



基于中值滤波和非均匀B样条的拉曼光谱基线校正算法

王昕 康哲铭 刘龙 范贤光

Baseline correction algorithm for Raman spectra based on median filtering and un-uniform B-spline

Wang Xin Kang Zhe-Ming Liu Long Fan Xian-Guang

引用信息 Citation: *Acta Physica Sinica*, 69, 200701 (2020) DOI: 10.7498/aps.69.20200552

在线阅读 View online: <https://doi.org/10.7498/aps.69.20200552>

当期内容 View table of contents: <http://wulixb.iphy.ac.cn>

您可能感兴趣的其他文章

Articles you may be interested in

电场条件下氧化锌结晶特性及极化产物的拉曼光谱分析

Crystallization characteristics of zinc oxide under electric field and Raman spectrum analysis of polarized products

物理学报. 2019, 68(7): 070701 <https://doi.org/10.7498/aps.68.20181961>

非贵金属表面增强拉曼散射基底的研究进展

Recent research progress of non-noble metal based surface-enhanced Raman scattering substrates

物理学报. 2020, 69(19): 190701 <https://doi.org/10.7498/aps.69.20200788>

SiC纤维增强Ti17合金复合材料轴向残余应力的拉曼光谱和X射线衍射法对比研究

Comparison between axial residual stresses measured by Raman spectroscopy and X-ray diffraction in SiC fiber reinforced titanium matrix composite

物理学报. 2018, 67(19): 197203 <https://doi.org/10.7498/aps.67.20181157>

混沌信号自适应协同滤波去噪

An adaptive denoising algorithm for chaotic signals based on collaborative filtering

物理学报. 2018, 67(6): 060501 <https://doi.org/10.7498/aps.67.20172470>

外电场极化对纳米氧化锌拉曼活性及气敏性能的影响

Polarization effect of external electric field on Raman activity and gas sensing of nano zinc oxide

物理学报. 2018, 67(14): 140701 <https://doi.org/10.7498/aps.67.20180182>

双光子荧光与相干反斯托克斯拉曼散射显微成像技术的实验研究

Experimental study on two-photon fluorescence and coherent anti-Stokes Raman scattering microscopy

物理学报. 2017, 66(10): 104204 <https://doi.org/10.7498/aps.66.104204>

基于中值滤波和非均匀 B 样条的 拉曼光谱基线校正算法*

王昕¹⁾²⁾ 康哲铭¹⁾ 刘龙¹⁾ 范贤光^{1)2)†}

1) (厦门大学仪器与电气系, 厦门 361005)

2) (福建省高等院校传感器技术重点实验室, 厦门市光电子传感器技术重点实验室, 厦门 361005)

(2020 年 4 月 15 日收到; 2020 年 6 月 19 日收到修改稿)

基线校正是拉曼光谱数据预处理的关键步骤之一, 是消除荧光干扰的有效方法. 传统的多项式拟合和均匀 B 样条拟合算法原理简单、易于实现, 但拟合阶数和内节点的不确定性限制了其灵活性. 因此, 本文提出了一种基于中值滤波和非均匀 B 样条的拉曼光谱基线校正算法. 该算法首先通过平滑预处理、差分计算和设置阈值筛选波谷点, 并根据光谱数据的波谷位置自适应地选择非均匀 B 样条的内节点; 接着利用中值滤波算法对光谱数据进行处理, 使非均匀 B 样条算法能够更好地拟合基线. 该算法克服了传统 B 样条算法需要根据不同的拉曼光谱手动选择内节点的缺点, 同时避免了光谱数据中的随机噪声对基线拟合的影响, 且进一步提高了光谱基线校正效果. 实验结果表明, 该算法能较好地消除拉曼信号基线漂移, 且不存在过拟合和欠拟合现象. 因此, 该算法可以为光谱数据的进一步分析提供更准确、可靠的信息.

关键词: 基线校正, 拉曼光谱, 非均匀 B 样条, 中值滤波, 自适应节点

PACS: 07.50.Qx, 07.07.Df

DOI: 10.7498/aps.69.20200552

1 引言

拉曼光谱由印度物理学家 Raman 于 1919 年从水分子的散射现象中首次发现, 并将其发表于杂志《Nature》上^[1]. 拉曼光谱作为一种用于分析分子化学成分、结构等信息的检测技术, 具有无侵入、特异性高、无标记、无电离辐射、不受水的干扰等优点, 被广泛应用于生物医学、材料生产、化学化工等领域^[2,3]. 然而由于荧光干扰等因素的存在, 基线漂移现象通常存在于光谱信号中, 而基线校正算法是解决该现象的必要手段^[4–7].

