

## HPLC 结合多元化学计量学鉴别水洗和日晒咖啡

杨浩澜<sup>1</sup>, 包亮英<sup>1</sup>, 刘实琪<sup>1</sup>, 刘亦菲<sup>1</sup>, 朱炎贞<sup>2</sup>, 刘芳<sup>1,3\*</sup>

<sup>1</sup>湖南中医药大学药学院 长沙 410208

<sup>2</sup>湖南中医药大学中西医结合学院 长沙 410208

<sup>3</sup>湖南省中药饮片标准化及功能工程技术研究中心 长沙 410208)

**摘要** 建立了一种通过筛选差异成分准确鉴别水洗和日晒咖啡的方法。采集咖啡样本的高效液相(HPLC)色谱图,使用 Kennard-Stone 算法将两类不同加工方式咖啡样品的全谱原始信号和共有峰峰面积数据分别按 7:3 的比例划分为校正集和测试集,首先建立主成分分析(PCA)及偏最小二乘判别分析(PLS-DA)模型,再通过多元算法对 PLS-DA 模型进行优化,最后将最佳模型判别函数中  $X$  变量的相关系数 $\times 100^{-1}$  标记于色谱图上,标识为两类咖啡的差异成分,并对其中含量最高的成分进行定性。结果表明,采用共有峰峰面积数据无法建立合格的预测模型以区分水洗和日晒咖啡,而采用 90 个区间下的反向区间偏最小二乘法(BiPLS)优化后的全谱原始信号建立的模型则具有良好的鉴别能力,其对测试集样本预测的准确度达 83.33%, $Q^2$  值为 0.813。标准品定性结果提示绿原酸为两类咖啡的重要差异成分之一。相较于以单个或多个化合物的含量信息作为样品差异评价标准的传统方法,本法可准确挖掘产品的差异信息,为整体评判产品质量提供了新思路。

**关键词** 高效液相色谱法; 化学计量学; 咖啡加工方式; 差异成分

**文章编号** 1009-7848(2022)09-0304-08 **DOI:** 10.16429/j.1009-7848.2022.09.032

咖啡是茜草科咖啡属的一种灌木<sup>[1]</sup>,广泛种植于热带地区,我国的主要分布地区为海南省和云南省<sup>[2]</sup>。研究发现,定期喝咖啡可预防心血管疾病、2 型糖尿病、肥胖症和某些类型的癌症,并降低中风的风险<sup>[3]</sup>。此外,因其独特的口感及提神醒脑的效果而受到世界各地消费者的追捧。据统计,全球每天约消费 22.5 亿杯咖啡<sup>[4]</sup>。随着咖啡热度的不断增加,咖啡品质相关的负面报道使得人们越来越重视咖啡的质量<sup>[5]</sup>。

水洗和日晒是咖啡最常见的两种加工方式<sup>[6]</sup>。两类咖啡在口感上存在较明显的区别,然而,其化学成分种类及含量的差异较小,凭借常规方法较难实现差异成分的筛选及对加工方式的预测。高效液相色谱(HPLC)指纹图谱能够较全面反映物质的成分和含量信息,在药物和食品的质量监测中发挥着重要作用<sup>[7-8]</sup>。面对 HPLC 图谱中蕴藏的繁杂数据,如何进行有效分析成为一大研究热点,

因化学计量学可同时控制大量变量并最大限度地提取不同类别间的差异信息<sup>[9]</sup>,故越来越被认为是实现 HPLC 图谱价值的有力补充<sup>[10]</sup>。化学计量学与 HPLC 图谱相结合的分析方法在产品类别及相应差异成分的预测方面具有独特优势,逐渐被应用于药物和食品的质量控制,现有研究思路大致可分为两类,一类采用主要化合物色谱峰面积信息建立主成分分析(PCA)或偏最小二乘判别分析(PLS-DA)模型,以探究化合物含量与质量间的关系<sup>[11-13]</sup>,然而常忽略图谱中微弱波动可能具有的重要作用,模型预测效果一般,且难以全面挖掘潜在质量标志物;另一类通过 HPLC 全谱数据信息建立产品类别的预测模型<sup>[14-16]</sup>,然而现有文献报道仅获得 HPLC 图谱对产品质量的预测模型,尚未深入探究化合物成分与类别间的潜在关联。本研究旨在建立多元化学计量学结合 HPLC 全谱原始信号寻找差异成分的方法,并与共有峰峰面积信息建立的模型进行对比分析。以水洗咖啡和日晒咖啡为例,建立两类咖啡的鉴别模型,分别探究通过全谱信息和共有峰峰面积建立的 PLS-DA 模型的效果差异,以及多元算法对 HPLC 鉴别模型的贡献,为药品、食品特征差异成分的确认和相关质量标准的建立提供一个更为精确、全面的新方法,以

