不同生长期当归红外光谱的偏最小二乘分析

李四海,潘新波,任真,顾志荣,王亚丽*
(甘肃中医学院当归研究所,兰州 730000)

[摘要] **目的**:研究偏最小二乘(PLS)方法在不同生长期当归傅里叶变换红外光谱(FT-IR)分析中的应用。**方法**:使用正交信号校正及小波压缩(OSCW)对原始 FT-IR 信号进行预处理,然后对预处理后的光谱信号进行 PLS 分析,提取前 2 个主成分对 3 种不同生长期的当归进行聚类。**结果**:3 种生长期的当归被正确地分为 3 类,聚类结果与当归的生长期密切相关,反映了不同生长期当归在主要化学成分含量上存在一定的差异。**结论**:对光谱信号进行正交信号校正及小波压缩处理能够有效降低光谱信号的噪声,有助于提高聚类性能。

[关键词] 当归; 傅里叶变换红外光谱; 正交信号校正; 偏最小二乘
[中图分类号] R284.1 [文献标识码] A [文章编号] 1005-9903(2013)12-0132-04
[doi] 10.11653/syfj2013120132

PLS Analysis of FT-IR Spectrum of *Angelica sinensis* at Different Growth Stages

LI Si-hai, *PAN Xin-bo*, *REN Zhen*, *GU Zhi-rong*, *WANG Ya-li**
(*Gansu College of Traditional Chinese Medicine*, *Research Institute of Angelia Sinensis*, *Lanzhou* 730000, *China*)

[**Abstract**] **Objective**: To study application of partial least squares in fourier transforn infrared spectroscopy (FT-IR) spectrum analysis for *Angelica sinensis* at different growth stages. **Method**: Pretreatment method of orthogonal signal correction and wavelet compression (OSCW) was used to reject uncorrelated variables in the original spectra before partial least squares (PLS) analysis. the first two principal components were employed to cluster samples of *A. sinensis* at different growth stages. **Result**: All samples are properly classified into three categories according to their growth stage, the results of clustering was closely related to their growing period, which reflect the differences in relative content of main chemical composition among the samples from various growing period. **Conclusion**: The proposed method that established with orthogonal signal correction plus wavelet compression can decrease noise of FT-IR spectrum and help to improve the clustering performance.

[**Key words**] *Angelica sinensis*; fourier transform infrared spectroscopy; orthogonal signal correction; partial least squares

当归为伞形科植物当归的干燥根,分布于甘肃、云南、四川、青海、陕西、贵州等地,甘肃为主产区,仅定西地区岷归的种植面积和产量就占全国的 70% 左右[1]。当归味甘、辛、苦,性温,具有补血、活血、调经止痛、润燥滑肠之功效。

傅里叶变换红外光谱(FT-IR)具有无需样品的制备、快速、非破坏性等特点[2]。近年来,随着计算机技术和化学计量学的快速发展,FT-IR 技术已广泛用于药品、食品的定性、定量分析中[3]。由于 FT-IR 产生的数据维数很高,对其谱图的解析和分析需要结合相关的模式识别方法,常用的方法主要有无

[收稿日期] 20130115(013)
[基金项目] 国家自然科学基金项目(30960037);甘肃省发改委战略新兴产业和产业技术研究与开发专项项目
[第一作者] 李四海,硕士,讲师,从事模式识别及中药制剂与质量控制研究,Tel:0931-8765455,E-mail:lshroom@163.com
[通讯作者] *王亚丽,博士,教授,博士生导师,从事化学计量学及代谢组学研究,Tel:0931-8765470,E-mail:cnwyl1166@hotmail.com

监督的 PCA 及 K-近邻算法,主要用于谱图的聚类分析;有监督的方法如偏最小二乘(PLS),OPLS,PLS-DA,ANN,SVM 等,可用于对光谱进行多组分建模,以进行多种成分含量的快速无损测定[4-6]。

