基于广义Whittaker平滑器的拉曼光谱基线校正方法

杨桂燕 李 路 陈 和* 陈思颖 张寅超 郭 磐

北京理工大学光电学院, 北京 100081

摘要 拉曼光谱的频率、强度及偏振特性表征散射物质的独特性质,是研究物质结构及组成成分的特征光谱,因而得到了广泛应用。但是,基线漂移现象会给拉曼光谱的定量分析带来不利影响。为了校正拉曼光谱基线漂移,提出了一种结合导数谱峰检测与Whittaker平滑器的基线校正算法。利用拉曼二阶导数光谱检测并标定谱峰区域;Whittaker平滑器结合标定信息计算非谱峰区域的拟合曲线,并同时对谱峰区域进行平滑插值,最终得到整个光谱的基线估计。将该算法应用于模拟和实际拉曼光谱进行基线校正,结果表明,算法可以同时实现光谱去噪与基线估计,而主成分分析结果的改善进一步验证了该算法的有效性。

关键词 光谱学; 基线校正; Whittaker平滑器; 拉曼光谱; 二阶导数谱; 主成分分析

中图分类号 O433.4 **文献标识码** A

doi: 10.3788/CJL201542.0915003

Baseline Correction Method for Raman Spectra Based on Generalized Whittaker Smoother

Yang Guiyan Li Lu Chen He Chen Siying Zhang Yinchao Guo Pan

School of Opto-Electronics, Beijing Institute of Technology, Beijing 100081, *China*

Abstract Raman spectrum is the characteristic spectrum for material structure and composition research as its frequency, intensity and polarization can characterize the peculiar properties of the scatterer, and thus it has been widely applied in many fields. However, the baseline drifting often occurs in the Raman spectrum and adversely affects its quantitative analysis. In order to eliminate the baseline drifting, a correction algorithm is presented, which combines the peak detection method based on derivative spectrum and the Whittaker smoother for baseline estimation. The spectral peak region is detected and identified utilizing the second derivatives. Then the Whittaker smoother calculates the curve fitting of the non-peak region combining the identification information, and interpolates the peak region baseline smoothly at the same time. The whole baseline estimation of the spectrum is obtained. The algorithm has been applied to simulated and actual Raman signals to correct the baseline drifting. The results show that the algorithm realizes spectral denoising and baseline estimation simultaneously. The improved analytical results of principal component analysis also verify its effect.

Key words spectroscopy; baseline correction; Whittaker smoother; Raman spectra; second derivative spectra; principal component analysis

OCIS codes 300.6450; 200.4560; 300.2530

1 引 言

拉曼光谱是一种通过测量分子振动、转动引起的散射光谱来表征物质分子结构的分析技术,具有快速、简单、可重复以及无损伤、无电离辐射的优点,因此在物理、化学、生物、考古等多个领域得到了广泛的应用[1-4]。然而,在光谱采集过程中由于样品粒度、环境温度、外界振动以及仪器自身工作状况的影响,拉曼光谱仪测

收稿日期: 2015-04-13; **收到修改稿日期**: 2015-05-07

基金项目: 总装预研基金(513210604)

作者简介: 杨桂燕(1989—),女,硕士研究生,主要从事激光光谱技术方面的研究。E-mail: xgy0910yi@163.com

导师简介: 张寅超(1962—),男,博士,教授,主要从事激光雷达技术和激光光谱技术等方面的研究。E-mail: ychang@bit.edu.cn

*通信联系人。E-mail: shinianshao@gmail.com

得的光谱常常会发生基线漂移现象，这样就不能准确提取有效光谱信息，从而限制了定性和定量分析的效果[5]。因此在分析拉曼光谱之前，必须先进行基线校正以消除背景谱的影响，从而提高分析模型的预测精度。传统的基线校正大多是利用最小二乘多项式拟合方法对光谱信号的手动选点进行曲线拟合[6]，该方法简单有效，但是容易出现过拟合或欠拟合现象，而且需要定义合适的拟合阶数，具有较大的主观性和时间消耗。惩罚最小二乘法是一种运算快速、连续可控、具有自动插值功能且易于进行交叉验证的信号平滑方法[7]，不仅考虑了最小二乘拟合信号对原始信号的保真度，同时也兼顾了拟合信号的平滑度。该方法首先由Whittaker[8]于1922年提出，随后英国统计学者Silverman[9]在发展粗糙度惩罚方法的研究中，对其理论框架进行了完善。在化学数据分析方面，Eilers等[10-11]较早地将该方法应用于对间断数据的处理上，以实现丢失数据的自动连续插值。其提出的基线校正方法是先通过迭代修正权重因子逐渐消除谱峰的影响，继而再应用Whittaker平滑器估计光谱信号的基线背景。后来，Zhang等[7]基于Eilers的工作也提出了一种自适应循环加权的惩罚最小二乘法，该方法修改了Eilers的权重因子表达式，能够进一步提高基线的估计精度。

在光谱分析中，导数光谱技术由于能够增强诸如肩峰等微弱光谱的检测，常常用来提高光谱分辨率，降低背景干扰，并提供比原始光谱更易分辨的光谱特征[12]。自20世纪50年代提出以来，导数光谱得到了越来越多的研究，相关程序也已被现代光谱分析仪器广为采用[13]。特别是二阶导数光谱，其峰谷位置恰对应于原谱的中心位置。因此，不同于上述循环迭代方法，本文提出的拉曼光谱基线校正方法是利用原始数据的二阶导数谱来检测谱峰分布区间，再应用修正的Whittaker平滑器估计整个光谱的基线，最后通过模拟信号和实际拉曼光谱的处理实验，验证了该方法的有效性。

