《品种葡萄酒识别技术导则》国家标准编制说明
(征求意见稿)

一、工作简况

1、 任务来源

根据国家标准化管理委员会下达的2010国家标准制定计划，《品种葡萄酒识别技术导则》（计划编号：20101053-T-607）列入制定计划，本标准由中国轻工业联合会提出，由全国酿酒标准化技术委员会归口，中国食品发酵工业研究院等单位起草。

2、 目的意义

葡萄酒品种是葡萄酒产品质量特征的重要基础，决定了葡萄酒风格特征，单品种葡萄酒由于独特的风味品质逐步受到消费市场青睐。目前我国《葡萄酒》国家标准中对品种葡萄酒进行了明确定义，但缺乏配套的检测方法和判别依据等关键技术标准，仅仅依靠企业自律和现有的标准管理体系无法有效规范市场。亟需建立中国品种葡萄酒识别技术导则，对规范我国葡萄酒市场竞争秩序，促进葡萄酒向区域化、特色化、高档化良性发展具有重要意义。

本标准根据国内外文献资料和国际相关组织的技术文件中葡萄酒产地识别的指导原则，建立可用于中国品种葡萄酒识别的一般原则和程序，有助于进一步完善我国葡萄酒技术标准体系，充分发挥关键技术标准在行业规范化发展中的支撑作用。

3、 简要编制过程

2011年1月-2011年7月，计划下达后，为了使本标准更具有先进性、科学性，起草工作组开展多项基础研究工作，查阅了大量国内外文献资料及相关技术法规，通过对多种品种葡萄酒识别方法的特点进行对比分析，总结出各种品种葡萄酒识别方法通用的优缺点和一般性程序，初步建立可用于品种葡萄酒真实性识别的一般原则和程序。

2011年-2016年，起草组开展收集了不同企业的不同品种葡萄酒，同时开展了品种葡萄酒的不同鉴别技术研究：(1)采用固相微萃取测定不同品种葡萄酒中几十种挥发性成分，采用化学计量学对测定的数据进行分析,开展葡萄酒挥发性组分在品种葡萄酒鉴别方面的可行性研究；(2)开展分子生物学技术在葡萄酒分子鉴定方面的研究，进行葡萄酒中葡萄DNA提取、

DNA 质量评价和品种鉴定方面,开展基因检测技术在葡萄酒真伪和品质鉴定方面的应用研究；(3)采用液相色谱法测定不同品种的葡萄酒中多种花青素和有机酸，采用化学计量学对测定的数据进行分析，开展花青素及有机酸在品种葡萄酒鉴别方面的可行性研究。

2016-2018 年，起草组开展非目标一维核磁共振氢谱(^{1}H NMR)技术在品种葡萄酒中的研究应用，该技术作为一种快速、方便简单、无损检测方法,在国外已成功用于葡萄酒、果汁、蜂蜜等产品的真实性鉴别。起草组开展采用 ^{1}H NMR 测定不同品种葡萄酒，结合多元数据处理技术构建非目标 ^{1}H NMR 葡萄酒指纹图谱真实性鉴别技术模型，为解决中国葡萄酒品种鉴别提供技术支撑，该研究技术“基于非目标核磁共振氢谱指纹图谱技术的产地和品种葡萄酒真性鉴别”已经通过中国轻工联合会成果鉴定，国际领先水平。

2019 年，起草工作组选择采用非目标一维核磁共振氢谱、多种花青素和有机酸多种不同的技术参数结合作为品种葡萄酒识别的特征参数，综合考虑国内品种葡萄酒各产区生产情况，通过采集样品、构建模型、预测识别以验证该标准的实用性。

二 、标准编制原则和主要内容

1、标准编制原则

以科学技术和实验数据为依据，结合行业实际生产情况，经过科学研究而制定。本标准的制定充分考虑规范葡萄酒行业发展，促进葡萄酒行业技术进步，保证产品质量真实性；充分考虑国内相关的法规要求，结合国情和产品特点；与相关标准法规协调一致；确保标准的科学性、先进性、可操作性。

1）指导品种葡萄酒识别的一般性原则和操作程序，简明易懂，操作性强。

2）适用范围与当前检测技术水平相适应。

3）结合我国葡萄酒品种种植情况，与葡萄酒的生产流通相适应，适用于葡萄酒行业的推广应用。

2、主要内容

本标准规定了品种葡萄酒识别技术方案和识别程序。

本标准适用于对品种葡萄酒的识别。

三、主要试验（或验证）情况分析

（1）对以葡萄酒一维核磁共振氢谱（^{1}H NMR）作为品种葡萄酒识别技术进行了系统研究。

① 葡萄酒一维核磁共振氢谱（^{1}H NMR）测定方法的开发。起草工作组结合国外 ^{1}H NMR 技术成功用于检测分析葡萄酒、果汁、蜂蜜等产品的先进经验，采用合适的脉冲序列用于乙醇（δ= 1.2 ppm 和 3.6 ppm）和水（δ= 4.8 ppm）信号抑制，经过对脉冲参数的优化，有效的减弱水峰和乙醇峰对葡萄酒样品检测的干扰。通过稳定性、准确性等方法学研究，建立了葡萄酒 ^{1}H NMR 测定分析方法，结果符合要求。

② 建立了中国葡萄酒品种真实性的鉴别模型。通过非目标 ^{1}H NMR 指纹图谱技术，对红葡萄酒品种（赤霞珠、玫瑰蜜、蛇龙珠）和白葡萄酒品种（白玉霓，龙眼，霞多丽）样品进行了测定分析。对葡萄酒样品 ^{1}H NMR 图谱峰化学位移偏移进行了分段对齐，采用对 ^{1}H NMR 图谱进行分段积分及数据降维。分段积分值（bins）作为之后多元统计分析的输入变量。采用主成分分析结合线性判别分析建立的留一交叉验证葡萄酒品种鉴别模型，实现了红葡萄酒品种赤霞珠、玫瑰蜜和蛇龙珠的正确分类率分别为 68%，100% 和 81%，红葡萄酒的平均正确分类率为 82%；白葡萄酒品种白玉霓、龙眼和霞多丽的正确分类率为 93%，95% 和 94%，白葡萄酒的平均正确分类率为 94%。以主成分分析结合随机森林建立的葡萄酒品种鉴别模型，红葡萄酒品种赤霞珠、玫瑰蜜和蛇龙珠的正确分类率分别为 75%，100% 和 82%，红葡萄酒的平均正确分类率平均为 85%。白葡萄酒品种白玉霓、龙眼和霞多丽的正确分类率为 93%，81% 和 89%，白葡萄酒的平均正确分类率平均为 89%。为避免葡萄酒品种鉴别模型出现过拟合现象，采用外部重复双随机交叉验证方法对鉴别模型的有效性进行了验证。结果表明，葡萄酒非目标 ^{1}H NMR 指纹图谱结合多变量统计分析技术，可以有效鉴别品种葡萄酒的真实性。

（2）对以葡萄酒中花青素及有机酸含量作为品种葡萄酒识别的参考指标进行了系统研究。

① 葡萄酒中有机酸及花青素测定方法的开发。起草工作组建立了 HPLC 法测定葡萄酒中莽草酸、草酸和酒石酸三种有机酸和 Po-3-cug、Mv-3-cugl、Po-3-cugl、Mv-3-cug、峰 30、峰 61、峰 RT25.35 七种花色苷含量的分析方法。

