doi: 10.7690/bgzdh.2021.12.019

基于 SEV/SEC-iPLS 的近红外光谱波长优选方法

曾庆松，刘洋君，王菊香，邢志娜
（海军航空大学岸防兵学院，山东 烟台 264001）

摘要：为实现建模区间的波长优选，将 iPLS 法与验证标准差/校正标准偏差（standard error of verification/standard error of calibration，SEV/SEC）相结合，以喷气燃料为研究对象，对其初馏点进行定量分析。结果表明：iPLS 法结合 SEV/SEC 的组合评价指标用所选光谱区间建立的校正模型性能得到较大改善，模型的决定系数 R^2、SEV 分别为 0.94 和 1.05，较全谱区模型分别提升 55.70%和 36.23%，较传统 iPLS 法分别提升 44.44%和 17.50%，并通过配对 t 检验法验证，证明了该波长优选方法的可行性与有效性。

关键词：间隔区间偏最小二乘法；喷气燃料；初馏点；波长优选；t 检验
中图分类号：TJ01　　**文献标志码：**A

Wavelength Optimization Method for Near Infrared Spectrum Based on SEV/SEC-iPLS

Zeng Qingsong, Liu Yangjun, Wang Juxiang, Xing Zhina
(*College of Coast Defense*, *Naval Aviation University*, *Yantai* 264001, *China*)

Abstract: In order to optimize the wavelength of modeling interval, iPLS method and SEV/SEC (standard error of verification/standard error of calibration) were combined to quantitatively analyze the initial boiling point of jet fuel. The results show that the performance of the correction model established by iPLS method combined with SEV/SEC combination evaluation index using the selected spectral interval is greatly improved, the coefficient of determination (R^2) and standard error of verification (SEV) of the model are 0.94 and 1.05 respectively, which are 55.70% and 36.23% higher than the full spectral region model, and 44.44% and 17.50% higher than the traditional iPLS method, respectively. The feasibility and effectiveness of this method are verified by t test method.

Keywords: interval partial least squares method; jet fuel; initial boiling point; wavelength optimization; t test

0　引言

近红外光谱技术（near infrared spectroscopy，NIR）是光谱测量技术、化学计量学等学科的有机结合，因其分析高效快速、绿色环保、无需前处理等独特优势，已在食品、化工等领域得到了广泛的应用。近红外光谱谱区为 780～2 526 nm，全谱波长数量较多，存在信息强度弱、信噪比低、谱峰重叠严重等有效信息率低的问题。相关研究表明[1-3]：应用全谱数据进行建模不仅数据量庞大，而且存在多重共线信息甚至干扰信息，导致模型的预测能力未能达到最佳。在利用近红外技术进行检测分析时，对光谱进行波长优选，筛选出对某种评估标准最优的波长集合是很有必要的[4]，其主要目标是减少建模数据量，筛选出蕴含信息丰富的波长集合，进而提高光谱信噪比，优化模型的预测性能。

由于全谱区内存在与预测性质无关的干扰信息。利用全谱区建模时，这些干扰信息会影响所建立模型的精度和稳健性；因此，选择合适的区间对模型的质量有重要影响。为实现建模区间的优选，近年来在化学计量学领域基于 PLS 算法先后发展了 iPLS 法、移动窗口偏最小二乘法（moving window partial least squares，MWPLS）、组合区间偏最小二乘法（combined interval partial least squares，SiPLS）等波长区间优选方法。

其中，iPLS 法具有简便快捷、计算量小的优点，在实际应用中也取得了良好效果。冼瑞仪等[5]将 iPLS 法应用于橄榄油中掺杂煎炸老油的定量分析，效果与 BiPLS 法和 SiPLS 法相近；宋宁等[6]将正交信号校正算法（orthogonal signal correction，OSC）与 iPLS 相结合，消除了粉末状样品散射光的影响，提高了烃源岩红外光谱预测模型的准确性。由于 iPLS 法选取区间的评价指标是交互验证标准偏差（standard error of cross-validation，SECV）最小，所选区间具有唯一性，因此 iPLS 法是一种“单变量”

