ELSEVIER

Journal of Chromatography A, 749 (1996) 5–11

JOURNAL OF
CHROMATOGRAPHY A

Artificial neural networks aided deconvolving overlapped peaks in chromatograms

Huajian Miao*, Menghuai Yu, Shangxu Hu

Laboratory for Intelligent Information Engineering, Department of Chemical Engineering, Zhejiang University, Hangzhou 310027, China

Received 26 March 1996; accepted 15 April 1996

Abstract

A novel method for deconvolving overlapped peaks in chromatograms is proposed. The basic idea of this method consists of finding a set of parameters which characterize the shape of the overlapped peaks and using a multi-layered perceptron network for quantitatively correlating the parameters with the percentage area of an individual peak. The proposed method performs very well with high accuracy and less computing time compared to other conventional methods.

Keywords: Peak overlap; Neural networks, artificial

1. Introduction

In chromatograms each signal peak corresponds to some kind of substance and its quantity can be calculated from the area under the peak. However, many substances have their peaks located very close to each other and they are overlapped in the chromatogram. Hence, one of the important tasks of chromatogram data processing is to deconvolve the overlapped peaks, and find the percentage area belonging to each individual peak.

At present, the methods for deconvolving overlapped peaks can be categorized into three kinds: methods based on geometry, algebra and pattern recognition. The geometrical methods, such as vertical line splitting, tangent line splitting and triangle approximation, are based on simple principles, and the computing times needed are very short; therefore, these methods are often used in real-time processing in spite of their poor accuracy. As for the algebraic methods, the most conventional approach is based on curve fitting, the principle of which is to represent peaks by certain analytical functions with some undetermined parameters and optimize these parameters to approximate the actual chromatogram curve, and the individual peak area can be calculated with the optimized parameters after the precision of approximation is satisfied. However, the optimization procedure requires relatively long computing times such as a couple of seconds, so that this method type can hardly be employed in real-time processing, where people would like to cut the processing time down to some milliseconds. Moreover, the precision of the algebraic methods can not be guaranteed. The methods based on pattern recognition have good prospects and have been developed quickly in recent years. The method proposed in this paper is one of

*Corresponding author.

PII S0021-9673(96)00349-4

these and is aimed at both high accuracy and a reduced computing time requirement.

2. Basic concepts

Generally, the peaks in chromatograms can be described by an asymmetric gaussian distribution (AGD) function, which contains four parameters indicating the peak height, width, position and asymmetry, and is usually expressed as follows:

$$h(t) = H \exp\left[-\frac{1}{2}\left(\frac{t-T}{\sigma}\right)^2 \right] \tag{1}$$

where t is the abscissa, $h(t)$ is the ordinate, i.e. the peak intensity which changes with t, H is the extreme intensity or height of the peak, T is the position of the peak extremity on the abscissa, σ is the parameter denoting the peak width and asymmetry, in which $\sigma=\alpha$ when $t<T$, and $\sigma=\beta$ when $t\geq T$, where α and β are the horizontal distances from the left or right inflection point to the vertical line through the peak extremity; $\alpha+\beta$ is proportional to the peak width, whereas α/β determines the degree of asymmetry.

According to Eq. (1), every peak can be exclusively determined by four parameters, H, T, α and β. Therefore, two overlapped peaks can be determined by eight parameters, i.e., H_1, T_1, α_1, β_1, and H_2, T_2, α_2, β_2. Since the start point of the abscissa and the scale of the coordinates are taken arbitrarily, it is convenient to set $H_1=1000$, $T_1=0$, and $\alpha_1=100$, and only the other five parameters are left undetermined.

In fact, taking account of some particular hypotheses, the number of free parameters can be further reduced. For example, in the chromatogram the width and asymmetry of adjacent peaks can be assumed to take the same values, so only three of the above five parameters are free.

