

基线对傅里叶变换红外光谱的目标因子分析结果的影响研究

王丽 邵利民*

(中国科学技术大学化学与材料科学学院, 合肥 230026)

摘要 目标因子分析(TFA)在不经任何预处理的情况下,能否成功地从实验数据矩阵中提取目标光谱主要取决于相应组分浓度的方差。当氨的浓度方差较低、基线漂移较严重时,TFA可能给出假阴性结果,因此,需要校正基线,以提高TFA性能。本研究采用6组不同浓度方差、不同基线漂移程度的大气开放光路傅里叶变换红外(OP/FT-IR)光谱数据,比较标准正态变换(SNV)、多元散射校正(MSC)、线性拟合、小波变换(WT)和自适应迭代重加权惩罚最小二乘(airPLS)法去除基线的效果。结果表明,小波变换能够有效去除基线,从而使氨和甲烷光谱信息提前出现在PCA的特征向量中,显著提高了目标因子分析的定性能力。

关键词 目标因子分析;主因子数;基线校正;小波变换

近年来,开放光路傅里叶变换红外光谱(Open path Fourier transform infrared spectroscopy, OP/FT-IR)技术已广泛应用于大气动态实时监测与质量检测^[1-3]。OP/FT-IR无需取样、操作简单快速、且能够无损测量和无人值守监测^[4-6]。然而,大气中痕量成分的测量常受到水蒸汽和二氧化碳吸收、基线漂移与噪声的干扰,给大气的定性分析带来困难,因此,需要借助化学计量学方法提高分析效率和可靠性。

在氨的定性分析过程中,从OP/FT-IR光谱中寻找氨的光谱特征是一种简单而直接的方法,然而,低浓度氨无明显的光谱特征,该方法在实际中难以施行。目标因子分析(Target factor analysis, TFA)是Malinowski^[7]提出的一种化学因子分析方法,已被用于多种物质定性分析。如Schweitzer等^[8]结合欧几里得距离、目标因子分析和自适应余弦估计方法与光谱药物库进行比较,自动识别非法药物的化学成像;Msimanga等^[9]结合紫外光谱数据和目标因子分析减少红葡萄酒儿童抗感冒止咳露分析中的干扰;Reháková等^[10]利用纤维光学反射光谱法和目标因子分析法鉴别历史绘画中的油墨。本研究通过TFA法提取OP/FT-IR光谱数据矩阵中的氨的信息,进而判断大气中是否存在氨。

将TFA应用于OP/FTIR光谱的优点是无需进行预处理如补偿水线吸收或基线校正^[11],然而,对于浓度方差小、基线漂移大的光谱数据,TFA可能给出假阴性结果,因此需要基线校正。基线校正方法有标准正态变换法(Standard normal variate, SNV)、多元散射校正法(Multiplicative scatter correction, MSC)、导数法^[12]、线性拟合法^[13]、小波变换法(Wavelet transform, WT)^[14]和自适应迭代重加权惩罚最小二乘法(Adaptive iterative reweighted penalty least squares, airPLS)^[15]等。SNV和MSC均用于校正样品中粒子散射引起的光谱误差,在理想情况下,MSC主要用于消除线性散射的影响,由于SNV对每个频谱进行校正,一般认为SNV比MSC具有更强的校正能力。线性拟合方法对严重漂移的光谱处理能力较弱。WT具有时频定位特性好、小波基选择灵活、去相关性好、自动化程度高、鲁棒性强等优点,已经在傅里叶变换红外光谱的数据压缩、数据平滑去噪、基线和背景校正等方面得到了广泛应用^[16]。airPLS方法不需要任何预先的信息,如用户干预或峰值检测,已成功应用于高效液相色谱、拉曼光谱和核磁共振信号的基线校正^[15],但在OP/FTIR光谱的应用未见报道。

本研究采用正交投影最小二乘(Orthogonal projection approach least squares, OPALS)法判断主因子数,通过SNV、MSC、线性拟合、WT和airPLS法校正大气光谱的基线,并比较了5种方法对TFA定性能力的影响。

1 原理

1.1 目标因子分析

目标因子分析的原理详见文献[11]。在TFA过程中,通过主成分分析(PCA)和最小二乘法,基于

参考光谱 S_0 构建重构谱 S_R , 计算参考光谱与重构谱的加权相关系数(WCC)^[17], 以表征二者的相似性; 当 $WCC > 0.80$ 时, 可以判断目标对象存在。

1.2 OPALS 法判断主因子数

目标因子分析中, 首先需要判断主因子数。OPALS 法是正交投影法和最小二乘法的结合, 通过比较原始和重构数据的特征向量, 并用一致性函数判别主因子数^[18]。

