文章编号：1001-3806(2017)01-0133-05

随机抽样一致性算法在激光光谱中的应用研究

谢珊珊，王哲强，黄 河，陈宝宝，汪 培，李劲松*
（安徽大学 物理与材料科学学院，合肥 230601）

摘要：为了解决波长调制激光光谱技术探测大气痕量气体浓度中信号处理算法的不足，提出了一种基于随机抽样一致性算法的气体浓度反演算法。以大气甲醛分子的仿真信号和实际测量信号为例，进行了理论分析和实验研究，并与传统的最小二乘法相比较。结果表明，该算法具有较强的抗噪声和异常点干扰能力，尤其是在低信噪比的条件下，精确度可提高 1 个量级，体现出较高的可靠性和优越性。

关键词：信号处理；随机抽样一致性；最小二乘法；气体浓度反演；激光光谱
中图分类号：O433.1 **文献标志码**：A **doi**：10.7510/jgjs.issn.1001-3806.2017.01.027

Applications of random sample consistency algorithm on laser spectroscopy

XIE Shanshan, *WANG Zheqiang*, *HUANG He*, *CHEN Baobao*, *WANG Pei*, *LI Jingsong*
(School of Physics and Materials Science, Anhui University, Hefei 230601, China)

Abstract: In order to solve the insufficient of signal process algorithms during the detection of atmospheric trace gas concentration by wavelength modulation laser spectroscopy technique, a new method of gas concentration inversion based on the random sample consistency (RANSAC) algorithm was proposed. By choosing the simulation signal and the actual measurement signal of formaldehyde in the atmosphere as examples, theoretical analysis and experimental study were carried out and compared with the traditional least square method. The results show that the proposed algorithm has better immunity to noises and outliers. Especially under the conditions of low signal-to-noise ratio (SNR), the measurement accuracy can be improved by one order of magnitude. The algorithm shows better reliability and superiority.

Key words: signal processing; random sample consistency; least square method; gas concentration inversion; laser spectroscopy

引 言

自 20 世纪 80 年代以来，随着经济的发展，全球环境问题日益突出，如温室效应、臭氧层破坏、雾霾、酸雨和光化学烟雾等，严重威胁着全球人类的生存和发展。发展大气污染物的监测技术，为环境污染的监控、治理以及环境科学问题的研究提供必要的实验数据和理论支持，已成为环境科学家们的首要任务。可调谐半导体激光吸收光谱(tunable diode laser absorption spectroscopy，TDLAS)作为一种新型的非接触式气体浓度诊断技术，具有高分辨率和响应速度快等特点，从而被广泛地应用大气温室气体及其它大气痕量气体的测量研究[1-2]。TDLAS 的测量方法主要基于朗伯-比尔定律，通过检测气体分子对特定波长的激光吸收而产生的光强衰减量，结合一定的算法模型和已知实验参量，即可直接反演出被测分子的浓度、温度和速度等信息。该方法相对简单，但易受噪声影响，测量精度和灵敏度有限。鉴于半导体激光器的可调谐特性，1981 年，REID 和 LABRIE 提出将波长调制技术应用于 TDLAS[3]，利用高频正弦信号叠加到低频激光波长扫描信号中调制激光器，使得探测器探测到信号含有高频的谐波分量，进而通过锁相放大器进行提取，能有效地抑制 1/f 噪声，从而实现更高精确度的测量。

波长调制光谱中二次谐波探测技术灵敏度较高，因而被广泛地用于气体浓度的测量，但该技术首先需要对系统进行标定后才能反演出被测样品的浓度信息[4-5]。鉴于二次谐波信号与样品浓度之间的线性关系，目前采用的反演算法主要包括二次谐波信号峰值比值法和二次谐波信号整个线型轮廓线性回归分析法[6]。前者采用样品信号和参考信号之间单个峰值

基金项目：国家自然科学基金资助项目(61440010)；安徽大学创新训练计划资助项目(J18511120)；安徽省高等学校省级质量工程资助项目(2014tszy004)；安徽大学材料物理专业综合改革试点项目(2014zy007)

