第32卷，第3期
2012年3月

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

Vol. 32, No. 3, pp594-597
March, 2012

基于 Voigt 峰的快速光谱建模算法

李津蓉[1,2]，戴连奎[1*]

1. 浙江大学工业控制技术国家重点实验室，浙江 杭州 310027
2. 浙江科技学院自动化与电气工程学院，浙江 杭州 310023

摘 要 间接硬建模算法是近年来提出的一种新型光谱定量分析技术，适用于分析混合物光谱与待测成分之间的非线性关系，并能够很好地解决光谱波段上的重叠峰问题。这种算法在建立定量分析模型之前先要对光谱曲线进行数学建模，即通过多个 Voigt 峰函数的叠加形式来描述所测得的光谱曲线。光谱建模的精确程度直接影响定量分析模型的准确性，而光谱曲线往往需要由几十个甚至上百个 Voigt 峰的叠加而成，因此这一过程实际是一个高维寻优问题，需要较大的运算开销并有可能令优化问题呈现“病态性”。为了降低优化问题维数，文章通过对重叠峰的判断提出了一种改进型的光谱建模算法，实验表明改进型方法与传统方法相比具有运行时间短、模型精确度高等优点。

关键词 光谱分析；光谱建模；间接硬建模；Voigt 峰

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2012)03-0594-04

引 言

光谱技术作为一种非破坏性、高可靠性的快速检测技术，近年来在工业领域中得到越来越广泛的应用。光谱相关的定量分析方法大多基于 Lambert-Beer 定律[1]，假设成分浓度与光谱强度成线性比例关系，如多元线性回归、主成分回归[2]和偏最小二乘[3,4]等。但由于混合物中不同成分分子之间的相互作用使得混合物的光谱数据与成分浓度之间呈现非线性关系，如光谱峰的漂移、形变等。针对于这种非线性情况，常用建模算法包括局部线性建模[5]、支持向量机[6,7]、小波分析等。

上述线性和非线性建模均属于“软建模”方法，只是简单地把光谱数据看作输入向量，并没有考虑光谱数据本身的波形结构，以及波形结构与样本的待测参数之间所固有的物理关系。因此，利用这些方法所建立的回归模型具有共同的局限性，即在建模时需要较多的训练样本，且回归模型基本不具有外推性。

为了解决这些问题，Alsmeyer 于 2004 年提出了间接硬建模(indirect hard modeling，IHM)[8]方法，这种方法首先对纯物质溶液光谱进行“硬建模”，即通过 Voigt 峰函数的叠加形式对光谱曲线进行描述，这种数学描述称为光谱模型。然后再通过优化拟合的方法将混合溶液光谱分解成纯物质光谱模型的加权和形式，最后利用纯物质的权值对该物质在混合溶液中的浓度进行线性建模。这种方法利用简单的线性模型而得到较高的“外推性”。其后，Kriesten 又在 IHM 方法的基础上提出了补峰硬建模法(complemental hard modeling，CHM)和硬模型因子分析法(hard modeling factor analysis，HMFA)方法[9]，进一步扩大了硬建模算法的应用领域。

在 IHM 算法中的一个关键步骤就是对测量光谱进行数学建模。在这个过程中，需要将测量所得的光谱曲线表示为多个 Voigt 峰函数的累加和形式，并且达到令人满意的拟合误差。光谱建模的准确性对其后所建的定量或定性分析模型的预测精度具有至关重要的作用。由于这一步骤需要对大量 Voigt 峰的参数同时进行优化，而高维的优化问题会导致优化算法的运行开销较大，且容易陷入“病态”，而无法得到最优解[10]。本文针对于这个问题，提出了一种新的快速光谱建模算法，这种方法通过判断峰的重叠性来降低每次迭代中需要进行优化的 Voigt 峰的参数，从而可以大大降低优化算法的运行开销，并有效避免了优化算法的“病态性”。