1.3 SNV、MSC 和线性拟合的基线校正

SNV 和 MSC 的校正公式如下:

$$\mathbf{x}_{\text{SNV}} = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\sqrt{\frac{\sum_{k=1}^p (\mathbf{x}_k - \bar{\mathbf{x}})^2}{p-1}}} \quad (1)$$

$$\mathbf{x}_{\text{MSC}} = \frac{(\mathbf{x} - b)}{m} \quad (2)$$

其中, \mathbf{x}_{SNV} 和 \mathbf{x}_{MSC} 分别为经过 SNV 和 MSC 校正之后的光谱向量; $\bar{\mathbf{x}}$ 为光谱向量的平均; p 为波数, $k=1, 2, 3, \dots, p$; m 为相对偏移因子; b 为偏移量。通过对光谱和平均光谱进行一元线性回归得到。

线性拟合方法是将含有线性趋势的信号拟合出基线, 从原始光谱中减去拟合基线, 得到去除基线的光谱。

1.4 基于小波变换的基线校正

小波变换的基线校正基于多分辨率信号分解(MRSD)算法^[19]。利用小波变换对原始光谱进行分解, 得到离散近似和离散细节, 即低频和高频的光谱成分。基线频率总是显著低于光谱信息的频率, 因此将离散近似置零后进行小波逆变换, 即可实现基线校正。

1.5 airPLS 的基线校正

通过对拟合基线与原始光谱的平方和误差的权值进行迭代变换, 自适应地获得权值, 从而拟合出合理基线, 从原始光谱中减去基线, 实现基线校正。

2 实验部分

本研究的光谱数据在美国 Idaho 州南部完成测量。FTIR 光谱仪为 MDA (Atlanta, GA) 产品; 干涉仪为 Bomem Michelson 10; 检测器为汞镉碲(MCT); 望远镜直径为 31.5 cm; 立方角阵列反射镜; 仪器控制和数据采集使用 GRAMS 7.00 (Thermo Galactic, Salem, NH)。

光谱处理过程中, 填零因子为 8, 切趾函数为 Nonon-Beer (Medium)^[20], 对 MCT 检测器进行非线性响应校正^[21], 计算干涉谱获得分辨率为 1 cm^{-1} 的吸收光谱。通过此方法得到不同时间、不同基线漂移程度的 6 组吸收光谱数据集, 命名为数据集 1~6。数据集 1~3 的波数范围为 $1250 \sim 750 \text{ cm}^{-1}$, 表示氨的吸收光谱; 数据集 4~6 的波数范围为 $3200 \sim 2850 \text{ cm}^{-1}$, 表示甲烷的吸收光谱, 数据集的基本信息见表 1。采用偏最小二乘法得到氨和甲烷的预测浓度, 最终得到各数据集的浓度方差。

表 1 开放光路傅里叶变换红外光谱连续监测的数据集信息

Table 1 Data sets information for continuous open path Fourier transform infrared spectroscopy monitoring and measurement

序号 No.	时间 Time	位置 Location	光谱数 Number of spectrum
1	June	奶牛场 Dairy farm	331
2	June	养猪场 Hog farm	150
3	January	农场 Farm	692
4	June	养猪场 Hog farm	150
5	March	奶牛场 Dairy farm	550
6	June	奶牛场 Dairy farm	331

本研究所有的光谱计算和数据处理均在 MATLAB 2015b (The MathWorks, Inc., Natick, MA) 系统环境中完成。

3 结果与讨论

3.1 确定主因子数

OPALS 方法将选取的关键变量数设为原光谱数的 $1/3$, 规定相关系数大于 0.9 的点数为主因子数。经过基线校正后的数据依然使用原始数据得到的主因子数, 这是因为原始数据比基线校正数据的主因子数更大, 包含更多的光谱信息, 这在复杂的大气环境中更加合理。

3.2 基线校正

户外测量会受到各种因素的影响, 当一些极端条件引起严重漂移时, TFA 给出假阴性结果, 此时对数据进行预处理非常必要。本研究采用 MSC、SNV、线性拟合、WT 和 airPLS 方法分别对漂移的基线进行校正, 并比较 TFA 的定性结果。

研究了数据集 1、2、3, 主因子数计算结果分别为 4、6、5。数据集 1 的光谱如图 1 所示, 基线波动差别极大, 原始光谱图非常分散; 在 $1000 \sim 900 \text{ cm}^{-1}$ 范围内可见微弱的氨特征峰。数据集 3 的光谱基线波动差别次之, 数据集 2 的光谱基线波动差别最小, 而且均未见氨的特征峰, 这 3 个数据集包含了在实际大气测量中遇到的典型基线漂移情况。