本研究根据甘肃产 3 种不同生长期当归的 FT-IR 信号,在 SIMCA-P 平台下,对原始光谱信号进行正交信号校正及小波压缩预处理,然后进行 PLS 分析,取前两个主成分进行聚类。结果发现,所有样本按照生长期的不同被正确地聚为 3 类。通过对聚类结果的进一步分析,指认了不同生长期当归在一些主要化学成分上的差异性,对研究当归药材生长过程中所含物质的变化具有一定的参考意义。

1 仪器与样品

Bruker 公司的 a-ALPH-T 傅里叶变换红外光谱仪,DTGS 检测器。德国科恩 ABT 80-4M 型电子天平,溴化钾为分析纯。

当归样本全部采自甘肃岷县麻子川乡,共 30 个样本。其中一年期样本 4 个,二年期样本 15 个,三年期样本 11 个。均经甘肃中医学院中药鉴定教研室鉴定为 *Angelica sinensis* (Oliv.) Diels。样本采集及编号情况见表 1。

表 1 当归样本实验编号

No.	采收日期	No.	采收日期
1	一年期 8 月 30 日	16	二年期 10 月 5 日
2	一年期 9 月 13 日	17	二年期 10 月 14 日
3	一年期 10 月 1 日	18	二年期 10 月 24 日
4	一年期 10 月 8 日	19	二年期 11 月 7 日
5	二年期 3 月 29 日	20	三年期 4 月 30 日
6	二年期 4 月 4 日	21	三年期 5 月 29 日
7	二年期 5 月 29 日	22	三年期 6 月 28 日
8	二年期 6 月 15 日	23	三年期 7 月 26 日
9	二年期 6 月 28 日	24	三年期 8 月 30 日
10	二年期 7 月 12 日	25	三年期 9 月 13 日
11	二年期 7 月 26 日	26	三年期 9 月 30 日
12	二年期 8 月 15 日	27	三年期 10 月 29 日
13	二年期 8 月 27 日	28	三年期 3 月 14 日
14	二年期 9 月 14 日	29	三年期 3 月 30 日
15	二年期 9 月 25 日	30	三年期 4 月 27 日

注:采收部位均为根。

分别将当归植物样品干燥粉碎,过 200 目筛。取样品粉末与 KBr 粉末(1∶100)混合研磨充分均匀,压片,采用 Bruker 公司的 a-ALPH-T 傅里叶变换红外光谱仪,DTGS 检测器扫描测定,获得一维红外谱图,光谱范围为 4 000 ~ 400 cm^{-1},扫描信号累加次数为 16 次。扫描时实时扣除水和 CO_2 的干扰。不同生长期当归的傅里叶变换红外光谱图见图 1。

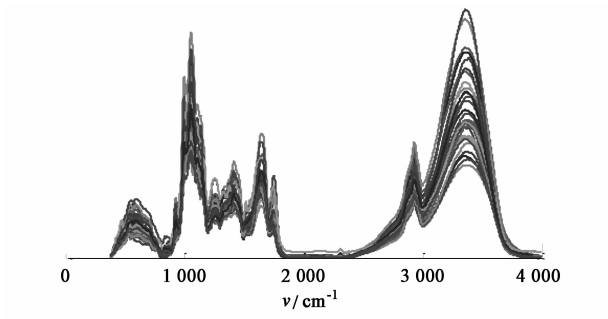

图 1 当归傅里叶变换红外光谱

2 方法

2.1 FT-IR 谱图分析 当归的生长期可分为 3 个阶段:第一年为育苗期,第二年为成药期,第三年为留种期。成药期是当归生长过程中的主要物候期,当归传统的采收期一般为第二年的 10 月上、中旬。

随着生长期的不同,当归中一些主要化学成分的含量会有不同,这些差异会反映在其红外谱图中。

由图 1 可知,3 种生长期当归的红外谱图比较相似,重合程度较高。其中一年期当归与其他两类当归的差异较为明显,而二年期和三年期当归的谱图重合程度较高。仅从一维谱图上很难准确指认不同生长期当归在化学成分上的差异性,需要对光谱进行预处理,降低噪声对有效信号的干扰,提高信噪比,以利于对光谱信号的进一步分析。光谱信号的预处理对光谱的定性、定量分析十分重要,通常要根据分析的目的选择适当的预处理方法。本文选择对聚类分析针对性较强的正交信号校正法进行光谱预处理。