② 建立了基于有机酸及花青素的中国葡萄酒品种真实性的鉴别模型。起草工作组通过HPLC法对赤霞珠、黑比诺、蛇龙珠、玫瑰蜜和梅鹿辄5个品种葡萄酒样品的莽草酸、草酸、酒石酸和七种花色苷进行了测定分析。采用方差分析、主成分分析和判别分析对数据进行了统计分析，建立了基于有机酸及花青素的中国葡萄酒品种真实性的鉴别模型，实现了对赤霞珠、黑比诺、蛇龙珠、玫瑰蜜和梅鹿辄5个品种葡萄酒样品的有效区分。线性判别分析模型按照2:1比例进行建模和预测，对赤霞珠、黑比诺、蛇龙珠、玫瑰蜜和梅鹿辄5个品种葡萄酒的预测能力可达100%。结果表明，葡萄酒中有机酸及花青素数据结合多元统计学技术，可现实对中国葡萄酒品种的真实性鉴别，该技术检测成本较低，方法成熟，模型简单，正确识别率高，具有很好推广应用价值。

四、 标准中涉及的专利

无。

五、产业化情况、推广应用论证和预期达到的经济效果等情况

该标准的实施，将填补我国品种葡萄酒识别技术方法标准的空白，对葡萄酒品种溯源具有十分重要的意义，为保护我国高端特色品种葡萄酒提供有效技术支撑。

六、 采用国际标准和国外先进标准情况，与国际、国外同类标准水平的对比情况，国内外关键指标对比分析或与测试的国外样品、样机的相关数据对比情况。

无。

七、 与现行相关法律、法规、规章及相关标准，特别是强制性标准的协调性

该标准从我国葡萄酒行业的实际情况出发，参考了国内外相关资料，体现了科学性、先进性和可操作性原则，在制定过程中充分考虑国内相关的法规要求，结合葡萄酒行业的特点；与相关标准法规包括强制性标准协调一致。

八、 重大分歧意见的处理经过和依据

无重大分歧意见。

九、 标准性质的建议说明

《品种葡萄酒识别技术导则》为推荐性国家标准。

十、 贯彻标准的要求和措施建议

在本标准通过审核、批准发布之后，由相关部门组织力量对本标准进行宣贯，在行业内进行推广。建议本标准自发布6个月之后开始实施。

十一、 废止现行相关标准的建议

无。

十二、 其它应予说明的事项

该标准从我国葡萄酒行业的实际情况出发，参考了国内外相关资料，体现了科学性、先进性和可操作性原则，综合评定达到了国际水平。

全国酿酒标准化技术委员会

2019年12月

附件一 品种葡萄酒识别技术的方法学验证

酿造高品质葡萄酒取决于很多要素，其中重要的因素之一就是合适的葡萄品种的选择。据报道，在中国葡萄酒行业存在许多利益驱动的虚假葡萄酒欺诈事件，例如通过旧瓶装新酒、仿造酒标等手段冒充具有高价值产区的葡萄酒，破坏葡萄酒市场的声誉。葡萄酒真实性面临的挑战，诸如产地、年份和品种造假等，以目前现有的官方检测技术很难应对。虽然我国《葡萄酒》国家标准中对品种葡萄酒进行了明确定义，但缺乏配套的检测方法和判别依据等关键技术标准，仅仅依靠企业自律和现有的标准管理体系无法有效规范市场。因此，亟需建立中国品种葡萄酒识别技术导则，对规范我国葡萄酒市场竞争秩序，促进葡萄酒向区域化、特色化、高档化良性发展具有重要意义。

根据国外相关文献报道，品种葡萄酒识别技术采用的主要分析策略是分析葡萄酒中目标特征成分主要包括矿物元素、挥发性成分、莽草酸、氨基酸、酚类或其它提取物、乙醇 $^{13}C/^{12}C$ 和水 $^{18}O/^{16}O$ 等。最新的研究成果表明一维核磁共振氢谱（^{1}H NMR）指纹图谱结合多元统计学技术已经成功应用于葡萄酒品种的识别。葡萄酒 ^{1}H NMR 指纹图谱主要优点是核磁共振信号能够提供挥发性化合物、酚类、有机酸、花色苷、氨基酸等高通量数据，其中包括那些无法定量或鉴别的物质。

鉴于此，标准起草工作组综合采用了葡萄酒核磁共振氢谱、有机酸和花青素等品种指纹特征信息，结合多元统计学手段，构建品种真实性鉴别模型，实现了对赤霞珠、黑比诺、玫瑰蜜、梅鹿辄、蛇龙珠、霞多丽、龙眼、白玉霓 8 个品种葡萄酒样品的有效识别，为《品种葡萄酒识别技术导则》国家标准的制定提供了理论和实践技术依据，对完善我国葡萄酒标准体系、促进行业规范具有重要意义，也为行业及企业构建品种葡萄酒质量信息数据库提供技术支撑。

1. 核磁共振技术识别品种葡萄酒

1.1 葡萄酒样品信息

本实验所有葡萄酒样品均由不同葡萄酒企业提供，2010-2015 年，共收集 170 个葡萄酒样品，其中 99 个红葡萄酒（赤霞珠 45 个，玫瑰蜜 24 个，蛇龙珠 30 个），71 个白葡萄酒（白玉霓 29 个，龙眼 21 个，霞多丽 21 个）。其中 123

个葡萄酒样品来自沙城产区，其余的样品来自云南弥勒、宁夏、新疆产区。所有葡萄酒样品均存放在4℃冰箱，避光冷藏。葡萄酒样品均为单一葡萄品种。

1.2 仪器与试剂

核磁共振波谱仪（400 MHz，Bruker，Germany）；pH计（SevenCompact™ S210，Mettler-Toledo，Switzerland）；叠氮化钠（超级纯，NaN_3，Merck，Germany）；3-（三甲基硅基）氘代丙酸钠（TSP）和磷酸二氢钾（98%，Merck，Germany）；重水（D_2O，98%，Merck，Germany）；氢氧化钠和盐酸（99%，Sigma-Aldrich，USA）；涡流振荡器（Biosan Ltd，Latvia）；5 mm带帽NMR测试管（Wilmad Labglas Inc，USA）。

1.3 样品前处理

采用D_2O，NaN_3，0.1% TSP，1 mol/L磷酸二氢钾配置磷酸盐缓冲液。采用磷酸或磷酸二氢钾准确调节磷酸盐缓冲液的pH值至2.0。100 μL磷酸盐缓冲液加入900 μL葡萄酒样品，采用涡流振荡器振荡30 s，混合均匀，然后采用2 mol/L的氢氧化钠或盐酸准确调整混合物的pH值至3.0±0.02。准确移取600 μL的混合物转移至5 mm NMR测试管，测试管随后采用^{1}H NMR技术进行检测。