收稿日期：2021-08-15；**修回日期：**2021-09-20
作者简介：曾庆松（1996—），男，福建人，硕士，从事兵器工程专业研究。E-mail: 1823742621@qq.com。

方法，没有考虑光谱区域之间的协同作用，L.Muncka 等[7]通过实验验证了多区间建模效果优于单个区间。仅以某个参数作为评选指标过于片面，代表性不强，并不能保证挑选出最优波长区间。针对该问题，笔者对 iPLS 法的区间评价指标进行了改进，结合 SEV 与 SEC 作为新评价指标，并以喷气燃料为研究对象，以其初馏点为例进行实验分析。

1 实验部分

1.1 基础数据测定及光谱采集

实验共采集 35 个喷气燃料样本，初馏点的基础数据按照《GB/T 6536—2010 石油产品常压蒸馏特性测定法》的方法测得。经测量得到 35 个样本的初馏点情况如表 1 所示。

表 1 样本初馏点

样本数量	最大值	最小值	平均值
35	167.9	151.2	159.0

光谱采集系统波长检测范围为 1 150～2 500 nm，波长的分辨率为 1 nm；采用光程为 5 mm 的石英比色皿测量样本吸光度，对每个样本扫描 3 次，积分时间为 30 ms，保证每次扫描波长的重复性不超过 5×10^{-4}，最终取平均光谱作为样本原始光谱。光谱采集时用恒温箱对样本进行保温，温度设置为 25 ℃，分辨率为 0.1 ℃。按上述操作及实验条件扫描得到 35 个喷气燃料样本的光谱数据。

1.2 样本集划分

Kennard/Stone(K/S)方法是一种应用广泛的样本选取方法，通过计算样本之间的欧式距离，可选出代表性强、分布范围均匀的样品作为校正集样品进行建模[8]。笔者采用 K/S 算法以 6:1 的比例选取了 30 个样本作为校正集，5 个样本作为验证集，样本集划分具体情况如表 2 所示。从表 2 可见，校正集的样本覆盖范围大于验证集，满足建模样本的要求。

表 2 K/S 划分样本集

	校正集	验证集
样本编号	2，3，4，5，6，7，8，9，10，11，14，15，16，17，18，19，20，22，23，24，25，26，27，28，29，30，32，33，34，35	1，12，13，21，31
最大值/℃	167.9	166.8
最小值/℃	151.2	156.3
平均值/℃	158.6	161.1

1.3 方法原理

iPLS 法是 L.Norgaard 等[9]提出的一种波长区间选择方法，其原理是首先将全谱区以固定宽度划分成若干等宽子区间，在每个子区间上分别进行 PLS 回归，建立局部校正模型。采用 SECV 作为衡量指标来评价各个局部校正模型的精度。SECV 越小，说明该区间的光谱数据与性质值的相关性越强，最终选取 SECV 最小的局部模型所对应的子区间作为优选区间[10]。除了 SECV 外，模型的评价指标还包括： SEC、SEV 和 R^2 等，每个指标都表示了模型某方面的性能。仅以 SECV 作为指标挑选出的波长区间未必能最大限度的包含有用信息；因此，iPLS 方法在局部模型评价参数的选择方面需进行改进，选出信息含量最丰富的波长区间。

SEV 通常要大于 SEC，但二者差距不能过大。SEC 过大时，模型对校正集样本的回归性较差，不宜采用；SEV 远大于 SEC 时，说明验证集样本的选择不具有代表性[11-13]。结合以上分析，为了对光谱进行更有效的波长筛选，选取近年由相关机构提出的分析标准 SEV/SEC 作为新评价指标。该指标可将被选择模型的关键参数 SEC 和 SEV 保持在合格的范围内。

