基于局部对称重加权惩罚最小二乘的拉曼基线校正

赵恒[1*]，陈娱欣[2]，续小丁[1]，胡波[1]

[1]西安电子科技大学生命科学技术学院，陕西 西安 710126；

[2]西安电子科技大学通信工程学院，陕西 西安 710071

摘要 拉曼光谱分析技术具有快速响应、非接触、检测限制小、灵敏度高的优点，广泛应用于生产生活的众多领域。然而实际测得的原始拉曼光谱总会有不同程度的基线漂移，严重影响光谱分析的有效性和准确性。针对现有基线校正方法容易造成估计基线偏低、校正后光谱抬升的问题，提出了一种基于局部对称重加权惩罚最小二乘(LSRPLS)的基线校正算法，该算法在非对称惩罚最小二乘的基础上，使用 softsign 函数引入局部对称加权的思想，对光谱中无谱峰的基线区域赋予相近的权重，并通过迭代调整估计基线的权重。在模拟和实际拉曼光谱上分别进行了验证。实验结果表明：LSRPLS 基线校正算法不仅能对不同类型的光谱基线进行校正，而且与现有的基线校正方法相比，具有更高的准确度和稳定性。基线校正后的光谱在主成分空间上的聚集度得到提升，模型的分类准确性明显提高，说明 LSRPLS 算法在去除基线的同时，能够保留光谱的有效信息，为拉曼光谱的进一步分析提供了依据。

关键词 光谱学；基线校正；惩罚最小二乘；拉曼光谱；softsign 函数

中图分类号 O433.4 **文献标识码** A **doi**：10.3788/CJL201845.1211001

Baseline Correction for Raman Spectra Based on Locally Symmetric Reweighted Penalized Least Squares

Zhao Heng[1*], Chen Yuxin[2], Xu Xiaoding[1], Hu Bo[1]

[1]*School of Life Science and Technology, Xidian University, Xi′an, Shaanxi* 710126, *China*;

[2]*School of Telecommunications Engineering, Xidian University, Xi′an, Shaanxi* 710071, *China*

Abstract Raman spectroscopy has advantages of rapid response, non-contact, less detection restrictions and high selectivity, which make it widely used in many fields of production and life. However, the actual measured Raman spectra contain varying degrees of baseline drift, which seriously affects the validity and accuracy of spectral analysis. In order to solve the issues that the final baseline is underestimated in the no peak region and the height of peaks might be overestimated in existing baseline correction methods, we propose a novel correction algorithm, which is named locally symmetric reweighted penalized least squares (LSRPLS). Based on asymmetrical least squares, the method works by iteratively adjusting weights of the difference between the fitted baseline and the original signal, introducing the idea of local symmetric weighting by a softsign function. The algorithm is applied to the simulated and the actual Raman spectra to correct the baseline drifting. The results show that the LSRPLS algorithm can not only correct different types of baselines, but also has good advantages in accuracy and stability compared with the existing baseline correction methods. In addition, after baseline correction, the distribution of samples in principal component spaces becomes concentrated, and the classification accuracy of the model is significantly improved. This indicates that the LSRPLS algorithm can retain the spectral information effectively while removing the baseline, which provides a basis for further analysis of Raman spectroscopy.

Key words spectroscopy; baseline correction; penalized least squares; Raman spectrum; softsign function

OCIS codes 300.2530; 300.6450; 200.4560

收稿日期：2018-06-07；**修回日期**：2018-08-02；**录用日期**：2018-08-10

基金项目：国家重点研发计划(2016YFC0102004)

* **E-mail**：hengzhao@mail.xidian.edu.cn

1 引　言

拉曼光谱基于物质分子的散射而产生，可以反映分子的振动信息与转动信息，因而具有对物质的“指纹”识别的特性[1]。除了具有快速无损、低成本等其他光谱检测方法的优点外，拉曼光谱还具有检测范围广、检测方式灵活、适用于水溶液测量等独特优势[2]。近年来，拉曼光谱已被广泛应用于生物医学、食品安全、环境保护、材料分析等领域[3-6]。然而，由于实验条件、样品状态以及仪器测量稳定性等原因，实际测得的拉曼光谱会不可避免地受到噪声和荧光背景等干扰信号的影响。荧光光谱的存在往往导致拉曼光谱变得模糊，甚至淹没重要的特征峰信息，使后续的定性、定量分析结果出现严重偏差。为了保证分析模型的准确性，需要使用预处理技术有效去除光谱中的噪声和荧光背景干扰，以提高光谱的信噪比，强调样品的光谱特征。

基线校正是拉曼光谱预处理中的关键步骤。现有的基线校正方法主要有求导、滤波、多项式拟合、惩罚最小二乘和形态学方法等。通常，一阶微分能够去除光谱中的常量偏移，二阶微分能够去除常量偏移和线性偏移[7]。但求导方法容易导致光谱的峰形失真，并且求导后的光谱难于解释。Tatarkovič等[8]利用快速傅里叶变换滤波对人血浆样本的拉曼光谱进行处理，结合荧光猝灭和光漂白技术，减少了约90%的荧光背景。Hu等[9]提出一种基于复小波变换的光谱基线校正法，并采用该方法解决了实小波变换信息丢失的问题。值得注意的是：频域滤波可能会引起光谱畸变，造成谱峰位置和强度信息的损失；小波重构信号可能会出现无意义的负值，且结果依赖于分解函数和分解级数的选取。传统的人工多项式拟合法需要人为在光谱上选取一系列属于背景信号的极小值点来拟合基线，操作繁琐费时，且准确度主要取决于操作者的经验。Lieber等[10]通过迭代的多项式拟合方法(ModPoly)实现了基线校正的半自动化，避免了人工方法的主观性干扰。Cao等[11]提出自适应极小极大基线拟合方法(AdaptMinmax)，根据估计基线与光谱的比值自适应确定拟合多项式的阶数。Wang等[12]在ModPoly的基础上提出一种基于迭代多项式平滑的基线校正方法(IPSA)，该方法的准确性较高，计算速度较快。

基于惩罚最小二乘的基线校正综合考虑了拟合基线对真实基线的保真度以及拟合基线自身的平滑度[13]，常被用于各种光谱的预处理，并发展出许多改进算法。Eilers等[14]使用非对称加权惩罚最小二乘法(AsLS)对多种光谱进行基线校正，通过优化平滑参数和非对称参数实现了拟合基线的灵活调节。鉴于AsLS的代价函数仅考虑了基线二阶导数的平滑性，He等[15]对其增添一阶导数平滑约束，得到了改进的非对称最小二乘(IAsLS)，提高了基线估计的准确性。针对AsLS的估计基线易受较大峰值影响的问题，Oller-Moreno等[16]提出一种峰值信号的非对称最小二乘算法，该算法通过一个衰减指数来调控非对称参数 p 的大小，有效减小了算法在大峰值区域上的基线估计误差。Zhang等[17]提出自适应迭代重加权惩罚最小二乘算法(airPLS)，该算法利用一个指数函数对拟合基线与光谱之间差值的权重进行迭代更新，加速了算法的收敛速度。Park等[18]提出了一种非对称重加权惩罚最小二乘法(arPLS)，并计算了拟合基线在可用参数范围内的保真度和平滑度，实现了最优参数的自动选择。Li等[19]将形态学方法与惩罚最小二乘相结合，提出形态学加权惩罚最小二乘(MPLS)算法，该算法将开运算得到的局部极小值作为惩罚最小二乘的输入，进一步优化了估计基线。

