葡萄酒的评价

摘　要

本文针对葡萄酒的评价问题，利用 K-S 非参数检验方法并建立了基于α信度系数的可靠性评价模型，判断了两组评酒员的评价结果有无显著性差异以及评分的可靠性的问题；基于主成分分析法的酿酒葡萄分类模型，解决了酿酒葡萄的分类问题；建立了基于粒子群优化算法的偏最小二乘回归分析模型，分析酿酒葡萄与葡萄酒的理化指标之间的联系；建立了基于灰色关联度分析的葡萄酒质量排序模型，解决了酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响问题，并论证了能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量问题。

针对问题一，利用 K-S 非参数检验方法并建立了基于α信度系数的可靠性评价模型，判断了两组评酒员的评价结果有无显著性差异以及评分的可靠性的问题。首先，初步筛选掉某些明显偏离其他正常数据的异常值，并利用 K-S 非参数检验方法判断样本是否属于正态分布；其次用 t 检验方法判段两组数据之间是否存在显著性差异；最后建立基于α信度系数的可靠性评价模型，判断两组专家对两种酒的可靠性程度。最终得到两组评酒员的结果都无显著性差异；第一组和第二组白葡萄酒评价结果均是可靠的，但第二组的评价结果更可靠。

针对问题二，基于主成分分析法的酿酒葡萄分类模型，解决了酿酒葡萄的分类问题。首先对数据进行标准化，消除量纲影响。其次对各个理化指标进行相关性分析并提取主成分；最后根据提取出的主成分对酿酒葡萄进行聚类分析，并将结果与品酒员的打分相比较，验证模型的可靠性。最终得到红葡萄酒可分为四类，样品编号分别为 1、2、3、9、23，4、5、6、7、8、12、13、14、15、16、17、18、19、20、21、22、24、25、26、27，10，11；白葡萄酒分为 4 类，样品编号分别为：1、6、7、13、15、18，2、3、4、5、8、9、10、11、12、14、17、19、20、21、22、23、24、25、26、28，16，27。

针对问题三，建立了基于粒子群优化算法的偏最小二乘回归分析模型，分析酿酒葡萄与葡萄酒的理化指标之间的联系。首先，利用第二问的方法提取出葡萄酒理化指标的主成分；其次建立偏最小二乘回归分析模型得到酿酒葡萄与葡萄酒理化指标之间的相关系数；最后利用粒子群算法对已求得的系数进行优化，并对模型进行精度检验。最终得到在红葡萄酒中总酚与单宁、柠檬酸、果穗质量、颜色、酒石酸；色泽与葡萄总黄酮、柠檬酸的的相关性较大；在白葡萄酒中总酚与柠檬酸、褐变度、黄酮醇、干物质含量；白藜芦醇与花色苷、总酚、白藜芦醇；色泽与柠檬酸、黄酮醇相关性较大。优化后模型的精度大大提高。

针对问题四，建立了基于灰色关联度分析的葡萄酒排序模型，解决了酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响问题，并论证了能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量。首先分析评分标准与理化指标相关程度，并将指标进行分类；其次根据每类评分所占百分比确定每类指标的权重；然后计算理化指标的灰色关联度，并据此葡萄酒按照质量进行排序；最后将我们的排序结果与专家评分的排序结果对比，验证模型的可靠性。最终得到黄酮醇、总糖、白藜芦醇、酒石酸、百粒质量、褐变度和 1-庚醇会对红葡萄酒的质量存在显著的联系；总糖、总酚、花色苷、PH 值、干物质含量和辛酸 3-甲基丁酯会对白葡萄酒的质量存在显著联系。葡萄和葡萄酒的理化指标可以评价葡萄酒的质量。

关键词：K-S 非参数检验方法　聚类分析　粒子群优化算法　最小二乘回归分析　灰色关联度

一、问题重述

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。

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。请尝试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？

附件 1：葡萄酒品尝评分表（含 4 个表格）
附件 2：葡萄和葡萄酒的理化指标（含 2 个表格）
附件 3：葡萄和葡萄酒的芳香物质（含 4 个表格）

二、问题分析

2.1 问题一的分析

分析两组评酒员的评价结果有无显著性差异，就是要对评价结果进行显著性检验。首先我们要利用 K-S 方法分析题目中所给的相关数据是否服从正态分布，而显著性检验的常见方法有 t 检验、U 检验、方差检验[1]等。对于服从正态分布的数据，我们一般对其进行 t 检验[2]，观察其假设方差是否相等，根据其假设方差的情况，找出均值方程的 t 检验中的 Sig 值，判定数据之间是否存在显著性差异。接着，我们需要分析评价过程中可能产生误差的原因[3]才能对其结果的可靠性进行判断。在这种情况下，可靠性判断通常通过α信度系数大小和评价结果的方差[4]来判断，以此对红、白葡萄酒的第一、二组数据进行可靠性分析。

2.2 问题二的分析

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，通常采用的分级方法有聚类分析法、判别分析法等[5]。首先，我们要对多次测量的数据取平均值，为了便于研究，还要将数据进行标准化。数据标准化常用的方法有 min-max 标准化、Z 标准化等。我们选择将其进行 Z 标准化因为所给数据理论上应符合正态分布[6]。其次，由于原始变量多而杂，我们选择对题目中所给的变量进行主成分分析提取主要因素以减少研究的复杂度。根据提取出的主成分对红、白葡萄酒进行聚类分析[7]。一般情况下我们选用系统聚类进行研究，根据树状图将葡萄酒进行分类。最后，利用主成分分析对红、白葡萄酒进行打分，与品酒员的打

分情况进行比较，验证模型的可靠性。

2.3 问题三的分析

分析酿酒葡萄与葡萄酒的理化指标之间的联系，实际上就是要找到酿酒葡萄与葡萄酒之间的确切关系式。首先，我们根据之前的主成分分析结果，能够得到酿酒葡萄与葡萄酒的主要理化指标，通过简化的变量进一步找出它们之间的联系。其次，我们要求出酿酒葡萄与葡萄酒的理化指标之间的系数。对于这种问题我们通常采用的是偏最小二乘回归分析法[8]，找出变量之间的近似关系。最后，再对所得的参数进行优化，得到更加准确的结果。

2.4 问题四的分析

想要合理分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，首先我们需要对酿酒葡萄和葡萄酒的理化指标进行提取，找出其主要因素。其次，找出指标与葡萄酒质量之间的关联度。一般情况下我们选用灰色关联度算法[9]进行求解，得出其关联系数。最后，根据所得结果与附件一中的品酒员打分进行对比，验证方法的可行性。

三、基本假设

1、葡萄酒中芳香物质具有很大的挥发性，性质不稳定；
2、测量数据的过程中忽略温度等环境因素对葡萄酒质量的影响；
3、各酒样品的酿酒工艺技术相同；
4、品酒员打分过程中互不影响，并能够反应主观印象。

四、符号说明

α	信度系数
$f(x)$	样本概率密度函数
μ	样本均值
σ	样本标准差
β	显著性水平
r_{ij}	相关系数
P_i	粒子
$J(\theta)$	适应度函数
x_i^T	位置矢量
v_i^{T+1}	速度矢量
Δs^T	粒子的最小允许间距
$p_{best\ i}^T$	个体极值点

g_{best}^{T}	全局极值点
ω	权重
p_i	偏差值
$\xi_i(k)$	关联系数
ρ	分辨系数
r_i	灰色加权关联度

五、模型的建立与求解

5.1 数据预处理

由于葡萄酒的测评是专家根据其自身的口味、喜好等进行评价的，因此过程中会存在由于专家的主观因素而产生的误差，使得给出的评价分数过高或过低。并且数据也可能存在读数误差以及录入错误。因此，我们首先对附件中的数据进行预处理，初步筛选掉某些明显偏离其他正常数据的异常值。

5.2 问题一模型的建立与求解

首先，我们采用显著性检验的方法分析两组评酒员的评价结果有无显著性差异。利用 K-S 方法和 t 检验分析题目中所给的相关数据，再根据其假设方差的情况，找出均值方程的 t 检验中的 Sig 值，判定数据之间是否存在显著性差异。其次，我们找出评价过程中误差产生的原因。最后，基于α信度系数的大小和评价结果的方差大小，判断红、白葡萄酒的第一、二组数据的可靠性程度。

