文章编号：1672-8785(2024)10-0038-07

基于 iPLS-KPCA 的高温燃气红外光谱特征提取方法研究

席剑辉 许壮壮

（沈阳航空航天大学自动化学院，辽宁 沈阳 110136）

摘 要：高温燃气红外光谱特征是判断燃气成分和浓度的有效途径。针对高温燃气红外辐射特性复杂、建模难度高的问题，研究了一种基于间隔偏最小二乘(interval Partial Least Squares，iPLS)和核主成分分析(Kernel Principal Component Analysis，KPCA)的特征提取算法。首先通过 iPLS 进行预筛选，确定具有最优预测能力的特征光谱波段，避免单个子区间建模过程中有用吸收峰信息的遗失；其次，利用 KPCA 降低数据维度，保留贡献率高的关键特征，降低成分预测模型的复杂度。仿真结果表明，经过 iPLS-KPCA 方法特征提取后，预测模型的复杂度大幅下降，且预测能力显著提升。

关键词：高温燃气；间隔偏最小二乘；核主成分分析；特征提取

中图分类号：TN219 **文献标志码**：A **DOI**：10.3969/j.issn.1672-8785.2024.10.006

Research on Infrared Spectral Feature Extraction Method for High Temperature Gas Based on iPLS-KPCA

XI Jian-hui，XU Zhuang-zhuang

(*College of Automation，Shenyang Aerospace University，Shenyang 110136，China*)

Abstract：The infrared spectrum characteristic of high temperature gas is an effective way to judge the composition and concentration of gas. Aiming at the problems of complex infrared radiation characteristics and high modeling difficulty of high-temperature gas，a feature extraction algorithm based on interval partial least squares (iPLS) and kernel principal component analysis (KPCA) is studied. Firstly，the characteristic spectral bands with the best prediction ability are determined by pre-screening with iPLS to avoid the loss of useful absorption peak information in the process of single subinterval modeling. Secondly，KPCA is used to reduce the data dimension，retain the key features with high contribution rate，and reduce the complexity of the component prediction model. The simulation results show that after feature extraction by iPLS-KPCA method，the complexity of the prediction model is greatly reduced，and the prediction ability is significantly improved.

Key words：high-temperature combustion gas；interval partial least squares；kernel principal component analysis；feature extraction

收稿日期：2024-03-28

基金项目：辽宁省自然科学基金项目(2015020069；2015020061)；沈阳市科技创新团队项目(src201204)

作者简介：席剑辉(1975-)，女，辽宁沈阳人，教授，博士，主要研究方向为复杂系统模型辨识、故障检测与诊断、红外辐射测试与分析等。E-mail：xihui_01@163.com

0 引言

高温燃气成分和浓度检测在工业和国防领域具有重要意义。例如，航空发动机尾焰高温燃气的浓度分布情况可以直接反映发动机的燃油燃烧效率，进而可以了解发动机的推力水平等参数[1]。此外，高温燃气尾焰光谱特征信息还是告警系统探测目标的重要依据。在高温燃气组成成分中，对燃气光谱的红外辐射特性分析主要集中在 H_2O、CO_2 气体[2-3]。因此，通过分析 CO_2 或 H_2O 气体的浓度即可推断出发动机当时的某种状态，从而为发动机的故障诊断、状态监测提供数据支持。

红外光谱在测量方面具有灵敏度高、稳定性强的优势[4]，在气体的定性和定量识别方面应用广泛，尤其在高温情况下不受采样条件限制[5]。通过目标的红外辐射光谱特性与细节特征，可以更好地进行识别。陶冶等人[6]利用高温燃气喷焰红外光谱特征分析，成功实现了火箭飞行高度辨识；刘尊洋[7]和舒锐[8]等人依据高温燃气尾焰红外光谱的双峰辐射特性确定了卫星的探测波段；苑智玮[9]等人采用改进的向前和向后 iPLS 法建立了特征波段提取模型，实现了目标与特征光谱数据库的匹配。

