Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Chromatography A, 1096 (2005) 133–145

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals Part I: Peak detection

G. Vivó-Truyols[a], J.R. Torres-Lapasió[a,*], A.M. van Nederkassel[b], Y. Vander Heyden[b], D.L. Massart[b]

[a] *Department of Analytical Chemistry, Universitat de València, c/Dr. Moliner 50, 46100 Burjassot, Spain*

[b] *Department of Pharmaceutical and Biomedical Analysis, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*

Available online 4 May 2005

Abstract

A series of two papers describing a procedure for automated peak deconvolution is presented. The goal is to develop a package of routines that can be used by non-experienced users. Part I (this paper) concerns peak detection, whereas Part II is dedicated to the deconvolution itself. In this first part, the most interesting features of the peak detection algorithms, which precede the deconvolution step, are outlined. High-order derivatives provide valuable information to assess the number of underlying compounds under a given peak cluster. A smoothing technique was found essential to compute properly the derivatives, since the noise is amplified when differences are calculated. The Savitsky–Golay smoother was applied in combination with the Durbin–Watson criterion to automate the window size selection. This strategy removed the noise without loosing valuable information. In some cases, it was found preferable to split the chromatogram in different elution regions, and apply the Durbin–Watson test and the Savitsky–Golay smoother to each region, separately. The derivatives allowed obtaining estimates of both peak parameters and the corresponding ranges for each eluting compound to be used in the deconvolution. An algorithm oriented to compare peaks from different chromatograms is also presented to perform deconvolution, using information from several related chromatograms.

Keywords: Overlapped peaks; Chromatographic one-way signals; Peak detection; Smoothing

1. Introduction

The resolution capability of chromatographic methods is limited. It frequently happens that the best separation still does not allow a satisfactory detection and quantitation of all components. Chemometrics can help to fill this gap. It not only assists the chromatographer in the design of experiments, the search for the best separation conditions and the analysis of the gathered data, but also provides solutions for partial (or even full) overlap of peaks. The partial separation achieved in the chromatographic domain can be completed –at least in some extent– by mathematical means.

Partial chromatographic selectivity can be tackled, for instance, with multi-channel detectors, which provide additional data orders. To take advantage of it, a new family of chemometric techniques has emerged, focused on the extraction of chromatographic information. For a review, see Ref. [1]. Some other deconvolution techniques for single-channel detection have been developed. They are based on the assumption that the underlying individual peak profiles confounded within the gross chromatographic signal can be described through mathematical peak models. This has led to an increased interest in the development of better peak models [2–7].

The use of these deconvolution techniques requires adapting the method to the characteristics of each chromatogram, in order to optimise the performance of the treatment and assure the quality of the results. This makes the routine application of deconvolution to non-experienced users difficult. Many decisions comprising the full process must be made in each step, before the deconvolution itself can be

[*] Corresponding author. Tel.: +34 963543003; fax: +34 963544436.
E-mail address: jose.r.torres@uv.es (J.R. Torres-Lapasió).

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.03.092

carried out. These decisions deal with peak detection, selection of the most appropriate deconvolution tool and peak model, and setting both the initial estimates and, in some instances, the searching ranges for the model parameters. Additionally, new considerations arise when several related chromatograms are treated altogether in order to reinforce the available information [8]. The main objective of this work is to develop a deconvolution program, able to be run with minimal user interaction and focused on complex multi-peak chromatograms. The goal is that a non-experienced user (e.g. a laboratory technician) should be able to perform peak deconvolution with minimal knowledge of both the sample and the chemometric tools applied, with reasonable expectations of success.

This work is divided in two parts. Part I describes an algorithm to analyse and prepare the data, in order to apply efficiently deconvolution either to a single sample or to a set of them. This involves peak detection, setting appropriate initial estimates and searching boundaries for each parameter, which is required in some deconvolution algorithms. Part II describes different deconvolution methods and an algorithm oriented to assess the complexity of the data, in order to select automatically the best mathematical tool and the most suitable peak model for each (set of) chromatogram(s).

This first part starts with an automated procedure to assess correctly the number of peaks. Peak detection algorithms often have difficulties in detecting the presence of more than one peak when several compounds coelute, yielding shoulders on the main peak [9–11]. Deconvolution constitutes an attractive possibility, especially in these situations, and therefore, a method to evaluate automatically the number of peaks and assess correctly the initial parameter estimates should be developed for cases of strong overlap. To detect the underlying peaks, the derivatives of the chromatographic signal are inspected. The *n*-order derivatives are usually computed through the well-known Savitsky–Golay (SG) method [12]. This method not only determines the derivatives but it also smoothes the chromatographic signal, to compensate the effect of noise amplification when the derivatives are computed. One of the critical parameters in the smoothing technique to be set is the window size (i.e. the number of neighbouring points used to fit the polynomial), which depends on the signal properties. In this work, the Durbin–Watson (DW) test is applied to automatically select the adequate window size. The method was tested with both simulated and experimental chromatograms.