作者简介：谢珊珊(1995-)，女，硕士研究生，主要从事数字信号处理算法研究。
* 通讯联系人。E-mail：jingsong_li@ahu.edu.cn

收稿日期：2015-12-23；收到修改稿日期：2016-01-15

的比值,在低浓度的情况下(吸收信号较弱,噪声干扰明显)测量误差较大。后者利用整个二次谐波信号的有效吸收线型,结合最小二乘法拟合算法(least square method,LSM)[7-8],反演数据点增加,可有效降低测量误差,但是仍然受到非相关的噪声影响,导致拟合结果有所偏差。

本文中针对当前波长调制光谱技术中信号处理方法存在的问题,提出了一种改进的反演气体浓度算法,结合随机抽样一致性(random sample consensus,RANSAC)算法,用于高精度的大气痕量气体测量。首先介绍了相关算法的基本理论,通过将自行建立的RANSAC 算法模型应用到理论模拟和实验测量的数据处理中,对该算法模型的可靠性进行系统地评估。

1 RANSAC 算法原理

1.1 RANSAC 算法

RANSAC 算法最早由 FISCHLER 和 BOLLES 于 1981 年提出[9],作为一种迭代方法,用来在一组包含离群的被观测数据中估算出数学模型的参量。其主要原理是:输入一组实验数据,通过迭代反复选择数据中的一组随机子集(局内点),排除噪声(局外点),给出一个模型,最大概率的适用于局内点,算法流程见图 1[10-11]。

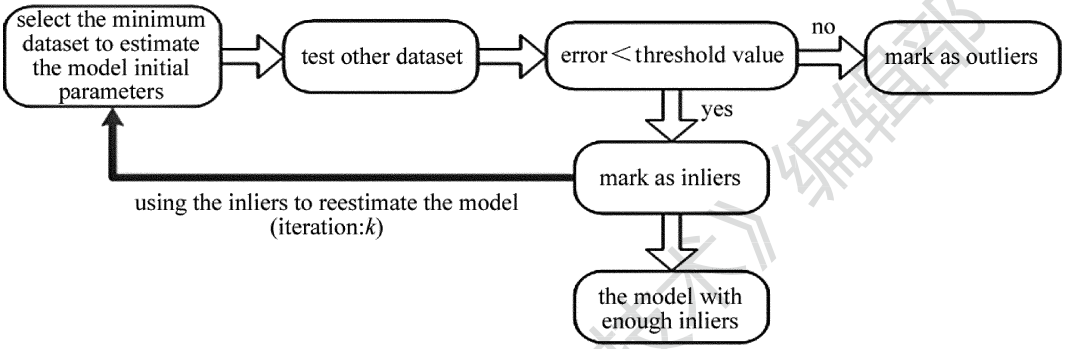

Fig. 1 The flowchart of RANSAC algorithm

RANSAC 算法计算过程中,要确定的参量有判断为内点的阈值 t、使得模型足够合理的内点数目 d,被重复执行上述流程的次数 k(迭代次数)可以从理论结果推断出。从估计模型参量时,用 p 表示在迭代过程中从数据集内随机选取出的点均为局内点的概率,用 w 表示每次从数据集中选取一个局内点的概率:w = 局内点的数目/数据集的数目。

假设估计模型需要选定 n 个点,w^n 是所有 n 个点均为局内点的概率;$(1-w^n)$ 是 n 个点中至少有一个点为局外点的概率,此时表明从数据集中估计出了一个不好的模型。$(1-w^n)^k$ 表示算法永远都不会选择到 n 个点均为局内点的概率,它和 $(1-p)$ 相同,即:$1-p=(1-w^n)^k$。两边取对数,即得到迭代次数:

$$k = \frac{\lg(1-p)}{\lg(1-w^n)} \tag{1}$$

阈值 t 的选取很重要,直接影响内点外点的判断[12]。因为在判断有效点的时候,若选取的 t 较小,则会放弃应该选择的有效点;而选取的 t 较大,则可能将异常点或误差点误判为有效点。针对该问题,本文中采用绝对中位差(median absolute deviation,MAD)D_{MAD}来估计数据的方差。假设选取的数据子集为 y_i,则其表达式为:

$$D_{\mathrm{MAD}} = \mathrm{median}_i(|y_i - \mathrm{median}_j(y_i)|) \tag{2}$$

式中,median 为求数组的中值函数,$|\cdot|$ 为求绝对值符号,i 和 j 分别为数据子集位置。阈值 t 取实验数据的绝对中位差,再用模型去测试其它实验数据,若数据点到直线的距离小于 t 时,此点被认为是内点,反之则为外点。