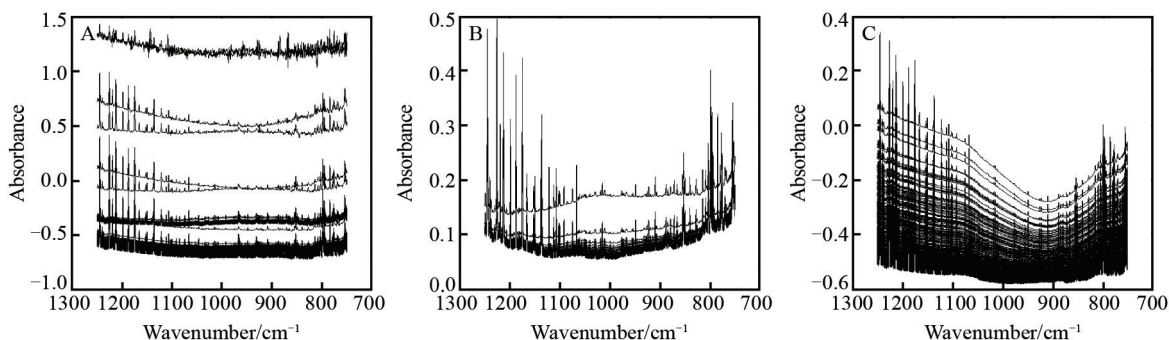


图 1 (A) 数据集 1、(B) 数据集 2 和 (C) 数据集 3 的原始光谱

Fig. 1 Original spectra of (A) data set 1, (B) data set 2 and (C) data set 3

图 2 是通过 TFA 方法得到的重构谱和参考谱, 数据集 1、2、3 的 WCC 分别为 0.75、0.60、0.66, 均小于 0.80, 浓度方差分别为 226.8、0.45、4.3。结果表明, 数据集 1 的浓度方差大、基线漂移十分严重; 数据集 2 的浓度方差非常小、基线漂移不严重; 数据集 3 的浓度方差小, 基线漂移严重。上述 3 种情况下, TFA 均无法检测出氨的存在。

对数据集 1、2、3 分别应用 MSC、SNV、线性拟合、WT 和 airPLS 方法进行基线校正。为了抵消这些校正方法引起的光谱信息畸变, 同样需要对参考光谱进行基线校正。

在 airPLS 方法中, 当参数 $\lambda = 100$ 时, 基线校正效果最好。经过对比, WT 和 airPLS 校正方法去基线效果最好, 如图 3 所示, 经过 WT 和 airPLS 校正后的 3 组光谱基线波动均很小, 非常整齐, 但 TFA 定性结果差异很大。当使用小波变换时, 重构谱与参考谱具有很高的重合度, 如图 4 所示, WCC 分别为 0.99、0.91、0.99, 均大于 0.80, 确定了氨的存在。当使用 airPLS 时, 重构谱与参考谱无重合 (见图 5), WCC 均小于 0.80。

其它基线校正方法的 TFA 结果列于表 2, MSC、SNV 和线性拟合方法不适用所有类型的基线漂移校正, 所以会出现有的数据集结果好、有的数据集结果差的情况。airPLS 虽然能基本去除基线, 但是也造成了部分光谱信息的损失, 导致此方法没有效果。表 2 给出 3 组甲烷数据集通过不同方法校正后的 TFA 结果, 与氨得出的结果一致。