**收稿日期:** 2021-09-28

**基金项目:** 国家自然科学基金项目(81704064);湖南省教育厅科学研究优秀青年项目(20B434);湖南省中药学一流学科资助项目(校行科字[2018]3号)

**作者简介:** 杨浩澜(2000—),男,本科

**通信作者:** 刘芳 E-mail: fliu0825@126.com

弥补传统方法的缺陷。

## 1 材料与方法

### 1.1 材料与试剂

供试的 40 批咖啡来源于淘宝(豆兽咖啡)和当地咖啡厅,其产地分别为巴拿马、肯尼亚、埃塞俄比亚、哥伦比亚、巴西、牙买加、卢旺达、危地马拉和洪都拉斯,每批 25 g,粒径和色泽均匀,密封包装。

甲醇(色谱纯),德国 Merck 公司;磷酸(分析纯),国药集团化学试剂有限公司;绿原酸(色谱纯),成都克洛玛生物科技有限公司;试验用水为来自 Direct-QR 8 UV Remote 纯水系统(ZRQSVR 800,美国密理博)的超纯水。

### 1.2 仪器与设备

Waters Alliance e2695 型高效液相色谱仪(四元梯度洗脱系统、配 2998 PDA Detector 紫外检测器、Empower 3 操作软件),美国 Waters 公司;LE204E/02 分析天平,梅特勒-托利多仪器上海有限公司;KM-500DE 型超声波清洗器,昆山美美超声仪器有限公司。

### 1.3 软件

使用“中药色谱指纹图谱相似度评价系统(2004A 版)”分析各样本指纹图谱间的相似度;The Unscrambler X 10.4(64-bit)划分校正集与测试集;SPSS 26 用于进行斯皮尔曼相关系数分析;SIMCA 14.1 建立 PLS-DA 模型和实现重要投影变量(VIP)值的评价;Matlab 用于反向区间偏最小二乘法(BiPLS)、连续投影算法(SPA)、无信息变量消除法-偏最小二乘法(UVE-PLS)、竞争性自适应重加权算法-偏最小二乘法(CARS-PLS)分析;数据结果图均由 Origin2020 生成。

### 1.4 HPLC 分析

1.4.1 样品及标准品溶液的配制 精密称取 1.0000 g 咖啡样品,置于具塞锥形瓶中,加入甲醇-0.1%磷酸水溶液(1:1)50.0 mL,超声提取 20 min,静置冷却后过 0.45  $\mu\text{m}$  滤膜,备用。

精密称取(0.0200 $\pm$ 0.0002)mg 绿原酸标准品,以 0.4%磷酸甲醇溶液溶解并稀释,制得质量浓度为(0.20 $\pm$ 0.002) g/L 的绿原酸标准溶液。

1.4.2 色谱条件 HPLC 分析采用 Waters Al-

liance 高效液相色谱仪(Waters2695 分离模块,Waters2998PDA 检测器),Waters Symmetry C18 色谱柱(250 mm $\times$ 4.6 mm,5  $\mu\text{m}$ );流动相 A 为甲醇,流动相 B 为 0.1%磷酸水溶液。梯度洗脱程序:0.0~1.78 min,11%~20% A;1.78~3.18 min,20%~26% A;3.18~8.09 min,26%~38% A;8.09~10.01 min,38% A;10.01~11.67 min,38%~60% A;11.67~20 min,60%~60% A。流速:1.0 mL/min;检测波长:210,254,320 nm;柱温:30  $^{\circ}\text{C}$ ;进样体积:10  $\mu\text{L}$ 。

### 1.5 多元数据分析

1.5.1 数据预处理 为考察咖啡产地差异是否可作为与其加工方式无关的随机干扰,采用 SPSS 26 计算咖啡产地与加工方式的斯皮尔曼相关系数,采用 Kennard-Stone 算法分别将 40 批样本的全谱信息和共有峰峰面积数据按 7:3 的比例划分为校正集和测试集,以保证校正集与测试集样本划分的合理性<sup>[7]</sup>。通过“中药色谱指纹图谱相似度评价系统(2004A 版)”进行分析,时间窗宽度设为 0.10,以平均数生成对照指纹图谱,各样本指纹图谱与对照指纹图谱间相似度的平均数为 0.992 提示各样本指纹图谱间无显著性差异,为尽可能地保留与峰形、化合物自身性质及化合物间相互影响等相关的原始信息,保证分析所得 PLS-DA 模型中各变量的相关系数能准确、真实地反映指纹图谱原始信息与各类别间的联系,故未对原始图谱信息进行其它处理。