在获取拉曼光谱的过程中，常伴随着各种噪声干扰，如CCD探测器散粒噪声、暗电流噪声、激光的发射噪声等。虽然可以通过平滑滤波等方法去除部分噪声，但是过度平滑也会损失光谱信息，因此在拉曼光谱基线校正中应考虑噪声的影响。基于此，本课题组提出了一种基于局部对称重加权惩罚最小二乘(LSRPLS)的基线校正算法——LSRPLS基线校正算法，该算法通过迭代改变softsign权重函数，实现了对拉曼光谱中无谱峰基线区域的局部对称加权和有谱峰信号区域的非对称加权，并通过对模拟和实际拉曼光谱的处理分析，验证了该算法的有效性。

2 LSRPLS基线校正算法

非对称加权惩罚最小二乘基线校正[14]是在惩罚最小二乘(PLS)的基础上，对保真度引入一个非对称的权重 $\boldsymbol{w}=[w_1,w_2,\cdots,w_N]$，通过求解惩罚最小二乘函数，最终获得基线的有效估计。设原始信号为 $\boldsymbol{y}$，其长度为 N。用一平滑序列 $\boldsymbol{z}$ 表示待估计的基线向量。则 $\boldsymbol{z}$ 对于 $\boldsymbol{y}$ 的保真度可以用它们之间的残差平方和来表征，即

$$F=\sum_{i=1}^{N}(y_i-z_i)^2。\tag{1}$$

序列 $\boldsymbol{z}$ 的粗糙度定义为

$$R=\sum_{i=2}^{N}(z_i-z_{i-1})^2=\sum_{i=1}^{N-1}(\Delta z_i)^2。\tag{2}$$

序列 $\boldsymbol{z}$ 越平滑，其对于原始信号 $\boldsymbol{y}$ 的保真度就越低。为了平衡这两者之间的冲突，引入保真度权重 w_i 和平滑参数 λ，得到如下的代价函数：

$$Q=\sum_i w_i(y_i-z_i)^2+\lambda\sum_i(\Delta^2 z_i)^2。\tag{3}$$

其矩阵形式可表示为

$$Q=(\boldsymbol{y}-\boldsymbol{z})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{z})+\lambda\boldsymbol{z}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{z},\tag{4}$$

式中：$\boldsymbol{D}$ 为 2 阶差分矩阵，$\boldsymbol{D}\boldsymbol{z}=\Delta^2\boldsymbol{z}=(z_i-z_{i-1})-(z_{i-1}-z_{i-2})=z_i-2z_{i-1}+z_{i-2}$；权重矩阵 $\boldsymbol{W}=\mathrm{diag}(w_1,w_2,\cdots,w_N)$。为使代价函数 Q 最小化，在(4)式中对向量 $\boldsymbol{z}$ 求偏导，并令导数为零，得到：

$$\boldsymbol{z}=(\boldsymbol{W}+\lambda\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{W}\boldsymbol{y}。\tag{5}$$

在 AsLS 算法中，权重 w_i 根据 $\boldsymbol{y}$ 与 $\boldsymbol{z}$ 之间的残差信号进行赋值：当光谱信号低于拟合基线时，认为其属于基线部分，故赋予较大的权重；当光谱信号高于拟合基线时，认为其属于光谱部分，故赋予较小的权重。如此，便可实现对信号保真度的非对称加权。该规则可概括如下

$$w_i=\begin{cases}p, & y_i>z_i\\ 1-p, & y_i\leqslant z_i\end{cases}。\tag{6}$$

式中：p 为非对称参数，是一个接近于 0 的较小值。在光谱的基线校正中，通常设置 $0.001\leqslant p\leqslant 0.1$ 以及 $10^2\leqslant\lambda\leqslant10^9$，然后根据校正效果来确定最佳的参数[14]。由此可知，当光谱叠加噪声时，非对称惩罚加权最小二乘法将低于拟合基线的噪声也赋予了较大的权重，导致最终的估计基线比实际位置偏低，基线校正后的光谱会被抬升，从而影响校正结果的准确性。airPLS 算法虽然使用指数形式的加权，但其权重值在略小于拟合基线的光谱区域近似等于或略大于 1，仍会低估无谱峰光谱区域基线的位置。MPLS 算法将局部最小值权重设为 1，非局部最小值权值设为任意小的正值，实际上是对开运算结果的非对称加权，从而导致估计基线沿着噪声信号的底部切入，而不是沿着中部切入；形态学结构选择不合理也可能会引起估计基线的严重失真。针对上述存在的问题，提出了一种基于 LSRPLS 的拉曼光谱基线校正算法。

2.1　算法原理

为了解决非对称惩罚最小二乘法在噪声情况下容易导致光谱抬升的问题，所提算法采用“局部对称加权”与“非对称加权”相结合的策略。所谓的“局部对称加权”就是对于光谱中无谱峰的信号区域，认为噪声在基线上下是均匀分布的，因而对这些信号赋予相近的权重。用 $\boldsymbol{y}$ 表示光谱向量，$\boldsymbol{z}$ 表示估计的基线向量，它们的长度均为 N，i 为光谱数据点。在一定的光谱强度范围内，即无论是 $y_i<z_i$，还是 $y_i\geqslant z_i$ 的，认为它们对估计基线的影响是相同的。所谓的“非对称加权”与 AsLS 算法一致，即当光谱信号 y_i 远大于拟合基线 z_i 时，认为其属于拉曼谱峰的一部分，此时将其权重赋值为零。然后通过迭代不断调整权重函数，当达到迭代终止条件或最大迭代次数时，输出最终的拟合基线。用原始拉曼光谱减去拟合基线，就可以实现基线校正。

为了满足上述要求，选择 softsign 函数作为权重函数。softsign 函数为

$$\mathrm{softsign}(x)=\frac{1}{2}\left(1-\frac{x}{1+|x|}\right)。\tag{7}$$

设初次迭代的权重为 1，第 t 次迭代的权重可以表示为

$$w_i^t=\begin{cases}\dfrac{1}{2}\left\{1-\dfrac{10^t[d_i-(2s_{d^-}-m_{d^-})]/s_{d^-}}{1+|10^t[d_i-(2s_{d^-}-m_{d^-})]/s_{d^-}|}\right\}, & y_i>z_i\\ 1, & y_i\leqslant z_i\end{cases},\tag{8}$$

式中：$d_i=y_i-z_i$，为光谱信号与拟合基线之间的残差信号；下标 $\boldsymbol{d}^-$ 表示残差信号的负值部分；m_{d^-} 和 s_{d^-} 分别表示 $\boldsymbol{d}^-$ 的均值和标准差。

