第32卷，第4期
2012年4月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 32, No. 4, pp1038-1042
April, 2012

直接正交信号校正算法在烷烃类多组分气体定量分析中的应用

李玉军[1,2]，汤晓君[2]，刘君华[2]

1. 西安理工大学自动化与信息工程学院，陕西 西安 710048
2. 西安交通大学电气工程学院，陕西 西安 710049

摘 要 针对烷烃类多组分混合气体中红外光谱存在的基线漂移问题，提出一种直接正交信号校正算法用于光谱数据预处理。实验中采用傅里叶变换红外光谱仪采集了936组混合气体样本的光谱数据，混合气体主要由不同浓度范围的七种组分气体组成。将直接正交信号校正算法与导数算法进行了对比分析，采用偏最小二乘回归方法建立了各组分气体定量分析模型，并对模型参数(主元个数、导数步长及正交分量的个数)进行了遍历优化选取最优分析模型。结果表明直接正交信号校正算法对于中红外光谱基线校正效果最好，直接正交信号校正算法用于烷烃类混合气体中红外光谱基线校正可行，效果良好，具有一定的实用和研究价值。

关键词 基线漂移；直接正交信号校正；导数算法；偏最小二乘回归；定量分析；混合气体

中图分类号：O657.33 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2012)04-1038-05

引 言

由于傅里叶变换红外光谱仪在使用过程中会受到各种外界环境干扰因素的影响，使所获取的红外光谱数据偏离基线而产生漂移，这对于后续分析非常不利，因此针对烷烃类混合气体的红外光谱数据在分析建模前进行基线校正格外重要，目的是通过对光谱数据的适当处理与变换，减弱以至消除各种非目标因素对光谱的不利影响，尽可能多地去除各种无关信息变量，提高分辨率与灵敏度，从而提高模型的分析准确度及稳健性[1]。导数算法简单易实现，可用于消除光谱信号中背景噪声、基线平移及漂移(散射)对建立分析模型的不利影响，在化学计量学[2]、食品科学、医药医学、生物学等众多领域都得到了广泛应用，且取得了较好效果。但此方法仅对光谱阵进行处理，并未考虑浓度阵的影响，有可能损失部分有用信息。而直接正交信号校正算法[4-7]能够很好的解决上述问题，虽比导数算法复杂，但由于其将光谱阵与浓度阵正交处理之后进行主成分提取，滤除了光谱阵中的冗余信息使得处理效果更好。目前该算法在烷烃类多组分混合气体定量分析领域未见报道。

针对七组分烷烃类混合气体中红外光谱数据中存在的光谱基线漂移问题，提出利用直接正交信号校正算法进行基线校正，采用偏最小二乘回归方法建立各组分气体分析模型，并将其与一阶[2,8]和二阶导数[3]校正算法进行比较，实验结果表明该方法能够更多的保留有用信息，提高信噪比，基线校正后各组分气体分析模型输出计算结果的总体平均相对误差比基线校正前降低66.80%，比一阶导数校正算法降低51.51%，比二阶导数算法降低56.30%，说明该方法能够有效的对中红外光谱数据进行基线校正，具有一定的实用性。

1 偏最小二乘回归算法简介

偏最小二乘回归(partial least-squares regression，PLSR)算法是由Wold及Albano等在1983年为了解决化学样本分析中自变量存在多重相关问题或自变量数量多于样本点数量的实际问题而提出的一种多元分析方法[9]。

针对于七组分烷烃类混合气体的建模，偏最小二乘算法主要是将训练集中的浓度矩阵$\boldsymbol{Y}_{\text{train}}$和相对应的训练集中的光谱阵$\boldsymbol{X}_{\text{train}}$同时进行主成分分解，其数学模型如下[10,11]

$$\boldsymbol{X}_{\text{train}} = \boldsymbol{TP} + \boldsymbol{E} \tag{1}$$

$$\boldsymbol{Y}_{\text{train}} = \boldsymbol{UQ} + \boldsymbol{F} \tag{2}$$

式中$\boldsymbol{T}$和$\boldsymbol{U}$分别是光谱阵$\boldsymbol{X}_{\text{train}}$和浓度阵$\boldsymbol{Y}_{\text{train}}$的得分阵，$\boldsymbol{P}$和$\boldsymbol{Q}$分别是其载荷阵，而$\boldsymbol{E}$和$\boldsymbol{F}$则分别是其偏差。由式(1)可推得训练集得分阵可由下式近似获取