2　Whittaker平滑器基本原理

设 $\boldsymbol{y}$ 为待分析均匀采样的光谱信号，$\boldsymbol{f}$ 为平滑信号，则惩罚最小二乘法的目标函数为

$$\boldsymbol{f}=\underset{f}{\operatorname{argmin}}\left[\sum_i (y_i-f_i)^2+\lambda\sum_i(\Delta^k f_i)^2\right], \tag{1}$$

式中 $\sum_i (y_i-f_i)^2$ 为一般最小二乘法的目标函数，反映了拟合信号 $\boldsymbol{f}$ 对原始信号 $\boldsymbol{y}$ 的保真度；$\lambda\sum_i(\Delta^k f_i)^2$ 为惩罚项，采用的是Tikhonov正则化方法[14]，反映了拟合信号 $\boldsymbol{f}$ 的平滑度。其中，Δ^k 为 k 阶微分算子，$\lambda(>0)$ 为正则参数，用于平衡平滑信号的保真度和平滑性。为计算简便，引入矩阵代数，则(1)式变为

$$\boldsymbol{f}=\underset{f}{\operatorname{argmin}}\left[(\boldsymbol{y}-\boldsymbol{f})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{f})+\lambda(\boldsymbol{D}_k\boldsymbol{f})^{\mathrm{T}}(\boldsymbol{D}_k\boldsymbol{f})\right], \tag{2}$$

进而求得最小二乘解为

$$\boldsymbol{f}=\left(\boldsymbol{I}+\lambda\boldsymbol{D}_k^{\mathrm{T}}\boldsymbol{D}_k\right)^{-1}\boldsymbol{y}, \tag{3}$$

式中 $\boldsymbol{I}$ 为单位矩阵，$\boldsymbol{D}_k$ 为替代微分算子 Δ^k 的 k 阶差分矩阵以用于数值计算。记 $\boldsymbol{L}=\left(\boldsymbol{I}+\lambda\boldsymbol{D}_k^{\mathrm{T}}\boldsymbol{D}_k\right)^{-1}$，则(3)式可写为 $\boldsymbol{f}=\boldsymbol{L}\boldsymbol{y}$，由傅里叶变换分析可知，$\boldsymbol{L}$ 具有低通滤波器特性，λ 取值越大，滤波器带宽越窄，输出信号越光滑，因此称 $\boldsymbol{L}$ 为Whittaker平滑器[10,15-16]。

对于有元素缺失的不连续数据片段，Eilers[10]通过引入权重矢量 $\boldsymbol{w}$，得到了广义化的Whittaker平滑器，从而可以对缺失数据实现自动平滑性插值。权重矢量 $\boldsymbol{w}$ 由0或1元素构成，数据缺失片段对应的权重值为0，其他有效片段对应的值为1，则(1)式和(3)式分别变为

$$\boldsymbol{f}=\underset{f}{\operatorname{argmin}}\left[\sum_i w_i(y_i-f_i)^2+\lambda\sum_i(\Delta^k f_i)^2\right], \tag{4}$$

$$\boldsymbol{f}=\left(\boldsymbol{W}+\lambda\boldsymbol{D}_k^{\mathrm{T}}\boldsymbol{D}_k\right)^{-1}\boldsymbol{W}\boldsymbol{y}, \tag{5}$$

式中 $\boldsymbol{W}=\mathrm{diag}(\boldsymbol{w})$ 是以 w_i 为对角元素的对角阵。分析(4)式可知，该方法实质如下：对于有效测量的数据片段，要求拟合信号相对原始数据同时具有保真度和平滑度；但对于未知的缺失片段，只能对拟合信号提出平滑度要求，而最终的效果则是对缺失部分自动进行了连续且平滑的插值。若数据虽无缺失但却发生了突变，可仍将突变部分权重置为0，则其输出效果与数据缺失情况相同(图1)。类似地，记 $\boldsymbol{M}=\left(\boldsymbol{W}+\lambda\boldsymbol{D}_k^{\mathrm{T}}\boldsymbol{D}_k\right)^{-1}\boldsymbol{W}$，则称 $\boldsymbol{M}$ 为广义Whittaker平滑器。

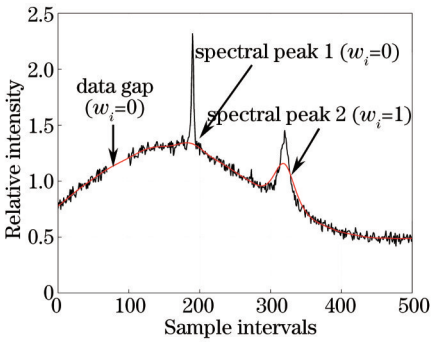

图1 广义Whittaker平滑器对不连续数据的拟合效果

Fig.1 Fitting effect of generalized Whittaker smoother for discontinuous data

图1中，黑色曲线表示测量的不连续数据(包括缺失和突变)，红色曲线是应用广义Whittaker平滑器对测量数据的拟合结果。由此可见，平滑器对连续数据区间(权重置1)进行了平滑，而对不连续区间(权重置0)则进行了连续性插值，本文提出的拉曼光谱基线校正方法正是利用了这一特性。另外还要注意，如果突变区域的权重仍置为1，则Whittaker平滑器会输出此区域数据的拟合结果，实际就是平滑结果，并不是为了保持连续性而根据区域前后数据输出的插值结果。

