2009 年 5 月
May 2009

色　　谱
Chinese Journal of Chromatography

Vol. 27 No. 3
351 ~ 355

自动化色谱谱图解析——谱峰的自动识别与快速解析

刘明明，　夏炳乐*，　杨　俊

（中国科学技术大学烟草与健康研究中心，安徽 合肥 230052）

摘要：结合基于高阶导数的谱峰识别方法和面积重现法，建立了一种完全自动化的对色谱曲线进行分割、识别与快速解析的方法。其中，DW（Durbin-Watson）测试的引入和区分信号与噪声判据的采用减少了在色谱解析过程中的人为干预，降低了对操作人员专业知识和经验的要求，为实现色谱解析的自动化奠定了基础。通过对模拟色谱和实验色谱的比较，验证了该方法是一个很有用的工具，可以为色谱分析工作提供有力的帮助。

关键词：面积重现法；谱峰识别；色谱解析；DW 测试

中图分类号：O658　　**文献标识码**：A　　**文章编号**：1000-8713（2009）03-0351-05　　**栏目类别**：研究论文

Automatic peak recognition and rapid resolution of chromatographic signals with a self-compiling program

LIU Mingming, XIA Bingle*, YANG Jun

(*Research Center of Tobacco and Health, University of Science and Technology of China, Hefei 230052, China*)

Abstract: Area reproduction method was introduced in combination with peak recognition algorithm based on high-order derivatives to automate the chromatogram division, peak recognition and rapid resolution. Durbin-Watson method and the criterion to distinguish the signal and noise were adopted to reduce the user interaction. The objective was that the operators should be able to perform this method with minimal experience and professional knowledge. The method is a useful tool by applying it to the resolution of model and real chromatographic signals.

Key words: area reproduction; peak recognition; resolution of chromatographic signals; Durbin-Watson method

随着色谱技术的进步，色谱的分辨能力亦日益提高，在分析检测领域的应用也逐渐广泛。近年来，人们提出了许多色谱解析的方法，常见的有傅里叶自去卷积法[1]、小波变换法[2]、神经网络法[3]、正交投影法[4]、曲线拟合法[5]、遗传算法[6]、免疫算法[7]等。这些方法大部分需要人工干预，提高了对操作者的经验要求，影响了色谱解析技术的实用性，且多数方法在解析的过程中耗时较长，解析效率较低，不能满足短时间内解析大量谱图的需求。

2007 年，Boe[8] 报道了一种称为面积重现法的色谱模拟方法，它以修正的泊松模型为基础，对色谱曲线进行直接测量，通过计算模型参数，可以实现对色谱曲线的快速模拟。对于分离的色谱峰，面积重现法可以取得很好的效果；但对于重叠的色谱峰，谱峰的参数难以测量，Boe 以相同的峰形参数拟合所有的谱峰，忽略了不同组分之间色谱峰形状的差异，给解析结果带来了很大的偏差。文献[9]则提出了另外一种基于色谱曲线高阶导数的谱峰识别方法。该方法是根据一阶导数确定组分的洗提区域，根据二阶导数确定谱峰的性质；但在色谱识别的过程中需要人工指定阈值以区分信号和噪声，这对于操作人员来说是一个很大的挑战。

为了降低对操作者的经验要求，同时迅速地解析含有重叠峰的色谱曲线，我们将基于高阶导数的色谱识别方法加以改进，结合面积重现法，建立了一种完全自动化的色谱解析方法。经过对模拟色谱和实验色谱的比较，证实了该方法可靠快速，可以为日益增多的色谱分析工作提供一定的帮助。

1　原理

1.1　信号与噪声的判据

在色谱曲线的自动处理过程中，区分信号与噪声是一个基本的问题。在传统的算法中，人们往往通过指定阈值的方法[9]来区分信号与噪声。当响

* 通讯联系人：夏炳乐，副教授. E-mail：xiabl@ustc.edu.cn.
基金项目：国家烟草专卖局科研基金项目（No. 1002001041）.
收稿日期：2008-10-04