$$\boldsymbol{T} = \boldsymbol{X}_{\text{train}}\boldsymbol{W} \tag{3}$$

式中$\boldsymbol{W}$为光谱权重矩阵，根据得分阵即可建立回归模型如

收稿日期：2011-08-10，**修订日期**：2011-12-10

基金项目：国家自然科学基金项目(50877056)资助

作者简介：李玉军，1977年生，西安交通大学电气工程学院博士生，讲师 e-mail：leowho@163.com

下

$$\boldsymbol{U}=\boldsymbol{TB} \tag{4}$$

式中 $\boldsymbol{B}$ 所表征的即为浓度阵和光谱阵之间的内在关系，称为回归矩阵

$$\boldsymbol{B}=(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{U} \tag{5}$$

由式(3)可得校验集浓度得分阵如下所示

$$\boldsymbol{T}_{\text{test}}=\boldsymbol{X}_{\text{test}}\boldsymbol{W} \tag{6}$$

根据回归阵 $\boldsymbol{B}$ 及校验光谱得分阵，即可按照下式求得校验集计算浓度阵

$$\boldsymbol{U}_{\text{test}}=\boldsymbol{T}_{\text{test}}\boldsymbol{B} \tag{7}$$

2　导数校正算法简介

导数校正算法是被广泛应用于光谱分析特别是近红外光谱分析中的光谱预处理算法，其一阶导数算法(first derivative algorithm，FDA)及二阶导数算法(second derivative algorithm，SDA)校正式分别如式(8)和(9)所示

$$\frac{\mathrm{d}T}{\mathrm{d}\lambda}\approx\frac{T_{\lambda_{i+n}}-T_{\lambda_i}}{\lambda_{i+n}-\lambda_i} \tag{8}$$

$$\frac{\mathrm{d}^2T}{\mathrm{d}^2\lambda}\approx\frac{T'_{\lambda_{i+n}}-T'_{\lambda_i}}{\lambda_{i+n}-\lambda_i}\approx\frac{T_{\lambda_{i+n}}-2T_{\lambda_i}+T_{\lambda_{i-n}}}{(\lambda_{i+n}-\lambda_i)^2} \tag{9}$$

式中 $\lambda_i(i=1,2,3,\cdots,m,\ m$ 为数据点总数)是混合气体中红外光谱数据中第 i 个数据点对应的中红外光波数，T_{λ_i} 是中红外光通过混合气体样本时在波数 λ_i 处的透过率，$n=1,2,3,\cdots,m$ 表示导数步长，式中采用了一阶后向差分及二阶后向差分来近似代替一阶导数及二阶导数。

3　直接正交信号校正算法简介

直接正交信号校正算法(direct orthogonal signal correction，DOSC)是 Westerhuis[5] 等在正交信号校正算法(orthogonal signal correction，OSC)[12] 基础上提出的一种改进型算法，基本思想就是在建立分析模型前，将光谱阵与浓度阵正交，滤除光谱阵中与浓度阵无关的信息，然后再建立分析模型，达到简化模型，提高模型分析准确度的目的[1]。其算法较之 OSC 更加简单有效，简单介绍如下[4-7]：

(1) 首先求取训练集浓度阵 $\boldsymbol{Y}_{\text{train}}$ 在光谱阵 $\boldsymbol{X}_{\text{train}}$ 所张开的线性空间的投影，$\boldsymbol{Y}_{pj}=\boldsymbol{X}'_{\text{train}}((\boldsymbol{X}'_{\text{train}})^{-1})'\boldsymbol{Y}_{\text{train}}$；

(2) 计算 $\boldsymbol{X}_{\text{train}}$ 在 $\boldsymbol{Y}_{pj}$ 的正交补空间的投影，$\boldsymbol{A}_{yx}=\boldsymbol{X}_{\text{train}}-\boldsymbol{Y}_{pj}\boldsymbol{Y}_{pj}^{-1}\boldsymbol{X}_{\text{train}}$；

(3) 对 $\boldsymbol{A}_{yx}\boldsymbol{A}'_{yx}$ 进行主成分提取，取前 n 个需正交处理的主成分的得分阵 $\boldsymbol{T}$；

(4) 通过回归计算权重矩阵 $\boldsymbol{w}=\boldsymbol{X}_{\text{train}}^{-1}\boldsymbol{T}$；

(5) 计算新的得分矩阵 $\boldsymbol{T}_{\text{new}}=\boldsymbol{X}_{\text{train}}\boldsymbol{w}$；