1 光谱拟合算法

红外或拉曼光谱曲线可由多个不同波段上的 Voigt 峰函

收稿日期：2011-06-15，**修订日期**：2011-10-08

基金项目：国家(863 计划)项目(2009AA04Z123)资助

作者简介：李津蓉，女，1977 年生，浙江大学控制系博士研究生，浙江科技学院讲师 e-mail：lijinrong_hz@yahoo.cn

* 通讯联系人 e-mail：lkdai@iipc.zju.edu.cn

数的叠加来近似表示[11]。Voigt 峰函数可近似表示为 Gaussian 函数和 Lorentzian 函数的线性叠加形式，即

$$V(\nu)=\theta\alpha\exp\left[-\frac{4\ln2(\nu-\omega)^2}{\gamma^2}\right]+(1-\theta)\alpha\frac{\gamma^2}{(\nu-\omega)^2+\gamma^2} \tag{1}$$

其中变量 ν 表示波数，一个 Voigt 函数包括 4 个峰参数，分别为：α(峰高)、ω(峰的中心位置)、γ(峰的半宽)、θ(Gauss-Lorentz 系数)。

测量光谱可以表示成 L 个 Voigt 峰及拟合残差的累加和形式，即

$$A(\nu)=\sum_{i=1}^{L}V_i(\nu,\phi_i)+r(\nu) \tag{2}$$

其中 $A(\nu)$表示测量光谱；$V_i(\nu,\phi_i)$表示第 i 个 Voigt 峰函数，ϕ_i 为峰的参数向量 $\phi_i=(\alpha_i,\omega_i,\gamma_i,\theta_i)^{\mathrm{T}}$；$r(\nu)$表示拟合残差。在这里，我们假设测量光谱已经过去基线处理，即不考虑基线的影响。定义 $A^*(\nu,\Phi)$为测量光谱的数学模型，它可以表示为 L 个 Voigt 峰函数的叠加形式，其中 $\Phi=(\phi^1,\cdots,\phi_L^T)^T$ 表示拟合模型中所有峰函数的参数，即

$$A^*(\nu,\Phi)=\sum_{i=1}^{L}V_i(\nu,\phi_i) \tag{3}$$

使用非线性最小二乘算法，如 Levenberg-Marquardt[12]，对拟合模型中每个峰的参数进行寻优，使得拟合残差 $r(\nu)$最小，其目标式如下

$$\min_{\Phi}\|A(\nu)-A^*(\nu,\Phi)\| \tag{4}$$

对于以上寻优问题，需要寻优的参数个数为 $4L$ 个。例如，若测量所得的光谱曲线中包含 40 个峰，在不考虑基线的情况下需要进行寻优的 Voigt 峰参数就达到 160 个，如此高维的寻优问题一方面会带来较大的运行开销，另外还会导致寻优算法无法给出正确的最优解。Alsmeyer 提出了一种迭代补峰的拟合优化算法[13]，这种算法的每次迭代过程中在光谱拟合残差的最高点处添加一个峰，然后对新峰及所有已添加峰的参数同时进行优化，直至收敛条件满足。该算法随着迭代次数的增多，同样会带来寻优参数过多的问题。

为了解决这种参数过多而导致的优化问题，我们提出了一种改进的迭代拟合算法。对于 Voigt 峰函数而言，在距其中心点 4 倍半宽之外的位置，Voigt 函数值基本接近于零，如图 1 所示。根据 Voigt 函数的这一特征，对于任意两个 Voigt 峰 $V_1(\nu,\alpha_1,\omega_1,\lambda_1,\theta_1)$和 $V_2(\nu,\alpha_2,\omega_2,\lambda_2,\theta_2)$，当两个峰的中心点间距小于这两个峰的半宽较大值的 4 倍时，即，$|\omega_1-\omega_2|<4\max(\lambda_1,\lambda_2)$则可认为它们为重叠峰；否则，这两个峰相互独立，它们之间无重叠部分。当使用Voigt峰函数对测量光谱曲线进行建模拟合时，对于相互重叠的峰，它们的参数需要同时调整，以达到最优拟合度，而对于相互之间不存在重叠部分的峰，其参数可以分别进行调整。