3　基于Whittaker平滑器的基线校正算法

根据拉曼光谱特征，若将其基线部分视为连续信号，谱峰区域视为基线突变信号，则由广义Whittaker平滑器特性可知，只要检测出光谱谱峰位置区域，并将其权重置为0，即可应用(5)式估计其整体光谱基线。根据这一思想，提出了一种基于广义Whittaker平滑器的基线校正算法(GWSBLC)。

3.1 二阶导数谱光谱检测方法

在GWSBLC方法中，准确检测出光谱谱峰区域位置是进行基线校正的关键，Eilers[11]和Zhang等[7]都是通过循环迭代的方法给谱峰设置很小的权重，逐渐将峰区数值拉低到与基线同等量级。在这类方法中，设计合理的权重修正方案非常重要，否则，即便循环迭代已满足终止条件，谱峰仍作为突变信号给基线估计带来干扰。与上述循环方法不同，这里采用原始光谱的二阶导数谱检测出谱峰分布区域，这主要是利用了导数谱不受基线背景影响、更能反映光谱特征的性质。

关于在基线校正中利用导数谱检测谱峰位置，已有一阶导数谱的研究报导，其主要依据就是一阶导数谱的零点位于谱峰及左右边界，并且谱型具有左正右负的特点[17-18]。然而，实际信号很容易受噪声干扰而不能准确找到零点；另外，一阶导数谱的峰-谷位置与原始光谱宽度也没有必然联系(图2)。因此采用二阶导数谱来构建谱峰检测模型，设拉曼光谱线型由Lorentzian函数表示：

$$L(\omega)=\frac{\gamma/2\pi}{(\omega_0-\omega)^2+\gamma^2/4}, \tag{6}$$

则其二阶导数为

$$\frac{\mathrm{d}^2L}{\mathrm{d}\omega^2}=\frac{\gamma}{\pi}\cdot\frac{3(\omega_0-\omega)^2-\gamma^2/4}{\left[(\omega_0-\omega)^2+\gamma^2/4\right]^3}, \tag{7}$$

式中ω_0为光谱谱峰中心位置，γ为半峰全宽(FWHM)。

如图2所示，二阶导数谱的两个局部极大值峰点对应于原谱的半峰全宽点，而局部极小值谷点则与原谱谱峰中心位置重合。因此，原始光谱的中心位置及半峰全宽可以用二阶导数谱的峰-谷-峰区域确定。另外引入展宽因子α，将二阶导数谱的峰-谷-峰区域向两侧延展，使其频率区间包括原谱90%以上的信息，即可得到谱峰检测结果。

3.2 拉曼光谱基线校正步骤

根据以上论述，GWSBLC基线校正算法的操作如下：

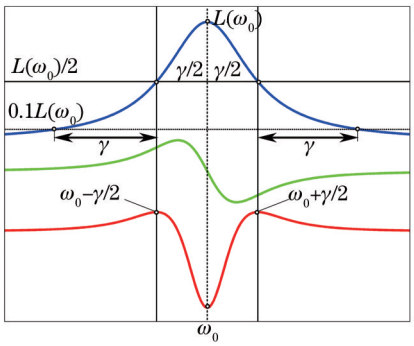

图 2　Lorentzian 光谱线型及其一阶、二阶导数谱

Fig.2　Lorentzian spectral profiles and its first-, second-order derivatives

1) 载入原始测量光谱数据 $\boldsymbol{y}$ 并对其进行平滑，预先减少噪声对后续导数谱的影响。为了体现程序复用性，采用(5)式所示的 Whittaker 平滑器，其中 $\boldsymbol{w}=1$，$\boldsymbol{D}_k=\boldsymbol{D}_2$，可以调节正则参数 λ_a 来控制平滑效果。

2) 利用差分方法计算 $\boldsymbol{y}$ 的二阶导数谱 $\boldsymbol{y}_{d2}$，检测出二阶导数谱所有的极大值(峰点)和极小值点(谷点)对应的位置，并分别存储到 $\boldsymbol{p}$ 和 $\boldsymbol{v}$ 数组中。

3) 设置导数谱阈值限 g_{thres}。无论噪声还是光谱的二阶导数，一般均是两个极大值之间存在一个极小值点，则光谱仍不易被检测出来。但得益于拉曼光谱的锐利特征，其二阶导数谱(负谱)也非常尖锐，很容易突破噪声干扰，尤其是经过预先平滑后，二阶谱的特征更加明显。设置 g_{thres} 的方法是首先计算二阶导数谱所有极小值的均方根(RMS)以表征其整体分布水平；然后再将 RMS 乘以一个负值小因数 a_c 便得到 g_{thres}。取 $a_c<0$ 是因为二阶导数谱的大多数极小值都是小于零的。

4) 从 $\boldsymbol{v}$ 中剔除分布在零坐标轴和阈值限 g_{thres} 之间的极小值点对应的位置坐标得到 $\boldsymbol{v}'$，则剩下的数据以二阶导数谱的极小值位置为主。