1.2 对比分析

为验证两种算法的可靠性和稳健性,对同一组含有相同误差和异常点的数据进行模拟,当数据(inliers)中不添加异常点和添加 50 个异常点(outliers)时,LSM 和 RANSAC 算法的拟合结果如图 2 所示。

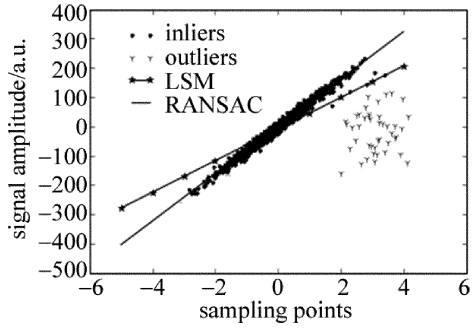

Fig. 2 Fitness results of the same data sets by applying LSM and RANSAC

结果显示,最小二乘法对异常值较敏感,当实验数据中出现异常点时,最小二乘法拟合的直线极大地偏离原直线;而 RANSAC 则可有效地排除异常点的影响,拟合结果非常接近原模型,具有较好的稳健性。相比于最小二乘法,RANSAC 算法在计算参量的迭代次数没有上限,其优点是它能鲁棒性地估计模型参量,即

使是对于存在一定显著数量的异常值的数据集,也可以高精度的估计参量,因而被广泛地应用于图像处理。实际应用中,通过最佳化 RANSAC 模型参量可找到最大内点集[13-15],减小误差概况概率,提高数据处理的精确度。

2 实验结果和讨论

2.1 仿真实验

本文中首先通过采用 Python 程序语言自行编程对波长调制二次谐波信号进行仿真研究。假设不含噪声的体积分数为 10^{-7} 的甲醛二次谐波信号为参考信号 X(信号幅值为相对值),体积分数为 2×10^{-7} 的二次谐波信号为待分析信号 Y,并通过对待分析信号 Y 添加不同幅值 A 的噪声(部分信号如图 3a ~ 图 3c 所示),进而对以上两种线性拟合模型进行评估。以图 3a ~ 图 3c 仿真二次谐波信号(横坐标为采样点数,无单位)中每个 X 信号为横坐标,Y 信号为纵坐标,画出的图形及相应线性拟合结果分别展示在对应的图 3d ~ 图 3f 中,最终拟合结果统计如表 1 所示。从表 1 可以看出,对具有较高信噪比(signal-to-noise ratio, SNR)R_{SNR}的谐波信号进行线性拟合,两种算法的拟合结果具有很好的一致性;随着噪声的增加,对具有较低信噪比的二次谐波信号进行线性拟合时,RANSAC 算法明显比 LSM 更具有优越性,拟合结果的线性相关度 R^2 要明显高于 LSM, 且拟合的比值(slope)(即线性拟合的斜率,Y 与 X 的比值)更接近真实值 2.0。

