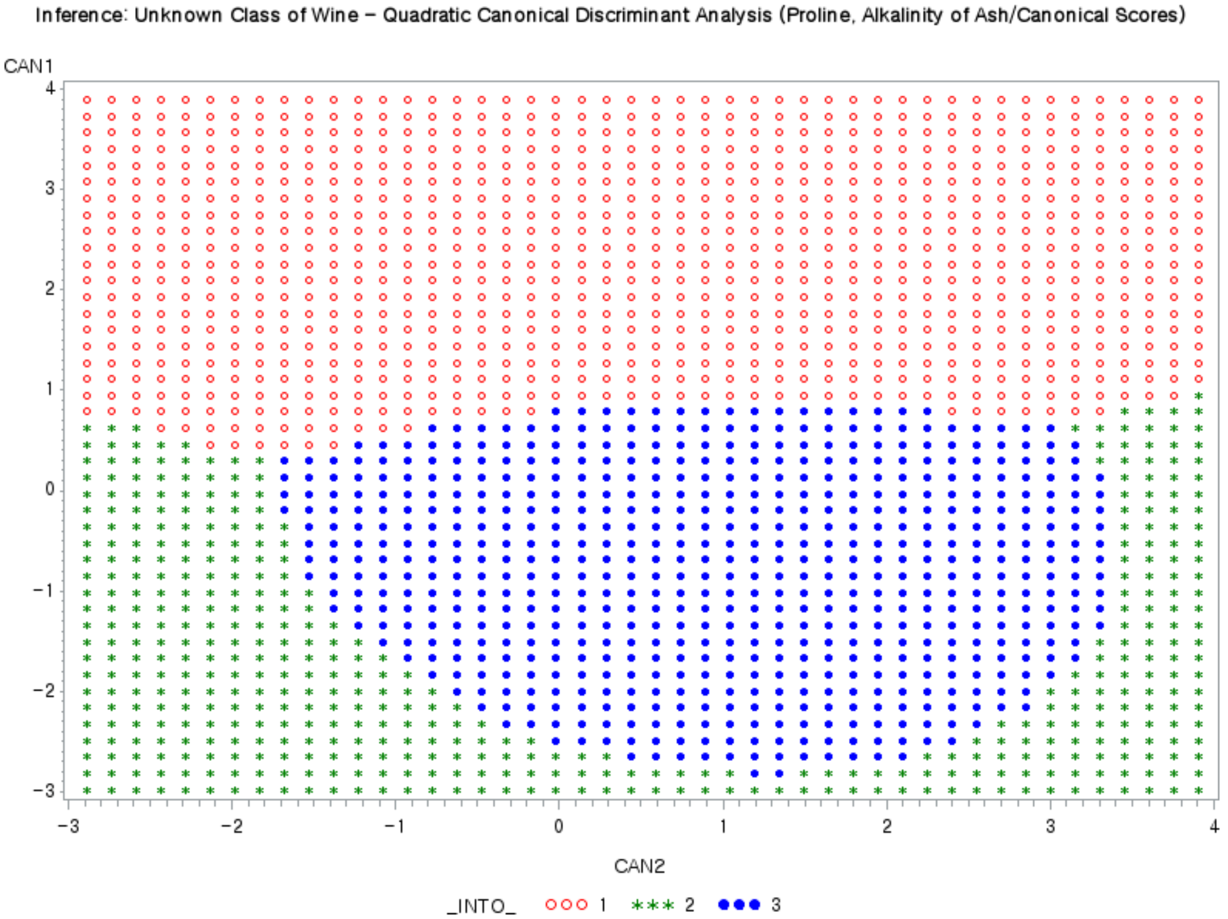
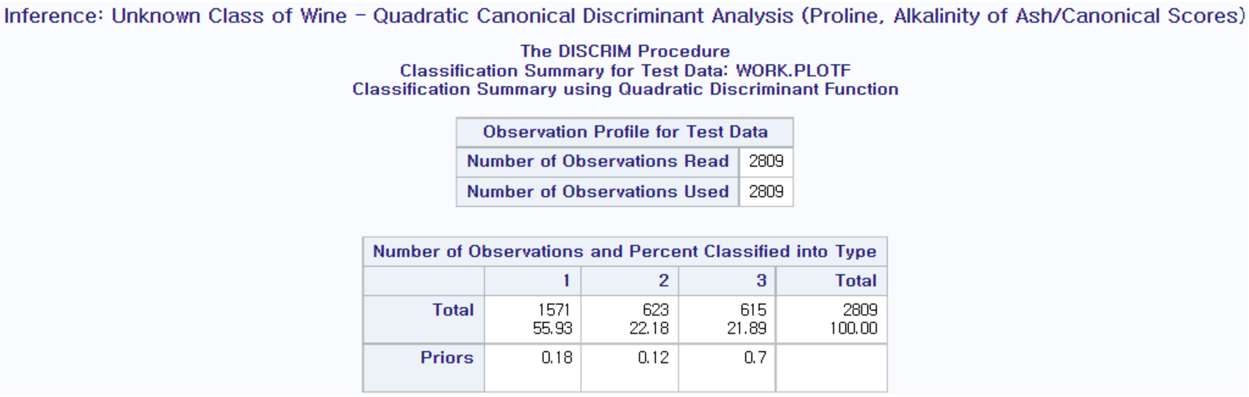


Table 27: Classification Summary



Inference: Unknown Class of Wine - Quadratic Canonical Discriminant Analysis (Proline, Alkalinity of Ash/Canonical Scores)