1.4 ^{1}H NMR实验流程及仪器参数设置

^{1}H NMR检测一般实验流程包括调谐和匹配、温度控制、锁定、匀场、选择序列、参数设置以及最后样品检测。其中调谐和匹配、温度控制、锁定及匀场过程均高度自动化，操作简单。仪器实验条件设置如下：检测温度300 K（±0.2）。采用5 mm ^{1}H/D探头，^{1}H NMR的共振频率为400.13 MHz，Bruker Top Spin 2.1软件用于^{1}H NMR图谱数据的采集，检测信号时间3.9846 s，弛豫延迟为4 s，每次自由感应衰减扫描次数为32，空扫描次数为4，谱宽20.5525 ppm，采样点数为65536（64K），谱线加宽因子为0.3 Hz。以3-（三甲基硅基）氘代丙酸钠（TSP，δ= 0）作为内标，化学位移的单位（ppm）。标准的脉冲序列（NOESYGPPS）用于乙醇（δ= 1.2 ppm和3.6 ppm）和水（δ= 4.8 ppm）的信号抑制。以上各项参数设置好之后，样品检测前，需要等待5分钟用来平衡温度。

1.5 多元统计分析

采用 MestReNova 9.1 软件（Mestrelab Research S.L.，MestReNova（Mnova）NMR，USA） 首先对 ^{1}H NMR 图谱进行傅里叶变换，基线矫正、调整相位。然后对 ^{1}H NMR 图谱峰化学位移偏移进行校正对齐，最后对 ^{1}H NMR 图谱进行分段积分。分段积分值作为之后统计分析的输入变量。为了便于数据处理，葡萄酒样品属于同一品种分为一组。多元统计分析包括无监督主成分分析（principal conponent analysis，PCA）和有监督的线性判别分析（linear discriminant analysis，LDA）以及随机深林（random forest，RF）。

PCA 作为多元探索性数据降维分析的第一步，根据解释 ^{1}H NMR 数据集内 95%方差来确定主成分矩阵。主成分得分矩阵作为 LDA 以及 RF 的输入变量，由于主成分之间相互正交，因此，可以避免 ^{1}H NMR 数据集里的变量之间高度共线性问题。LDA 可以实现因变量组间差异的最大化，组内样品距离的最小化，提高模型的有效性和解析能力。随机深林基于多个决策树的分类器，输出的分类判别结果由决策树输出的类别的众数决定，这能够有效的避免模型的过拟合现象。

PCA、LDA 和 RF 数据分析均由 R 软件 3.2.3 版本处理完成。PCA、LDA 以及 RF 分析采用的 R 软件具体的数据包分别是 mdatools、mass 以及 RandomForest。PCA、LDA 和 RF 分析之前，所有数据由 R 软件采用帕累托（Pareto）方法进行标准化。应用帕累托方法即每个变量除以其标准差（SD）的平方根，有利于降低数据矩阵中变量具有很大的数值的相对重要性而不增加基线噪声。随后采用重复双随机交叉验证和留一交叉验证方法对建立的 PCA/LDA 模型的有效性进行验证，均由 R 软件编程完成。

1.6 结果与分析

1.6.1 ^{1}H NMR 数据质量控制

所有 170 个葡萄酒样品采用 ^{1}H NMR 进行了测定。不同品种之间葡萄酒样品记录的氢谱看起来非常相似。由于采用标准的脉冲序列（NOESYGPPS）用于水（4.8 ppm）以及乙醇（1.2 ppm，3.6 ppm）信号压制，葡萄酒中微量成分的信号强度得到了明显的提高。将所有葡萄酒样品的核磁共振氢谱叠加，进一步检查发现一些氢谱出现了基线较大的偏移，可能是由于水和乙醇的抑制不当，pH 调节不

准确，系统误差或其它原因。因此，需要对这些葡萄酒样品进行重复测定，以保证数据质量的可靠性。

1.6.2 ^{1}H NMR 葡萄酒图谱主要代谢物的归属

在保证图谱数据质量的基础上，对 ^{1}H NMR 图谱的主要葡萄酒代谢物/成分（氨基酸，糖，有机酸，乙醇）进行了归属。通过化学物质标品验证主要代谢物的归属是否正确。本研究提供了一个典型中国红葡萄酒 ^{1}H NMR 图谱，并且在图谱上具体给出了葡萄酒主要代谢物/成分的共振信号（见图 1.1）。

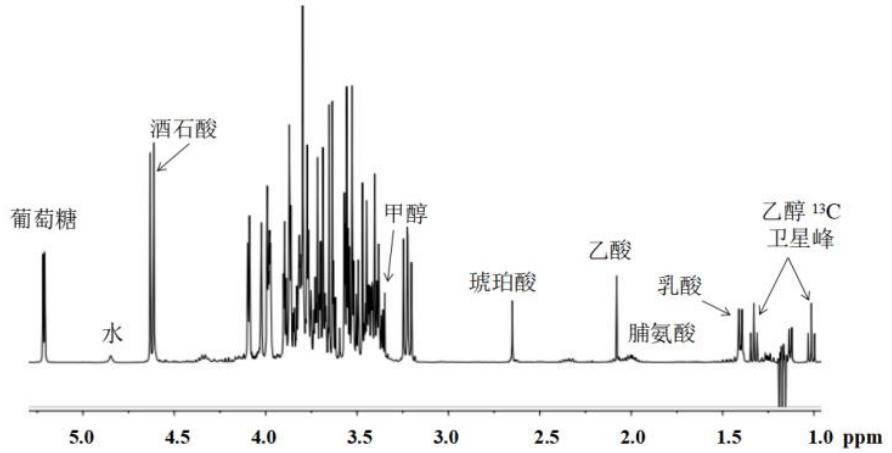

图 1.1 一个典型的标有主要葡萄酒代谢物的中国红葡萄酒 ^{1}H NMR 图谱（其中果糖、葡萄糖和甘油信号出现在 3～4.2 ppm 高度重叠区域）

1.6.3 ^{1}H NMR 图谱峰化学位移对齐以及数据降维

由于样品 pH 值微小的变化会造成 ^{1}H NMR 图谱峰化学位移偏移，本研究葡萄酒氢谱一些典型峰信号出现的化学位移偏移如图 1.2（A）和图 1.2（B）所示。采用适当的峰化学位移偏移校正对齐方法对随后的多元变量统计分析相当重要。化学位移偏移会引起化学计量学模型样本错误的分组。为了解决这一问题，至今为止已经报道了多种峰化学位移校正对齐方法。MestReNova 采用的是分段对齐的方法，该方法能够允许用户采用交互式页面，灵活的选择峰化学位移校正对齐的区域。每段的平均图谱作为参考图谱进行对齐，依次进行下去，直到得到满意的 ^{1}H NMR 图谱。图 1.2（A）和图 1.2（B）采用分段对齐后的图谱分别如图 1.2（C）和图 1.2（D）所示。由图 1.2 可知，分段对齐方法较好的解决了 ^{1}H NMR 图谱峰化学位移的偏移问题。在 MestReNova 完成 ^{1}H NMR 图谱整个峰化学位移

对齐以及分段积分之后，整个 ^{1}H NMR 图谱被细分为多个区域，每个区域所有点的积分值累加起来可以抽象的代表整个原始图谱，这些区域的积分值随后作为统计建模的输入变量，从而实现对每 ^{1}H NMR 图谱约六万个高共线性数据点的有效降维。

