基于自适应加窗 spline 曲线拟合的拉曼光谱去基线方法

刘 龙[1]，范贤光[1,2*]，康哲铭[1]，吴 怡[1]，王 昕[1,2*]

1. 厦门大学航空航天学院仪器与电气系，福建 厦门 361005
2. 传感技术福建省高等学校重点实验室，福建 厦门 361005

摘 要 拉曼光谱是一种无损快速检测技术，可以提供材料的定性和定量信息，因而在医药、化工等诸多领域得到了广泛的应用。但是，由于样品荧光背景噪声的影响，造成拉曼光谱信号出现基线漂移现象，这给拉曼光谱的特征峰识别和拉曼成像带来十分严重的影响。目前，改进实验方法和数值处理是解决该问题的两种重要手段。改进实验方法上，有偏振调制法和高频调制法等，但存在实验设备复杂，检测技术难度大等缺点；数值处理上，有多项式拟合和小波变换等，但容易出现欠拟合和过拟合等现象。本文在不改换高精密设备的前提下，针对传统基线校正的方法进行了改进，提出一种基于自适应加窗 spline 曲线拟合的拉曼光谱去基线方法。首先，基于谱峰识别算法和初始搜索步长求得谷值的最优搜索间距，并利用谱谷识别算法完成谷值曲线的拟合；其次，利用最优搜索间距和谱峰识别算法，求得谷值曲线峰值位置，并在该位置处对称添加自适应矩形窗函数去除峰值，重新划分整个区间，拟合谷值曲线；再次，逐点比较拟合曲线与原拉曼光谱信号，取较小值，拟合曲线；最后，重复加窗去除峰值操作，直至自适应窗函数宽度低于阈值，完成拉曼光谱信号的基线拟合。在实验中，选用乙酸丁酯、聚甲基丙烯酸甲酯(polymethyl methacrylate，PMMA)作为实验样品，利用该方法对其拉曼光谱信号进行了基线校正，观察并比较该方法和传统方法的校正结果。实验结果表明，该方法能够有效地消除拉曼光谱信号的基线漂移，较好的保留一些较弱的拉曼特征峰，且不易出现欠拟合和过拟合的现象，获得了良好的基线校正效果，为进一步分析光谱数据和实现拉曼成像提供准确可靠的信息。

关键词 拉曼光谱；基线漂移；基线校正；最优搜索

中图分类号：O657.37 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)01-0111-05

引 言

拉曼散射又称拉曼效应，是一种基于激光的光谱技术，由印度物理学家拉曼(Raman)[1]于1928年首先发现。作为一种鉴定分子结构的重要手段，拉曼光谱可提供有关分子振动的定量信息，可用于研究组织和细胞内分子的化学组成和结构。同时，拉曼光谱又具备无损伤、无需标记等优点，在生物医药、食品监测以及各种疾病诊断等多个领域得到广泛应用[2-3]。

然而，由于自发拉曼信号很弱，仅为原始激发光信号强度的10^{-8}左右，在利用拉曼光谱仪测试中，不可避免的受到荧光背景干扰，出现基线漂移现象，严重影响拉曼光谱的分析应用能力。因此，减少基线漂移，提高拉曼光谱信号的信噪比，变得至关重要。目前，解决该问题的主要策略分为两大类：改进实验方法和数值处理。改进实验方法中，有偏振调制法，高频调制法和门控法等[4-6]。虽然可以在一定程度上减少基线漂移，但是其结构复杂、造价昂贵，一般由实验室自行搭建，用于前沿科学研究，推广难度相对较大。数值处理中，有频域滤波、小波变换和曲线拟合等[7-10]。频域滤波是对拉曼信号进行傅里叶变换，然后设计合适滤波器进行滤波处理的方法。尽管频域滤波有一定的效果，但是这种方法可能造成拉曼光谱的人为扭曲，且参数设计复杂。小波变换是对拉曼信号作分解处理，得到一系列不同频率正弦波，实现基线去除的目标。然而，不同拉曼光谱的噪声和基线频率不尽相同，寻找一个通用的分解方法比较困难，且计算量和计算复杂度也相对较高。曲线拟合是将拉曼信号中的基线通过多项式拟合出来，然后从拉曼光谱中去除。但是，拟合