5.2.1 模型的建立

1、两组评酒员的评价结果有无显著性差异

显著性差异是统计学中对大量数据中存在的某些具有显著性差异的数据的评价。当数据之间存在显著性差异时，说明参与对比得数据不是来自于同一总体或来自同一总体的数据实验处理条件不一样。显著性差异分析方法就是基于大量实验数据的统计结果设定一个静态阈值从而达到对异常信号的预警作用[10]。

(1)显著性分析前提：

利用显著性分析的方法对葡萄酒评价样本数据进行处理、分析需要满足以下前提：

① 参与统计的样本数据必须来源于同一总体；

② 样本数据分布基本符合正态分布规律，即样本概率密度函数应满足正态分布函数[11]其计算公式如下：

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

其中，x 为样本值；f 为样本在该值出现的概率密度； μ 为样本均值；σ为样

本标准差； μ 和σ的计算公式如下：

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

根据以上要求，我们进行显著性分析的步骤如图 1 所示：

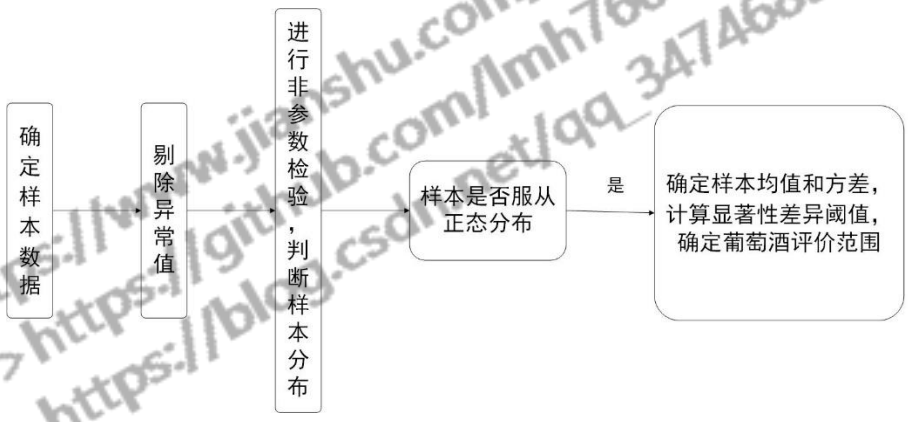

图 1 显著性检验步骤

(2) 非参数检验

样本的非参数检验采用 Kolmogorov-Smirnov 检验(K-S 检验)方法进行[12]。在统计学中，Kolmogorov-Smirnov 检验基于累计分布函数，用以检验两个经验分布是否不同或一个经验分布与另一个理想分布是否不同[13]。K-S 检验比较样本数据的累计频数分布和特定理论分布，若两者间的差距小于要求值，则推论该样本取自该分布。单样本 K-S 检验可以将一个样本的实际频数分布与正态分布、均匀分布、泊松分布、指数分布进行比较。样本的非参数检验过程如下：

首先建立假设：

$$H_0: F(x) = F_0(x)$$

其中，$F_0(x)$为连续型分布函数，对于正态分布$F_0(x)$为：

$$F_0(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

检验真伪的统计量：

$$D_n = \underset{-\infty<x<\infty}{\mathrm{SUP}} F_n(x) - F_0(x)$$

其中，$F_n(x)$为容量是 n 的样本的经验分布函数。

$$F_n(x) = \begin{cases} 0 & x < X_{(1)} \\ \dfrac{i}{n} & X_{(i)} < x < X_{(i+1)} \\ 1 & x < X_{(n)} \end{cases}$$

其中，$X_{(1)}, X_{(2)}, \ldots, X_{(i)}$是样本$(x_1, x_2, \ldots, x_n)$的次序统计量。

K-S 检验是在样本的每个次序统计量$X_{(i)}$内，求样本经验分布函数与假设的

分布函数之间的偏差中最大的一个，即：

$$d_i = \max\left\{F_0(X_{(i)}) - \frac{i-1}{n}, \frac{i}{n} - F_0(X_{(i)})\right\}, i = 1,2,\dots,n$$

求得的 n 个d_i中最大的一个，就是 K-S 检验统计量D_n的取值。

K-S 检验的检验规则为：

$D_n > D_{(n,\beta)}$时，拒绝原假设H_0；

$D_n < D_{(n,\beta)}$时，接受原假设H_0。

其中，$D_{(n,\beta)}$为假设检验 K-S 检验的临界值，$D_{(n,\beta)}$可通过查表获取，K-S 检验的显著性水平β一般取 0.05。

如样本数据 K-S 检验结果为正态分布，假设$\frac{x-\mu}{\sigma} = u$，则公式可变为：

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

即 u 服从标准正态分布即u~N(0,1)。在这里我们只需要考虑曲线的单边。

参考 GB/T4883-2008[14]，通过对离群值的分析要求，样本的显著性差异条件确定如下：

$$\frac{x-\mu}{\sigma} > u_{1-\alpha}$$

其中，α为样本的显著性水平，一般取 0.01。分布情况如图 2 所示：

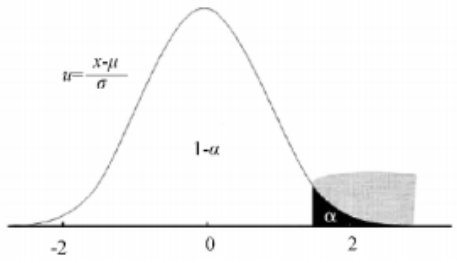

图 2 标准正态分布置信区间图

经查 U 分布表获取在显著水平α值下的置信区间上限值为 2.58。

(3) 显著性检验

如果样本数据全部符合正态分布，那么我们利用 t 检验对红白葡萄酒进行显著性检验。以葡萄酒的酒号分类，分别对第一组和第二组该酒号的相关数据进行 t 检验。若 t 检验中 Levene 检验的 Sig 值小于 0.01，则我们可以认为两组数据的假设方差不相等。在这种情况下观察均值方程的 t 检验中的假设方差不相等的 Sig 值，若 Sig 值大于 0.05，则我们可以认为两组数据无显著性差异。其他情况下我们则可以认为两组数据存在显著性差异。

2、评价结果可靠性判断

在对葡萄酒评价结果的可信性分析中，品酒员的选取是评价结果可信性或者评价公平性的一个重要因素，如果选择不当，评价结果的准确性和可信性就会受到影响。由于评酒方案及评价因素本身的模糊性以及品酒员所掌握信息的有限性和事物认知上的局限性导致品酒员存在“信息结构性偏差”和“认知结构性偏差”

[15]的问题。此外，品酒员主体有追求自身利益偏好等原因，造成评价结果具有随机性和不稳定性。评价结果的可信度实际取决于品酒员的评价能力、素质和问题的难度[16]等。

因此，我们认为可能产生误差的原因主要有以下两个方面：

① 品酒员主体对问题的认知不同；

② 问题本身的模糊性。

鉴于以上误差可能存在的原因，我们认为，要想使评酒的结果可靠，则：

① 不同品酒员对同一种酒的评价基本一致；

② 品酒员对不同酒的评价差异越大越好。

在这里，我们选择利用 SPSS 对葡萄酒的评价结果进行可靠性分析，目前最常用的是α信度系数法。信度评价的是测量结果的前后一致性，反映的是测验受随机误差影响的程度。一般而言，两个或两次测量结果越是一致，则误差越小，信度也就越高。一般情况下主要考虑量表的内在信度——项目之间是否具有较高的内在一致性。α信度系数在 0-1 之间，若$\alpha > 0.7$，则我们认为其信度很好，反之则需要修正。

5.2.2 模型的求解

1、显著性差异分析结果

(1)样本校验

利用 K-S 检验方法对样本进行正态分布校验。校验时，一般用D_n和 n 反推接受原假设显著性水平β值。若$\beta > 0.05$，表明样本接受原假设，服从正态分布[17]。将附件中所给数据分为 4 组，分别为第一组红葡萄酒、第二组红葡萄酒、第一组白葡萄酒、第二组白葡萄酒。利用 K-S 检验方法对剔除异常值后的样本进行正态分布校验，将其校验结果的最大β值整理如表 1 所示：