1.5.2 PLS-DA 模型的建立 PLS-DA 模型可提取与样本类别相关度最高的数据信息,即最大限度地提取不同类别数据之间的差异<sup>[8]</sup>,从而判别咖啡的加工方式。分别以 HPLC 全谱和两类咖啡共有峰为数据源建立 PLS-DA 模型,采用 7 折交叉验证和置换检验(置换次数为 200 次)检查模型,并通过模型的预测能力(Q<sup>2</sup>)、校正集均方根误差(RMSEE)、交叉验证均方根误差(RMSE<sub>cv</sub>),以及测试集的准确度、敏感度、特异度对模型质量进行评价。其中 Q<sup>2</sup> 值用于衡量模型的预测能力,其值大于 0.5 提示模型具备良好的预测能力<sup>[8]</sup>。RMSEE 与 RMSE<sub>cv</sub> 值越小,则表明模型与数据的偏差越小、模型交叉验证结果越准确。准确度(Accuracy)、敏感度(Sensitivity)、特异度(Specificity)用于评价模型对测试集中各样本所属类别

的预测能力<sup>[19]</sup>。计算公式如下:

$$\text{RMSEE} = \sqrt{\frac{1}{n-A-1} \sum_{i=1}^n (Y_{ip} - Y_i)^2} \quad (1)$$

$$\text{RMSEcv} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_{ip} - Y_i)^2} \quad (2)$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN}) \quad (3)$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \quad (4)$$

$$\text{Specificity} = \text{TN} / (\text{FP} + \text{TN}) \quad (5)$$

式中:  $Y_{ip}$  和  $Y_i$  分别表示校正集或交叉验证中的第  $i$  个样本类型的预测值和实测值 (日晒及水洗判别式中的实测  $Y$  值均默认为 1);  $A$  为因子数;  $n$  为样本集数;  $\text{TP}$  和  $\text{FN}$  分别为测试集中日晒类别的真阳性或假阳性数目;  $\text{TN}$  和  $\text{FP}$  则分别为测试集中水洗类别的真阳性或假阳性数目。

**1.5.3 模型优化及差异成分的确定** 数据中存在的噪音变量不但干扰样品类别的判定, 而且影响模型质量, 故应当筛选变量以消除噪音干扰, 提高模型的准确性。通过筛选 VIP 值大于 1 的变量以及 BiPLS、SPA、UVE-PLS、CARS-PLS 等变量优选算法分别对两类数据建立的 PLS-DA 模型进行优化<sup>[20-21]</sup>, 其中 BiPLS 算法的区间数设为 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, CARS 算法的迭代采样次数设置为 100 次。由于 BiPLS 算法是基于区间进行模型优化的, 最佳区间组合中可能仍存在干扰变量, 因而采用 SPA、UVE、CARS 等算法对效果最佳的 BiPLS 模型继续进行优化, 以消除可能残余的干扰变量。为进一步探寻优选变量与咖啡类别间的关系, 将最佳模型判别函数中  $X$  变量的相关系数  $\times 100^{-1}$  标记于色谱图上, 生成相关系数标记图, 充分挖掘两类咖啡的差异成分及其与加工方式之间的关联。

## 2 结果与分析

### 2.1 HPLC 分析

综合比较了 210, 254, 320 nm 波长下 40 批咖啡样品的 HPLC 全谱, 选择信息量最大的 254 nm 波长下的数据进行化学计量学分析, 结果如图 1 所示。从中发现, 日晒咖啡中各主要峰的峰面积均相对较大, 提示日晒咖啡中主要成分的含量可能高于水洗咖啡, 但仅凭峰面积值的大小不能科学、合理地判断两类咖啡间的差异。为筛选出反映咖

啡加工方式的差异成分, 分别导出全谱数据及共有峰数据进行多元化学计量学分析, 其中, 共有峰数据选择信号值较强且峰形较好的 17 个峰 (图 1)<sup>[28]</sup>。由表 1 可知, 斯皮尔曼相关系数分析所得 Sig. (双尾) 值为 0.496, 提示本研究咖啡样本的产地虽存在差异, 但该差异属于随机干扰, 与其加工方式间并无显著关联。

表 1 咖啡产地与加工方式的斯皮尔曼相关系数分析  
Table 1 Analysis of Spearman correlation coefficient between the producing area and processing mode of coffee

属性	参数	产地	加工方式
产地	相关系数	1	-0.111
	Sig. (双尾)	-	0.496
	$N$	40	40
加工方式	相关系数	-0.111	1
	Sig. (双尾)	0.496	-
	$N$	40	40

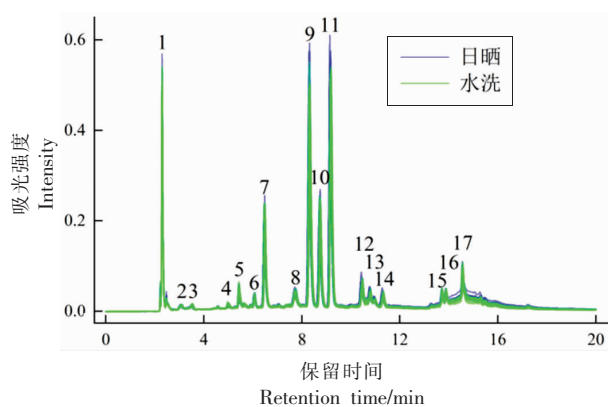


图 1 咖啡样品的 HPLC 全谱及两类咖啡的共有峰  
Fig.1 HPLC full chromatograms and common peaks between the two types of coffee

