

2012 高教社杯全国大学生数学建模竞赛

承 诺 书

我们仔细阅读了中国大学生数学建模竞赛的竞赛规则.

我们完全明白, 在竞赛开始后参赛队员不能以任何方式(包括电话、电子邮件、网上咨询等)与队外的任何人(包括指导教师)研究、讨论与赛题有关的问题。

我们知道, 抄袭别人的成果是违反竞赛规则的, 如果引用别人的成果或其他公开的资料(包括网上查到的资料), 必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺, 严格遵守竞赛规则, 以保证竞赛的公正、公平性。如有违反竞赛规则的行为, 我们将受到严肃处理。

我们授权全国大学生数学建模竞赛组委会, 可将我们的论文以任何形式进行公开展示(包括进行网上公示, 在书籍、期刊和其他媒体进行正式或非正式发表等)。

我们参赛选择的题号是(从 A/B/C/D 中选择一项填写): A

我们的参赛报名号为(如果赛区设置报名号的话): S46027

所属学校(请填写完整的全名): 郑州轻工业学院

参赛队员(打印并签名): 1. 任 静

2. 李 栋

3. 毛新梅

指导教师或指导教师组负责人(打印并签名): 指导教师组

日期: 2012 年 9 月 10 日

赛区评阅编号(由赛区组委会评阅前进行编号):

2012 高教社杯全国大学生数学建模竞赛

编 号 专 用 页

赛区评阅编号（由赛区组委会评阅前进行编号）：

赛区评阅记录（可供赛区评阅时使用）：

评 阅 人										
评 分										
备 注										

全国统一编号（由赛区组委会送交全国前编号）：

全国评阅编号（由全国组委会评阅前进行编号）：

葡萄酒的评价

摘 要

本文研究的是葡萄酒的评价问题。根据已有数据,采用方差分析和主成分分析等方法,分析了酿酒葡萄和葡萄酒的理化指标与葡萄酒质量的相关关系,给出了较合理的葡萄酒质量评价指标。

针对问题一,根据每组评酒员对每个葡萄酒样品的打分,采用平均值方法,计算出葡萄酒样品的质量得分,对每类葡萄酒的两组质量得分进行单因子方差分析。分析结果显示,两组评酒员对两类葡萄酒评价结果都有显著差异。在此基础上,根据各组评酒员的评价结果,分别计算其方差和,均得出第二组的方差较小,即第二组评酒员的评价结果更可信。

针对问题二,首先,根据资料^{[8][7]}选取 30 个理化指标中有代表性的 12 个,同时选取第二组的评价结果作为葡萄酒的质量,采用主成分分析法,得出关于主成分 F_{13} , F_{12} , F_{11} , F_{10} , F_9 , F_8 和 F_7 的简化系统,并验证了系统的可行性。其次,采用计算各个样品的主成分值的方法,建立关于 7 个主成分值的综合评价模型,得出葡萄的分级如下表所示。

级别	酿酒葡萄(白)	酿酒葡萄(红)
优	26,21,2,5	9,23,3
良	23,25,10,14,19, 28,22,12,24,16	14,13,10,25,20,26,2,19,5,17
中	9,4,27,20,7,17,3,1,11,8,6	8,6,27,4,21,15,24,16,22,1,18
差	18,15,13	12,11,7

针对问题三,首先,根据附表中的 30 个酿酒葡萄的理化指标,采用主成分分析方法,建立关于酿酒葡萄和葡萄酒的典型相关性分析模型,得出葡萄酒和酿酒葡萄的各理化指标之间存在一定的相关性,具体见正文。

针对问题四,首先,根据第三问的模型分析出酿酒葡萄与葡萄酒理化指标间既有正相关也有负相关的结果,采用多元回归分析法,分别建立酿酒葡萄和葡萄酒的理化指标关于葡萄酒质量的线性拟合模型,得出酿酒葡萄和葡萄酒的理化指标完全可以代替人工评分评定葡萄酒的质量,并加以论证。

关键词:葡萄酒评价 主成分分析 单因子方差分析 理化指标

一 问题重述

1.1 问题背景

葡萄酒的质量问题是一个很受人关注的问题，确定葡萄酒质量时一般是通过聘请一批有资质的评酒员进行品评。每个评酒员在对葡萄酒进行品尝后对其分类指标打分，然后求和得到其总分，从而确定葡萄酒的质量。酿酒葡萄的好坏与所酿葡萄酒的质量有直接的关系，葡萄酒和酿酒葡萄检测的理化指标会在一定程度上反映葡萄酒和葡萄的质量。

1.2 问题提出

该题给出了我们三个附件，其中附件 1 给出了某一年份一些葡萄酒的评价结果，附件 2 和附件 3 分别给出了该年份这些葡萄酒的和酿酒葡萄的成分数据。我们尝试利用这些数据对葡萄酒的质量进行分析，建立数学模型讨论下列问题：

1. 分析附件 1 中两组评酒员的评价结果有无显著性差异，哪一组结果更可信？
2. 根据酿酒葡萄的理化指标和葡萄酒的质量对这些酿酒葡萄进行分级。
3. 分析酿酒葡萄与葡萄酒的理化指标之间的联系。
4. 分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并论证能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量？

二 模型假设

1. 所有的评酒员对葡萄酒的品质的打分客观公正；
2. 附件中的葡萄酒样品与酿酒葡萄样品是一一对应的关系；
3. 酿制红葡萄酒和白葡萄酒的酿酒葡萄是相互对立的；
4. 附件中错误的数据，缺失的数据对样品总体没影响，可以删去。

三 符号说明

b_i : 葡萄酒的原始质量；
 a_i : 标准化后的第 i 种葡萄酒的质量；
 n : 葡萄酒样品的数量；
 X : 酿酒葡萄的指标矩阵；
 x_i : 酿酒葡萄的第 i 个指标的观测值；
 U : 7 个主成分的各个指标的单位化特征向量矩阵；
 y_{ji} : 第 j 组的第 i 个样品的第 k 个评酒员的评分结果；
 s_{ji} : 第 j 组的第 i 个样品的评分结果的方差；
 S_j : 第 j 组的所有样品的评分结果的方差和；
 W_i : 第 i 个酿酒葡萄样品的综合得分；
 x_i : 第 i 个酿酒葡萄的指标的观测值；
 l_{ii} : 第 i 个酿酒葡萄的指标的组内方差和；
 F_j : 单位化特征向量矩阵的第 j 个主成分向量；
 a_{ji} : x_i 在 F_j 上的载荷；
 A : 载荷矩阵；
 λ_i : 第 i 个酿酒葡萄指标的特征值；
 R : 相关系数矩阵；
 μ : 单位化特征值向量矩阵；
 μ_{ij} : 第 i 个指标关于第 j 个指标的单位化特征向量；