选取了 0.01，0.015，0.02，0.025，0.03，0.035 和 0.04 ppm 作为每个分段积分的区域宽度的取值，研究结果表明 0.01ppm 能够保持足够的数据分辨率以及能够最大限度地减少图谱信息的损失。由于 ^{1}H NMR 中的水和乙醇信号的区域与本研究调查的问题无关，所以这些区域被排除在外。分段积分的有效区域是在 0.8 ppm ~ 9.5 ppm（其中不包含被特定脉冲序列压制极度变形的水峰δ（H_2O）= 4.65 - 5.10 ppm）和乙醇峰（δ（CH_3）= 1.05 - 1.21 ppm，δ（CH_2）= 3.60 - 3.72 ppm），分段积分后大约能够获得 700 个分段积分值，这些分段积分值随即作为统计分析的输入变量。

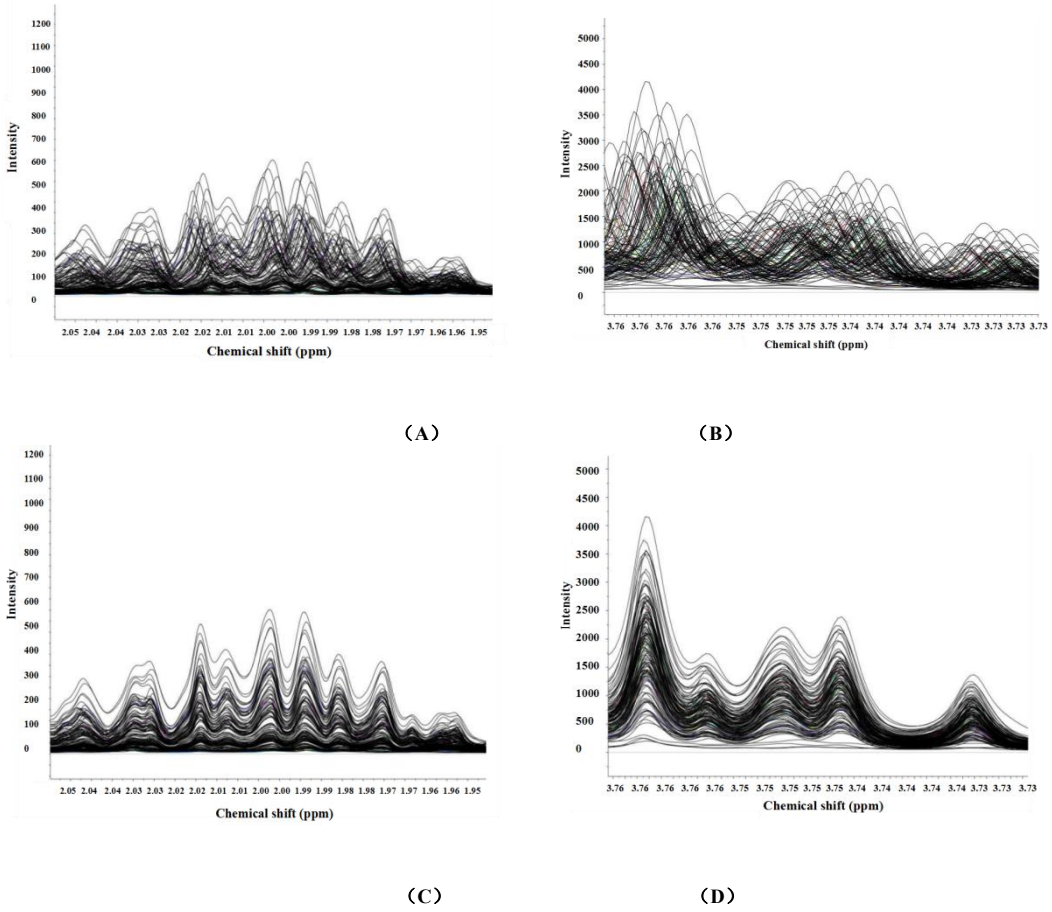

图 1.2 葡萄酒样品一维核磁共振叠加氢谱化学位移在（1.95 - 2.05）ppm 及（3.73 - 3.76）ppm 区域对齐前后的比较（A 与 C，B 与 D）

1.6.4 变质葡萄酒快速识别

在化学计量学数据分析之前，所有的葡萄酒样品进行是否变质筛查。葡萄酒样品会由于不恰当的储存条件以及超过葡萄酒本身的货架期而导致变质或腐败。通过叠加所有葡萄酒样品的核磁共振氢谱，结果发现一些葡萄酒样品图谱在 2.07 ppm 处的乙酸信号峰的绝对积分值明显高于其它大多数葡萄酒样品。因此，对这些葡萄酒样品需要进行仔细检查以发现潜在的变质。根据 GB15037-2006 规定了葡萄酒的挥发性酸含量的最大值为 1.2 g/L，这意味着葡萄酒中乙酸的含量不能超过 1.2 g/L。基于外部标准的 PULCON（pulse-length-based concentration）方法对所有葡萄酒样品中的乙酸含量进行定量。结果表明 28 个红葡萄酒样品（17 个赤霞珠，9 个蛇龙珠，2 个玫瑰蜜），3 个霞多丽白葡萄酒样品，乙酸含量已经超过 1.2 g/L，因此做进一步的统计分析建模之前需要排除这些酸败样品。

然而 Godelmann 等直接删除乙酸信号峰建模，在本研究中则是保留了乙酸信号峰（乙酸含量低于 1.2 g/L），与 Papotti 等有关于葡萄酒产地的研究的 ^{1}H NMR 图谱前处理相一致。本研究很好的保持非目标指纹图谱方法特点，而不是预先忽略乙酸或其它腐败参数对结果的影响。尽管乙酸与葡萄品种几乎没有关联，并且在随后建立的品种 PCA / LDA 以及 PCA/RF 模型， 发现乙酸信号峰区域并没有对葡萄酒品种的区别能力有重要贡献。

1.6.5 探索性数据分析

主成分分析利用降维的思想，通过线性变换把多指标转为少数几个综合指标，且综合指标完全不相关，综合指标按照所得方差从高到低的顺序依次排列。在数学变换中依旧保持变量的总方差不变，使第一变量具有最大的方差，称为第一主成分，第二变量的方差次大，称为第二主成分。主成分得分图是对数据集可能存在分组可视化有效的工具。

将 ^{1}H NMR 谱图进行峰对齐、分段积分等数据预处理操作后，将包含红、白葡萄酒整个数据集导入 R 数据分析软件中进行主成分分析。由主成分分析可得图 1.3，图 1.3 初步反映了红、白葡萄酒样品在主成分 1（PC 1）和主成分 2（PC2）组成的二维坐标体系中的无监督的分组情况，同时也可以得到 PCA 模型前两个主成分解释数据的方差累积贡献率约为 26%，PC1 占总变量的 17%。PC2 占总变量

的 9%。通过对图 1.3 进一步观察发现，红葡萄酒样品相对白葡萄酒样品聚集的较为分散；在置信区间为 95%时，红葡萄酒和白葡萄酒样品区分的趋势比较明显，结果表明 PCA 模型能够实现对葡萄酒类型（红/白）进行区分。

下一步将整个葡萄酒数据集分为包含三个红葡萄酒品种的数据集以及包含三个白葡萄酒品种的数据集。分别对这两个数据集进行 PCA 分析。为此，对 PCA 的主成分进行了计算，直到数据集中至少有 95%的方差得到解释。对于红葡萄酒品种模型，前 15 个主成分解释了数据集中 95.4%的方差；对于白葡萄酒品种模型，前 12 个主成分解释了数据集中 95.6%的方差。根据红、白葡萄酒数据集分别建立的 PCA 模型，结果表明葡萄品种都没法实现区分。