收稿日期：2019-12-02，**修订日期**：2020-04-08

基金项目：国家自然科学基金项目(21874113，21974118)，国家重大科研仪器研制项目(21627811)资助

作者简介：刘 龙，1989年生，厦门大学航空航天学院仪器与电气系博士研究生 e-mail：944618294@qq.com

*通讯作者 e-mail：xinwang@xmu.edu.cn；fanxg@xmu.edu.cn

阶数不易确定，且容易导致欠拟合或过拟合现象的发生。阶数选择过少，会导致欠拟合；阶数选择过多，会导致过拟合。为实现拉曼信号基线的完美去除，需要对信号进行大量的尝试，计算量相对较大且耗时。本文在不增加实验设备成本的前提下，针对传统基线校正的方法进行了改进，提出一种基于自适应加窗 spline 曲线拟合的拉曼光谱去基线方法，该方法不仅可以够克服拟合阶数不易确定和计算复杂的难题，而且还具备样条曲线平滑去噪的优点。有效地消除拉曼光谱信号的基线漂移，较好的保留一些较弱的拉曼特征峰，为进一步分析光谱数据和实现拉曼成像提供准确可靠的信息。

1 算 法

1.1 spline 曲线拟合原理

spline 曲线拟合是将一些指定点连接成一条光顺曲线，具有样条曲线平滑和计算相对简单的优点，广泛应用于船体和机翼外形设计等对光滑性要求较高的造型中[11-12]。其中，3 次样条函数，不仅有着较高的精度，而且方便操作。在本文中，将其用于拉曼光谱基线的拟合。

待拟合区间$[a, b]$分为 n 段：$a=x_0<x_1<\cdots<x_n=b$，且函数 $S(x)$在$[a, b]$的每一个子区间上是三次多项式。若函数 $S(x)$在节点 x_j 上值等于给定的函数值，则称 $S(x)$是区间$[a, b]$的一个 3 次样条函数，即

$$S(x)=S(x_j)=a_j+b_jx+c_jx^2+d_jx^3,\quad x\in[x_j, x_{j+1}]\ (j=0, 1, \cdots, n-1) \tag{1}$$

其中，共有 $4n$ 个待定系数，且满足如下的条件。

$$\begin{cases} S_j-1(x_j)=S_j(x_j) \\ S'_{j-1}(x_j)=S'_j(x_j) \\ S''_{j-1}(x_j)=S''_j(x_j) \end{cases}\quad (j=1, 2, \cdots, n-1) \tag{2}$$

考虑到本文是实现拉曼光谱的基线拟合，首尾两端处需要具备有连续性和光滑性，所以在首尾两端满足第一种边界条件：给定 $y=f(x)$在端点的一阶导数。

$$\begin{cases} S'(x_0)=f'(x_0) \\ S'(x_n)=f'(x_n) \end{cases} \tag{3}$$

联立式(1)，式(2)和(3)，即可求得三次样条函数 $S(x)$。

1.2 基于自适应加窗 spline 曲线拟合算法

3 次样条函数 $S(x)$具备样条曲线平滑的特点，使用其拟合拉曼光谱基线的同时，又可实现对基线的平滑作用，达到一定的去噪功能。故此，本文基于 3 次样条函数 $S(x)$，提出了基于自适应加窗 spline 曲线拟合算法校正基线，其原理如图 1 所示。利用 3 次 spline 函数拟合算法，通过自适应加窗去基线峰值循环迭代，不断逼近光谱信号基线，原始信号扣线基线后，即可实现基线校正后的光谱。信号处理方法的基本步骤如下：