表 1 校验结果

	第一组红葡萄酒	第二组红葡萄酒	第一组白葡萄酒	第二组红葡萄酒
β的最大值	0.091	0.062	0.066	0.058

通过表 1 我们发现，校验结果的β值分别为 0.091、0.062、0.066、0.058，都大于 0.05，即我们认为样本全部服从正态分布。

(2)显著性检验

由于样本数据全部符合正态分布，接着我们利用 t 检验对红白葡萄酒进行显著性检验。

以红葡萄酒 25 号为例，我们对第一组和第二组的相关数据进行分析整理，结果如表 2 所示：

表 2 红葡萄酒 25 号显著性检验结果

		Levene 检验 Sig.	均值方程的 t 检验 Sig.
var001	假设方差相等	.047	.264
	假设方差不相等		.266
var002	假设方差相等	.759	.517
	假设方差不相等		.517
var003	假设方差相等	.215	.749
	假设方差不相等		.750
var004	假设方差相等	.564	.841

	假设方差不相等		.841
var005	假设方差相等	.561	.548
	假设方差不相等		.548
var006	假设方差相等	.104	.639
	假设方差不相等		.640
var007	假设方差相等	.383	.844
	假设方差不相等		.844
var008	假设方差相等	.867	1.000
	假设方差不相等		1.000
var009	假设方差相等	.533	.799
	假设方差不相等		.799
var010	假设方差相等	.398	.342
	假设方差不相等		.344

通过表 2 我们可以看出，10 个变量中 Levene 检验的 Sig 值均大于 0.01，则我们观察其假设方差相等的 Sig 值。通过观察我们发现，10 个变量假设方差相等的 Sig 值分别为 0.264、0.517、0.749、0.841、0.548、0.639、0.844、1.000、0.799、0.342，均大于 0.05，即我们可以认为红葡萄酒 25 号的两组评价数据无显著性差异。

即红葡萄酒 25 号的两组评价数据中无显著性差异。

接着，我们对白葡萄酒 26 号进行显著性检验，结果如表 3 所示：

表 3 白葡萄酒 26 号显著性检验结果

		Levene 检验 Sig.	均值方程的 t 检验 Sig.
var001	假设方差相等	.044	.109
	假设方差不相等		.113
var002	假设方差相等	.483	.342
	假设方差不相等		.343
var003	假设方差相等	.001	.024
	假设方差不相等		.031
var004	假设方差相等	.007	.004
	假设方差不相等		.006
var005	假设方差相等	.396	.094
	假设方差不相等		.094
var006	假设方差相等	.543	.331
	假设方差不相等		.331
var007	假设方差相等	.882	.063
	假设方差不相等		.063
var008	假设方差相等	.232	.210
	假设方差不相等		.211
var009	假设方差相等	.952	.865
	假设方差不相等		.865
var010	假设方差相等	1.000	.660
	假设方差不相等		.660

同理，通过表 3 我们可以看出，10 个变量中 Levene 检验的 Sig 值均大于

0.01，则我们观察其假设方差相等的 Sig 值。通过观察我们发现，10 个变量假设方差相等的 Sig 值分别为 0.109、0.342、0.024、0.004、0.094、0.331、0.063、0.210、0.865、0.660，基本大于 0.05，即我们可以大致认为白葡萄酒 26 号的两组评价数据无显著性差异。

整理红、白葡萄酒各酒号的显著性结果如表 4、表 5 所示：

表 4 红葡萄酒显著性检验结果

酒号	显著性检验	酒号	显著性检验	酒号	显著性检验
25	无	18	无	8	无
27	无	6	无	12	无
7	无	4	无	5	无
10	无	13	无	23	无
11	无	22	无	15	无
20	无	17	无	26	无
16	无	1	无	9	无
24	无	2	无	21	无
19	无	3	无	14	无

表 5 白葡萄酒显著性检验结果

酒号	显著性检验	酒号	显著性检验	酒号	显著性检验	酒号	显著性检验
26	无	11	无	10	无	22	无
5	无	15	无	7	无	8	无
4	无	14	无	1	无	17	无
23	无	12	无	3	无	9	无
20	无	18	无	16	无	25	无
19	无	13	无	2	无	24	无
28	无	21	无	6	无	27	无

通过表 4、表 5 的数据可以知道，在红、白葡萄酒的评价过程中，两组评酒员的结果都无显著性差异。

2、评价结果可靠性判断

对于红葡萄酒，我们得到其第一组和第二组的可靠性分析结果如表 6、表 7 所示：

表 6 第一组红葡萄酒可靠性

Cronbach's Alpha	基于标准化项的 Cronbachs Alpha	项数
.854	.882	10

表 7 第二组红葡萄酒可靠性

Cronbach's Alpha	基于标准化项的 Cronbachs Alpha	项数
.747	.781	10

通过表 6、表 7 我们发现，第一组和第二组红葡萄酒的α信度系数都大于 0.7，即我们可以认为第一组和第二组红葡萄酒评价结果的可信度较高。接着，我们计算第一组和第二组红葡萄酒评价结果的方差，结果分别如表 8、表 9 所示：

表 8 第一组红葡萄酒方差

	均值	方差
项的均值	7.310	14.889
项方差	2.390	4.174

表 9 第二组红葡萄酒方差

	均值	方差
项的均值	7.057	13.979
项方差	1.466	1.727

通过表 8、表 9 我们可以看出，第一组红葡萄酒评价结果的方差为 4.174，而第二组红葡萄酒评价结果的方差为 1.727，小于 4.174。因此，我们认为第二组红葡萄酒的评价结果更可靠。

同理，第一组和第二组白葡萄酒的可靠性分析结果分别如表 10、表 11 所示：

表 10 第一组白葡萄酒可靠性

Cronbach's Alpha	基于标准化项的 Cronbachs Alpha	项数
.756	.872	10

表 11 第二组白葡萄酒可靠性

Cronbach's Alpha	基于标准化项的 Cronbachs Alpha	项数
.823	.857	10

因此其第一二组白葡萄酒的α信度系数都大于 0.7，我们可以认为第一组和第二组白葡萄酒评价结果是可信的。接着，第一组和第二组白葡萄酒评价结果的方差分别如表 12、表 13 所示：

表 12 第一组白葡萄酒方差

	均值	方差
项的均值	7.482	15.446
项方差	4.254	34.041

表 13 第二组白葡萄酒方差

	均值	方差
项的均值	7.671	17.185
项方差	1.398	1.773

通过表 12、表 13 我们可以看出，第一组白葡萄酒评价结果的方差为 34.041，而第二组白葡萄酒评价结果的方差为 1.773，小于 34.041。因此，我们认为第二组白葡萄酒的评价结果更可靠。

综上所述，可以认为第一组和第二组白葡萄酒评价结果是可靠的，第二组的评价结果更可靠。

5.3 问题二的模型建立与求解

通过观察我们发现，附件中所给的数据多而复杂，因此，我们首先对数据进行标准化。其次，要想根据酿酒葡萄的理化指标和葡萄酒的质量对酿酒葡萄进行分级，就要对各因素进行相关性分析并提取主成分。最后，将葡萄酒进行聚类分析，并将结果与品酒员的打分相比较，验证可靠性。

5.3.1 模型的建立

在处理问题前，我们首先对题目中所给数据进行分析处理。通过对葡萄酒理化指标资料的查找，我们知道白藜芦醇由反式白藜芦醇、顺式白藜芦醇、反式白藜芦醇苷和顺式白藜芦醇苷等物质组成；黄醇酮由杨梅黄酮、槲皮素、山奈酚和异鼠李素组成；氨基酸由苏氨酸、丝氨酸等氨基酸组成；还原糖由葡萄糖和果糖

组成[18]，同时，又因为我们通过对题目中所给二级指标的数据求和后，与一级指标数据进行比较发现其值相同，故最终我们认为建模中只需考虑一级指标的理化性质即可。并且，葡萄酒中的芳香物质具有很大的挥发性，其性质不稳定[19]且在一定程度上也会影响实验的研究，故我们在这里不考虑葡萄酒中芳香物质对实验的影响。