给定均值和方差，softsign 权重函数的示意图如图 1 所示。可以看到：softsign 权重函数值随着信号强度的增加而逐渐减小；在拉曼光谱与拟合基线的差值 $\boldsymbol{d}<\boldsymbol{0}$ 的部分，仅使用 softsign 函数便可以表示(6)式中的分段形式；当差值信号 d_i 小于估计噪声的均值时，softsign 函数能够对低于或高于基线的信号赋予近似相等的权重。在高斯噪声假设下，根据三倍标准差准则可知，距离噪声均值 3 倍标准差的范围能够覆盖 99.7%的噪声，对应到图 1，此时 softsign 函数权重值仍保持为较小的权重值。随着差值信号继续增大，直至光谱信号远大于拟合基

线时，可认为该部分为谱峰，此时的 softsign 函数权重值为 0。

从图 1 还可以看出：随着迭代次数的增加，softsign 函数逐渐趋向于一个经过平移和反转的单位阶跃函数的形式，等价于 AsLS 算法的权重形式；这与多次迭代后基线趋于稳定的情况相符；随着迭代次数不断增加，上一次迭代的估计基线与下一次迭代的估计基线的差异逐渐减小；当达到一定的迭代次数后，估计基线几乎不再改变，因而也不必考虑对光谱信号的局部对称加权。

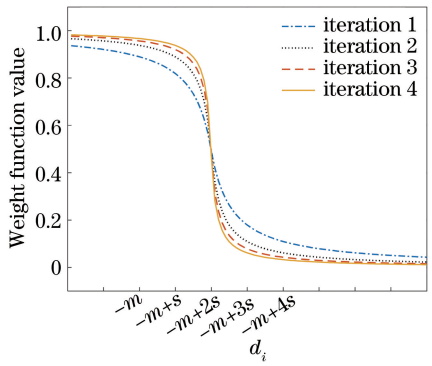

图 1　softsign 权重函数示意图

Fig. 1　Schematic of softsign weighting function

2.2　算法实现

LSRPLS 基线校正算法的步骤可概括如下。

步骤 1)：载入待基线拟合的原始光谱数据 $\boldsymbol{y}$，设置平滑参数 λ。

步骤 2)：对权重进行初始化 $\boldsymbol{w}^0=[1,1,\cdots,1]$，则权重矩阵 $\boldsymbol{W}$ 是一个稀疏对角矩阵，$\boldsymbol{W}=\mathrm{diag}(w_1,w_2,\cdots,w_N)$，$N$ 为光谱数据点个数。

步骤 3)：将初始权重 $\boldsymbol{W}^0$ 代入(5)式，计算初始基线，$\boldsymbol{z}^0=(\boldsymbol{W}^0+\lambda\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{W}^0\boldsymbol{y}$。

步骤 4)：判断是否满足迭代终止条件，若不满足转步骤 5)，否则转至步骤 7)。

步骤 5)：计算拉曼光谱与拟合基线之间的残差信号 $\boldsymbol{d}=\boldsymbol{y}-\boldsymbol{z}$，取残差信号的负值部分 $\boldsymbol{d}^-$，并计算其均值 $m_{\boldsymbol{d}^-}$ 和标准差 $s_{\boldsymbol{d}^-}$。

步骤 6)：使用 softsign 函数和迭代次数 t 对权重进行迭代更新得到 $\boldsymbol{W}^t$，计算第 t 次迭代的拟合基线，$\boldsymbol{z}^t=(\boldsymbol{W}^t+\lambda\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{W}^t\boldsymbol{y}$，转步骤 4)。

步骤 7)：迭代终止，用原始光谱减去最终的拟合基线即可实现拉曼光谱的基线校正。

综上所述，可以得到 LSRPLS 基线校正算法流程图，如图 2 所示。

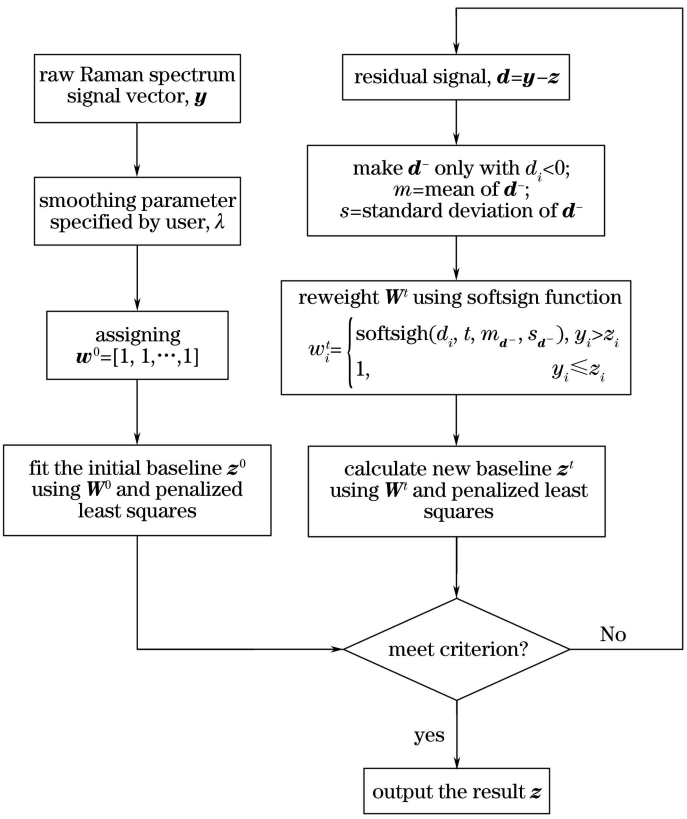

图 2　LSRPLS 基线校正算法的流程图

Fig. 2　Flow chart of LSRPLS baseline correction algorithm

将 LSRPLS 基线校正算法的迭代终止条件设置为连续两次迭代的权重不再改变或者改变量很小，即：

$$|\boldsymbol{w}^{t+1}-\boldsymbol{w}^{t}|/\boldsymbol{w}^{t}<R_{\mathrm{w}}。\tag{9}$$

通常，比值 R_{w} 的取值范围为 $10^{-6}\sim10^{-2}$，即可达到有效的估计精度。

3　实验结果及分析

3.1　LSRPLS 基线校正算法在模拟拉曼光谱上的仿真

1）模拟数据

模拟光谱由三部分构成：模拟谱峰信号、模拟基线和随机噪声。本工作使用多个 Lorentzian 峰进行叠加生成模拟光谱的谱峰信号。Lorentzian 表达式如下

$$y=\sum_{i=1}^{N_{\mathrm{um}}}\frac{2A_{oi}}{\pi}\frac{\omega_{oi}}{4(r-r_{oi})^{2}+\omega_{oi}^{2}},\tag{10}$$

式中：r 为拉曼位移；r_{oi} 为拉曼谱峰的中心位置；A_{oi} 为拉曼谱峰面积；ω_{oi} 为谱峰半峰全宽；N_{um} 为谱峰总个数。实验中设置光谱长度为 3000，Lorentzian 峰共 5 个，其参数为 $r_{oi}=\{300,\ 600,\ 990,\ 1200,\ 2000\}$，$\omega_{oi}=\{30,\ 23,\ 10,\ 20,\ 40\}$，$A_{oi}=\{9900,\ 11000,\ 2900,\ 5200,\ 14000\}$。