气体光谱测量过程不仅会受仪器本身精度的影响，还会受到环境等因素干扰。同时，高温燃气红外辐射增强，光谱变化更加复杂，需要对光谱数据进行预选，保留有效信息丰富、数据干扰少的样本。Norgaard L 等人[10]提出通过用 iPLS 去除冗余信息来简化模型。但该算法采用单一区间建模，可能会丢失有用特征信息。本文沿用 iPLS 思路并对其进行改进，通过分组建模获取多个区间的光谱信息，然后结合 KPCA[11]对气体进行红外光谱特征提取。首先使用 iPLS 进行预筛选，确定具有最优预测能力的特征光谱波段，接着利用 KPCA 方法进行进一步降维，在减少数据冗余信息的同时避免了单一区间建模带来的信息缺失。

1 燃气光谱偏最小二乘模型

偏最小二乘（Partial Least Squares，PLS）[12-13]利用大量已知浓度的光谱建立定量校正模型，通过校正模型来预测未知光谱中组分的浓度信息。

气体光谱 PLS 定量校正模型[14]的建立步骤如下：

（1）对浓度矩阵 $Y_{m\times l}$ 与吸光度矩阵 $Y_{m\times n}$ 进行分解：

$$X_{m\times n} = T_{m\times d}P_{d\times n} + E_{m\times n} \tag{1}$$

$$Y_{m\times l} = U_{m\times d}Q_{d\times l} + F_{m\times l} \tag{2}$$

式中，吸光度矩阵 $X_{m\times n}$ 表示每个光谱样本中含有 n 个波数点，浓度矩阵 $Y_{m\times l}$ 表示包含 l 个吸光组分、m 个光谱校正样本，T 和 U 称为矩阵 X 与矩阵 Y 的得分矩阵，P 和 Q 分别为 X 和 Y 矩阵的载荷矩阵，E 和 F 分别为 X 和 Y 矩阵的残差矩阵。

（2）建立得分矩阵 T 与 U 之间的回归模型：

$$U_{m\times d} = T_{m\times d}B_{d\times d} \tag{3}$$

式中，B 称作回归系数矩阵，代表了矩阵 X 与矩阵 Y 之间得分矩阵的关系。B 的估计矩阵为

$$\hat{B}_{d\times d} = (T^T_{(m\times d)}T^{(-1)}_{(m\times d)})T^T_{(m\times d)}U_{(m\times d)} \tag{4}$$

（3）建立矩阵 X 与自矩阵 Y 之间的校正模型：

$$Y_{m\times l} = X_{m\times n}P^{-1}_{d\times n}\hat{B}_{d\times d}W_{d\times l} \tag{5}$$

当预测样本 m' 的浓度未知时，可以利用吸光度矩阵 $X'_{m\times n}$ 和回归系数矩阵 $\hat{B}_{d\times d}$ 来对浓度矩阵 $Y'_{m\times l}$ 进行预测。通过已知浓度校正数据集，建立了 PLS 回归模型。

模型的验证包括两种方式：内部验证利用参与建模的观测值来评价模型的稳定性，其评价指标为交叉验证均方根误差（Root Mean Square Error of Cross Validation，RMSECV）[15]；外部验证利用验证集来检验模型预测值和真实值之间的误差大小，其评价指标为预测集的预测值均方根（Root Mean Square of Prediction，RMSEP）。此外，决定系数 R^2 也是判断模型性能和稳定性的重要指标[16]。相关评

价指标的公式如下：

$$RMSECV=\sqrt{\frac{\sum_{i=1}^{n}(c_i-\hat{c}_i)^2}{n}} \tag{6}$$

$$RMSEP=\sqrt{\frac{\sum_{i=1}^{p}(c_i-\hat{c}_i)^2}{p}} \tag{7}$$

$$R^2=1-\frac{\sum_{i=1}^{p}(c_i-\hat{c}_i)^2}{\sum_{i=1}^{p}(c_i-\bar{c}_i)^2} \tag{8}$$