(1) 输入原始拉曼信号 R(n 维向量)和拉曼峰值搜索的初始步长 step1；

(2) 利用谱峰识别算法，对 R 以 step1 进行初始搜索，得到 R 峰值坐标集合。并以概率统计的方法，估算峰值出现频率最多的频段位置，得到优化的峰值搜索步长 step2。再以 step2 重复上述操作，得到拉曼光谱谷值搜索的优化步长 step3；

(3) 借助谱谷识别算法，对 R 以 step3 进行搜索，得到 R 谷值坐标集合，利用 spline 函数拟合基线 r。再以概率统计的方法，估算谷值出现频率最多的频段位置，得到初始加窗函数的宽度 Win_step；

(4) 再次利用谱峰识别算法，对 r 以 step3 进行搜索，得到 r 峰值坐标集合，并在峰值位置，对称加窗去峰值。为防止出现边缘效应，对 r 两端附近出现的峰值加半窗处理。然后利用 spline 函数拟合基线，得到 r_1。逐点比较 r_1 和 R，取较小的点赋值给 r_1；

(5) 返回步骤(4)继续执行，并重新赋值 step3＝step3/i，Win_step＝Win_step/i，其中 i 是循环次数。直至自适应窗函数宽度 Win_step 低于阈值 Win_min；

(6) 校正后的光谱信号 $R_{correct}=R-r_1$。

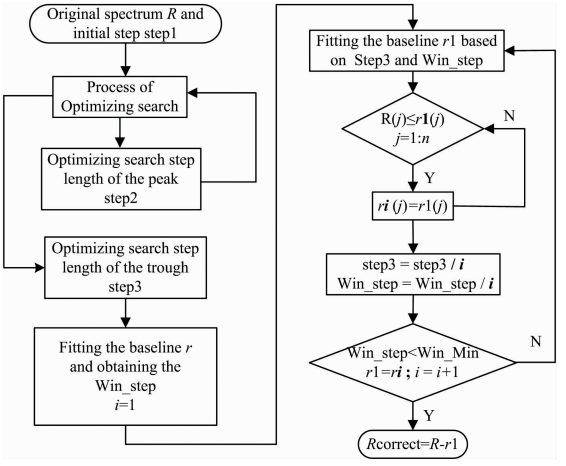

图 1 自适应加窗 spline 曲线拟合去基线流程图

Fig. 1 Process of baseline fitting by adaptive windowed spline fitting

2 实验部分

2.1 材料和仪器

选用乙酸丁酯、PMMA 作为实验样品。实验仪器选用由 QE65Pro，拉曼光纤探头，激光器组成的模块化拉曼系统。其中，叉光纤一端接 785nm 激光器，另一端接光谱仪。

2.2 方法

将实验样品乙酸丁酯，PMMA 分别置于比色皿和自封袋中，设定激光功率为 500 mW，积分时间为 10 s，利用实验室搭建的模块化拉曼系统，完成样品测试，获得原始拉曼光谱数据 R。其中光谱拉曼位移范围为 200～3 300 cm^{-1}，光谱分辨率为 1 cm^{-1}。在拉曼光谱拟合基线前，首先设定合适的拉曼峰值搜索的初始步长 step1。基于谱峰识别算法，经优化搜索得：step1 设定在 40～80cm^{-1}范围最优。本文中选用 step1 为 70 cm^{-1}，完成初始基线 r 的拟合，如图 2 所示。

然后，利用本文算法完成对初始基线的进一步拟合，如图 3 所示。由图 3 可知，拟合基线能够很好的通过原始拉曼信号各谷值点，同时拟合基线在拉曼光谱信号特征峰集中的位置区间能够很好地捕捉到各个特征峰基点；在特征峰分散