设置了 4 种不同形式的模拟基线，分别为线性函数型、正弦曲线型、高斯曲线型和指数函数型。生成函数如下：

$$\begin{cases}b_1=0.0374x+123.5\\ b_2=250\sin(7.5\times10^{-4}x)\\ b_3=500\exp\left[-\dfrac{(x-500)^2}{2\times750^2}\right]+\\ \qquad 2000\exp\left[-\dfrac{(x-2200)^2}{2\times350^2}\right]\\ b_4=573\exp\left(-\dfrac{x}{250}\right)\end{cases}。\tag{11}$$

将上述基线分别与 Lorentzian 谱峰进行叠加，得到无噪声的理想拉曼光谱。模拟光谱的噪声水平统一设置信噪比为 15。图 3 为不同基线类型的模拟拉曼光谱图。

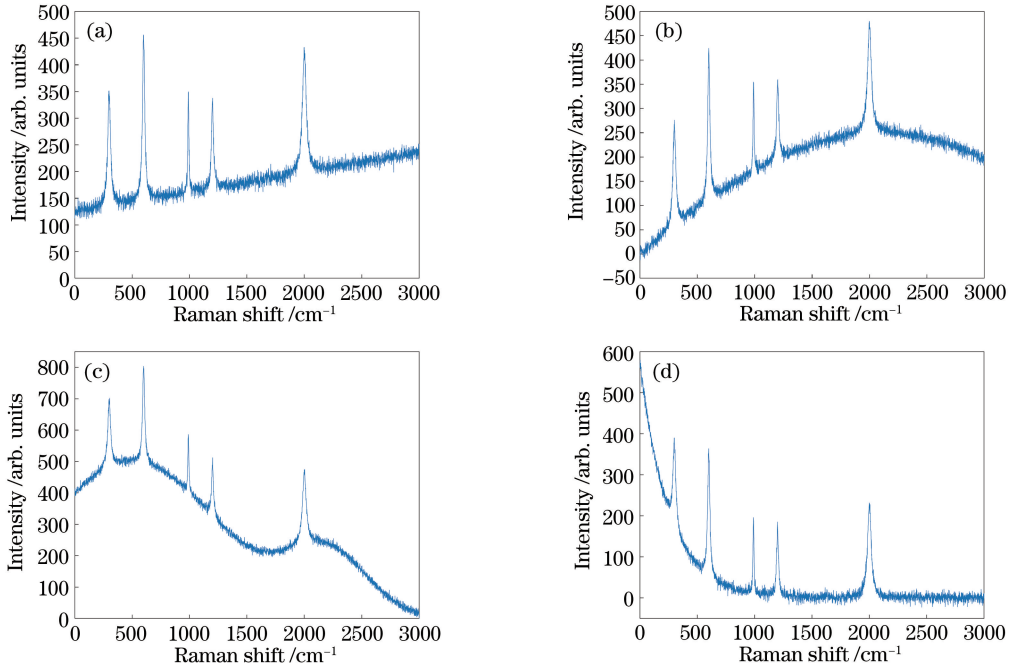

图 3　不同基线类型的模拟拉曼光谱（信噪比为 15）。

（a）线性函数型基线；（b）正弦曲线型基线；（c）高斯曲线型基线；（d）指数函数型基线

Fig. 3　Simulated Raman spectra with different types of baselines (signal to noise ratio of 15). (a) Linear baseline; (b) sinusoidal baseline; (c) Gaussian baseline; (d) exponential baseline

2）方法对比与结果讨论

为了验证 LSRPLS 基线校正的效果，采用理想基线 $b(i)$ 与估计基线 $\hat{b}(i)$ 的均方根误差 e_{RMSE} 进行评价。e_{RMSE} 的计算公式为

$$e_{\mathrm{RMSE}}=\sqrt{\frac{\sum_{i=1}^{N}[b(i)-\hat{b}(i)]^{2}}{N}},\tag{12}$$

均方根误差的值越小，说明基线校正的准确性越高。因此，实验中调节平滑参数，选取均方根误差值最小时对应的 λ 为最优平滑参数。通常，参数 λ 的取值范围为 $10^2 \sim 10^9$，选取原则可以参考 Eilers 等[14]的论述。

将(11)式中的 4 种基线进行叠加可以得到一个组合基线，用以模拟实际中较为复杂的荧光背景。使用 LSRPLS 算法对该组合基线进行拟合，图 4(a)显示了当信噪比为 15 时不同 λ 所对应的均方根误差，可以看到：当 $\lg\lambda=6.5$ 时，均方根误差最小，为 5.9281。图 4(b)为 $\lg\lambda=6.5$、$R_w=10^{-4}$ 时对应的基线校正结果，可以看到拟合基线与理想基线符合得较好，表明该算法能够准确、有效地进行拉曼光谱的基线校正。

为了评估算法在不同信噪比下的适用性，将 Lorentzian 谱峰与基线进行叠加得到无噪声的理想拉曼光谱。模拟光谱的噪声水平，设置信噪比为 20、30。分别选取合适的 λ 进行基线校正，实验结果如图 4(c)、(d)所示。进一步计算可知，当信噪比为 20 时，$e_{\mathrm{RMSE}}=5.2524$；当信噪比为 30 时，$e_{\mathrm{RMSE}}=6.0309$。这说明 LSRPLS 基线校正算法对不同信噪比的拉曼光谱均具有较好的基线校正能力。应当注意的是，LSRPLS 基线校正算法主要是通过对无谱峰部分的噪声信号进行近似对称加权来避免噪声对估计基线的不良影响，而对于噪声信号的估计，则是通过计算原始光谱与估计基线差值信号 $\boldsymbol{d}^-$ 的均值得到的。可见，算法对于弱的拉曼散射光谱的适用性需要结合光谱中所含的噪声水平进行考虑：当弱的拉曼散射光谱的谱峰强度小于均值 m_{d^-} 时，即谱峰信号被噪声信号淹没时，算法将难以区分该部分的谱峰区域和噪声信号。实际上，如果测得光谱的谱峰被噪声淹没，就将难以提取拉曼特征谱峰的有效信息，通常可认为该谱峰为无效谱峰，按照噪声信号进行处理。可见，所提算法在实际应用中依然具有良好的适用性。