四 模型的建立与求解

4.1 问题一的分析

问题一的关键是明确显著性差异的因子，以此建立单因子方差分析模型。首先，根据附表中的数据，去掉两组评分员对每种葡萄酒评分总和的最大值和最小值再求平均值得出葡萄酒的质量，再将数据进行标准化处理，以此为因子建立单因子方差分析模型，进行显著性分析。其次，根据所得的葡萄酒的质量，分析各组品酒员的品酒结果的波动情况，因此采取求各组评定的葡萄酒的质量的方差，分析得出更为可信的评酒员小组。

4.1.1 各组评酒员的评价结果

评价员的评价结果即为葡萄酒的质量。葡萄酒的质量是每个评酒员在对葡萄酒进行品尝后对其分类指标打分，然后求和得到的总分。为了使得到的数据更为合理，采用去掉评分结果的最大值和最小值再求平均值的方法来确定葡萄酒的质量。各种葡萄酒的质量如附表 1（2）所示。

分析数据时发现第一组四号评酒员评价红葡萄酒样品 20 时，没有给出具体数据，同样的该组的七号评酒员评价白葡萄酒样品 3 时，给出了远超出浓度的数据，为保证分析的严谨，我们把红葡萄酒样品 20 和白葡萄酒样品 3 剔除，因此只剩下了 26 种红葡萄酒样品和 27 种白葡萄酒样品。

4.1.2 两组评价结果的显著性分析

由于各种葡萄酒的质量本身是不同的，若要对此进行数据分析，必须将数据标准化。为此建立数据的标准化模型

$$a_i = \left| b_i - \frac{\sum_{i=1}^n b_i}{n} \right|, \quad (1.1)$$

其中： a_i 为标准化后的第 i 种红（白）葡萄酒的质量， b_i 为红（白）葡萄酒的原始质量， n 为红（白）葡萄酒样品的数量， $n = 27, 28$ 分别为红、白葡萄酒样品的数量。

将表 1 中的数据代入公式（1.1）标准化的数据见附表 2（2）。

该题中，要比较的是两组的评价结果对葡萄酒的影响，为此，把各种酒标准化后的数值作为因子，记为 A ，第一组的评价结果作为因子水平 A_1 ，第二组的评价结果作为因子水平 A_2 ，第 j 组对第 i 种葡萄酒的评价结果用 y_{ji} 表示， $i = 1, 2, \dots, 27$ 或 28 ，目的是比较两组评价的结果是否相同，为此，把研究的问题归结为一个统计问题，用方差分析法进行分析。

建立方差分析模型为

$$\left\{ \begin{array}{l} T_j = \sum_{i=1}^n y_{ji} \\ T = \sum_{j=1}^2 T_j \\ S_T = \sum_{j=1}^2 \sum_{i=1}^n y_{ji}^2 - \frac{T^2}{n} \\ S_A = \frac{1}{n} \sum_{j=1}^2 T_j^2 - \frac{T^2}{n} \\ S_e = S_T - S_A \end{array} \right. \quad (1.2)$$

将数据代入公式（1.2）中，用 EXCEL 计算出各偏差平方和，自由度，填入方差分析表，并继续计算得到各均方以及 F 比， p 值，见表 1 所示。

表1. 单因子的方差分析表

来源	平方和	自由度	均方	F 比	p 值	F crit
因子 A	71.69427	1	71.69427	6.437288	0.012659	3.932438
误差 e	1158.283	104	11.13734			
总计 T	1229.978	105				

若取 $\alpha = 0.05$ ，则 $F_{0.95}(1, 104) = 3.9324$ ，由于 $F = 6.437288 > 3.9324$ ，故认为因子 A 是显著的，即两组的评价结果有显著差异。

4.1.3 两组可信度的分析

根据显著性差异的分析，可知两组的评价结果有显著性差异，因此计算每组的成员对各种酒样品的评价结果的方差的总和（即成员之间评价的波动性强弱）来判断各组评价结果的可信度。

建立可信度的模型为

$$\left\{ \begin{array}{l} s_{ji} = \sum_{k=1}^{10} \left(y_{jik} - \frac{\sum_{k=1}^{10} y_{jik}}{10} \right)^2 \\ S_j = \sum_{i=1}^n s_{ji} \end{array} \right. \quad (1.3)$$

其中： y_{jik} 为第 j 组的第 i 个样品的第 k 个评酒员的评分结果， s_{ji} 为第 j 组的第 i 个样品的评分结果的方差， S_j 为第 j 组的所有样品的评分结果的方差和。

将附件 1 中的数据代入公式（1.3）中得到数据如附表 1 所示。分别求得的两组对于红酒和白酒的评分结果的方差和如表 2 所示。

表2. 两组的方差和

	第一组	第二组
红葡萄酒	1424.45	821.11
白葡萄酒	3255.57	1411.69

由于把每组的成员评价各种酒的评价结果的方差的总和来作为各组评价结果的可信度，根据方差的性质，方差越小，稳定性越好，可信度越高。因此分析上表的结果可知，无论是红酒还是白酒，都是第二组的可信度更高，即为第二组的评价结果更为可信。

4.2 问题二的分析

问题二的关键是确定评价指标，建立综合评价模型。首先，根据资料^[1]筛选附表中的数据，确定对酿酒葡萄影响较大的理化指标，由于第二组的评价结果更可信，选取第二组的评价结果作为葡萄酒的质量。综上，确定了评价指标。其次，将数据进行标准化处理，采用主因子分析法，找出主因子，以简化评价指标。最后，根据以所得的主因子为新的评价指标，建立对酿酒葡萄分级的综合评价模型，对酿酒葡萄的优劣进行分级。

4.2.1 酿酒葡萄分级的评价系统的简化

由于附表的数据量太大，必须对数据进行初步的筛选。根据资料^{[7][8]}，酿酒葡萄中的氨基酸，蛋白质，花色苷，有机酸，酚类，醇类，还原糖，果穗，出汁率，多酚氧化酶活力，DPPH 自由基，可溶性固形物这几个理化指标是对酿酒葡萄影响较大的理化指标。其次，由于第二组的评价结果更为可信，故选用第二组的评价结果作为酿酒葡萄的葡萄酒的质量指标。综上所述，初步确定了酿酒葡萄分级的指标。

由于选取的指标太多，而且主观性较强，容易使它们提供的整体信息发生重叠，不易得出简明的规律。因此，采用主成分分析法，将多指标问题化为较少的综合指标问题，不但保证了各指标的不相关性，又反映了原来多指标的信息。

