第 28卷第 5期
2009年 10月

红 外 与 毫 米 波 学 报
J. Infrared Millim. Waves

Vol. 28, No. 5
October, 2009

文章编号: 1001 - 9014(2009)05 - 0357 - 05

基于净分析物预处理算法的绿茶中儿茶素的近红外光谱定量分析

陈全胜, 郭志明, 赵杰文, 欧阳琴
(江苏大学 食品与生物工程学院,江苏 镇江 212013)

摘要:由于原始近红外光谱数据中含有与待测组分不相关的噪音及冗余信息,增加了偏最小二乘法(PLS)模型的复杂程度.为了简化儿茶素的预测模型,采用净分析物预处理法(NAP)对近红外光谱进行预处理,提取出待测组分的净分析物信号,然后利用 PLS建立绿茶中三种儿茶素(EGCG、ECG和 EGC)含量的(NAP/PLS)模型.在模型建立过程中,通过交互验证的方法优化 NAP因子数及模型的主成分因子数,并且将 NAP的结果与经典的标准正态变量(SNV)光谱预处理结果相比较.比较结果显示,经过 NAP与 SNV光谱预处理后,模型的预测结果相差不大,但是经过净分析物预处理后,模型的主成分因子数大大降低.研究结果表明,NAP光谱预处理算法能在保证精度的前提下有效地简化绿茶中儿茶素含量的预测模型.

关 键 词:近红外光谱;净分析物预处理法;偏最小二乘;儿茶素
中图分类号: O657.33 **文献标识码**: A

QUANTITATIVE ANALYSIS OF THE CATECHINS CONTENTS IN GREEN TEA WITH NEAR INFRARED SPECTROSCOPY AND NET ANALYTE PREPROCESSING ALGORITHM

CHEN Quan-Sheng, GUO Zhi-Ming, ZHAO Jie-Wen, OUYANG Qin
(School of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: Complex degree of partial least squares(PLS) model is often increased due to the noise and redundant information in raw near infrared spectra. In order to simplify PLS model, net analyte preprocessing (NAP) algorithm was used to extract some useful net analyte signals from the raw spectra, then three NAP/PLS models of EGCG, ECG and EGC were constructed. The number of NAP factors and the number of PLS components were optimized by cross-validation. The spectral preprocessing result of NAP algorithm was compared with that of the classical standard normal variate(SNV). The predicting result of NAP is almost similar to that of SNV, but the number of PLS components factor by NAP is much less than that by SNV. This work demonstrates that NAP pretreatment can simplify the prediction models of catechin content in green tea.

Key words: near infrared spectrum; net analyte preprocessing; partial least squares(PLS); catechin

引言

绿茶中含有大量的儿茶素,具有抗癌防癌和抗氧化等功效,正受到越来越多的关注.绿茶中儿茶素类物质成分复杂,数量与种类繁多,其中最主要的三类儿茶素分别是表没食子儿茶素没食子酸酯(epigallocatechin gallate, EGCG)、表儿茶素没食子酸酯(epicatechin gallate, ECG)和表没食子儿茶素(epigallocatechin, EGC).它们也是形成茶汤苦涩、收敛和鲜爽等滋味的重要因子.随着绿茶消费量的增加,绿茶的品质质量的控制得到了更多的关注,其中,EGCG、ECG和 EGC含量通常是衡量绿茶品质的重要参数[1].

绿茶中儿茶素含量的测定通常情况下采用高效液相色谱[2]和毛细血管电泳[3]等理化检测方法,这些方法费时费力,而且属于破坏性检测,不适合绿茶

收稿日期: 2008 - 08 - 12,**修回日期**: 2009 - 03 - 02 **Received date:** 2008 - 08 - 12, **revised date:** 2009 - 03 - 02
基金项目:国家自然科学基金项目(30800666);江苏省博士后科研资助计划项目(0901048C)和江苏省高校自然科学基金项目资助(08KJB550003)
作者简介:陈全胜(1973-),男,安徽桐城人,副教授,博士,主要从事基于光谱和图像分析技术的食品、农产品品质快速无损检测研究.