目前, 常用的基线校正算法有多项式拟合算法 (polynomial fitting)、B 样条拟合算法 (B-spline fitting)、小波变换 (wavelet transform) 和自适应

迭代重加权惩罚最小二乘 (adaptive iterative reweighted penalty least square, airPLS) 等算法^[8–13]. 多项式拟合算法简单易实现, 但是在处理不同的光谱信号时难以确定拟合阶数. 小波变换则是通过用一系列的正弦波进行叠加来代替每条拉曼光谱, 能够有效地将拉曼光谱的低频和低频进行区分. 但是, 由于不同的光谱具有不同的频率和噪声, 很难找出一个适用于所有拉曼光谱的分解办法, 且其计算量相比其他算法也相对较高. 均匀 B 样条拟合算法在面对基线漂移较为平缓的拉曼光谱信号时, 其基线校正效果良好. 然而由于均匀 B 样条内节点的选择缺乏目的性, 随着基线漂移程度越来越剧烈, 通常导致最终基线校正效果一般. 而非均匀选择内节点, 则需要针对每条光谱信号进行手动选择, 工作量也会随之增加, 且普适性和灵活性较差.

* 国家自然科学基金 (批准号: 21874113, 21974118) 资助的课题.

† 通信作者. E-mail: fanxg@xmu.edu.cn

范贤光等^[14]通过小波变换找到光谱数据的峰值并作为内结点,该方法可以解决内结点选择的问题,但其计算复杂度和计算量较高,且过拟合和欠拟合仍不能完全消除.

因此,本文提出一种基于中值滤波和非均匀 B 样条 (median filtering and un-uniform B-spline, MF-UUB) 的拉曼光谱基线校正算法. 该算法不仅可以根据不同的拉曼光谱信号自适应地选择内结点,且通过使用中值滤波和非均匀 B 样条算法能获得更好的基线校正效果.

2 原理

2.1 非均匀 B 样条理论

B 样条算法因其具有低阶平滑性而广泛应用于曲线和曲面拟合中^[15]. 在本文中,将该算法用于拉曼光谱的基线拟合.

假设输入拉曼光谱数据 \mathbf{r} , 用节点符号 t 将数据点列在 x 轴上进行区间划分, 点序列如下所示:

$$t_{-k+1} \leq \cdots \leq t_{-1} \leq t_0 < t_1 < \cdots < t_N < t_{N+1} \leq t_{N+2} \leq \cdots \leq t_{N+k}, \quad (1)$$

其中 t_1 到 t_N 为内节点, 其余的为外节点. 如果内节点在区间内均匀划分则为均匀 B 样条, 反之, 则为非均匀 B 样条^[16]; 而外节点对算法的影响较小, 其区间大小一般取内节点区间的平均值. N 为区域内内节点数, k 为 B 样条的阶数, 且一般取 4. x 的 k 阶 B 样条曲线如 (2) 式所示:

$$y = \sum_{i=-k+1}^N c_i B_i^k(x), \quad (2)$$

式中 c_i 为控制系数列向量, B 样条曲线的控制多边形就是由这些控制点依次连接而成; $B_i^k(x)$ 为 B 样条基函数, 定义为^[17]:

$$B_i^1(x) = \begin{cases} 1, & t_i \leq x \leq t_{i+1}, \\ 0, & x < t_i, x > t_{i+1}, \end{cases} \quad (3)$$

$$B_i^k(x) = \frac{x-t_i}{t_{i+k-1}-t_i} B_i^{k-1}(x) + \frac{t_{i+k}-x}{t_{i+k}-t_{i+1}} B_{i+1}^{k-1}(x). \quad (4)$$

2.2 局部加权线性回归理论

局部加权线性回归 (locally weighted linear regression, LWLR)^[18] 是线性回归方法的一个改进, 可有效应用于光谱信号的平滑. 因此, 本文通

过使用局部加权线性回归对原始光谱数据进行平滑预处理, 以便后续有效地进行内节点的选取. 它通过预测数据与输入数据的距离赋予不同权重, 距离越近则权重越大, 然后再通过最小二乘法进行计算求解. 该方法的本质是基于加权的均方误差最小化进行求解:

$$\sum_{i=1}^n w_i (r_i - \theta y_i) \rightarrow \min, \quad (5)$$

式中 \mathbf{r} 为输入的光谱数据; \mathbf{y} 为预测数据; θ 为回归系数; w_i 为权重, 该值可通过权重函数计算得到, 一般选用高斯函数作为权重函数, 权重函数如下所示:

$$w_i = \exp\left(-\frac{\|y_i - r_i\|^2}{2\delta^2}\right), \quad (6)$$

式中 δ 为带宽参数.

2.3 自适应选取内节点方法

通过非均匀 B 样条算法进行基线拟合, 其拟合效果的好坏主要取决于原始拉曼光谱的基线能否尽可能地被内节点划分为几个平缓的波段. 因此, 如何根据不同的拉曼光谱自适应地选取内节点, 则为非均匀 B 样条的核心问题. 为保证扣除基线后仍能最大程度的保留原始光谱的特征峰, 且尽可能地在基线变化较为剧烈的位置处选择内节点, 同时有效地将基线划分为几个平缓的波段. 在本文中, 内节点的选取则根据原始光谱信号的波谷点进行确定, 因此, 内节点的选取步骤如下.

1) 使用局部加权线性回归对原始光谱数据进行平滑预处理. 由于光谱数据中伴随着随机噪声, 若直接对光谱数据进行筛选波谷点, 则会导致筛选后的数据点中存在无效数据点. 局部加权线性回归的窗口长度可通过设置窗口数量计算得到, 窗口数量一般取 10^2 个. 原始光谱数据与平滑后的数据如图 1 所示.