下面根据上述分析来进行主成分分析。

首先，要对各个指标的观测值进行标准化。各个指标的观测值的具体值见附表 2。

$X = (x_1, x_2 \cdots x_{13})^T$ ，建立数据标准化模型

$$\begin{cases} l_{ii} = \sum (x_i)^2 - \frac{1}{n} (\sum x_i)^2 \\ x_i' = \frac{x_i - \bar{x}_i}{\sqrt{l_{ii} / (n - 1)}} \end{cases} \quad i = (1, 2, \cdots 13),$$

其中： x_i 为第 i 个指标的观测值， l_{ii} 为组内方差和。

标准化后的数据见附表 3。

建立各个指标的相关系数模型

$$\mu_{ij} = \frac{\sum_{i=1}^{13} \sum_{j=1}^{13} (x_i - \bar{x})(x_j - \bar{x}_j)}{\sqrt{\sum_{i=1}^{13} (x_i - \bar{x})^2 \sum_{j=1}^{13} (x_j - \bar{x}_j)^2}},$$

将附表 3 中的数据代入公式中，得到相关矩阵为

$$R = \begin{pmatrix} 1 & 0.12 & -0.06 & 0.50 & -0.33 & 0.21 & 0.18 & 0.15 & 0.41 & 0.45 & -0.09 & -0.25 & 0.32 \\ 0.12 & 1 & -0.45 & -0.08 & -0.28 & 0.10 & 0.53 & 0.37 & 0.03 & -0.04 & 0.19 & -0.10 & -0.04 \\ -0.06 & -0.45 & 1 & 0.12 & 0.35 & -0.30 & -0.22 & -0.20 & -0.23 & -0.24 & 0.07 & -0.06 & -0.26 \\ 0.50 & -0.08 & 0.12 & 1 & -0.09 & -0.10 & 0.08 & 0.25 & 0.10 & 0.28 & -0.18 & -0.01 & 0.35 \\ -0.33 & -0.28 & 0.35 & -0.09 & 1 & -0.43 & -0.32 & -0.20 & -0.31 & -0.14 & -0.02 & 0.18 & -0.28 \\ 0.21 & 0.10 & -0.30 & -0.10 & -0.43 & 1 & 0.38 & 0.16 & 0.27 & 0.30 & -0.06 & -0.25 & 0.31 \\ 0.18 & 0.53 & -0.22 & 0.08 & -0.32 & 0.38 & 1 & 0.55 & -0.05 & -0.07 & 0.27 & -0.08 & -0.07 \\ 0.15 & 0.37 & -0.20 & 0.25 & -0.20 & 0.16 & 0.55 & 1 & 0.13 & 0.03 & 0.07 & -0.26 & 0.02 \\ 0.41 & 0.03 & -0.23 & 0.10 & -0.31 & 0.27 & -0.05 & 0.13 & 1 & 0.59 & -0.41 & -0.41 & 0.29 \\ 0.45 & -0.04 & -0.24 & 0.28 & -0.14 & 0.30 & -0.07 & 0.03 & 0.59 & 1 & -0.67 & -0.52 & 0.48 \\ -0.09 & 0.19 & 0.07 & -0.18 & -0.02 & -0.06 & 0.27 & 0.07 & -0.41 & -0.67 & 1 & 0.40 & -0.47 \\ -0.25 & -0.10 & -0.06 & -0.01 & 0.18 & -0.25 & -0.08 & -0.26 & -0.41 & -0.52 & 0.40 & 1 & -0.16 \\ 0.32 & -0.04 & -0.26 & 0.35 & -0.28 & 0.31 & -0.07 & 0.02 & 0.29 & 0.48 & -0.47 & -0.16 & 1 \end{pmatrix}$$

相关矩阵 R 的特征值及相应的单位化特征向量为

$$\lambda = 0.16 \quad 0.21 \quad 0.24 \quad 0.34 \quad 0.39 \quad 0.56 \quad 0.58 \quad 0.81 \quad 0.96 \quad 1.17 \quad 1.44 \quad 2.51 \quad 3.64$$

$$\mu = \begin{pmatrix} 0.10 & 0.32 & -0.45 & 0.10 & -0.20 & 0.02 & -0.35 & 0.43 & -0.22 & -0.14 & -0.40 & 0.00 & 0.33 \\ -0.18 & 0.24 & 0.34 & -0.31 & -0.07 & 0.43 & -0.22 & 0.26 & 0.40 & 0.08 & 0.07 & -0.45 & 0.13 \\ -0.36 & 0.08 & 0.04 & -0.47 & 0.06 & 0.25 & 0.04 & -0.08 & -0.45 & 0.33 & -0.39 & 0.24 & -0.22 \\ 0.22 & -0.04 & 0.56 & 0.14 & 0.20 & -0.03 & 0.07 & -0.09 & 0.14 & -0.22 & -0.67 & 0.11 & 0.19 \\ 0.19 & -0.01 & 0.07 & -0.16 & -0.33 & -0.33 & -0.60 & -0.18 & 0.33 & 0.27 & -0.10 & 0.24 & -0.29 \\ 0.01 & 0.37 & 0.36 & 0.00 & -0.06 & -0.20 & -0.27 & -0.40 & -0.51 & -0.08 & 0.28 & -0.17 & 0.29 \\ 0.22 & -0.44 & -0.28 & -0.24 & 0.42 & -0.02 & -0.27 & -0.24 & -0.07 & 0.13 & -0.16 & -0.50 & 0.14 \\ -0.28 & 0.23 & -0.19 & 0.07 & -0.26 & -0.37 & 0.38 & -0.28 & 0.24 & 0.33 & -0.29 & -0.34 & 0.19 \\ 0.15 & -0.16 & 0.17 & -0.51 & -0.01 & -0.48 & 0.22 & 0.44 & -0.05 & 0.13 & 0.14 & 0.13 & 0.37 \\ -0.65 & -0.30 & 0.04 & 0.26 & 0.19 & -0.07 & -0.32 & 0.02 & 0.12 & 0.10 & 0.07 & 0.26 & 0.41 \\ -0.20 & -0.47 & 0.24 & 0.16 & -0.47 & -0.13 & -0.05 & 0.24 & -0.31 & -0.13 & -0.13 & -0.39 & -0.27 \\ -0.36 & 0.19 & -0.11 & -0.30 & 0.27 & -0.38 & -0.08 & 0.01 & 0.13 & -0.63 & -0.06 & -0.09 & -0.30 \\ -0.02 & -0.26 & -0.15 & -0.35 & -0.48 & 0.27 & 0.09 & -0.39 & 0.10 & -0.42 & 0.00 & 0.17 & 0.33 \end{pmatrix}$$