流通过程中的快速检测.近红外光谱检测技术由于具有快速、无损以及可以同时检测多种品质的优点,近年来越来越广泛地被应用到茶叶及其它农产品品质的检测中[4~8].但采集得到的原始光谱数据常含有因外界环境不稳定造成的噪音信息以及与待测品质不相关的冗余信息等,在模型的校正过程中,这些信息的介入势必会增加模型的复杂程度,影响模型的精度和稳定性.常规的光谱数据预处理方法有标准正态变量(SNV)、多元散射校正(MSC)、一阶导数(FOD)和二阶导数(SOD)等,它们可以滤除因仪器或外界环境不稳定造成的噪音信号或基线漂移,但是这些方法不能消除光谱信号中与待测成分不相关的冗余信息[9].前期研究工作发现,在利用近红外光谱技术检测绿茶中儿茶素含量时,进行常规预处理后,建立的PLS模型往往过于复杂(模型的主成分因子数过高),影响了模型的稳定性.鉴于此,本研究尝试采用一种新的光谱数据预处理方法——净分析物预处理法(net analyte preprocessing, NAP)对绿茶的近红外原始光谱进行预处理.NAP通过空间正交的途径最大程度地剔除了原始光谱中与儿茶素含量不相关的信息,并结合PLS方法建立EGCG、ECG和EGC含量的预测模型,以期获得较为理想的简化模型.

1 材料与方法

1.1 试验材料

试验所用的材料分别来自中国江苏、安徽、浙江、云南、福建和河南等6个省份11种不同品牌的绿茶,所有茶叶的出厂日期为07年3月至5月.试验前用咖啡粉碎机将绿茶样品粉碎,过40目筛,随机称取1g作为一个样本,对所得到的绿茶粉末样本进行光谱扫描.试验选用11种绿茶,每种绿茶选择10个样本,共110个绿茶样本,选择其中75个绿茶样本建立PLS校正模型,其余的35个作为预测集来验证模型的稳健性.

1.2 光谱采集

试验采用Antaris Ⅱ傅里叶变换近红外光谱仪(Thermo Scientific公司,美国)进行茶叶的近红外光谱数据采集,该仪器采用InGaAs检测器和漫反射积分球检测装置.光谱扫描范围为10000~4000cm^{-1};扫描次数为32次;采样间隔为1.928 cm^{-1},这样每条光谱就有3112个变量.试验时,将1g左右的绿茶样本均匀地倒入样品杯(仪器的标准配件)中,然后将其置于积分球上方进行光谱采集.每个样本在不同时间测定四次,取其平均光谱作为该样本的原始光谱.试验过程中尽量保持试验室内的温度和湿度基本一致.

1.3 参考检测方法

本试验采用高效液相色谱(HPLC)作为参考方法来检测绿茶样本中的3种儿茶素(EGCG、EGC和ECG)含量.检测采用的仪器是LC-20A高效液相色谱仪(Shimadzu,日本),并配有C18色谱柱(VP-ODS,250×4.6mm,5μm)和紫外可见检测器(Prominence SPD-20A),梯度系统(LC-20AT).分析采用的试剂如下:EGCG、ECG和EGC标准品为色谱纯(Sigma公司,美国),乙腈为色谱纯(国药集团化学试剂有限公司),水为Mili-Q超纯水(美国Millipore公司),其它试剂均为分析纯.采集完光谱后,将每个绿茶样本经浸泡、超声、离心、过0.45μm微孔滤膜等一系列预处理后,所得提取液用HPLC测定EGCG、ECG和EGC的含量,色谱条件参考ISO-14502-2006.HPLC测得的EGCG、ECG和EGC结果如表1所列.

1.4 净分析物预处理法

净分析物预处理法由Goicoechea等人[10]于2001年首先提出,该法基于净分析物信号(NAS)理论,主要用于提取混合物光谱中某一纯组分的光谱信息.其基本思想是:利用数学上空间正交的原理,将原始光谱矩阵中待测组分的净分析物信号提取出来[11,12].绿茶光谱的净分析物预处理(NAP)算法如下:设绿茶校正集的原始光谱矩阵为$X(I\times J)$,儿茶素含量的实测值向量为$y(I\times 1)$.在运用净分析物预处理法时将绿茶的近红外原始光谱矩阵分为两部分,其中一部分是与儿茶素含量相关的信息,而另一部分是与儿茶素含量不相关的所有干扰信息(包括来自绿茶内部以及来自环境的干扰信息)的综合,

表1 HPLC测定的绿茶样本中EGCG、ECG和EGC的含量(%)

Table 1 Contents(% w/w) of EGCG, ECG and EGC in green tea samples by HPLC

成分	单位(%)	训练集75个样本			预测集35个样本		
		范围	均值	标准方差	范围	均值	标准方差
EGCG	g/g	7.340~14.304	11.236	1.824	7.651~14.088	11.184	1.764
ECG	g/g	1.764~3.784	2.595	0.552	1.845~3.743	2.695	0.544
EGC	g/g	2.126~5.428	3.873	0.809	2.336~5.392	3.779	0.799

即

$$X = X_{SC} + X_{-SC} \quad , \qquad (1)$$

式中,X_{SC}表示绿茶光谱中与儿茶素含量相关的信息,X_{-SC}则表示光谱中儿茶素之外的所有其它干扰信息的综合.