5) 将数组 $\boldsymbol{p}$ 和 $\boldsymbol{v}'$ 连接起来并重新排序得到数组 $\boldsymbol{e}$。随后，根据峰-谷-峰规则确定二阶导数谱两个对称峰的间距 $\boldsymbol{R}_{pp}$。如图 2 所示，向两边扩展 $\boldsymbol{R}_{pp}$ 得到 $\boldsymbol{R}_{ex}=3\alpha\cdot\boldsymbol{R}_{pp}$，$\alpha\in(0, 2]$。这样，$\boldsymbol{R}_{ex}$ 基本上可表征原始光谱谱峰的分布区域。

6) 将各个 $\boldsymbol{R}_{ex}$ 对应的权重因子 $\boldsymbol{w}$ 置零，得到由 0-1 二元元素构成的新权重数组 $\boldsymbol{w}'$，$\boldsymbol{Wy}$ 即是扣除了光谱谱峰的剩余信号[$\boldsymbol{W}$=diagonal($\boldsymbol{w}'$)]，而这些剩余信号片段代表了基线漂移的趋势。最后，应用(5)式即可估计出整个光谱的平滑基线。同样，通过调节新的正则参数 λ_b 来控制估计基线的保真度和平滑度。

在上述基线校正算法中，共引入了 4 个调节参数：λ_a，a_c，α 和 λ_b。一般情况下，各个参数的选取需根据实际拉曼谱线分布情况和噪声量级具体分析，在对醇类、油类、不饱和烃类等物质的信噪比为 0~30 dB 的拉曼光谱进行大量实验后，给出各参数取值范围为：$1\leqslant\lambda_a\leqslant10^2$，$10^3\leqslant\lambda_b\leqslant10^9$，平滑因子的选取原则可以参考 Eilers 等的论述[11]；a_c 取决于二阶导数谱的信噪比，实际上间接反映了原始光谱平滑后的信噪比，需要配合 λ_a 进行调节；对于 Lorentzian 线型，展宽因子 α 的理论值为 1，而对于实际情况，α 值可以灵活选取。另外，在进行峰检测之前，对原谱和二阶导数谱进行平滑操作虽然可以滤除高频噪声，但有效光谱的高频成分同时也被滤除，从而导致平滑后光谱线宽有一定程度的展宽，因而此时 α 的取值应小于或等于 1；如果光谱线型采用高斯函数模拟，则 α 的理论值变为 $3^{1/2}=1.732$。综上所述，α 的取值范围推荐为 $0<\alpha\leqslant2$。

通常，4 个参数按顺序进行调节，即先选择正则参数 λ_a，随后根据二阶导数谱的极小值确定阈值限系数 a_c，再由测量的局部峰-谷-峰范围 $\boldsymbol{R}_{pp}$ 选择展宽因子 α 的取值，最后调节正则参数 λ_b 得到合理的基线估计 $\boldsymbol{f}$。

4　光谱数据实验

为了验证 GWSBLC 方法的效果，将其分别应用于模拟光谱和实际拉曼光谱的基线校正。

4.1　模拟光谱信号

模拟光谱由三部分构成：分析信号、模拟基线和随机噪声。其中，分析信号是应用(6)式生成的 5 条 Lorentzian 型谱线，相关参数如表 1 所示；模拟基线则由非线性函数生成；随机噪声为高斯白噪声(信号整体信

噪比为22 dB)。模拟光谱如图3所示。

表1 5条Lorentzian模拟谱线的系数

Table 1 Coefficients of 5 Lorentzian spectral profiles

Peak index	ω_0	γ
Peak 1	200	12
Peak 2	500	10
Peak 3	750	15
Peak 4	900	13
Peak 5	1150	11

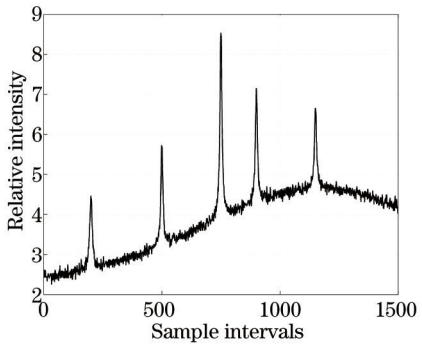

图3 具有非线性基线的模拟光谱

Fig.3 Simulated spectra with non-linear baseline

为了检验GWSBLC算法的有效性和在不同噪声环境下的适应性,对算法处理不同模拟光谱的结果进行了对比(模拟信号整体信噪比分别为22 dB和4 dB)。基线校正效果如图4所示,可以看出在不同噪声环境下,估计基线都是平滑地穿过噪声带中间,并且与模拟基线符合得很好,表明该算法得到的基线估计是可靠的,并具有良好的噪声适应性。