由于后七个主成分的累计贡献率

$$\eta_{(7)} = \frac{0.58 + 0.81 + 0.96 + 1.17 + 1.44 + 2.51 + 3.64}{13} = 85.3915\% > 85\%$$

故取后七个主成分:

$$\begin{aligned} F_{13} &= 0.33x_1 + 0.13x_2 - 0.22x_3 + 0.19x_4 - 0.29x_5 + 0.29x_6 + 0.14x_7 + 0.19x_8 + 0.37x_9 + 0.41x_{10} - 0.27x_{11} - 0.30x_{12} + 0.33x_{13}, \\ F_{12} &= 0.00x_1 - 0.45x_2 + 0.24x_3 + 0.11x_4 + 0.24x_5 - 0.17x_6 - 0.50x_7 - 0.34x_8 + 0.13x_9 + 0.26x_{10} - 0.39x_{11} - 0.09x_{12} + 0.17x_{13}, \\ F_{11} &= -0.40x_1 + 0.07x_2 - 0.39x_3 - 0.67x_4 - 0.10x_5 + 0.28x_6 - 0.16x_7 - 0.29x_8 + 0.14x_9 + 0.07x_{10} - 0.13x_{11} - 0.06x_{12} + 0.00x_{13}, \\ F_{10} &= -0.14x_1 + 0.08x_2 + 0.33x_3 - 0.22x_4 + 0.27x_5 - 0.08x_6 + 0.13x_7 + 0.33x_8 + 0.13x_9 + 0.10x_{10} - 0.13x_{11} - 0.63x_{12} - 0.42x_{13}, \\ F_9 &= -0.22x_1 + 0.40x_2 - 0.45x_3 + 0.14x_4 + 0.33x_5 - 0.51x_6 - 0.07x_7 + 0.24x_8 - 0.05x_9 + 0.12x_{10} - 0.31x_{11} + 0.13x_{12} + 0.10x_{13}, \\ F_8 &= 0.43x_1 + 0.26x_2 - 0.08x_3 + 0.09x_4 - 0.18x_5 - 0.40x_6 - 0.24x_7 - 0.28x_8 + 0.44x_9 + 0.02x_{10} + 0.24x_{11} + 0.01x_{12} - 0.39x_{13}, \\ F_7 &= -0.35x_1 - 0.22x_2 + 0.04x_3 + 0.07x_4 - 0.60x_5 - 0.27x_6 - 0.27x_7 + 0.38x_8 + 0.22x_9 - 0.32x_{10} - 0.05x_{11} - 0.08x_{12} + 0.09x_{13}. \end{aligned}$$

F_{13} 主要反映了 x_{10} (可溶性固物) 的信息, 它们的系数为 0.41, 它代表了 x_{10} 对 F_{13} 作用的权数, 同理可分析得出以下结论。

F_{12} 主要综合了 x_2 (蛋白质)、 x_7 (酚类)、 x_{11} (果穗质量) 的信息;

F_{11} 主要综合了 x_1 (氨基酸) 和 x_4 (有机酸) 的信息;

F_{10} 主要综合了 x_{12} (出汁率) 和 x_{13} (葡萄酒的质量) 的信息;

F_9 主要综合了 x_3 (花色苷鲜重) 和 x_6 (DPPH 自由基) 的信息;

F_8 主要反映了 x_9 (还原糖) 的信息;

F_7 主要反映了 x_5 (多酚氧化酶活力) 和 x_8 (醇类) 的信息。

4.2.2 酿酒葡萄分级的评价系统可行性的验证

为了验证用 F_{13} , F_{12} , F_{11} , F_{10} , F_9 , F_8 和 F_7 简化系统的可行性, 必须验证该简化系统的可行性。

把主成分分析的结果用矩阵的形式表示出来, 用 X 表示各种指标的信息矩阵, 因此可以将 X 用如下形式表示出来。

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{13} \end{pmatrix} = \begin{pmatrix} \mu_{11} & \mu_{21} & \cdots & \mu_{m1} \\ \mu_{12} & \mu_{22} & \cdots & \mu_{m2} \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{1,13} & \mu_{2,13} & \cdots & \mu_{13,13} \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_{13} \end{pmatrix} = U_1 F_1 + U_2 F_2 + \cdots U_{13} F_{13}$$

为了分析简化系统的可行性, 要求出 x_i 在 F_j 上的载荷

$$a_{ji} = \mu_{ji} \sqrt{\lambda_j}$$

即 F_j 上承载了 x_i 标准差的多少。则上述矩阵可转化为

$$X = \begin{pmatrix} \mu_{11} & \mu_{21} & \cdots & \mu_{m1} \\ \mu_{12} & \mu_{22} & \cdots & \mu_{m2} \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{1,13} & \mu_{2,13} & \cdots & \mu_{13,13} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_{13}} \end{pmatrix} \begin{pmatrix} F_1 / \sqrt{\lambda_1} \\ F_2 / \sqrt{\lambda_2} \\ \vdots \\ F_{13} / \sqrt{\lambda_{13}} \end{pmatrix}$$

其中 $A = U \Lambda^{\frac{1}{2}} = (a_1, a_2, \cdots a_{13}) = (\sqrt{\lambda_1} U_1, \sqrt{\lambda_2} U_2, \cdots, \sqrt{\lambda_{13}} U_{13})$ 为 X 在 F 上的载荷阵。将数据代入公式中, 计算出因子载荷阵 A , 以阐明各 x_i 的方差在各主成分上的载荷:

$$A = U \Lambda^{\frac{1}{2}} = \begin{pmatrix} 0.04 & 0.15 & -0.22 & 0.06 & -0.12 & 0.01 & -0.27 & 0.38 & -0.22 & -0.15 & -0.47 & 0.01 & 0.63 \\ -0.07 & 0.11 & 0.17 & -0.18 & -0.04 & 0.32 & -0.16 & 0.24 & 0.39 & 0.09 & 0.08 & -0.71 & 0.24 \\ -0.14 & 0.04 & 0.02 & -0.28 & -0.04 & 0.18 & -0.03 & -0.07 & -0.44 & 0.36 & -0.46 & 0.39 & -0.42 \\ 0.09 & -0.02 & 0.27 & 0.08 & 0.12 & -0.02 & 0.05 & -0.09 & 0.14 & -0.24 & -0.81 & 0.17 & 0.36 \\ 0.07 & -0.01 & 0.03 & -0.09 & -0.20 & -0.25 & -0.46 & -0.16 & 0.32 & 0.29 & -0.12 & 0.39 & -0.55 \\ 0.00 & 0.17 & 0.18 & 0.00 & -0.04 & -0.15 & -0.21 & -0.36 & -0.50 & -0.09 & 0.33 & -0.28 & 0.55 \\ 0.09 & -0.20 & -0.14 & -0.14 & 0.26 & -0.01 & 0.20 & -0.22 & -0.07 & 0.14 & -0.19 & -0.79 & 0.27 \\ -0.11 & 0.11 & -0.09 & 0.04 & -0.16 & -0.28 & 0.29 & -0.25 & 0.23 & 0.35 & -0.34 & -0.54 & 0.36 \\ 0.06 & -0.08 & 0.08 & -0.30 & 0.00 & -0.36 & 0.17 & 0.40 & -0.05 & 0.14 & 0.17 & 0.21 & 0.70 \\ -0.26 & -0.14 & 0.02 & 0.15 & 0.12 & -0.05 & -0.25 & 0.02 & 0.11 & 0.11 & 0.08 & 0.41 & 0.78 \\ -0.08 & -0.22 & 0.12 & 0.09 & -0.29 & -0.10 & -0.04 & 0.22 & -0.30 & -0.14 & -0.16 & -0.62 & -0.51 \\ -0.14 & 0.09 & -0.05 & -0.18 & 0.17 & -0.28 & -0.06 & 0.01 & 0.13 & -0.67 & -0.07 & -0.14 & -0.57 \\ -0.01 & -0.12 & -0.07 & -0.21 & -0.30 & 0.20 & 0.07 & -0.35 & 0.10 & -0.46 & 0.00 & 0.27 & 0.63 \end{pmatrix}$$

由于 x_i 的方差均为 1, 则用下面模型

$$\begin{cases} v_{ij} = \mu_{ji}^2 \lambda_j \\ v_i(l) = \sum_{j=1}^l \mu_{ji}^2 \lambda_j \end{cases} (i = 1, 2, \cdots 13)$$

表示 F_j 对 x_i 和前 l 个主成分对 x_i 的方差贡献率。

用上述模型计算出 F_{13} , F_{12} , F_{11} , F_{10} , F_9 , F_8 , F_7 分别承载了 x_{13} 方差的

$$v_{13,1} = 0.63^2 = 39.6\% \quad v_{13,2} = 0.01^2 = 0.01\% \quad v_{13,3} = (-0.47)^2 = 22.5\% \quad v_{13,4} = (-0.15)^2 = 2.25\%$$

$$v_{13,5} = 0.22^2 = 4.84\% \quad v_{13,6} = 0.38^2 = 14.7\% \quad v_{13,7} = 0.27^2 = 7.26\%$$

$F_{13}, F_{12}, F_{11}, F_{10}, F_9, F_8, F_7$ 共同承载了 x_{13} 方差的 $v_{13(7)} = 91.1\%$ 。同理可计算 $v_{12(7)} = 81.4\%$ $v_{11(7)} = 86.6\%$ $v_{10(7)} = 89.6\%$ $v_{9(7)} = 88.3\%$ $v_{8(7)} = 91.7\%$ $v_{7(7)} = 84.5\%$ 这就说明 $F_{13}, F_{12}, F_{11}, F_{10}, F_9, F_8, F_7$ 不但能反映 X 变化信息的 85.3915%，而且各 x_i 的方差在 $F_{13}, F_{12}, F_{11}, F_{10}, F_9, F_8, F_7$ 上的载荷均很高，从而用 $F_{13}, F_{12}, F_{11}, F_{10}, F_9, F_8, F_7$ 简化的系统是可以的。

4.2.3 酿酒葡萄分级的评价系统的求解

用主成分可计算各样品的每个主成分值,将每个样品的所有主成分值加到一起即为该样品的综合得分。

建立样品关于 7 个主成分的主成分值的综合评价模型为

$$W_i = \sum_{j=1}^{28} (X * U)_{ij} \quad i = (1, 2, \dots, 28)$$

其中： x 为 28 个酿酒葡萄（白）的标准化结果矩阵， U 为 7 个主成分的各个指标的单位化特征向量矩阵。 w_i 为第 i 个样品的综合得分。

由于酿制红葡萄酒和白葡萄酒的酿酒葡萄的理化指标相同，因此用同样的方法建立类似的模型来分析 27 个酿酒葡萄（红）的综合得分。用 MATLAB 计算结果如表 3 所示。

表3. 27 个酿酒葡萄（红）和 28 个酿酒葡萄（白）的综合得分

样品序号	酿酒葡萄（白）	酿酒葡萄（红）
1	-2.7726	-3.4819
2	4.0311	1.8823
3	-2.23	4.0794
4	-0.049	-0.9115
5	4.0297	1.2
6	-3.6802	-0.0854
7	-1.5322	-6.4839
8	-3.2648	-0.0269
9	-0.0265	6.4997
10	1.8045	2.6692
11	-3.187	-5.0333
12	0.5718	-4.9215
13	-7.1146	3.4764
14	1.7614	3.6958
15	-5.009	-2.1414
16	0.198	-2.538
17	-1.7178	0.0982
18	-4.666	-3.9351
19	1.2693	1.4832
20	-0.1677	2.0683
21	4.5765	-1.6575
22	0.882	-2.8137
23	3.8774	5.2773
24	0.3947	-2.3314
25	3.6651	2.2013
26	7.6107	2.0088