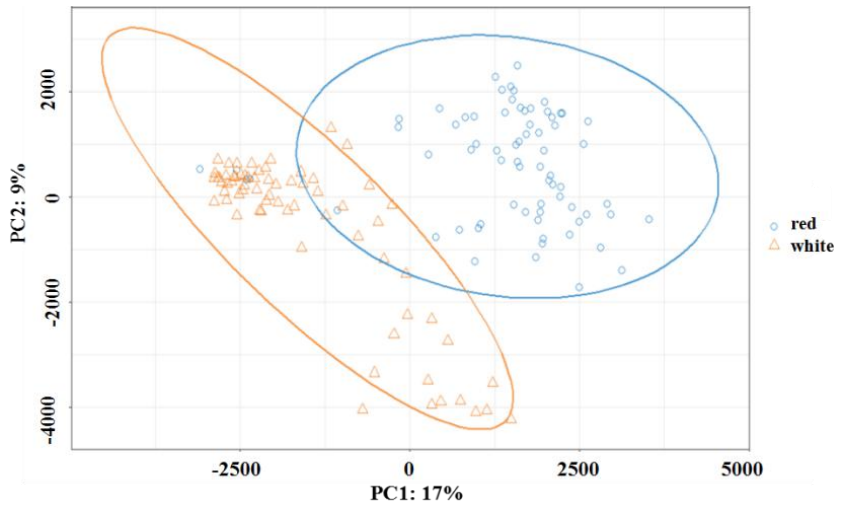

图 1.3 红、白葡萄酒样品 PCA 得分图（实线表示 95%的置信区间）

1.6.6 线性判别分析（LDA）以及随机深林（RF）

为了解决 PCA 没法区分葡萄酒品种，引入了 LDA 以及 RF 这两种分类方法进一步数据分析。由于 ^{1}H NMR 图谱原始数据既有对分类起作用的差异变量，同时包含大量对分类无作用的变量，变量之间也存在高度共线性问题。为了解决这一难题，本研究根据 PCA 所得的解释红、白葡萄酒数据集内至少解释 95%方差，分别对应的相互独立的 15 个、12 个主成分，作为 LDA 和 RF 的输入变量。红、白葡萄酒品种则作为分类变量分别建模。

通过 PCA/LDA 建立的线性判别函数能够使这样具有高维（多变量）、小样本数据分类可视化效果最大化。图 1.4 和图 1.5 表示的是红、白葡萄酒品种 PCA/LDA 建立的线性判别函数 1 和 2 得分图。由图 1.4 和图 1.5 可以看出葡萄酒

样品组内的离散的程度减少，组间的区分更加清晰，表明 PCA/LDA 模型能够很好的分别对红葡萄酒品种和白葡萄酒品种进行有效的区分。红、白葡萄酒品种内部留一交叉验证 PCA/LDA 模型的分类结果（见表 1.1）。留一交叉验证方法的原理：假设数据集有 N 个样本，随机选取 N-1 个样本作为训练样本，剩下的一个样本作为测试样本。这样每个样品将会被模型预测一次，因此，得到的 N 个分类结果可以用来保守的衡量模型的分类性能。由表 1.1 可知红葡萄酒品种赤霞珠、玫瑰蜜和蛇龙珠的正确分类率分别为 68%，100% 和 81%。白葡萄酒品种白玉霓、龙眼和霞多丽分别实现了 93%，95% 和 94%正确分类。红、白葡萄酒品种平均分别实现了 82%和 94%正确分类。赤霞珠与蛇龙珠这两个品种葡萄酒的正确分类率比较低的主要原因是赤霞珠与蛇龙珠葡萄品种在一定程度上具有高度相似性。

随机深林算法的输出的分类结果是基于 200 个决策树输出的类别的众数投票决定，从而得到红、白葡萄酒品种内部留一交叉验证 PCA/RF 模型的分类结果（见表 1.2）。由表 1.2 可知红葡萄酒品种赤霞珠、玫瑰蜜和蛇龙珠的正确分类率分别为 75%，100% 和 82%。白葡萄酒品种白玉霓、龙眼和霞多丽的正确分类率为 93%，81% 和 89%。红、白葡萄酒品种的平均正确分类率分别为 85%和 89%。