寻求一个与 X_{-SC}正交的 $J \times J$阶矩阵 F_{NAP}(即 $X_{-SC} F_{NAP} = 0$),使式(1)两边同乘以 F_{NAP}后有 $XF_{NAP} = X_{SC} F_{NAP}$成立,这一步是该算法的关键步骤.矩阵 F_{NAP}的求解过程为:

1) 将原始光谱矩阵 X向儿茶素含量的实测值向量 y作正交投影得到 $X_{-SC} = [I - y(y^{T}y)^{-1}y^{T}]X$(式中 I为 $J \times J$阶单位矩阵);

2) 求出平方矩阵 $[(X_{-SC})^{T}X]_{-SC}$的特征向量矩阵 U(U为 $J \times A$阶矩阵,U中的每一列为一个净分析物预处理因子);

3) 构造矩阵 $F_{NAP} = I - UU^{T}$(式中 I为 $J \times J$阶单位矩阵).

这样即可求出经 A个净分析物预处理因子处理后的光谱 $X^{*}_{SC} = XF_{NAP} = X(I - UU^{T})$,式中 X^{*}_{SC}为经净分析物预处理法处理后得到的光谱矩阵,即儿茶素的净分析物信号矩阵.

预测集绿茶光谱 X_{UN}的净分析物预处理按式 $X^{*}_{UN,SC} = X_{UN}[I - UU^{T}]$进行,$X^{*}_{UN,SC}$为预测集绿茶光谱中儿茶素含量的净分析物信号矩阵.

1.5 数据处理及分析

所有的试验数据分析都是基于 MATLAB V7.0软件平台.数据处理首先按照校正集中的 75个样本光谱值及相应的参考值通过多变量校正的方法建立预测模型,然后通过预测集中的 35个样本来验证模型的可靠性.以校正样品的相关系数(R_c^2)和校正均方根偏差(RMSEC)以及预测样品的相关系数(R_p^2)和预测均方根偏差(RMSEP)来对模型进行综合评价.R_c^2和 R_p^2越高,RMSEC和 RMSEP越小,预测模型的性能就越好.

2 结果与讨论

2.1 光谱区域的选择

绿茶中儿茶素类物质的含氢基团(如 C-H、O-H、S-H和 N-H等)在近红外区域都能产生倍频与合频吸收,它们的一级倍频近红外光谱带位于 7200～5500cm^{-1}处;二、三、四级倍频位于 12800～8300cm^{-1}处;合频位于 5000～4000cm^{-1}处.图 1是绿茶样本的原始光谱图,绿茶的原始近红外光谱在 5155cm^{-1}和 6944cm^{-1}附近有一个明显的吸收峰.因为纯水中的 O-H伸缩振动的一级倍频位于 6944cm^{-1}附近,它的一个合频区位于 5155cm^{-1}附近,在这两个波长附近是水分吸收的敏感区[9].从图 1中可以看出在这两个区域,干茶中的水分对近红外光谱的吸收峰影响很大.为了减小水分的影响,分析时,选择光谱波长范围尽量避开水分吸收峰的特征波长区.本研究有比较地选用了各段的波长进行了分析,结果显示在一级倍频区选用 6500～5500cm^{-1}范围内的光谱数据既避开了水分的影响又取得了较好的试验结果.