(6) 计算载荷矩阵 $\boldsymbol{P}=\boldsymbol{X}'_{\text{train}}\boldsymbol{T}/(\boldsymbol{T}'\boldsymbol{T})$；

(7) 去除正交信号之后即可得到新的光谱阵为：$\boldsymbol{X}_{\text{new}}=\boldsymbol{X}_{\text{train}}-\boldsymbol{T}\boldsymbol{P}'$；

对于校验集光谱数据 $\boldsymbol{X}_{\text{test}}$ 根据载荷矩阵 $\boldsymbol{P}$ 及权重矩阵 $\boldsymbol{w}$ 即可求出校正后的校验集光谱 $\boldsymbol{T}_{\text{test}}=\boldsymbol{X}_{\text{test}}\boldsymbol{w}$，$\boldsymbol{X}_{\text{osctest}}=\boldsymbol{X}_{\text{test}}-\boldsymbol{T}_{\text{test}}\boldsymbol{P}$。

3　实验数据的获取

实验中所用光谱数据均采用 Tensor27 中红外傅里叶光谱仪获取，该光谱仪由德国 Bruker 公司生产。采用该光谱仪分别对浓度为 1%的甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷、正戊烷单组分气体进行标定，获得的光谱数据合放在同一张图中，得到波数分布在 4 000～400 cm^{-1} 范围内的中红外光谱数据如图 1 所示。

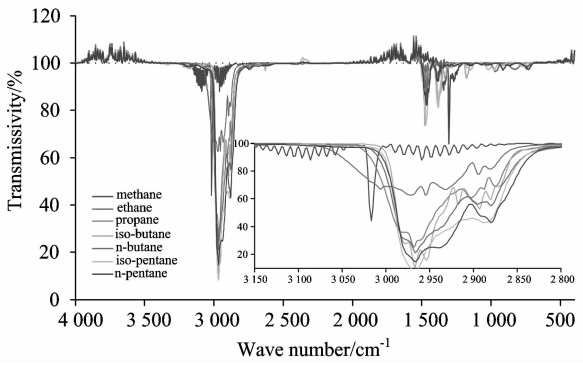

Fig. 1　Spectrum of single component gas

由图 1 中主吸收峰区域的光谱放大图可以看出各组分气体吸收峰除了甲烷组分气体外基本重合在一起，吸收峰交叠严重，因此对于七组分烷烃类混合气体的定量分析是非常困难的。实验中采用该光谱仪对浓度分布范围在 0.01%～0.1%的甲烷、乙烷，0.01%～0.15%的丙烷，0.0%～0.1%异丁烷、正丁烷，0.0%～0.05%异戊烷、正戊烷组成的混合气体进行标定得到共计 936 组样本数据，取其中的一半数据做训练样本，全部做校验样本。由于在获取混合气体样本光谱数据过程中，光谱仪会受到外界环境干扰因素影响而产生基线漂移，如图 2 所示。

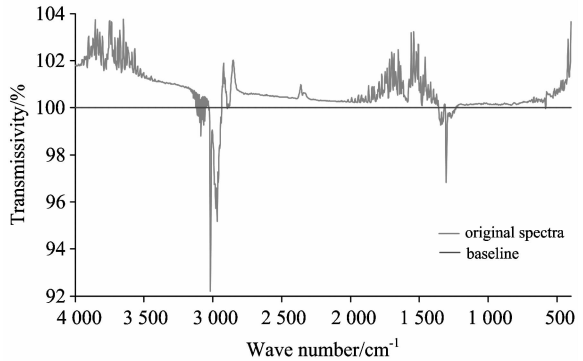

Fig. 2　Baseline departure spectrum of gas mixture

由图 2 可以看出该混合气体光谱偏离基准线幅度较大，漂移是很严重的。混合气体光谱若存在漂移偏离基准线，则无法正确反映其真实透过率，从而对建立的分析模型的准确度产生不利影响，因此对光谱数据进行基线校正则意义重大。

5 结果与分析

5.1 光谱数据校正前模型优化

为了对比光谱基线校正前后的处理效果，首先采用 PLSR 方法根据基线校正前光谱数据建立各组分气体分析模型，为了建立最优分析模型，实验中选取分析模型计算结果平均相对误差(mean relatively error，MRE)作为衡量指标，计算公式如下

$$\mathrm{MRE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-y'_i}{y'_i}\right| \tag{10}$$