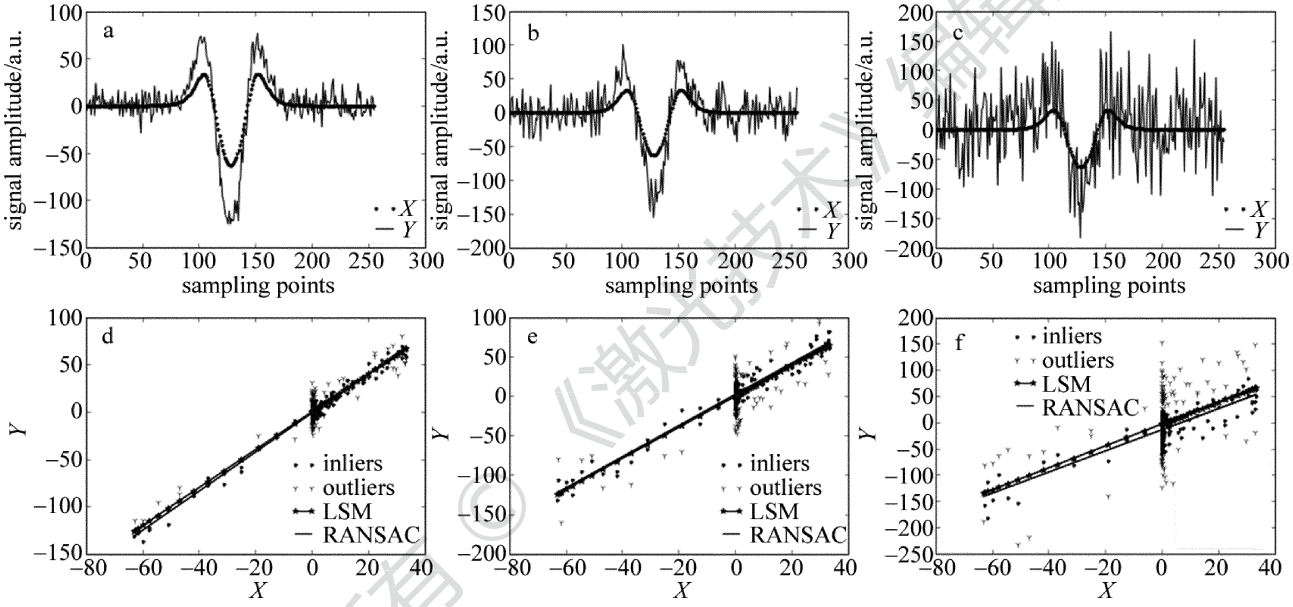

Fig. 3 a ~ c—the simulated second harmonic signal of formaldehyde with noise level A = 10, 20, 50, respectively e ~ f—the corresponding fitted results by using LSMand RANSAC algorithms, respectively

Table 1 Fitting results of two harmonic signal of formaldehyde simulation with different SNR (adding Gaussian noise with different amplitude A to Y, while keeping X unchanged)

experimental results	R^2		slope	
	LSM	RANSAC	LSM	RANSAC
$A=0$ ($R_{SNR}=\infty$)	1.0	1.0	2.0	2.0
$A=5$ ($R_{SNR}=8.87$)	0.9833	0.9935	1.9845	2.0059
$A=10$ ($R_{SNR}=6.00$)	0.9371	0.9831	2.0183	1.9997
$A=20$ ($R_{SNR}=3.54$)	0.8045	0.9412	2.0252	2.0003
$A=50$ ($R_{SNR}=0.64$)	0.2654	0.7482	1.6848	1.9083

2.2 实验数据分析

为了进一步对两种拟合算法进行评估,将两种拟合模型应用到实验中记录的大气甲醛二次谐波信号处理中,实验测量系统如参考文献[16]中所述。大气中甲醛含量极低,因此,实验中测量的光谱信号质量较差。实验上获得二次谐波信号 $I_{2,f}$ 与气体分子浓度 C 之间满足以下关系[17]:

$$I_{2,f} \propto I_0 \alpha C L \tag{3}$$

式中,I_0 为激光初始光强,α 为分子吸收系数,L 为有效吸收光程。因此,通过将未知浓度的样品信号与已知参考样品的信号进行对比分析,即可消除初始光强的影响,从而获得未知样品的浓度信息。本文中主要是通过已知浓度的甲醛信号,对相关算法的可靠性进行初步的评估。图 4a 是不同甲醛体积分数的两个二次谐波信号(signal_1:42×10^{-9};signal_2:35×10^{-9}),信号基线部分受到采集系统噪声的严重干扰。类似于图 3 处理方法,以 signal_2 的数据点为横坐标和 signal_1 的数据点为纵坐标时,给出如图 4b 中符号“·”所示的依赖关系(包含 inliers 和 outliers),图中符号“–”描述的分别为 LSM 和 RANSAC 算法线性拟合的结果。由此图可见,LSM 算法处理的对象为整个数据点集

