

乙醇含量拉曼光谱检测中噪声与背景同时消除方法研究

韩庆阳, 周鹏骥

中国科学院长春光学精密机械与物理研究所, 吉林 长春 130033

摘要 乙醇含量拉曼光谱检测中, 拉曼光谱信号中的各种噪声及光谱荧光造成的基线漂移和样品池背景等, 影响了校正模型的预测精度。利用总体平均经验模态分解, 将光谱信号分解成若干无模态混叠的内在模式分量, 根据排列熵的信号随机性检测判断出代表背景信息和噪声信息的内在模式分量, 将其置零即可同时消除拉曼光谱中的噪声与背景。将总体平均经验模态分解与排列熵相结合的预处理方法应用于乙醇含量的拉曼光谱检测中, 并与小波变换和平均平滑滤波做了对比。实验结果表明: 应用总体平均经验模态分解与排列熵相结合的方法能够有效的同时消除乙醇含量拉曼光谱检测中的噪声和背景信息, 提高校正模型的预测精度, 且使用简便, 无需参数设置, 对乙醇含量拉曼光谱检测具有实用价值。

关键词 乙醇; 拉曼光谱; 噪声; 背景; 总体平均经验模态分解; 排列熵

中图分类号: O657.3 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2015)12-3406-04

引言

拉曼光谱作为一种快速、无损检测新技术, 因其散射强度与样品浓度存在关联性, 且受水分含量影响非常小, 被认为是分析水溶液样品的理想工具^[1]。乙醇含量拉曼光谱检测中, 拉曼光谱的噪声主要有两类: 一是仪器的热噪声, 二是环境干扰; 背景则主要是: 拉曼光谱荧光造成的基线漂移以及样品池背景等^[2]; 从拉曼光谱信号中提取有用信息, 对乙醇含量检测精度的提高至关重要。

有效的对拉曼光谱进行预处理, 为校正模型预测精度的提高打下基础。目前, 对拉曼光谱的预处理方法有: 平均平滑滤波和小波变换等^[3]。平均平滑滤波操作简单, 但精度较差; 小波变换精度较高, 但是需要选择分解层数和小波基函数。基于总体平均经验模态分解(ensemble empirical mode decomposition, EEMD)的预处理方法^[4], 把拉曼光谱信号分解成若干无模态混叠的内在模式分量(intrinsic mode function, IMF), 根据检测时间序列的随机性的排列熵(permutation entropy, PE)^[5], 判断分别代表噪声和背景的 IMF 分量, 将其置零后信号重构, 即实现拉曼光谱中噪声与背景的同时消除。

将 EEMD 与 PE 相结合的预处理方法应用于乙醇含量拉曼光谱检测, 搭建实验测量平台进行实验验证, 并与平均平滑滤波和小波变换的预处理方法进行对比。实验结果表明该方法有效的同时消除了拉曼光谱的噪声和背景, 乙醇含量的检测精度得到提高。

1 算法原理

在拉曼光谱信号 $A(\lambda^{-1})$ 中添加白噪声, 则信号变为

$$A'(\lambda^{-1}) = A(\lambda^{-1}) + c_i \text{noise}_i(\lambda^{-1}) \quad (1)$$

λ^{-1} 为拉曼光谱的波数, 并且满足 $1 \leq \lambda^{-1} \leq n$; c_i 为白噪声的幅值, 选取原则为其标准差一般原始信号的 0.1~0.2 倍, 对信号进行 EMD 分解, 然后重复添加幅值为 c_i 的白噪声再分解, 如此重复 d 次, 平均次数; 一般取 $d=100 \sim 200$ 。则信号可以表示成

$$A'(\lambda^{-1}) = \sum_{i=1}^n \text{IMF}_i(\lambda^{-1}) + r_n(\lambda^{-1}) \quad (2)$$