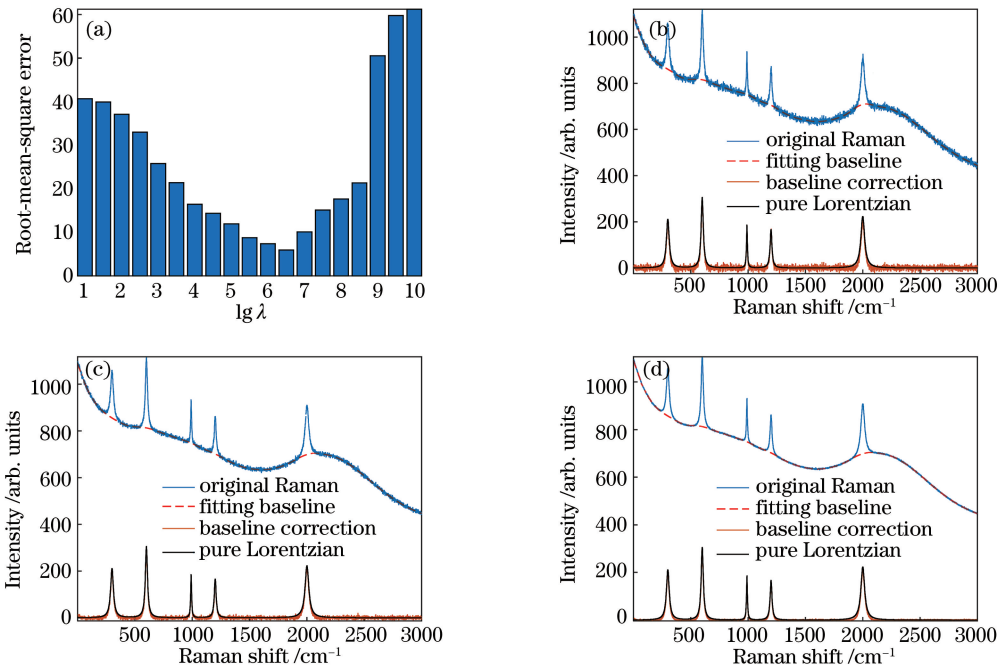

图 4　组合基线的光谱校正结果。(a)信噪比为 15 时，不同 λ 对应的均方根误差；(b)信噪比为 15 时，基线校正前后的光谱图；(c)信噪比为 20 时，基线校正前后的光谱图；(d)信噪比为 30 时，基线校正前后的光谱图

Fig. 4　Spectrum correction results with combination baseline. (a) Root-mean-square error for various λ when signal to noise ratio is 15; (b) Raman spectra before and after baseline correction when signal to noise ratio is 15; (c) Raman spectra before and after baseline correction when signal to noise ratio is 20; (d) Raman spectra before and after baseline correction when signal to noise ratio is 30

为了检验基线校正对光谱相对强度的影响，以上述信噪比为 20 时的模拟光谱为例，将原始 Lorentzian 谱峰强度和校正后的谱峰强度(PH)进行对比，结果如表 1 所示。从表 1 中可以看到：基线校正后光谱的强度略高于或低于原始光谱强度，会对谱峰间的相对强度产生一定影响；对于较简单的基线形态，如线性曲线和正弦曲线两种基线，校正后的光谱强度变化很小；随着基线形态的复杂化，校正

后的光谱强度变化也增大。应当注意的是，尽管所提算法进行基线校正后谱峰间的相对强度发生了改变，但这其中也受到噪声信号的干扰。例如组合型基线光谱，结合图 4(c)可以看到，基线校正后的光谱在整体上能够较好地与原始 Lorentzian 谱峰峰形、谱峰强度保持一致。此外，与其他常用基线校正方法相比，所提算法有效避免了光谱的抬升，因此校正后的光谱强度的改变量要小于其他方法。

表 1　模拟光谱基线校正前后的谱峰强度对比

Table 1　Comparison of peak heights before and after simulated spectrum baseline correction

Baseline function	Index	Peak number				
		1	2	3	4	5
Linear	Estimated PH	211.54	309.38	187.50	157.72	223.91
	Peak error	−0.95	−4.26	−2.05	8.21	−1.03
Sinusoidal	Estimated PH	209.27	307.16	180.15	163.47	221.13
	Peak error	1.32	−2.04	5.30	2.47	1.76
Gaussian	Estimated PH	203.33	295.17	186.00	160.90	210.04
	Peak error	7.26	9.95	−0.55	5.04	12.84
Exponential	Estimated PH	194.16	300.85	184.19	160.89	211.38
	Peak error	16.43	4.26	1.25	5.05	11.50
Combination	Estimated PH	202.81	303.99	185.23	158.39	209.92
	Peak error	7.78	1.12	0.21	7.54	12.97
Actual PH		210.59	305.12	185.45	165.94	222.88

为进一步测试所提算法对不同基线类型、不同荧光水平的拉曼光谱的基线校正能力，使用所提算法对图 3 中的各组光谱进行处理，并与常用多项式拟合方法——多项式拟合基线校正(ModPoly)及 AdaptMinmax、AsLS、airPLS 和 MPLS 进行比较。模拟光谱的荧光水平用信基比 e_{RSB} 进行描述，信基比定义为拉曼光谱的最大强度值与荧光信号幅度的比值[20]。对于各组模拟光谱，设置 $e_{RSB}=\{0.1, 0.2, \cdots, 1\}$。实验结果如图 5 所示。为了比较基线校正的平均性能，表 2 给出了各方法在不同基线类型的模拟光谱上的均方根误差的均值。

由图 5(a)可以看出：LSRPLS 基线校正算法的均方根误差最小，具有最高的基线估计准确度；airPLS 算法次之，之后是 MPLS 算法，但这两种算法的估计误差均随着信基比的增大而增大；AdaptMinmax 算法是一种自动基线校正方法，无需调节参数，其性能与 AsLS 算法类似；ModPoly 算法由于将高于拟合基线的噪声认为是拉曼信号，并使用拟合多项式的估计值进行替换，因而导致结果存在较大偏差。

由图 5(b)可知：对于正弦曲线型基线，三种改进的非对称加权惩罚最小二乘法与原始 AsLS 算法相比，性能都得到了提升，能够获得良好的基线校正结果；从估计基线的准确性来看，LSRPLS 算法的准确性最高，airPLS 与 MPLS 算法的准确度相当；从基线估计的稳定性来看，airPLS 算法的误差曲线有较为明显的波动，而 LSRPLS 算法与 MPLS 算法则相对稳定。

由图 5(c)可以看出：对于高斯曲线型基线，AdaptMinmax 算法的估计误差明显增大，不能进行有效的基线校正，说明该算法虽然可以省去调节参数的步骤，但欠缺对复杂形态基线的处理能力。为了凸显各方法之间的差异，略去均方根误差过大的 AdaptMinmax 算法的实验结果，并对坐标轴进行纵向放大处理，得到图 5(e)。由图 5(e)可知：LSRPLS 算法的结果依然保持着最高的准确度，性能优于 airPLS 算法和 MPLS 算法；AsLS 算法与 ModPoly 算法的表现略差些，但仍能满足基线校正的要求。

由图 5(d)可以看出：对于指数函数型基线，AdaptMinmax 算法不能很好地进行处理，并且其他各类方法的准确度与图 5(a)～(c)中的结果相比，也都有不同程度的下降。尤其是 MPLS 算法，其在信基比较低时的均方根误差明显增大。原因是边界处较为陡峭的指数函数型基线与拉曼光谱尖锐的谱峰形态类似，而信基比较低时光谱中的谱峰特征又不够明显，使得 MPLS 算法的形态学结构元素误将陡峭的基线部分作为拉曼谱峰进行运算处理，导致该部分的估计基线偏离真实基线。

尽管如此，LSRPLS 算法的均方根误差仍平稳地保持在较小的范围内，说明所提算法在指数曲线型基线上也具有较高的准确度和良好的稳定性。而 airPLS 与 AsLS 算法的性能相近，没能体现出自适应迭代加权的优越性。

