# 特征变量筛选在近红外光谱测定绿茶汤中茶多酚的应用*

吴瑞梅[1,2]　岳鹏翔[3]　赵杰文[1]　黄星奕[1]　陈全胜[1]

(1. 江苏大学食品与生物工程学院，镇江 212013；2. 江西农业大学工学院，南昌 330045；
3. 大闽食品(漳州)有限公司，漳州 363005)

【摘要】 利用化学计量学方法从绿茶汤近红外光谱中提取茶多酚光谱信息，建立茶多酚近红外光谱定量分析模型。光谱采集使用 5 mm 光程的石英比色皿，利用联合区间偏最小二乘法(siPLS)筛选特征光谱区间，然后在筛选的光谱区间内进一步利用遗传算法(GA)优选特征变量。结果表明，siPLS 筛选的特征光谱区间避开了水的强吸收峰影响，利用 GA 在筛选的特征光谱区间内优选出 166 个特征变量建立 PLS 模型，模型预测集均方根误差为 0.685%，相对标准差为 5.26%，相对分析误差为 3.22，所建模型能达到精度要求，可用于实际检测。

**关键词**：茶多酚　近红外光谱　联合区间偏最小二乘法　遗传算法

**中图分类号**：O657.33　　**文献标识码**：A　　**文章编号**：1000-1298(2011)12-0154-04

# Application of Characteristic Variables Selection in Determination of Polyphenols Content in Green Tea Infusion by Near Infrared Spectroscopy

Wu Ruimei[1,2]　Yue Pengxiang[3]　Zhao Jiewen[1]　Huang Xingyi[1]　Chen Quansheng[1]

(1. *School of Food and Biological Engineering*, *Jiangsu University*, *Zhenjiang* 212013, *China*
2. *College of Engineering*, *Jiangxi Agricultural University*, *Nanchang* 330045, *China*
3. *Damin* (*Zhangzhou*) *Food Co.*, *Ltd.*, *Zhangzhou* 363005, *China*)

**Abstract**

The feasibility to extract the polyphenols spectra information from near infrared spectroscopy of green tea infusion using chemometrics methods was attempted. A cylindrical quartz glass tube with an optical path of 5 mm was used. First, synergy interval PLS (siPLS) was implemented to select efficient spectral regions from SNV preprocessed spectra. Then, optimal variables were selected using genetic algorithm (GA) from these selected spectral regions by siPLS. This study showed that the selected characteristic spectral intervals were not within the range of the strong absorbance for water. When 166 variables selected by GA and eight PLS factors were included, the optimal model (siPLS－GA) was achieved with $P_{RMSEP}=0.685\%$, $P_{RSD}=5.26\%$ and $P_{RPD}=3.22$ in the prediction set. The results showed that the performance of siPLS－GA model could be employed to measure polyphenols content in green tea infusion by near infrared spectroscopy in practice.

**Key words**　Polyphenols, Near infrared spectroscopy, Synergy interval PLS, Genetic algorithm

## 引言

茶多酚是决定茶汤浓度的主要滋味物质[1]，还具有降血脂、防辐射、防动脉粥样硬化及血管硬化、抗癌和抗突变等药理功效[2]。茶多酚常规检测方法主要有可见分光光度法[3]、高锰酸钾滴定法[4]等，这些检测方法属于化学方法，检测步骤烦琐、耗时长、费用高、测定结果可靠性差。近红外光谱技术

收稿日期：2011－03－01　修回日期：2011－04－06

* 国家自然科学基金资助项目(30971685)、江苏省自然科学基金资助项目(BK2009216)、福建漳州市蓝火计划资助项目(Z2010016)和江苏大学博士创新计划资助项目(CX10B_020X)