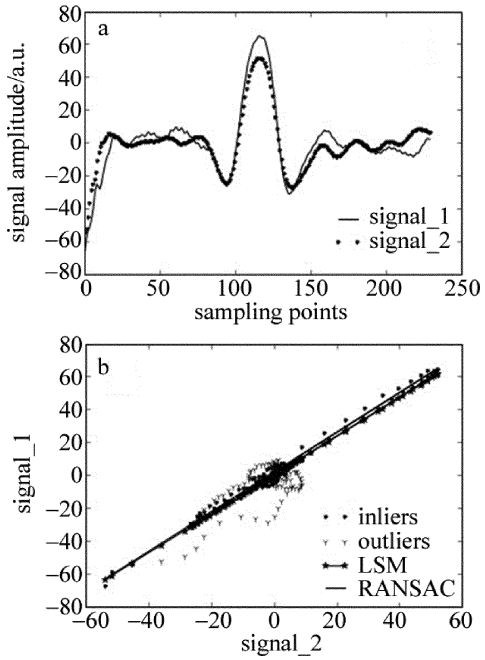

Fig. 4 a—the experimentally measured second harmonic signal of formaldehyde with different concentrations b—the fitting results by using LSM and RANSAC algorithms, respectively

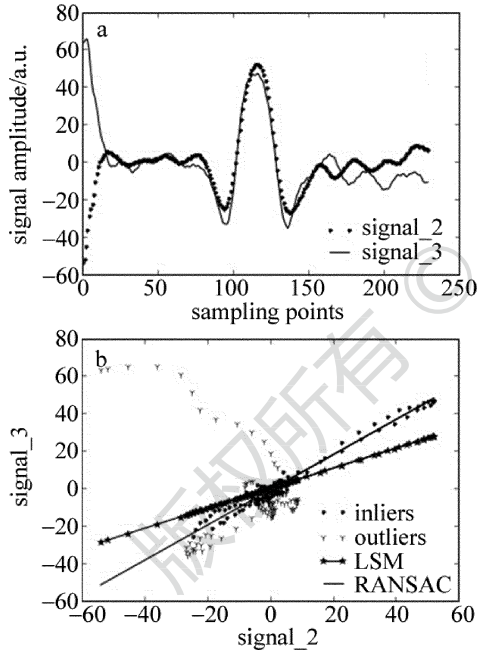

Fig. 5 a—the experimentally measured second harmonic signal of formaldehyde with same concentration b—the corresponding fitted results by using LSM and RANSAC algorithm, respectively

(inliers + outliers), 而 RANSAC 算法通过排除含有噪声干扰的数据点(outliers), 只对有效数据(inliers)进行拟合分析, 从而使得拟合结果的可靠性得到显著提高。图 5a 为甲醛体积分数为 35×10^{-9} 不变的情况下, 长时间连续测量二次谐波信号时不同时刻下选取的两个二次谐波信号(signal_2 和 signal_3), 由于受系统的稳定性和背景噪声的干扰, 使得信号的峰峰值出现明显的上下波动。同理, 图 5b 中给出了 signal_2 和 signal_3 之间的依赖关系(如“·”所示), 及相应 LSM 和 RANSAC 算法线性拟合结果(如“-”所示), 拟合结果相关的参量统计归纳在表 2 中。

Table 2 Linear fitting results of formaldehyde spectra under different experimental conditions

formaldehyde	sample 1		sample 2	
	LSM	RANSAC	LSM	RANSAC
actual ratio	1.20	1.20	1.0	1.0
fitted value	1.1818	1.1818	0.5342	1.0272
correlation coefficient R^2	0.8923	0.9853	0.2111	0.9743
error/%	1.517	-0.18	46.58	-2.72

从拟合结果可以看出, LSM 在信噪比较低情况下, 极易受异常数据的影响, 使拟合模型明显偏离, 线性相关度较低, 拟合的体积分数误差高达 47%。而 RANSAC 算法通过设置阈值来区分内外点, 可以很好地排除仪器系统噪声(光学干涉噪声和电子学噪声)的影响, 使得拟合线性相关度提高, 反演的气体体积分数误差较小。

通过以上对仿真信号和实验数据的分析处理可见, 当光谱数据信噪比较高的时候, 两种模型拟合结果保持很好的一致性, 当光谱数据信噪比较差的时候, 尤其是光谱信号受到采集系统噪声的严重干扰, RANSAC 算法比 LSM 更能鲁棒性地估计模型参量, 提高线性相关度, 减小气体浓度反演的误差。