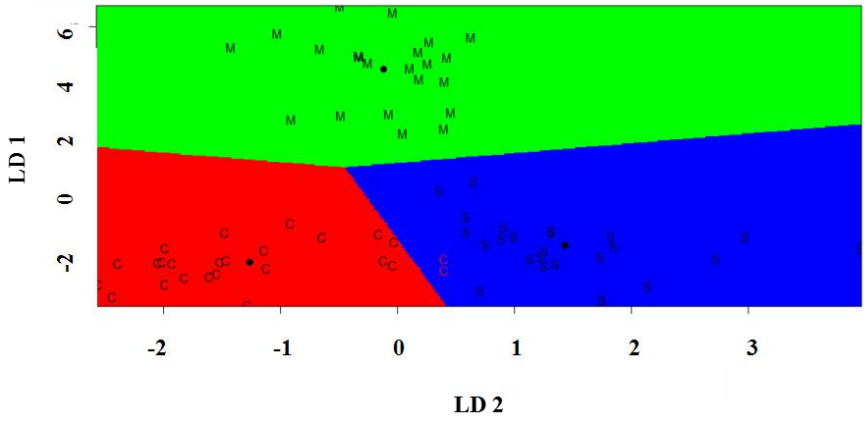

图 1.4 PCA/LDA 模型对红葡萄酒品种（赤霞珠（C）、玫瑰蜜（X）和蛇龙珠（S））线性判别函数（LD 1）和线性判别函数（LD2）得分图

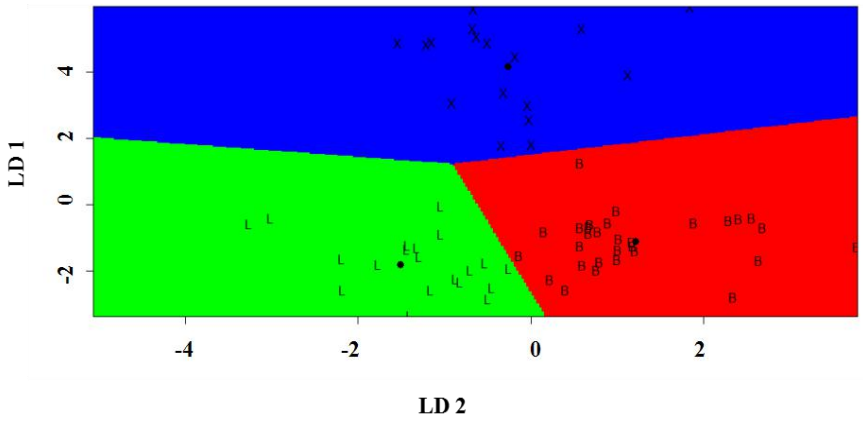

图 1.5 PCA/LDA 模型对白葡萄酒品种（白玉霓（B）、龙眼（L）和霞多丽（X））线性判别函数（LD 1）和线性判别函数（LD2）得分图

1.6.7 PCA/LDA 及 PCA/RF 模型有效性验证

在实际数据分析过程中 PCA/LDA 及 PCA/RF 模型通常容易产生过拟合现象，从而得到较好的分类判别效果的假象。多元统计模型的验证对于任何指纹图谱的方法的评估和评价至关重要。为此本研究采用了外部重复双随机交叉验证来验证 PCA/LDA 及 PCA/RF 模型分类结果的可靠性。重复双随机交叉验证主要的原理是首先采用随机分层重复抽样方式（重复抽样 100 次），将原始数据集分为两部分:（1）80%数据作为训练集用于建模；（2）20%数据作为外部验证集用于预测。由于对整个数据集随机外部验证集抽样 100 次，于是就得到了 100 个不同的外部验证集。对这 100 个外部验证集样本正确分类率的平均值、标准偏差、中位数以及中位数的绝对标准偏差进行统计分析（见表 1.3 和表 1.4），用这 4 项指标评估模型的有效性。正确分类率的平均值及中位数这两个指标数值越接近 1，标准偏差和中位数绝对标准偏差值越小，表明建立的模型正确分类能力越强。

从表 1.3 可以看出，PCA/LDA 模型红葡萄酒品种赤霞珠、玫瑰蜜和蛇龙珠的平均正确分类率分别为 71%，100% 和 77%。白葡萄酒品种白玉霓、龙眼和霞多丽的正确分类率为 97%，82% 和 91%。红、白葡萄酒品种的平均正确分类率分别为 83%和 90%。PCA/LDA 模型外部验证集红、白葡萄酒品种样本正确分类率的平均值和中位数均大于 0.7，且平均值及中位数相对应的标准偏差均较小；对比表 1.1 和表 1.3 可以发现，PCA/LDA 内部留一交叉验证与外部重复双随机交叉验证获得的红、白葡萄酒品种的平均正确分类率统计学上无显著性差异（P >

0.05），进一步表明建立的 PCA/LDA 模型的有效性。

从表 1.4 可以看出，PCA/RF 模型红葡萄酒品种赤霞珠、玫瑰蜜和蛇龙珠的平均正确分类率分别为 74%，98% 和 71%。白葡萄酒品种白玉霓、龙眼和霞多丽的正确分类率为 97%，85% 和 94%。红、白葡萄酒品种的平均正确分类率分别为 81%和 92%。PCA/RF 模型外部验证集红、白葡萄酒品种样本正确分类率的平均值和中位数均大于 0.7，且平均值及中位数相对应的标准偏差均较小；对比表 1.2 和表 1.4 可以发现，PCA/RF 内部留一交叉验证与外部重复双随机交叉验证获得的红、白葡萄酒品种的平均正确分类率，依然在统计学上无显著性差异（P > 0.05），进一步表明建立的 PCA/RF 模型的有效性。

表 1.1 红、白葡萄酒品种内部留一交叉验证 PCA/LDA 模型的分类结果

葡萄酒品种	PCA/LDA （留一交叉验证）		
	正确分类样品（个）	错误分类样品（个）	P （%）
赤霞珠	19	9	68[a]
玫瑰蜜	22	0	100[a]
蛇龙珠	17	4	81[a]
红葡萄酒	58	13	82[b]
白玉霓	27	2	93[a]
龙眼	20	1	95[a]
霞多丽	17	1	94[a]
白葡萄酒	64	4	94[b]

[a] 每个葡萄酒品种正确分类的百分比；[b] 红、白葡萄酒的平均正确分类百分比。

表 1.2 红、白葡萄酒品种内部留一交叉验证 PCA/RF 模型的分类结果

葡萄酒品种	PCA/LDA （留一交叉验证）		
	正确分类样品（个）	错误分类样品（个）	P （%）
赤霞珠	21	7	75[a]
玫瑰蜜	22	0	100[a]
蛇龙珠	18	4	82[a]
红葡萄酒	61	11	85[b]
白玉霓	27	2	93[a]
龙眼	17	4	81[a]
霞多丽	16	2	89[a]
白葡萄酒	60	8	89[b]

[a] 每个葡萄酒品种正确分类的百分比；[b] 红、白葡萄酒的平均正确分类百分比。

表 1.3 红、白葡萄酒品种外部重复双随机交叉验证 PCA/LDA 模型对测试集的分类结果（测试集是随机选取 20%整个数据集，重复循环 100 次）

葡萄酒品种	平均值（%）	标准偏差（%）	中位数（%）	中位数绝对偏差（%）
赤霞珠	71	±20	80	±30
玫瑰蜜	100	± 0	100	± 0
蛇龙珠	77	± 20	75	± 37
红葡萄酒	83[a]	± 13[a]	85[a]	± 22[a]
白玉霓	97	± 7	100	± 0
龙眼	82	± 21	88	± 19
霞多丽	91	± 17	100	± 0
白葡萄酒	90[a]	± 15[a]	96[a]	± 6[a]

[a]红、白葡萄酒的平均正确分类百分比

表 1.4 红、白葡萄酒品种外部重复双随机交叉验证 PCA/RF 模型对测试集的分类结果（测试集是随机选取 20%整个数据集，重复循环 100 次）

葡萄酒品种	平均值（%）	标准偏差（%）	中位数（%）	中位数绝对偏差（%）
赤霞珠	74	± 21	80	±29
玫瑰蜜	98	± 7	100	± 0
蛇龙珠	71	± 22	75	± 37
红葡萄酒	81[a]	± 17[a]	85[a]	± 22[a]
白玉霓	97	± 7	100	± 0
龙眼	85	± 2	100	± 0
霞多丽	94	± 16	100	± 0
白葡萄酒	92[a]	± 8[a]	100[a]	± 0[a]

[a]红、白葡萄酒的平均正确分类百分比

1.7 本章小结

本章节探讨了非目标 ^{1}H NMR 指纹图谱技术结合多元统计分析手段验证葡萄酒品种真实性的可行性，研究结果表明，PCA/LDA 和 PCA/RF 模型能够实现葡萄酒品种比较好的区分，进一步表明 ^{1}H NMR 指纹图谱技术可以作为一项非常有效的验证葡萄酒品种真实性手段。建立的相关非目标 ^{1}H NMR 指纹图数据库，能够推动我国葡萄酒品种真实性溯原制度的建立与完善，进一步保障葡萄酒市场稳定健康发展，维护消费者的合法权益。然而，本研究需要指出的是依然存在一些缺陷，例如，大部分葡萄酒品种样品主要来自沙城区域，仅有少数样品来自云南弥勒、新疆、宁夏地区等。在后续研究中，应该收集更多具有代表性不同品种、不同年份、不同产地且来源真实可靠的葡萄酒样品。

2. 葡萄酒中花青素及有机酸特征指纹识别品种葡萄酒

2.1 葡萄酒样品信息

本实验所有葡萄酒样品均由不同葡萄酒企业提供，2009-2010 年，共收集 68 个红葡萄酒样品，其中赤霞珠 18 个，玫瑰蜜 12 个，玫瑰香 13 个，蛇龙珠 13 个，梅鹿辄 12 个。葡萄酒样品来自沙城、烟台、昌黎、新疆、天津、云南和宁夏产区。所有葡萄酒样品均存放在 4℃冰箱，避光冷藏。葡萄酒样品均为单一葡萄品种。