2.2 正交信号校正 目前,对光谱信号的预处理方法主要有平滑、多元散射校正(MSC)、标准正态变量校正(SNV)、一阶、二阶导数、正交信号校正(OSC)及小波去噪和压缩等。其中,正交信号校正能够滤除与因变量无关的信息,保留有用信息,具有较强的特征提取能力[7-8]。

正交信号校正(orthogonal signal correction, OSC)方法由 Wold 等于 1998 年提出,并应用于光谱计量学领域。其目的是去除原始光谱矩阵中与因变量无关的信息,保留有用信息,增强对光谱分析的目的性和针对性。

设光谱矩阵为 X,因变量矩阵为 Y(品质、含量、浓度等)。OSC 方法希望从光谱矩阵 X 中提取出正交主成分 T^{**},使其与因变量矩阵 Y 的相关性最

小,同时求得载荷向量 P^{**},然后去除这些正交主成分[9],即:

$$X = T^{**}P'^{**} + E$$

$$X_{osc} = E$$

其中 E 为残差矩阵,X_{osc}为正交信号校正后的光谱矩阵,将其做为新的自变量矩阵再进行相关分析。

3 结果

3.1 光谱预处理 对原始全光谱信号进行正交信号校正,剔除与生长期类别无关的信息,保留有用信息,然后选用 db6 小波进行分解,分解水平为 6,将光谱点压缩为原始光谱点的 $1/2^6$,为 40 维,此时能够保留原始光谱 95% 的能量信息。预处理后的光谱信号见图 2。

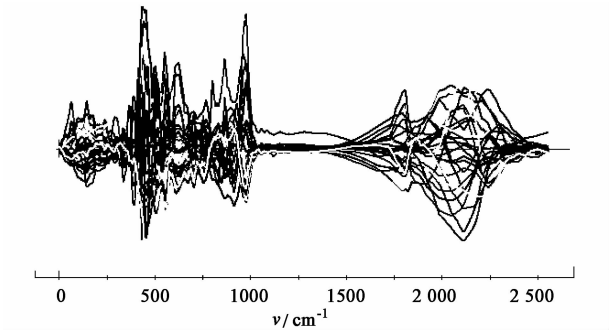

图 2 OSCW 预处理后的光谱信号

3.2 聚类分析 在 SIMCA-P 软件中,主成分得分图反映了原始样本在以主成分变量组成的新坐标空间中的分布状况,又称为样本的散点图。通常选取前 3 个主成分就可以直观显示样本在二维或三维空间中的聚集和离散程度,发现样本之间的差异。其中第一个主成分包含了原始光谱数据的最大差异。

对经过预处理后的光谱信号分别进行 PCA 及 PLS 分析,比较无监督和有监督的模式识别方法的聚类结果。其中,主成分的个数根据 R^2 和 $Q^2$2 个参数选取,通常 $R^2, Q^2 \geqslant 0.5$,且二者差值不易过大。

图 3 显示了 PCA 的前 2 个主成分 t_1 和 t_2 的得分图,前两个主成分对自变量的累计解释水平为 68.6%。其中,$Q^2 = 0.574$。从得分聚类图可以看出:二年期和三年期当归样本没有完全区分开。

图 4 显示了 PLS 的前 2 个主成分 t_1 和 t_2 的得分图,前 2 个主成分对自变量和因变量的累计解释水平分别为 $R_X^2 = 0.631$, $R_Y^2 = 0.916$。其中,$Q^2 = 0.875$,表明模型交叉验证结果显著。从得分聚类图可以看出,3 类当归样本已经被完全正确地区分开,类别之间的分类间隔相对较大,表明模型具有较好的泛化能力。比较 2 种模式识别方法的结果可知,