3. Characteristic parameters

For two adjacent peaks, represented by H_1, T_1, α_1, β_1, and H_2, T_2, α_2, β_2 respectively, there are four inflection points on the overlapped peaks curve, or four extreme points A, B, C, D on the first-derivative curve, correspondingly, whose coordinates are (t_A, d_A), (t_B, d_B), (t_C, d_C), and (t_D, d_D) respectively, where d denotes the intensity of the first derivatives. Generally speaking, the overlapping formed by two peaks can be classified into three kinds, namely:

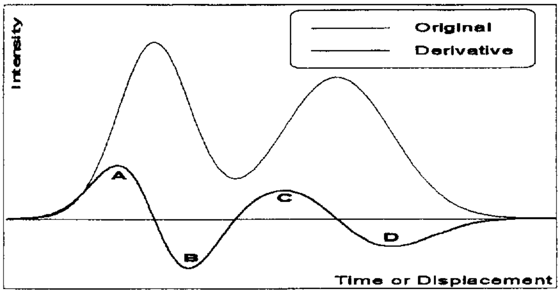

Fig. 1. Distinguishably overlapped peaks and corresponding first-derivative curve.

1. Relatively distinguishable, i.e., d_A, $d_C>0$, and d_B, $d_D<0$, as shown in Fig. 1,
2. Front-shoulder overlapped, i.e., d_A, d_B, $d_C>0$ and $d_D<0$, as shown in Fig. 2, and
3. Rear-shoulder overlapped, i.e., $d_A>0$ and d_B, d_C, $d_D<0$, as shown in Fig. 3.

Different kinds of overlapped peaks may have diverse shapes with one or three extreme points; however, their first-derivative curves have similar shapes, i.e., they show a "max.-min.-max.-min." sequence as illustrated in Figs. 1–3.

In order to work out a general solution, it is better to find the area percentage of any one of the overlapped peaks, rather than to evaluate the absolute values of areas, since the absolute peak areas are

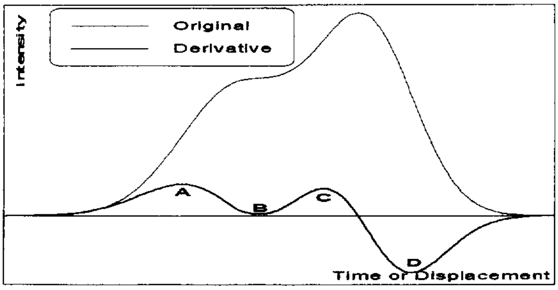

Fig. 2. Front-shoulder overlapped peaks and corresponding first-derivative curve.

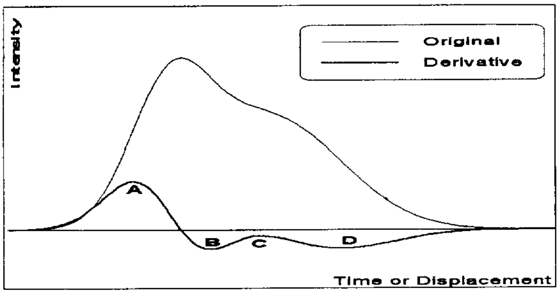

Fig. 3. Rear-shoulder overlapped peaks and corresponding first-derivative curve.

determined by the actual coordinate scales in different cases. Obviously, the relative area should have a relationship with the relative widths and heights of the two overlapped peaks. Two dimensionless parameters defined as

$$p_1 = \frac{t_B - t_A}{(t_B - t_A) + (t_D - t_C)},$$

$$p_2 = \frac{t_C - t_B}{(t_B - t_A) + (t_D - t_C)}$$

are used to denote the relative width, and three dimensionless parameters defined as

$$p_3 = \frac{d_A}{d_A - d_D}, \quad p_4 = \frac{d_B}{d_A - d_D}, \quad p_5 = \frac{d_C}{d_A - d_D}$$

are used to denote the relative heights of the overlapped peaks. These dimensionless groups are called the characteristic parameters.

There must exist some relationship between these characteristic parameters and the area percentage of an individual peak, but the explicit expression of the relationship can hardly be derived. Considering the requirements of simulating an unknown implicit function, the artificial neural network (ANN) was selected to express this kind of relationship. For convenience, the multi-layered feed-forward (MLF) network is used and trained according to the error back-propagation (EBP) strategy.