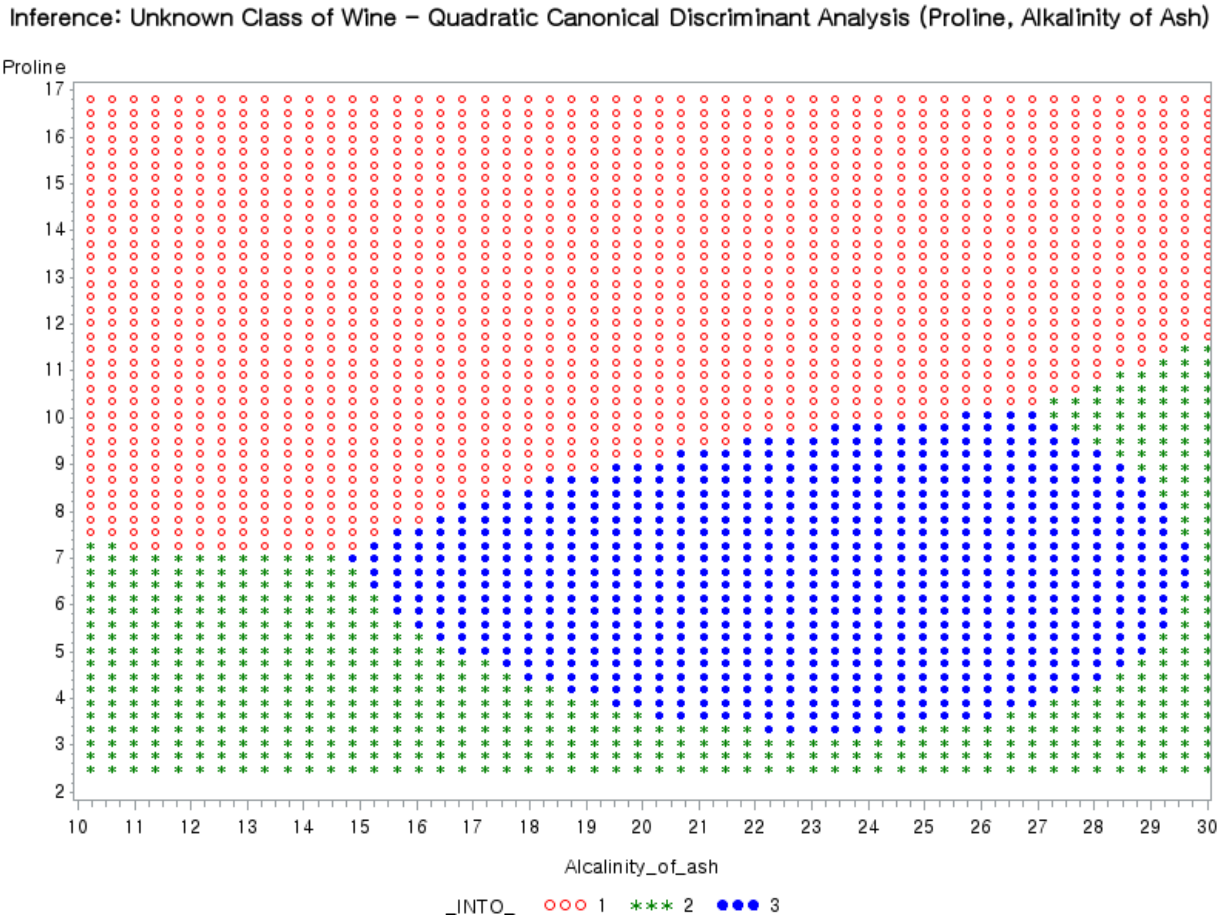
The DISCRIM Procedure
Classification Summary for Test Data: WORK.PLOTF
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	2809
Number of Observations Used	2809

Number of Observations and Percent Classified into Type				
	1	2	3	Total
Total	1571 55.93	623 22.18	615 21.89	2809 100.00
Priors	0.18	0.12	0.7	

We have checked that the first canonical variable is equivalent to proline and the second canonical variable is alkalinity of ash. In the Table 25, we can see 13% for the error count estimates. Also, when we have used only two main variables (Proline and Alkalinity of Ash) to do classification, the classification results are exactly same as the results using canonical variables. Here is the classification plot by using proline and alkalinity of ash only.

Figure 8: Plot of each type of wine (Classified class by Proline & Alk. of ash only)



Most results look normal, but a strange thing from the classification results is we have observed there are a lot of misclassified observations in Type 2. Also, we have checked the unusual classification plot showing in Figure 7 and 8. Based on the plot above, we can infer several hypotheses on the next page.

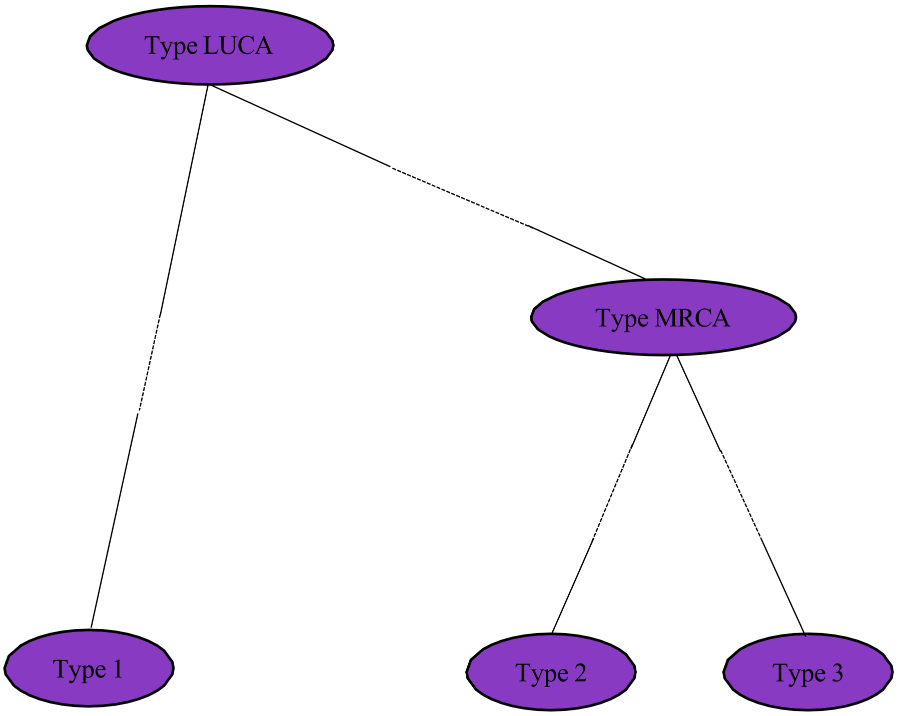
<**Special Case**: Interpretation and Hypotheses for the Unusual Classification Result>

Interpretation of the classification results:

Based on the results above, we have checked there are many misclassified observations between Type 2 and 3. From Figure 7, the first canonical variable which is heavily weighted on Proline is the key variable to classify three types and the second canonical variable "Alkalinity of ash" does not give enough evidence to classify each type of wine. Firstly, we can say that there is a specific border line between Type 1 (Relatively large amount proline in Type 1) vs Type 2 & 3 (Relatively small amount proline in Type 2 and 3) explained by Proline. However, Type 2 and 3 seem too ambiguous to distinguish each other clearly. It implies that these two types are not significantly different explained by the most important variable "Proline", so we can infer the relationship between Type 2 and 3 might be:

Hypothesis 1: [Taxonomy: Classification] Let all three cultivars have the last universal common ancestor. Then, based on the result of classification, we can say Type 2 and 3 cultivar has the most recent common ancestor. (See Figure 9)

Figure 9: LUCA & MRCA



Hypothesis 2: [Molecular Biology: Genetic mutation] For a long ago, there was a mutation by an unknown reason on the specific DNA sequence determines the amount of Proline in the second cultivar. As a result of mutation, the third cultivar was classified and then the third cultivar has been proliferated.

Hypothesis 3: [Ecology: Natural selection] Let assume the type 2 and 3 cultivar be same cultivar originally. (Let's say these cultivars was the type A cultivar.) Let some of the type A cultivar (Let's say this cultivar is the type 3 cultivar) be moved and started growing in different environment and location. By the ecological phenomenon, even though the characteristic of proline in the cultivar A was same before, after the cultivar 3 started growing in different locations, the characteristic of amount of proline in the cultivar 3 has been selected naturally to adapt environments of the new location. Then, the cultivar A cultivar has been classified as Type 2 and Type 3.