2.2 模型的校正与优化

在采用偏最小二乘法建立模型前,校正集和预测集中样本的光谱将分别经过净分析物预处理.经过净分析物预处理后建立的偏最小二乘模型简称为 NAP/PLS模型,该模型所采纳的主成分因子数以及预处理过程中所采用的 NAP因子数对模型的预测性存在一定的影响.因此,在 NAP/PLS模型校正过

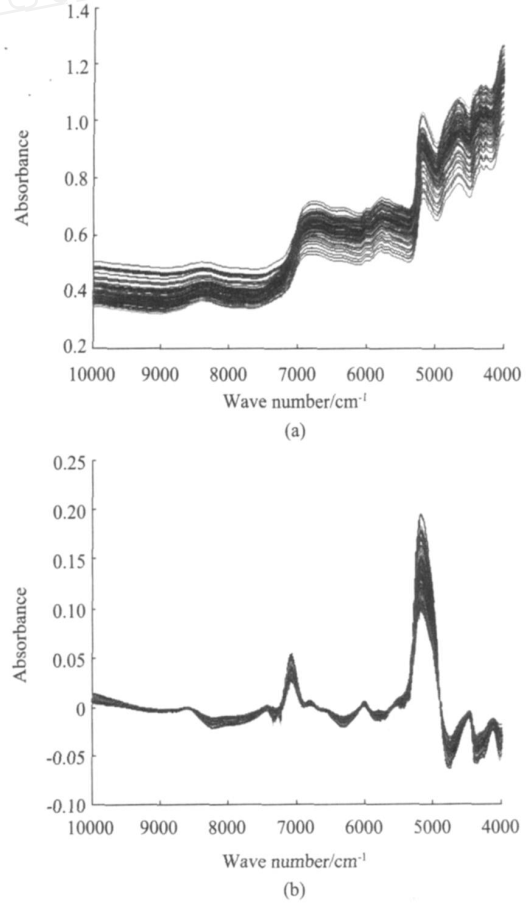

图 1 绿茶样本的(a)原始光谱(b)净分析物预处理后的光谱

Fig 1 Spectra of green tea obtained from (a) raw data and (b) NAP data

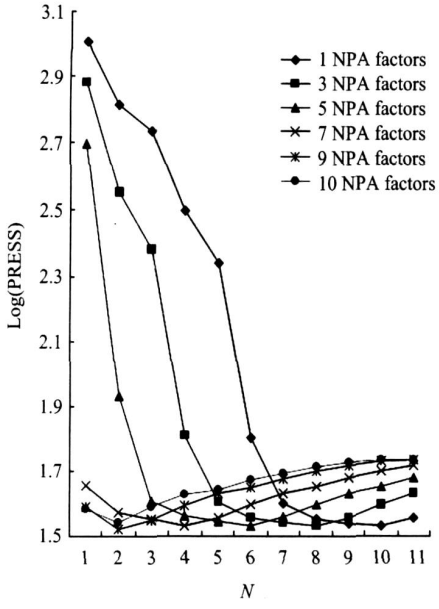

图 2 EGCG模型中 Log(PRESS)值与 PLS模型因子数 N 的关系图

Fig. 2 Diagram of Log (PRESS) value vs number of PLS factors N in EGCG model

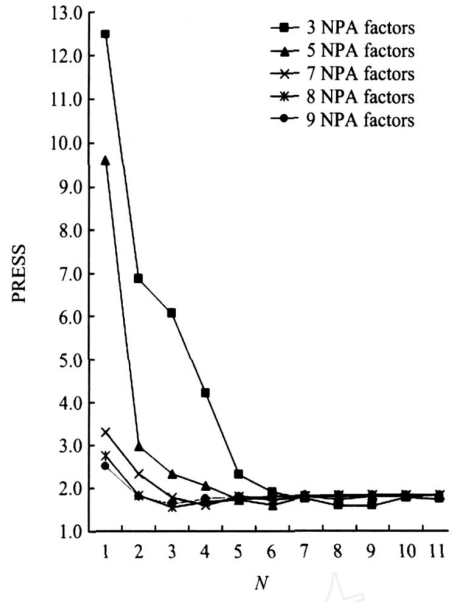

图 3 ECG 模型中 PRESS 值与 PLS 模型因子数 N 的关系图

Fig. 3 Diagram of PRESS value vs. number of PLS factors N in ECG model

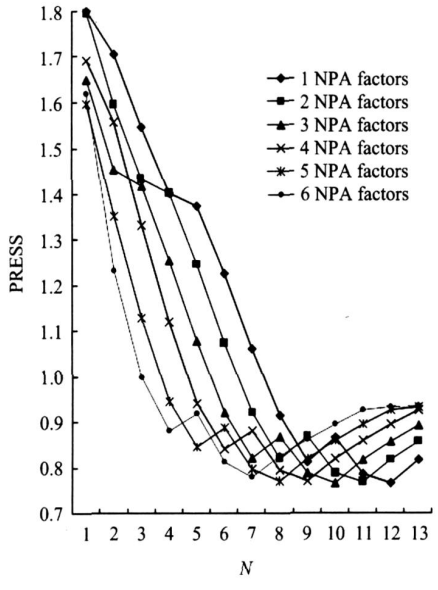

图 4 EGC 模型中 PRESS 值与 PLS 模型因子数 N 的关系图

Fig. 4 Diagram of PRESS value vs. number of PLS factors N in EGC model

程中有必要对模型的主成分因子数及 NAP因子数同时通过交互验证 (Cross-Validation)的方法来优化,即由最小的预测残差平方和 (PRESS)来确定.