式中, i 为整数, $1 \leq i \leq n$, $r_n(\lambda^{-1})$ 为余项, $\text{IMF}_1(\lambda^{-1}) \sim \text{IMF}_n(\lambda^{-1})$ 为分解得到的 IMF 分量, 对每个长度为 N 的 IMF 分量进行空间重构得到如下序列

收稿日期: 2014-10-23, 修订日期: 2015-02-04

基金项目: 国家自然科学基金项目(60977001), 吉林省科技厅项目(201026015, 20125092), 吉林省与中国科学院合作长吉图开发开放先导区科技创新合作专项(2011CJT0004)资助

作者简介: 韩庆阳, 1988 年生, 中国科学院长春光学精密机械与物理研究所助理研究员 e-mail: lanbohesky@126.com

$$\left. \begin{aligned} \text{IMF}_i(1) &= \{\text{IMF}_i(1), \text{IMF}_i(1+\delta), \dots, \\ &\quad \text{IMF}_i(1+(\alpha-1)\delta)\} \\ \text{IMF}_i(h) &= \{\text{IMF}_i(h), \text{IMF}_i(h+\delta), \dots, \\ &\quad \text{IMF}_i(h+(\alpha-1)\delta)\} \\ &\vdots \\ \text{IMF}_i(N-(\alpha-1)\delta) &= \{\text{IMF}_i(N-(\alpha-1)\delta), \\ &\quad \text{IMF}_i(N-(\alpha-2)\delta), \dots, \text{IMF}_i(N)\} \end{aligned} \right\} \quad (3)$$

其中 $1 \leq h \leq N-(\alpha-1)\delta$, α 是嵌入维数, δ 是时间延迟, 将 $\text{IMF}_i(h)$ 的 α 个向量: $\text{IMF}_i(h) = \{\text{IMF}_i(h), \text{IMF}_i(h+\delta), \dots, \text{IMF}_i(h+(\alpha-1)\delta)\}$ 按照升序排列, 即

$$\text{IMF}_i(h) = \{\text{IMF}_i(h+(j_1-1)\delta) \leq \dots \leq \text{IMF}_i(h+(j_\alpha-1)\delta)\} \quad (4)$$

若 $\text{IMF}_i(h+(j_{h1}-1)\delta) = \text{IMF}_i(h+(j_{h2}-1)\delta)$, 则按照 j 的大小进行排列, 所以, 任意 IMF 分量都可以得到一组符号序列 $S(g) = [j_1, j_2, \dots, j_m]$, 其中 $g=1, 2, \dots, h, h \leq \alpha$ 。 α 个不同的符号 $[j_1, j_2, \dots, j_m]$ 有 $\alpha!$ 种排列, $S(g)$ 是 $\alpha!$ 种符号的一种。计算每种符号出现的概率 P_g , 概率的和为 1。

IMF 分量的排列熵为

$$PE_i(\alpha) = - \sum_{g=1}^h P_g \ln P_g \quad (5)$$

当 $P_g = \frac{1}{\alpha}$ 时, $PE_i(\alpha)$ 达到最大值 $\ln(\alpha!)$, 将 $PE_i(\alpha)$ 通过 $\ln(\alpha!)$ 标准化处理后得

$$PE_i = \frac{PE_i(\alpha)}{\ln(\alpha!)} \quad (6)$$

所以信号分解后的 IMF 分量 IMF_i 的排列熵为 PE_i (i 为整数, $1 \leq i \leq n$), 在计算排列熵时需要确定嵌入维数 α 和时间延迟 δ , 根据经验值一般取 $\alpha=6$, $\delta=1$; PE_i 值越大说明该 IMF 分量越随机, 反之说明越规则。

确定排列熵的阈值, 判断出代表高频噪声的 IMF 分量为 $\text{IMF}_1(\lambda^{-1}) \sim \text{IMF}_m(\lambda^{-1})$ (m 为整数, $1 \leq m \leq n$), 和代表基线及背景的 IMF 分量为 $\text{IMF}_p(\lambda^{-1}) \sim \text{IMF}_n(\lambda^{-1})$ (p 为整数, $1 \leq p \leq n$); 则基线及背景为

$$\text{Base} = r_n(\lambda^{-1}) + \sum_{k=p}^n \text{IMF}_k(\lambda^{-1}) \quad (7)$$