2. Theory

2.1. Signal smoothing and derivative calculation

The smoothed chromatogram and up to third-order derivatives were calculated according to the SG algorithm [12]. This algorithm allows computing in a single step both the smoothed signal and the derivatives of the fitted polynomial [13]. Some conditions are required to apply SG. One of the requirements of the original SG algorithm is that the signal should be sampled at a constant rate. However, in situations of fast elution, sudden variation of the sampling frequency throughout the analysis may occur, since lower retention times require higher frequencies to avoid undersampling. For this reason, the SG algorithm was applied independently to zones where the program detects a constant sampling rate so that it was possible to tackle different sampling frequencies within the same chromatogram. In order to detect changes in sampling rate, the user should input to the program not only the signal vector(s), but also the corresponding time vector(s). To get more flexibility, the convolution coefficients required to smooth or differentiate the signals were computed within the program in all cases.

Two parameters must be selected to apply the SG technique: the polynomial degree and the window size (i.e. the number of neighbouring points used to fit the polynomial). These parameters determine the flexibility of the smoothing procedure and should be chosen with care. A too flexible smoothing (i.e. high polynomial degrees and small window sizes) yields noisy chromatograms, and noisy and biased derivatives. On the other hand, low polynomial degrees and large window sizes generate smoothed chromatograms with flattened peaks and again, biased derivatives [14]. An ideal smoother should remove the noise though preserving the valuable chromatographic information.

The selection of the best polynomial degree and window size is difficult to automate. The most adequate values depend strongly on the sampling frequency (i.e. the number of points per second), the noise, and the peak width. These features can change from sample to sample, and in some instances, within a given chromatogram. Since the most influential parameters is the window size, we decided to simplify the procedure by keeping the less critical one (i.e. the polynomial degree) fixed, applying in all instances a second-degree polynomial, so that only the most adequate window size was determined. Nevertheless, the polynomial degree can be changed manually by the user, but the window size was always automatically selected. This was done by applying the DW test [15] to the residuals obtained from the difference between the original ($\boldsymbol{y}_{\text{exp}}$) and smoothed ($\boldsymbol{y}_{\text{smd}}$) chromatogram. This test is based on the computation of the DW statistic:

$$\mathrm{DW} = \frac{\sum_{i=2}^{n}[(y_{\text{exp},i} - y_{\text{smd},i}) - (y_{\text{exp},i-1} - y_{\text{smd},i-1})]^2}{\sum_{i=1}^{n}(y_{\text{exp},i} - y_{\text{smd},i})^2} \quad (1)$$

where n is the number of points in the chromatogram, and $y_{\text{exp},i}$ and $y_{\text{smd},i}$ the ith values of the original and smoothed signals, respectively. Eq. (1) requires a final correction to account that the numerator includes one measurement less than the denominator (this adjustment is particularly required for low n values):

$$\mathrm{DW} = \frac{\sum_{i=2}^{n}[(y_{\mathrm{exp},i}-y_{\mathrm{smd},i})-(y_{\mathrm{exp},i-1}-y_{\mathrm{smd},i-1})]^2}{\sum_{i=1}^{n}(y_{\mathrm{exp},i}-y_{\mathrm{smd},i})^2} \times \left(\frac{n}{n-1}\right) \qquad (2)$$

It should be noted that the DW test is not applied here as explained in Ref. [16]. The DW statistic determines if consecutive points in a signal that oscillates around zero (i.e. a mean-centred signal) have often the same sign. If they do, then they are called correlated signals. Here, the considered signal is, in fact, the difference between y_{exp} and y_{smd} at consecutive points in the chromatogram. If these differences (residuals) have the same sign, then it means that y_{smd} diverges from y_{exp} always in the same direction (i.e. there is a systematic difference, which is undesirable). It can be concluded that the smoothing technique did not remove the noise only. In contrast, uncorrelated residuals denote that only the noise has been eliminated. The purpose of the DW criterion here is to determine which window size yields differences between y_{exp} and y_{smd} that are as little correlated as possible. If there is no correlation between residuals (i.e. they are randomly distributed, and therefore, an optimal smoothing is obtained), the DW value converges to 2 [17].