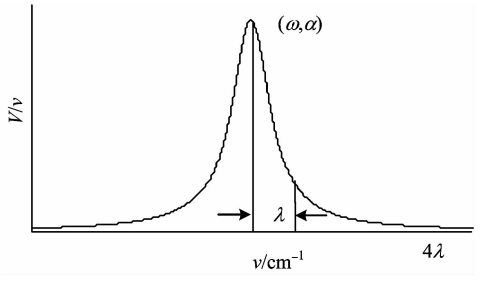

Fig. 1　Voigt peak

基于这一思想，在每一次迭代过程中，当新增一个 Voigt 峰时，根据此峰函数的参数初值在已添加的峰集合中判断哪些峰与这个新峰相互重叠，若存在一组峰与新峰有重叠关系，则需继续在已添加峰的集合中判断还有哪些峰与这一组峰存在重叠关系，最终在可以得到一个相互重叠的 Voigt 峰的集合 V_{overlap}。由于 V_{overlap}集合中的峰相互重叠且与新峰具有重叠关系，因此在这一次迭代中需要对新峰的参数和集合 V_{overlap}中所有峰的参数同时进行寻优，才能得到最优的拟合效果。而对于那些在集合 V_{overlap}之外的已添加峰，由于它们与新峰及 V_{overlap}集合中的所有峰相独立，因此不需要对其修改参数。具体算法步骤如下：

Step1：令 $r(\nu)=A(\nu)$，集合 $V_{\mathrm{added}}=\{\ \}$，$i=1$；

Step2：寻找光谱 $r(\nu)$的最大值 α_i，及其位置 ω_i，估计新峰 V_i 的半宽值 γ_i，设定新峰 V_i 的参数初值 $\phi_{0i}=(\alpha_i,\omega_i,\gamma_i,0.5)$；

Step3：在集合 V_{added}中寻找与新峰 V_i 相重叠的峰集合，并记为 $V_{\mathrm{overlap}}=\{VO_j\}$，$j=1,\cdots,J_i$，$J_i$ 为与新峰 V_i 相重叠的峰的个数。由于集合 V_{overlap}中所有峰的参数都需要重新调整，因此需要在拟合残差上重新补上这些峰的曲线。即令 $r(\nu)=r(\nu)+\sum_{j=1}^{J_i}VO_j$；

Step4：利用非线性最小二乘算法调整峰 V_i 和集合 $\{VO_j\}$，$j=1,\cdots,J_i$ 共 $4\times(J_i+1)$个峰的参数，优化目标式为

$$\min_{\phi_j^*,\phi_i}\frac{1}{\nu_N}\sum_{i=1}^{\nu_N}\|r(\nu_k)-\sum_{j=1}^{J_i}VO_j(\nu_k,\phi_j^*)-V_i(\nu_k,\phi_i)\| \tag{5}$$

Step5：记录新峰 V_i 的参数为 ϕ_i，并更新集合 V_{overlap}中峰 VO_j 的参数为 ϕ_j^*，$j=1,\cdots,J_i$；令 $V_{\mathrm{added}}=V_{\mathrm{added}}\cup V_i$；计算新的拟合残差

$$r(\nu)=r(\nu)-\sum_{j=1}^{J_i}VO_j(\nu,\phi_j^*)-V_i(\nu,\phi_i) \tag{6}$$

Step6：判断拟合残差是否满足收敛条件，若满足收敛条件，则算法结束，集合 V_{added}即为所有的拟合峰；否则令 $i=i+1$，转到 Step2。

2　实验部分

2.1　数据来源

数据来自实验室中配制的间二甲苯(m-xylene)、邻二甲苯(o-xylene)和对二甲苯(p-xylene)混合溶液。拉曼光谱仪的检测范围为 0～2 600 $\mathrm{cm^{-1}}$，采用中心波长为 785 nm 的激光器作为激发光源。由于所用的光谱仪受 CCD 检测器阵列物理特性的限制，测得的拉曼光谱两端包含无效谱段。因此，选取有效谱段为 150～1 700 $\mathrm{cm^{-1}}$的拉曼谱段数据进行分析建模，其原始拉曼光谱及进行基线扣除和均值归一化处理后的谱图如图 2 所示。所有程序在 Matlab 7.1 环境下编写，在