式中 k 为整数, $p \leq k \leq n$, 将代表高频噪声的 IMF 分量和代表基线漂移及背景的 IMF 分量置零, 信号重构得到 $A(\lambda^{-1})''$

$$A(\lambda^{-1})'' = \sum_{l=m}^p \text{IMF}_l(\lambda^{-1}) \quad (8)$$

式中, p, m 和 l 为整数, 且 $1 \leq p \leq n$, $l \leq m \leq p$, $m \leq l \leq n$ 。

2 实验部分

样品制备: 无水乙醇(分析纯)和去离子水, 利用 100 mL 的容量瓶和 100 mL 的移液管, 采用国家标准溶液配方法, 即先采用 100 mL 的移液管准确量取所需乙醇的体积, 然后加入去离子水在 100 mL 的容量瓶进行定容, 得到相应浓度的乙醇水溶液。共配制乙醇水溶液 14 个样本, 浓度范围为 5%~70%, 间隔为 5%。实验温度为常温 25 °C。

实验系统搭建: 采用美国海洋光学(Ocean Optics,

America)公司生产的 QE65000 科研级拉曼光谱仪(包括单色器和检测器系统)和 Laser-785 拉曼光谱激发激光器以及 RIP-RPS-785 光纤探头。PC 机作为记录系统, 光谱仪与 PC 机通过 USB 口进行通讯。

图 1 是由该实验系统采集得到的拉曼光谱, 由图可以看出: 所采集的拉曼光谱信号中存在由荧光引起的基线漂移, 同时也存在由仪器热噪声引起的高频噪声和样品池的背景。这些因素影响乙醇浓度的检测精度。

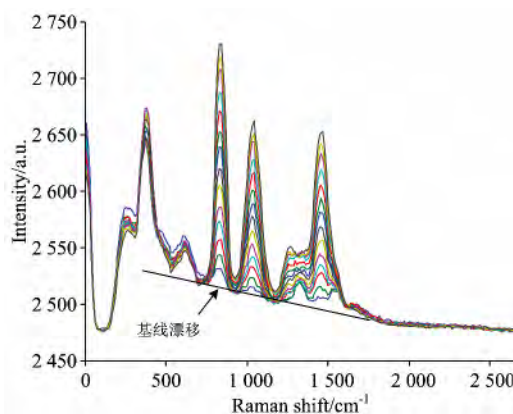


Fig 1 Original Raman spectra

3 结果与讨论

首先, 对测量得到的拉曼光谱进行 EEMD 分解; 其次, 计算每个分解得到 IMF 分量的 PE 值; 第三, 选取经验 PE 阈值, 经多次实验, 选取 $0.1 \leq PE \leq 0.2$, PE 值大于 0.2 的 IMF 分量代表噪声; PE 值小于 0.1 的 IMF 分量背景信息; 第四, 将代表噪声和背景的 IMF 分量置零后信号重构得到同时消除噪声和背景的拉曼光谱。

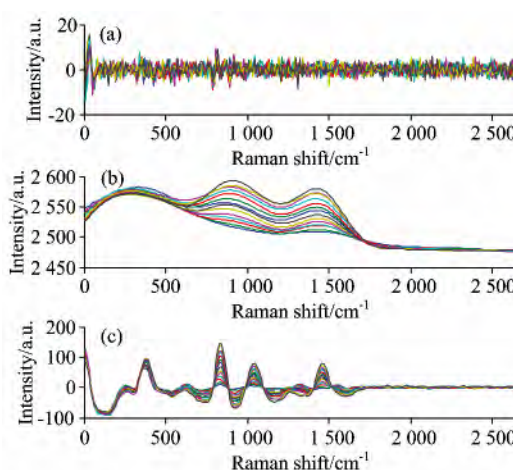


Fig 2 Noise(a), baseline(b) and useful information(c)