的位置区间，又能够很好地与原始光谱信号逼近，且基线整体变化平缓。

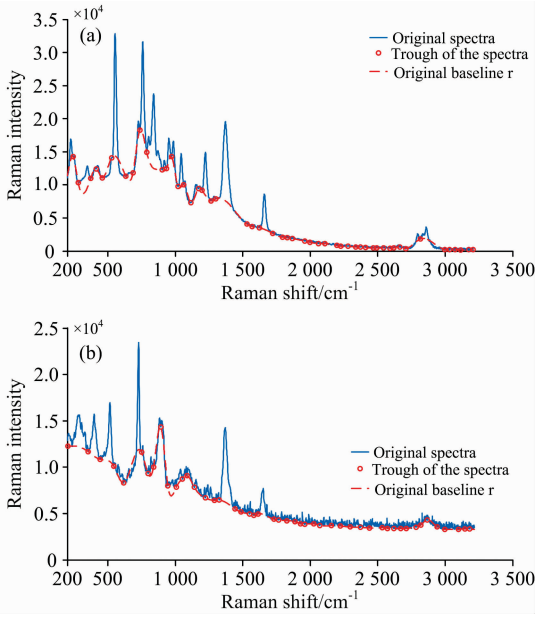

图 2　原始拉曼光谱和初始基线 *r*

(a)：乙酸丁酯；(b)：PMMA

Fig. 2　Original spectra and its original baseline *r*

(a)：*n*-Butyl acetate；(b)：PMMA

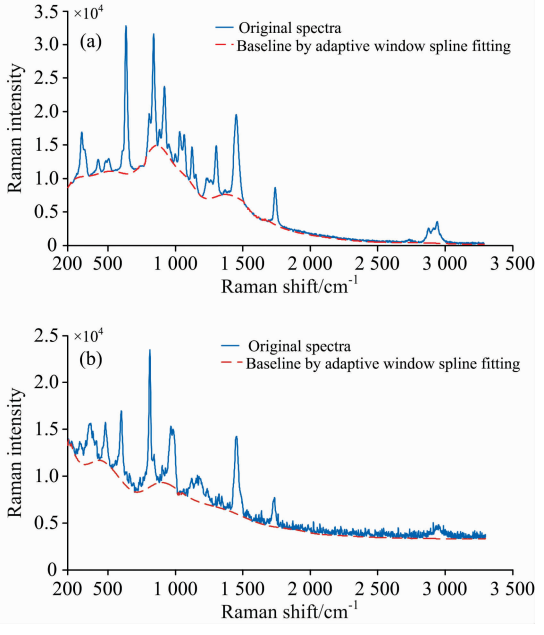

图 3　原始拉曼光谱和自适应加窗 spline 曲线拟合基线

(a)：乙酸丁酯；(b)：PMMA

Fig. 3　Original spectra and its baseline by adaptive window spline fitting

(a)：n-Butyl acetate；(b)：PMMA

最后，基于求得的拟合基线，完成拉曼光谱信号基线的校正，如图 4 所示。由图 4 可知，本文算法校正基线后的拉曼光谱，很好地保留了光谱信号的特征峰段信息。同时，没有出现多余的波峰，且较好的保留一些较弱的拉曼特征峰，可以用于进一步的消噪平滑以及特征峰的识别和匹配。这为本文算法的可行性和良好性能提供了有力的证明。