2) 筛选光谱数据的波谷点. 对平滑后的光谱信号的每个数据点逐个进行差分计算, 并设置两波谷点间的最小距离阈值 ξ_L , 从而初步筛选出波谷点 (图 1 中红色数据点) 并设为 $\mathbf{K} = [k_0, k_1, \cdots, k_n]$. 最小距离阈值 ξ_L 通过内节点预设值计算得到, 如 (7) 式:

$$\xi_L = \text{Round}\left(\frac{P}{s_{ik}}\right), \quad (7)$$

式中 s_{ik} 为内节点数量预设值, 一般设为 30; P 为原始光谱数据点数; $\text{Round}[\cdot]$ 为四舍五入取整. 将从平滑后的光谱数据中筛选的波谷点 K 对应至原始光谱数据上. 由于对应点并非都准确地位于原始数据的波谷上, 因此对每个数据点进行加窗, 一般窗口距离取 0.2 倍的最小距离阈值 ξ_L , 并求得每个窗口内的最小值. 求得的数据点则为原始光谱数据的波谷点 (图 1 中蓝色数据点), 并设为 $P = [p_0, p_1, \dots, p_n]$.

3) 扣除无效的数据点. 计算两波谷点区间内最大值与最小值之差, 设为 h_i , 并设置最小高度阈值 ξ_H , 其值一般取 0.05 倍的原始光谱数据中最大值与最小值之差:

$$\xi_H = \alpha (\max(r) - \min(r)), \quad (8)$$

式中 α 值一般取 0.05, r 为原始拉曼光谱数据. 若波谷点 p_i 和 p_{i+1} 之间的高度差 h_i 小于阈值 ξ_H , 则扣除数据点 p_{i+1} . 无效的数据点 (图 2 中绿色数据点) 如图 2 所示, 剩余的有效数据点设为 $d = [d_0, d_1, \dots, d_n]$.

4) 插入额外数据点. 计算两数据点 d_i 和 d_{i+1} 间

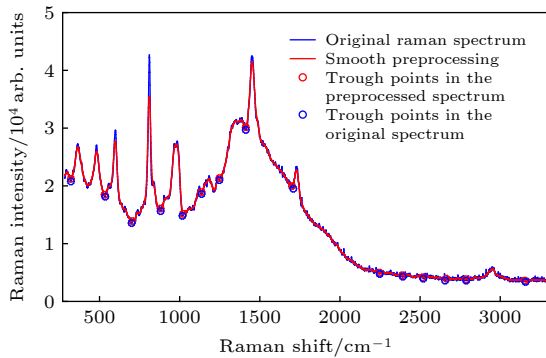


图 1 波谷点选择策略

Fig. 1. Strategy of selection of trough points.

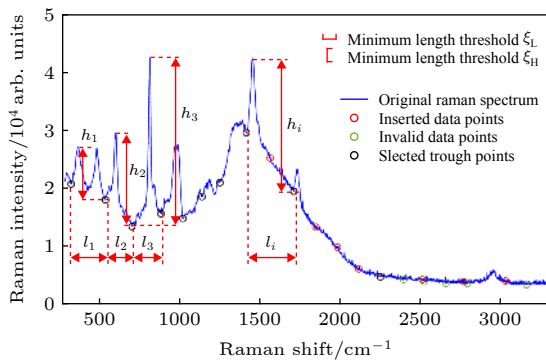


图 2 内节点选择策略

Fig. 2. Strategy of selection of internal knots.

的距离, 并设为 l_i ; 若 l_i 大于内节点距离阈值 (3 倍的最小距离阈值 ξ_L), 则在数据点 d_i 和 d_{i+1} 间插入另一数据点 q_i . 循环执行该步骤, 直至两数据点间的距离均小于内节点距离阈值. 插入的数据点 (图 2 中红色数据点) 设为 $q = [q_0, q_1, \dots, q_n]$, 如图 2 所示.

5) 使用确定的数据点作为非均匀 B 样条的内节点. 数据点 d 和 q 之和得到的数据点 $T = [t_0, t_1, \dots, t_n]$ 则作为内节点 (图 2 中黑色与红色数据点之和).

2.4 中值滤波

中值滤波 (median filtering)^[19] 是一种非线性的平滑滤波技术, 它通过把一串数字序列或者数字图像中的某一个点的值用其相邻区域内所有值的中值进行代替. 在本文中, 利用中值滤波算法对原始拉曼光谱数据进行处理, 使光谱数据中的波峰趋于平缓, 从而优化 B 样条曲线的控制系数列向量 c_i , 以便后续的非均匀 B 样条能在波峰过渡到平滑波段的位置处更好地进行基线拟合. 使其拟合的基线在该位置处变化更加平缓, 能有效地减少欠拟合现象的发生. 中值滤波的公式为:

$$y_j = \text{Med}[x_{j-N}, \dots, x_j, \dots, x_{j+N}], \quad (9)$$

式中 $\text{Med}[\cdot]$ 为取输入数据点列的中值. 此外, 中值滤波的窗口数量越少则拟合的基线越趋于平缓, 但窗口数量过少则会影响拟合基线的准确性. 因此, 窗口数量一般取与内节点数量一致.