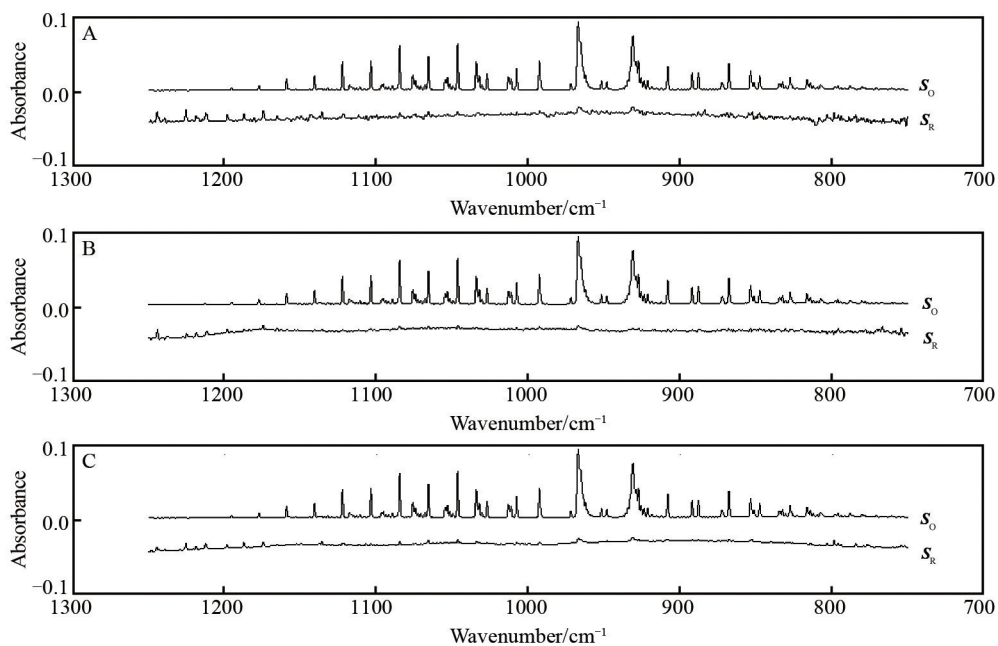


图2 (A) 数据集1、(B) 数据集2和(C) 数据集3的TFA结果

S_0 和 S_R 分别表示参考光谱和重建光谱。为了便于比较, S_R 向下移动0.04单位。

Fig.2 Target factor analysis (TFA) results for (A) data set 1, (B) data set 2 and (C) data set 3

S_0 and S_R represent the reference spectrum and the reconstruction spectrum

For clarity, S_R is moved downwards by 0.04 unit.

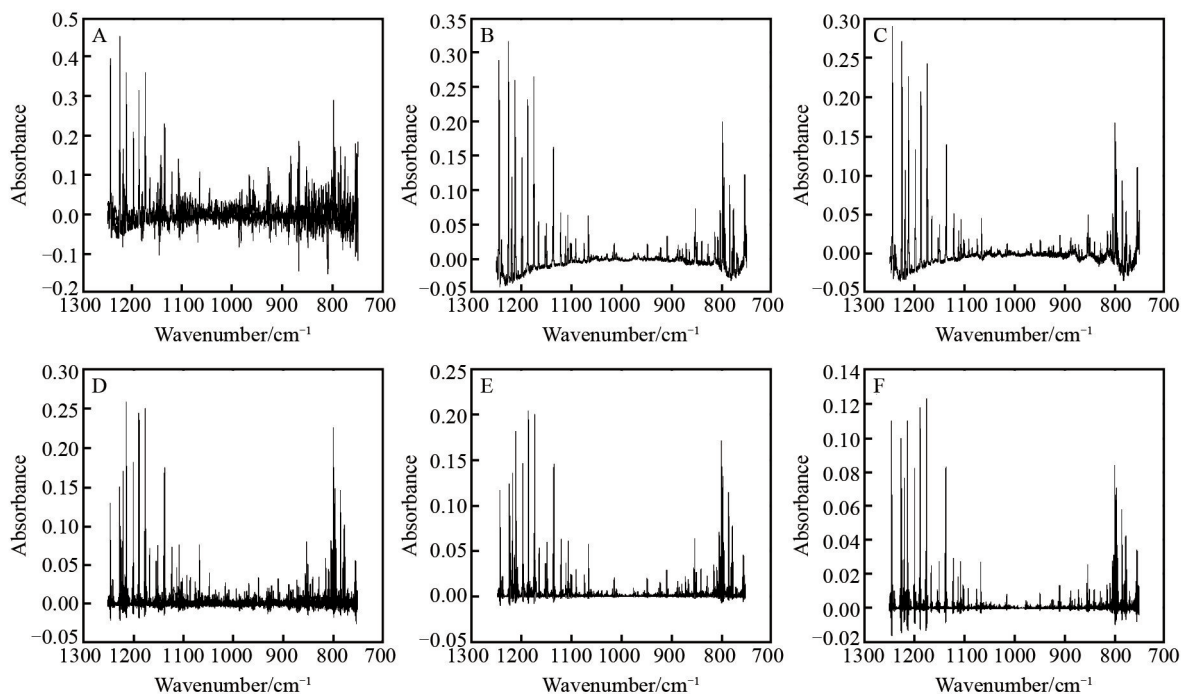


图3 3个数据集基线校正后的光谱图:(A) 数据集1、(B) 数据集2和(C) 数据集3的WT校正结果;
(D) 数据集1、(E) 数据集2和(F) 数据集3的airPLS校正结果

Fig.3 Spectra of the three data sets after baseline correction; results of wavelet transform (WT) of (A) Data set 1, (B) data set 2 and (C) data set 3; Results of adaptive iterative reweighted penalty least squares (airPLS) of (D) data set 1, (E) data set 2 and (F) data set 3