由于 SEV/SEC 是在食品领域多年的近红外分析实践中总结得到的经验参数，在喷气燃料近红外光谱分析中是否适用还有待实践证实。针对该问题，笔者采用 SEV/SEC 作为区间优选的评价参数改进 iPLS 方法，对比不同区间选择方法的效果，验证该评价参数的可行性以及建模区间选择方法的优劣。改进局部模型评价指标的 iPLS 方法具体流程如图 1 所示。

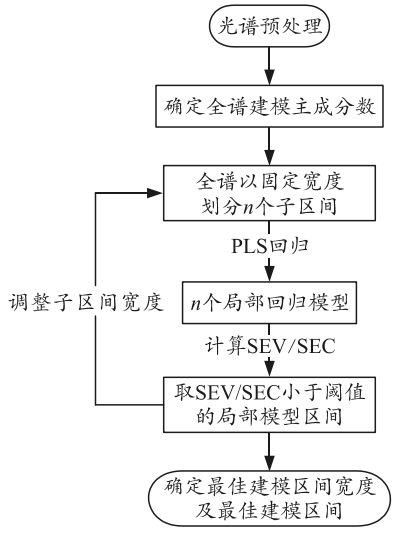

图 1 改进的 iPLS 波长区间筛选流程

最终即可选择出与性质值关联最大的波长区间，利用区间内的光谱即可建立高精度的校正模型。

2　实验结果与分析

2.1　光谱预处理

Savitzky-Golay(S-G)卷积平滑法是目前应用较为广泛的预处理方法，可有效消除光谱中的噪声，提高信噪比；S-G 卷积平滑也可用于求取导数光谱，可有效消除基线和其他背景干扰，分辨重叠峰，提高分辨率和灵敏度[14-15]。笔者采用 S-G 卷积平滑和 S-G 卷积导数的方法对原始光谱进行预处理，在提高光谱信噪比的同时消除光谱在吸光度方向上的基线平移。原始光谱和 S-G 处理之后的光谱分别如图 2、3 所示。

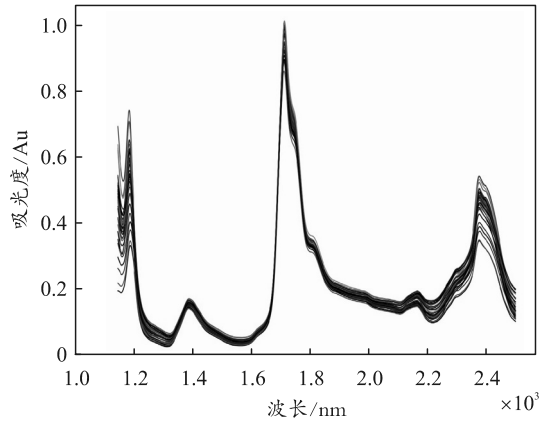

图 2　样本原始光谱

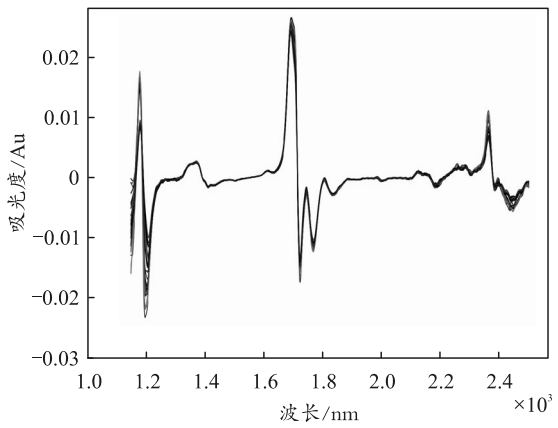

图 3　Savitzky-Golay 平滑后光谱

2.2　iPLS 窗口宽度确定

窗口宽度作为 iPLS 的首要参数，对波长筛选起着关键作用。将 iPLS 结合 SEV/SEC 所选取的区间作为优选的建模区间，合并优选区间用 PLS 算法建立的初馏点校正模型。通过改变子区间宽度，对比不同宽度下校正模型的预测能力，选择预测精度相对较高的校正模型所对应的子区间宽度。实验中窗口宽度变化区间为 50～75，步长取 1，不同子区间宽度下所建立的校正模型预测性能如图 4 所示。