式中 y_i 是分析模型计算结果，y'_i 是分析模型的期望输出，n 为样本数。对 PLSR 算法中提取主元个数由 1～30 进行了遍历优化选取，可得各组分最优分析模型计算结果误差如表 1 所示。

Table 1 Error result of optimal analysis model

组分气体	所用时间/s	提取主元个数	测试结果/%
甲烷	28.58	4	28.22
乙烷	28.97	4	49.30
丙烷	28.98	5	56.17
异丁烷	29.02	7	51.51
正丁烷	29.42	7	49.06
异戊烷	29.08	4	47.06
正戊烷	29.08	4	68.22

从表 1 中可以看出，虽然经过了模型优化，最终模型的分析准确度还很差。

5.2 导数算法处理效果

采用导数算法对光谱数据进行基线校正，利用校正之后光谱数据采用 PLSR 方法建立分析模型，为了得到最优分析模型同样需要对导数步长及 PLSR 算法中的主元个数进行优化选取，实验中分别对提取的主元个数由 1～30，导数步长由 1～30 进行遍历优化选取，可得经过一阶及二阶导数算法校正之后最优分析模型计算结果误差如表 2 和表 3 所示。

由表 2 和表 3 可以看出，经过导数算法校正之后模型的分析准确度得到一定程度的改善，但效果不是特别显著，为了便于分析将算法校正前后分析模型的计算误差同放于一张表，如表 4 所示。

Table 2 Error result of optimal analysis model that pretreated by first derivative algorithm

组分气体	所用时间/s	导数步长	提取主元个数	测试结果/%
甲烷	844.14	27	3	23.48
乙烷	847.83	10	7	32.12
丙烷	844.19	29	21	35.53
异丁烷	861.94	24	11	27.31
正丁烷	862.86	5	27	29.21
异戊烷	841.77	3	3	36.90
正戊烷	831.44	3	3	54.78

Table 3 Error result of optimal analysis model that pretreated by second derivative algorithm

组分气体	所用时间/s	导数步长	提取主元个数	测试结果/%
甲烷	857.28	28	3	26.40
乙烷	825.17	20	7	34.05
丙烷	827.42	27	12	36.85
异丁烷	829.28	22	3	28.92
正丁烷	852.34	22	3	33.30
异戊烷	819.14	25	2	43.66
正戊烷	816.41	25	2	62.40

Table 4 Comparison of pretreatment effect of derivative spectra

组分气体	处理前 MRE/%	FDA 处理后		SDA 处理后	
		MRE/%	误差降低/%	MRE/%	误差降低/%
甲烷	28.22	23.48	16.80	26.40	6.45
乙烷	49.30	32.12	34.85	34.05	30.93
丙烷	56.17	35.53	36.75	36.85	34.40
异丁烷	51.51	27.31	46.98	28.92	43.86
正丁烷	49.06	29.21	40.46	33.30	32.12
异戊烷	47.06	36.90	21.59	43.66	7.22
正戊烷	68.22	54.78	19.70	62.40	8.53

由表 4 可以看出，一阶导数算法(FDA)对烷烃类混合气体光谱数据处理效果比二阶导数算法(SDA)好，采用 FDA 处理后分析模型的计算结果相对误差平均比处理前降低了 31.02%，而采用 SDA 处理之后仅比处理前降低了 23.36%。FDA 在处理丁烷时效果最好，但总体上其误差水平还是偏高。为了便于观察 FDA 对光谱数据处理效果，将图 2 中的原始光谱与经过 FDA 处理之后的光谱对比如图 3 所示(选取导数步长为 27)。

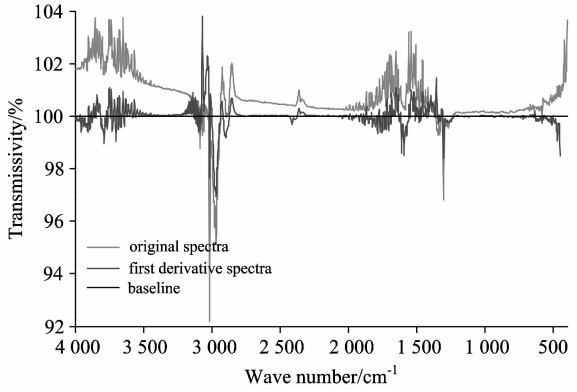

Fig. 3 First derivative spectrum of mixture gas

由图 3 可以看出，原始光谱的基线漂移得到校正，但光谱数据的透光率有所减弱，说明光谱数据中的有用信息有损失。

5.3 直接正交信号校正算法处理效果

采用直接正交信号校正(DOSC)算法进行光谱数据基线校正，将校正之后的光谱数据利用 PLSR 方法建立组分气体

分析模型，为了得到最优分析模型，同样需要对算法中的参数进行优化选取。分别对 DOSC 算法中剔除的正交成分个数由 1～30，PLSR 算法中提取的主元个数由 1～30 进行遍历优化选取，可得经过 DOSC 算法校正后分析模型误差结果如表 5 所示。