27 -0.1371 -0.2778
28 0.8823

将酿酒葡萄的综合得分按降序排列后用 EXCEL 做出酿酒葡萄的综合条形图如图 1 和图 2 所示。

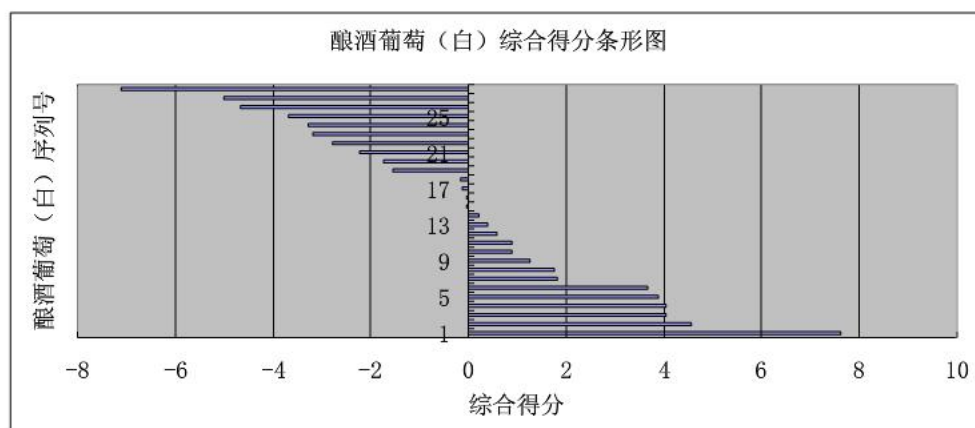


图1. 酿酒葡萄（白）综合得分条形图

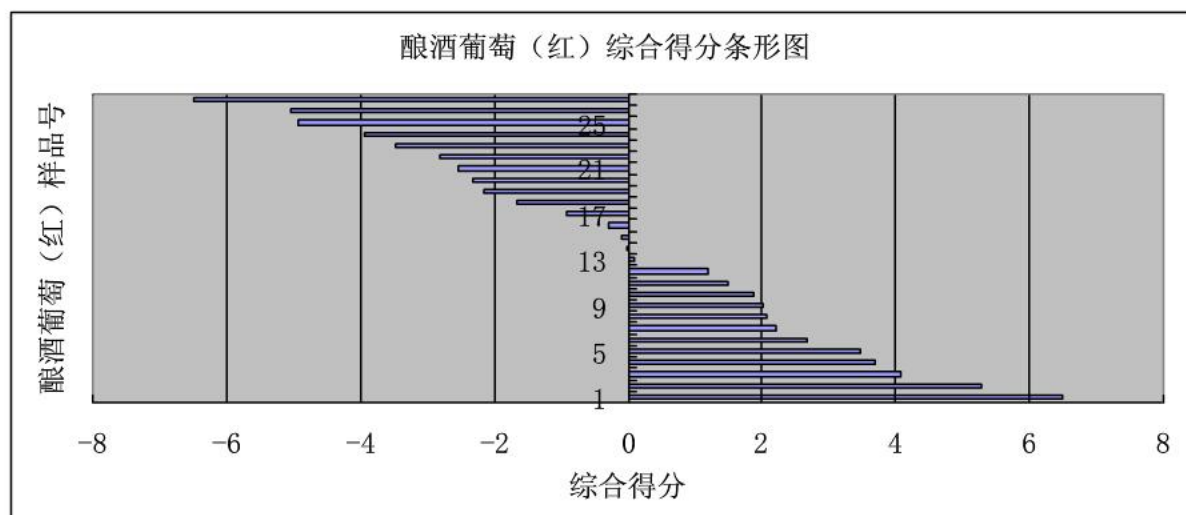


图2. 酿酒葡萄（红）综合得分条形图

分析图 1 和图 2，发现大部分的酿酒葡萄的质量落在 $[-4,4]$ 之间，只有少部分的葡萄在 $[-8,-4]$ 和 $[4,8]$ 之间。这也正验证了现实的酿酒葡萄的种植情况，极好的葡萄由于难以种植因此数量较少，极差的葡萄由于某些葡萄酒的理化指标需要又必须种植，一般的葡萄由于成本和品质同时可以满足需求，故会大量种植。

综上分析可以将酿酒葡萄分为四个等级：优 $[4,8]$ 、良 $[0,4]$ 、中 $[-4,0]$ 、差 $[-8,-4]$ 。则对表 5 中的酿酒葡萄的得分进行分级，即得酿酒葡萄的分级如表 4 所示。

表4. 酿酒葡萄的分级

级别	酿酒葡萄（白）	酿酒葡萄（红）
优	26,21,2,5	9,23,3
良	23,25,10,14,19, 28,22,12,24,16	14,13,10,25,20,26,2,19,5,17
中	9,4,27,20,7,17,3,1,11,8,6	8,6,27,4,21,15,24,16,22,1,18
差	18,15,13	12,11,7

4.3 问题三的分析

问题三要分析酿酒葡萄与葡萄酒的理化指标的联系，关键是将葡萄酒的理化指标与酿酒葡萄的理化指标对应起来，进行相关性分析。首先，采用第二问的模型，分析酿酒葡萄的各个理化指标的相关性。然后，对酿酒葡萄和葡萄酒的共有的理化指标进行相关性分析，再结合已求得的酿酒葡萄的理化指标之间的相关性，即可得出酿酒葡萄与葡萄酒理化指标间的相关性。

4.3.1 酿酒葡萄指标的分类

建立同第二问类似的相关系数的模型，对酿酒葡萄（红）的 30 个理化指标进行相关性分析，相关系数见附表 4。

同样的建立同第二问类似的主成分分析的模型，得出 F_{30} 、 F_{29} 、 F_{28} 、 F_{27} 、 F_{26} 、 F_{25} 、 F_{24} 、 F_{23} 、 F_{22} ，并且这七个主成分可以反映这 30 个酿酒葡萄理化指标的 86.2%，可以用于分析这 30 个理化指标的相关性。

F_{30} 主要反映了总酚、花色苷鲜重、DPPH 自由基、单宁的信息；

F_{29} 主要反映了干物质含量、总糖、还原糖、可溶性固形物的信息；

F_{28} 主要反映了白藜芦醇、果皮颜色 A 的信息；

F_{27} 主要反映了褐变度、PH 值、苹果酸、多酚氧化酶活力、氨基酸总量的信息；

F_{26} 主要反映了果穗质量的信息；

F_{24} 主要反映了果皮质量、黄酮醇的信息；

F_{23} 主要反映了酒石酸、柠檬酸、果皮颜色 L 的信息；

F_{22} 主要反映了 VC 含量、出汁率、葡萄总黄酮的信息；

因此可以按主成分将 30 个理化指标分成 8 类。如表 5 所示。

表5. 指标的分类

类别	指标的种类
1	总酚、花色苷鲜重、DPPH 自由基、单宁
2	干物质含量、总糖、还原糖、可溶性固形物
3	白藜芦醇、果皮颜色 A
4	褐变度、PH 值、苹果酸、多酚氧化酶活力、氨基酸总量
5	果穗质量
6	果皮质量、黄酮醇
7	酒石酸、柠檬酸、果皮颜色 L
8	VC 含量、出汁率、葡萄总黄酮

4.3.2 酿酒葡萄与葡萄酒指标的相关性分析

下面分析这 8 类指标与葡萄酒指标之间的相关性。

首先，将酿酒葡萄与葡萄酒的指标中的相同的指标找出来，分别为：花色苷、单宁、总酚、DPPH 半抑制体积（自由基）、葡萄（酒）总黄酮、白藜芦醇、果皮颜色（色泽）L、果皮颜色（色泽）A、果皮颜色（色泽）B。