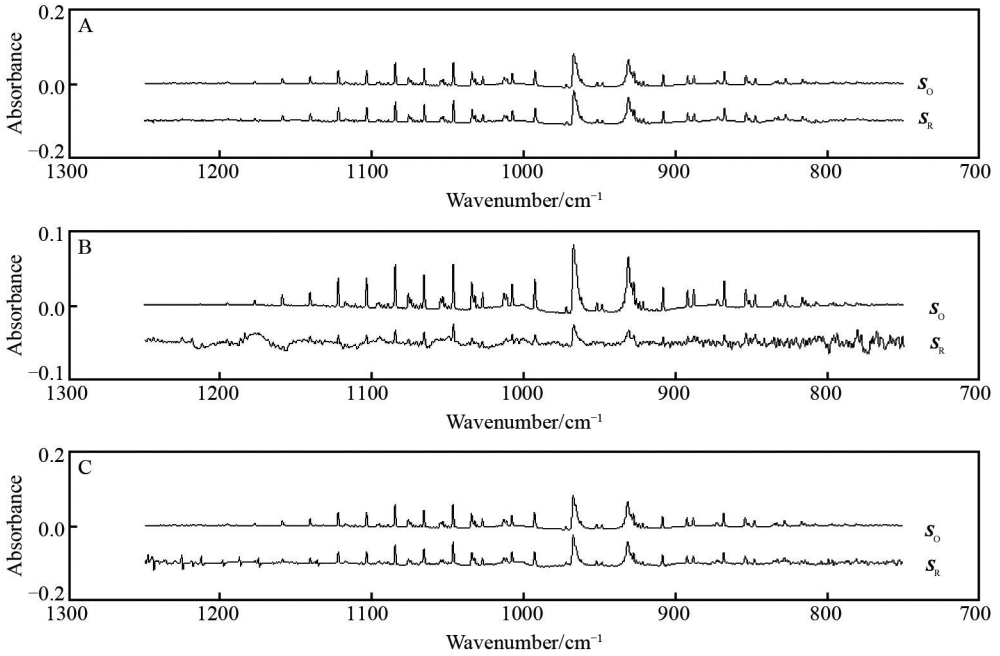


图 4 WT 校正后的(A)数据集 1、(B)数据集 2 和(C)数据集 3 的 TFA 结果

S_0 和 S_R 分别表示参考光谱和重建光谱。为了便于比较,(A)和(C)中的 S_R 向下移动 0.1 单位,(B)中的 S_R 向下移动 0.05 单位。

Fig. 4 TFA results for (A) data set 1, (B) data set 2 and (C) data set 3 after WT

S_0 and S_R represent the reference spectrum and the reconstruction spectrum, respectively. For clarity, S_R in (A) and (C) are moved downwards by 0.1 unit, S_R in (B) is moved downwards by 0.05 unit.

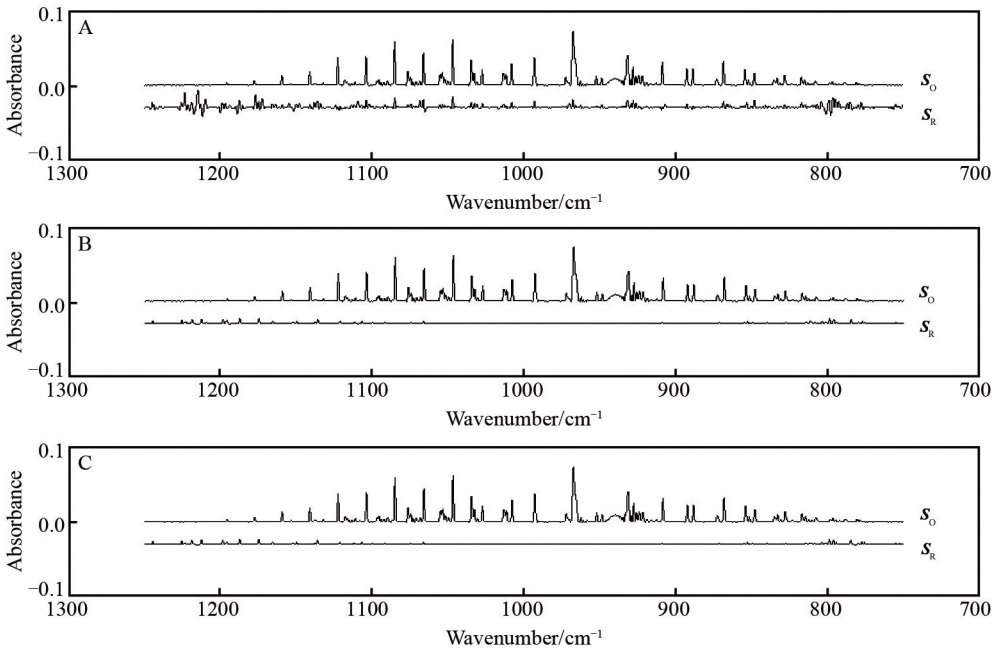


图 5 airPLS 校正后的(A)数据集 1、(B)数据集 2 和(C)数据集 3 的 TFA 结果

S_0 和 S_R 分别表示参考光谱和重建光谱。为了便于比较, S_R 向下移动 0.03 单位。

Fig. 5 TFA results for (A) data set 1, (B) data set 2 and (C) data set 3 after airPLS

S_0 and S_R represent the reference spectrum and the reconstruction spectrum, respectively. For clarity, S_R is moved downwards by 0.03 unit.