应大于阈值时为信号,否则为噪声。阈值的选取依赖操作人员的经验,并且需要进行多次尝试。本方法中采用了一种依据数据曲线的判据,它依赖数据自身的特点,不需要人为的指导。其主要思想可以通过以下两个公式描述:

$$\rho = \frac{1}{n}\sum |f_i| \tag{1}$$

$$\varepsilon = \frac{n}{\sum \frac{1}{|f_i| + \rho}} - \rho \tag{2}$$

即

$$\frac{1}{\varepsilon + \rho} = \frac{1}{n}\sum \frac{1}{|f_i| + \rho}$$

而

$$x_i = \frac{1}{2m + 1}\sum_{k=i-m}^{i+m} f_k \tag{3}$$

其中 f_i 表示第 i 个数据点;n 为数据点的总数;ρ 为数据点绝对值的平均值;ε 描述了数据曲线的加权平均,响应越大的数据点,对 ε 的贡献越小,因此 ε 主要表现为噪声的绝对平均;x_i 表示第 i 个数据点前后 m 个点的平均值,表现了第 i 个数据点附近数据点的性质。当 x_i 的绝对值大于 ε,则表示第 i 个数据点为信号;否则表示噪声。

1.2 色谱曲线的分割

对于一条多组分的色谱曲线,对其进行分割是一个十分重要的步骤,它不仅可以减小数据量,提高色谱曲线的解析速度,而且可以降低曲线平滑所带来的误差。依赖区分信号与噪声的算法,色谱曲线可以视为由信号单元与噪声单元交替组成,信号单元是由响应大于 ε 的连续数据点构成,是进一步解析的基本数据结构。