Intel Core 2 微机 Windows XP 环境下运行。

2.2 光谱拟合

利用第 1 节改进的迭代拟合算法对混合溶液的拉曼光谱进行 Voigt 峰函数的建模，并与 Alsmeyer 的算法进行比较。由于当通过多个 Voigt 峰的叠加来拟合光谱曲线时，解空间不具有唯一性，而从优化的角度，我们期望能够通过尽可能少的 Voigt 峰对光谱进行建模并达到尽可能小的拟合误差。因此对拟合算法的评价指标包括算法运行时间、算法运行结束时参与建模的 Voigt 峰总数以及拟合误差的均方和 RMSE。RMSE 定义如下

$$\mathrm{RMSE}=\frac{1}{N}\sum_{n=1}^{N}[A(\nu_n)-M(\nu_n)]^2 \tag{7}$$

波数 940～1 120 cm^{-1}之间的拉曼光谱峰重叠较为严重，图 3 给出了两种算法在此光谱范围拟合结果，从图中可以看出，在峰相互重叠较严重的区域，使用改进的迭代拟合算法，参与拟合的 Voigt 峰的个数较少且拟合效果更好。两种方法的拟合性能比较如表 1 所示。

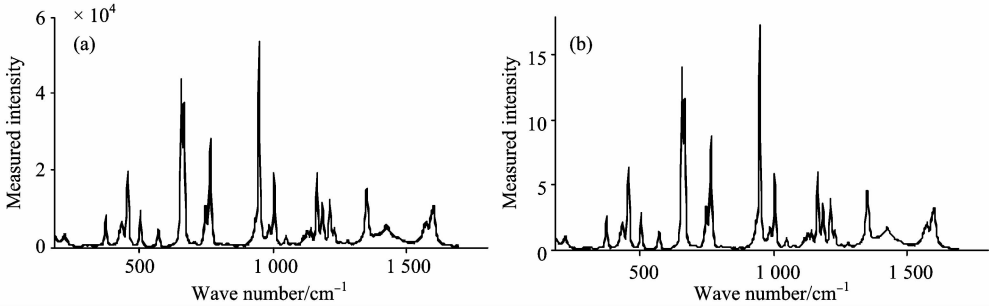

Fig. 2 Original spectra and pre-processed spectra

(a): Measured spectrum; (b): Preprocessed spectrum

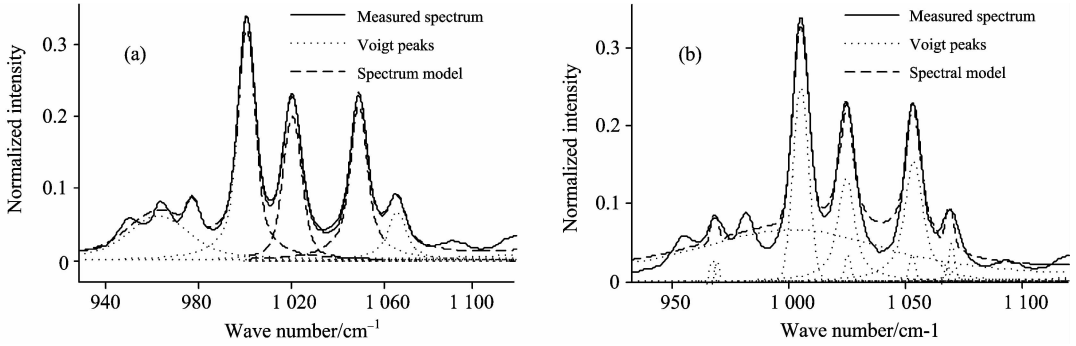

Fig. 3 Comparison of fitting models built by two methods

(a): New method; (b): Alsmeyer's method

Table 1 Comparison of two modeling methods

	Run time /s	Number of peaks	RMSE
New mothod	13.75	31	7.35×10^{-5}
Alsmeyer's method	81.4	42	1.65×10^{-4}

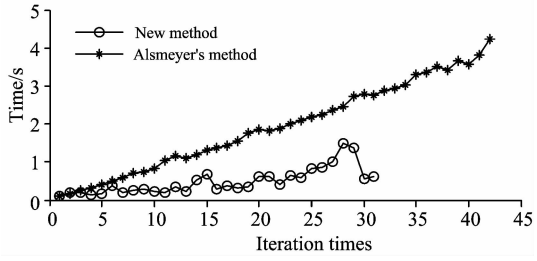

Fig. 4 Runtime of two methods

图 4 中对两种算法的运行时间开销进行了比较，可以看出 Alsmeyer 算法的运行时间随着迭代次数(峰的个数)的增加而迅速上升，而改进的迭代拟合算法在每次迭代中仅对重叠峰的参数进行寻优，因此时间开销相对于迭代次数的增长趋势相对平缓。