Problem:

However, for now, we are facing a huge problem we have only one information that three types of wine are produced by three different cultivars in Italy. Since the data provider did not present the specific cultivar and name of three wine types, it is impossible defining what cultivar of wines exactly is. In this restriction, we must infer what characteristic contains in each type of wine and must rely on only our statistical results what we have done. Since we know that the most important chemical variable was proline, we need to test something related to Proline which can explain and classify each class of wine. There might be several ways to figure out what characteristic contains in, but we are going to present one chemical test for Proline.

< Ninhydrin Test: Detection of Amino Acid – Short Explanation>

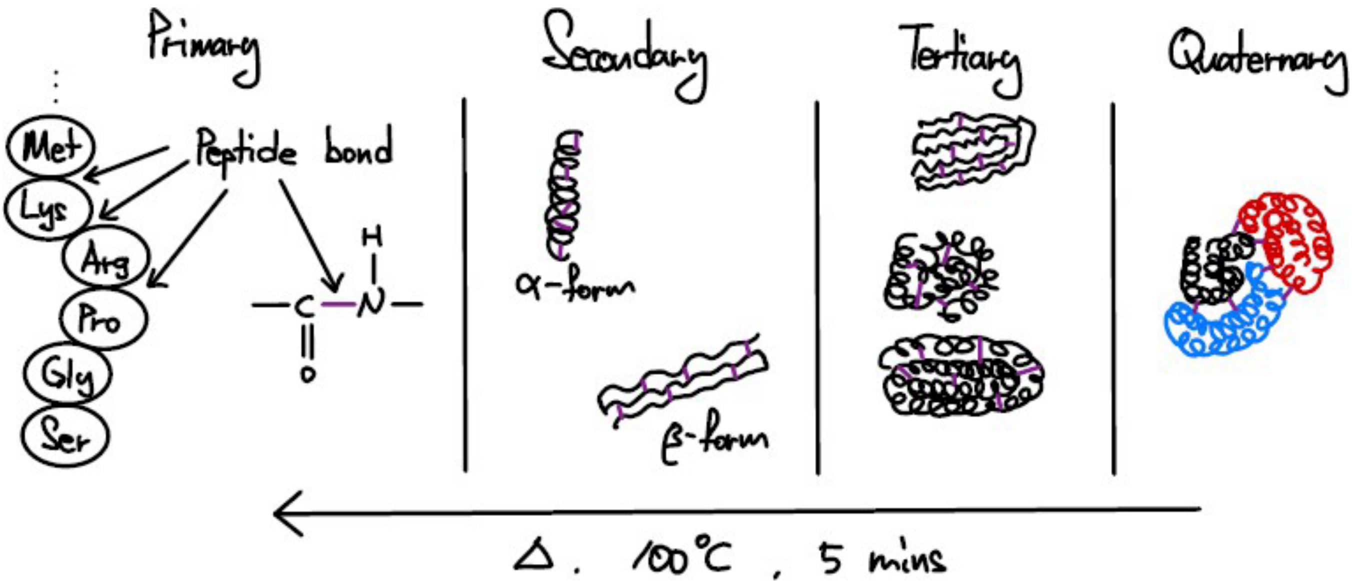
We are presenting this test because there is no way to distinguish three unknown wine types. Furthermore, even though we do not know the exact name of three cultivars, we might be able to figure out the characteristic of each cultivar. The expected result could be inaccurate since the dataset has lack of information in chemical compounds. The specific chemical experiment processes will not be presented. The result may only be explained theoretically.

1) Even though several proteins are contained in wine, since our data has only one amino acid (Proline), let assume there is only one amino acid in all types of wine OR assume we have done by the SDS-PAGE test (One of the tests for separating protein by electricity and mass) and know what kinds of proteins in each type of wine. [SDS-PAGE: *Sodium Dodecyl Sulfate – Polyacrylamide Gel Electrophoresis*]

2) In fact, since all amino acids are reacted by Ninhydrin (*2,2 – Dihydroxyindoane – 1,3 –dione*), the only one neutral amino acids "Proline" is turned out the yellow product at the end of reaction. The other amino acids are turned out the purple product at the end of reaction.

3) Since almost all amino acids are formed as a protein structure in the interior of the body, plant or food (not heated/cooked), especially, most proteins have the tertiary (Three-dimensional folding pattern) or quaternary (Group of tertiary protein structures) protein structure. (See Figure 10)

Figure 10: Type of protein structures



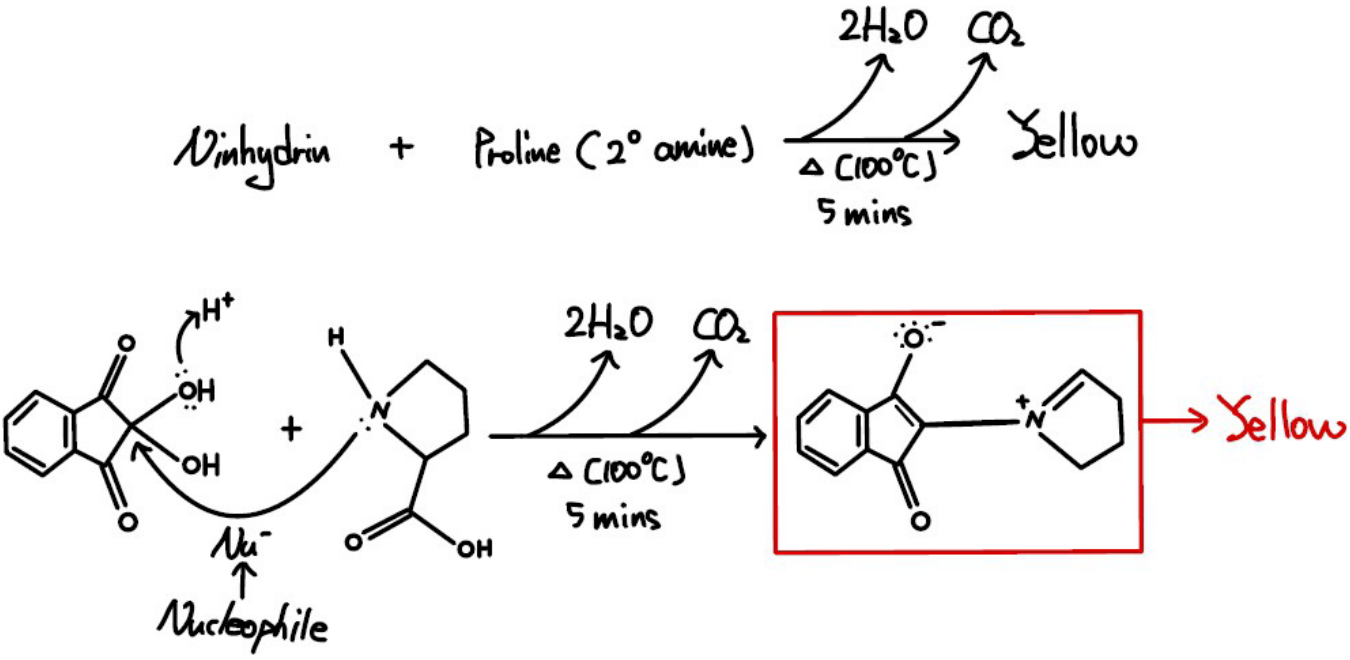
4) Since Ninhydrin does not react with any protein structure, we need to break all bonds of the protein structure. To break bonds in protine, 100°C heat during 5 minutes must be needed to each type of wine solution.