1.3 色谱峰的分区与识别

2005 年,文献[9]报道了一种色谱峰识别方法,它是根据色谱曲线的一阶导数对曲线进行分区,再根据曲线的二阶导数进行色谱峰的识别。具体过程如图 1 所示。

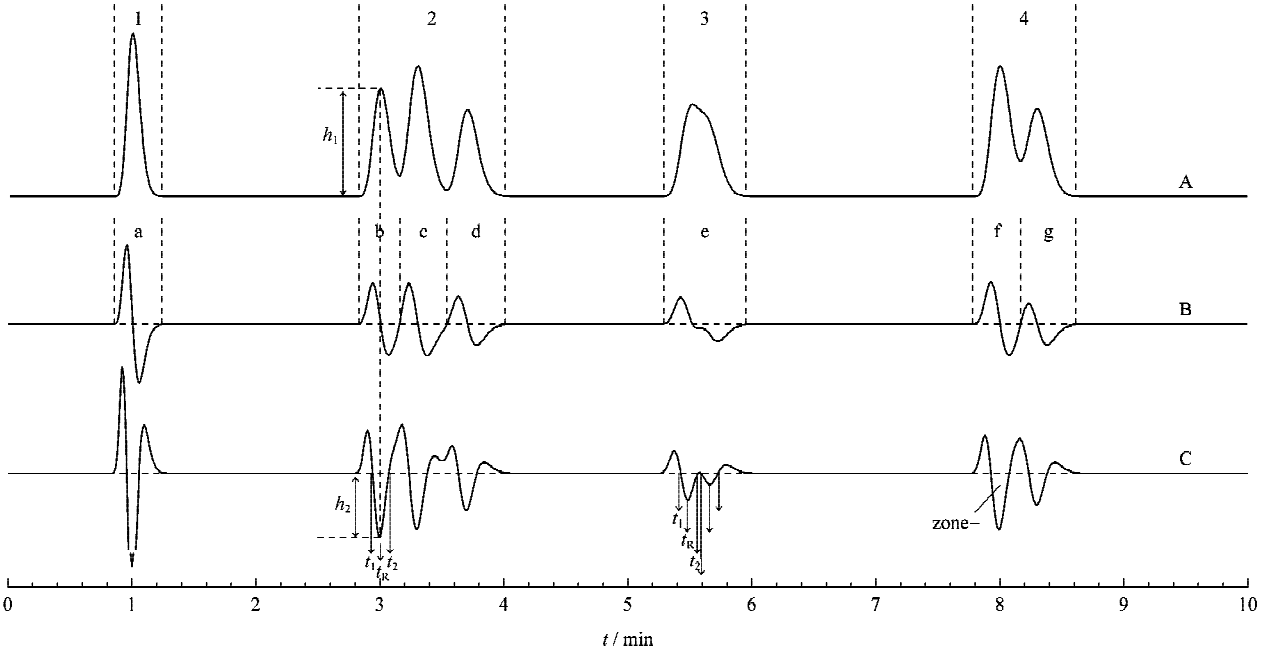

图 1 模拟色谱图及其导数

Fig. 1 Model chromatogram and its derivatives

A: chromatographic signal; B: first-order derivatives; C: second-order derivatives. 1 – 4: message units. a – g: peak regions.

根据色谱曲线的一阶导数,可以将色谱曲线划分为一系列的时间区域,即谱峰区域(peak region),每一个谱峰区域在一阶导数上由相邻的一个正区和一个负区组成。根据色谱曲线的二阶导数,可以得到该色谱曲线含有的组分数:一般来讲,根据二阶导数上负区域(zone –)的数目可确定组分数,即每个负区域对应一个组分;同时,可以得到这个组分的一系列的峰形参数:负区域最小值对应的保留时间 t_R,二阶导数正负改变所对应的时间 t_1 和 t_2,二阶导数的负区域高度 h_2 以及 t_R 时色谱曲线的高度 h_1(换句话说,h_1 描述了单组分色谱峰的峰高上限)。

结合公式(4)即可对色谱峰进行初步的拟合:

$$h(t) = h_{max}\exp\left(-\frac{1}{2}\left(\frac{t - t_R}{s_0 + s_1(t - t_R)}\right)^2\right) \tag{4}$$

其中公式(4)的主要参数由公式(5) ~ (7)确定:

$$h_{max} = h_2\left(\frac{t_2 - t_1}{2}\right)^2 \tag{5}$$

$$s_0 = \frac{t_2 - t_1}{2} \tag{6}$$

$$s_1 = \frac{\dfrac{t_2 - t_R}{t_R - t_1} - 1}{\dfrac{t_2 - t_R}{t_R - t_1} + 1} \tag{7}$$

在信号求导之前,滤波去噪通常是一个必要的操作,经典的 Savitzky-Golay 多项式平滑方法[10](SG 方法)可以很好地实现这一过程。在 SG 方法中有两个参数需要选择:多项式的指数和窗口的尺寸。文献[9]通过 Durbin-Watson 方法(又称 DW 测试)对 SG 方法中的参数进行自动选取,取得了良好的效果。具体方法如下:采用二次多项式的 SG 方法作为色谱曲线平滑去噪的基本方法。调节窗口的大小,分别对信号进行 SG 滤波,通过公式(8)求取 DW 指数[11]。DW 指数描述了平滑信号的残差与信号的相关性,随着相关性的减小,DW 指数收敛于 2。因此,DW 指数最接近 2 的窗口尺寸是最合适的。

$$\mathrm{DW} = \frac{n}{n-1} \cdot \frac{\sum_{i=2}^{n} [(h_{0,i} - h_{s,i}) - (h_{0,i-1} - h_{s,i-1})]^2}{\sum_{i=1}^{n} (h_{0,i} - h_{s,i})^2} \tag{8}$$

其中 h_0 为原始信号,h_s 为平滑信号,n 为信号中数据点的数目。

1.4　色谱曲线的解析

面积重现法[8]是根据对图谱的测量直接确定色谱模型的形状参数,从而实现对色谱曲线的快速模拟。采用修正的泊松函数可以很好地描述色谱曲线[12-15],而且具有形式简单、模型参数少、参数之间无关联的优点。常用的泊松函数为:

$$f(t) = h_{\max} \exp(-k(t - t_R)) \left[1 + \frac{k}{n}(t - t_R)\right]^n \tag{9}$$

其中 t_R 为保留时间,$h_{\max}$为峰的高度,n 和 k 是泊松模型的峰形参数。

面积重现法以修正的泊松模型为基础,采取如下色谱模型:

$$h(t) = \begin{cases} f(t), & t \geq t_R - \dfrac{n}{k} \\ 0, & t < t_R - \dfrac{n}{k} \end{cases} \tag{10}$$