表2 不同方法校正后的目标因子分析(TFA)结果

Table 2 TFA results corrected by different methods

序号 No.	浓度方差 Concentration variance	正交投影 最小二乘 OPALS	加权相关系数 WCC					自适应迭代 重加权惩罚 最小二乘 airPLS
			原始数据 Original data	多元散射校正 MSC	标准正态变换 SNV	线性拟合 linear fitting	小波变换 WT	
1	226.8	4	0.75	0.96	0.99	0.97	0.99	0.68
2	0.45	6	0.60	0.60	0.67	0.64	0.91	0.012
3	4.3	5	0.66	0.95	0.93	0.62	0.99	0.62
4	8.4	7	0.79	0.95	0.96	0.79	0.83	0.52
5	134610	2	0.094	0.12	0.85	0.94	0.94	0.49
6	18663	2	-0.14	0.20	0.38	0.89	0.93	0.73

WCC, weighted correlation coefficient; OPALS, orthogonal projection approach least squares; MSC, multiplicative scatter correction; SNV, standard normal variate.

以上结果表明,不同类型的基线经 WT 校正后,TFA 的定性能力均可提高,不再出现假阴性结果,原因是基线校正后氨的光谱信息提前出现在 PCA 的特征向量中。如图 6 所示,数据集 1 的浓度方差大,