4. Training the artificial neural networks

In order to make application practical, the following section will be restricted to the overlapped peaks which are composed of two individual peaks only. We hypothesize that the chromatography peaks can be described by the AGD function and denote the two individual peaks as peak 1 and peak 2, where peak 1 is regarded as the reference peak and we assign its parameters as $T_1 = 0$, $H_1 = 1000$, $\alpha_1 = 100$, and $\beta_1 = 100\Phi_1$, where Φ_1 is a variable factor depicting the asymmetry of peak 1. As mentioned above, the width and asymmetry of two adjacent chromatography peaks can be assumed to be equal, and are expressed in following equations:

$$\alpha_1 + \beta_1 = \alpha_2 + \beta_2, \text{and } \alpha_1/\beta_1 = \alpha_2/\beta_2$$

From the above relationships we obtain:

$$\alpha_2 = \alpha_1 = 100, \text{and } \beta_2 = \beta_1 = 100\Phi_1$$

Since the area of an AGD peak is $A = \sqrt{\pi/2}H(\alpha + \beta)$, the percentage area of peak 1 to the whole overlapped peak area is:

$$Q_1 = A_1/(A_1 + A_2) = H_1/(H_1 + H_2)$$

and, hence, it follows that:

$$H_2 = H_1(1 - Q_1)/Q_1$$

According to the above equations, all the parameters for the two individual peaks can be determined when the asymmetry factor of peak 1, Φ_1, the percentage area of peak 1, Q_1, and the position of peak 2, T_2, are given. In this way, the first-derivative curve of the two overlapped peaks can be readily evaluated.

To form a set of data for training the MLF, assign $\Phi_1 = 0.1$, 0.2, 0.5, 1, 2, 5, 10 and $Q_1 = 0.1$, 0.2,, 0.3,..., 0.9. Since two peaks are overlapped completely when $T_2 = T_1$, and almost not overlapped when $T_2 = T_1 + 3(\beta_1 + \alpha_2)$, the space between 0 and $300(1 + \Phi_1)$ was partitioned and we got ten well-distributed values of T_2. In this way, 630 patterns were obtained, each pattern a vector composed of p_1, p_2, p_3, p_4, p_5, and the corresponding Q_1 is its counterpart. For training the ANN we use the vector $(p_1, p_2, p_3, p_4, p_5)$ as the network input and Q_1 as the target output.

The MLF network used consists of 5 nodes in the input layer, 10 nodes in the hidden layer, and 1 node in the output layer. By systematic tests, we found the accuracy of the final results was not very much related to the number of hidden layer nodes such as 10, 15, 20 or more, so the lower number of 10 was selected. The processing function of the node is the Sigmoid function, i.e.,

$$r_j = 1/(1 + \exp(-s_j + \theta_j))$$

where s_j and r_j are the input and the output of node j respectively, and θ_j is the threshold of node j. The input of node j is the sum of the multiplication of r_k and w_{jk},

$$s_j = \sum_k r_k w_{jk}$$

where r_k is the output of the last layer node k and w_{jk} is the weight connecting nodes j and k. The BP algorithm [1] was used to train the weights and thresholds in the above network. We initialized the net weights at random, and set the momentum rate and learning rate to 0.9 and 0.7, respectively. In training, we randomly selected 500 out of the above 630 patterns to train the net and used the other 130 patterns to cross-validate the net performance. After 3000 iterations, the average absolute error between the target and the network output decreased to a small value, 0.00007, and then the training was ended.

Denoting the input, hidden, and output layers as A, B and C, all the weights and thresholds obtained after training are shown in Table 1. As far as the overlapped peaks of chromatography are concerned, it has been already pointed out that only three characteristic parameters are necessary for determining the individual peak percentage area; the others are redundant. Therefore, a test was made such that only three characteristic parameters, p_3, p_4 and p_5, were used as the network inputs, and the network was then trained in the same way as above. The average absolute error between the target and the network output decreased to 0.00033. Though the error is still small, it is much bigger than the one generated when five parameters are used. Accordingly, it may be deduced that, though the other two characteristic parameters are redundant in a physical sense, they help to train the network. To test the robustness, we added random noises to the five inputs with a maximal 10% relative error, the relative errors of the network outputs never exceeded 6%; this reveals the stability of net against noises, which is one of the advantageous features of ANN. A more valuable test performed was that even when five characteristic parameters from various overlapped peaks synthesized by the exponentially modified Gaussian (EMG) model [2,3] were taken, the relative error of the network outputs seldom exceeded 4%; this shows the insensitivity of the ANN method to the peak model.

It is worth noting that the way of selecting the characteristic parameters is not casual. In fact, when p_3 was substituted by another dimensionless group $d_A/(d_C - d_B)$, no matter how the network structure is reformed, the average absolute error reached is always greater than 2.7, and the network output is very different from the target, which reveals that the network does not depict the expected relationship.