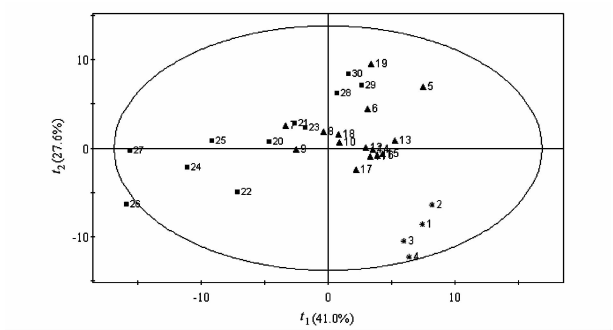

图 3 OSCW 滤噪后 PCA 得分聚类分析

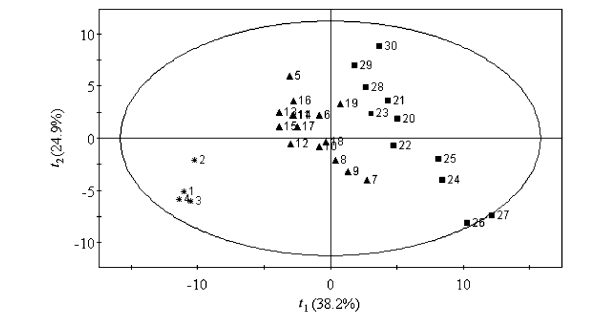

图 4 OSCW 滤噪后 PLS 得分聚类分析

PCA 所得到的主成分仅能代表原始变量信息,其对光谱信号的降维处理并没有考虑因变量信息,而 PLS 分析对原始光谱的降维处理考虑了因变量信息,从对当归的类别分析结果来看,其效果明显好于 PCA。

从图 4 的聚类结果看,所有样本被聚为 3 类,且均在置信圆内(置信度为 0.95),聚类效果较好。其中一年期样本与其他两类样本距离最远,显示出较大的差异性。进一步分析发现,采收期相近的样本,在聚类图上的位置也相近,如,三年期样本中的 20 ~23,28 ~30 共 7 个样本的采收期集中在 3 ~7 月,在聚类图中其位置彼此相近,形成一个集中聚集区域;24 ~25 两个样本的采收期为 8 ~9 月,这两个样本在聚类图上位置也十分相近;26 ~27 号样本的采收期最晚,在聚类图上其位置距离其他同类样本也最远。值得注意的是 19 号样本,该样本的采收期在二年期同类样本中时间最晚(第 2 年 11 月 7 日),其在得分图上的位置位于二年期和三年期两类样本之间,对模型的泛化能力有一定影响,是一个潜在的特异点。总之,样本在聚类图中的位置与样本的采收期密切相关,说明正交信号校正很好地保留了当归光谱中与生长期类别有关的特征信息,剔除了无关信息,对提高聚类效果的作用显著。

3.3 变量的 VIP 曲线分析 为对聚类结果进行一

些定量分析,图 5 给出了变量的 VIP(variable importance for the projection, VIP)曲线。

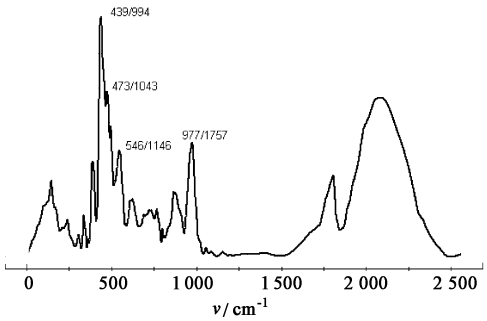

图 5 变量的 VIP 曲线

VIP 曲线反映了变量对自变量的解释水平及与因变量的相关程度,VIP 值越大,该变量越重要。通过 VIP 曲线的峰值点,可以发现不同生长期当归在化学成分上的一些差异。

