动态移动最小二乘多项式平滑的拉曼全自动基线校正算法

一. 成果完成人及所在单位：马真，电气工程及其自动化学院　　指导教师：熊飞兵

二. 成果简介：

基线漂移是目前光谱仪收集拉曼光谱时难以避免的现象。在应用拉曼光谱进行定性、定量分析过程中，基线漂移会给光谱分析带来不利影响，为增强分析结果的准确性和稳定性，需要对拉曼光谱进行基线校正预处理。为了校正基线漂移，提出一种动态移动最小二乘多项式平滑（DMSG）的方法拟合背景基线，自动识别原始光谱的谱峰区域和非谱峰区域的子区间，从而在每个拉曼光谱子区间中自动控制平滑迭代次数与窗口宽度，逐渐将拉曼谱峰数值逐渐拉低到与基线同等量级，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，结果表明该算法无论对模拟的拉曼光谱还是真实拉曼光谱，都有较好的基线校正效果，为进一步分析光谱数据提供更准确的信息，是一种可行的基线校正方法。这项设计对改善基线矫正方法均有理论指导意义。

三. 主要研究内容及创新点：

针对 Savitzky-Golay 平滑算法能对过于偏离正常趋势曲线的部分进行平滑，用拟合值代替原始数值，从而达到平滑原数据序列的特点，提出一种基于移动窗口最小二乘多项式（DMSG）的基线校正算法：拉曼光谱都是由基线，波峰及抖动噪声组成，根据拉曼光谱特征，可以将整个光谱分为谱峰子区间和基线子区间，并将其基线部分视为连续信号，谱峰区域视为基线突变信号，则由移动窗口最小二乘多项式平滑的特性可知，只要找到正确的谱峰子区间并对此区域进行迭代平滑，逐渐消除谱峰的影响，直至最后一次的迭代结果与上次的结果无重叠点，则这个平滑的结果就是这谱峰子区间的基线，将各个子区间的基线拼接就构成了整个拉曼光谱的基线。

DMSG 算法的关键是找到正确的基线子区间和谱峰子区间。目前的方法一般是通过采用原始光谱的二阶导数检测出谱峰分布区域，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。在这类方法中，实际信号容易受噪声干扰二不能准确的找到零点，从而影响判断谱峰的宽度导致基线不准确。与上述方法不同，这里通过判断区间的平滑前与平滑后的光谱差值 X 的范围来寻找基线子区间。通过模拟计算表明，对于长度大于 L（取 L=20）的区间，当经过 m（取 m=35）次 S-G 迭代平滑后，平滑前和平滑后的光谱差值 X 小于阈值 T_1（取 T_1=0.03H，H 是此区间中最大峰高），则此区间存在基线子区间，同时考虑噪声影响，对区间作进一步限定 $T_2<X<T_1$（取 T_2=0.003H），其对应的区间可视作基线子区间 SS。另外由 S-G 平滑算法特性可知，对于 S-G 平滑后的光谱数据首尾两端各 M 个点会保存其光谱值不变，因此在光谱 XX 的两端各添加一条利用光谱两端各 n 个光谱点利用最小二乘法（LSM）拟合而成的直线，所以 DMSG 算法是针对于光谱 XX 的两端添加两直线后的延长光谱 XL 来寻找基线子区间。因此拉曼光谱基线校正步骤如下：

1.载入原始测量光谱数据 S 并对此进行 S-G 平滑滤波，预先减少噪声对后续处理的影响，得到平滑后的拉曼光谱 XX。

2.在拉曼光谱 XX 两端的 n 个光谱点分别用最小二乘法拟合两条直线，得到延展光谱 XL。

3.对延展光谱 XL 进行 m 次 S-G 迭代平滑得到 XX'，将 XX 与 XX' 相减计算得到光谱差值 X。找出所有长度大于 L 的区间中光谱差值 X 满足 $T_2<X<T_1$ 限定条件的基线子区间 SS。

4.对于由基线子区间 SS 将整个拉曼光谱分割成多个谱峰子区间 S_p，分别在各个子区间利用迭代 S-G 算法计算出子区间的基线。

5.在迭代过程中，多项式拟合阶次 d=4，平滑窗口宽度 W（初值 W=5）逐渐增加，窗口宽度步进值为 2，并设宽度最大值 W_{max}，为确保当 $W=W_{max}$ 时子区间两端数据也能平滑，在每个拉曼谱峰子区间 S_p 两端添加 W_{max} 个数据，其中 W_{max} 个数据可由延展光谱 XL 中得到。考虑到噪声影响，当迭代中