Table 5　Error result of optimal analysis model that pretreated by DOSC algorithm

组分气体	所用时间/s	剔除正交成分个数	提取主元个数	测试结果/%
甲烷	1 079.34	7	1	6.80
乙烷	1 081.84	12	3	6.72
丙烷	1 081.20	27	7	10.36
异丁烷	1 080.13	21	1	17.04
正丁烷	1 075.39	6	5	22.38
异戊烷	1 071.19	27	14	20.26
正戊烷	1 097.42	12	1	32.49

由表 5 所示，经过 DOSC 算法校正后建立的分析模型准确度得到较大幅度的提升，但由于算法较导数算法复杂一些，所以优化时间较长。将图 2 中的原始光谱与经过 DOSC 算法处理之后的光谱对比如图 4 所示(选取剔除主元个数为 7)。

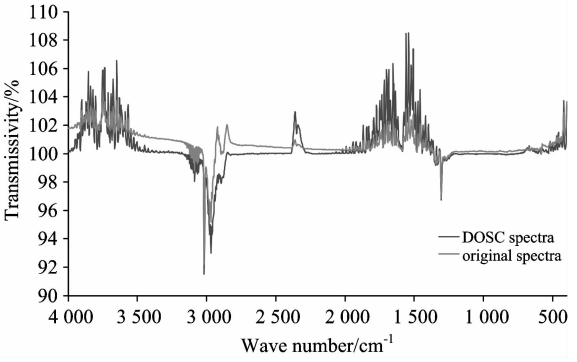

Fig. 4　DOSC spectrum of mixture gas

由图 4 可以看出，原始光谱数据的基线漂移得到有效校正，但背景噪声(水蒸气及二氧化碳吸收波长区域)的信号也有所放大，背景噪声主要是由于扫描背景时气室内残留有一定浓度二氧化碳气体及水蒸气所造成。由于背景噪声所在光谱区域与组分气体吸收峰区域光谱区别明显，且可以采用其他技术手段进行处理，所以对分析建模影响较小。对比图 4 与图 3 可以看出，经过 DOSC 算法校正之后光谱数据在组分气体吸收峰区域的有用信息在被校正的同时损失较少，表明此时有用光谱数据的信噪比要好于一阶导数算法处理得到的光谱数据。为了便于比较，将 FDA，SDA 及 DOSC 算法处理前后分析模型的误差结果对比列于表 6。

Table 6　Comparison of spectra calibration effect

组分气体	处理前 MRE/%	FDA 处理后 MRE/%	SDA 处理后 MRE/%	DOSC 处理后 MRE/%
甲烷	28.22	23.48	26.40	6.80
乙烷	49.30	32.12	34.05	6.72
丙烷	56.17	35.53	36.85	10.36
异丁烷	51.51	27.31	28.92	17.04
正丁烷	49.06	29.21	33.30	22.38
异戊烷	47.06	36.90	43.66	20.26
正戊烷	68.22	54.78	62.40	32.49
平均误差	49.93	34.19	37.94	16.58

由表 6 对比可以看出，DOSC 算法对光谱数据的处理效果要明显优于导数算法，经过 DOSC 算法处理后建立的各组分分析模型的平均误差为 16.58% 比光谱数据处理前的 49.93%下降 66.80%，比 FDA 处理后平均误差 34.19%下降 51.51%，表明 DOSC 算法具有明显的优势。

6　结　论

由以上对比分析可以看出，DOSC 算法用于含烃类混合气体中红外光谱数据基线校正是完全可行的，处理效果较好。由于 DOSC 算法在对光谱数据进行处理过程中引入了浓度阵，使得光谱数据中有用信息损失较少，基线漂移得到有效校正，提高了有用光谱数据信噪比，但同时背景噪声有所放大，虽然水蒸气及二氧化碳所引起的背景噪声不会对组分分析模型建立产生直接不利影响，但也会有一定影响，所以对该光谱数据的进一步处理也是有必要的，例如选取合适波段光谱数据或采用滤波算法对光谱数据进行处理可能会进一步提高组分气体分析模型准确度。

References

[1] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立，袁洪福，陆婉珍). Progress in Chemistry(化学进展), 2004, 16(4): 528.