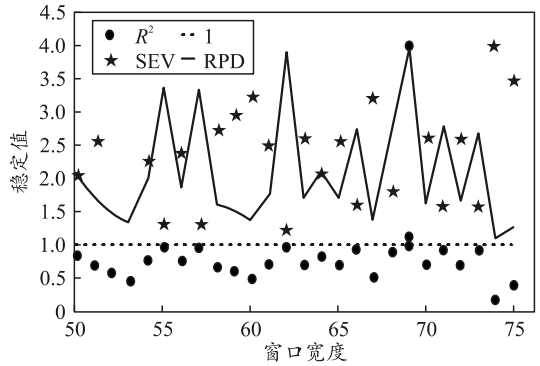

图 4　不同区间宽度下模型评价指标值

由上图可以看出：当子区间宽度为 69 个波长点时，SEV 达到最小，R^2 达到最大，此时校正模型的预测性能最佳，说明 69 为窗口宽度的最优值。

2.3　SEV/SEC 阈值确定

光谱区间的选取是以阈值为依据，需要先确定最佳的阈值，才能选出最佳的波长区间。参考食品领域的经验参数 SEV/SEC≤1.2，实验的阈值变化区间为 0.8～1.8，步长为 0.1，选取小于阈值的光谱区间作为最终的建模区间。为确定 SEV/SEC 的最佳阈值，利用不同阈值下改进 iPLS 方法所选波长区间建立校正模型。计算出不同阈值下校正模型的 SEV 和 R^2 并绘出曲线图，结果如图 5 所示。

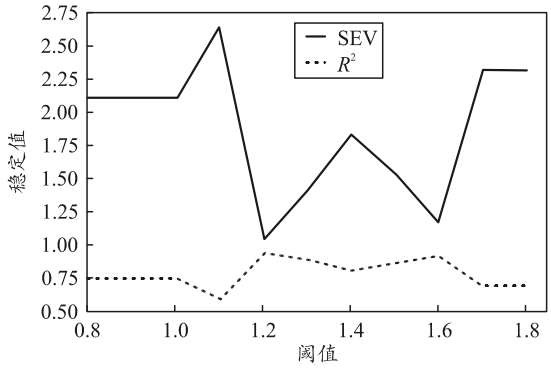

图 5　不同阈值下改进 iPLS 区间选择效果

由上图可见：随着阈值的增大，校正模型的预测性能并非单调变化，阈值在 1.0 之前和 1.7 之后校正模型性能变化不大；在区间 1.0～1.7 内预测精度先下降然后上升，再以平缓的趋势下降；当阈值取 1.2 时模型的 R^2 和 SEV 均达到全局最优，模型预测效果好，说明改进 iPLS 方法的评价指标 SEV/SEC 的阈值取 1.2 时所选出的波长区间最优。此时该方法选择的波长点及其对应的光谱如图 6 所示，选择的波长区间分别为 1 307～1 376 nm，1 652～1 721 nm 和 1 997～2 135 nm，波长点总个数为 276。

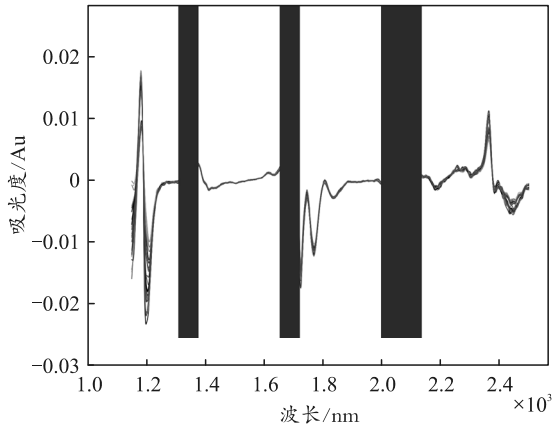

图 6 改进局部模型评价指标后的 iPLS 法所选区间

2.4 所选光谱区间建模结果分析

主成分数作为 PLS 模型的首要参数，对模型的预测性能影响很大。在进行 PLS 回归建模时，若主成分数选取太小将会丢失原始光谱较多的有用信息，造成欠拟合；太大则会将测量噪声过多地包括进来，造成过拟合。实验中采取“留一法”计算预测残差平方和 PRESS 来确定最佳的主成分数，绘出 PRESS 值与主成分数目的曲线图如图 7 所示，选取曲线图最低点对应的主成分数作为最佳主成分数。