5) Ninhydrin test: Chemical equation

$$\boldsymbol{Ninhydrin}\ \ (reagent) + \boldsymbol{Proline}\ (in\ wine\ solution) \rightarrow Yellow\ product$$

To add, here is the short mechanism in ninhydrin reaction test. (See Figure 11)

Figure 11: Ninhydrin reaction: Short mechanism



6) If the wine type 1 has more proline than the wine type 2, then we will be able to check that the product of the wine type 1 has a strong yellow color product than the type 2. (Since we assume there is only one amino

acid in all types of wine)

<Statistical Results for Ninhydrin Reaction>

1) Based on the result of canonical correlation and canonical discriminant analysis with canonical variables using Proline and Alkalinity of ash, we have checked that the canonical variable 1 (Can1) has a significant strong correlation, explains approximately 98% variability and the canonical variable 1 is heavily weighted on Proline.

2) From the Figure 7, we have inferred that there is not much difference between Type 2 and Type 3, but we have checked that Type 1 and 2 & 3 are significantly different defined by the canonical variable 1. Since that, we can guess that the wine type 1 contains more proline than the other types. Therefore, we would be able to conclude the chemical product of the wine type 1 has a strong yellow color product rather than the other types.

<Expected Result: Color Brightness - Yellow>

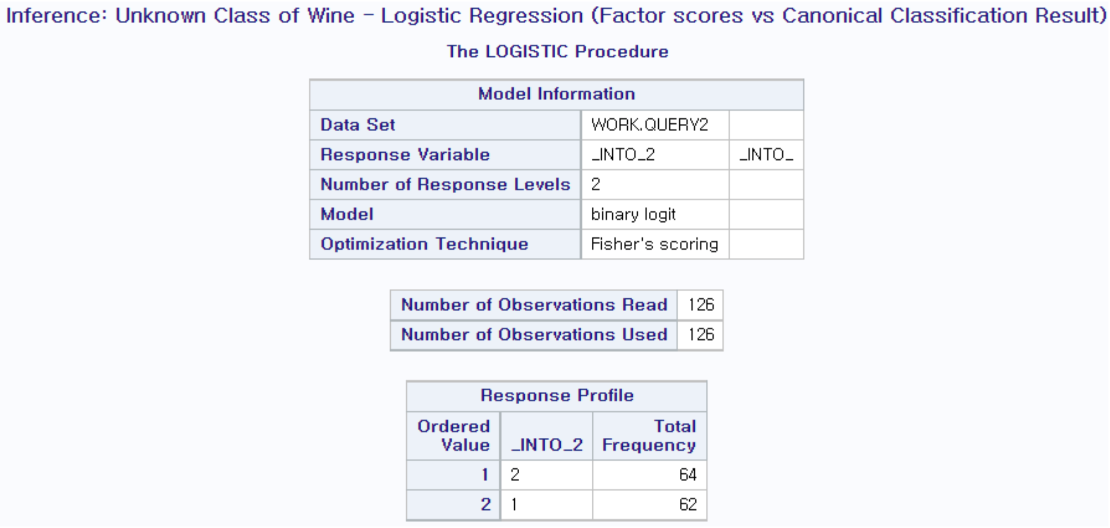
If we do the Ninhydrin test in same environment and conditions for each type of wine solution, the expected result will be:

Relatively strong = Type 1 ≥ Type 2 ≈ Type 3 = Relatively weak

3. Multivariate Analysis (C. Logistic Regression Analysis)

Through this analysis, we are going to check whether our classification results are well-classified. For the first logistic regression analysis, we are going to put three factor scores into our logistic regression model. These two results are based on the three factor scores compared to Type 1 vs Type 2. (See Table 28 and 29)

Table 28: Description - Type 1 vs Type 2



Inference: Unknown Class of Wine - Logistic Regression (Factor scores vs Canonical Classification Result)

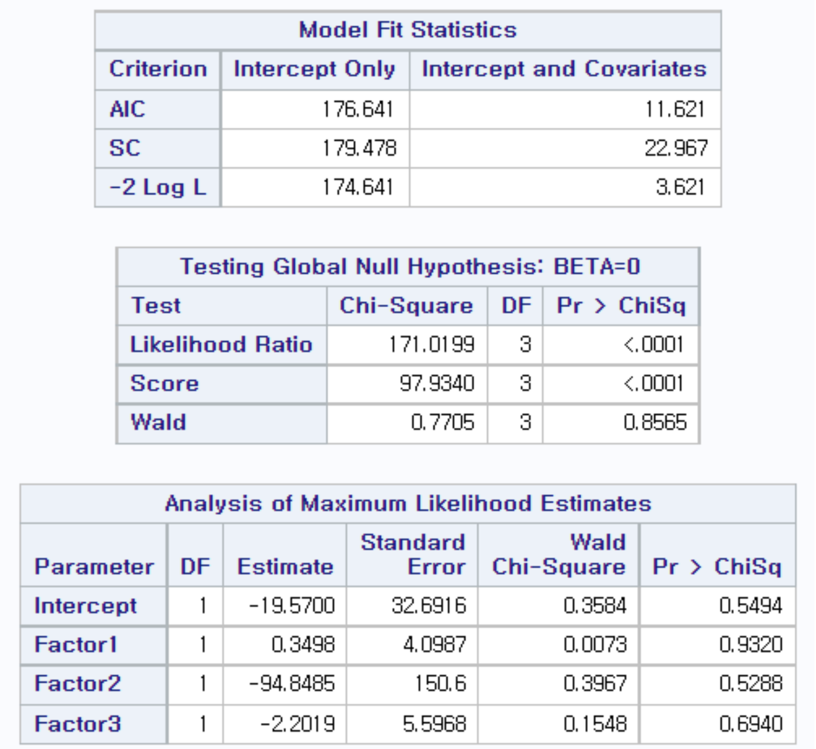
The LOGISTIC Procedure

Model Information		
Data Set	WORK.QUERY2	
Response Variable	_INTO_2	_INTO_
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	126
Number of Observations Used	126

Response Profile		
Ordered Value	_INTO_2	Total Frequency
1	2	64
2	1	62

Table 29: Logistic regression model and odd ratio - Type 1 vs Type 2



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	176.641	11.621
SC	179.478	22.967
-2 Log L	174.641	3.621

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	171.0199	3	<.0001
Score	97.9340	3	<.0001
Wald	0.7705	3	0.8565

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-19.5700	32.6916	0.3584	0.5494
Factor1	1	0.3498	4.0987	0.0073	0.9320
Factor2	1	-94.8485	150.6	0.3967	0.5288
Factor3	1	-2.2019	5.5968	0.1548	0.6940

In the table 29, at first, we can see that all parameters are not significant at 0.05 by the Wald test. The likelihood ratio test statistic is 171.0199 which can be calculated by:

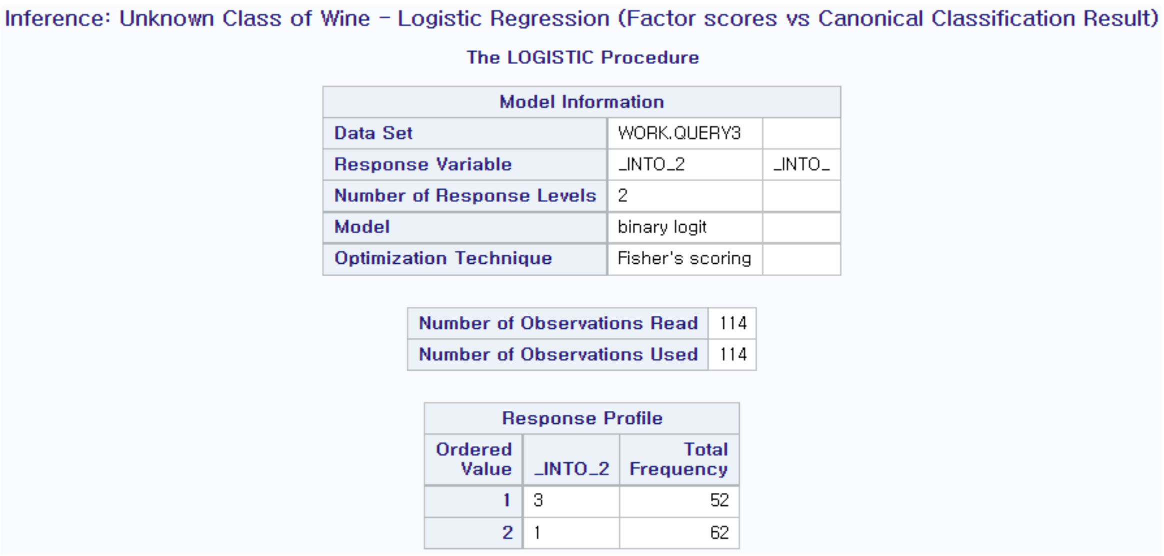
$$-2logL(b_0) - (-2logL(b_0, b_1)) = 174.641 - 3.621 = 171.0199 \rightarrow Enough\ to\ reject\ H_0$$

Under the null hypothesis (H_0: $All\ coefficients\ are\ zero\ vs\ H_1$: $At\ least\ one\ coeff\ is\ not\ zero$), we can say that at least one coefficient is not zero at 0.05 due to the high LR statistic. However, we can see that the Wald statistic is closed to 0, so we fail to reject the null hypothesis by the Wald test. It means our logistic regression model for Type 1 vs Type 2 is not valid. We have only the Wald test results in the table of analysis of maximum likelihood estimates, so we conclude that our first logistic regression model can be expressed:

$$g_1(\underline{x}) = 0$$

Now, we can move on the next model. (See Table 30 and 31)

Table 30: Description - Type 1 vs Type 3



Inference: Unknown Class of Wine - Logistic Regression (Factor scores vs Canonical Classification Result)

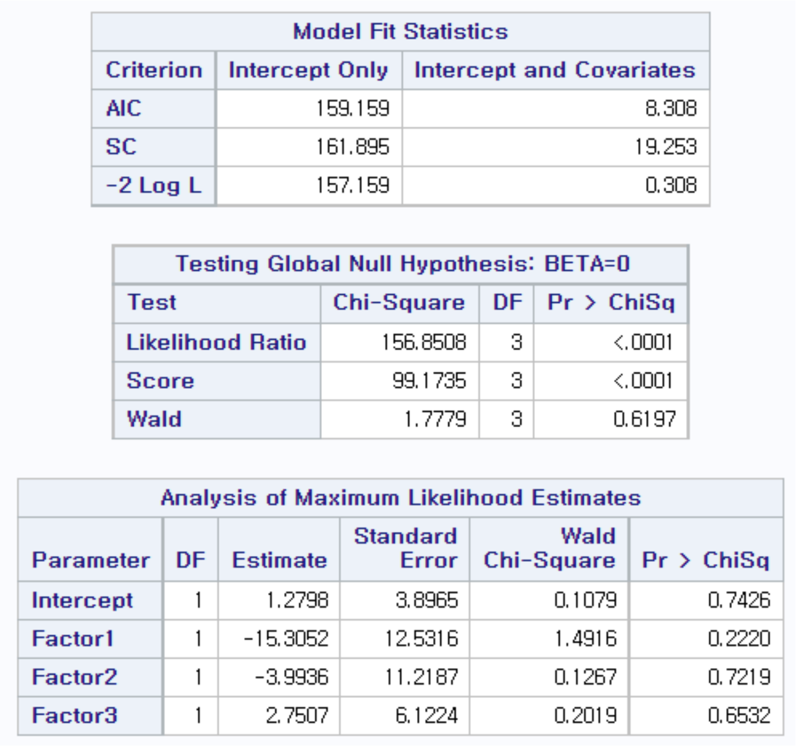
The LOGISTIC Procedure

Model Information		
Data Set	WORK.QUERY3	
Response Variable	_INTO_2	_INTO_
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	114
Number of Observations Used	114

Response Profile		
Ordered Value	_INTO_2	Total Frequency
1	3	52
2	1	62

Table 31: Logistic regression model and odd ratio - Type 1 vs Type 3



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	159.159	8.308
SC	161.895	19.253
-2 Log L	157.159	0.308

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	156.8508	3	<.0001
Score	99.1735	3	<.0001
Wald	1.7779	3	0.6197

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2798	3.8965	0.1079	0.7426
Factor1	1	-15.3052	12.5316	1.4916	0.2220
Factor2	1	-3.9936	11.2187	0.1267	0.7219
Factor3	1	2.7507	6.1224	0.2019	0.6532

In the table 31, at first, we can see that the likelihood ratio test statistic is

$$-2logL(b_0) - (-2logL(b_0, b_2)) = 157.159 - 0.308 = 156.8508 \rightarrow Enough\ to\ reject\ H_0$$

Similarly, under the null hypothesis, we can say that at least one coefficient is not zero at 0.05, so our logistic regression model for Type 1 vs Type 3 is valid. However, the results of the Wald test, all coefficients equal to zero. Also, in the table of analysis of maximum likelihood estimates, all coefficients are not significant by the Wald test. As like a first comparison, we conclude that our second logistic regression model can be expressed $g_2(\underline{x}) = 0$. Lastly, since two logistic regression models we have computed equal to 0, we can still compute the fitted probabilities even though these fitted probabilities are not meaningless. (See Table 32)