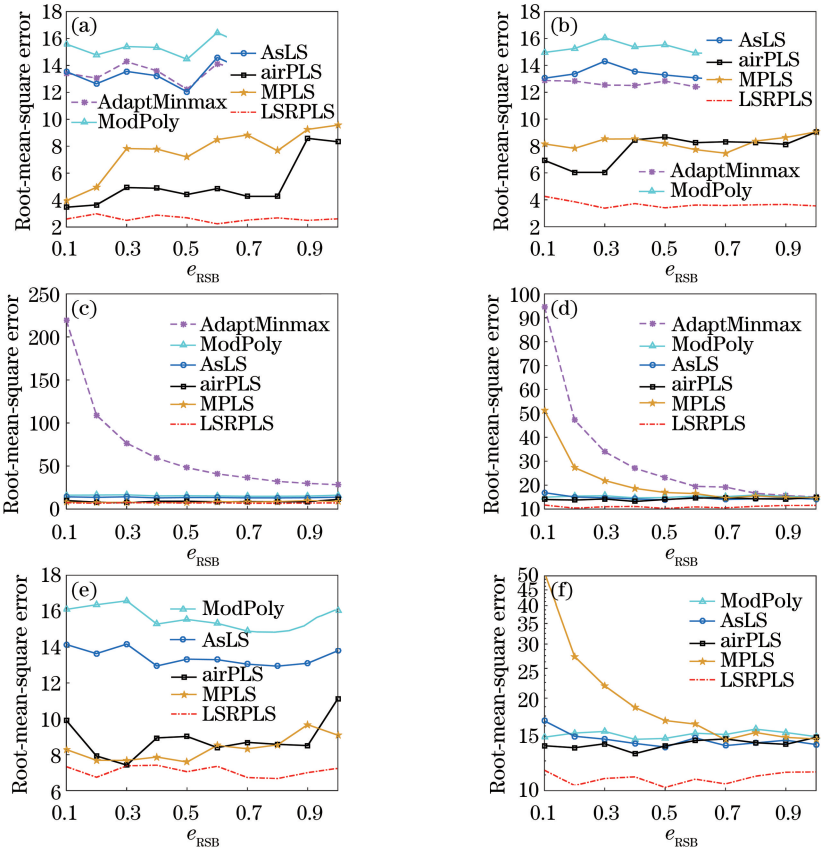

图 5　模拟拉曼光谱的实验结果。(a)线性函数型基线的估计误差曲线；(b)正弦曲线型基线的估计误差曲线；(c)高斯曲线型基线的估计误差曲线；(d)指数函数型基线的估计误差曲线；(e)高斯曲线型基线纵向放大的估计误差曲线；(f)指数函数型基线纵向放大的估计误差曲线

Fig. 5　Experimental results of simulated Raman spectra. (a) Estimated error curves for linear baseline; (b) estimated error curves for sinusoidal baseline; (c) estimated error curves for Gaussian baseline; (d) estimated error curves for exponential baseline; (e) vertically enlarged error curves for Gaussian baseline; (f) vertically enlarged error curves for exponential baseline

表 2　不同基线校正方法的均方根误差的均值

Table 2　Mean root-mean-square error using different baseline correction methods

Method	Mean root-mean-square error				
	Linear	Sinusoidal	Gaussian	Exponential	Combination
AdaptMinmax	13.6634	12.7126	67.9442	31.2003	60.9522
ModPoly	15.4735	15.3248	15.5909	15.2304	15.4193
AsLS	13.4091	13.4045	13.4393	14.6637	14.4587
airPLS	5.1672	7.8164	8.8518	14.1867	10.9666
MPLS	7.5553	8.2472	8.3297	21.2369	14.6842
LSRPLS	2.6808	3.6650	7.0933	10.9928	11.2270

综上所述，LSRPLS算法在不同基线形态的模拟拉曼光谱都具有最佳的基线校正准确度，说明所提算法使用的局部对称加权思想可以有效减小噪声对校正结果的影响，使估计基线更接近于真实基线。因此，在拉曼光谱基线校正方法中考虑噪声的影响，是一种提高校正结果准确度的有效手段。

3.2　LSRPLS 基线校正算法在实际拉曼光谱上的应用

1) 实验数据

使用北京卓立汉光仪器有限公司生产的 Finder One 型微区激光拉曼光谱仪采集罗丹明 6G（Rhodamine 6G，R6G）、4-巯基苯甲酸（4-Mercaptobenzoic Acid，4-MBA）和 4-巯基吡啶（4-

Mercaptopyridine，4-MPY）的表面增强拉曼光谱（SERS）。设置激光波长为 532 nm，积分时间为 0.5 s，功率为 50 mW。以银纳米星为增强基底，用移液枪取 10 μL 银纳米星溶液，将其滴到预先洗净的硅片上，风干后作为基底。取 10 μL 样品溶液，滴在已风干的基底上，然后采集其 SERS 光谱，每种样品采集 50 条有效光谱。

2）R6G 的 SERS 光谱基线校正

利用 LSRPLS 算法对由强荧光背景引起基线漂移的 R6G 的 SERS 光谱进行处理，设置参数：$\lambda=10^7$，$R_w=10^{-4}$。选取与所提算法相关的三种基于惩罚最小二乘的基线校正算法进行对比，这三种算法分别为 AsLS、airPLS 和 MPLS。实验结果如图 6 所示。图 6(a)中的蓝色实线为 R6G 的原始 SERS 光谱，三种不同颜色的虚线分别表示 AsLS、airPLS 和 MPLS 算法所得的估计基线，红色实线为 LSRPLS 算法所得的估计基线。图 6(b)为采用 LSRPLS 算法基线校正后 R6G 的 SERS 光谱。

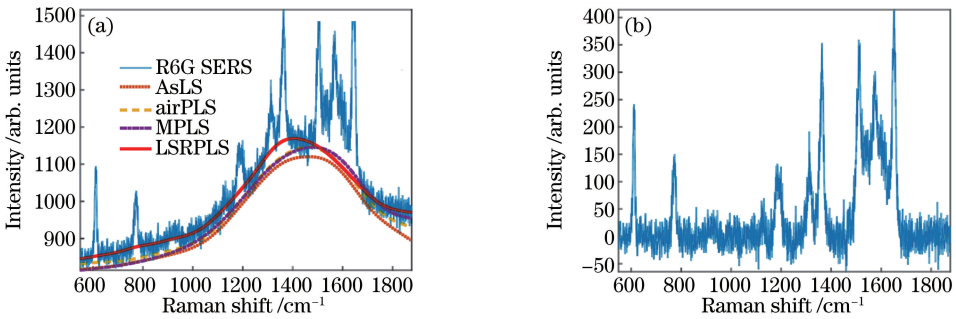

图 6　R6G 的 SERS 光谱估计基线与校正结果。(a)原始 R6G 拉曼光谱及各校正算法的估计基线；(b) LSRPLS 基线校正算法的基线校正结果

Fig. 6　SERS spectra estimated baselines and correction results of R6G. (a) Original R6G Raman spectrum and estimated baselines with different correction algorithms; (b) baseline correction result using LSRPLS algorithm