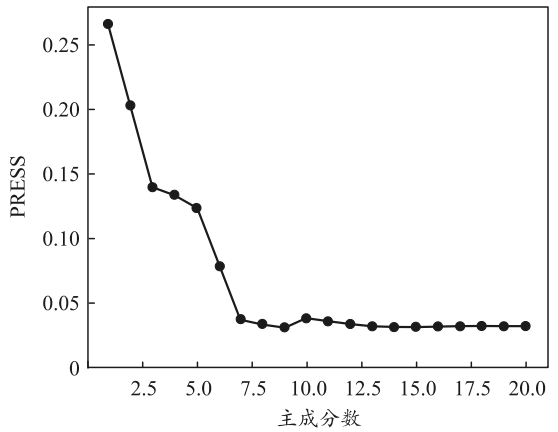

图 7 模型 PRESS 与主成分数关系

由上图可见，当主成分数为 9 时 PRESS 值最小，即模型的最佳主成分数为 9。

采用由改进 iPLS 方法选择的波长区间进行 PLS 建模，然后对验证集进行预测，结果如表 3、图 8 所示。

表 3 校正集样本初馏点 PLS 模型预测结果

验正集样本	实测值	预测值	残差
1	161.6	163.0	1.4
12	166.8	165.6	1.2
13	163.1	163.9	0.8
21	157.6	158.3	0.7
31	156.3	156.6	0.3

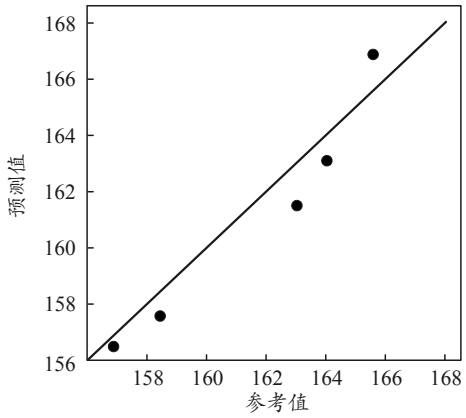

图 8 验证集样本的预测值与参考值的关系

如上图可见，利用改进的 iPLS 方法选出的波长区间所建模型的预测效果较好，预测值与参考值之间的误差较小。

对预测结果进行配对 t 检验，得统计量 t=0.21，低于显著性水平 α=0.05，自由度为 5 时的临界值 2.571，说明模型预测的准确度较高，预测结果可信度较高。

2.5 不同波长选择方法建模结果的比较

采用不同方法选择的波长进行建模，模型的预测性能如表 4 所示。

表 4 不同区间选择方法所建模型评价参数

波长选择方法	选择波长区间/nm	SEC	SEV	R^2
全谱区	1 100～2 500	2.94	2.37	0.69
iPLS	1 320～1 375	1.77	1.89	0.80
改进 iPLS	1 307～1 376, 1 652～1 721, 1 997～2 135	1.01	1.05	0.94

由上表可知：经过波长优选后建立的校正模型预测性能均优于采用全谱建立的校正模型，改进后的 iPLS 方法对校正模型性能的提升效果最好。与全谱模型相比，iPLS 方法模型的 SEC、SEV 降幅分别为 39.80%、20.25%，R^2 的增幅为 15.94%；改进 iPLS 方法模型的 SEC、SEV 降幅分别为 65.65%、55.70%，R^2 的增幅为 36.23%。全谱区所建立的校正模型精度不及改进的 iPLS 法，原因是全谱区建模引入了许多与预测性质无关的干扰信息，影响了所建立的校正模型预测的准确性；而传统 iPLS 法只选择一个子区间进行建模，在波长选择过程中舍弃了部分相关信息，导致模型精度稍差。另外，iPLS 和改进 iPLS 选取的波长数为 55 和 276，分别为全谱的 3.93%，19.71%；说明波长选择还可大幅减少建模的数据量，进而提升建模的效率，降低模型的复杂度。

