



Department of Statistics and Data Science
Jahangirnagar University
Course Title: Multivariate Analysis
Course No.: STAT-403
Course Teacher: Prof. Dr. Rumana Rois

Assignment 1

Wine quality data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 11 constituents found in each of the types of wines. There are 4898 rows and 12 columns in this dataset. The wine quality is on a scale that ranges from 0 (very bad) to 10 (excellent). Now, you need to reallocate the quality of wine as Average quality ≤ 5 and Good quality >5 . This data is available in the file **winequality-white.csv**.

Table 1: Wine Quality Data Set.

Fixed acidity	volatile acidity	Citric acid	Residual sugar	chlorides	Free sulfur dioxide	Total sulfur dioxide	density	pH	sulphates	alcohol	quality
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5.5	0.29	0.3	1.1	0.022	20	110	0.988	3.34	0.38	12.8	7
6	0.21	0.38	0.8	0.02	22	98	0.989	3.26	0.32	11.8	6

1. Examine the multivariate normality of **Fixed acidity, volatile acidity, Citric acid, Residual sugar, chlorides, Free sulfur dioxide, Total sulfur dioxide, density, pH, sulphates, and alcohol** variables of the Wine Quality data. Is there any outlier in those variables? Explain.
2. Construct the principal component analysis using the sample covariance matrix S for all the predictor variables in the Wine Quality data.
 - (i). Determine the sample principal components and their variances for the covariance matrix S .
 - (ii). Decide how many principal components to retain for this data. Hence, construct a Boxplot for each adequate principal component and interpret those selected principal components.
 - (iii). Calculate the correlation between the first principal component and pH predictor variables.
3. Conduct the factor analysis for the Wine Quality data and answer the following questions:
 - (i). Find the adequate number of common factors for this data set.
 - (ii). Hence, estimate the matrix of specific variances for the selected common factors and define the most significant variable that fits neatly into our factors model.
 - (iii). Find the estimated factor loadings and communalities. Interpret the estimated factor loadings.
 - (iv). What proportion of the total population variance is explained by the selected common factors? Hence, construct a Boxplot for each adequate factor score and interpret those selected common factors.

4. Calculate Fisher's linear discriminant function for classifying the Wine Quality data into Average quality ≤ 5 or Good quality > 5 . Classify the Wine Quality data into Average or Good quality using the following characteristics:

Fixed acidity	volatile acidity	Citric acid	Residual sugar	chlorides	Free sulfur dioxide	Total sulfur dioxide	density	pH	sulphates	Alcohol
9	0.22	0.38	0.8	0.12	22	98	0.99	3.26	0.32	11.8
8	0.22	0.26	1.2	0.035	18	97	0.99	3.12	0.41	9.7

5. Calculate Fisher's linear discriminant function for classifying the Wine Quality data into Average quality ≤ 5 or Good quality > 5 .
- based on the selected principal components in part (2)
 - based on the selected factor scores in part (3)

Classify the Wine Quality data into Average or Good quality using the Fisher's linear discriminant function in part (4), part (5.i) and part (5.ii) for the following dataset. Hence, calculate the apparent error rate (APER) and find the best discriminant function.

Fixed acidity	volatile acidity	Citric acid	Residual sugar	chlorides	Free sulfur dioxide	Total sulfur dioxide	density	pH	sulphates	alcohol	quality
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	Good
8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	Good
8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	Average
8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	Average
7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	Average
6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	Good
8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	Average
6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	Good
6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	Good
6.5	0.31	0.14	7.5	0.044	34	133	0.9955	3.22	0.5	9.5	Average
6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	Good
6.4	0.31	0.38	2.9	0.038	19	102	0.9912	3.17	0.35	11	Good
6.8	0.26	0.42	1.7	0.049	41	122	0.993	3.47	0.48	10.5	Good
7.6	0.67	0.14	1.5	0.074	25	168	0.9937	3.05	0.51	9.3	Average
7.2	0.32	0.36	2	0.033	37	114	0.9906	3.1	0.71	12.3	Good
7.4	0.25	0.36	2.05	0.05	31	100	0.992	3.19	0.44	10.8	Good

- Good Luck -