1、数据标准化

由于题目中所给葡萄和葡萄酒的理化指标多而杂，于是我们选择对原始数据进行数据标准化，以消除量纲的影响。在数据标准化之前，我们先对多次测量的理想化指标取平均值。在这里，我们选择对数据进行 Z 标准化。Z 标准化的公式如下：

$$x'_{ij} = \frac{x_{ij} - \bar{x_j}}{S_j}$$

其中，$\bar{x_j}$为第 j 列的平均值，S 为标准差。

2、主成分分析

主成分分析也称为主分量分析，旨在利用降维的思想，把多 转化为少数的几个综合指标。在实际问题研究中，为了全面、系统的分析问题，我们必须考虑众多因素的影响。这些涉及的因素一般称为指标，在多元统计分析中也称为变量。因为每个变量都在不同程度上反映了所研究问题的某些信息，并且指标之间彼此有一定的相关性，因而所得的统计数据反映的信息在一定程度上有重叠。在用统计方法研究多变量问题时，变量太 多会增加计算量和增加分析问题的复杂性，因此，我们通常选用主成分分析法筛选出主要因素，减少变量。因此我们选择主成分分析法对葡萄的理化指标进行处理。

(1) 首先，我们对之前准备的标准化后的数据进行 KMO 和 Bartlett 球形检验，判断标准化后的数据是否能够进行主成分分析。
当 KMO 的检验系数>0.5 并且 Bartlett 球形检验的 P 值<0.05 时，该数据才能够进行主成分分析。

(2) 第二步，选择分析的变量。选择变量时需要用定性分析和定量分析的方法，主成分分析的前提条件是观测变量之间有较强的相关性，因为如果变量之间无相关性或相关性较小，则它们之间就不会存在相关性。因此我们计算所选变量的相关系数矩阵 R 来判断各元素之间的相关系数，其计算公式如下：

$$\mathrm{R} = (r_{ij})_{m\times m}$$

$$r_{ij} = \frac{\sum_{k=1}^{n} \widetilde{a_{kl}} \times \widetilde{a_{kJ}}}{n-1}, \mathrm{i,j} = 1,2,\dots,\mathrm{m}$$

其中，$r_{ii} = 1, r_{ij} = r_{ji}, r_{ij}$是第 i 个指标与第 j 个指标的相关系数。

两个指标之间的相关系数，反映了两个指标之间的相关性。相关系数越大，两个指标反映的信息相关性就越高。

(3) 第三步，进行方差分析提取。在方差主成分分解表中取其特征值大于 1 的因子，再按照累计方差贡献率来确定因子，一般认为累计贡献率达到 60%才能符合要求。

(4) 第四步，根据其成份得分系数矩阵，分析哪些变量可以大致代替原筛选过的变量进行研究分析。

(5) 最后，出主成分表达式，进而求出影响葡萄酒的主要因素。

主成分的表达式如下：

$$F = \frac{\sum_{i=1}^{n} \lambda_i \times F_i}{\sum_{i=1}^{n} \lambda_i}$$

3、聚类分析

聚类分析指将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程。它是一种重要的人类行为。聚类分析的目标就是在相似的基础上通过收集数据来分类。在这里，我们采用系统聚类和树状图对红、白葡萄酒进行聚类分析。

系统聚类的步骤一般是首先根据一批数据或指标找出能度量这些数据或指标之间相似程度的统计量;然后以统计量作为划分类型的依据，把一些相似程度大的变量(或样品)首先聚合为一类，而把另一些相似程度较小的变量(或样品)聚合为另一类，直到所有的变量(或样品)都聚合完毕，最后根据各类之间的亲疏关系，逐步画成一张完整的分类系统图，又称谱系图。其相似程度由距离或者相似系数定义。进行类别合并的准则是使得类间差异最大，而类内差异最小。由此，我们对红、白葡萄酒分别进行聚类分析。其步骤如图 3 所示：

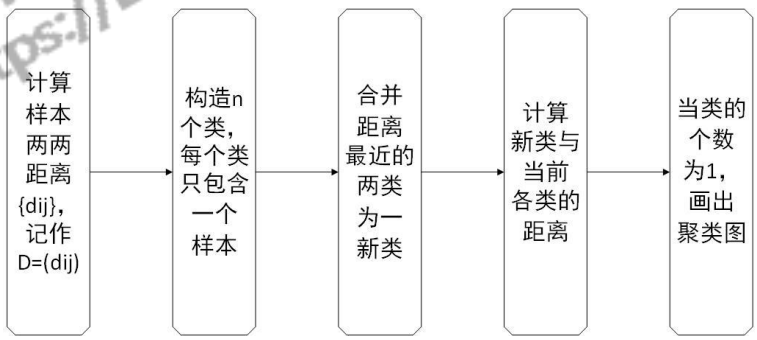

图 3 聚类分析步骤

(1) 计算 n 个样品两两间的距离;
(2) 构造 n 个类，每个类只包含一个样品;
(3) 合并距离最近的两类为一新类;
(4) 计算新类与各当前类的距离;
(5) 重复步骤(3)、(4)，合并距离最近的两类为新类，直到所有的类并为一类为止;
(6) 画聚类谱系图;
(7) 决定类的个数和类。

5.3.2 模型的求解

1、主成分分析结果

(1)相关系数矩阵分析

将标准化处理的数据经过公式计算得到相关系数矩阵,通过该矩阵我们发现蛋白质与 DPPH 自由基、蛋白质与总糖蛋白质与可溶性固形物、蛋白质与 PH 值、花色苷与总酚、DPPH 自由基和总糖、DPPH 自由基和可滴定酸、总酚和单宁、总酚和葡萄总黄酮、单宁和葡萄总黄酮黄酮醇和果梗比、总糖和可溶性固形物、还原糖和干物质含量、可溶性固形物与 PH 值、PH 值与固酸比、可滴定酸与干物质含量、固酸比与出汁率、干物质含量与出汁率、果穗质量与百粒质量、出汁

率与果皮颜色、果皮质量与果皮颜色等之间都有较高的相关性。

(2)方差分析提取

红葡萄酒的方差分解主成分提取表如表 14 所示：

表 14 方差分解主成分提取表

成份	特征值	每个贡献率%	累计贡献率%
1	7.277	25.094	25.094
2	4.766	16.434	41.528
3	3.155	10.880	52.408
4	2.926	10.090	62.498
5	1.973	6.803	69.300
6	1.767	6.094	75.394
7	1.342	4.629	80.023

由表 14 我们可以看出，前 7 个成份的初始特征值都大于 1，且累计贡献率为 80.023%，大于 80%，因此我们认为，可以提取前 7 个因素作为影响葡萄酒质量的主要因素。

(3)计算成份矩阵

经过我们得到的成份矩阵，根据不同指标在 7 个不同主成分上具有的不同载荷，7 个主成分基本上可以反映出全部指标的信息，即可以用 7 个新变量代替原来的 14 个变量。即我们得出红葡萄的理化指标为还原糖 、单宁、葡萄总黄酮、柠檬酸、果穗质量、颜色、酒石酸。