When three or more peaks are overlapped, according to the AGD function, $4n$ parameters are needed to determine the overlapped peaks, where n is the number of individual peaks. As mentioned above, we can set the position, height and width of the first

Table 1
Weights and thresholds of the ANN after training

	A_1	A_2	A_3	A_4	A_5	Thresholds	C_1
B_1	−0.713133	−8.344479	0.510332	−0.193877	−0.160070	−1.085222	−5.721345
B_2	−0.413454	−5.244912	−0.041913	−0.082980	−5.798004	−0.249638	5.869261
B_3	−4.122913	0.550064	−0.330824	−7.180707	−0.839873	−0.514240	−2.247042
B_4	7.087600	0.112555	−2.498871	1.212016	0.705067	−0.434963	0.531882
B_5	−7.317007	−0.448445	0.410964	0.585086	8.079370	−1.042037	−1.492455
B_6	0.160550	1.529279	−1.183294	−0.502343	−0.465771	−2.952801	−4.029602
B_7	1.117842	1.872316	−0.581536	0.484731	−3.781395	0.933481	1.138486
B_8	−0.554405	−0.690276	1.121064	1.230760	−9.053371	4.847096	4.478941
B_9	4.334997	0.112802	−2.150803	6.066226	−0.373994	−0.439049	1.255853
B_{10}	8.816632	−0.511345	−0.904907	−1.065835	−1.260757	0.101028	0.697668
Thresholds	–	–	–	–	–	–	0.561587

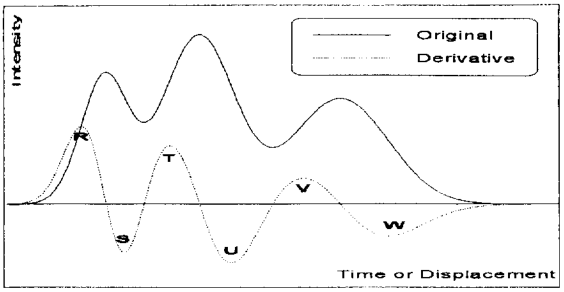

Fig. 4. Three overlapped peaks and corresponding first-derivative curve.

individual peak arbitrarily, and hypothesize that all individual peaks are asymmetric and all widths are equal; thus, only $4n-3-2(n-1)=2n-1$ parameters are free and the same number of characteristic dimensionless parameters should be evaluated from the overlapped peaks to determine the area percentage of an individual peak. Usually, there exists $2n$ inflection points on the curve composed of n overlapped peaks, so in the intensity direction $2n-1$ dimensionless parameters from the $2n$ extreme points on the first-derivative curve can be obtained to meet the requirements of the characteristic parameters, just as the p_3, p_4 and p_5 do in the case when two peaks are overlapped. For example, when three peaks are overlapped, we can obtain five dimensionless parameters taking the following form $(q_1, \ldots, q_5)$ from the six extreme points on the first-derivative curve shown in Fig. 4. Thus, the proposed method can also be used in cases when two more peaks are overlapped, but this paper only highlights the evaluation of the individual peak area in the case when only two peaks are overlapped.

$$q_1=\frac{d_R}{d_R-d_W},\quad q_2=\frac{d_S}{d_R-d_W},\quad q_3=\frac{d_T}{d_R-d_W},\quad q_4=\frac{d_U}{d_R-d_W},\quad q_5=\frac{d_V}{d_R-d_W}$$

5. Comparison between different methods

In order to compare the accuracy of the ANN method with other conventional methods such as vertical line splitting and curve fitting, a series of experiments was conducted. The experimental data was taken on a chromatographic meter by injecting pure reagent twice with different known amounts in a short time to make artificially overlapped peaks. Because the overlapped two peaks correspond to the same substance under identical operating conditions, the area proportion of the overlapped two peaks should be equal to the known proportion of dosage injected. A general comparison among the three kinds of methods was made after performing the same set of experiments by using different samples under different operation conditions. To save space here, only one set of these results is listed in Table 2.

From the experiments we conducted, the following conclusions can be reached.

5.1. Vertical line splitting method

Splitting the overlapped peaks via a vertical line through the overlapped peaks valley point, the accuracy of this method varies in different situations and becomes rather poor when two peaks are severely overlapped, are very different in size, or are quite asymmetrical [4–6]. Moreover, when there is only one maximal point on the overlapped peaks, this method is wholly incompetent due to its simple principle.