式中，n 为校正集数，p 为验证集数，c_i 和 $\hat{c}_i$ 分别表示第 i 个样本的真实值和预测值，$\bar{c}_i$ 表示样本集中目标真值的平均值。当 RMSECV 及 RMSEP 值越小，R^2 越接近于 1 时，模型的预测能力越强，稳定性越好。

2 燃气光谱 iPLS-KPCA 模型

2.1 基于 iPLS 的样本优化

iPLS 的主要思想就是进行分组，并且建立各组的 PLS 模型。通过比较局部光谱与全光谱 PLS 模型 RMSECV 的大小，将局部 PLS 模型中 RMSECV 值大于全光谱的部分去掉，并保留剩余波段来组成一个新的特征光谱。具体步骤如下：

(1)建立全谱 PLS 校正模型(见式(1)～式(5))，并且计算其 RMSECV 值。

(2)将全光谱分成 n 个等宽的子区间。对于全谱波长点不满足等分的情况，当整除完剩余波长不超过子区间长度的一半时，将其并入最后一个子区间，否则作为单独的区间[17]。

(3)各个子区间分别建立 PLS 模型，并计算相应的 RMSECV 值。

(4)比较全谱 PLS 模型和局部 PLS 模型的 RMSECV 值，并选取局部 RMSECV 值小于全谱的波段来组合成新的光谱。

2.2 基于 iPLS-KPCA 的成分预测模型

由于高温燃气呈现局部线性、整体非线性的特点，即光谱数据包含许多互相重叠的谱带和峰，因此通过引入 KPCA 映射将输入空间中的数据点转换到一个特征空间中，使得在特征空间中的数据点能够更容易地被处理和分析。

基本原理如下：样本 X 是由 M 个样本(红外光谱组别)、N 维(特征波长数量)向量组成的矩阵($x_i\in R^{M\times N}$ $(i=1,2,\cdots,M)$)，假设映射 $\Phi(x_i)$ 已经满足中心化要求，即

$$\sum_{i=1}^{M}\Phi(x_i)=0 \tag{9}$$

那么在特征空间中，协方差矩阵可表示为

$$C=\frac{1}{M}\sum_{i=1}^{M}\Phi(x_i)\Phi(x_i)^T \tag{10}$$

协方差矩阵 C 的特征向量 V_i 和特征值 λ_i 的求解方程为

$$CV_i=\lambda_iV_i \tag{11}$$

则特征向量可以表达为

$$V_i=\sum_{i=1}^{M}\alpha_i\Phi(x_i) \tag{12}$$

式中，α_i 为线性系数矩阵。联立式(10)、式(11)、式(12)可得

$$\frac{1}{M}\sum_{i=1}^{M}\Phi(x_i)\Big(\sum_{i=1}^{M}\alpha_{jl}\Phi(x_i)^T\Phi(x_l)\Big)\Phi(x_i)^T=\lambda_j\sum_{l=1}^{M}\alpha_{jl}\Phi(x_l) \tag{13}$$

定义一个 $M\times M$ 的 K 矩阵：

$$K_{ij}=K(x_i,x_j)=(\Phi(x_i)^T,\Phi(x_j)) \tag{14}$$

则式(13)就变为

$$\frac{1}{M}\sum_{i=1}^{M}\Phi(x_i)\Big[\sum_{i=1}^{M}\alpha_{jl}K(x_i,x_l)\Big]\Phi(x_i)^T=\lambda_j\sum_{l=1}^{M}\alpha_{jl}\Phi(x_l) \tag{15}$$