同理，我们得到白葡萄的理化指标为花色苷、柠檬酸、褐变度、总酚、白藜芦醇、黄酮醇、PH 值、干物质含量、颜色。

2、聚类分析结果

将红、白葡萄酒分别进行聚类分析，其结果如下：

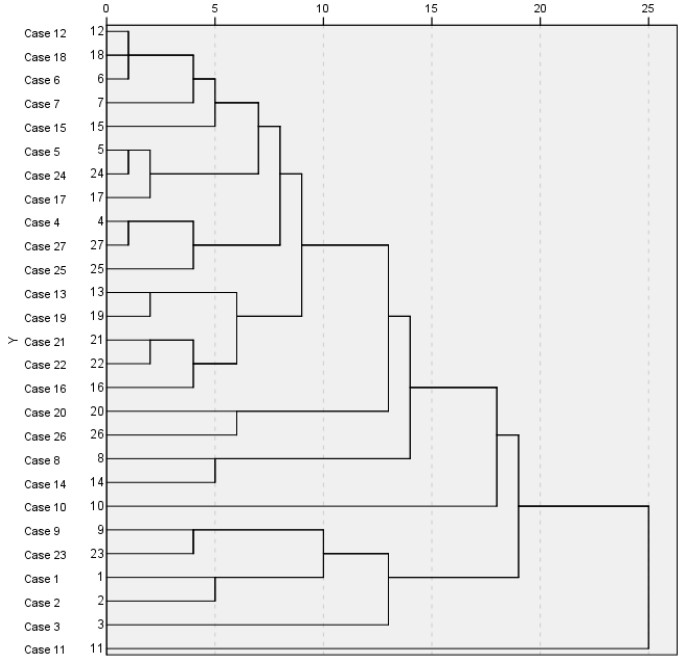

图 4 红葡萄酒聚类分析结果

由图 4 我们可以发现，红葡萄酒可分为四类，即：

第一类：样品 1,2,3,9,23

第二类：样品 4,5,6,7,8,12,13,14,15,16,17,18,19,20,21,22,24,25,

26,27

第三类：样品 10

第四类：样品 11

分类得分情况如表 15 所示：

表 15 红葡萄酒分类结果

	第一类	第二类	第三类	第四类
均分	78.04	71.98	74.2	70.1

通过表 15，我们可以得出结论：第一类酒样品的等级最高，平均得分为 78.04 分；第三类酒样品的等级较高，均分为 74.2 分；第二类酒样品的等级较低，均分为 71.98.分；第四类酒样品的等级最低，均分为 70.1 分。

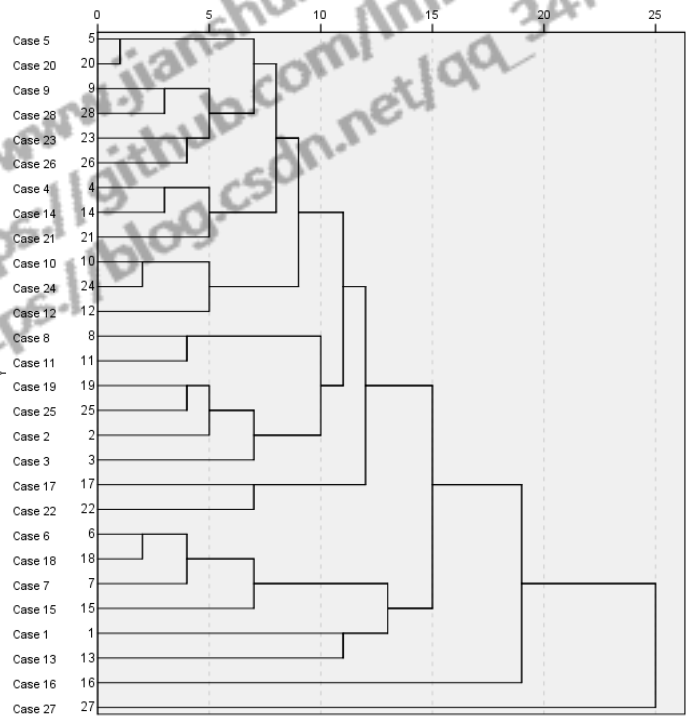

图 5 白葡萄酒聚类分析结果

同理，由图 5 我们将白葡萄酒分为 4 类，即：

第一类：样品 1,6,7,13,15,18

第二类：样品 2,3,4,5,8,9,10,11,12,14,17,19,20,21,22,23,24,25,26，28

第三类：样品 16

第四类：样品 27

得到白葡萄酒的得分分类情况如下：

表 16 白葡萄酒分类情况

	第一类	第二类	第三类	第四类
均分	73.7	75.66	74.3	65.2

通过表 16，我们可以得出结论：第二类酒样品的等级最高，平均得分为 75.66 分；第三类酒样品的等级较高，均分为 74.3 分；第一类酒样品的等级较低，均分为 73.7 分；第四类酒样品的等级最低，均分为 65.2 分。

5.4 问题三的模型建立与求解

作为酿酒原料，葡萄与葡萄酒之间的关系密切，其理化指标会在一定程度上反映葡萄酒的质量。首先，我们根据第二问的主成分分析，提取酿酒葡萄与葡萄酒的主要因素。其次，通过引入偏最小二乘回归分析，找出酿酒葡萄与葡萄酒理化指标之间的相关系数。最后，基于粒子群算法(PSO 算法)对以上求得的系数进行优化，分析酿酒葡萄、葡萄酒理化指标对葡萄酒质量的影响。

5.4.1 模型的建立

要分析酿酒葡萄与葡萄酒之间理化指标之间的联系，要先对红、白葡萄酒的指标进行简化。根据问题二中的主成分分析法的结果，红葡萄酒提取出 7 个理化指标，白葡萄酒提取出 9 个理化指标，红、白葡萄分别提取出分别为 2 个、3 个理化指标。在此基础上，我们采用偏最小二乘回归分析法，对葡萄与葡萄酒之间的相关系数进行求解。

1、偏最小二乘回归分析

偏最小二乘回归是一种改进的多元线性回归分析的方法。它可以解决多重共线性造成的问题，如进行判定时没有十分可靠的检验方法等。当两个组变量的个数很多，且都存在多重相关性，而观测数据（样本量）又较少时，用偏最小二乘回归建立的模型具有传统经典回归分析等方法所没有的优点。

由于在实际应用中，判断确定过程中存在一定的不确定性和模糊性，因此在计算出葡萄与葡萄酒之间的相关系数之后，我们在基于粒子群的算法(PSO 算法)对计算的系数进一步优化，克服品酒员决策中的不确定性以及测量数据的不确定性。其具体步骤如下：

(1) 分别提取两个变量的第一对成分，并使之关联度达到最大。由两组变量标准化的数据矩阵 A 和 B，可以计算第一对成分的得分向量，记为$\hat{u}_1$和$\hat{v}_1$:

$$\hat{u}_1 = A\rho^{(1)} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} \alpha_{11} \\ \vdots \\ \alpha_{1m} \end{bmatrix}$$

$$\hat{v}_1 = B\gamma^{(1)} = \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nm} \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \vdots \\ \beta_{1m} \end{bmatrix}$$

(2) 建立$y_1, \dots, y_p$对u_1的回归及$x_1, \dots, x_m$对u_1的回归；

(3) 用残差阵A_1和B_1代替 A 和 B，重复以上步骤；

(4) 设n × m数据阵 A 的秩为r ≤ min(n − 1, m)，则存在 r 个成分$u_1, \dots, u_r$使得：

$$\begin{cases} A = \hat{u}_1\sigma^{(1)T} + \cdots + \hat{u}_r\sigma^{(1)T} + A_r \\ B = \hat{u}_1\tau^{(1)T} + \cdots + \hat{u}_r\tau^{(1)T} + B_r \end{cases}$$

即得到 p 个因变量的片最小二乘回归方程式为：

$$y_j = c_{j1}x_1 + \cdots + c_{jm}x_m, j = 1, \dots, p$$

2、PSO 优化算法

在此处采用 PSO 优化算法来实现对回归模型的参数估计，其中心思想是将回归模型中的一组参数看做一个粒子$P_i = (\alpha_{i,0}, \beta_{i,1}, \beta_{i,2}, \dots, \beta_{i,n}), i = 1,2, \dots, k$，通

过跟踪当前最优粒子搜索最优解，最优解即为回归模型中最优的参数估计。PSO优化算法中根据适应度函数的计算结果对每个粒子进行评价，J(θ)为适应度函数，最优解为适应度函数最小的粒子。

为了避免粒子过分聚集，采用斥力因子的位置更新方法使粒子均匀分散于搜索范围。该方法的思想是当粒子的间距小于最小允许间距时，存在一个斥力将各粒子推至大于或等于最小允许间距。带斥力因子的位置更新方程如下所示：

$$x_i^{T+1}=\begin{cases} x_i^T+v_i^{T+1}+2\Delta s^T & \sum_{i\neq j,j=1}^{N}\left\|x_i^T-x_j^T\right\|<\Delta s^T \\ x_i^T+v_i^{T+1} & \sum_{i\neq j,j=1}^{N}\left\|x_i^T-x_j^T\right\|>\Delta s^T \end{cases}$$