### 2.2 模型优化及差异成分的确定

**2.2.1 基于多元算法的模型优化** 采用 BiPLS、VIP、SPA、CARS、UVE 等算法处理后的共有峰峰面积数据仍无法建立有效的 PLS-DA 模型, 而处理后的全谱数据建立的 PLS-DA 模型却发生了极大改善, 结果见表 2, 与原始模型相比, 经过变量优选后的模型均获得了较好的效果, 其中采用 90 个区间建立的 BiPLS 模型准确度最高, RMSEcv 值最小, 以及远远大于 0.5 的  $Q^2$  值, 分别为 83.33%,

0.217,0.813,提示此模型具有较好的预测能力,为变量优选后的最优模型。由优化后的模型得分图(图 2)可知,水洗与日晒两类加工方式的咖啡样本在独立贡献度最高的 3 个主成分上区分较好,该模型中日晒咖啡类别的实际值与预测值拟合图则清晰地展现了模型预测效果(图 3),各样本坐标分布越集中于曲线 X=Y 附近,则模型效果越

好。采用算法优选后的变量分布情况见图 4,从中可见经 BiPLS 算法优选后的残余变量数目较多,连续性好且主要分布于化合物峰上,能较好的保留化合物信息,而经其它算法优选后的残余变量数目较少,分布零散,经 SPA、UVE 算法优选后仅极少变量分布于化合物峰上,由此可见,采用 Bi-PLS 算法进行变量优选的意义大于其它算法。

表 2 基于多元算法的全谱变量优选结果

Table 2 Full graphs variable optimization results based on multivariate algorithm

算法类型	区间	模型评价指标						
		A	Q2	准确度/%	敏感度/%	特异度/%	RMSEE	RMSEcv
BiPLS	10	2	0.253	58.33	37.50	100.00	0.418	0.496
	20	16	0.755	66.67	75.00	50.00	0.051	0.479
	30	5	0.748	58.30	50.00	75.00	0.189	0.338
	40	4	0.576	58.33	50.00	75.00	0.228	0.337
	50	9	0.734	33.33	37.50	25.00	0.082	0.305
	60	8	0.611	75.00	75.00	75.00	0.151	0.349
	70	4	0.308	58.33	37.50	100.00	0.382	0.440
	80	4	0.568	66.67	62.50	75.00	0.228	0.333
	90	6	0.813	83.33	87.50	75.00	0.192	0.217
	100	5	0.608	75.00	75.00	75.00	0.251	0.310
VIP	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
SPA	(-)	19	0.889	33.00	25.00	50.00	0.024	0.457
CARS	(-)	14	0.633	67.00	75.00	50.00	0.076	0.460
UVE	(-)	8	0.654	58.33	50.00	75.00	0.166	0.359

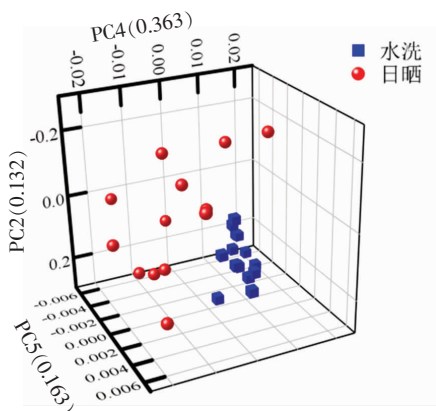


图 2 PLS-DA 模型得分图

Fig.2 The score of PLS-DA model

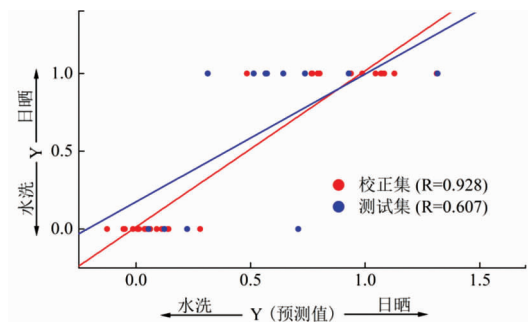


图 3 日晒类别的实际值与预测值拟合图

Fig.3 Matching diagram of actual values and predicted values of sun exposure category