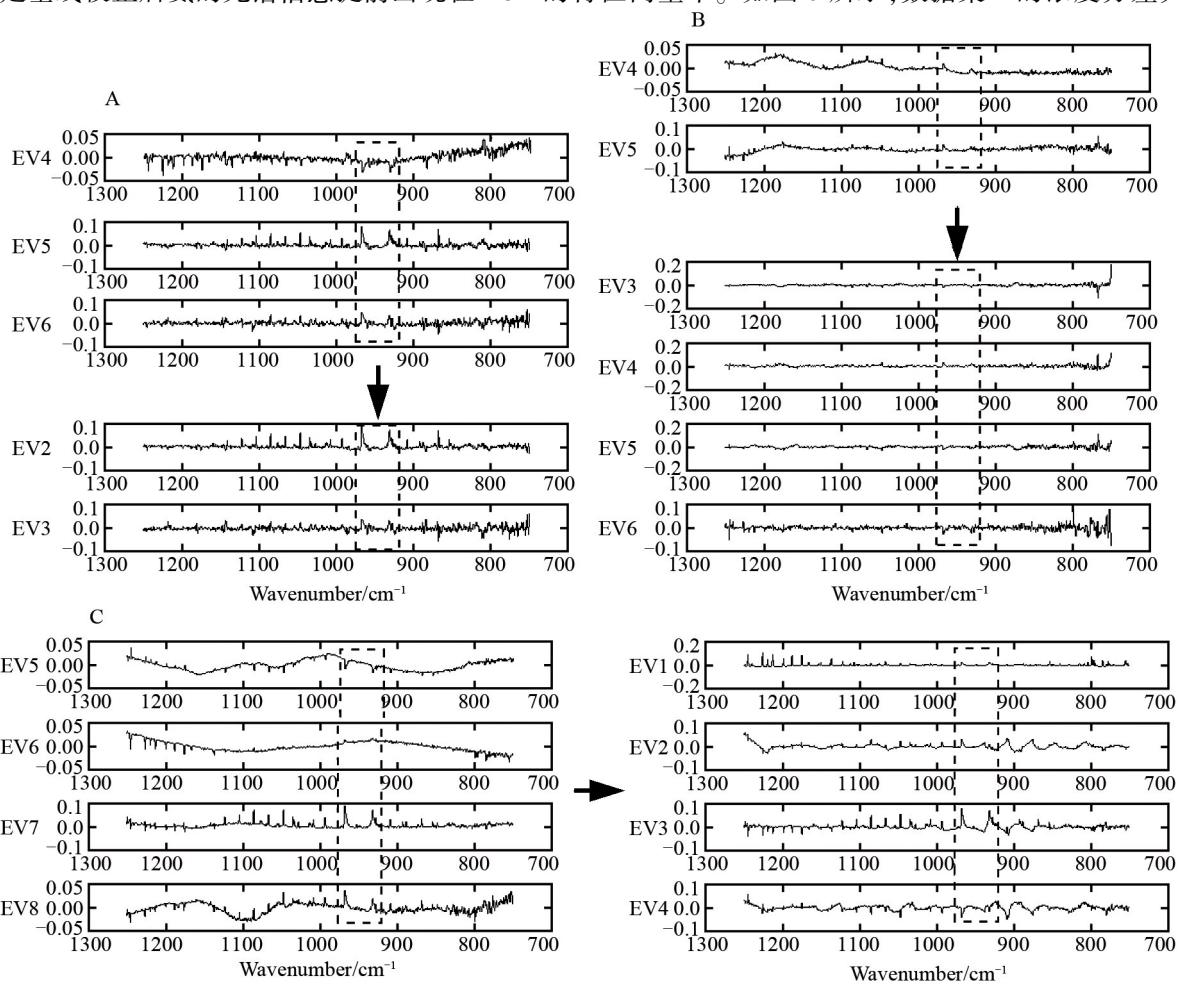


图6 (A) 数据集 1、(B) 数据集 2 和(C) 数据集 3 在小波变换校正前后的 PCA 的特征向量
箭头由原始光谱特征向量指向小波变换后的特征向量,其中虚线框标明氨的光谱特征。

Fig. 6 Eigenvectors of principal component analysis (PCA) of (A) data set 1, (B) data set 2 and (C) data set 3 before and after WT

The arrow points from the eigenvectors of the original spectra to the eigenvectors after WT, in which dotted boxes indicate the spectral characteristics of ammonia.

基线漂移严重,经过 WT 校正后,氨的光谱信息由 EV4、EV5、EV6 移到 EV2、EV3;数据集 2 的浓度方差极小,基线漂移小,经过 WT 校正后,氨的光谱特征由 EV4、EV5 移到 EV3、EV4、EV5、EV6;数据集 3 的浓度方差小,基线漂移较严重,经过 WT 校正后,氨的光谱信息由 EV5、EV6、EV7、EV8 移到 EV1、EV2、EV3、EV4。由于基线信息被 WT 移除,氨的光谱信息进入位置靠前的特征向量,因而被 TFA 成功提取,从而避免了假阴性结果。

4 结论

TFA 在大气 OP/FTIR 光谱定性分析中具有重要的作用。本研究选择基线校正的方法减少 TFA 中出现的假阴性结果,首先选择 3 组不同浓度方差、不同漂移程度的氨的数据集,再采用 OPALS 方法判断主因子数,最后分别使用 SNV、MSC、线性拟合、airPLS 和 WT 方法进行基线校正,并使用 3 组甲烷光谱进行验证。结果表明,使用小波变换进行预处理可以进一步提高 TFA 的定性分析能力。本方法为多组分体系的快速定性分析提供了一种有效途径。

References

- [1] SAUER U, BORSODORF H, DIETRICH P, LIEBSCHER A, MOLLER I, MARTENS S, MOLLER F, SCHLOMER S, SCHUTZE C. *Environ. Monit. Assess.*, 2018, 190(3): 114.
- [2] CHEN M J, DUH J M, SHIE R H, WENG J H, HSU H T. *Indoor Air*, 2016, 26(3): 457-467.
- [3] LIN C H, GRANT R H, HEBER A J, JOHNSTON C T. *Atmos. Meas. Tech.*, 2019, 12(6): 3403-3415.
- [4] PHILLIPS F A, NAYLOR T, FOREHEAD H, GRIFFITH D W T, KIRKWOOD J, PATON-WALSH C. *Atmosphere*, 2019, 10(4): 208.
- [5] JIAO L Z, GUO Y M, CHEN J, ZHAO X D, DONG D M. *Food Res. Int.*, 2019, 119: 968-973.
- [6] BYRNE B, STRONG K, COLEBATCH O, YOU Y, WUNCH D, ARS S, JONES D B A, FOGAL P, MITTERMEIER R L, WORTHY D, GRIFFITH D W T. *Atmos. Ocean*, 2020, 58(1): 25-45.
- [7] MALINOWSKI E R. *Factor Analysis in Chemistry*. New York, 2002.
- [8] SCHWEITZER R C, TREADO P J, OLKHOVYK O, ZBUR L. *J. Chemom.*, 2018, 32(10): e3038.
- [9] MSIMANGA H Z, LAM T T H, LATINWO N, SONG M K, TAVAKOLI N. *Spectrochim. Acta, Part A*, 2018, 192: 159-167.
- [10] REHAKOVA M, GAL L, BELOVICOVA M, ORAVEC M, DVONKA V, STOJKOVICOVA D, CEPPAN M. *J. Cult. Herit.*, 2017, 27: 137-142.
- [11] SHAO L M, GRIFFITHS P R. *Anal. Chem.*, 2007, 79(5): 2118-2124.
- [12] LEGER M N, RYDER A G. *Appl. Spectrosc.*, 2006, 60(2): 182-193.
- [13] LAN T G, FANG Y H, XIONG W, KONG C. *Chin. Opt. Lett.*, 2007, 5(10): 613-616.
- [14] PAN Zhong-Xiao, SHAO Xue-Guang, ZHONG Hong-Bo, LIU Wei, WANG Hong, ZHANG Mao-Sen. *Chin. J. Anal. Chem.*, 1996, 24(2): 149-153.
潘忠孝, 邵学广, 仲红波, 刘卫, 王洪, 张懋森. *分析化学*, 1996, 24(2): 149-153.
- [15] ZHANG Z M, CHEN S, LIANG Y Z. *Analyst*, 2010, 135(5): 1138-1146.
- [16] SHAO X G, LEUNG A K M, CHAU F T. *Acc. Chem. Res.*, 2003, 36(4): 276-283.
- [17] SHAO L M, GRIFFITHS P R. *Appl. Spectrosc.*, 2009, 63(8): 916-919.
- [18] HAO S L, SHAO L M. *Chemom. Intell. Lab. Syst.*, 2015, 149: 17-23.
- [19] MALLAT S G. *IEEE. Trans. Pattern Anal. Machine Intell.*, 1989, 11(7): 674-693.
- [20] ZHU C J, GRIFFITHS P R. *Appl. Spectrosc.*, 1998, 52(11): 1403-1408.
- [21] SHAO L M, GRIFFITHS P R. *Anal. Chem.*, 2008, 80(13): 5219-5224.