[2] CHEN Hong-lei, CHEN Yuan-cai, ZHAN Huai-yu, et al(陈洪雷，陈元彩，詹怀宇，等). Journal of South China University of Technology • Natural Science Edition)(华南理工大学学报 • 自然科学版)), 2009, 37(10): 150.

[3] CHEN Jie-mei, PAN Tao, CHEN Xing-dan(陈洁梅，潘　涛，陈星旦). Optics and Precision Engineering(光学精密工程), 2006, 14(1): 1.

[4] Luypaert J, Heuerding S, Massart D L, et al. Analytica Chimica Acta, 2007, 582(1): 181.

[5] Westerhuis J A, de Jong S, Smilde A K. Chemometrics and Intelligent Laboratory Systems, 2001, 56(1): 13.

[6] Zhu D, Ji B, Meng C, et al. Chemometrics and Intelligent Laboratory Systems, 2008, 90(2): 108.

[7] ZHU Shi-ping, WANG Gang, YIN Xiong, et al(祝诗平，王　刚，尹　雄，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2008, 39(4): 104.

[8] DU Shu-xin, DU Yang-feng, WU Xiao-li(杜树新, 杜阳锋, 武晓莉). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2010, 30(12): 3268.
[9] WANG Hui-wen, WU Zai-bin, MENG Jie(王惠文, 吴载斌, 孟 洁). Partial Least Squares Regression-Linear and Nonlinear Methods(偏最小二乘回归的线性与非线性方法). Beijing: National Defense Industry Press(北京:国防工业出版社), 2006. 1.
[10] GUO Chun-xiao(郭纯孝). Computational Chemistry(计算化学). Beijing: Chemical Industry Press(北京: 化学工业出版社), 2004. 306.
[11] XU Lu, SHAO Xue-guang(许 禄, 邵学广). Methods of Chemometrics(化学计量学方法). Beijing: Science Press(北京:科学出版社), 2004. 166.
[12] Wold S, Antti H, Lindgren F, et al. Chemometrics and Intelligent Laboratory Systems, 1998, 44(1-2): 175.

Application of Direct Orthogonal Signal Correction Algorithm in Multi-Component Alkane Quantitative Analysis

LI Yu-jun[1,2], TANG Xiao-jun[2], LIU Jun-hua[2]
1. Faculty of Automation & Information Engineering, Xi'an University of Technology, Xi'an 710048, China
2. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Abstract According to the baseline departure of multi-component alkane gas mixture spectra, direct orthogonal signal correction (DOSC) algorithm was proposed to pretreat the infrared spectra data. Fourier transform infrared (FTIR) spectrometer was used to sample 936 spectra data of seven components gas mixture, including methane, ethane, propane, iso-butane, n-butane, iso-pentane and n-pentane gases. The concentration of each component ranges from 0.01% to 0.1%, 0.01% to 0.1%, 0.01% to 0.15%, 0.0% to 0.1%, 0.0% to 0.1%, 0.0% to 0.05%, and 0.0% to 0.05%, respectively. For analyzing intuitively, partial least square regression (PLSR) was introduced to build the component gas quantitative analysis model. In experiment, DOSC method was compared with first derivative algorithm (FDA) and second derivative algorithm (SDA). In order to get the optimal model, ergodic optimization method was used to select the optimal parameters of the model, i.e. the step of the derivative algorithm, the number of the primary component of the PLSR and the number of orthogonal components of the DOSC algorithm. The experiment results show that DOSC algorithm has the better effect in the field of infrared spectra data pretreating. The average mean relative error (MRE) of the component gas analysis models is 16.58%, which declined by 66.80% from the average MRE before data pretreating 49.93%. Compared with DA, the average MRE declined by 51.51% from 34.19% after pretreated by FDA, and declined by 56.30% from 37.94% after pretreated by SDA. So DOSC method is feasible to pretreat the IR spectra data, and has definite practical and investigation value.

Keywords Baseline departure; Direct orthogonal signal correction; Derivative algorithm; Partial least square regression; Quantitative analysis; Gas mixture

(Received Aug. 10, 2011; accepted Dec. 10, 2011)