其中，x_i^T为第 T 次迭代后第 i 个粒子的位置矢量；v_i^{T+1}为第 T+1 次迭代后第 i 个粒子的速度矢量；Δs^T为第 T 次迭代后粒子的最小允许间距且随迭代次数的增加而不断减少。

速度更新方程如下所示：

$$v_i^{T+1}=\omega v_i^T+c_1 rand_1(p_{best\ i}^T-x_i^T)+c_2 rand_2(g_{best}^T-x_i^T)$$

其中，c_1,c_2为学习因子；$p_{best\ i}^T$为第 T 次迭代后第 i 个粒子的个体极值点；g_{best}^T为第 T 次迭代后的全局极值点。

回归模型中参数估计的粒子群优化算法步骤如下：

(1) 初始化各类参数。搜索空间的下限L_d，和上限L_u，学习因子c_1,c_2，算法的最大迭代次数T_{max}，粒子速度范围；随机初始化搜索点的位置及其速度；

(2) 评价粒子。计算每个粒子的适应度函数，设每一个初始化粒子P_i为粒子的个体极值点$p_{best\ i}^T$，使适应度函数最小的粒子为全局极值点g_{best}^T；

(3) 判断迭代是否终止。若终止，转向(5);否则，转向(4)；

(4) 更新惯性权重和粒子状态，转向(2)；

(5) 输出g_{best}^T为最优的回归模型中参数估计。

在(4)中的惯性权重ω的选择必须平衡算法的全局搜索与局部搜索，防止粒子在运动过程中陷入局部最优。常用的非线性递减惯性权重策略有 3 种，分别表示如下:

(1)$\omega(\mathrm{T})=\omega_0-(\omega_0-\omega_{end})(\frac{T}{T_{max}})^2$；

(2)$\omega(\mathrm{T})=\omega_0-(\omega_0-\omega_{end})[\frac{2T}{T_{max}}-\left(\frac{T}{T_{max}}\right)]^2$

(3) $\omega(\mathrm{T})=\omega_{end}(\frac{\omega_0}{\omega_{end}})^{1/(1+cT/T_{max})}$

在这里我们采用(1)来进行权重的更新，且最小允许间距Δs^T的变化规律与(1)类似，如下所示：

$$\Delta s^T=\Delta s_0-(\Delta s_0-\Delta s_{end})(\frac{T}{T_{max}})^2$$

5.4.2 模型的求解

1、偏最小二乘回归分析求得结果

根据问题二中的主成分分析法的结果，选取红、白葡萄的指标如表 17、表 18 所示：

表 17 选取红、白葡萄指标

红葡萄	还原糖	单宁	葡萄总黄酮	柠檬酸	果穗质量	颜色	酒石酸		
白葡萄	花色苷	柠檬酸	褐变度	总酚	白藜芦醇	黄酮醇	PH 值	干物质含量	颜色

表 18 选取红、白葡萄酒的理化指标

红葡萄酒	总酚	色泽	
白葡萄酒	总酚	白藜芦醇	色泽

接着，我们利用偏最小二乘法计算红、白葡萄与红、白葡萄酒理化指标的相关系数如表 19 和表 20 所示：

表 19 红葡萄与红葡萄酒相关系数

	总酚	色泽
还原糖	0.4464	-0.2255
单宁	-0.0332	0.5015
葡萄总黄酮	0.2265	0.1714
柠檬酸	0.0828	-0.024
果穗质量	-0.1062	-0.021
颜色	-0.16534	-0.0383
酒石酸	0.1352	0.3916
常数项	0.0019	0.0025

表 20 白葡萄与白葡萄酒相关系数

	总酚	白藜芦醇	色泽
花色苷	0.112	-0.031	0.506
柠檬酸	0.291	0.098	0.198
褐变度	-0.007	0.005	0.208
总酚	-0.004	0.002	0.07
白藜芦醇	0.11	-0.039	-0.229
黄酮醇	0.047	-0.013	0.157
PH 值	0.198	0.063	0.15
干物质含量	0.06	-0.017	0.246
颜色	-0.164	0.055	0.082
常数项	0.00	0.00	0.014

通过表 19、20 我们可以得出：

在红葡萄酒中，总酚与还原糖、葡萄总黄酮、果穗质量、颜色、酒石酸；色泽与还原糖、单宁、葡萄总黄酮、酒石酸的相关系数的绝对值较大，因此相关性较大。同理在白葡萄酒中，总酚与柠檬酸、PH 值；白藜芦醇与柠檬酸、PH 值；色泽与花色苷、干物质含量相关性较大。

2、PSO 优化算法求得结果

通过 PSO 优化算法解得的红、白葡萄与葡萄酒之间的相关系数如表 21、表

22 所示：

表 21　PSO 算法红葡萄相关系数

	总酚	色泽
还原糖	0.0453	-0.0286
单宁	0.4917	-0.2542
葡萄总黄酮	0.0121	0.4729
柠檬酸	0.2718	0.1428
果穗质量	0.1281	-0.0526
颜色	-0.16534	-0.0496
酒石酸	-0.1201	-0.0669

表 22 PSO 算法白葡萄相关系数

	总酚	白藜芦醇	色泽
花色苷	-0.0009	0.0691	-0.1015
柠檬酸	0.1109	0.0382	0.4048
褐变度	0.2897	-0.0285	-0.2993
总酚	-0.0076	0.0736	0.1063
白藜芦醇	-0.0044	0.0710	-0.0315
黄酮醇	0.1086	0.0305	-0.3306
PH 值	0.0457	0.0556	0.0555
干物质含量	0.1966	0.0060	0.0484
颜色	0.0589	0.0523	0.1445

通过表 21 和表 22，我们得到经过 PSO 优化算法的红、白葡萄与葡萄酒之间的参数估计值分别为：

$$\{\alpha_0,\beta_1,\beta_2,\dots,\beta_7\}=(0.0453,0.4917,0.0121,$$
$$0.2718,0.1281,-0.0608,-0.1201,0.1826)$$
$$\{\alpha_1,\beta_1,\beta_2,\dots,\beta_7\}=(-0.0286,-0.2542,0.4729,$$
$$0.1428,-0.0526,-0.0496,-0.0669,0.2999)$$
$$\{\alpha_0,\beta_1,\beta_2,\dots,\beta_9\}=(12.7951,-0.0009,0.1109,0.2897,$$
$$-0.0076,-0.0044,0.1086,0.0457,0.1966,0.0589)$$
$$\{\alpha_1,\beta_1,\beta_2,\dots,\beta_9\}=(19.3377,0.0691,0.0382,-0.0285,$$
$$0.0736,0.0710,0.0305,0.0556,0.0060,0.0523)$$
$$\{\alpha_2,\beta_1,\beta_2,\dots,\beta_9\}=(26.3799,-0.1015,0.4048,-0.2993,$$
$$0.1063,-0.0315,-0.3306,0.0555,0.0484,0.1445)$$

由此，我们发现，在红葡萄酒中，总酚与单宁、柠檬酸、果穗质量、颜色、酒石酸；色泽与葡萄总黄酮、柠檬酸的相关系数的绝对值较大，因此相关性较大。同理在白葡萄酒中，总酚与柠檬酸、褐变度、黄酮醇、干物质含量；白藜芦醇与花色苷、总酚、白藜芦醇；色泽与柠檬酸、黄酮醇相关性较大。

5.4.3 模型的检验

为了验证所用 PSO 优化算法的准确性，我们分别计算利用偏最小二乘法和 PSO 优化算法所得结果与题目中所给数据的偏差值p_i，其计算公式为：

$$p_i = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_0)^2}{n}}$$

其中，y_0为题目中原始数据。

其结果如表 23 所示：

表 23 两种算法偏差值的对比

	偏最小二乘回归分析	PSO 优化算法
红葡萄酒偏差值	0.134043	0.132344
	0.246756	0.244823
白葡萄酒偏差值	0.127951	0.127812
	0.193377	0.176302
	0.263799	0.253425

由于偏差值越小代表计算结果越接近理论值，即计算结果越精确。通过表 23 我们发现，基于 PSO 优化算法计算得到的偏差值与基于偏最小二乘回归分析计算得到的偏差值相比都较小，因此，PSO 算法较偏最小二乘回归分析更精确。