3 结 论

通过理论和实验研究了 RANSAC 算法在波长调制吸收光谱数据处理中的应用。结果表明, 与传统的 LSM 相比, RANSAC 算法可以很好地适应光谱数据中各种异常情况, 对带有误差和异常值的数据集进行拟合并得到线性相关度较高的拟合结果。尤其是在吸收光谱信号较弱(气体浓度较低)、背景噪声影响显著的情况下, RANSAC 算法作为一种鲁棒性的线性拟合算法, 可很好地排除异常数据的干扰, 有效且可靠地反演出样品浓度信息, 体现出其在激光光谱高精度测量大气温室气体及其它大气痕量气体应用研究方面的潜力。

参 考 文 献

[1] PHILIPPE L C, HANSON R K. Laser diode wavelength-modulation spectroscopy for simultaneous measurement of temperature, pressure, and velocity in shock-heated oxygen flows [J]. Applied Optics, 1993, 32(30): 6090-6103.

[2] LI J S, YU B L, ZHAO W X, *et al*. A review of signal enhancement and noise reduction techniques for tunable diode laser absorption spectroscopy [J]. Applied Spectroscopy Reviews, 2014, 49(8): 666-691.

[3] REID J, LABRIE D. Second-Harmonic detection with tunable diode

lasers-comparison of experiment and theory [J]. Applied Physics, 1981, B26(3):203-210.

[4] LI J, REIFFS A, PARCHATKA U, *et al*. In situ measurements of atmospheric CO and its correlation with NO_x and O_3 at a rural mountain site [J]. Metrology and Measurement Systems, 2015, 22(1): 25-38.

[5] CAI Y, WU Sh Q, WU A, *et al*. Study on calculation method of detection limit based on wavelength modulation spectroscopy [J]. Laser Technology, 2012, 36(3):390-393(in Chinese).

[6] XU Y Z, GUO J Q, GAO X R, *et al*. Effect of temperature on absorption spectral lines of carbon monoxide [J]. Laser Technology, 2010, 34(6):778-780(in Chinese).

[7] PLACKETT R L. The discovery of the method of least squares [J]. Biometrika, 1972, 59(2):239-251.

[8] JIA X Y, XU C S, BAI X. The invention and way of thinking on least squares [J]. Joumal of Northwest University, 2006, 36(3):507-511 (in Chinese).

[9] FISCHLER M A, BOLLES R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography [J]. Communications of the ACM, 1981, 24(6): 381-395.

[10] ZHOU C L, ZHU H H, LI X J. Research and application of robust plane fitting algorithm with RANSAC [J]. Computer Engineering and Applications, 2011, 47(7):177-179(in Chinese).

[11] CAO Y, FENG Y, YANG Y T, *et al*. Application of estimation algorithm based on RANSAC in road points cloud optimization [J]. Infrared and Laser Engineering, 2012, 41(11): 3108-3112 (in Chinese).

[12] WEI Y Z, LIU X L. Robust plane fitting of clouds based on RANSAC [J]. Journal of Beijing University of Technology, 2014, 40(3):400-403(in Chinese).

[13] ZHEN Y, LIU X J, WANG M Zh. An improved RANSAC of fundamental matrix estimation method [J]. Bulletin of Surveying and Mapping, 2014(4): 39-43(in Chinese).

[14] ZHANG H M, ZHENG Z. An improvement of the adjacent probability random sampling consistency algorithm [J]. Laser Journal, 2013, 34(5):29-30(in Chinese).

[15] HAST A, NYSJÖ J, MARCHETTI A. Optimal RANSAC-towards a repeatable algorithm for finding the optimal set [J]. Journal of WSCG, 2013, 21(1): 21-30.

[16] LI J S, PARCHATKA U, FISCHER H. A formaldehyde trace gas sensor based on a thermoelectrically cooled CW-DFB quantum cascade laser [J]. Analytical Methods, 2014, 6(15): 5483-5488.

[17] LI J, PARCHATKA U, FISCHER H. Development of field-deployable real time QCL spectrometer for simultaneous detection of ambient N_2O and CO [J]. Sensors and Actuators, 2013, B182(3): 659-667.