The application of the DW criterion implies monitoring the statistic with different window sizes for the SG smoothing. The window size yielding a smoothed chromatogram with a DW value closest to 2 is considered to be the optimal. In this way, an automatic selection of this parameter is possible, even when different chromatograms are being processed. According to these results, first-, second- and third-order derivatives can be computed as explained above. Since a second-degree polynomial is used, the third-order derivative cannot be computed directly. To overcome this problem, the third-order derivative was computed from the second-order derivative by applying to it the first-order SG derivative.

Another problem arises with large chromatograms containing peaks with different band broadenings, since band broadening for the low retained solutes can be significantly smaller than for compounds with longer retention times. This effect is particularly conspicuous in chromatograms obtained under isocratic conditions in low efficiency columns. Under these conditions, a unique window size is not optimal for the whole chromatogram: peaks at the beginning of the chromatogram will tend to be more distorted, whereas peaks at the end of the chromatogram will not be properly filtered. To tackle cases like this, a temporary smoothed signal together with its derivatives are calculated in a first step, using the same window size for the whole chromatogram. Then, the chromatogram is split in several blocks (see Section 2.4), for each of which an optimal window size is determined by applying the methodology exposed above.

2.2. *Peak detection*

The implemented algorithm makes use of both the derivatives and the input signal. Fig. 1 illustrates the shape of the derivatives for a single experimental peak. The chromatogram of triphenylene injected at 85% methanol in water (Fig. 1a) is depicted, together with the first-, second- and third-order derivatives (Fig. 1b–d, respectively). The computation of the derivatives was performed according to Section 2.1 (in this case, a five-point window was found optimal for the SG smoothing with a second-degree polynomial).

As can be seen, a single-positive peak in the input signal yields two bands in the first derivative, with a positive band at the left side and a negative band at the right side. The peak region is found by considering the times at which the first derivative is below a certain threshold (e.g. five-fold the noise) at both sides of the retention time (Fig. 1b). The second derivative presents a negative region, together with two positive regions around it (Fig. 1c), and four bands are observed in the third derivative (Fig. 1d). When only one compound is eluting, as is the case in the figure, three changes in sign (labelled as 1, 2 and 3 in Fig. 1d) are detected within the elution region.

The peak detection algorithm is based on finding negative regions in the second derivative. However, not all negative regions are due to the elution of a compound. Peaks can also be due to noise. For accounting it, the noise ($\varepsilon_{\mathrm{sd}}$) in the second derivative is computed to establish a cut-off value that allows distinguishing real peaks from noise. For each point p_i of the second derivative, the distance, h_i, from this point to the mean of its neighbouring points (p_{i-1} and p_{i+1}) is computed (obviously, the h_i values for the first and last points cannot be calculated). The noise is defined as the median of the absolute h_i values. This result is used to calculate a threshold value (e.g. five-fold the noise), called $\mathrm{thr}_{\mathrm{sd}}$, which is depicted as a horizontal line overlaid in Fig. 1c. The time ranges where a negative value of the second derivative falls below this threshold are considered as domains where a compound was eluting.

In some cases, the magnitude of this second derivative threshold, $\mathrm{thr}_{\mathrm{sd}}$, is not sufficiently selective, and can lead to the misidentification of an incidental deviation as a peak. For this reason, two additional conditions were also imposed before accepting a perturbation as a peak. Both require that the value of the input signal at the retention time (h_1, Fig. 1a) should be higher than a threshold value. The first signal threshold, $\mathrm{thr}_{\mathrm{h1}}$, is equal to three-fold the noise. The second one, $\mathrm{thr}_{\mathrm{h2}}$, is selected by the user, and requires that the response should be higher than this value. The need for the second signal threshold can be seen in Fig. 1a, where the value of $\mathrm{thr}_{\mathrm{h2}}$ was set at 10 mAU. In this case, $\mathrm{thr}_{\mathrm{h1}}$ is too low, which means that some variations of the signal will be wrongly identified as eluting compounds if only $\mathrm{thr}_{\mathrm{h1}}$ were considered.

How the algorithm detects peaks when they are confounded depends on the overlapping degree and is described in the following sections.