图 2(a)是从拉曼光谱中剔除的噪声, 图 2(b)是剔除的背景信息, 图 2(c)是乙醇含量检测有用的信息即消除噪声和背景的拉曼光谱。图 3 是原始拉曼光谱和消除噪声和背景的拉曼光谱, 由图可知: 乙醇的三个特征峰, 分别为 870, 1 050 和 1 450 cm^{-1} , 得到了很好保留, 对测量精度有影响的

无用信息得到消除。

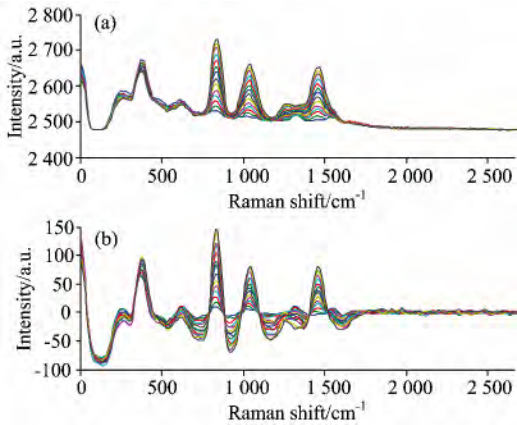


Fig 3 Original Raman spectrum and proceeded Raman spectrum

为了进一步说明该预处理方法对预测结果的影响，分别采用平均平滑滤波和小波变换的预处理方法，然后对所有的模型处理后的拉曼光谱数据采用偏最小二乘回归 (partial least regression, PLS) 的留一交互验证方式建立校正模型。其中模型精度采用交互验证预测均方根 (root mean square error of cross validation, RMSECV) 表示，模型的拟合程度采用复相关系数 R^2 表示，模型的预测能力采用预测均方根误差 (root mean square error of prediction, RMSEP) 和最大绝对预测误差 ($\max |e|$) 表示。其计算公式如下

$$\text{RMSECV}(P) = \sqrt{\frac{\sum_{\gamma=1}^M (y_{\gamma} - \hat{y}_{\gamma})^2}{M}}$$

References

[1] WU Bin, LUO Xiao-sen, LU Jian, et al(吴斌, 骆晓森, 陆建, 等). Optics and Precision Engineering(光学精密工程), 2011, 19(2): 392.

[2] WANG Xin, FAN Xian-guang, XU Ying-jie, et al(王昕, 范贤光, 许英杰, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2014, 34(8): 2118-2121.

[3] FANG Cheng-xuan, LI Jian-hua, LIANG Yi-zeng(房承宣, 李建华, 梁逸曾). Journal of Instrumental Analysis(分析测试学报), 2012, 31(5): 541.

[4] ZHENG De-jin, CHENG Jun-sheng, YANG Yu(郑德近, 程军圣, 杨宇). Journal of Vibration and Shock(振动与冲击), 2013, 32(21): 22.

[5] FENG Fu-zhou, RAO Guo-qiang, SI Ai-wei, et al(冯辅周, 饶国强, 司爱威, 等). Journal of Vibration Engineering(振动工程学报), 2012, 25(2): 221.

$$R^2 = 1 - \frac{\sum_{\gamma=1}^M (y_{\gamma} - \hat{y}_{\gamma})^2}{\sum_{\gamma=1}^M (y_{\gamma} - \bar{y})^2} \tag{9}$$

$$\max |e| = \max |y_{\gamma} - \hat{y}_{\gamma}| \quad \gamma = 1, 2, \dots, M$$