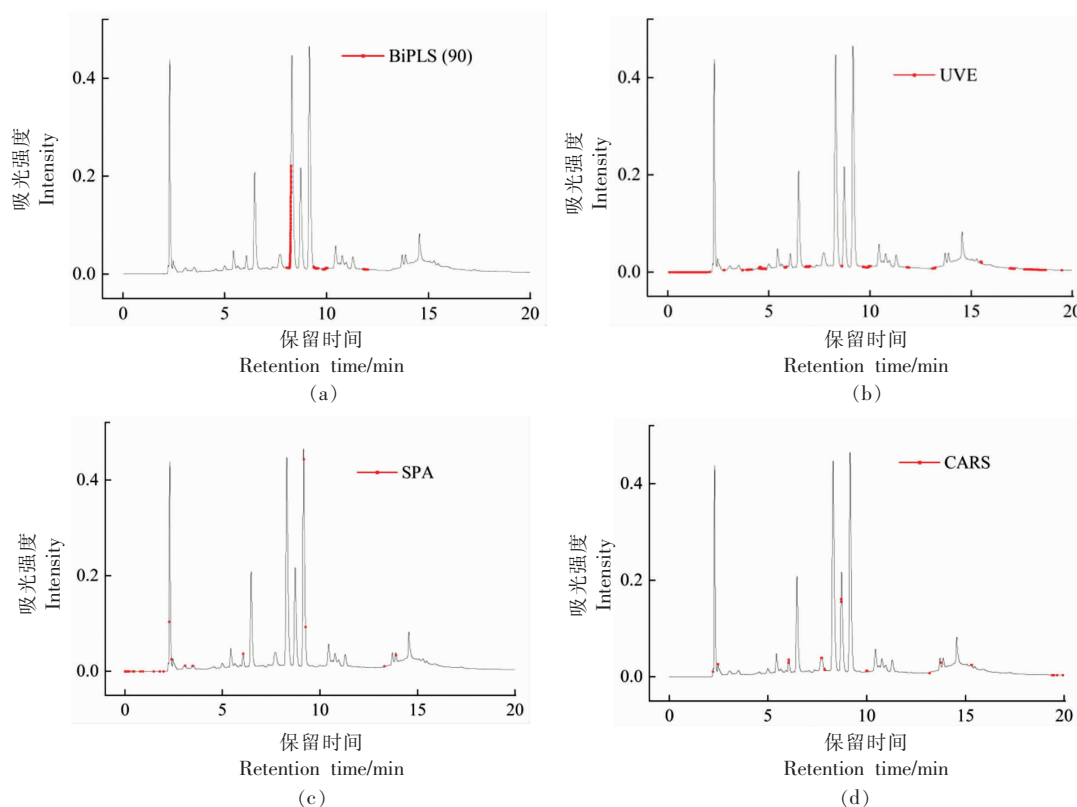


图4 经 BiPLS (a)、UVE (b)、SPA (c)及 CARS (d)算法优选后的变量分布图

Fig.4 Variable distribution maps optimized by BIPLS (a), UVE (b), SPA (c) and CARS (d) algorithm

2.2.2 基于组合算法的变量优选 采用 VIP、SPA、UVE、CARS 等算法对效果最佳的 BiPLS 模型进行再次优化,结果见表 3,经各类组合算法优化后建模效果反而降低,其中通过 SPA 算法处理后的模型 Q2 值下降最为显著,相较处理前最佳的 BiPLS 模型,其 Q2 值由 0.813 降为 0.0812,推测可

能由于过度优选变量破坏了 BiPLS 算法建立的区间连续性,从而使模型失去了对色谱峰信息的描述能力。综合比较后选择模型效果最好且信息保留较多的 90 个区间下的 BiPLS 模型作为两类咖啡样本的鉴别模型。

表 3 基于组合算法的变量优选结果

Table 3 Variable optimization results based on combination algorithm

算法类型	模型评价指标						
	A	Q2	准确度/%	敏感度/%	特异度/%	RMSEE	RMSEcv
BiPLS-VIP	6	0.784	75.00	75.00	75.00	0.207	0.232
BiPLS-SPA	2	0.0812	50.00	50.00	50.00	0.460	0.478
BiPLS-CARS	3	0.673	66.67	62.50	75.00	0.220	0.289
BiPLS-UVE	3	0.313	58.33	37.50	100.00	0.364	0.417

2.2.3 差异成分的确定 最佳模型中优选变量的相关系数标记图见图 5,该图详细反映了每个差异变量与两类咖啡间的关系,其中系数为正的变量值越高,样品为日晒咖啡的可能性则越高。若单

个峰中的大部分变量值为正,则提示该峰对应的化合物能在日晒咖啡样品中检测出较大含量,为日晒咖啡的特征成分;反之,色谱峰包含的变量相关系数为负数,则提示该峰为水洗咖啡的特征成