**作者简介**：吴瑞梅，博士生，江西农业大学副教授，主要从事农产品品质无损检测研究，E-mail：wuruimei036@163.com

**通讯作者**：赵杰文，教授，博士生导师，主要从事食品、农产品品质无损检测研究，E-mail：zhao_jiewen@ujs.edu.cn

具有检测速度快、检测成本低、重现性好等优点，已被广泛应用于液态样品品质的快速检测中[5~7]。

茶汤的近红外光谱谱峰重叠严重，并含有高强度的背景信息。茶汤内茶多酚含量很低，相对于水的强吸收，茶多酚的吸收非常微弱。姜礼义等[8]研究了不同光程对绿茶汤中茶多酚近红外光谱分析模型的影响，得出选用 1 mm 的光程可减少水的强吸收影响。但透射附件的光程太小，清洗困难，甚至会因清洗不干净产生严重的测定误差；另外透射附件光程太小，制作工艺要求及成本也高。在以往研究中，对液态样品的近红外分析，除通过选择合适的光程减少水的强吸收影响外，通过一些光谱变量筛选方法（如联合区间偏最小二乘法等）选择合适的光谱区，以避开水的强吸收干扰也有所应用[9]。本文利用联合区间偏最小二乘法（synergy interval PLS，简称 siPLS）结合遗传算法（genetic algorithm，简称 GA）从绿茶汤的近红外光谱中提取微弱茶多酚光谱信息，建立绿茶汤中茶多酚近红外光谱定量分析模型。

## 1　材料与方法

### 1.1　样本收集与茶汤制备

从茶叶市场收集不同等级的绿茶样 104 个，每个茶样 3 g，茶样原产地为国内重要绿茶产区：江西、江苏、河南、安徽、浙江等，将茶样编号后置于 4℃冰柜中保存。试验时，将每个茶样分别在室温下放置 12 h，达到室温平衡后放入审评杯中，用 150 mL 沸蒸馏水加盖冲泡 5 min，将茶汤倒出，用滤纸过滤，将滤液冷却到室温待用。

### 1.2　近红外光谱仪与茶汤光谱采集

茶汤光谱采集使用 Antaris Ⅱ傅里叶变换近红外光谱仪（Thermo Scientific，USA），光谱仪带有透射样品池附件。仪器工作参数：光谱波数范围 10 000 ~4 000 $cm^{-1}$，波数间隔为 3.856 $cm^{-1}$，扫描次数为 32 次。光谱采集时，将备用茶汤注入 5 mm 光程的石英比色皿中，每个样本采集光谱后旋转一定角度重复采集 3 次，求平均值作为原始光谱。

### 1.3　茶多酚含量的化学分析

对于每个茶汤样本，用吸管分别吸取 10 mL 茶汤备用液到 25 mL 容量瓶中，用蒸馏水稀释到刻度，以防止茶汤浓度过高。根据酒石酸亚铁比色法（GB/T 8313—2002）测量各稀释后的茶汤中的茶多酚含量。从原始样本中随机选出 74 个组成校正集，用来建立校正模型；余下 30 个组成预测集，用来预测模型性能。

### 1.4　数据分析及模型评价

模型建立过程中，采用联合区间偏最小二乘法（siPLS）筛选光谱特征区间，然后在筛选的特征区间上进一步利用遗传算法优选特征变量，利用交互验证法（cross-validation）优化模型参数。以校正集相关系数 $R_C$、校正集交互验证均方根误差 $C_{RMSECV}$ 作为模型校正效果的评价指标，以预测集均方根误差 $P_{RMSEP}$、相对标准差 $P_{RSD}$ 和相对分析误差 $P_{RPD}$ 评价模型的预测性能及预测精度。如果相对标准差大于 10% 或相对分析误差小于 2.5，说明该成分难以利用近红外光谱进行定量分析；如果相对标准差小于 10% 且相对分析误差在 2.5 ~3.0 之间，说明对该成分进行近红外光谱定量分析可行，但模型预测精度有待提高，只能用于实际估测；如果相对标准差小于 10% 且相对分析误差大于 3.0，则说明模型预测性能良好，预测精度高，可用于实际检测[10~12]。所有数据分析基于 Matlab V7.8.0 平台完成。

## 2　结果分析与讨论

### 2.1　茶汤近红外原始光谱及光谱预处理

图 1a 为茶汤近红外原始光谱图。图中波数 5 155 $cm^{-1}$附近和 6 900 ~7 140 $cm^{-1}$波段的强吸收峰，是茶汤中水的 O—H 伸缩振动的合频和一级倍频吸收[13]，若这些波数变量参与模型建立，会严重