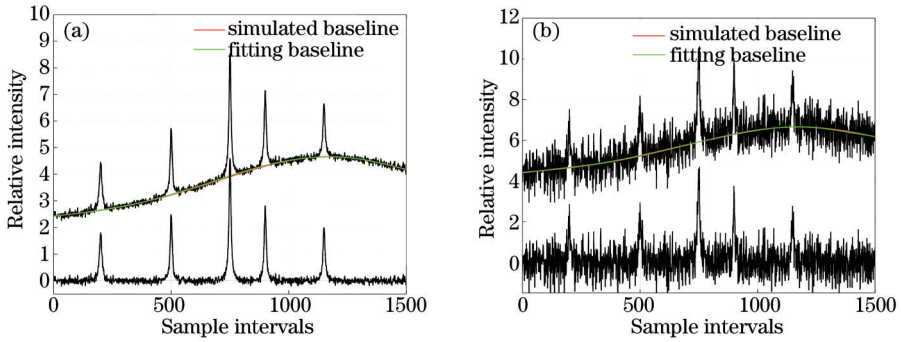

图4 模拟拉曼光谱的基线校正结果。(a) 低噪声情况; (b) 高噪声情况

Fig.4 Baseline correction results of simulated Raman spectra. (a) Low noise; (b) high noise

4.2 实际拉曼光谱

为了进一步检验GWSBLC方法,将其应用于由荧光背景引起基线漂移的β-胡萝卜素($C_{40}H_{56}$)拉曼光谱处理中,并与非对称最小二乘(AsLS)[11]和自适应迭代惩罚最小二乘方法(airPLS)[7]的基线校正结果进行比较。样本的原始拉曼光谱由532 nm激光器激发,并使用AvaSpec-2048L型光纤光谱仪测量得到。光谱仪的光谱测量范围为539~770 nm,光谱分辨率为0.8 nm,数据采集积分时间根据光谱信噪比设为1000 ms。AsLS和airPLS基线校正结果如图5所示,GWSBLC方法的基线校正结果如图6所示。

从图5、6可以看到,对于发生严重背景漂移的β-胡萝卜素拉曼光谱,三种算法都能估计出光谱的整体基线。然而,当光谱信号中存在由CCD像素延迟电流引起的伪谱线时(4200 pixel附近),由于AsLS和airPLS估计基线沿着噪声底部,不能很好地处理伪线情况,甚至使校正后的光谱产生假峰;相反,GWSBLC估计基线则是沿着噪声中部,伪线对算法不构成影响,因而能有效降低荧光背景干扰,同时使拉曼光谱峰形和峰位都得到了保持。

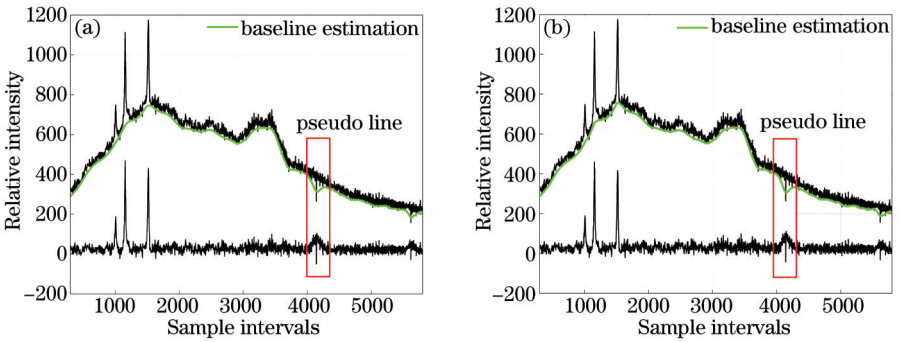

图5 β-胡萝卜素拉曼光谱基线校正结果(积分时间1000 ms)。(a) AsLS方法; (b) airPLS方法

Fig.5 Raman spectra of beta-carotene before and after baseline correction (integral time is 1000 ms). (a) AsLS method; (b) airPLS method

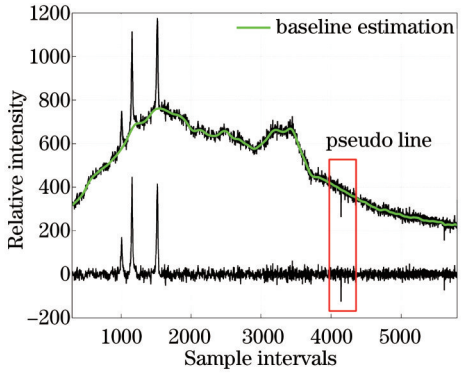

图6 应用GWSBLC方法对β-胡萝卜素拉曼光谱的基线校正结果(积分时间1000 ms)

Fig.6 Raman spectra of beta-carotene before and after baseline correction using GWSBLC (integral time is 1000 ms)

5　算法效果的讨论

5.1　边缘峰及重叠峰光谱处理效果

在拉曼光谱基线校正中,边缘峰和重叠峰的处理一直是一个难题,本节将考察GWSBLC对这两种情况的处理效果。

当原谱中存在边缘峰时,得到的拟合曲线在边缘处会出现偏离基线的现象,这是由于边缘峰外侧的先验基线信息不足甚至缺少,上述二阶导数谱不能确定谱峰在外侧的截止位置,对这一区域的基线估计,广义Whittaker平滑器实际上起到了向外插值作用,而不是一般的向内插值。对于这种情况,GWSBLC采用了一种经验做法,即不加区分地将二阶导数谱的两个端点设为峰点并作为一个强制边界条件,然后再应用广义Whittaker平滑器进行基线估计。实验表明,这一边界条件可以令GWSBLC较好地处理边缘峰情况,并且对于无边缘峰情况,该条件同样适用。