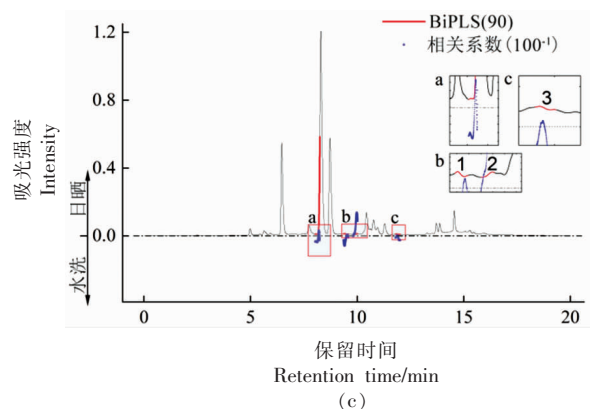
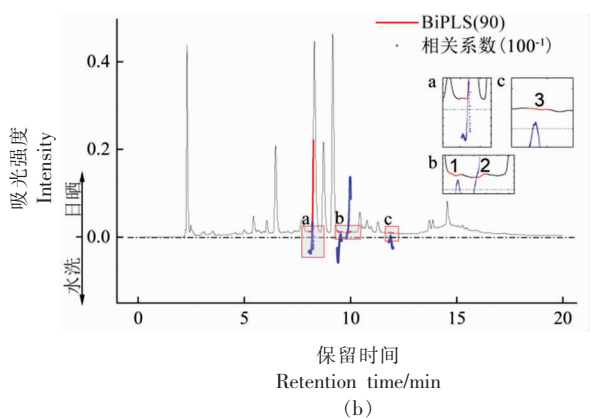
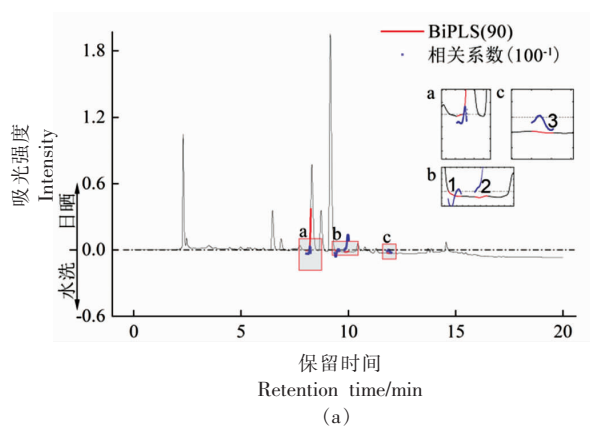


图 5 210nm(a)、254nm(b)及 320nm(c)各变量的相关系数及局部放大图(a-c)

Fig.5 Correlation coefficients and partial enlarged of drawing (a-c) in 210 nm(a), 254 nm(b) and 320 nm(c)

分,以此有效区分水洗和日晒两类咖啡,由此可判断 1~3 号波动依次为水洗、日晒、水洗加工方式的潜在特征成分(图 5a~5c 中的 5b、5c)。在本试验选定的色谱洗脱梯度和 210 nm 或 254 nm 检测波长下,1~3 号相应信号仅仅为较小的细微波动,但变

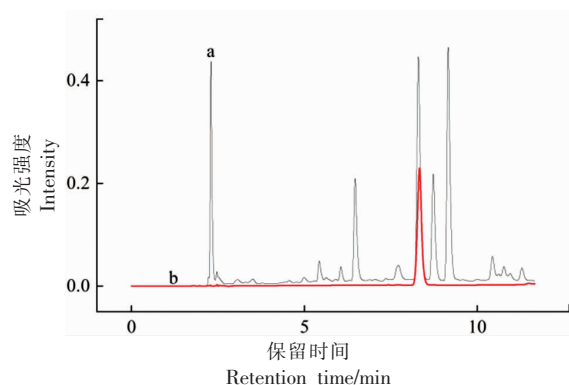


图 6 咖啡样品(a)和绿原酸标准品(b)的 HPLC 色谱图  
Fig.6 HPLC chromatogram of coffee sample (a) and chlorogenic acid standard (b)

量优选及建模结果提示这些波动对咖啡类别的区分具有重要意义,故图中 1~3 号信号为可反映两类咖啡内在特性的差异成分。从图 5c 可见,320 nm 下的 3 处波动信号值较 254 nm 及 210 nm 均出现明显提升,且峰形改善,提示通过优化色谱条件可改善最佳模型中仅为细微波动的差异成分响应值及峰形,为后续对该峰进行定性定量分析提供途径,最终全面准确地挖掘样品中可能存在的内在差异成分,体现样品的特征性。综合分析相关系数标记图,结合标准品定性分析,确认最佳模型中优选变量较多且含量最高的化合物为绿原酸(图 6),绿原酸含量对咖啡加工方式的影响为其峰上原始数据的综合反应,而非单纯正负相关(图 7a~7c 中的 7a),这正是在积分过程中舍去单个数据点与类别间关系的共有峰数据无法实现对两类样本进行鉴别预测的主要原因。