5.2. Curve-fitting method

A good function model should be first selected to depict the peak, but different functions may lead to different deconvolution results. In fact, this method is so sensitive to the model selected that if the model slightly disagrees with the real peak shape the deconvolution results will diverge. Even if the function selected is the same but a different optimization algorithm or different initial values of the parameters are used, the deconvolution results may be much different because of the local minimal points in the object function [7,8]. Therefore, the curve-fitting method is hardly applicable in practice for general purposes. In this paper, the AGD function was used to model the chromatography peak and the Marquardt algorithm [9] was used to optimize the function parameters. Because the initial values of the function parameters are difficult to assign properly, the curve-fitting method did not only cost a consider-

Table 2
Comparison of three kinds of methods for deconvolving overlapped peaks

Proportion of injection dosage	Area proportion by the ANN	Relative error (%)	Area proportion by vertical line splitting	Relative error (%)	Area proportion by curve fitting	Relative error (%)
10.00	9.61	3.9	2.33	76.7	2.38	76.2
9.00	8.63	4.1	5.48	39.1	3.98	55.8
8.00	8.21	2.6	4.24	47.0	3.21	59.9
7.00	6.83	2.4	3.81	45.6	3.14	55.1
6.00	6.15	2.5	4.17	30.5	4.11	31.5
5.00	5.23	4.6	3.04	39.2	4.82	3.6
4.00	4.21	5.3	2.30	42.5	2.06	48.5
3.00	2.91	3.0	2.15	28.3	1.94	35.3
2.00	2.07	3.5	1.36	32.0	1.36	32.0
1.00	1.04	4.0	0.40	60.0	0.73	27.0
0.50	0.52	4.0	0.29	42.0	0.23	54.0
0.33	0.32	3.0	0.25	24.2	0.20	39.4
0.25	0.27	8.0	0.24	4.0	0.36	44.0
0.20	0.22	10.0	0.17	15.0	0.24	20.0
0.17	0.16	5.9	0.18	5.9	0.28	64.7
0.14	0.14	0.0	0.14	0.0	0.19	35.7
0.13	0.12	7.7	0.13	0.0	0.50	284.6
0.11	0.10	9.1	0.12	9.1	0.15	35.4
0.10	0.10	0.0	0.11	10.0	0.21	110.0
Average error		4.4		29.0		58.6

able amount of computing time for the iterations, but it also sometimes gave results which were not reasonable; this may be due to the fact that the optimization of the parameters is easily trapped in the local minimal point.

5.3. ANN method

The accuracy of the ANN method is much better no matter how severe the degree of overlap of the two individual peaks is. Because the trained ANN needs only a little computing time to process the input pattern, the ANN deconvolving method is suited for use in real-time applications. In order to put the ANN method into practice, the positions and heights of the inflection points on the first-derivative curve should be determined exactly at first. Fortunately, the peak-detecting method based on the quasi-second derivatives [10] and the high-fidelity filtering method based on medians [11], both proposed by the same authors, will satisfy this requirement.

Acknowledgments

This project was supported by the National Science Foundation of China.

References

[1] D.E. Rumelhart, G.E. Hinton and R.J. Williams, in D.E. Rumelhart and J.L. McClelland (Editors), Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. 1: Foundations, MIT Press, Cambridge, MA, 1986, pp. 318–362.

[2] Mark S. Jeansonne and Joe P. Foley, J. Chromatogr., 461 (1989) 149–163.

[3] Joe P. Foley, Anal. Chem., 59 (1987) 1984–1987.

[4] Walerian J. Chromatogr. Sci., 19 (1981).

[5] Andrew N. Papas and Terrence P. Tougas, Anal. Chem., 62 (1990) 234–239.

[6] Joe P. Foley, J. Chromatogr., 384 (1987) 301–313.

[7] J. Grimalt, H. Iturriaga and J. Olive, Anal. Chim. Acta., 201 (1987) 193–205.

[8] Rajeev A. Vaidya and Roger D. Hester, J. Chromatogr., 287 (1984) 231–244.

[9] P.R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1969, p. 235 and p. 269.

[10] Huajian Miao and Shangxu Hu, Chinese J. Anal. Chem., 22 (1994) 247–250.

[11] Huajian Miao and Shangxu Hu, J. Chemistry of Chinese Universities, 16 (1995) 1020–1023.