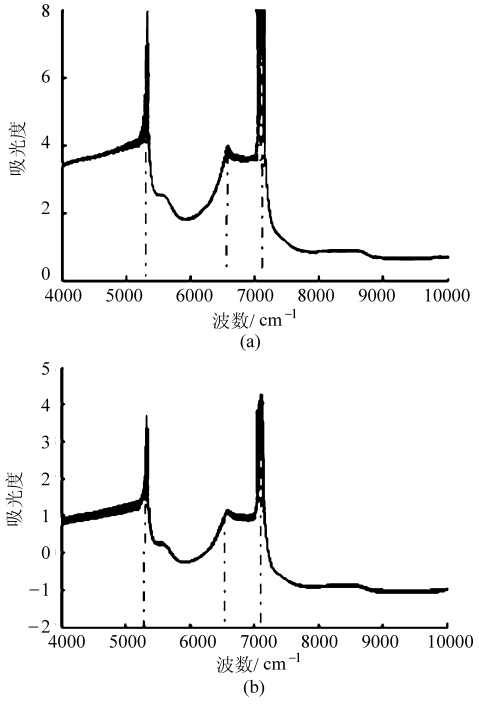


图 1　茶汤原始光谱图和标准正态变量变换预处理后光谱图

Fig. 1　NIR original spectra and SNV preprocessing spectra of tea infusion

（a）原始光谱图　（b）SNV 预处理后光谱图

影响模型的精度。本研究尝试使用特征变量筛选方法筛选出特征光谱区间,以去除水峰影响。为消除光谱在采集过程中受到仪器、环境等因素引起的噪声及其他无关信息的影响,采用标准正态变量变换(standard normal variate transformation,简称 SNV)对原始光谱进行预处理,预处理后的光谱图如图 1b 所示。

### 2.2 siPLS 筛选特征光谱区间及模型建立

联合区间偏最小二乘法是一种有效的波长筛选方法[14],该方法将整个光谱区等分成若干个相等的子区间,然后分别联合其中的 2、3 个或更多的子区间建立模型,并对各模型的预测误差进行比较,误差最小的模型所对应的联合子区间即是要筛选的特征光谱区间。为考察不同子区间数的划分及联合的子区间数对特征子区间优选的影响,并考虑到计算时间,试验将整个光谱区划分为 10 ~ 25 个子区间,在相同子区间划分的情况下,又分别联合其中的 2、3 和 4 个子区间建立模型。在筛选过程中,由交互验证法优化模型参数,由最小的交互验证均方根误差值作为衡量标准,筛选结果如表 1 所示。

**表 1 由 siPLS 划分不同间隔数的特征光谱区间筛选结果**

**Tab. 1 Results of feature spectral regions selected by siPLS**

| 区间数 | 主成分数 | 被选子区间 | $C_{RMSECV}$/% |
|---|---|---|---|
| 10 | 7 | 8、10 | 1.180 |
| 11 | 11 | 4、7、9 | 0.749 |
| 12 | 7 | 9、10、12 | 1.170 |
| 13 | 11 | 5、8、10、12 | 0.724 |
| 14 | 8 | 5、10、12 | 0.730 |
| 15 | 10 | 5、6、12、13 | 0.855 |
| 16 | 9 | 6、10、11、13 | 0.839 |
| 17 | 8 | 6、13、14 | 0.767 |
| 18 | 7 | 6、7、14、18 | 0.766 |
| 19 | 8 | 7、14、15、19 | 0.783 |
| 20 | 9 | 7、12、19 | 0.802 |
| 21 | 8 | 7、8、13、20 | 0.714 |
| 22 | 7 | 8、16、17、21 | 0.755 |
| 23 | 7 | 8、13、23 | 0.796 |
| 24 | 9 | 8、9、14、19 | 0.790 |
| 25 | 8 | 9、15、24 | 0.815 |

由表 1 看出,当整个光谱区被划分成 21 个子区间,使用 8 个主成分数,联合其中的 4 个子区间(分别为第 7、8、13 和 20 个)建立模型的 $C_{RMSECV}$ 最小。图 2 中的虚线带区是筛选出的 4 个特征子区间,各子区间对应的波数范围分别为 5 723.7 ~ 6 005.2 $cm^{-1}$、6 009.1 ~ 6 290.7 $cm^{-1}$、7 436.2 ~ 7 717.7 $cm^{-1}$和 9 434.1 ~ 9 715.6 $cm^{-1}$,共 296 个变量。从图 2 中看到,用 siPLS 优选出来的波数范围不在水的强吸收峰内,避免了水峰对待测组分的干扰。