5.5 问题四的模型建立与求解

在众多数据的基础上分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，首先我们需要分析出与评分标准可能相关的因素。其次，通过灰色关联度法分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响程度。最后，将计算所得结果与品酒员评分标准相对比，验证模型的可靠性，即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。

5.5.1 模型的建立

由于葡萄酒在制作的过程当中，受到很多复杂因素的制约。其发酵过程是由基质成分 、气候条件、环境等诸多因素影响和控制的微生物区系衍变过程 ,其中各类参量和因素彼此依赖又相互制约, 因而过程非常复杂,制作出来的葡萄酒质量也不尽相同[20]。为了便于我们分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，我们首先将影响葡萄酒质量的因素分为三类，即外观、香气、口感。在葡萄与葡萄酒的理化指标当中，通过查阅资料我们发现：花色苷、褐变度等都会对外观造成影响；而影响葡萄酒香气的是芳香物质；单宁、总酚、酒石酸等会影响葡萄酒的口感。根据以上条件，我们利用灰色关联分析法分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响。

灰色关联度分析法是一种多因素统计分析方法，它是以各因素的样本数据为依据用灰色关联度来描述因素间关系的强弱、大小和次序，若样本数据反映出的两因素变化的态势(方向、大小和速度等)基本一致，则它们之间的关联度较大；反之，关联度较小。灰色关联度分析的具体计算步骤如下：

(1) 确定比较对象(评价对象)和参考数列(评价标准)

设评价对象有 m 个，评价指标有 n 个，参考数列为$x_0 = \{x_0(k)|k = 1,2,\dots,n\}$，比较数列为$x_i = \{x_i(k)|k = 1,2,\dots,n\}, \mathrm{i} = 1,2,\dots,\mathrm{m}$

(2) 确定各指标值对应的权重

可用层次分析法等确定各指标对应的权重$\mathrm{w} = [w_1, w_2, \dots, w_n]$。其中，

$w_k(k=1,2,\dots,n)$为第 k 个评价指标的对应权重。

(3) 计算灰色关联度系数

比较数列x_i对参考数列x_0在第 k 个指标上的关联系数公式如下：

$$\xi_i(k)=\frac{\min\limits_s\min\limits_t|x_0(t)-x_s(t)|+\rho\max\limits_s\min\limits_t|x_0(t)-x_s(t)|}{|x_0(k)-x_i(k)|+\rho\max\limits_s\min\limits_t|x_0(t)-x_s(t)|}$$

其中，$\rho\in[0,1]$为分辨系数。称$\min\limits_s\min\limits_t|x_0(t)-x_s(t)|,\max\limits_s\min\limits_t|x_0(t)-x_s(t)|$分别为两级最小差和两级最大差。

一般来讲，分辨系数ρ越大，分辨率越大；ρ越小，分辨率越小。

(4) 计算灰色加权关联度

其公式为：

$$r_i=\sum_{k=1}^{n}w_i\xi_i(k)$$

其中，r_i为第 i 个评价对象对理想对象的灰色加权关联度。

(5) 评价分析

根据灰色加权关联度的大小，对各评价对象进行排序，可建立评价对象关联序，关联度越大，其评价结果越好。

5.5.2 模型的求解

根据附件一的评分标准，我们可得知，外观方面占总分的 15%，香气方面占总分的 30%，口感方面占总分的 44%，因此得到各个类别的权重$w=(0.15,0.30,0.44)$。

根据灰色关联度分析法，我们计算得到灰色关联度系数如表 24、25 所示：

表 24 红葡萄酒的灰色关联度系数

序号	关联度系数	序号	关联度系数	序号	关联度系数
1	0.5122	10	0.4742	19	0.5003
2	0.5366	11	0.5754	20	0.4588
3	0.6314	12	0.5295	21	0.6278
4	0.4803	13	0.4819	22	0.5184
5	0.6146	14	0.6593	23	0.5610
6	0.4783	15	0.5410	24	0.5462
7	0.5582	16	0.5117	25	0.4854
8	0.5922	17	0.5344	26	0.5488
9	0.6054	18	0.5019	27	0.4881

根据红葡萄酒的关联度系数，我们对其进行排序，结果如表 25 所示：

表 25 红葡萄酒灰色关联度排序结果

序号	酒样品	序号	酒样品	序号	酒样品
1	14	10	26	19	18
2	3	11	24	20	19
3	21	12	15	21	27
4	5	13	2	22	25

5	9	14	17	23	13
6	8	15	12	24	4
7	20	16	22	25	6
8	23	17	1	26	10
9	7	18	16	27	11

同理可得白葡萄酒的灰色关联度系数和排序结果如表 26、表 27 所示：

表 26 白葡萄酒的灰色关联度系数

序号	关联度系数	序号	关联度系数	序号	关联度系数	序号	关联度系数
1	0.4386	8	0.4763	15	0.5116	22	0.5438
2	0.4856	9	0.5080	16	0.4875	23	0.5364
3	0.5562	10	0.4584	17	0.4851	24	0.5279
4	0.5006	11	0.4776	18	0.5015	25	0.5097
5	0.5645	12	0.4743	19	0.4822	26	0.5978
6	0.4919	13	0.4634	20	0.4814	27	0.6041
7	0.4932	14	0.5482	21	0.5210	28	0.5493

表 27 白葡萄酒灰色关联度排序结果

序号	酒样品	序号	酒样品	序号	酒样品	序号	酒样品
1	27	8	23	15	4	22	20
2	26	9	24	16	7	23	11
3	5	10	21	17	6	24	8
4	3	11	15	18	16	25	12
5	28	12	25	19	2	26	13
6	14	13	9	20	17	27	10
7	22	14	18	21	19	28	1

由表 24 可知，黄酮醇、总糖、白藜芦醇、酒石酸、百粒质量、褐变度和 1-庚醇会对红葡萄酒的质量存在显著的联系。由表 25 可知，总糖、总酚、花色苷、PH 值、干物质含量和辛酸 3-甲基丁酯会对白葡萄酒的质量存在显著联系。

接着，我们将灰色关联度计算所得排序与品酒员所评红、白酒样品排序相对比，分别如图 6 和图 7 所示：

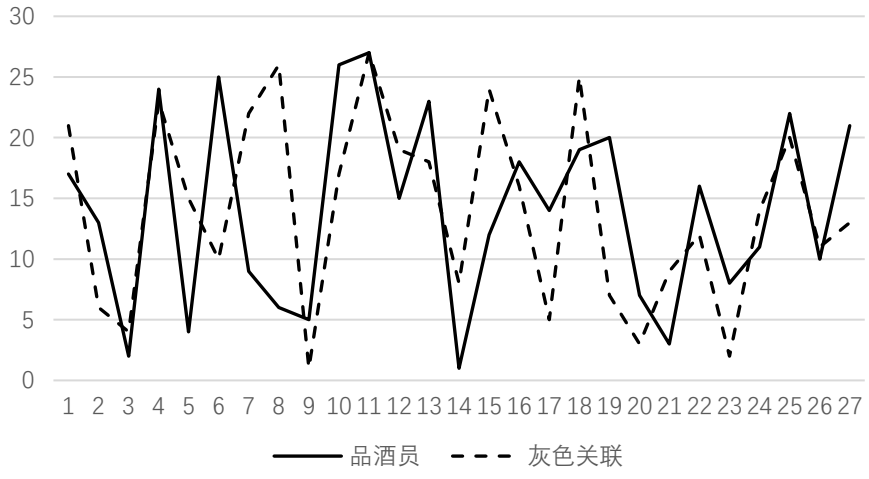

图 6 红葡萄酒灰色关联结果验证

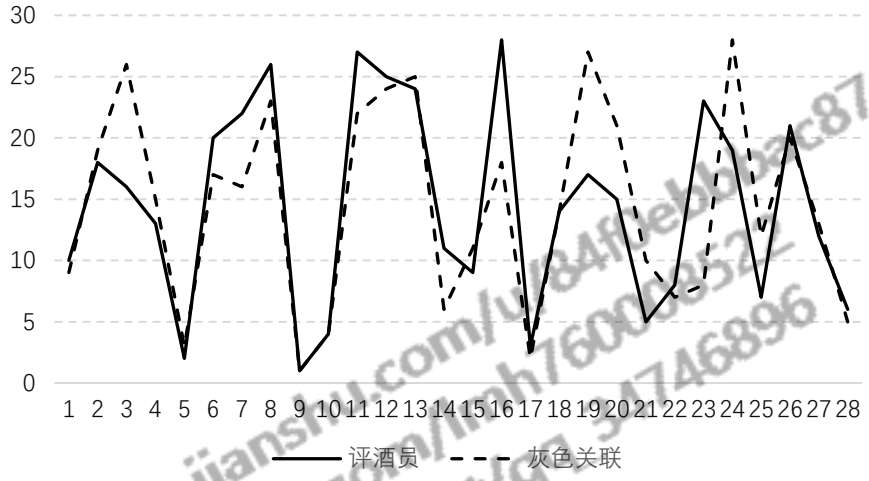

图 7 白葡萄酒灰色关联结果验证

通过图 6、图 7 我们可以发现，通过灰色关联得出的葡萄酒排序与品酒员打分得出的葡萄酒排序的结果虽然有一定的误差，但是大致排序的区间是一样的。故我们认为可以用葡萄和葡萄酒的理化指标来评价葡萄酒的质量。