2.5 基于中值滤波和非均匀 B 样条的基线校正算法

将内节点选取方法、中值滤波算法和非均匀 B 样条拟合等算法应用于拉曼光谱的荧光背景扣除, 该算法步骤如下:

1) 使用局部加权线性回归对原始光谱数据 r 进行平滑预处理, 得到数据 z ;

2) 对数据 z 进行逐点差分计算和设置最小距离阈值 ξ_L 得到数据 z 的波谷点, 并通过加窗处理得到数据 r 的波谷点 $P = [p_0, p_1, \dots, p_n]$;

3) 利用最小高度阈值 ξ_H 扣除无效波谷点, 并设置内节点距离阈值进行插值, 得到内节点 $T = [t_0, t_1, \dots, t_n]$;

4) 将内节点数量作为中值滤波窗口数量, 对数据 r 进行中值滤波处理, 得到数据 u ;

5) 对数据 u 进行非均匀 B 样条拟合, 得到拟合基线 v ;

6) 计算 u 和 v 的误差绝对值, 若误差值大于所设定的阈值 ξ (阈值 ξ 一般取 0.05), 则将 v 赋值给 u 并回到步骤 4); 误差绝对值公式为

$$\xi = \sum_{i=1}^n \frac{u_i - v_i}{u_i}; \quad (10)$$

7) 从原始拉曼光谱 r 中扣除基线 v .

由上述流程可知, 该算法需要调节的参数仅包含最小距离阈值 ξ_L 和局部加权线性回归窗口数量. 其中最小距离阈值 ξ_L 通过设置内节点数量预设值 s_{ik} 计算得到. 且内节点数量预设值 s_{ik} 一般取 30 个, 局部加权线性回归窗口数量一般取 10^2 个即可获得良好的基线校正效果.

3 实验与讨论

3.1 实验材料

本文选用聚甲基丙烯酸甲酯 (polymethyl methacrylate, PMMA) 和正辛烷 (N-OCTANE) 作为待测样品. N-OCTANE 拉曼光谱在光谱仪积分时间为 1 s 下获得, PMMA 拉曼光谱在光谱仪积分时间为 10 s 下获得.

3.2 实验结果

图 3(a),(b) 分别是 N-OCTANE 和 PMMA 的原始拉曼光谱, 其中 N-OCTANE 拉曼光谱的基线漂移程度较小, 而 PMMA 拉曼光谱的基线漂移程度则较为剧烈, 尤其在拉曼位移为 1500 cm^{-1} 处, 基线的存在会淹没原本就微弱的谱峰, 这些现象使得光谱信号难以被直观地辨识. 通过本文所提的 MF-UUB 基线校正算法对两种样品的拉曼光谱分别进行基线拟合, 得到的拟合基线同样如图 3 所示, 基线校正后的拉曼光谱如图 4 所示. 其中, 局部加权线性回归的窗口数量设为 10^2 个, 内节点数量预设值设为 30 个, 因此对于两条拉曼光谱分别得到的内节点数量为 14 个和 17 个; 中值滤波的窗口数量分别与各自光谱数据的内节点数量一致. 由图 3 可知, 该算法能有效地随波峰的变化进行基线拟合, 且在线性波段和曲线波段的拟合中均有较好的拟合效果, 这证明了 MF-UUB 算法的灵活性. 由图 4 可知, 基线校正后的拉曼光谱能有效地保留

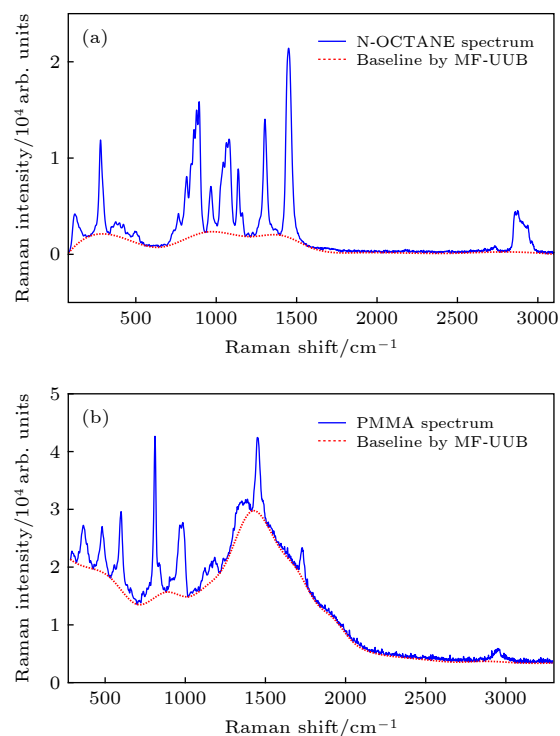


图 3 原始拉曼光谱和基于 MF-UUB 算法拟合的基线 (a) N-OCTANE 拉曼光谱; (b) PMMA 拉曼光谱

Fig. 3. Original Raman spectra and fitted baseline by MF-UUB: (a) N-OCTANE Raman spectrum; (b) PMMA Raman spectrum.