因为所提算法采用“局部对称加权”与“非对称加权”相结合的策略，引入 softsign 函数实现局部对称加权的思想。“局部对称加权”就是认为光谱中无谱峰的信号区域的噪声在基线上下是均匀分布的，因而对这些信号赋予相近的权重，而当光谱信号远大于拟合基线时，则认为其属于拉曼谱峰的部分，此时将其权重赋值为零，然后通过迭代不断调整权重函数，当达到迭代终止条件或最大迭代次数时，输出最终的拟合基线，因而有效避免了因为未采取数据滤波而出现的估计基线偏低、校正后光谱抬升的现象。由图 6 可知，LSRPLS 基线校正算法所得的估计基线能够平滑地穿过噪声带，估计基线与 SERS 光谱的真实基线最为接近，验证了 LSRPLS 基线校正算法使用局部对等加权来避免噪声影响的有效性。而 AsLS、airPLS 和 MPLS 算法所得估计基线存在着不同程度的位置偏低的现象，尤其是直接使用常数加权的 AsLS 算法，估计基线沿着噪声信号的底部，而不是沿着拉曼光谱真实基线的底部，这将造成校正后光谱向上抬升，导致光谱强度偏大，从而影响对拉曼光谱的后续分析，特别是定量分析结果的准确性。

3）药品 SERS 光谱基线校正前后的定性分析

实验测得 4-MBA 和 4-MPY 的 SERS 光谱如图 7(a)、(c)所示，可以看到原始光谱存在明显的噪声干扰和基线漂移，需要对其进行预处理。采用 Savitzky-Golay 滤波法对光谱进行去噪处理，设置参数为 2 阶多项式和 21 大小的窗口宽度；采用 LSRPLS 基线校正算法进行基线校正，设置参数 $\lambda=10^4$，$R_w=10^{-5}$，对 50 条拉曼光谱进行批量校正。图 7(b)、(d)分别为 4-MBA 和 4-MPY 基线校正后的 SERS 光谱。可以看出：虽然基线光谱在形态和强度上有所不同，但使用相同参数的 LSRPLS 基线校正算法均能较好地完成基线校正，说明 LSRPLS 基线校正算法具有一定的自适应性；相比传统的人工多项式拟合法，LSRPLS 基线校正算法更适合用于快速、准确的基线批处理。

为了更直观地显示，将两类药品的原始光谱和基线校正后的光谱分别进行主成分分析（PCA），并取第一主成分（PC1）和第二主成分（PC2）绘制散点图，如图 8(a)和图 8(b)所示。可以看到：4-MBA 和 4-MPY 的原始光谱存在较大混叠，难以直接区分；经过 LSRPLS 基线校正处理后，两类药品的光谱在主成分空间中的聚集度明显提高，可以通过线性判别很好地进行分类。

利用原始数据和预处理后的数据分别构建基于主成分分析的线性判别模型（PCA-LDA），实现两类

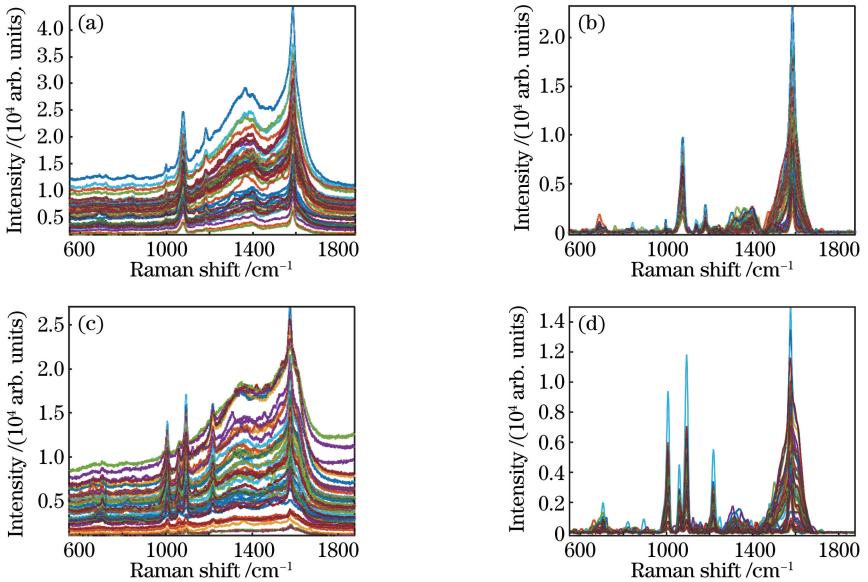

图 7　两类药品 SERS 光谱的基线校正结果。(a) 4-MBA 的原始光谱;(b) 4-MBA 的基线校正光谱;(c) 4-MPY 的原始光谱;(d) 4-MPY 的基线校正光谱

Fig. 7　SERS spectra of two samples before and after baseline correction. (a) Original spectra of 4-MBA; (b) corrected spectra of 4-MBA; (c) original spectra of 4-MPY; (d) corrected spectra of 4-MPY

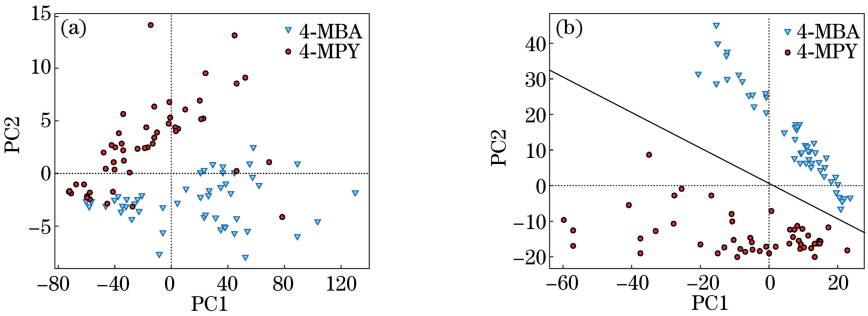

图 8　两类药品拉曼光谱的 PCA 得分图。(a)原始拉曼光谱的 PCA 得分;(b)基线校正光谱的 PCA 得分

Fig. 8　Plots of PCA scores for Raman spectra of two samples. (a) PCA score of original Raman spectra; (b) PCA score of baseline corrected spectra

药品混合拉曼光谱的定性分析。从拉曼光谱中随机选取 30 个用于训练、20 个用于测试,因此 PCA-LDA 分类模型的训练集共计 60 个样本,测试集共计 40 个样本。

表 3 中列出了 PCA-LDA 模型的分类结果。其中 R_{CC} 为正确分类率,R_{CC} 越高,则分类性能越好,其计算公式为

$$R_{CC}=\frac{T_P+T_N}{(T_P+F_N)+(F_P+T_N)},\tag{13}$$

式中:T_P 为真正类,表示正确划分为正类的个数;F_P 为假正类,表示错误判别为正类的个数;T_N 为真负类,表示正确判别为负类的个数;F_N 为假负类,表示错误判别为负类的个数。