在式(15)两侧同时左乘 $\Phi(x_k)^T$ 可得：

$$\frac{1}{M}\sum_{i=1}^{M}\Phi(x_k)^T\Phi(x_i)\Big[\sum_{i=1}^{M}\alpha_{jl}\Phi(x_i)^T\Phi(x_i)\Big]\Phi(x_i)^T=\lambda_j\sum_{l=1}^{M}\alpha_{jl}\Phi(x_k)^T\Phi(x_l) \tag{16}$$

由式(14)和式(16)可得：

$$K\alpha_j=M\lambda_j\alpha_j \tag{17}$$

由式(17)可知，$M\lambda$ 是由核函数求得的矩阵 K 的特征值，特征向量为

$$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_M)^T \tag{18}$$

对特征向量 V 作归一化处理需满足：

$$V^T\cdot V=1 \tag{19}$$

联立式(17)、式(19)可得：

$$\lambda_j M\alpha_j^T\alpha_j = 1 \tag{20}$$

原始空间中 x 的点映射到主成分形成的新空间的值为

$$t_k = \Phi(x)^T V = \sum_{i=1}^{M}\alpha_{ji}K(x,x_i) \tag{21}$$

当式(9)不成立时，需要进行核矩阵 K 中心化操作，即

$$\tilde{\Phi}(x_k)=\Phi(x_i)-\frac{1}{M}\sum_{i=1}^{M}\Phi(x_k)x_i,\ i=1,2,\cdots,M \tag{22}$$

那么

$$\begin{aligned}\tilde{K}(x_i,x_j) &= [\tilde{\Phi}(x_i)^T\cdot\tilde{\Phi}(x_j)]\\ &= K(x_i,x_j)-\frac{1}{M}\sum_{k=1}^{M}K(x_i,x_k)\\ &\quad-\frac{1}{M}\sum_{k=1}^{M}K(x_k,x_j)\\ &\quad+\frac{1}{M^2}\sum_{l,k=1}^{M}K(x_l,x_k)\end{aligned} \tag{23}$$

因此

$$\tilde{K}_{ij} = K - I_M K - K I_M + I_M K I_M \tag{24}$$

式中，I_M 为 $M\times M$ 的矩阵，每一个元素都为 $1/M$。则特征向量为

$$\tilde{K}\alpha_i = \lambda_i\alpha_i \tag{25}$$

原数据映射到高维空间中的投影为

$$y_j = \sum_{i=1}^{M}\alpha_{ji}K(x,x_i),\ j=1,\cdots,d \tag{26}$$

对所得特征向量进行单位化处理，并利用特征值累计贡献率选取主元成分，满足：

$$\frac{\sum_{j=1}^{N}\lambda_j}{\sum_{i=1}^{M}\lambda_i}\geqslant F \tag{27}$$

式中，$\sum_{j=1}^{N}\lambda_j$ 代表前 N 个特征值的累计量，F 为阈值。为确保所提取的成分信息能保留足够多的特征信息，前 N 个特征值累计量与 M 个特征值累计量的关系应满足式(27)。

iPLS-KPCA 步骤如下：

(1)输入经 iPLS 选取的新的特征光谱数据，并构建标准核矩阵。

(2)根据式(17)计算核矩阵 K，并利用式(23)对核矩阵进行中心化处理得到 $\tilde{K}$。

(3)根据式(25)计算特征空间中的特征向量 a_i 和特征值 λ_i。

(4)根据式(26)计算原数据映射到高维空间上的投影 y_i。

(5)根据式(27)特征值累积贡献率选取光谱主成分 N。

3 仿真实例

测量高温燃气红外光谱数据 80 组，波长范围为 2～5 μm，全波段共有 3749 个波长数据点。

在进行特征波长选取以及定量回归建模之前，使用 Kennard-Stone 方法[18]以 3∶1 的比例将数据集进行划分。其中，校正集有 60 组，验证集有 20 组。