先分析红酒的 30 个酿酒葡萄的理化指标的相关系数表里可得出各类指标内各指标

的相关性，同时结合酿酒葡萄与葡萄酒中相同的指标，选取有葡萄酒的理化指标的四类，从附表 3 的相关系数表可以得到这四类指标间的各个指标间的相关系数。如表 6 所示。

表6. 各类内各指标的相关性

相关系数	花色苷鲜重	单宁	总酚	DPPH 自由基
花色苷鲜重	1	0.688131653	0.727828016	0.654607258
单宁	0.688131653	1	0.754693327	0.645011318
总酚	0.727828016	0.754693327	1	0.85712
DPPH 自由基	0.654607258	0.645011318	0.85712	1

相关系数	果皮颜色 L	相关系数	葡萄总黄酮	相关系数	白藜芦醇
酒石酸	-0.19709	VC 含量	-0.12923	果皮颜色 A	0.738243
柠檬酸	-0.13178	出汁率	0.568325		

为了分析酿酒葡萄与葡萄酒的理化指标之间的关系，还要找出两者之间的相关性，求得的相关系数如表 7 所示。

表7. 酿酒葡萄与葡萄酒的共有的理化指标的相关系数

葡萄酒指标（红）	酿酒葡萄指标（红）	相关系数
花色苷	花色苷鲜重	0.922628519
单宁	单宁	0.718040214
总酚	总酚	0.875139176
DPPH 半抑制体积（IV50）	DPPH 自由基 1/IC50	0.778280088
酒总黄酮	葡萄总黄酮	0.822833317
白藜芦醇	白藜芦醇	0.013515055
色泽 L*(D65)	果皮颜色 L	0.493648232
色泽 A*(D65)	果皮颜色 A(红绿)	-0.541953482
色泽 B*(D65)	果皮颜色 B（黄蓝）	0.025494318

分析表 9 可知，葡萄酒的花色苷与酿酒葡萄的花色苷呈强正相关，分析表 8 可知，酿酒葡萄的花色苷与酿酒葡萄的单宁、总酚、DPPH 自由基呈强正相关，因此可得到葡萄酒的花色苷与酿酒葡萄的花色苷鲜重、单宁、总酚、DPPH 自由基呈正相关。同理可分析出葡萄酒（红）的理化指标与酿酒葡萄的理化指标之间的相关性，如表 8 所示。

表8. 葡萄酒（红）的与酿酒葡萄的理化指标之间的相关性

葡萄酒（红）	酿酒葡萄	相关性
花色苷	花色苷鲜重	强正相关
	单宁	强正相关
	总酚	强正相关
	DPPH 自由基	强正相关
酒总黄酮	葡萄总黄酮	强正相关
	VC 含量	弱负相关
	出汁率	正相关
单宁	单宁	强正相关
	花色苷鲜重	强正相关
	总酚	强正相关
	DPPH 自由基	强正相关
白藜芦醇	白藜芦醇	弱正相关
	果皮颜色 A	强正相关
总酚	总酚	强正相关

	花色苷鲜重	强正相关
	单宁	强正相关
	DPPH 自由基	强正相关
色泽 L	果皮颜色 L	正相关
	酒石酸	弱负相关
色泽 A	果皮颜色 A	负相关
	白藜芦醇	负相关
色泽 B	果皮颜色 B	弱正相关
DPPH 半抑制体积	DPPH 自由基	强正相关
	总酚	强正相关
	花色苷鲜重	强正相关
	单宁	强正相关

采用同样的方法分析出葡萄酒（白）的理化指标与酿酒葡萄的理化指标之间的相关性，如表 9 所示。

表9. 葡萄酒（白）的与酿酒葡萄的理化指标之间的相关性

葡萄酒	酿酒葡萄	相关性
单宁	单宁	强正相关
	出汁率(%)	强负相关
总酚	总酚	强正相关
	葡萄总黄酮	强正相关
酒总黄酮	葡萄总黄酮	强正相关
	总酚	强正相关
白藜芦醇	白藜芦醇	弱负相关
	酒石酸	弱负相关
DPPH 半抑制体积	DPPH 自由基 1/IC50	强正相关
色泽 L	果皮颜色 L	负相关
色泽 A	果皮颜色 A(红绿)	负相关
	果皮颜色 B（黄蓝）	正相关
	固酸比	负相关
	VC 含量	弱正相关
	可滴定酸	正相关
色泽 B	果皮颜色 B（黄蓝）	正相关
	果皮颜色 A(红绿)	负相关
	固酸比	负相关
	VC 含量	弱正相关
	可滴定酸	正相关

4.4 问题四的分析

问题四要分析酿酒葡萄与葡萄酒的理化指标的对葡萄酒质量的影响，关键是明确葡萄酒的质量与葡萄酒和酿酒葡萄的理化指标的相关性。首先，根据第三问的模型分析出酿酒葡萄与葡萄酒理化指标间既有正相关也有负相关的结论，用多元回归分析法分别分析葡萄酒和酿酒葡萄的理化指标对葡萄酒质量的影响，即可得出葡萄酒和酿酒葡萄的理化指标对葡萄酒的质量的影响，并对结论加以论证。

4.4.1 酿酒葡萄和葡萄酒的理化指标对葡萄酒质量影响的分析

由问题一的结论可知，第二组专家的评定较使人信服，因此可以将第二组专家的评分结果作为葡萄质量的标准，再采用多元回归分析法定量地分析葡萄及葡萄酒的理化指标对葡萄酒的质量影响。

首先，分析酿酒葡萄（红）的理化指标与葡萄酒质量之间的关系。记第二组评酒员对每种葡萄酒的评分结果总分为 y ，将第三问中得出的酿酒葡萄（红）主成分 F_{30} 、 F_{29} 、 F_{28} 、 F_{27} 、 F_{26} 、 F_{25} 、 F_{24} 、 F_{23} 、 F_{22} 为 x_9 、 x_8 、 x_7 、 x_6 、 x_5 、 x_4 、 x_3 、 x_2 、 x_1 。然后，用 SAS 软件的“分析家”对它们进行带参数的拟合，得到此模型的 F 统计量的值为 4.98，其 p 值 $<0.0001 < \alpha = 0.05$ ，所以模型的作用是显著的。在 SAS 得出的汇总信息中的参数估计部分列举了回归方程中两个参数的值以及有关的显著性检验的结果，详见附件 5.1。拟合的回归方程为

$$y = 73.99777 + 0.35113 x_1 - 0.09471 x_2 + 0.05248 x_3 + 0.30948 x_4 + 0.59193 x_5 + 1.19042 x_6 - 0.58373 x_7 - 0.60545 x_8 - 0.69916 x_9,$$