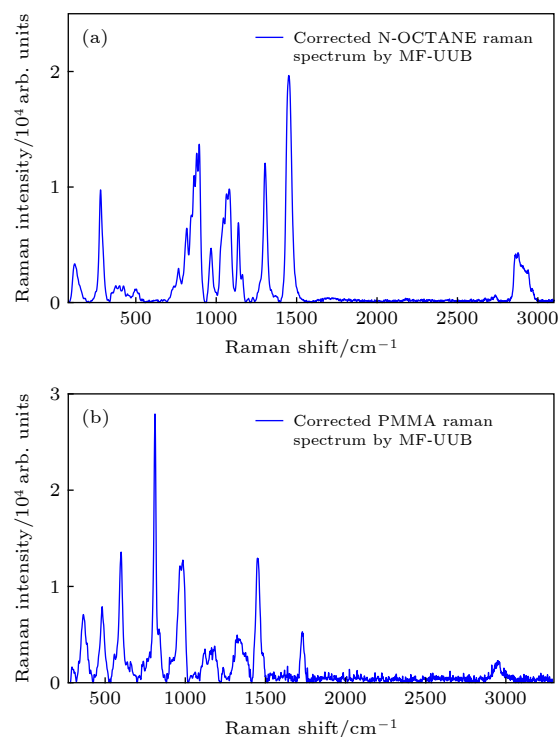


图 4 基于 MF-UUB 算法基线校正后的光谱 (a) N-OCTANE 拉曼光谱; (b) PMMA 拉曼光谱

Fig. 4. Corrected Raman spectra by MF-UUB: (a) N-OCTANE Raman spectrum; (b) PMMA Raman spectrum.

所有特征峰信息,且没有因为荧光背景的扣除而出现多余的无效特征峰或凸包.对于大部分的拉曼光谱,使用固定的参数设置,便能获得较好的基线拟合效果.在后续对拉曼信号的进一步分析中,如拉曼组分识别的应用,识别的过程主要根据谱峰的位置等特点进行识别,而若无法有效且完整地扣除荧光背景,将使得某些谱峰的位置发生偏移,或在一定程度上改变特征峰形状及大小,这将影响对物质的有效识别.而本文所提算法则能有效地避免上述问题,能为后续的进一步分析提供更可靠有效的信息.

3.3 中值滤波算法有效性验证

为验证中值滤波算法是否能有效地提升拟合基线的准确性,通过使用不加中值滤波处理的非均匀 B 样条拟合算法进行基线拟合,拟合的基线如图 5 所示.由图 5(a)可知,拟合的基线与本文所提算法拟合的基线在整体上基本一致.但是在波峰过渡到平滑波段的位置,如拉曼位移为 $1400\text{--}2000\text{ cm}^{-1}$ 等处,拟合的基线存在一定的欠拟合现象.而通过

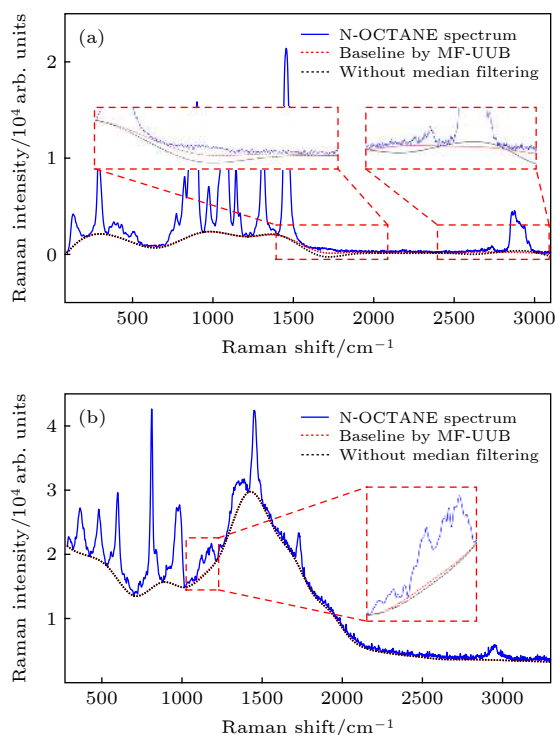


图 5 原始拉曼光谱和基于非均匀 B 样条算法拟合的基线 (a) N-OCTANE 拉曼光谱; (b) PMMA 拉曼光谱

Fig. 5. Original Raman spectra and fitted baseline by non-uniform B-spline algorithm: (a) N-OCTANE Raman spectrum; (b) PMMA Raman spectrum.

使用中值滤波算法处理后,则可以有效地改善这一缺陷.这是因为通过中值滤波处理后的拉曼光谱的波峰会趋于平缓,使得拟合的基线变化幅度变小,从而能更好地在波峰过渡到平滑波段的位置处进行拟合.

此外,如果通过非均匀 B 样条拟合算法便能获得较好的拟合基线,则中值滤波算法不会对拟合的基线有太大的改变.由图 5(b)可知,该光谱的基线拟合效果好于图 5(a),且不存在欠拟合现象.因此,通过中值滤波进行处理后再进行基线拟合的效果则提升相对有限,仅在拉曼位移为 $1000\text{--}1300\text{ cm}^{-1}$ 处存在一定程度的改善.