当拉曼谱峰为孤立且尖锐的线型时,很容易得到高精度的基线估计,但当出现光谱相互重叠的情况时,由于谱峰之间以及二阶导数谱之间的相互影响,使得基线校正变得相对困难,不准确的基线估计可能会偏离实际基线,甚至会穿过谱峰区域。通过对二阶导数谱峰间距$\boldsymbol{R}_{\mathrm{pp}}$向左右两边进行非对称展宽的措施进行修正,即将对称展宽因子α替代为左向展宽因子α_{left}和右向展宽因子α_{right},而最终扩展间距为$\boldsymbol{R}_{\mathrm{ex}}=(1+\alpha_{\mathrm{left}}+\alpha_{\mathrm{right}})\boldsymbol{R}_{\mathrm{pp}}$,两因子推荐取值范围为$\alpha_{\mathrm{left}}\in(0, 2]$,$\alpha_{\mathrm{right}}\in(0, 2]$。这也是一种经验方法,虽然实验表明能够较好地处理重叠峰的情况,但无疑增加了程序调节难度,降低了算法的自适应性,至于更完善的处理方案仍需要作进一步的研究。

图7(a)、(b)是利用Mazet[19]提供的程序生成的模拟光谱,而图7(c)、(d)是实际测量的四氯化碳(CCl_4)和正丙醇(C_3H_7OH)的拉曼光谱。对4种光谱应用修正的GWSBLC算法进行基线校正,可以发现,算法能够很好地估计孤立窄带光谱区域的基线;对于边缘峰[图7 (a)、(c)]或重叠峰[图7 (b)、(d)]情况,且光谱线宽变动比较大时,算法也能够准确标识出谱峰分布区域,最终利用残余基线片段信息估计出合理的漂移基线。

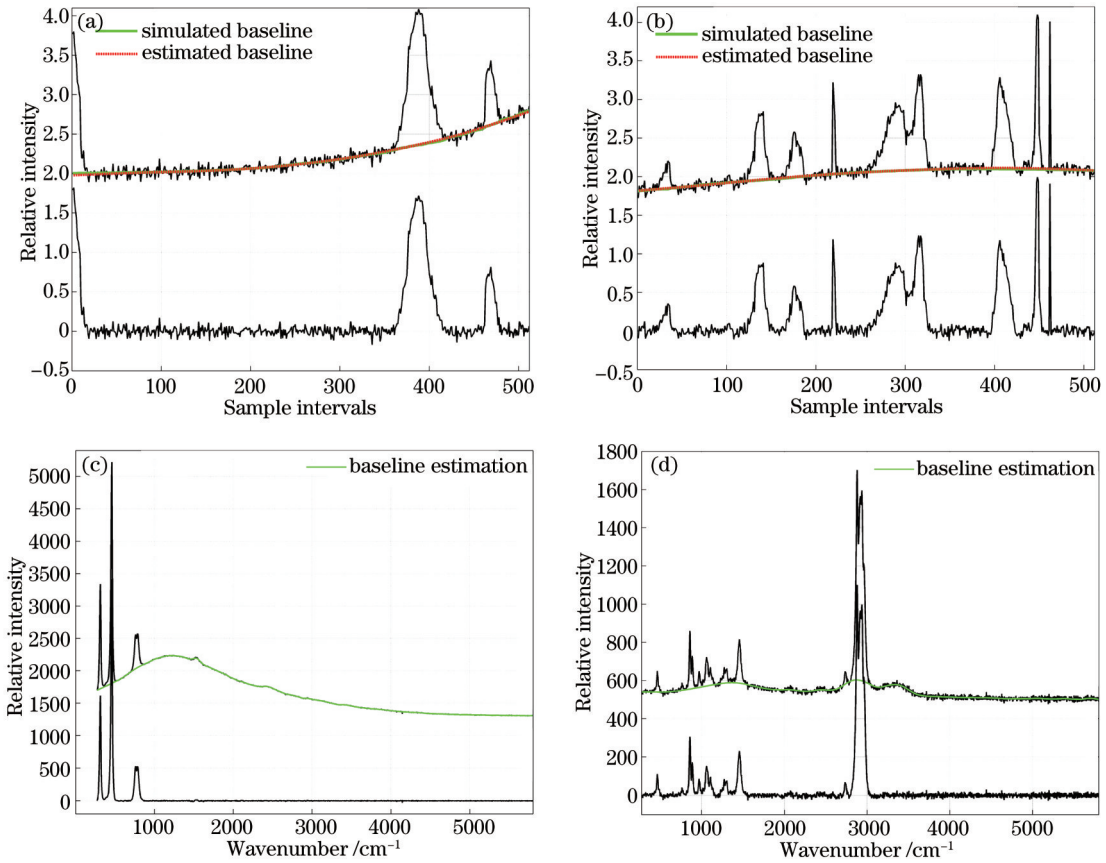

图 7　修正 GWSBLC 算法对边缘峰和重叠峰情况的处理结果

Fig.7　Processing results of edge peaks and overlapping peaks by the modified GWSBLC algorithm

5.2　基线校正对主成分分析的改进

主成分分析(PCA)是一种特征变量法，目的是对数据进行降维，它对原始光谱数据所包含的多个自变量进行线性拟合，以新的低维变量代替原始高维变量，常用于提取光谱特征信息[20-21]。为了验证经 GWSBLC 基线校正后数据的可靠性以及算法对后续光谱处理的影响，对校正光谱进行了主成分分析的分类处理。