### 3 结论

通过联立 HPLC 全谱与多元化学计量学,构建一个全面准确鉴别水洗咖啡和日晒咖啡内在特征差异成分的方法。研究表明,共有峰的数据在积分过程中失去单个数据点与加工方式之间独立的相关性,故而难以区分两类咖啡,而 HPLC 全谱中丰富的信息不仅包含反映化合物含量的峰面积积分信息,还包含了各种含量较低的化学成分信息以及每个峰的峰形信息,可全面反映各个化合物的内在特征及各化合物之间可能存在的相互作用,最终有效识别两类样本,故相较于以单个或

多个共有峰评判产品质量的传统方法,采用 HPLC 全谱结合多元化学计量学的方法建立的鉴别模型准确度更高、描述能力更强且优化空间更大。通过多元化学计量学分析方法比较发现,BiPLS 算法能建立良好的预测模型并保留丰富的连续信息,使用组合算法反而破坏了 BiPLS 算法的预测效果。本研究最终通过相关系数标记图确认了多个水洗咖啡与日晒咖啡的特征差异成分,经标准品定性确认含量最高的重要差异成分为绿原酸。

本文构建的方法全面体现了样品内在的丰富信息,最终全面挖掘所有可能的特征差异成分,该方法具有极高的应用性和适用性,通过将 Y 值替换为药品质量等级、炮制程度或药理学参数等其它数据,可在质量控制领域发挥作用,从而更好地监管药品食品质量,维护消费者的健康与权益。

### 参 考 文 献

- [1] CHENG K, DONG W, LONG Y, et al. Evaluation of the impact of different drying methods on the phenolic compounds, antioxidant activity, and in vitro digestion of green coffee beans[J]. *Food Science & Nutrition*, 2019, 7(3): 1084-1095.
- [2] DONG W, TAN L, ZHAO J, et al. Characterization of fatty acid, amino acid and volatile compound compositions and bioactive components of seven coffee (*coffea robusta*) cultivars grown in Hainan Province, China[J]. *Molecules*, 2015, 20(9): 16687-16708.
- [3] SOCAŁA K, SZOPA A, SEREFKO A, et al. Neuroprotective effects of coffee bioactive compounds: A review [J]. *International Journal of Molecular Sciences*, 2020, 22(1): 107.
- [4] CORDOBA N, PATAQUIVA L, OSORIO C, et al. Effect of grinding, extraction time and type of coffee on the physicochemical and flavour characteristics of cold brew coffee[J]. *Scientific Reports*, 2019, 9(1): 8440-8452.
- [5] KIYAMA R. Estrogenic activity of coffee constituents [J]. *Nutrients*, 2019, 11(6): 1401.
- [6] CAMPOS G, SAGU S T, CELIS P S, et al. Comparison of batch and continuous wet-processing of coffee: changes in the main compounds in beans, by-products and wastewater[J]. *Foods*, 2020, 9(8): 1135-1153.
- [7] YUAN B, BYRNES D R, DINSSA F F, et al. Identification of polyphenols, glycoalkaloids, and saponins in solanum scabrum berries using HPLC-UV/Vis-MS[J]. *Journal of Food Science*, 2019, 84(2): 235-243.
- [8] SUN M, WU H, HE M, et al. Integrated assessment of medicinal rhubarb by combination of delayed luminescence and HPLC fingerprint with emphasized on bioactivities based quality control [J]. *Chinese Medicine*, 2020, 15(1): 845-854.
- [9] ZHU J, YANG Y, DUAN S, et al. The antialgal mechanism of luteolin-7-o-glucuronide on *phaeocystis globosa* by metabolomics analysis[J]. *International Journal of Environmental Research and Public Health*, 2019, 16(17): 3222-3232.
- [10] SAHU P K, RAMISETTI N R, CECCHI T, et al. An overview of experimental designs in HPLC method development and validation[J]. *Journal of Pharmaceutical and Biomedical Analysis*, 2018, 147: 590-611.
- [11] SU C, ZHENG X, ZHANG D, et al. Investigation of sugars, organic acids, phenolic compounds, antioxidant activity and the aroma fingerprint of small white apricots grown in Xinjiang[J]. *Journal of Food Science*, 2020, 85(12): 4300-4311.
- [12] CAI Z, WANG C, CHEN C, et al. Quality evaluation of *Lonicerae Japonicae Flos* and *Lonicerae Flos* based on simultaneous determination of multiple bioactive constituents combined with multivariate statistical analysis [J]. *Phytochemical analysis: PCA*, 2021, 32(2): 129-140.
- [13] CAO X, SUN L, DI L, et al. Quality evaluation of *phellodendri chinensis cortex* by fingerprint-Chemical pattern recognition[J]. *Molecules*, 2018, 23(9): 2307-2317.
- [14] VERA D N, JIMÉNEZ-CARVELO A M, CUADROS-RODRÍGUEZ L, et al. Authentication of the geographical origin of extra-virgin olive oil of the Arbequina cultivar by chromatographic fingerprinting and chemometrics[J]. *Talanta*, 2019, 203: 194-202.
- [15] BIKRANI S, JIMÉNEZ-CARVELO A M, NECHAR M, et al. Authentication of the geographical origin of margarines and fat-spread products from liquid chromatographic UV-Absorption fingerprints and chemometrics[J]. *Foods*, 2019, 8(11): 588.