再进行 t 检验，得出的结果表明 x_9 、 x_8 、 x_7 、 x_6 、 x_5 、 x_4 、 x_3 、 x_2 、 x_1 的系数显著不为 0，即可以得出红葡萄的理化指标对葡萄酒的质量有很大的影响。

同样可以得出第三问得出的白葡萄的主成分与葡萄酒质量之间的非线性拟合回归方程为

$$y = 77.30179 - 0.10958 x_1 + 0.43095 x_2 - 0.91572 x_3 - 0.10765 x_4 - 0.09802 x_5 + 0.40723 x_6 + 0.03551 x_7 - 0.22512 x_8 + 0.51521 x_9 - 0.03911 x_{10} + 0.73002 x_{11},$$

且其 F 统计量的值为 1.42，其 p 值 $<0.0001 < \alpha = 0.05$ ，所以模型的作用是显著的。再进行 t 检验，得出的结果表明 x_{11} 、 x_{10} 、 x_9 、 x_8 、 x_7 、 x_6 、 x_5 、 x_4 、 x_3 、 x_2 、 x_1 的系数显著不为 0，即可以得出白葡萄的理化指标对葡萄酒的质量有很大的影响。

设花色苷、单宁、总酚、酒总黄酮、白藜芦醇、DPPH 半抑制体积、色泽中的 L、色泽中的 A、色泽中的 B 分别为自变量 x_9 、 x_8 、 x_7 、 x_6 、 x_5 、 x_4 、 x_3 、 x_2 、 x_1 。然后采用 SAS 软件可以得出其 F 统计量的值为 2.48，其 p 值 $<0.0001 < \alpha = 0.05$ ，所以模型的作用是显著的。且其非线性拟合回归方程为

$$y = 82.08040 - 0.01827 x_1 + 0.79292 x_2 - 0.46037 x_3 + 0.57254 x_4 + 0.44832 x_5 - 9.30909 x_6 - 0.17805 x_7 - 0.06118 x_8 - 0.06544 x_9,$$

再进行 t 检验，得出的结果表明 x_9 、 x_8 、 x_7 、 x_6 、 x_5 、 x_4 、 x_3 、 x_2 、 x_1 的系数显著不为 0，即可以得出红葡萄酒的理化指标对葡萄酒的质量有很大的影响。

设花色苷、单宁、总酚、酒总黄酮、白藜芦醇、DPPH 半抑制体积、色泽中的 L、色泽中的 A、色泽中的 B 分别为自变量 x_8 、 x_7 、 x_6 、 x_5 、 x_4 、 x_3 、 x_2 、 x_1 。然后采用 SAS 软件可以得出其 F 统计量的值为 2.35，其 p 值 $<0.0001 < \alpha = 0.05$ ，所以模型的作用是显著的。且其非线性拟合回归方程为

$$y = 82.08040 + 0.01312 x_1 - 0.05596 x_2 - 0.50294 x_3 - 1.14900 x_4 + 29.34657 x_5 - 0.81747 x_6 + 0.79763 x_7 + 0.32694 x_8$$

再进行 t 检验，得出的结果表明 x_8 、 x_7 、 x_6 、 x_5 、 x_4 、 x_3 、 x_2 、 x_1 的系数显著不为 0，即可以得出白葡萄酒的理化指标对葡萄酒的质量有很大的影响。

综上所述可知酿酒葡萄与葡萄酒的理化指标对葡萄酒质量的评定有很大的影响，并

可以通过指标之间的线性组合得到葡萄酒质量的生成函数。

4.4.2 用酿酒葡萄和葡萄酒的理化指标评价葡萄酒质量的论证

由上面的分析可知酿酒葡萄与葡萄酒的理化指标对葡萄酒的质量有很大影响，这就说明，在对葡萄酒的质量进行评价时，是可以利用酿酒葡萄和葡萄酒的理化指标来实现。首先，由于现代科学技术的发展，测量仪器的精密性有力的保证了测量的葡萄酒和酿酒葡萄的理化指标的准确性。

其次，采用测定理化指标来分析葡萄酒的质量，不仅可以避免品酒员的主观性错误，还可以节省测定的时间，同时也可以减少劳动力的浪费。

最后，采用测定理化指标来评定葡萄酒的质量有效的避免了人为的操作性，使得结果更能令人信服。

五 模型优缺点及改进方向

5.1 模型的优点

1. 模型所运用了方差分析法、主成分分析法、相关性分析等统计方法对各种数据进行了全面分析。
2. 模型的实施可操作性强，可运用到类似的统计学问题中，模型中对主成分分析法的验证使分析方法更为合理。

5.2 模型的缺点

运用模型解决问题时，仅有小部分样品的数据，具有一定的局限性，另外不能保证参考文献的可信性，从而不能保证模型中简化指标的科学性。

5.3 模型的改进方向

我们的模型在保证指标能尽量完整的基础上简化理化指标的评价系统，使得问题的数学描述更为简单直观，但实际上还有部分指标与葡萄酒的质量密切相关，而且是应该加以考虑的。如选取更多的指标，进行更多的理化指标间的相关性分析等，但由于时间关系，我们只是讨论了可以代表所有指标的部分指标，选取更多的指标，进行更多的指标间的相关性分析，这是我们努力的方向。

六 参考文献

- [1] 韩中庚, 数学建模竞赛——获奖论文精选与点评, 北京: 科学出版社, 2007.5;
- [2] 袁志发, 周静芋, 多元统计分析, 北京: 科学出版社, 2002.9;
- [3] 戴明强 李卫军 杨鹏飞, 数学建模及其应用, 北京: 科学出版社, 2007, 169-186;
- [4] 史道济 张玉环, 应用数理统计, 天津: 天津大学出版社, 2008, 13-25;
- [5] 徐全智 杨晋浩, 数学建模, 北京: 高等教育出版社, 2008;
- [6] 汪远征, 徐雅静, SAS 软件与统计应用教程, 北京: 机械工业出版社, 2012;
- [7] 茆诗松, 程依明, 濮小龙, 概率论与数理统计教程, 北京: 高等教育出版社, 2004.7;
- [8] 李记明, 关于葡萄酒品质的评价指标, 百度文库, 712100: 1-4, 1999;
- [9] 中国葡萄酒旅游网, 葡萄酒的营养以及多酚物质的保健作用, <http://www.9winetour.com/Culture/2007-12-07/29791.html>, 2012.9.8。

七 附 录

附录见附件。