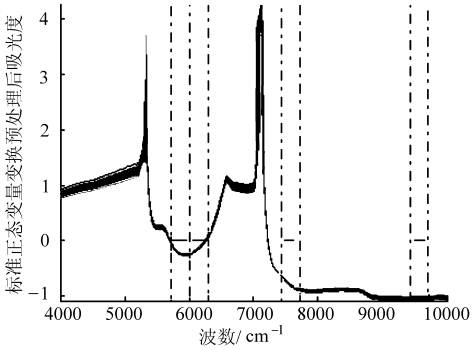


图 2 由 siPLS 筛选的特征光谱区间(7、8、13、20)

Fig. 2 Optimal spectral regions (7,8,13,20) selected by siPLS

由筛选的 296 个特征变量建立 PLS 模型,其校正集 $R_C$ 为 0.939,$C_{RMSECV}$ 为 0.714;用预测集样本检验模型精度,其 $P_{RMSEP}$ 为 0.771,$P_{RSD}$ 为 5.92%,$P_{RPD}$ 为 2.86,$P_{RSD}$ 小于 10%,但 $P_{RPD}$ 在 2.5 ~ 3.0 之间,表明用近红外光谱结合 siPLS 筛选特征变量定量分析茶汤中茶多酚含量是可行的,但模型预测精度有待进一步提高。

### 2.3 利用 GA 优选特征变量及 siPLS - GA 模型建立

由 siPLS 剔除了大量与茶多酚无关的信息,但在优选出的 4 个子区间内,相邻变量间可能存在高度相关性[15],即在筛选出的 296 个变量中仍存在大量冗余信息,这些冗余信息会降低模型精度。本文利用遗传算法[16]在优选的子区间 7、8、13、20 中进一步优选出与茶多酚光谱信息密切相关的特征变量建立模型,以提高模型精度。参数设置:初始群体大小为 50,交叉概率 $P_c$ 为 0.5,变异概率 $P_m$ 为 0.01,以 $C_{RMSECV}$ 构造适应度函数,遗传迭代次数为 100 次。迭代终止后,将被选频数较多的波数变量按被选频数高低逐一参与 PLS 模型建立,以最小的交互验证均方根误差值确定出最佳建模变量数。

由于遗传算法是一种随机搜索方法,其初始群体的选取及遗传操作算子的执行过程都带有较强的随机性,为防止随机性对结果的影响,本试验对该算法共运行了 6 次。表 2 为 6 次运算的结果,由表中

**表 2 6 次 GA - PLS 运行结果**

**Tab. 2 Six times results calculated by GA - PLS models**

| 模型序号 | 所选变量数 | 主成分数 | 校正集 | | 预测集 | |
|---|---|---|---|---|---|---|
| | | | $R_C$ | $C_{RMSECV}$/% | $R_P$ | $P_{RMSEP}$/% |
| 1 | 121 | 8 | 0.955 | 0.620 | 0.945 | 0.728 |
| 2 | 176 | 8 | 0.953 | 0.628 | 0.946 | 0.741 |
| 3 | 166 | 8 | 0.955 | 0.615 | 0.955 | 0.685 |
| 4 | 125 | 8 | 0.954 | 0.622 | 0.944 | 0.733 |
| 5 | 193 | 8 | 0.953 | 0.631 | 0.943 | 0.756 |
| 6 | 130 | 8 | 0.951 | 0.645 | 0.947 | 0.716 |

可看出，当所选变量为 166 个、主成分数为 8 时所建模型性能最好（即表中的第 3 号模型），所建 siPLS－GA 模型结果如表 3 所示。从表 3 可看出，siPLS－GA 模型所用的变量数减少，但精度得到提高，$P_{RMSEP}$为 0.685%，$P_{RSD}$为 5.26%，$P_{RPD}$为 3.22，大于 3.0，说明利用遗传算法进一步优选的 166 个变量，当主成分数为 8 时所建模型能达到精度要求，可用于实际检测。

**表 3　siPLS 模型与 siPLS－GA 模型结果**

**Tab.3　Results of siPLS and siPLS－GA model**

| 模型名称 | 变量数 | 主成分数 | 校正集 | | | 预测集 | | |
|---|---|---|---|---|---|---|---|---|
| | | | $R_C$ | $C_{RMSECV}$/% | $R_P$ | $P_{RMSEP}$/% | $P_{RSD}$/% | $P_{RPD}$ |
| siPLS | 296 | 8 | 0.939 | 0.714 | 0.940 | 0.771 | 5.92 | 2.86 |
| siPLS－GA | 166 | 8 | 0.955 | 0.615 | 0.955 | 0.685 | 5.26 | 3.22 |