3.1 iPLS 特征波长选取

根据前面的 iPLS 算法流程，首先建立全光谱 PLS 校正模型，结果如图 1 所示。其中，虚线表示全光谱 PLS 建模的 RMSECV 值 1.5124。然后将全光谱分成 n 个等宽的子区间。对于 iPLS 模型，如果区间数太小，就可能退化为全谱 PLS 模型；如果区间数太多，则会增加计算量。因此，本文用 iPLS 方法将全光谱等分为 20 个子波段。实验数据全谱共 3749 个波数点。在将其分为 20 个子区间的过程中，前 19 个子波段中波长变量为 187 个，最后一个子波段中波长变量为 196 个。对各波段分别建立 PLS 校正模型，柱状条表示每个子区间的 RMSECV 大小。由图 1 中的柱状条可以看出，子区间 1、4、5、6、15、16、17 的 PLS 模型的 RMSECV 值小于全光谱 RMSECV 值。因此，将经 iPLS 方法选出的 7 个子区间共 1309 个特征波数点重新组成新的光谱。新光谱波数点占全谱波长的 35%。然后对该光谱进行 PLS 建模(结果见表 1)。

表 1 不同特征波长选取方法的实验结果

	波长数量	*RMSECV*	R_c	*RMSEP*	R_p
全光谱	3749	1.5124	0.9026	1.6512	0.9010
iPLS	1309	0.8925	0.9445	0.9943	0.9340
iPLS-KPCA	12	0.5342	0.9478	0.5211	0.9512

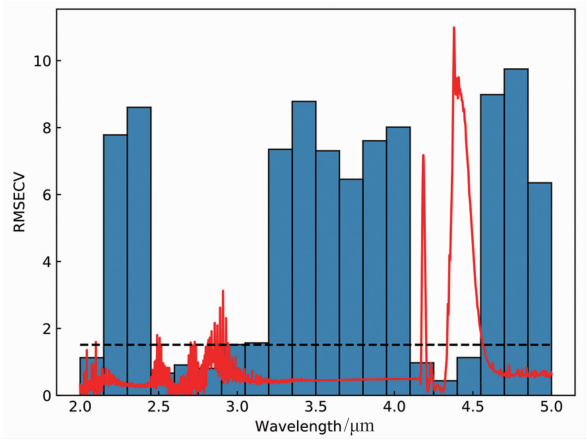

图 1 CO_2 光谱 iPLS 建模结果

3.2 iPLS-KPCA 特征波长选取

经过 iPLS 特征波长选取后，对新光谱数据样本再次进行核主成分特征提取。KPCA 的核函数选用高斯核函数，主元成分贡献率阈值选择 95%。

表 2 列出了利用 KPCA 算法获取的主元特征的贡献率计算结果。可以看出，气体的前 12 个主元的累积贡献率达到 95.2%，超过了 95%。

表 2 KPCA 各主元贡献率及累积贡献率

特征值序号	λ_i 值	贡献率/%	累计贡献率/%
1	5.326	75.57	75.57
2	0.372	5.27	80.84
3	0.203	2.88	83.72
4	0.165	2.34	86.06
5	0.127	1.8	87.86
6	0.108	1.53	89.39
7	0.094	1.33	90.72
8	0.085	1.2	91.92
9	0.072	1.02	92.94
10	0.066	0.93	93.87
11	0.051	0.72	94.59
12	0.043	0.61	95.2

利用提取的前 12 个主元特征数据进行 PLS 建模，所得实验结果如表 1 所示。表 1 中包含了全光谱波长以及经过不同特征提取算法处理后的红外光谱 PLS 建模的情况。可以看出，经过特征提取后模型的预测精度较全光谱波长下均有不同程度的提高。通过 iPLS 选出 7 个特征波段后，模型校正集的 RMSECV 由 1.5124 降至 0.8925，校正集相关系数 R_c 从 0.9026 变为 0.9445，预测集的 RMSEP 由 1.6512 降至 0.9943，相关系数 R_p 从 0.9010 提升至 0.9340；经过 iPLS-KPCA 特征提取后，特征波长数量由 3749 减至 12，校正集的 RMSECV 降至 0.5342，R_c 变为 0.9478，预测集 RMSEP 降至 0.5211，R_p 变为 0.9512，模型的复杂度大幅下降。