公式(10)具有如下性质:

$$P_{\mathrm{area}} = \int_{t_R - n/k}^{\infty} h(t)\,\mathrm{d}t \approx \sqrt{2n\pi}\,\frac{h_{\max}}{k} \tag{11}$$

P_{area}表示色谱峰的面积。Boe 的研究[8]表明,对于一般的色谱模型而言,公式(11)具有如下性质:

$$n = \left(\frac{0.778}{\log \alpha}\right)^2 \tag{12}$$

$$k = \frac{2.24\sqrt{n}}{w} \tag{13}$$

其中倾斜因子 $\alpha = w_l / w_r$;半峰宽 $w = w_l + w_r$,w_l 与 w_r 分别代表左半峰宽和右半峰宽。因此,只要测量峰高 $h_{\max}$、保留时间 t_R、w_l 和 w_r 即可解析色谱曲线。

1.5　自动化色谱谱图解析系统

在自动化色谱解析系统中,首先将多组分色谱曲线分割成一系列信号单元,通过 DW 测试找出最合适的 SG 参数,对信号单元进行去噪和求导;然后,根据导数曲线识别色谱峰,并进行初步的重构;最后,测量重构的色谱峰,结合面积重现法对色谱峰进行模拟,得到初步的谱峰参数。

对于两个峰形相同,即 n、k 不变的色谱峰,根据公式(11)可知其面积取决于 $h_{\max}$,那么校正因子 β 可以通过公式(14)求出:

$$\beta = \frac{A}{\sum A_i} \tag{14}$$

其中 A 表示某一信号单元中原始色谱曲线的面积,A_i 表示此信号单元中第 i 个组分拟合谱峰的面积。对于每一个组分,其校正峰高:

$$h_{\max,i} = \beta h_{\max,i} \tag{15}$$

如果某一组分的峰高校正之后大于 $h_{1,i}$,则使其峰高等于 $h_{1,i}$,在保证面积相等的条件下,根据公式(11)校正峰形参数 k_i,即可得到最终的谱峰参数,实现对色谱曲线的解析。

1.6　自动化色谱谱图解析系统算法程序

该算法程序主要执行了以下步骤:

(1)根据区分信号与噪声的判据,将原始谱图中的色谱曲线分割成一系列的具有时间先后顺序的信号单元;

(2)取出第 1 个信号单元;

(3)通过 DW 测试找出最合适的 SG 参数,对取出的信号单元进行去噪和求导;

(4)根据该信号单元中色谱曲线的一阶导数确定一系列谱峰区域时间段;

(5)在第 1 个谱峰区域的时间范围内,找出对应的色谱曲线二阶导数部分,根据二阶导数可确定该时间段内的谱峰数,同时得到相应的峰形参数;

(6)依次在该信号单元剩下的谱峰区域时间段重复步骤(5),就可以得到按时间顺序排列的各组分峰形参数;

(7)得到该信号单元的所有组分的峰形参数后,将其代入公式(4)~(7)可以初步重构各个组分的色谱曲线;

(8)测量初步重构出的各组分色谱曲线的峰高 $h_{\max,i}$、保留时间 $t_{R,i}$、$w_{l,i}$ 和 $w_{r,i}$,代入公式(12)和(13),获得初步的谱峰参数 $h_{\max,i}$,$t_{R,i}$,n_i,k_i;

(9)利用公式(11),(14)和(15)校正峰高 $h_{\max,i}$ 和峰形参数 k_i 即可得到最终的解析结果;

(10)取出原始谱图中第 2 个信号单元,重复步骤(3)~(9),即可解析出第 2 个信号单元中的色谱曲线;

(11)以此类推,依次取出剩下的信号单元进行解析,直至取完,就完成了对整个原始谱图的解析。

2 模拟与实验

2.1 实验条件

仪器与色谱条件:美国 Finnigan Trace GC 气相色谱仪,火焰离子化检测器(FID),检测器温度 260 ℃。升温程序:初始温度 40 ℃,保持 2 min,以 10 ℃/min 升至 170 ℃,再以 4 ℃/min 升至 250 ℃,保持 5 min。