3　结论

针对传统 iPLS 方法存在的不足，笔者对其进行改进，并应用于喷气燃料。实验结果表明：采用 SEV/SEC 作为新的衡量指标进行建模区间的优选，保证 iPLS 法优选波长区间所使用判据的合理性，在增加参与建模有效信息量的同时减少了建模数据量；采用改进的 iPLS 方法选出的波长区间建立的校正模型预测性能得到显著提升，证明新评价指标 SEV/SEC 在喷气燃料近红外光谱分析中具有可行性。

参考文献：

[1] 褚小立. 化学计量学方法与分子光谱分析技术[M]. 北京: 化学工业出版社, 2011: 79.

[2] BALABIN R M, SMIRNOV S V. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data[J]. Analytica Chimica Acta, 2011, 692: 63−72.

[3] DI W, CHEN X, ZHU X, et al. Uninformative variable elimination for improvement of successive projections algorithm on spectral multivariable selection with different calibration algorithms for the rapid and non-destructive determination of protein content in dried laver[J]. Analytical Methods, 2011, 3(8): 1790−1796.

[4] 邵学广, 蔡文生, 徐恒. 近红外光谱分析中波长选择的必要性[C]//中国化学会第 27 届学术年会第 15 分会场摘要集. 厦门: 中国化学会第 27 届学术年会, 2010.

[5] 冼瑞仪, 黄富荣, 黎远鹏, 等. 可见和近红外透射光谱结合区间偏最小二乘法(iPLS)用于橄榄油中掺杂煎炸老油的定量分析[J]. 光谱学与光谱分析, 2016, 36(8): 2462−2467.

[6] 宋宁, 徐晓轩, 王斌, 等. 基于正交信号校正和 iPLS 的烃源岩红外漫反射光谱定量分析[J]. 光谱学与光谱分析, 2009, 29(2): 378−381.

[7] MUNCKA L, NIELSENA J P, MOLLERA B, et al. Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics[J]. Analytica Chimica Acta, 2011, 446: 171−186.

[8] 詹雪艳, 赵娜, 林兆洲, 等. 校正集选择方法对于积雪草总苷中积雪草苷 NIR 定量模型的影响[J]. 光谱学与光谱分析, 2014, 34(12): 3267−3272.

[9] NORGAARD L, SAUDLAND A, WANGER 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(4): 413−419.

[10] 黄富荣, 李仕萍, 余健辉, 等. 基于 iPLS 的血清胆固醇、甘油三酯近红外定量分析[J]. 光谱实验室, 2011, 28(6): 2774−2778.

[11] 刘荣欣, 胡萍. 偏最小二乘法回归模型在分析毛涤混纺面料纤维含量中的应用[J]. 河南工程学院学报(自然科学版), 2019, 31(1): 8−12.

[12] 王赋腾, 孙晓荣, 刘翠玲, 等. 光谱预处理对便携式近红外光谱仪快速检测小麦粉灰分含量的影响[J]. 食品工业科技, 2017(10): 58−61.

[13] 李文采, 刘飞, 田寒友, 等. 基于高光谱成像技术的鸡肉菌落总数快速无损检测[J]. 肉类研究, 2017, 31(3): 35−39.

[14] 褚小立, 袁洪福, 陆婉珍, 等. 近红外分析中光谱预处理及波长选择方法进展与应用[J]. 化学进展, 2004, 16(4): 528−541.

[15] 赵环环, 严衍禄. 噪声对近红外光谱分析的影响及相应的数学处理方法[J]. 光谱学与光谱分析, 2006, 26(5): 842−845.