其中 y_{γ} 为样本的浓度, $\hat{y}_{\gamma}$ 为拟合或预测值, M 为样本个数。

表 1 是三种预处理方法处理后的拉曼光谱数据, 经过定量校正后得到 RMSECV, R^2 和 $\max |e|$; 由表可知: 三种方法中小波变换精度最高, 拟合程度最好; EEMD-PE 法相对小波变换效果略差; 平均平滑滤波法效果最差。但是, EEMD-PE 法无需设置参数和选择基函数, 其简便性是小波变换无法比拟的; 另外, 小波变换由于小波基函数形状的影响, 去噪后的光谱图上会看到一点小波基的形状, 相比于 EEMD-PE 法, 光谱的光滑度受到一定的影响。

Table 1 Results of the three methods

Method	RMSECV /%	R^2	$\max e $ /%
average smoothing filter	0.73	0.998 9	1.41
wavelet transform	0.57	0.999 7	1.17
EEMD-PE	0.62	0.999 5	1.35

4 结 论

利用总体平均经验模态分解结合排列熵的预处理方法, 对乙醇含量拉曼光谱检测实验系统采集的数据进行处理, 并与平均平滑滤波法和小波变换预处理方法对比, 实验结果表明: 该方法不仅能有效的同时消除拉曼光谱中的噪声和背景, 而且简便易实现, 对乙醇含量拉曼光谱检测具有重要实用价值。

致谢: 感谢北京航空航天大学精密光机电一体化技术教育部重点实验室在实验数据环节的共享与支持。

Research on the Method of Eliminating Noise and Background in the Meantime in Detecting Ethanol Content Based on Raman Spectra

HAN Qing-yang, ZHOU Peng-ji

Photoelectric Technology Research and Development Center, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

Abstract In the process of detecting ethanol content by Raman spectra, the precision of correction model prediction is affected by noise and baseline drift, which is caused by the spectral fluorescence and sample pool's background. Use ensemble empirical mode decomposition to decompose spectrum into several intrinsic mode functions, which are without aliasing. The permutation entropy is employed to judge the intrinsic mode functions. Set the intrinsic mode functions which are on behalf of noise and background to zero, and then the signal is without noise and background. In this paper combine ensemble empirical mode decomposition and permutation entropy, and apply to the Raman spectrum, which are used to detect ethanol content. At the same time compare with wavelet transform and average smoothing filter. The experimental result shows that the application of empirical mode decomposition and permutation entropy can effectively eliminate the noise and background. The precision of correction model prediction is improved. This method simply employs and doesn't need to set parameters, which has great value of application in the process of detecting ethanol content by Raman spectra.

Keywords Ethanol; Raman spectrum; Noise; Background; Ensemble empirical mode decomposition; Permutation entropy

(Received Oct. 23, 2014; accepted Feb. 4, 2015)

《光谱学与光谱分析》对来稿英文摘要的要求

来稿英文摘要不符合下列要求者,本刊要求作者重写,这可能要推迟论文发表的时间。

1. 请用符合语法的英文,要求言简意明、确切地论述文章的主要内容,突出创新之处。
2. 应拥有与论文同等量的主要信息,包括四个要素,即研究目的、方法、结果、结论。其中后两个要素最重要。有时一个句子即可包含前两个要素,例如“用某种改进的 ICP-AES 测量了鱼池水样的痕量铅”。但有些情况下,英文摘要可包括研究工作的主要对象和范围,以及具有情报价值的其他重要信息。在结果部分最好有定量数据,如检测限、相对标准偏差等;结论部分最好指出方法或结果的优点和意义。
3. 句型力求简单,尽量采用被动式,通常应有 2500 个印刷字符,400 个英文单词为宜,不能太短;也不要太长。用 A4 复印纸单面打印。
4. 摘要不应有引言中出现的内容,换言之,摘要中必须写进的内容应尽量避免在引言中出现。摘要也不要对论文内容作解释和评论,不得简单重复题名中已有的信息;不用非公知公用的符号和术语;不用引文,除非该论文证实或否定了他人已发表的论文。缩略语、略称、代号,除相邻专业的读者也能清楚地理解外,在首次出现时必须加以说明,例如用括号写出全称。