整个色谱解析方法采用 Matlab 7.0 编程,在 PC 机上运行。

2.2 模拟色谱

随意选定 8 个组分的模型参数,通过公式(10)产生模拟色谱图模型,并加上随机噪声,其色谱解析结果如图 2-A 所示。模拟色谱图模型中包含孤立峰和不同程度的重叠峰,比较适合对“1.6”节所述自动化色谱解析系统算法程序进行考察。

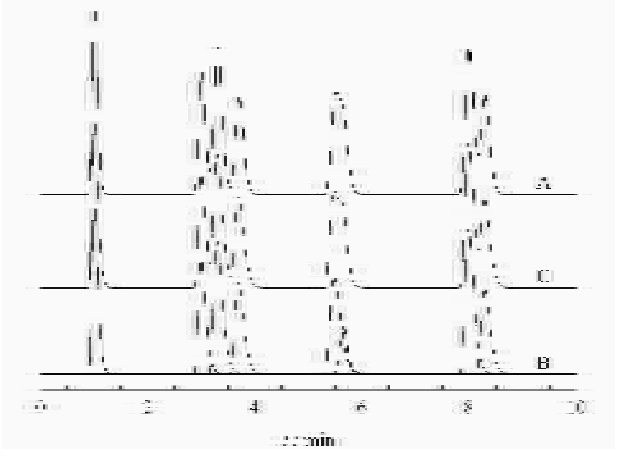

图 2 模拟色谱的解析结果

Fig. 2 Resolution result of the model chromatogram

A: model chromatogram; B: resolution result; C: sum of the resolution result.

2.3 实验色谱

为了验证本程序,我们对青岛卷烟烟气色谱谱图中较为复杂的部分进行了解析,结果见图 3。

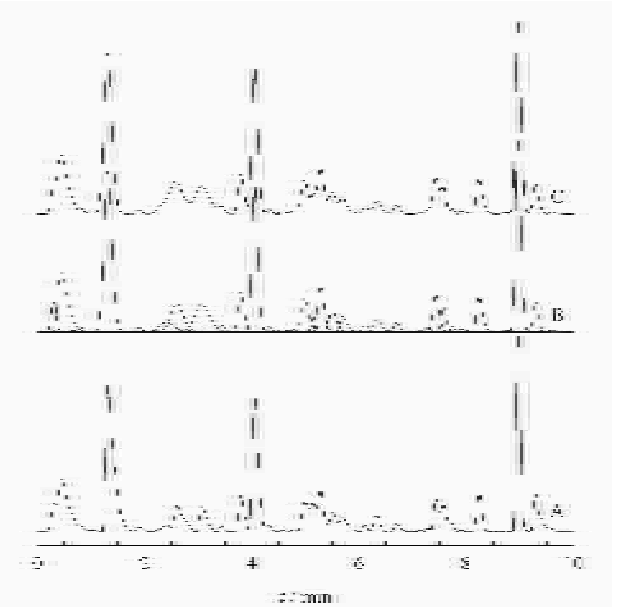

图 3 实验色谱的解析结果

Fig. 3 Resolution result of the experimental chromatogram

A: experimental chromatogram; B: resolution result; C: sum of the resolution result.

3 结果与讨论

3.1 信号单元的分区、识别与重构

本方法是通过 SG 滤波和求导,结合区分噪声与信号的判据,根据色谱曲线的一阶导数进行分区。图 1-B 显示的是一模拟色谱图的分区:根据正区与负区的组成,信号单元 1 划分出 1 个谱峰区域 a;信号单元 2 划分出 b,c,d 共 3 个谱峰区域;信号单元 3 划分出 1 个谱峰区域 e;信号单元 4 划分出了 f,g 共 2 个谱峰区域。

以信号曲线的二阶导数为基础,结合区分噪声与信号的判据对色谱峰进行识别,如对于谱峰区域 b,在其对应的二阶导数上,可搜索到 1 个负区域,同时得到 1 组峰形参数:t_1,t_2,t_R,h_1,h_2。对于谱峰区域 e,在其对应的二阶导数上,可搜索到 2 个负区域,同时得到 2 组峰形参数。同理可得其他谱峰区域的峰形参数,即信号单元 1,2,3,4 分别可得到 1 组、3 组、2 组、2 组峰形参数(见图 1-C),这样就全部识别出了模拟谱图中的 8 个组分,同时得到了各组分的谱峰参数。将信号单元的峰形参数代入公式(4)~(7)则可以初步重构各个组分的色谱曲线。