3 结 论

针对于 IHM 定量分析方法中对光谱的建模算法进行了改进，在每次迭代步骤中仅对相互重叠的峰参数进行同时优化，从而大大降低了优化目标的空间维度。实验表明与原算法相比，新方法在运行开销及拟合误差方面均有较大的改善。为进一步的光谱定量或定性分析奠定了可靠的基础。

References

[1] Manuela Zude, Michael Pflanz, Lorenzo Spinelli, et al. J. Food Eng., 2011, 103(1): 68.
[2] GUO Long-hai, YUAN Hong-fu, QIU Teng, et al(郭隆海, 袁洪福, 邱 藤, 等). Chemical Research in China University(高等学校化学学报), 2008, 29(6): 1255.
[3] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍, 袁洪福, 徐广通, 等). Modern Analytical Techniques in Near-Infrared Spectroscopy(现代近红外光谱分析技术). 2nd ed. Beijing: China Petrochemical Press(北京: 中国石化出版社), 2007. 306.
[4] Gaydou V, Kister J, Dupuy N. Chemometr. Intel. Lab. Syst., 2011, 106(2): 190.
[5] CHU Xiao-li, XU Yu-peng, LU Wan-zhen(褚小立, 许育鹏, 陆婉珍). Chinese J. Anal. Chem.(分析化学), 2008, 36(5): 702.
[6] BAO Xin, DAI Lian-kui(包 鑫, 戴连奎). Chinese J. Anal. Chem.(分析化学), 2008, 36(1): 75.
[7] Roman M Balabin, Ravilya Z Safieva, Ekaterina I Lomakina. Microchem. J., 2011, 98(1): 121.
[8] Alsmeyer F, Koß H J, Marquardt W. Appl. Spectrosc., 2004, 58(8): 975.
[9] Kriesten E, Mayer D, Alsmeyer F, et al. Chemometr. Intel. Lab. Syst., 2008, 93: 108.
[10] Kriesten E, Alsmeyer F, Bardow A, et al, Chemometr. Intel. Lab. Syst., 2008, 91: 181.
[11] Tommasi E De, Castrillo A, Casa G. J. Quant. Spectrosc. Radiat. Transfer, 2008, 109(1): 168.
[12] Stefan Finsterle, Michael B. Kowalsky. Computers & Geosciences, 2011, 37(6): 731.
[13] Alsmeyer F, Marquardt W. Appl. Spectrosc., 2004, 58(8): 986.

Fast Spectral Modeling Based on Voigt Peaks

LI Jin-rong[1,2], DAI Lian-kui[1*]

1. National Key Lab of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China
2. Zhejiang University of Science & Technology, Hangzhou 310023, China

Abstract Indirect hard modeling (IHM) is a recently introduced method for quantitative spectral analysis, which was applied to the analysis of nonlinear relation between mixture spectrum and component concentration. In addition, IHM is an effectual technology for the analysis of components of mixture with molecular interactions and strongly overlapping bands. Before the establishment of regression model, IHM needs to model the measured spectrum as a sum of Voigt peaks. The precision of the spectral model has immediate impact on the accuracy of the regression model. A spectrum often includes dozens or even hundreds of Voigt peaks, which mean that spectral modeling is a optimization problem with high dimensionality in fact. So, large operation overhead is needed and the solution would not be numerically unique due to the ill-condition of the optimization problem. An improved spectral modeling method is presented in the present paper, which reduces the dimensionality of optimization problem by determining the overlapped peaks in spectrum. Experimental results show that the spectral modeling based on the new method is more accurate and needs much shorter running time than conventional method.

Keywords Spectral analysis; Spectral modeling; Indirect hard modeling (IHM); Voigt peak

(Received Jun. 15, 2011; accepted Oct. 8, 2011)

* Corresponding author