## 3　结论

（1）利用 siPLS 方法筛选出与茶多酚光谱信息相关的特征光谱区间。结果显示所筛选的特征光谱区间避开了水峰影响；在筛选的光谱区间内建立的 PLS 模型可定量分析绿茶汤中茶多酚含量，但模型精度有待提高。

（2）利用 GA 在筛选出的特征光谱区间内进一步优选出与茶多酚信息最相关的特征变量建立 PLS 模型。结果表明，所建模型能达到定量分析的精度要求，可用于实际检测。与合适光程的选择方法相比，此方法简单、方便，建立的模型更简洁，甚至可避免利用小光程分析液体成分所产生的缺陷。

## 参　考　文　献


1　陆松侯，施兆鹏. 茶叶审评与检验[M]. 3 版. 北京：中国农业出版社，2001.

2　陈宗懋. 20 世纪茶与健康研究的主要进展[J]. 中国茶叶，2001，23(4)：8～10.
Chen Zongmao. Major progress in studies between tea and health in 20th century[J]. China Tea, 2001, 23(4): 8～10. (in Chinese)

3　GB/T 8313—2008　茶叶中茶多酚和儿茶素类含量的检测方法[S]. 2008.
GB/T 8313—2008　Determination of total polyphenols and catechins content in tea [S]. 2008. (in Chinese)

4　陈宗懋. 中国茶叶大辞典[M]. 北京：中国轻工业出版社，2000：407.

5　Leornardo S M, Flavia C C, Paulo A Z, et al. Determination of ethanol in fuel ethanol and beverages by Fourier transform (FT)－near infrared and FT－Raman spectrometries[J]. Analytica Chimica Acta, 2003, 493(2)：219～231.

6　Yu H Y, Ying Y B, Fu X P, et al. Quality determination of Chinese rice wine based on Fourier transform near infrared spectroscopy[J]. Journal of Near Infrared Spectroscopy, 2006, 14(1)：37～44.

7　Cozzolino D, Smyth H E, Gishen M. Feasibility study on the use of visible and near-infrared spectroscopy together with chemometrics to discriminate between commercial white wines of different varietal origins[J]. Journal of Agricultural and Food Chemistry, 2003, 51(26)：7 703～7 708.

8　姜礼义，刘福莉，陈华才，等. 绿茶汤中茶多酚近红外定量分析的光程选择[J]. 中国计量学院学报，2009，20(2)：135～138.
Jiang Liyi, Liu Fuli, Chen Huacai, et al. Selection of optical path length on nir analysis for tea polyphenols in green tea drinks[J]. Journal of China University of Metrology, 2009, 20(2)：135～138. (in Chinese)

9　朱向荣，李娜，史新元，等. 近红外光谱与组合的间隔偏最小二乘法测定清开灵四混液中总氮和栀子苷的含量[J]. 高等学校化学学报，2008，29(5)：906～911.
Zhu Xiangrong, Li Na, Shi Xinyuan, et al. Analysis of geniposide and total nitrogen conten in Qingkailing injection intermediate by near infrared spectroscopy and synergy interval partial least squares [J]. Chemical Journal of Chinese Universities, 2008, 29(5)：906～911. (in Chinese)

10　牛智有，韩鲁佳. 鱼粉中氨基酸近红外光谱定量分析[J]. 农业机械学报，2007，38(5)：114～117.
Niu Zhiyou, Han Lujia. Quantitative analysis of amino acid in fishmeal by near infrared reflectance spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery, 2007, 38(5)：114～117. (in Chinese)

11　Malley D F, Ronicke H, Findlay D L, et al. Feasibility of using near-infrared reflectance spectroscopy for the analysis of C, N, P and diatoms in lake sediments[J]. Journal of Paleolimnology, 1999, 21(3)：295～306.