Table 32: Fitted probability table

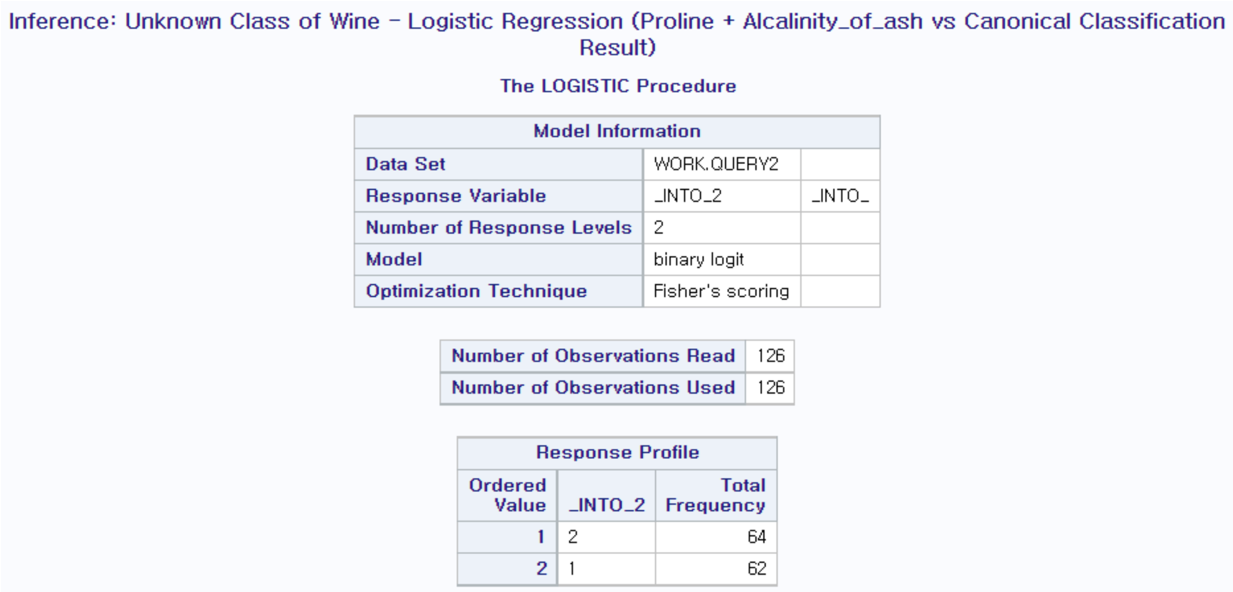
$x_1 = Factor1,\quad x_2 = Factor2,\quad x_3 = Factor3$

$\hat{P}(Y = 1 \| x_1, x_2, x_3)$	$\frac{1}{1 + e^0 + e^0} \approx 0.33$
$\hat{P}(Y = 2 \| x_1, x_2, x_3)$	$\frac{e^0}{1 + e^0 + e^0} \approx 0.33$
$\hat{P}(Y = 3 \| x_1, x_2, x_3)$	$\frac{e^0}{1 + e^0 + e^0} \approx 0.33$

It means that we cannot have a predicted classification result using factor scores by the logistic regression analysis.

For the next logistic analysis, we are going to use two important variables "Proline & Alkalinity of Ash" with the classification result of canonical variables. Here is our result of the logistic regression model of Type 1 vs Type 2 (See Table 33 and 34)

Table 33: Description - Type 1 vs Type 2



Inference: Unknown Class of Wine – Logistic Regression (Proline + Alcalinity_of_ash vs Canonical Classification Result)

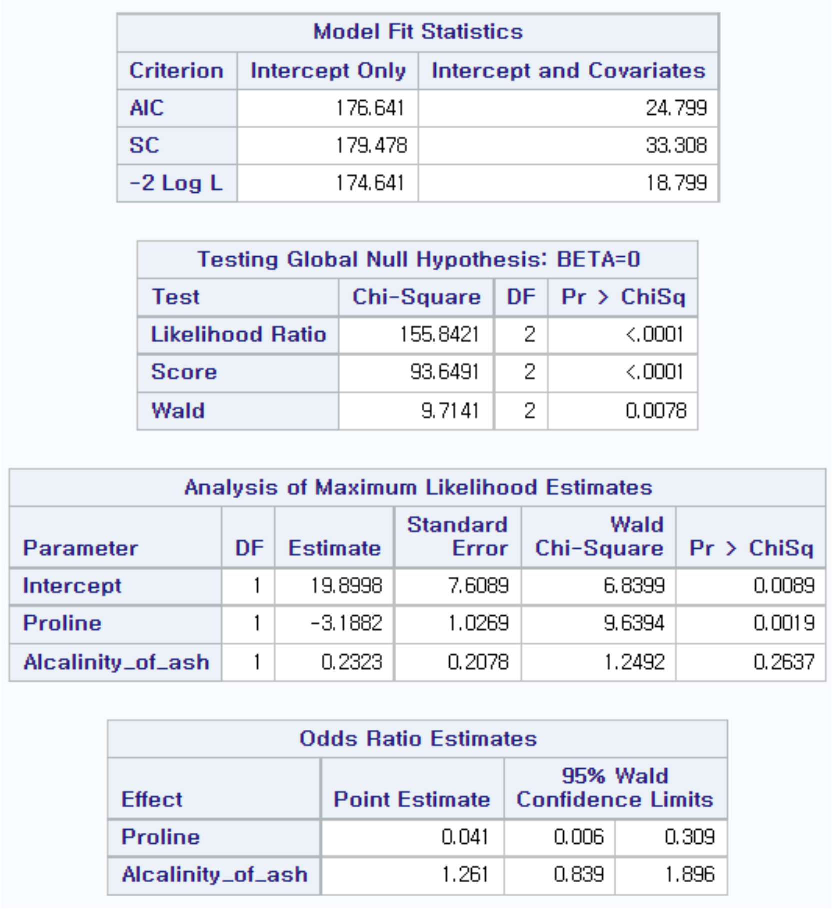
The LOGISTIC Procedure

Model Information		
Data Set	WORK.QUERY2	
Response Variable	_INTO_2	_INTO_
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	126
Number of Observations Used	126

Response Profile		
Ordered Value	_INTO_2	Total Frequency
1	2	64
2	1	62

Table 34: Logistic regression model and odd ratio - Type 1 vs Type 2



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	176.641	24.799
SC	179.478	33.308
-2 Log L	174.641	18.799

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	155.8421	2	<.0001
Score	93.6491	2	<.0001
Wald	9.7141	2	0.0078

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	19.8998	7.6089	6.8399	0.0089
Proline	1	-3.1882	1.0269	9.6394	0.0019
Alcalinity_of_ash	1	0.2323	0.2078	1.2492	0.2637

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Proline	0.041	0.006	0.309
Alcalinity_of_ash	1.261	0.839	1.896

In the table 29, at first, we can see that the likelihood ratio test statistic is 155.8421 which can be calculated by:

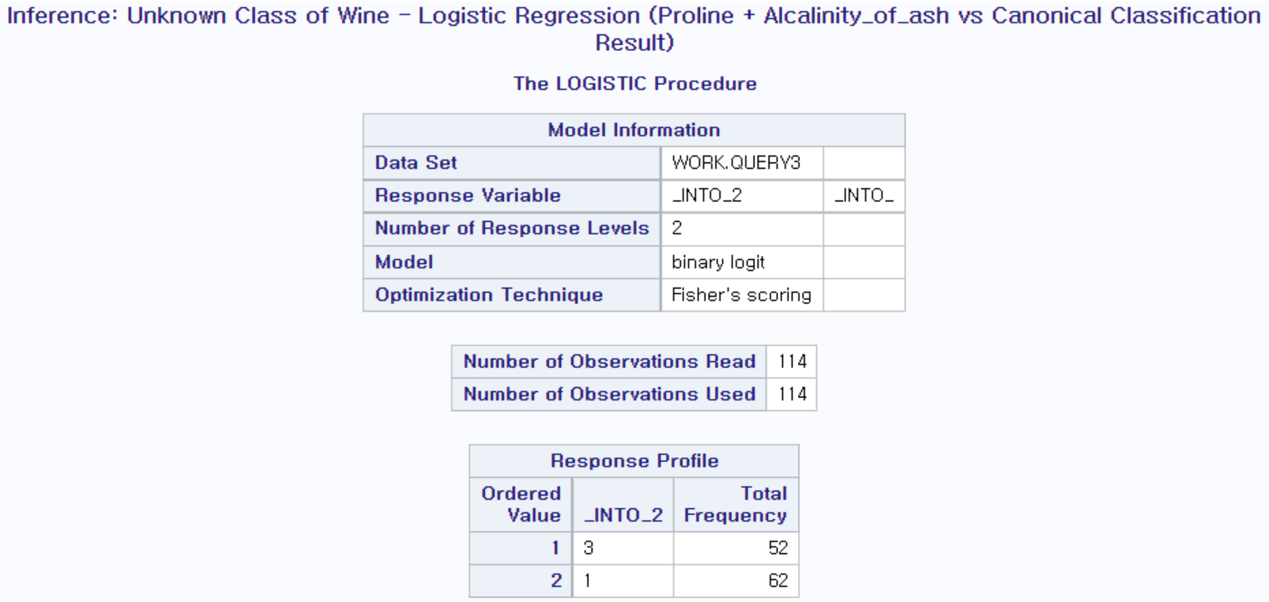
$$-2logL(b_0) - (-2logL(b_0, b_1)) = 176.641 - 18.799 = 155.8421 \rightarrow Enough\ to\ reject\ H_0$$

Under the null hypothesis (H_0: $All\ coefficients\ are\ zero\ vs\ H_1$: $At\ least\ one\ coeff\ is\ not\ zero$), we can say that at least one coefficient is not zero at 0.05, so our logistic regression model for Type 1 vs Type 2 is valid. Also, by the Wald test, we reject the null hypothesis. For the coefficients, by the Wald test, the coefficient of Alkalinity of ash is not significant, but the other parameters are significant, so we conclude that our first logistic regression model can be expressed:

$$g_1(\underline{x}) = 19.8998 - 3.1882x_1$$

The odd ratio of proline for Type 1 vs Type 2 is 0.041 and it implies that the probability to be classified into Type 1 is the product of 0.041 and the probability for Type 2 by proline. Now, we can move on the next model. (See Table 35 and 36)

Table 35: Description - Type 1 vs Type 3



Inference: Unknown Class of Wine - Logistic Regression (Proline + Alcalinity_of_ash vs Canonical Classification Result)

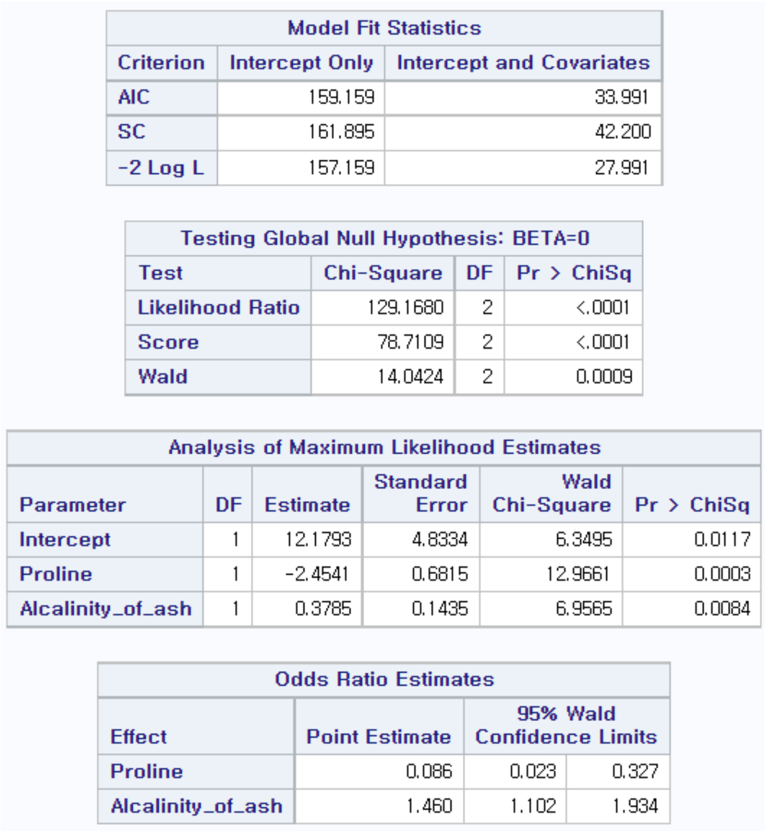
The LOGISTIC Procedure

Model Information		
Data Set	WORK.QUERY3	
Response Variable	_INTO_2	_INTO_
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	114
Number of Observations Used	114

Response Profile		
Ordered Value	_INTO_2	Total Frequency
1	3	52
2	1	62

Table 36: Logistic regression model and odd ratio - Type 1 vs Type 3



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	159.159	33.991
SC	161.895	42.200
-2 Log L	157.159	27.991

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	129.1680	2	<.0001
Score	78.7109	2	<.0001
Wald	14.0424	2	0.0009

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	12.1793	4.8334	6.3495	0.0117
Proline	1	-2.4541	0.6815	12.9661	0.0003
Alcalinity_of_ash	1	0.3785	0.1435	6.9565	0.0084

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Proline	0.086	0.023	0.327
Alcalinity_of_ash	1.460	1.102	1.934

In the table 36, at first, we can see that the likelihood ratio test statistic is

$$-2logL(b_0) - (-2logL(b_0, b_2)) = 157.159 - 27.991 = 129.1680 \rightarrow Enough\ to\ reject\ H_0$$

Under the null hypothesis (H_0: $All\ coefficients\ are\ zero\ vs\ H_1$: $At\ least\ one\ coeff\ is\ not\ zero$), we can say that at least one coefficient is not zero at 0.05 by the likelihood ratio test and the Wald test, so our logistic regression model for Type 1 vs Type 3 is valid. Also, based on the result of the Wald test, all coefficients are significant, so we conclude that our first logistic regression model can be expressed:

$$g_2(\underline{x}) = 12.1793 - 2.4541x_1 + 0.3785x_2$$

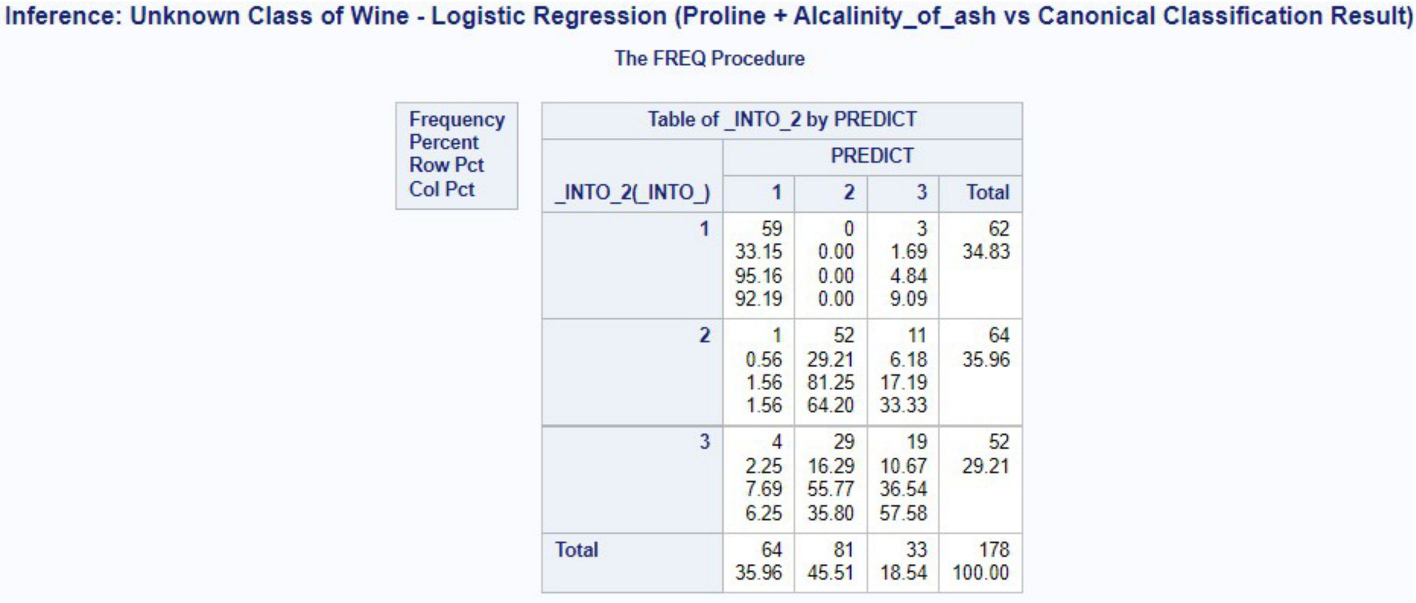
Lastly, the odd ratio of proline for Type 1 vs Type 3 is 0.086 and it implies that the probability to be classified into Type 1 is the product of 0.086 and the probability for Type 3 by proline. The odd ratio of alkalinity of ash is 1.46, so the probability to be classified into Type 1 is the product of 1.46 and the probability for Type 3 by alkalinity of ash. To calculate the fitted probability of the response given predictors, these two separated logistic models will be used. (See Table 37 and 38)

Table 37: Probability table

$x_1 = Proline, \qquad x_2 = Alkalinity\ of\ ash$

$P(Y = 1 \mid x_1, x_2)$	$\frac{1}{1 + e^{19.8998 - 3.1882x_1} + e^{12.1793 - 2.4541x_1 + 0.3785x_2}}$
$P(Y = 2 \mid x_1, x_2)$	$\frac{e^{19.8998 - 3.1882x_1}}{1 + e^{19.8998 - 3.1882x_1} + e^{12.1793 - 2.4541x_1 + 0.3785x_2}}$
$P(Y = 3 \mid x_1, x_2)$	$\frac{e^{12.1793 - 2.4541x_1 + 0.3785x_2}}{1 + e^{19.8998 - 3.1882x_1} + e^{12.1793 - 2.4541x_1 + 0.3785x_2}}$

Table 38: Classification result of logistic regression with Proline and Alkalinity of ash



Inference: Unknown Class of Wine - Logistic Regression (Proline + Alcalinity_of_ash vs Canonical Classification Result)

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of _INTO_2 by PREDICT

	PREDICT			
_INTO_2(_INTO_)	1	2	3	Total
1	59 33.15 95.16 92.19	0 0.00 0.00 0.00	3 1.69 4.84 9.09	62 34.83
2	1 0.56 1.56 1.56	52 29.21 81.25 64.20	11 6.18 17.19 33.33	64 35.96
3	4 2.25 7.69 6.25	29 16.29 55.77 35.80	19 10.67 36.54 57.58	52 29.21
Total	64 35.96	81 45.51	33 18.54	178 100.00

As before, based on the previous classification results in canonical discriminant analysis, we can check that misclassification frequencies between Type 2 and Type 3 are still quite high. Also, compared to Table 25, we can check the number of misclassified observations in logistic regression model are relatively smaller than the number the number of misclassified observations defined by canonical discriminant analysis using Proline & Alkalinity of Ash (same result with Can1 and Can2).

4. Summary & Conclusion

<Factor Analysis>

By the result of the analysis, we had three factor loadings (Factor 1: Phytochemical group, Factor 2: Proline group and Factor 3: Alkalinity of ash group). We expected the unique chemical components such as Proanthocyanins, Polyphenols and Total phenols in different cultivars determine a different class of wine. However, from the weighted communality table, we checked one of the neutral amino acids (Proline) and Alkalinity of ash mostly determine each class of wine.

<Canonical Discriminant Analysis>

We worked two tests. The first test was classification using factor scores. From the factor score, we had two canonical scores. The first canonical variable was heavily weighted on Factor 1 (Phytochemical group) and the second canonical variable was heavily weighted on Factor 2 (Proline group). The classification result said that Cultivar 1 and Cultivar 3 are classified by Phytochemical compounds and Cultivar 1 & Cultivar 2 and Cultivar 2 & Cultivar 3 are classified by Proline. This result gave us our original data was already well-classified. However, since we checked Proline and Alkalinity of ash have a relatively large proportion of the weighted communality in the Factor analysis, we also tried classification by original two variables and canonical scores of those two variables (Proline and Alkalinity of ash). After that, we figured out there is an unusual relationship between Cultivar 2 & Cultivar 3. Then, we inferred the unusual relationship using knowledges outside of Statistics. Since there is no information about the specific name of three cultivars, we realized that inferring the specific name of three wines are impossible. Based on that, we gave one Chemical & Biological experiment to figure out a characteristic of each type of wine, which can be explained by Proline. The expected results told us Cultivar 1 has larger amount of Proline than the other Cultivars relatively.

<Logistic Regression Analysis>

Through this analysis, we checked the classification results is well-classified. However, the result using factor scores totally were not significant, so we had no way to keep going ahead with this analysis. In other words, we could say using factor scores does not give enough evidence to classify each type of wine. For the second analysis, with Proline and Alkalinity of ash, we checked there is a massive relationship between Cultivar 2 and Cultivar 3 as we expected, but the performance of the Logistic regression model was more accurate than the classification result tested by Canonical discriminant analysis. We also checked that Proline is the most important variable to classify each class of wine.

5. Reference

(1) Wine Data Set. *UCI Machine Learning Repository: Wine Data Set*, https://archive.ics.uci.edu/ml/datasets/Wine. USA, Accessed Oct/20/2022

(2) A., Richard, et al. *Applied Multivariate Statistical Analysis*. 6th ed., Pearson Education, 2018.

(3) Reece, Jane B., et al. *Campbell Biology: United States Edition*. 9th ed., Pearson, 2010.