3.4 均匀 B 样条算法、多项式拟合算法和 airPLS 的基线校正情况

为进一步验证该算法能有效地进行基线拟合且优于传统算法,分别使用多项式拟合算法和均匀 B 样条拟合算法对 N-OCTANE 和 PMMA 原始拉曼光谱进行基线拟合.此外,由于多项式拟合算法的拟合阶数难以确定,因此同时选用 5 阶、7 阶和 9 阶多项式拟合算法对原始光谱进行基线拟合.而均匀 B 样条拟合算法则采用 4 阶 3 次 B 样条,且为保证验证的有效性,均匀 B 样条的内节点数与本文所提算法保持一致,分别设为 14 个和 17 个.基于均匀 B 样条算法和多项式拟合算法拟合的基线分别如图 6 和图 7 所示.

由图 6 可知,均匀 B 样条拟合算法在整体上能有效地进行基线拟合.但由于内节点的选择缺乏目的性,无法有效地在正确位置处将光谱数据的基线划分为几个平滑波段,这导致拟合的基线存在欠拟合现象,如图 6(a) 中拉曼位移为 $300\text{--}700\text{ cm}^{-1}$ 以及图 6(b) 中拉曼位移为 $1000\text{--}1200\text{ cm}^{-1}$ 等多处波段中.由于无法有效且完整地扣除荧光背景,将导致扣除荧光背景后的拉曼光谱在某些波段处出现无效的凸包,且在一定程度上改变特征峰的形状.

由图 7 可知,多项式拟合算法在 N-OCTANE 原始拉曼光谱的基线拟合中表现效果好于 PMMA 原始拉曼光谱,能相对有效地进行基线拟合,但整体上仍存在着一定的拟合缺陷.随着多项式拟合阶数的增加,拟合的基线在某些波段处会得到改善,如图 7(a) 中拉曼位移为 $700\text{--}1200\text{ cm}^{-1}$ 以及图 7(b) 中拉曼位移为 $1100\text{--}1800\text{ cm}^{-1}$ 处.但同时随着拟合阶数的增加,拟合的基线在某些波段处的拟合效果也会变差,如图 7(a) 中拉曼位移为 $1200\text{--}1600\text{ cm}^{-1}$

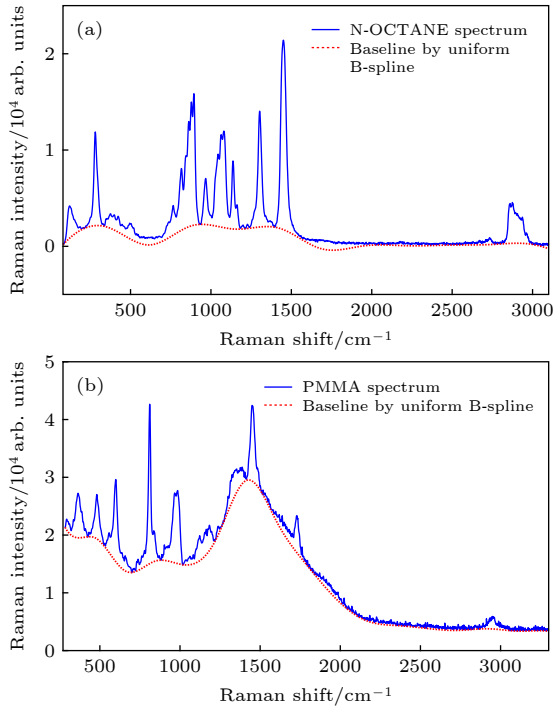


图6 原始拉曼光谱和基于均匀B样条算法拟合的基线
(a) N-OCTANE 原始拉曼光谱; (b) PMMA 原始拉曼光谱
Fig. 6. Original Raman spectra and fitted baseline by uniform B-spline algorithm: (a) N-OCTANE Raman spectrum; (b) PMMA Raman spectrum.

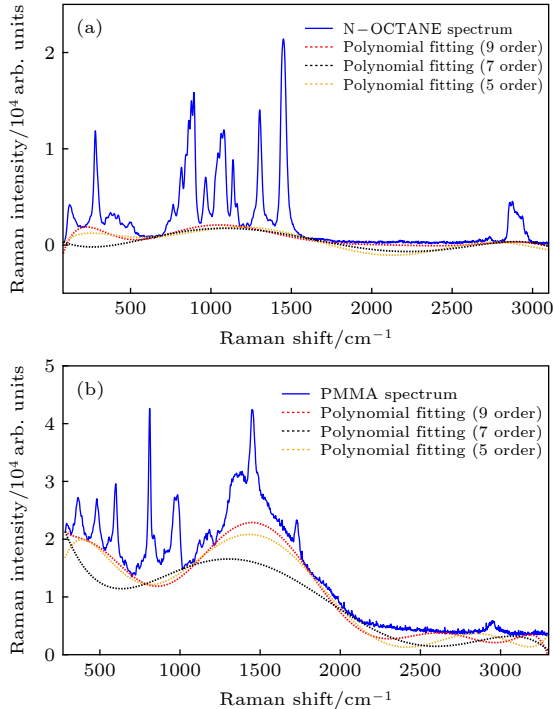


图7 原始拉曼光谱和基于多项式拟合算法拟合的基线
(a) N-OCTANE 拉曼光谱; (b) PMMA 拉曼光谱
Fig. 7. Original Raman spectra and fitted baseline by polynomial fitting algorithm: (a) N-OCTANE Raman spectrum; (b) PMMA Raman spectrum.