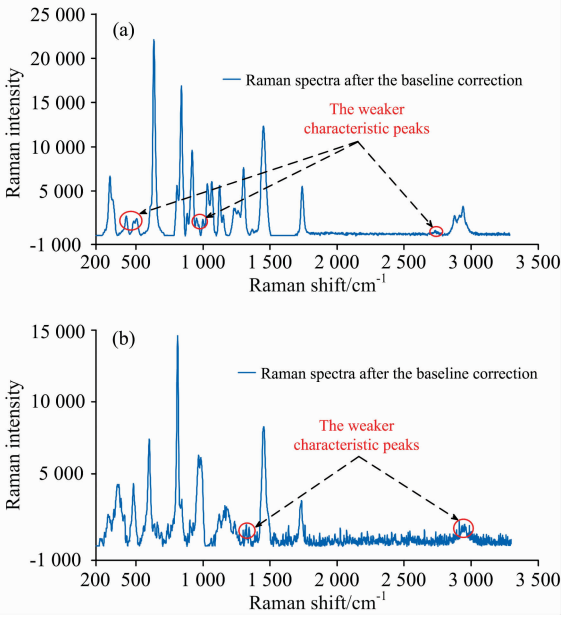

图 4　自适应加窗 spline 曲线拟合去除基线

(a)：乙酸丁酯；(b)：PMMA

Fig. 4　Raman spectra after the baseline correction by adaptive window spline fitting

(a)：n-Butyl acetate；(b)：PMMA

3　结果与讨论

为进一步验证本文算法的良好性能，选取传统多项式拟合方法，零相位高通滤波器[13]和 BEADS[14]算法(Baseline estimation and denoising with sparsity)进行比较，图 5 给出了三种算法校正乙酸丁酯基线的结果。其中，图 5(a)给出了多项式拟合基线的结果。由于多项式阶数对基线拟合结果有较大的影响，本文采用三阶和六阶作为拟合对照组。由图 5(a)可看出，对于样品乙酸丁酯的光谱信号，在使用三阶多项式拟合基线时，在拉曼位移为 2 200 cm^{-1}左右两侧出现了明显的过拟合和欠拟合现象，拟合基线基本上没有通过光谱信号的谷值点；六阶多项式拟合基线时，在一定程度上改善了过拟合现象，但是在拉曼位移为 2 200～3 000 cm^{-1}范围，欠拟合现象反而严重。因此，传统多项式拟合基线需要在阶数上进行优化，然而拟合的阶数又因样品不同会有所差异，造成多项式拟合基线算法的通用性能不佳。

图 5(b)给出了零相位高通滤波拟合基线的结果，其中滤波器设计采用 chebyschev1 型。由于滤波器的阶数和通带波纹 δ 影响拟合结果，本文采用阶数 1，2，通带波纹 δ 为 0.1，0.5 拟合四组基线。由图 5(b)可看出，在拉曼位移 1 600 cm^{-1}左右，基线发生了不同程度的过拟合和欠拟合现象。比较图 5(b)中的拟合基线 1 和 3，2 和 4 可知：阶数相同，δ 越大，在拉曼位移 1 600 cm^{-1}右侧的欠拟合现象有所改善；比