灵敏性表示分类模型将实际 g 类的样本判为 g 类的能力,其表达式为

$$S_{ens}=\frac{T_P}{T_P+F_N}。\tag{14}$$

特异性表示分类模型将非 g 类的样本判为非 g 类的能力,其表达式为

$$S_{pec}=\frac{T_N}{F_P+T_N}。\tag{15}$$

由表 3 可知:由于 4-MBA 和 4-MPY 光谱之间的相似性,以及噪声和荧光背景的干扰,模型对原始混合光谱的分类性能较差,对两类样本的误分类率较高;在光谱预处理之后,训练集和测试集上的灵敏性、特异性和正确分类率均达到了 100%,实现了完全正确的分类,说明 LSRPLS 基线校正算法能在去除基线漂移的同时保留光谱的有效信息,从而提升

表 3　PCA-LDA 对 4-MBA 和 4-MPY 的分类结果

Table 3　Classification results for 4-MBA and 4-MPY by PCA-LDA

True class			Cross-validation		Test	
			4-MBA	4-MPY	4-MBA	4-MPY
Raw spectrum	Assigned class	4-MBA	22	8	16	4
		4-MPY	6	24	1	19
	Sensitivity /%	73	80	80	95	
	Specificity /%	80	73	95	80	
	R_{CC}/%	77		88		
Pretreatment	Assigned class	4-MBA	30	—	20	
		4-MPY	—	30	—	20
	Sensitivity /%	100	100	100	100	
	Specificity /%	100	100	100	100	
	R_{CC}/%	100		100		

了光谱定性分析的准确性，进一步验证了该算法应用于实际拉曼光谱处理的有效性。

4　结　　论

提出了一种基于迭代非对称加权惩罚最小二乘(LSRPLS)的基线校正算法。与阶跃函数相比，所用 softsign 函数形式的权重可看作是一个局部平等加权，解决了最小二乘法普遍造成的估计基线偏低、校正结果不准的问题。所提算法分别在模拟拉曼光谱和实验拉曼光谱上进行了准确性和有效性验证，并与其他常用方法进行了比较。结果表明：LSRPLS 基线校正算法对模拟拉曼光谱和实测 SERS 光谱都具有良好的估计准确度，且性能稳定，不易受光谱信基比的影响；基线校正后光谱的聚集度提升，分类准确率得到明显改善，说明 LSRPLS 基线校正算法在实现基线校正的同时，能够有效保留原始光谱的有用信息，为光谱的进一步分析提供了依据。因此，所提算法能够较好地实现拉曼光谱的基线校正，并且使用方便，可作为一种有效的拉曼光谱基线校正算法。

参　考　文　献

[1] McCreery R L. Raman spectroscopy for chemical analysis[M]. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2000.

[2] Wang Y Q, Yan B, Chen L X. SERS tags: novel optical nanoprobes for bioanalysis [J]. Chemical Reviews, 2013, 113(3): 1391-1428.

[3] Wang S, Haishan Z. Real-time *in vivo* Raman spectroscopy and its clinical applications in early cancer detection [J]. Chinese Journal of Lasers, 2018, 45(2): 0207002.
王爽, Haishan Z. 实时拉曼光谱分析技术及其在临床早期癌症检测中的应用[J]. 中国激光, 2018, 45(2): 0207002.

[4] Zhao F, Peng Y K. Measurement of iodine value of pork′s subcutaneous adipose tissue by interval partial least square and Raman spectroscopy [J]. Chinese Journal of Lasers, 2017, 44(11): 1111001.
赵芳, 彭彦昆. 区间偏最小二乘法结合拉曼光谱测定猪肉皮下脂肪的碘值[J]. 中国激光, 2017, 44(11): 1111001.

[5] Liu G K, Zheng H, Lu J L. Recent progress and perspective of trace antibiotics detection in aquatic environment by surface-enhanced Raman spectroscopy[J]. Trends in Environmental Analytical Chemistry, 2017, 16: 16-23.

[6] Campos J L E, Miranda H, Rabelo C, *et al*. Applications of Raman spectroscopy in graphene-related materials and the development of parameterized PCA for large-scale data analysis[J]. Journal of Raman Spectroscopy, 2018, 49(1): 54-65.

[7] Huzortey A A, Anderson B, Owusu A. A composite algorithm for optimized baseline correction inRaman spectroscopy[C]. Frontiers in Optics 2017, Frontiers in Optics, Washington, D.C., 2017.

[8] Tatarkovič M, Synytsya A, Šťovíčková L, *et al*. The minimizing of fluorescence background in Raman optical activity and Raman spectra of human blood plasma[J]. Analytical and Bioanalytical Chemistry, 2015, 407(5): 1335-1342.

[9] Hu Y G, Zhou J J, Tang J, *et al*. The application of complex wavelet transform to spectral signals background deduction[J]. Chromatographia, 2013, 76(11/12): 687-696.

[10] Lieber C A, Mahadevan-Jansen A. Automated method for subtraction of fluorescence from biological Raman spectra[J]. Applied Spectroscopy, 2003, 57(11): 1363-1367.

[11] Cao A, Pandya A K, Serhatkulu G K, *et al*. A robust method for automated background subtraction of tissue fluorescence [J]. Journal of Raman Spectroscopy, 2007, 38(9): 1199-1205.

[12] Wang T, Dai L K. Background subtraction of Raman spectra based on iterative polynomial smoothing[J]. Applied Spectroscopy, 2017, 71(6): 1169-1179.

[13] Yang G Y, Li L, Chen H, *et al*. Baseline correction method for Raman spectra based on generalized Whittaker smoother[J]. Chinese Journal of Lasers, 2015, 42(9): 0915003.
杨桂燕, 李路, 陈和, 等. 基于广义 Whittaker 平滑器的拉曼光谱基线校正方法[J]. 中国激光, 2015, 42(9): 0915003.

[14] Eilers P H C, Boelens H F M. Baseline correction with asymmetric least squares smoothing [J/OL]. (2005-10-21)[2018-06-01]. https://zanran_storage.s3.amazonaws.com/www.science.uva.nl/ContentPages/443199618.pdf.

[15] He S X, Zhang W, Liu L J, *et al*. Baseline correction for Raman spectra using an improved asymmetric least squares method[J]. Anal Methods, 2014, 6(12): 4402-4407.

[16] Oller-Moreno S, Pardo A, Jiménez-Soto J M, *et al*. Adaptive asymmetric least squares baseline estimation for analytical instruments [C]. 2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14), 2014: 14281834.

[17] Zhang Z M, Chen S, Liang Y Z. Baseline correction using adaptive iteratively reweighted penalized least squares[J]. Analyst, 2010, 135(5): 1138-1146.

[18] Park A, Baek S J, Park J Q, *et al*. Automatic selection of optimal parameter for baseline correction using asymmetrically reweighted penalized least squares[J]. Journal of the Institute of Electronics and Information Engineers, 2016, 53(3): 124-131.

[19] Li Z, Zhan D J, Wang J J, *et al*. Morphological weighted penalized least squares for background correction[J]. Analyst, 2013, 138(16): 4483-4492.

[20] Schulze G, Jirasek A, Yu M M L, *et al*. Investigation of selected baseline removal techniques as candidates for automated implementation [J]. Applied Spectroscopy, 2005, 59(5): 545-574.