以及图 7(b) 中拉曼位移为 800—1100 cm^{-1} 处. 因此可以得出, 多项式拟合算法的阶数难以确定, 且欠拟合和过拟合现象难以控制和预料. 针对基线漂移较为剧烈的原始拉曼光谱, 通过多项式拟合算法拟合的基线效果较为一般, 无法有效且完整地扣除荧光背景.

此外, airPLS 是一种应用广泛且效果良好的基线校正算法^[20], 因此, 通过该算法拟合的基线如图 8 所示. 由图 8 可知, 该算法整体基线拟合效果良好, 在大部分波段处均能准确有效地进行基线拟合, 且拟合的基线与通过本文所提算法拟合的基线基本一致. 但是, 仍然有一些拟合缺陷产生, 如拟合的基线通过波谷点时存在着一定的过拟合现象.

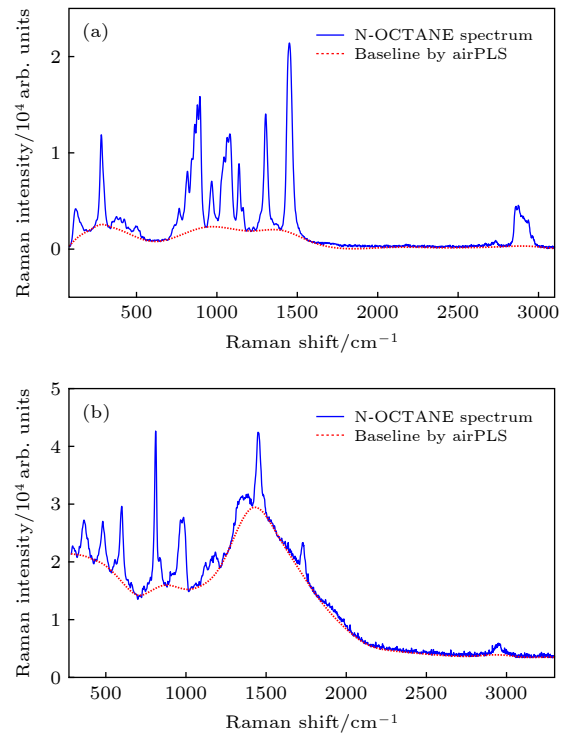


图8 原始拉曼光谱和基于 airPLS 拟合的基线 (a) N-OCTANE 拉曼光谱; (b) PMMA 拉曼光谱
Fig. 8. Original Raman spectra and fitted baseline by airPLS algorithm: (a) N-OCTANE Raman spectrum; (b) PMMA Raman spectrum.

综上所述, 本文提出 MF-UUB 算法能有效地进行基线校正. 与多项式拟合算法相比, 本文所提算法能以固定的阶数进行拟合; 与均匀 B 样条算法相比, 本文所提算法能根据不同的拉曼光谱自适应选择内节点; 与非均匀 B 样条算法和 airPLS 相比, 本文所提算法具有更好的基线拟合效果.

4 结 论

本文提出了一种简单灵活且有效的拉曼光谱基线校正算法. 该方法能自适应地根据不同光谱数据选取内节点, 且通过中值滤波算法优化 B 样条曲线控制系数, 使非均匀 B 样条算法对拉曼光谱数据进行拟合时能获得更好的基线拟合效果. 与多项式拟合, 均匀 B 样条和 airPLS 算法相比, 该算法具有更好的基线校正效果, 且不存在欠拟合和过拟合等缺陷. 此外, 该算法具有广泛的适用性和灵活性, 针对不同的基线漂移情况, 可以有效地进行基线拟合. 因此, 本文所提算法能为拉曼光谱信号的进一步分析提供更准确可靠的信息.