3.2 色谱图的解析

对初步重构出的各组分独立色谱曲线进行测量,得出半峰宽 $w_{l,i}$ 和 $w_{r,i}$,结合上一步的 $t_{R,i}$ 和 $h_{\max,i}$,代

入公式(12)和(13),就获得了公式(10)的各个参数。最后,利用公式(11),(14)和(15)校正峰高和倾斜因子即可得到最终的解析结果(见图 2,3)。图 2-A 为模拟原始谱图,图 2-B 为该谱图的解析结果,图 2-C 为解析结果的包络线。图 3-A 为烟气原始谱图的一部分,图 3-B 为该谱图的解析结果,图 3-C 为解析后曲线的包络线。模拟色谱图解析耗时 0.31 s;实验色谱图解析耗时 1.54 s。

4　结论

本自动化色谱谱图解析方法较好地实现了色谱曲线的自动快速解析,对于含有大量重叠峰的色谱曲线亦能迅速地解析出来,可作为较复杂色谱图的分析辅助工具。对于如烟气谱图这样具有重叠峰的谱图,该方法在一定程度上可帮助寻找隐藏在众多重叠峰里的未知成分,以及预测其大概含量。该方法适用于大部分正常的色谱曲线,但要注意以下情况:(1)修正的泊松模型不能很好地拟合“伸舌”峰形,因此“伸舌”峰的解析结果偏差较大;(2)对于重叠得非常严重甚至完全重叠的色谱峰,即使高阶导数亦不能将其识别,用本方法解析出的结果仍然是一个峰,因此本方法不能解析出完全重叠以及接近完全重叠的色谱峰;(3)如果原色谱图存在较大的基线,则应先校正一下基线,再用本方法进行解析。

参考文献:

[1] Xu X F, Zhao M S, Hu X Y. Computers and Applied Chemistry (徐晓峰, 赵明生, 胡鑫尧. 计算机与应用化学), 1998, 15(3):153

[2] Shao X G, Hou S Q, Zhao G W. Chinese Journal of Analytical Chemistry (邵学广, 侯树泉, 赵贵文. 分析化学), 1998, 26(12):1 428

[3] Li Y B, Huang X Y, Sha M, et al. Chinese Journal of Chromatography (李一波, 黄小原, 沙明, 等. 色谱), 2001, 19(2):112

[4] Chen D Z, Cui H. Chinese Journal of Chromatography (陈迪钊, 崔卉. 色谱), 2000, 18(2):100

[5] Lu X Q, Liu H D, Zhang M, et al. Chinese Journal of Analytical Chemistry (卢小泉, 刘宏德, 张敏, 等. 分析化学), 2003, 31(2):143

[6] Chen K, Li T H, Lu P C. Chinese Journal of Analytical Chemistry (陈开, 李通化, 卢佩章. 分析化学), 2003, 31(2):158

[7] Shao X G, Chen Z H, Lin X Q. Chinese Journal of Analytical Chemistry (邵学广, 陈宗海, 林祥钦. 分析化学), 2000, 28(2):152

[8] Boe B. J Chromatogr A, 2007, 1 139(1):1

[9] Vivo-Truyols G, Torres-Lapasio J R, van Nederkassel A M, et al. J Chromatogr A, 2005, 1 096(1/2):133

[10] Savitzky A, Golay M J E. Anal Chem, 1964, 36(8):1 627

[11] Rutledge D N, Barros A S. Anal Chim Acta, 2002, 454(2):277

[12] Li J W. Anal Chem, 1997, 69(21):4 452

[13] Dondi F. Anal Chem, 1982, 54(3):473

[14] Reh E. TrAC Trends Anal Chem, 1995, 14(1):1

[15] Pirogov A V, Obrezkov O N, Shpigun O A. J Chromatogr A, 1995, 706(1/2):31