图 2为绿茶经过不同的 NAP因子预处理后, EGCG模型所采纳的主成分因子数与对应的 Log (PRESS)值之间的关系.为了使图形看得清楚,图中只显示了采用 1、3、5、7、9和 10个 NAP因子的情形,从图中可以看出,随着所用 NAP因子个数的增加,最小的 Log(PRESS)值所对应的拐点逐渐向左移动,相应地模型的最佳因子数也逐渐减小.当 NPA因子增加到 9时, Log(PRESS)所对应的拐点值达到最低,此时模型的最佳主成分因子数等于 2,如果再继续增加 NAP因子数,模型的最佳主成分因子数仍为 2,但 Log(PRESS)所对应的拐点值反而略微上升.因此,优化后得到的结果是采用 9个 NAP因子和 2个主成分因子的 EGCG模型最佳.同理,图 3为绿茶经过不同的 NAP因子预处理后, ECG模型所采纳的主成分因子数与对应的 PRESS值之间的关系,从图中可以看出采用 8个 NAP因子和 3个主成分因子的 ECG模型可以得到最佳的优化结果.图 4为绿茶经过不同的 NAP因子预处理后, EGC模型所采纳的主成分因子数与对应的 PRESS值之间的关系,可以得到采用 5个 NAP因子和 8个主成分因子的 EGC模型结果最佳.

2.3 PLS校正结果比较与讨论

为了突出 NAP/PLS模型的优越性,将 NAP预处理后得到的结果与经典的 SNV光谱预处理后的结果相比较,比较的结果如表 2所示.从表 2可以看出,采用 SNV法对光谱进行预处理,偏最小二乘模型能较好地预测绿茶中 EGCG、ECG和 EGC的含量. 3个模型校正集的相关系数 (R_c^2)分别为 0.95405、0.96762和 0.96864,预测集的相关系数 (R_p^2)分别为 0.93258、0.94258和 0.94829.但模型

采纳的 PLS最佳因子数分别为 11、11和 13,模型明显过于复杂.当选择合适的 NAP因子进行 NAP光谱预处理后,EGCG、ECG和 EGCG三个模型采纳的 PLS最佳因子数分别为 2、3和 8;校正时,相关系数(R_c^2)分别为 0.95136、0.96723和 0.96773;预测时,相关系数(R_p^2)分别为 0.93036、0.93960和 0.94587.与 SNV预处理方法相比,模型经过 NAP光谱预处理后,模型的精确度几乎没有变化,但模型的复杂度(即模型采用的主成分因子数)却大大降低.总结图 2、3和图 4的结果可以得到,3个 NAP/PLS模型采纳的最佳因子数随着预处理过程中所用 NAP因子的增加而逐渐减少.特别是在最初阶段,NAP因子每增加 1个,NAP/PLS模型采纳的最佳因子数就减少 1,在这一过程中,每个 NAP/PLS模型的最佳因子数与模型所对应的 NAP因子数的和都保持不变.这说明绿茶光谱经 NAP法预处理后,模型中的部分主成分因子已转化为 NAP因子. NAP光谱预处理方法正是通过空间正交的途径最大程度地剔除了原始光谱中与儿茶素含量不相关的信息(包括来自绿茶内部的其它成分的信息以及来自外部环境的干扰信息).这样,NAP处理后的光谱中仅仅含有儿茶素含量的信息和少量干扰信息.因此,在原始光谱数据中,冗余信息和噪音信息的大量减少不但没有影响 PLS模型的精度,反而大大降低了模型的复杂度.

3　结论

净分析物预处理法是一种较新的光谱预处理算法,主要用于剔除光谱中与待测品质无关的信息.试验采用净分析物预处理法对绿茶的近红外光谱进行预处理,并利用偏最小二乘法分别建立绿茶中主要儿茶素 EGCG、ECG和 EGC的含量预测模型.结果表明,NAP光谱预处理算法在保证精度的前提下大大简化了儿茶素含量的预测模型.