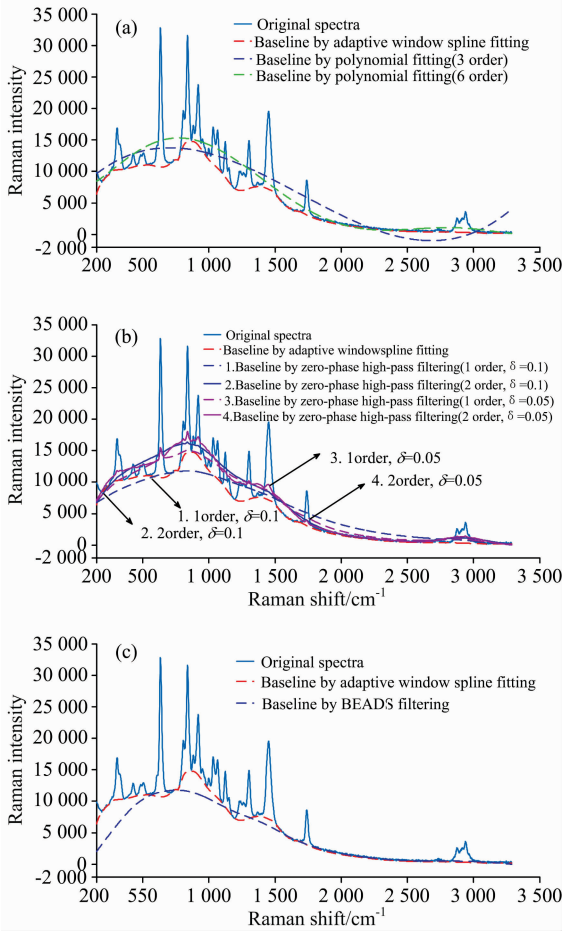

图 5 乙酸丁酯拉曼光谱及其基线

(a)：多项式拟合基线；(b)：零相位高通滤波拟合基线；(c)：BEADS 拟合基线

Fig. 5 Raman spectra of n-Butyl acetate and its baseline

(a): Baseline by polynomial fitting; (b): Baseline by zero-phase high-pass filtering; (c): Baseline by BEADS fitting

较图 5(b)中的拟合基线 1 和 2，3 和 4 可知：阶数越小，δ 相同，在拉曼位移 1 600 cm^{-1}左侧的过拟合现象有所改善。因此，滤波器拟合基线需要在多个参数之间进行耦合优化，计算较为复杂。

图 5(c)给出了 BEADS 拟合基线的结果。BEADS 算法常用于处理色谱信号，但是也可用于处理其他含有基线干扰的信号。这里用以对照，验证本文算法的性能。由图 5(c)可知：BEADS 算法整体拟合良好，仅在拉曼位移 1 170～1 210 cm^{-1}范围，出现了过拟合现象；在拉曼位移 600 cm^{-1}左侧，出现了一定的欠拟合现象。因此，BEADS 算法应用到拉曼光谱信号基线去除时，整体性能良好，但是局部拟合结果有待提高。

综上所述，采用基于自适应加窗 spline 曲线拟合的拉曼光谱去基线方法，充分利用了 spline 函数的光滑特性，克服了传统多项式拟合基线阶数不易确定的缺陷，滤波器拟合基线参数设计复杂的弊端，和 BEADS 算法拟合基线局部性能不佳的瑕疵。同时，在原始光谱信号出现严重基线漂移现象时，本文算法仍能够拟合出光滑的基线，且不易出现欠拟合和过拟合的现象，较好的保留一些较弱的拉曼特征峰，实现了较好的基线校正效果，为进一步分析光谱数据提供准确可靠的信息。

4 结 论

提出了一种基于自适应加窗 spline 曲线拟合的拉曼光谱去基线方法，首先利用谱峰、谱谷识别算法，借助优化搜索步长得到谷值拟合曲线，然后针对拟合曲线峰值自适应加窗去除并利用 spline 样条函数重新拟合基线，最后利用循环迭代的形式，直至窗宽小于阈值，从而实现对拉曼光谱信号基线校正。与传统多项式拟合基线和滤波器拟合基线等相比，本文算法克服了阶数难确定，参数复杂的缺陷，并且整体和局部拟合基线结果较好，通用性能强。同时，对于基线漂移较大的光谱信号，也能够获得较好的校正效果，能够很好地避免欠拟合和过拟合现象。因此，本文提出的算法可以作为一种有效的基线校正方法应用到实际中。