参考文献

- [1] Raman C V 1928 *Nature* **121** 619
- [2] Morris M D 2006 *Anal. Chem.* **78** 33
- [3] Geiman I, Leona M, Lombardi J R 2009 *J. Forensic Sci.* **54** 947
- [4] Liu H, Zhang Z L, Liu S Y, Yan L X, Liu T T, Zhang T X 2015 *Appl. Spectrosc.* **69** 1013
- [5] Cadusch P J, Hlaing M M, Wade S A, Mcarthur S L 2013 *J. Raman Spectrosc.* **44** 1587
- [6] Zhang R, Zhao X H, Hu Y J, Guo Y, Wang Z, Zhao Y, Li Z X, Wang Y 2014 *Acta Phys. Sin.* **63** 070702 (in Chinese) [张锐, 赵学珏, 胡雅君, 郭媛, 王喆, 赵迎, 李子晓, 汪曦 2014 物理学报 **63** 070702]
- [7] Pang Y, Deng L, Lin J Z, Li Z Y, Zhou Q N, Li G Q, Huang H W, Zhang Y, Wu W 2014 *Acta Phys. Sin.* **63** 098701 (in Chinese) [庞宇, 邓璐, 林金朝, 李章勇, 周前能, 李国权, 黄华伟, 张懿, 吴炜 2014 物理学报 **63** 098701]
- [8] Perez-Pueyo R, Soneira M J 2010 *Appl. Spectrosc.* **64** 595
- [9] Shao L M, Griffiths P 2007 *Environ. Sci. Technol.* **41** 7054
- [10] Zhao J H, Lui H, Mclean D I, Zeng H S 2007 *Appl. Spectrosc.* **61** 1225
- [11] Wang W P, Pottmann H, Liu Y 2006 *ACM Graphic.* **25** 214
- [12] Cai Y Y, Yang C H, Xu D G, Gui W H 2018 *Anal. Methods.* **10** 3525
- [13] Zhang Z M, Chen S, Liang Y Z 2010 *J. Raman Spectrosc.* **41** 659
- [14] Fan X G, Wang H T, Wang X, Xu Y J, Wang X F, Que J 2016 *Spectrosc. Spec. Anal.* **36** 724 (in Chinese) [范贤光, 王海涛, 王昕, 许英杰, 王秀芬, 阙靖 2016 光谱学与光谱分析 **36** 724]
- [15] Martin T, Cohen E, Kirby R M 2009 *Comput. Aided Geom. D.* **26** 648
- [16] Wang X, Fan X G, Xu Y J 2015 *Meas. Sci. Technol.* **26** 115503
- [17] Wang X, Fan X G, Xu Y J, Wu J L, Lian J, Zuo Y 2014 *Spectrosc. Spec. Anal.* **34** 2117 (in Chinese) [王昕, 范贤光, 许英杰, 吴景林, 梁骏, 左勇 2014 光谱学与光谱分析 **34** 2117]
- [18] Lu J D, Cuan K X, Zhang W F 2018 *Spectrosc. Spec. Anal.* **38** 3708 (in Chinese) [卢德俊, 爨凯旋, 张伟峰 2018 光谱学与光谱分析 **38** 3708]
- [19] Juhola M, Katajainen J, Raita T 1991 *IEEE T. on Signal Proces.* **39** 204
- [20] Zhang Z M, Chen S, Liang Y Z 2010 *Analyst* **135** 1138

Baseline correction algorithm for Raman spectra based on median filtering and un-uniform B-spline*

Wang Xin¹⁾²⁾ Kang Zhe-Ming¹⁾ Liu Long¹⁾ Fan Xian-Guang^{1)2)†}

1) (*Department of Instrumental and Electrical Engineering, Xiamen University, Xiamen 361005, China*)

2) (*Xiamen Key Laboratory of Optoelectronic Transducer Technology, Fujian Key Laboratory of Universities and Colleges for Transducer Technology, Xiamen 361005, China*)

(Received 15 April 2020; revised manuscript received 19 June 2020)

Abstract

As one of the key steps for data preprocessing of Raman spectra, baseline correction is an effective method to eliminate fluorescence interference. Traditional algorithms such as polynomial fitting and uniform B-spline fitting are simple and easy to implement, but the uncertain fitting order and internal knots limit their flexibility. In addition, the baseline correction results of traditional algorithms often occur over and under fitting phenomena. Therefore, we propose a baseline correction algorithm for Raman spectra based on median filtering and un-uniform B-spline. Firstly, the trough points of the spectral data are filtered by smoothing preprocess, difference calculation and threshold setting, and the internal knots of the un-uniform B-spline are adaptively selected by the trough positions of the spectral data. Then, the median filtering algorithm is used to process the spectral data so that the un-uniform B-spline has a better baseline fitting effect at the position where the signal changes from peak to smooth band. Finally, the un-uniform B-splines is used to fit the baseline by fitting the baseline iteratively. The proposed algorithm overcomes the shortcoming of traditional B-spline algorithm that the internal knots need to be selected manually based on different Raman spectra, and also avoids influencing the baseline fitting by random noise in the spectral data, and thus further improving the spectral baseline correction effect. The original Raman spectra of polymethyl methacrylate and normal octane are used for experimentally evaluating the baseline correction effect. Compared with the results from polynomial fitting, uniform B-spline and adaptive iteratively reweighted penalized least squares algorithms, the experimental results show that the proposed algorithm can well eliminate the Raman signal baseline drift effectively without over or under fitting phenomena, and it can perform better baseline correction for different baseline drift situations. Therefore, the proposed algorithm can provide more accurate and reliable information for the further analysis of spectral data.

Keywords: baseline correction, Raman spectrum, un-uniform B-spline, median filtering, adaptive knots

PACS: 07.50.Qx, 07.07.Df

DOI: 10.7498/aps.69.20200552

* Project supported by the National Natural Science Foundation of China (Grant Nos. 21874113, 21974118).

† Corresponding author. E-mail: fanxg@xmu.edu.cn