实验选取 270~4000 cm^{-1} β-胡萝卜素的拉曼光谱作为特征光谱分析区域，采集 8 组拉曼光谱数据，对拉曼光谱进行基线校正(图 8)，再对基线校正前后的拉曼光谱进行主成分分析，并将其得分矩阵的第一主成分和第二主成分作图进行分析。

应用 GWSBLC 算法对β-胡萝卜素拉曼光谱进行基线校正前后对应的 PCA 分析结果如图 9 所示。由图可见，基线校正后的拉曼光谱得分矩阵点聚合性要远好于校正前的结果。

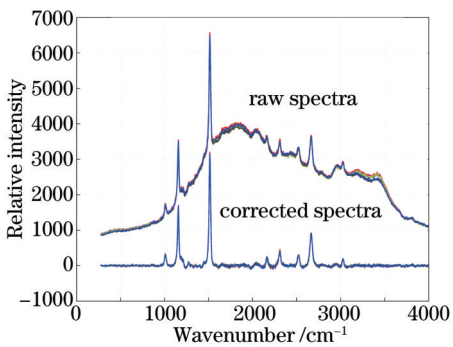

图 8　8 组β-胡萝卜素拉曼光谱的基线校正结果

Fig.8　Raman spectra of 8 groups of beta-carotene before and after baseline correction

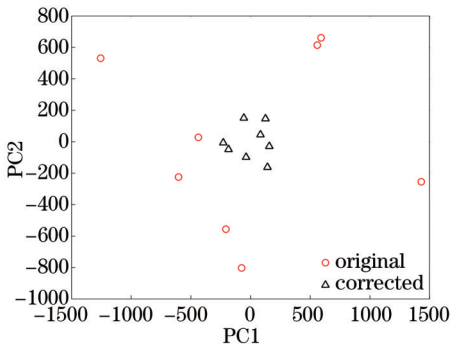

图 9　基线校正前后β-胡萝卜素拉曼光谱的主成分分析结果对比

Fig.9　PCA results of beta-carotene Raman spectra before and after baseline correction

同样选取270~4000 cm^{-1}无水乙醇和四氯化碳的拉曼光谱作为特征光谱分析区域，各采集9组拉曼光谱数据，利用GWSBLC算法对采集的光谱数据进行基线校正，结果如图10所示。分别对基线校正前后无水乙醇和四氯化碳的拉曼光谱数据进行主成分分析，并将其得分矩阵的第一主成分和第二主成分作图。

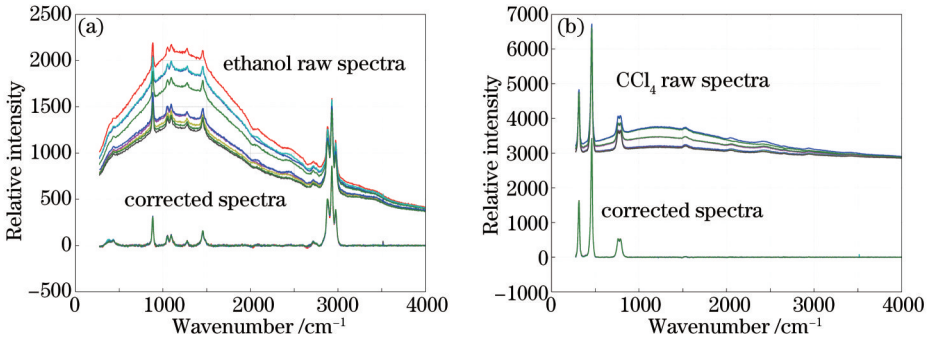

图10 无水乙醇和四氯化碳拉曼光谱的基线校正结果

Fig.10 Baseline correction results of ethanol and carbon tetrachloride Raman spectra

应用GWSBLC算法对无水乙醇和四氯化碳拉曼光谱进行基线校正前后对应的PCA分析结果如图11所示，由图可见，基线校正后的拉曼光谱的得分矩阵点可以很好地聚集，样本聚集度得到提高并且使两类样本可以完全分开。

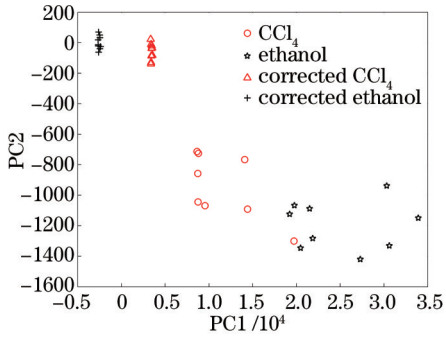

图11 无水乙醇和四氯化碳拉曼光谱基线校正前后的主成分分析结果对比

Fig.11 PCA analysis results of ethanol and carbon tetrachloride Raman spectra before and after baseline correction

由此可见，GWSBLC基线校正算法可以在保留有效信息的前提下将背景扣除，使原本在背景影响下不能达到良好聚类的物质实现了成功聚类。通过对基线校正前后的拉曼光谱做主成分分析验证了该算法的有效性和准确性。