3.3 不同特征提取方法的模型预测结果

将经过不同特征波长选取后的红外光谱数据建立 PLS 校正模型，并对预测集进行评估。

图 2 所示为利用不同特征提取方法建立 PLS 模型的预测浓度结果。从中可以直观地看到预测值与实际值之间的相关性。表 3 列出了具体的评价指标结果。与全光谱 PLS 模型的预测结果相比，iPLS 模型和 iPLS-KPCA 模型的预测误差明显减小。特别在 iPLS-KPCA 模型中，预测浓度与实际浓度之间的相关系数为 0.9512，与全光谱模型相比预测能力显著提升，证明了本文特征提取方法的有效性。

表 3 不同浓度预测模型的评价指标结果

	平均相对误差	相关系数
全光谱	0.155	0.9010
iPLC	0.109	0.9340
iPLS-KPCA	0.069	0.9512

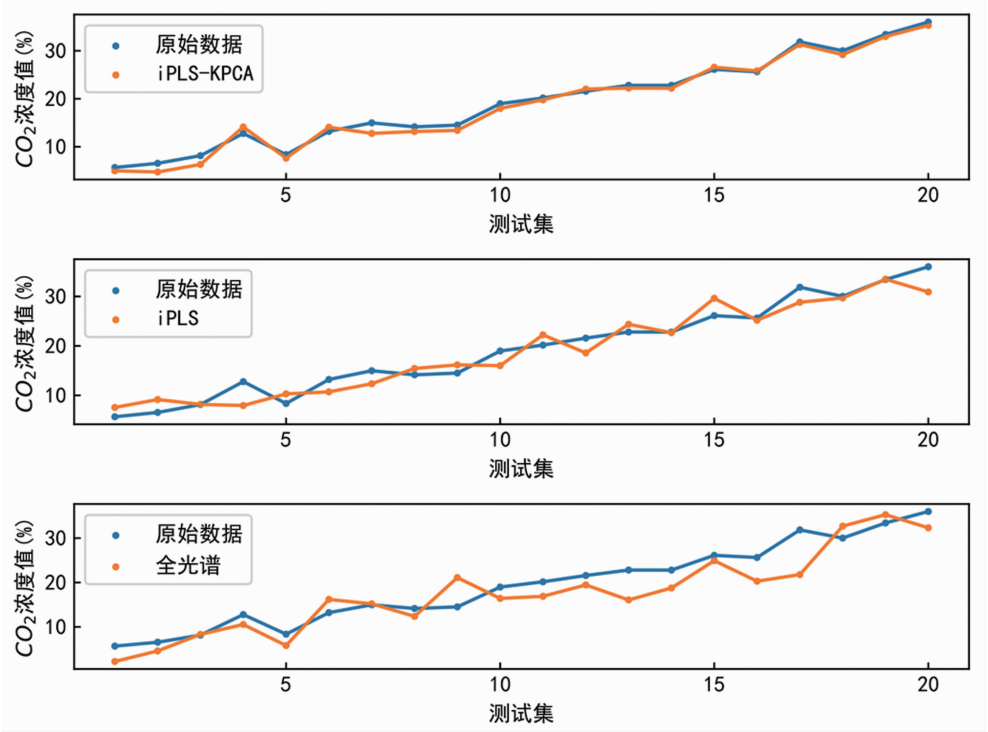

图 2 不同特征提取方法的 CO_2 预测精度

4 结束语

本文提出了一种基于 iPLS-KPCA 的高温燃气红外光谱特征提取算法。该算法将 iPLS 预选能力和 KPCA 降维能力相结合，利用特征提取降低了预测模型的计算复杂度，并通过仿真实验证明了该方法的有效性。实验结果表明，经过特征提取算法处理后，特征波长数量减少到全谱数量的 3.2%，PLS 校正模型中预测均方根误差 RMSEP 由 1.6512 减至 0.5211（减少 68.4%），预测相关系数 R_p 提升 4.46%，预测能力显著提高。未来将考虑进一步扩充数据集，从而验证该特征提取方法的有效性。

参考文献

[1] 韩畅，曾文，刘靖，等. 燃烧室燃烧与排放特性试验与数值计算 [J]. **航空发动机**，2023，**49**(6)：35-41.