- [16] SHEN T, YU H, WANG Y Z. Assessing geographical origin of gentiana rigescens using untargeted chromatographic fingerprint, data fusion and chemometrics[J]. *Molecules*, 2019, 24(14): 2562.
- [17] SANTOS M C D, MONTEIRO J D, ARAÚJO J M G, et al. Molecular fluorescence spectroscopy with multi-way analysis techniques detects spectral variations distinguishing uninfected serum versus dengue or chikungunya viral infected samples[J]. *Scientific Reports*, 2020, 10(1): 13758–13771.
- [18] RAZALI M T A, ZAINAL Z A, MAULIDIANI M, et al. Classification of raw stingless bee honeys by bee species origins using the NMR- and LC-MS-based metabolomics approach[J]. *Molecules*, 2018, 23(9): 1–18.
- [19] BALLABIO D, CONSONNI V. Classification tools in chemistry. Part 1: Linear models. PLS-DA[J]. *Analytical Methods*, 2013, 5(16): 3790–3798.
- [20] LI X, ZHOU R, SUN C, et al. Detection of lead chrome green illegally added in tea based on confocal Raman spectroscopy[J]. *Spectroscopy and Spectral Analysis*, 2017, 37(2): 461–466.
- [21] ZHANG D, XU L, WANG Q, et al. The optimal local model selection for robust and fast evaluation of soluble solid content in melon with thick peel and large size by Vis-NIR spectroscopy[J]. *Food Analytical Methods*, 2018, 12(55): 1–12.
- [22] DING Y, WU E, LIANG C, et al. Discrimination of cinnamon bark and cinnamon twig samples sourced from various countries using HPLC-based fingerprint analysis[J]. *Food Chemistry*, 2011, 127(2): 755–760.

### Identification of Washed and Sun-dried Coffee by HPLC Combined with Multi-chemometrics

Yang Haolan<sup>1</sup>, Bao Liangying<sup>1</sup>, Liu Shiqi<sup>1</sup>, Liu Yifei<sup>1</sup>, Zhu Yanzhen<sup>2</sup>, Liu Fang<sup>1,3\*</sup>

<sup>1</sup>College of Pharmacy, Hunan University of Chinese Medicine, Changsha 410208)

<sup>2</sup>College of Integrated Traditional Chinese and Western Medicine, Hunan University of Chinese Medicine, Changsha 410208)

<sup>3</sup>Center for Standardization and Functional Engineering of Traditional Chinese Medicine in Hunan Province, Changsha 410208)

**Abstract** A method was established to identify washed and sun-dried coffee comprehensively and accurately by screening differential components. High performance liquid chromatography (HPLC) of coffee samples were collected for multi-chemometrics analysis, and then the full raw signals and common peaks' area data that came from the two different processing types of coffee samples were divided into calibration set and test set respectively using the Kennard-Stone algorithm in the ratio of 7:3. Principal component analysis (PCA) and partial least squares-discriminant analysis (PLS-DA) models were established based on those data primitively. In the following work, the raw data were processed to optimize PLS-DA model by using the sole algorithm and combinatorial algorithm, such as variable importance in projection (VIP), backward interval partial least squares (BiPLS), successive projections algorithm (SPA), uninformative variable elimination-partial least squares (UVE-PLS), competitive adaptive reweighted sampling-partial least squares (CARS-PLS) separately, and BiPLS-VIP/CARS/SPA/UVE together. After that, the correlation coefficients  $\times 100^{-1}$  of X-variables in the discriminant function of the optimum model were labeled on the chromatogram, which was used for screening quality markers between the two types of coffee, and the highest correlation component was qualitative analyzed. The results demonstrated that a qualified prediction model could not be established using the full raw signals to distinguish washed and sun-dried coffee, while the BiPLS algorithm at 90 intervals was the best optimum proposal among all kinds of algorithms, which had good discriminatory ability with an accuracy of 83.33%, a Q2 value of 0.813 and an root mean square error of cross validation (RMSEcv) value of 0.217. The qualitative results of the standards illustrated that chlorogenic acid is one of the important differential components of coffee with two different processing methods. Above all, compared to the traditional method of using the single or multiple compounds contents information as quality evaluation criteria, this identification method could dig out the differential components entirely and accurately, which providing a new thought with reference value for the product quality assessment comprehensive.

**Keywords** high performance liquid chromatography; chemometrics; coffee processing; differential components