表 2　通过 NAP和 SNV预处理后 PLS模型在校正集和预测集中的结果比较

Table 2　Comparison of results of PLS models in calibration and prediction sets by SNV and NAP spectral preprocessing

预处理方法	成分	PLS因子数	校正集		预测集	
			R_c^2	RMSEC	R_p^2	RMSEP
SNV预处理	EGCG	11	0.95405	0.39102	0.93258	0.45803
	ECG	11	0.96762	0.09932	0.94258	0.13027
	EGC	13	0.96864	0.14336	0.94829	0.18168
NAP预处理	EGCG	2	0.95136	0.40120	0.93036	0.46520
	ECG	3	0.96723	0.09986	0.93960	0.13428
	EGC	8	0.96773	0.14539	0.94587	0.18588

REFERENCES

[1] YUAN Xian-Qiang, WANG Yue-Fei, CHEN Liu-Ji. *Tea Polyphenol Chemistry* [M]. Shanghai: Shanghai Science and Technology Press (杨贤强,王岳飞,陈留记. 茶多酚化学. 上海:上海科学技术出版社), 2003: 412—422.

[2] Friedman M, Levin, C E, Choi S H, *et al*. HPLC analysis of catechins, theaflavins, and alkaloids in commercial teas and green tea dietary supplements: comparison of water and 80% ethanol/water extracts [J]. *Journal of Food Science*, 2006, **71** (6): 328—337.

[3] Kotani A, Takahashi K, Hakamata H, *et al*. Attomole catechins determination by capillary liquid chromatography with electrochemical detection [J]. *Analytical Sciences*, 2007, **23** (2): 157—163.

[4] Chen Q S, Zhao J W, Liu M H, *et al*. Detection of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms [J]. *Journal of Pharmaceutical and biomedical Analysis*, 2008, **46** (3): 568—573.

[5] Luypaert J, Zhang M H, Massart D L. Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, Camellia sinensis (L.) [J]. *Analytica Chimica Acta*, 2003, **478** (2): 303—312.

[6] Chen Q S, Zhao J W, Zhang H D, *et al*. Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration [J]. *Analytica Chimica Acta*, 2006, **572** (1): 77—84.

[7] WU Di, FENG Lei, ZHANG Chuan-Qing, *et al*. Early Detecti On Of Gray Mold (Cinerea) On Eggplant Leaves Based On ViS/Near Infrared Spectra [J]. *J. Infrared Millim. Waves* (吴迪,冯雷,张传清,等. 基于可见/近红外光谱技术的茄子叶片灰霉病早期检测研究. 红外与毫米波学报), 2007, **26** (4): 269—273.

[8] LIU Yan-De, LUO Ji, CHEN Xing-Miao. Analysis of soluble solid content in Nanfeng mandarin fruit with visible near infrared spectroscopy [J]. *J. Infrared Millim. Waves* (刘燕德,罗吉,陈兴苗. 可见/近红外光谱的南丰蜜桔可溶性固形物含量定量分析. 红外与毫米波学报), 2008, **27** (2): 119—122.

[9] CHU Xiao-Li, YUAN Hong-Fu, LU Wan-Zhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique [J]. *Progress in Chemistry* (褚小立,袁洪福,陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用. 化学进展), 2004, **16** (4): 528—542.

[10] Hector C, Goicoechea, Alejandro C O. A comparison of orthogonal signal correction and net analyte preprocessing methods. Theoretical and experimental study [J]. *Chemometrics and Intelligent Laboratory Systems*, 2001, **56** (2): 73—81.

[11] ZHAO Jie-Wen, ZHANG Hai-Dong, LIU Mu-Hua. Preprocessing methods of near-infrared spectra for simplifying prediction model of sugar content of apples [J]. *Acta Optica Sinica* (赵杰文,张海东,刘木华. 简化苹果糖度预测模型的近红外光谱预处理方法. 光学学报), 2006, **26** (1): 136—140.

[12] Lee Y, Chung Hoeil, Arnold M. A Improving the robustness of a partial least squares (PLS) model based on pure component selectivity analysis and range optimization: Case study for the analysis of an etching solution containing hydrogen peroxide [J]. *Analytica Chimica Acta*, 2006, **572** (1): 93—101.