六、模型的评价与推广

6.1 模型的优点与缺点

针对问题三，偏最小二乘回归分析在建模的过程中集中了主成分分析和线性回归分析法的特点，除了提供了一个更为合理的模型外，还可以同时完成一些类似于主成分分析的研究内容，提供了更加丰富深入的信息。PSO 优化算法克服了品酒员决策中的不确定性以及测量数据的不确定性。针对问题四灰色关联度分析可以针对大量不确定因素及其相互关系，将定量和定性方法有机结合起来，使原本复杂的评价问题变得更加清晰简单，而且计算方便，并可以在一定程度上排除品酒员的主观任意性，得出的结论也比较客观，有一定的参考价值。

针对问题二，系统聚类分析在样本量较大时计算并不准确，如果根据距离或相似系数得出聚类分析的结果，显然是不恰当的。针对问题三，尽管芳香物质具有易挥发性，但是忽略芳香物质的求解过程存在一定误差。

6.2 模型的推广

粒子群优化算法在讨论生物系统——社会系统时，可以模拟系统利用局部信息预测可能产生的群体行为，可以推广到检验医疗方案的合理性、环境质量的检测等。灰色关联度不仅可以分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，在分析气象与农业之间、设备与工业之间关联度等方面也有很大的应用价值。整个模型也可以应用于医药医学方面，如检测药物含量与药物效果的关系等。

参考文献

[1]曹俊忠.一元线性回归显著性检验方法分析[J].西安工程科技大学学报,2015(6):

07-23
[2]范琳琳.王红瑞.宋乃琦.俞淞.王欣莉.基于 t 检验的水文时间序列 HHT 分析方法及应用[J].系统工程理论与实践,2015(5):12-15
[3]穆广杰.T 检验失效的原因及处理[J].统计与决策,2011(21):2-8
[4]赵安龙,李洪双. 基于概率支持向量机的可靠性分析与设计方法[J/OL]. 应用力学学报,2017,34(01):50-56+195. (2017-01-20)[2017-08-10]
[5] 赵媛. 基于类电磁机制算法的 SVM 决策树多分类策略研究[D].西安电子科技大学,2014.
[6] 李燕琴. 医院信息化建设的数据标准化研究[J]. 科技广场,2016,(09):43-45. [2017-08-10].
[7]王星华.陈卓优.彭显刚.一种基于双层聚类分析的负荷形态组合识别方法[J].电网技术,2016(5):16-18
[8]赵中军.刘善亮.游大鸣.王学忠.胡邦辉.程一帆.偏最小二乘回归模型在辽宁汛期降水预测中的应用[J].干旱气象.2015(6):22-24
[9]蒋诗泉.刘思峰.刘中侠.方志耕.基于面积的灰色关联决策模型[J].控制与决策,2015(3):2-6
[10]姚万业,杨金彭. 基于显著性差异分析的风机变桨电机故障预警[J]. 电力科学与工程,2016,32(02):50-54+65. [2017-08-10].
[11]E.Bartezzaghi,R.Verganti.Managing demand uncertainty through order overplanning,International Journal of Production Economics,2015(40):107-120
[12]DeJong R M,Amsler C,Schmidt P.A Robust Version of the KPSS Test Based on Indicators.Journal of Econometrics,2014,127:311-333
[13]孙翔.何文林.邱炜.基于显著性差异的油浸倒置式电流互感器氢气阈值分析[J].电力科学与工程,2015(6):20-25
[14]GB/T 4883-2008,数据的统计处理和解释正态样本离群值的判断和处理[S]
[15]赵亚娟.专家群评价结果可信度分析与检验[J].中国科学技术大学学报，2016(2)
[16]高波.王丽芳.基于可信度分析的高校科技创新能力评价[J].武汉科技大学学报，2015
[17]张宏哲. SPSS 非线性回归在沪铜期货价格预测中的应用[J].价值工程,2014(19)
[18]王其中.葡萄酒市场特点及其竞争力分析[J].技术与市场,2016,(4):35-37
[19]于静.李景明.吴继红.葛毅强.葡萄酒芳香物质研究进展[J].西北农业大学学报,2015(3):4-5
[20]YUHQ.LUJ.XIAOCB.Preparation and properties of novel hydrogels from oxidized konjac glucomannan cross-linked chitosan for in vitro drug delivery[J].Macromolecular Bioscience,2014,7(9/10):1100-1111

附 录

附录一：第三问代码

```
clear
format long g
ab0=load('C:\Users\ÃÀ°¬\Desktop\shumo\12\3红.txt');
mu=mean(ab0);sig=std(ab0);
ab=zscore(ab0);
a=ab(:,[1:7]);b=ab(:,[8:end]);
length=size(a,1);
ncomp=2;
[XL,YL,XS,YS,BETA,PCTVAR,MSE,stats]=plsregress(a,b,ncomp);
n=size(a,2); m=size(b,2);
BETA2(1,:)=mu(n+1:end)-
mu(1:n)./sig(1:n)*BETA([2:end],:).*sig(n+1:end);
BETA2([2:n+1],:)=(1./sig(1:n))'*sig(n+1:end).*BETA([2:end],:);
BETA2=BETA2'
format
```

附录二：粒子群优化代码

```
c1=1.4962;
c2=1.4962;
w=0.7298;
MaxDT=100;
D=8;
N=2;
num=20;
eps=10^(-7);          )
for i=1:N
    for j=1:D
        x(i,j)=BETA2(i,j);     v(i,j)=0.1*randn;
    end
end
for i=1:N
    p(i)=sphere0(x(i,:),D,N,a,b,length);
    y(i,:)=x(i,:);
    Pg(i,:)=x(1,:);
end
for i=1:N
    for t=1:MaxDT
        for j=1:D
            v(j)=0.1*(-1+(1-(-1))*rand(1,1));
        end
        if
```

```
sphere0(x(i,:),D,N,a,b,length)>sphere0(x(i,:)+v(j),D,N,a,b,length)
            x(i,:)=x(i,:)+v(j);
            p(i)=sphere0(x(i,:)+v(j),D,N,a,b,length);
        end
    end
end
disp('**********************************************************')
x
for i=1:N
sphere0(BETA2(i,:),D,N,a,b,length)
end
p
disp('**********************************************************')
```

附录三：sphere0函数

```
function result=sphere0(x,D,N,a,b,length)
y=0;
for i=1:length
    y0=0;
    for j=1:D
        if j==1
            y0=y0+x(j);
        else
            y0=y0+x(j)*a(i,j-1);
        end
    end
    y=y+abs(b(i)-y0);
end
result=y;
end
```

附录四：第四题代码

```
clc, clear
a=load('C:\Users\ÃÀ°¬\Desktop\shumo\12\4°×.txt');
a=a';
[m,n]=size(a);
cankao=max(a')'
t=repmat(cankao,[1,n])-a;
mmin=min(min(t));
mmax=max(max(t));
rho=0.5;
xishu=(mmin+rho*mmax)./(t+rho*mmax)
guanliandu=mean(xishu)
[gsort,ind]=sort(guanliandu,'descend')
```