（下转第 163 页）

置对称，分别与左、右槽壁发生热交换。TC1 与 TC2、TC4 温度一致是由于加热槽顶部有保温泡沫板，与 TC2、TC4 一样，只与槽壁发生热交换，故而它们温度一致。而 TC5 处于加热槽底角处，不仅与加热槽壁发生热交换，同时加热槽底面还吸热，因此 TC5 温度最低。

由图 8 可以看出，模拟预测的温度分布趋势与实验结果基本符合，TC5 点的预测结果与实验结果相差最大，为 2.2 K，其余 4 点温差较小，均小于 2.0 K。实验结果和模拟结果间的均方根误差为 1.86% <10% 在允许的误差范围内，因此用有限元法模拟豆浆在通电加热过程中的温度分布和变化情况是充分有效的。

## 4　结论

（1）用有限元法对通电加热过程中豆浆内部温度的分布和变化进行了模拟，在通电加热过程中其内部的温度分布比较均匀，但加热槽壁附近一直是其低温区，且随着加热时间的增加，低温区范围增加，豆浆内最大温差也会增加。

（2）模拟结果说明，当通电加热时间较短时，与加热槽对外部环境的散热相比，加热槽的吸热对豆浆的温度分布均匀性影响更大。

（3）对模拟结果进行了实验验证。在通电加热过程中，模拟结果和实验结果基本一致。

## 参　考　文　献


1　Berk Z. Technology of production of edible flours and protein products from soybeans [R]. FAO Agricultural Services Bullletin No. 97, Chapter 9, 1992.

2　周亚军，殷涌光，王淑杰，等. 食品欧姆加热技术的原理及研究进展[J]. 吉林大学学报：工学版，2004，34(2)：324～329.
Zhou Yajun, Yin Yongguang, Wang Shujie, et al. Mechanism and research progress of food ohmic heating technology [J]. Journal of Jilin University: Engineering and Technology Edition, 2004, 34(2): 324～329. (in Chinese)

3　Li Xingshu. Development of tofu processing system by the application of electrical impedance spectroscopy and ohmic heating [D]. Kobe: Kobe University, 2009.

4　長縄明大，三保茂之，秋山美展ら. ジュール加熱現象の可視化と有限要素解析[J]. 日本食品工学会誌，2006，7(2)：105～111.
Naganawa A, Miho S, Akiyama Y. Validation and FEM analysis of joule heating system [J]. Journal of Japanese Society of Food Science and Technology, 2006, 7(2): 105～111. (in Japanese)

5　Toyoda K, Li Xingshu. Monitoring of coagulation process of soymilk by an integrated electrical sensing and control system [M]. CIGR XVIIth World Congress, Québec City, Canada, 2010: 1～10.

6　Jaeger H, Meneses N, Knorr D. Impact of PEF treatment imhomogeneity such as electric field distribution, flow characteristic and temperature effects on the inactivation of E. coli and milk alkaline phosphatase [J]. Innovative Food Science and Emerging Technologies, 2009, 10(4): 470～480.

7　李修渠，李里特，辰巳英三. 豆浆的导电率[J]. 中国农业大学学报，1999，4(2)：103～106.
Li Xiuqu, Li Lite, Eizo Tatsumi. Study on electrical conductivity of soybean milk [J]. Journal of China Agricultural University, 1999, 4(2): 103～106. (in Chinese)

8　Singh P, Heldman D R. Introduction to food engingeering [M]. London: Academic Press, 2001: 10～128.

9　李法德，孙玉利，李陆星. 连续通电加热条件下豆浆的电导率[J]. 农业工程学报，2008，24(12)：275～278.
Li Fade, Sun Yuli, Li Luxing. Electrical conductivity of soybean milk during cooking with continuous ohmic heating device [J]. Transactions of the CSAE, 2008, 24(12): 275～278. (in Chinese)


(上接第 157 页)


12　Williams P C, Sobering D C. Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds [J]. Journal of Near Infrared Spectroscopy, 1993, 1(1): 25～32.

13　严衍禄，赵龙莲，韩东海，等. 近红外光谱分析基础与应用[M]. 北京：中国轻工业出版社，2005.

14　Norgaard L, Saudland A, Wagner J, et al. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy [J]. Applied Spectroscopy, 2000, 54(3): 413～419.

15　邹小波，朱曾，赵杰文. 基于间隔偏最小二乘法的农产品近红外光谱谱区选择方法[J]. 现代科学仪器，2007(1)：86～88.
Zou Xiaobo, Zhu Zeng, Zhao Jiewen. Selection of the efficient wavelength regions in agricultural product NIR spectroscopy based on interval partial least-squares (iPLS) [J]. Modern Scientific Instruments, 2007(1): 86～88. (in Chinese)

16　Leardi R. Application of genetic algorithm—PLS for feature selection in spectral data sets [J]. Journal of Chemometrics, 2000, 14(5～6): 643～655.