Influence of Baseline on Target Factor Analysis Results of Fourier Transform Infrared Spectroscopy

WANG Li, SHAO Li-Min*

(Department of Chemistry, University of Science and Technology of China, Hefei 230026, China)

Abstract The successful extraction of target spectra from the experimental data matrix by target factor analysis (TFA) without any preprocessing mainly depends on the concentration variance of the corresponding component. When the variance of ammonia concentrations is low and the baseline drift is serious, TFA may yield false negative results, so baseline correction is applied to improve TFA performance. In this study, six sets of data with different concentration variances and baseline drift were used to compare the baseline removal effects of standard normal variate (SNV), multiplicative scatter correction (MSC), linear fitting, wavelet transform (WT) and adaptive iterative reweighted penalty least squares (airPLS). The results showed that WT could effectively remove the baseline, so that more spectral information of ammonia and methane was allocated in the first eigenvectors of principal component analysis (PCA). The allocated information of ammonia and methane could be successfully extracted by TFA.

Keywords Target factor analysis; Principal factor number; Baseline correction; Wavelet transform

(Received 2020-10-21; accepted 2020-12-30)

安捷伦推出用于低丰度免疫细胞代谢分析的高灵敏度 XF 分析解决方案 提高数据质量、简化工作流程、使用更少细胞进行更多功能化的分析

2021年2月17日,北京——安捷伦科技公司(纽约证交所:A)今日推出专为安捷伦 Seahorse XF HS Mini 分析仪设计的安捷伦 Seahorse XF HS 迷你板,可用于提高免疫细胞代谢分析。

免疫学和疾病研究人员越来越多地使用稀有的体外基因工程细胞来建立更好的疾病模型。然而,此类细胞的生产数量有限,限制了研究人员可进行的细胞分析类型。

XF HS Mini 是安捷伦 Seahorse XF 平台系列的最新成员,可实时分析活细胞中的线粒体呼吸、糖酵解和 ATP 生成。这些代谢测量使研究人员能更充分地了解细胞的健康状况、功能和信号转导。

高度灵敏的 XF HS Mini 分析仪可提高性能和精度、减少每孔所需细胞数量、改善悬浮细胞工作流程一致性,并简化分析。这些改进使研究人员能从免疫细胞等数量有限或呼吸速率低的细胞类型中可靠地生成 XF 数据,进行以往无法完成的测量。

斯坦福大学干细胞移植与再生医学系儿科学副教授 Katja Weinacht 医学博士说道:“我们使用的是经过高度操纵的免疫细胞,其生命周期较短,生成成本较高且耗时费力。以更少的细胞数量获得更高的灵敏度是成败的关键,因此我们在疾病模型中使用安捷伦 Seahorse XF 技术。”

安捷伦细胞分析事业部高级总监 David Ferrick 博士表示:“随着我们的客户努力在更复杂、更特殊的体内环境中进行生物学探究,对稀有细胞群的研究需求已愈发明确。XF HS Mini 更高的灵敏度和精度将为客户开辟代谢分析的新领域。”