6 结　论

为了消除基线漂移对拉曼光谱分析的影响，提出了GWSBLC算法，这是一种将光谱谱峰检测与广义Whittaker平滑器相结合的基线校正方法，首先利用二阶导数谱判断谱峰分布区域，然后由广义Whittaker平滑器根据扣除谱峰的残余基线信号实现全局基线估计。通过模拟信号和实际拉曼光谱的应用实验，结果表明，GWSBLC算法能够同时实现光谱的信号平滑和基线校正。另外，对基线校正前后的光谱数据做主成分分析，其得分矩阵点聚合性的提升进一步验证了该算法的有效性，同时也说明了拉曼光谱基线漂移对数据分析会产生影响，在定量分析前进行基线校正是必要的。

参 考 文 献

1 E Smith, G Dent. Modern Raman Spectroscopy - A Practical Approach[M]. Chichester: John Wiley & Sons, 2005.

2 Ma Jing. Low-concentration detection of chlorobenzene based on laser Raman spectroscopy[J]. Chinese J Lasers, 2014, 41(2): 0215001.
马　靖. 基于激光拉曼光谱的氯苯低浓度探测[J]. 中国激光, 2014, 41(2): 0215001.

3 Han Xiaozhen, Guo Zhengye, Kang Yan, *et al*.. Application of Raman spectroscopy in certification of chicken-blood stones[J]. Acta Optica Sinica, 2015, 35(1): 0130003.

韩孝朕, 郭正也, 康　燕, 等. 拉曼光谱在鸡血石鉴定中的应用[J]. 光学学报, 2015, 35(1): 0130003.

4 Liu Zhaojun, Han Yunxia, Yang Rui, *et al*.. Micro-Raman analysis of the pigments in the mural paintings from a Ming Dynasty tomb [J]. Chinese J Lasers, 2013, 40(6): 0615003.

刘照军, 韩运侠, 杨　蕊, 等. 明代古墓葬壁画颜料的显微拉曼光谱分析[J]. 中国激光, 2013, 40(6): 0615003.

5 P R Griffiths, J A de Haseth. Fourier Transform Infrared Spectrometer (2nd Edition)[M]. Hoboken: John Wiley & Sons, 2007.

6 T J Vickers, R E Wambles, C K Mann. Curve fitting and linearity: Data processing in Raman spectroscopy[J]. Appl Spectrosc, 2001, 55(4): 389-393.

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

8 E T Whittaker. On a new method of graduation[J]. P Edinburgh Math Soc, 1922, 41: 63-75.

9 P J Green, B W Silverman. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach[M]. London: Chapman & Hall/CRC, 1994.

10 P H C Eilers. A perfect smoother[J]. Anal Chem, 2003, 75(14): 3631-3636.

11 P H C Eilers, H F M Boelens. Baseline Correction with Asymmetric Least Squares Smoothing [EBOL]. http://zanran_storage.s3.amazonaws.com/www.science. uva.nl/Cont entPages/443199618.pdf.

12 N Li, X Y Li, Z X Zou, *et al*.. A novel baseline-correction method for standard addition based derivative spectra and its application to quantitative analysis of benzo (a) pyrene in vegetable oil samples[J]. Analyst, 2011, 136(13): 2802-2810.

13 H Mark, J Workman Jr. Derivatives in spectroscopy, part i - the behavior of the derivative[J]. Spectroscopy, 2003, 18(4): 32-37.

14 A N Tikhonov, A Goncharsky, V V Stepanov, *et al*.. Numerical Methods for the Solution of Ill-Posed Problems[M]. New York: Springer, 1995.

15 M P Tarvainen, P O Ranta-aho, P A Karjalainen. An advanced detrending method with application to HRV analysis[J]. IEEE Trans Biome Eng, 2002, 49(2): 172-175.

16 S Chountasis, V N Katsikis, D Pappas, *et al*.. The Whittaker smoother and the Moore-Penrose inverse in signal reconstruction[J]. Appl Math Sc, 2012, 6(25): 1205-1219.

17 Ma Heng, Wang Zhibin, Zhang Jilong, *et al*.. Improve polynomial iterative fitting algorithm for baseline correction in infrared spectroscopy[J]. Laser Technology, 2013, 37(2): 223-226.

马　恒, 王志斌, 张记龙, 等. 改进多项式迭代拟合红外光谱基线校正方法[J]. 激光技术, 2013, 37(2): 223-226.

18 Xia Liangping, Li Huadong, Yin Shaoyun, *et al*.. Eliminating complex background noise of Raman spectrum based on configuration similarity comparing method[J]. Acta Optica Sinica, 2013, 33(5): 0530003.

夏良平, 李华栋, 尹韶云, 等. 基于形状相似性比较法消除拉曼光谱的复杂背景噪声[J]. 光学学报, 2013, 33(5): 0530003.

19 V Mazet, C Carteret, D Brie, *et al*.. Background removal from spectra by designing and minimizing a non-quadratic cost function[J]. Chemometr Intell Lab Syst, 2005, 76(2): 121-133.

20 Chen Yang, Zhang Taining, Guo Peng, *et al*.. Quantitative analysis for nonlinear fluorescent spectra based on principal component analysis[J]. Acta Optica Sinica, 2009, 29(5): 1285-1291.

陈　扬, 张太宁, 郭　澎, 等. 基于主成分分析的复杂光谱定量分析方法的研究[J]. 光学学报, 2009, 29(5): 1285-1291

21 T Hasegawa. Detection of minute chemical signals by principal component analysis[J]. Trend Anal Chem, 2001, 20(2): 53-64.

栏目编辑: 吴秀娟