[2] 孟夏莹，王彪，张玉涛，等. 一种基于尾焰特征光谱的发动机辨识方法：CN202211140952.4 [P]. 2024-03-04.

[3] 张宇辰，张海龙，姬文娟，等. 一种大气中 CO_2 精细吸收光谱特性的研究方法 [J]. **科学技术创新**，2021，**26**(33)：70-72.

[4] 贺书文，冯灿，王加熙，等. 基于光谱吸收技术的飞机座舱 CO_2 浓度测量研究 [C]. 西安：中国航空工业技术装备工程协会年会，2023.

[5] 沈英，邵昆明，吴靖，等. 气体光学检测技术及其应用研究进展 [J]. **光电工程**，2020，**47**(4)：3-18.

[6] 陶冶. 典型发动机喷焰多尺度光谱特征提取与分类研究 [D]. 哈尔滨：哈尔滨工业大学，2022.

[7] 刘尊洋，邵立，汪亚夫，等. 基于辐射通量表观对比度光谱的红外预警卫星探测波段选择方法 [J]. **红外与毫米波学报**，2014，**33**(5)：492-497.

[8] 舒锐，周彦平，卢春莲. 基于多光谱辐射特性差异的最佳探测波段的确定方法 [J]. **红外与激光工程**，2014，**43**(8)：2505-2512.

[9] 苑智玮，黄树彩，熊志刚，等. 尾焰特征光谱在主动段弹道目标识别中的应用 [J]. **光学学报**，2017，**37**(2)：306-313.

[10] Norgaard L, Saudland A, Wagner J. Interval partial least squares regression (iPLS): A comparative chemical study with an example from near in-

frared spectroscopy [J]. *Applied Spectroscopy*, 2000, **54** (3): 413-419.

[11] 石新发, 贺石中, 谢小鹏, 等. 基于核主成分的船舶柴油机磨损信息特征提取方法研究 [J]. **武汉理工大学学报(交通科学与工程版)**, 2022, **46**(6): 1039-1043.

[12] Lindberg W, Persson J A, Wold S. Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and lignin sulfonate [J]. *Analytical Chemistry*, 1983, **55**(4): 643-648.

[13] 张琳, 张黎明, 李燕, 等. 偏最小二乘法在傅里叶变换红外光谱中的应用及进展 [J]. **光谱学与光谱分析**, 2005, **25**(10): 76-79.

[14] 鞠薇. 环境污染气体的FTIR光谱特征提取及定性识别方法研究 [D]. 合肥: 合肥工业大学, 2019.

[15] 冯艳春, 张琪, 胡昌勤. 药品近红外光谱通用性定量模型评价参数的选择 [J]. **光谱学与光谱分析**, 2016, **36**(8): 2447-2454.

[16] 胥雪炎, 李补喜. 不同被解释变量选择对决定系数 R^2 的影响研究 [J]. **太原科技大学学报**, 2007, **28**(5): 363-365.

[17] 张优优, 陈伟豪, 唐志敏, 等. 区间偏最小二乘结合差分进化算法应用于鱼粉近红外光谱波长筛选 [J]. **分析测试学报**, 2020, **39**(11): 1392-1397.

[18] 李华, 王菊香, 邢志娜, 等. 改进的K/S算法对近红外光谱模型传递影响的研究 [J]. **光谱学与光谱分析**, 2011, **31**(2): 362-365.