2.2 试验数据

按照葡萄酒中花青素及有机酸检测方法对葡萄酒样品进行测定，通过多元统计学手段对数据进行处理，挖掘品种葡萄酒中的特征性成分。

2.3 结果与分析

2.3.1 单因素方差分析

方差分析主要用于两个及两个以上样本均数差别的显著性检验。采用上述检测方法测定 6 个品种葡萄酒样品，相关结果见表 2.1，表 2.1 为方差分析中 6 个不同品种挥发性成分的平均值及单因素方差结果。从 F 值和 p 值可见，葡萄酒中莽草酸、草酸、酒石酸量三种有机酸，Po-3-cugl(峰 56)、Mv-3-cugl(峰 57)、峰 30、峰 61、峰 RT25.35 等 7 种花青素等共 10 种组分在不同品种之间存在显著差异，显著水平达到 0.99 以上。但采用单一的成分，6 个品种葡萄酒均难以进行有效识别，还需要进一步采用多元数据处理模型进行有效识别。

表 2.1 不同品种葡萄酒挥发性成分的方差分析(平均值±标准误差)

	品种平均值					ANOVA	
目标成分/品种	赤霞珠	蛇龙珠	玫瑰香	玫瑰蜜	梅鹿辄	F 值	p 值
莽草酸	58.65	18.44	21.91	29.95	43.75	20.341	.000
草酸	184.00	770.97	38.77	1372.01	168.24	52.303	.000
酒石酸	1278.59	1388.24	3287.13	1626.23	1287.11	74.039	.000
De-3-gl(峰 13)	1.40	0.00	0.47	0.73	1.11	1.994	.092
Cy-3-gl(峰 15,16)	2.38	0.00	7.46	0.44	1.15	2.215	.064
Pt-3-gl(峰 17)	0.48	0.12	0.29	0.07	0.00	1.382	.243
Po-3-gl(峰 25)	2.77	0.16	2.66	0.86	0.43	.236	.945
Mv-3-gl(峰 26)	13.48	0.55	11.04	1.09	3.44	1.756	.136
Po-3-acgl(峰 44)	1.36	0.12	0.24	1.66	0.49	1.283	.283

Mv-3-acgl(峰 46)	2.44	0.04	0.71	1.11	0.00	.939	.462
Po-3-cugl(峰 56)	6.64	21.07	5.51	1.07	15.33	4.073	.003
Mv-3-cugl(峰 57)	1.50	7.44	0.98	25.49	0.00	30.348	.000
峰 30	34.94	4.52	23.70	19.63	26.28	7.552	.000
峰 51	0.83	3.92	1.24	0.67	0.21	.783	.566
峰 61	8.67	33.12	36.69	11.49	27.15	22.477	.000
峰 58	2.03	0.02	0.80	0.39	0.00	1.297	.277
峰 RT25.35	12.94	0.43	1.72	8.74	2.74	14.972	.000

2.3.2 主成分分析

为了简化数据结构，寻找综合因子从而达到减少变量维数的目的，根据因子间的相关性，对 6 个品种葡萄酒具有显著差异的 10 种成分进行主成分分析，按照指标组合 1 由 3 种有机酸组成，指标组合 2 由 3 种有机酸与 7 种花青素组成，进行多元统计分析，按照特征值的提取主成分原则，分别获得前 2 和前 3 个主成分，具体见表 2.2，其积累贡献度超过 80%，基本上可覆盖大部分数据信息。

表 2.2 品种鉴定的主成分特征值、贡献率及积累贡献率

	组合一		组合二	
主成分因子	贡献率(%)	积累贡献率(%)	贡献率(%)	积累贡献率(%)
1	45.684	45.684	34.196	34.196
2	34.384	80.068	30.622	64.818
3			16.785	81.603

不同品种葡萄酒的 PC1、PC2 的空间分布位置不一样，因而能够将不同品种的葡萄酒区分开来。图 2.1 为以 PC1 为横坐标，PC2 为纵坐标绘制的变量分布图。由图 2.2 可知不同品种的葡萄酒样品基本聚集在各自特定的区域内。对于组合 1，结合主成分的成分矩阵，PC1 正向量反映酒石酸的含量信息，PC2 的负方向反映莽草酸的含量信息，PC2 正向量反映草酸的含量信息。玫瑰蜜中酒石酸含量较高，黑比诺中草酸含量较高，梅鹿辄中草酸、酒石酸算量较高，赤霞珠中莽草酸含量较高。由图可看出，品种葡萄酒样品能够达到比较好的区分。

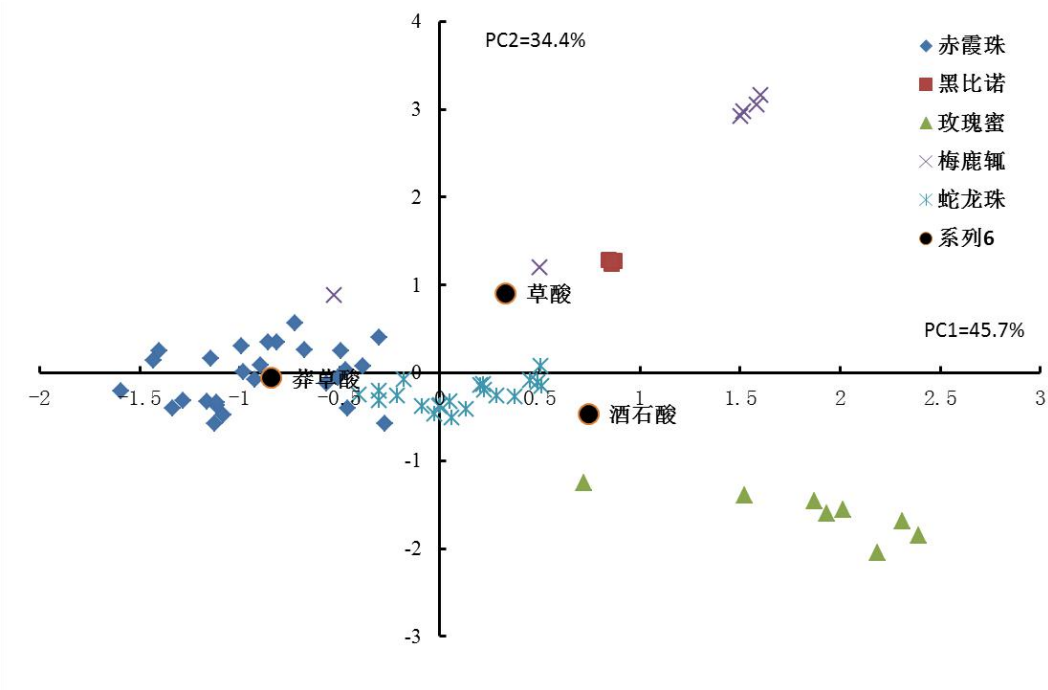

图 2.1 组合 1 的 PC1、PC2 散点图

对于组合 2，图 2.2 为以 PC1 为横坐标，PC2 为纵坐标绘制的变量分布图。结合图 2.2 可以看出，通过第 1 主成分和第 2 主成分作散点图，增加花青素作为区分的变量，并没有显著改进 6 个品种的区分，引入更多的指标变量不一定能改进品种的区分，需进一步通过判别分析进行研究。