References

[1] Chalmers J M, Uriffiths Peter R. Handbook of Vibrational Spectroscopy (Vol 1-5). Wiley, 2003.
[2] Christoph Krafft, Gerald Steiner, Claudia Beleites, et al. Journal of Biophotonics, 2009, 2: 13.
[3] HU Xue-tao, SHI Ji-yong, LI Yan-xiao, et al(胡雪桃，石吉勇，李艳肖，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(9): 2901.
[4] Arguello C A, Mendes G F, Leite R C. Applied Optics, 1974, 13(8): 1731.
[5] Bright F V, Hieftje G M. Applied Spectroscopy, 1986, 40(5): 583.
[6] Yaney P P. J. Opt. Soc. Am., 1972, 62(11): 1297.
[7] Huang Jie, Shi Tielin, Gong Bo, et al. Applied Spectroscopy, 2018, 72(11): 1632.
[8] Rehman S U, Alkanhal M A S. IET Microwaves Antennas & Propagation, 2019, 13(15): 2693.
[9] Gallo Crescenzio, Capozzi Vito, Lasalvia Maria, et al. Vibrational Spectroscopy, 2016, 83: 132.
[10] ZHU Gao-feng, ZHU Hong-qiu, QIAN Hao, et al(朱高峰，朱红求，钱 灏，等). Chinese Journal of Lasers(中国激光), 2018, 46(2): 211001.
[11] Gonzalez-Vidal J J, Perez-Pueyo R, Soneira M J. Journal of Raman Spectroscopy, 2017, 48(6): 878.

[12] Bertolazzi Enrico, Bevilacqua Paolo, Frego Marco. Mathematics and Computers in Simulation, 2020, 176: 57.
[13] Chambers S D, Choi T, Park S J, et al. Journal of Geophysical Research-Atmospheres, 2017, 122(24): 13525.
[14] Ning X R, Selesnick I W, Duval L. Chemometrics and Intelligent Laboratory Systems, 2014, 139: 156.

Baseline Correction Algorithm for Raman Spectroscopy Based on Adaptive Window Spline Fitting

LIU Long[1], FAN Xian-guang[1,2*], KANG Zhe-ming[1], WU Yi[1], WANG Xin[1,2*]
1. Department of Instrumental and Electrical Engineering, School of Aerospace Engineering, Xiamen University, Xiamen 361005, China
2. Fujian Key Laboratory of Universities and Colleges for Transducer Technology, Xiamen 361005, China

Abstract Raman spectroscopy is a non-destructive and rapid detection technology that can provide qualitative and quantitative information of the material. Therefore, it has been widely used in many fields such as medicine and chemical industry. However, the Raman spectrum suffers from the baseline drift due to the background fluorescence of the sample. Moreover, it has a serious impact on the identification of characteristic peaks of Raman spectra and the Raman imaging. At present, there are two methods to solve this problem, that is, improve the experimental methods and numerical processing. The improve the experimental methods include polarization modulation method and high frequency modulation method. However, they suffer from the disadvantages of complicated experimental equipment and difficult detection technology. The numerical processing includes polynomial fitting and wavelet transform. However, it is prone to suffer from the over and under-fitting. In order to solve this problem, we propose the baseline correction algorithm for Raman spectroscopy based on adaptive window spline fitting, which based on the existing equipment and the traditional baseline correction algorithm. Firstly, the optimal search interval of the trough value is obtained based on the peak recognition algorithm and the initial search step, and then the trough recognition algorithm is used to complete the fitting of the trough curve. Secondly, the peak position of the trough curve is obtained based on the optimal search interval and the peak recognition algorithm. Then, the adaptive rectangular window is symmetrically added at this position, in order to delete the peak, and fitting the trough curve. Thirdly, the fitting trough curve is compared with the original Raman spectrum, point by point, and taking the smaller value to fit a new trough curve. Finally, the operation above will continue until the width of the adaptive window is lower than the threshold. Afterwards, the baseline fitting of the Raman spectrum is completed. And then the baseline correction of the sample is obtained based on our algorithm and the traditional methods. It can be seen that our algorithm can effectively eliminate the baseline drift, and some weaker Raman characteristic peaks can be better remaining. Simultaneously, the over and under-fitting is avoided, and the result of baseline correction is good. Therefore, it provides reliable information on the further analysis of the Raman spectrum and the realization of the Raman imaging.

Keywords Raman Spectroscopy; Baseline drift; Baseline correction; Optimal search

(Received Dec. 2, 2019; accepted Apr. 8, 2020)

* Corresponding authors