通常认为 VIP > 1,说明该变量重要;VIP < 0.5,说明该变量不重要。在上图中,VIP > 1 的光谱点主要集中在全谱的前半部分,图中标记了 VIP > 1.5 的几个峰值点,其中峰值处标记的数值对表示图中的变量点与原始光谱波数之间的对应关系。在这些峰值点上,不同生长期当归的化学成分含量之间存在显著差异,其中在 994 cm^{-1}处差异最为明显。根据文献[10],对这 4 个峰值点进行指认和归属:在当归药材的红外光谱中,994 cm^{-1}为糖环的伸缩振动吸收峰,1 043 cm^{-1}处是 C-O 伸缩振动吸收峰,为当归中多糖类物质和苷类物质的特征峰;1 146 cm^{-1}处为酯类和纤维素类物质的 C-O 伸缩振动吸收峰,1 757 cm^{-1}是酯羰基、羧酸类及挥发油类物质中 C=O伸缩振动吸收峰。上述结果表明糖类、苷类、酯类及羧酸类物质是区分不同生长过程当归药材的主要差异物,为研究当归药材生长过程的所含物质变化提供参考和科学依据。

3.4 结果分析 由于将原始光谱矩形与生长期类别矩阵进行了正交运算,最大程度保留了与生长期类别有关的光谱信息,这种预处理方法对提高聚类效果的针对性较强。从聚类结果看,PLS 对预处理后光谱信号的聚类效果要优于 PCA。通过对 PLS 聚类结果的进一步分析,指认了不同生长期当归在化学成分含量上的不同,这些结果与已有的研究成果一致,说明了本文聚类方法的合理性和有效性。

4 讨论

正交信号校正能够剔除原始光谱中的无关变量信息,保留有用信息,是一种非常有针对性的光谱信号预处理方法。对预处理后的傅里叶变换红外光谱信号进行偏最小二乘分析,提取前两个主成分进行聚类,聚类结果与当归的生长期呈现出较强的相关性,生长期相近的当归在聚类图上的位置也接近。通过 VIP 解释水平曲线发现,不同生长期当归在糖类、苷类、酯类等物质的含量上的差异性较为明显。本文对于进一步研究当归生长过程中化学成分含量的动态变化具有一定的借鉴意义。

[参考文献]

[1] 王亚丽,唐红霞,朱书强. 甘肃不同产地当归的比较分类研究[J]. 中国中药杂志, 2009, 34(11): 1390.

[2] 王丹,卜海博,李向日. 红外光谱技术在中药炮制研究中的应用与展望[J]. 中国实验方剂学杂志,2011, 17(7):269.

[3] 卢红梅,梁逸曾. 代谢组学分析技术及数据处理技术[J]. 分析测试学报, 2008, 27(3): 325.

[4] Santos R N, Galvao R K, Araujo M C, et al. Improvement of prediction ability of PLS models employing the wavelet packet transform: A case study concerning FT-IR determination of gasoline parameters [J]. Talanta, 2007, 71(3): 1136.

[5] 郑娟梅,毛晓丽,李自达,等. FTIR 定量测定扶芳藤中卫矛醇的方法研究[J]. 中国实验方剂学杂志, 2012,18(6):60.

[6] 张福强,唐向阳,王俊全,等. 基于机器学习的红外光谱丹参聚类分析[J]. 计算机与应用化学,2010,27(9):1301.

[7] 尼珍,胡昌勤,冯芳. 近红外光谱分析中光谱预处理方法的作用及其发展[J]. 药物分析杂志, 2008, 28(5): 824.

[8] 张娴,袁洪福,郭峥,等. 正交信号校正应用于多元线性回归建模的研究[J]. 光谱学与光谱分析, 2011, 31(12):3228.

[9] 任芊,解国玲,董守龙,等. OSC-PLS 算法在近红外光谱定量分析中应用的研究[J]. 北京理工大学学报, 2005, 25(3): 272.

[10] 孙素琴. 中药红外光谱分析与鉴定[M]. 北京: 化学工业出版社, 2010.

[责任编辑 邹晓翠]