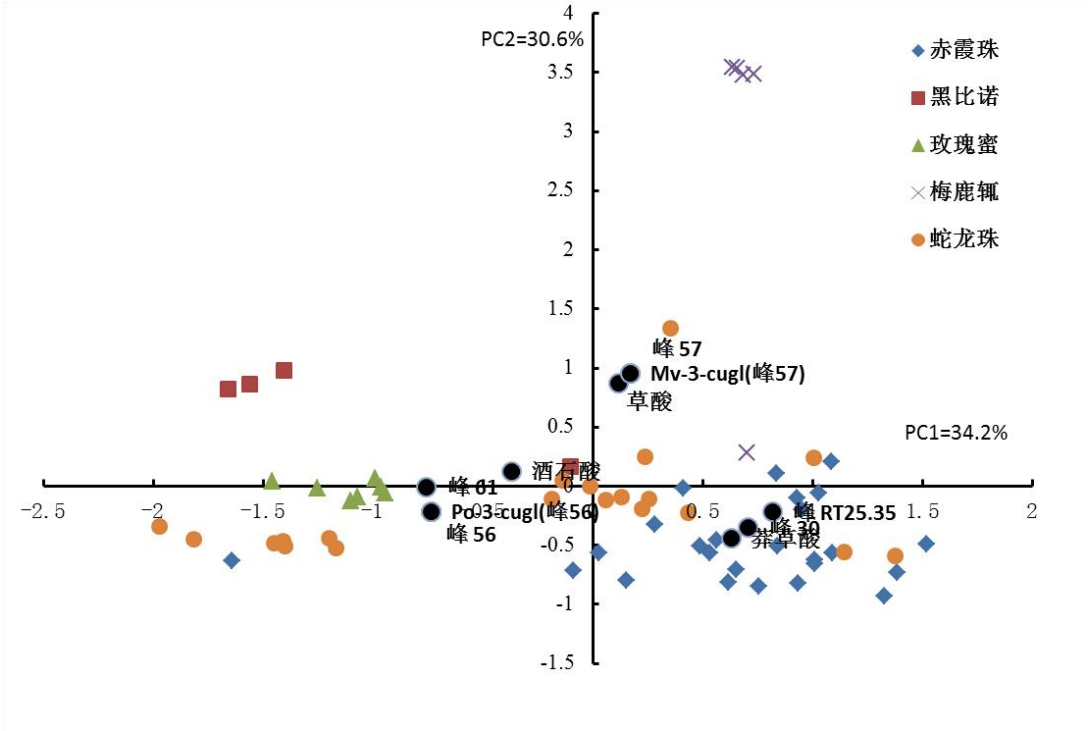

图 2.2 组合 2 的 PC1、PC2 散点图

2.3.3 判别分析

判别分析是根据表明事物特点的变量和它们所属的类，筛选出能够提供较多信息的重要变量，建立判别函数，并根据判别函数对未知所属类别的事物进行分类并使得错判率最小的一种多元统计分析方法。本文将采用逐步选择法进行不同品种葡萄酒的识别研究。采用 Fisher 系数线性逐步判别分析方法通过变量选择程序建立线性判别模型，判别分析采用 WiLḱs Lamda 方法，判别标准采用 F 值评价法，输出 Fisher 线性判别函数，筛选出对品种判别有效的变量，而剔除不必要的干扰因素，建立判别模型，并用于对新样本的判别。本文在不同品种葡萄酒中按照 2:1 比例将其分建模样品（47 个)和预测样品（21 个）；首先基于 47 个建模样品数据采用 SPSS 分类判别 Fisher 线性逐步判别分析，构建 Fisher 线性判别函数模型，结果见表 2.3 和表 2.4

表 2.3 组合一判别函数

特征成分	品种				
	赤霞珠	黑比诺	玫瑰蜜	梅鹿辄	蛇龙珠
莽草酸浓度	.238	-.003	-.229	.093	.038
草酸含量	-.004	.011	-.027	.023	-.008
酒石酸含量	.012	.013	.048	.009	.018
(常量)	-15.718	-14.613	-78.573	-24.046	-15.473

表 2.4 组合二判别函数

特征成分	品种				
	赤霞珠	黑比诺	玫瑰蜜	梅鹿辄	蛇龙珠
莽草酸浓度	.277	-.043	-.439	.104	.040
草酸含量	-.022	.013	-.055	.028	-.027
酒石酸含量	.021	.017	.069	.014	.030
Po-3-cugl(峰 56)	.536	.075	.053	.025	.426
Mv-3-cugl(峰 57)	.830	.452	1.183	1.136	.841
峰 30	.041	.034	.189	-.016	.015
峰 61	-.173	.624	.698	.383	.036
峰 RT25.35	.980	.038	.503	.428	.790
(常量)	-29.938	-30.779	-125.818	-55.456	-29.575

其次，再将 21 个预测样品代入模型进行预测评估，结果见表 2.5 和 2.6。对比表 2.5 和 2.6 可以看出，对于组合 1，黑比诺、玫瑰蜜、赤霞珠和梅鹿辄能够 100%判别正确，而蛇龙珠存在一定的误判；对于组合 2，能够对赤霞珠、黑比诺、玫瑰蜜、梅鹿辄和蛇龙珠实现 100%的判别。两者的平均判对率分别为 95%和 100%，基本符合建模要求。对于组合 2，增加了花青素指标，能够减少错判率，从而构建一个合适的判别模型对品种葡萄酒进行有效的区分。可根据实际情况，结合试验测试成本和精度要求，选择较为合适的指标组合对品种进行判别，以满足实际需要。

表 2.5 组合一识别结果

实际组	预测组					统计		
	赤霞珠	黑比诺	玫瑰蜜	梅鹿辄	蛇龙珠	验证	正确	正确率
赤霞珠	9					9	9	100%
黑比诺		1				1	1	100%
玫瑰蜜			2			2	2	100%
梅鹿辄				2		3	2	100%
蛇龙珠	1				6	7	6	86%
合计						21	20	95%

表 2.6 组合二识别结果

实际组	预测组					统计		
	赤霞珠	黑比诺	玫瑰蜜	梅鹿辄	蛇龙珠	验证	正确	正确率
赤霞珠	9					9	9	100%
黑比诺		1				1	1	100%
玫瑰蜜			2			2	2	100%
梅鹿辄				2		2	2	100%
蛇龙珠					7	7	7	100%
合计						21	21	100%

2.4 本章小结

项目基于葡萄酒中花青素及有机酸含量基础数据，通过多元统计学手段对数据进行处理分析，建立了方差分析、主成分分析和判别分析等数据处理模型，不断的对事实数据进行拟合分析，实现了对 5 个品种中 68 个葡萄酒样品的有效区分。采用判别分析模型按照 2:1 比例进行建模和预测，结果表明，选择组合合适的一些不同指标进行数据处理拟合分析，Fisher 线性判别模型对赤霞珠、黑比诺、蛇龙珠、玫瑰蜜和梅鹿辄 5 个品种葡萄酒的预测能力可达 100%。本研究构建了品种葡萄酒有效识别技术模型，检测成本较低，方法成熟，模型简单，具有很好推广应用价值。