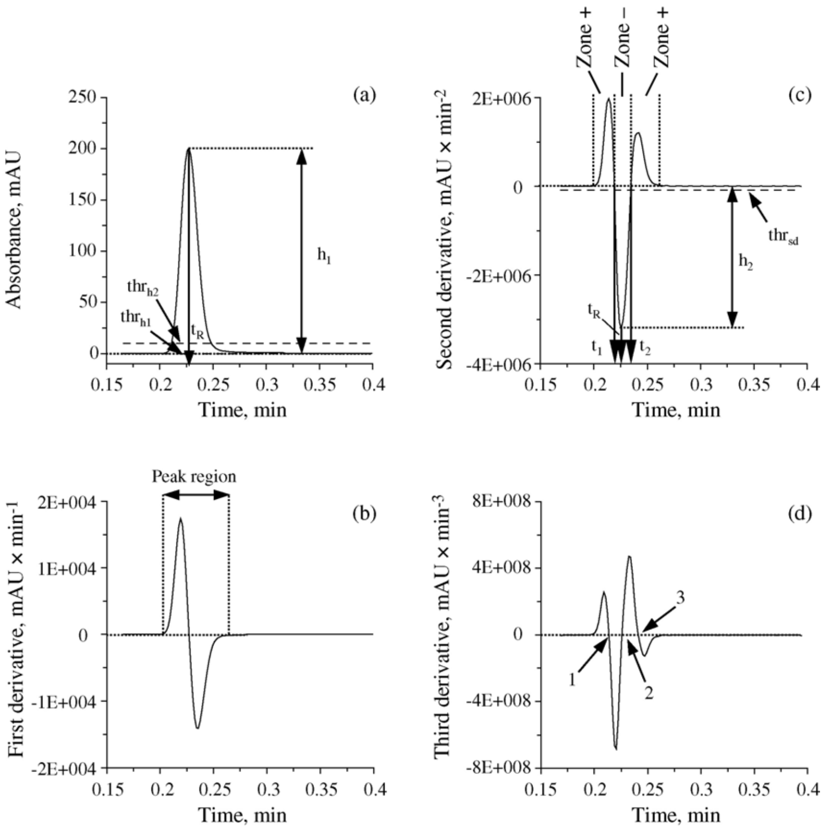

Fig. 1. Chromatogram (a) and first- (b), second- (c) and third-order (d) derivatives of an injection of triphenylene. Mobile phase: 85% methanol. Thresholds for peak detection in the input signal (thr_{h1} and thr_{h2}) and in second derivative (thr_{sd}) are also depicted (dashed lines). The peak region, the different zones of the second derivative (positive or negative), as well as the meaning of t_R, h_1, h_2, t_1 and t_2, are also indicated. Changes in sign of the third-order derivative are numbered from 1 to 3.

2.2.1. Moderate coelution (case i)

Figs. 2 and 3 consider two cases of moderate coelution of two compounds, namely toluene (Tol) and ethylbenzene (Eth), eluted with a mobile phase containing 80% methanol. An impurity ("Imp" in the figures) was also detected with the peak detection algorithm, but we will focus only on the detection of toluene and ethylbenzene.

The difference between the two examples is the peak height ratio of the compounds. In the case presented in Fig. 2, the peak height of the two compounds was the same. The first derivative indicates two separate peak regions for each compound (Fig. 2b). In the second case (Fig. 3), only one peak region is found. However, in both cases, two negative zones, each one corresponding to a single compound, are evident in the second derivative. A value of the second derivative below thr_{sd} (Figs. 2c and 3c) is found in each of these two negative zones, which confirms that the negative values are not due to noise. Further, the input signal is higher than thr_{h1} and thr_{h2} (Figs. 2a and 3a). Therefore, one can conclude that there are two peaks eluting (besides "Imp").

As in the example presented in Fig. 1, only one change in sign of the third derivative is detected within each sign zone (i.e. positive or negative: "Zone+" or "Zone−" of the second derivative). In Fig. 2d, zero values of the third derivative are found at "1Tol", "2Tol" and "3Tol" within the peak region of toluene, and "1Eth" and "2Eth" within the peak region of ethylbenzene, each one in a different zone of the second derivative. A similar situation is found in Fig. 3, with the difference that only a single elution region is found.

A case of coelution will be classified as a moderate overlap situation –case (i)– if the following condition is fulfilled:

$$n_3 \leq 2n_2 + 1 \tag{3}$$

where n_3 is the number of changes in sign of the third derivative within an elution region, in which n_2 significant negative regions are found.

2.2.2. Strong coelution (case ii)

When the overlap is high (strong coelution: case ii) and only a slight shoulder is found in the chromatogram, only one negative zone of the second derivative is detected, although more than one compound is eluting. In such a case, the third derivative allows to evaluate correctly the number of underlying peaks. Depending on the zone of the second derivative where additional changes in sign of the third derivative are detected, two different cases (ii-a and ii-b) can be distinguished.

Case ii-a is illustrated in Fig. 4. In this figure, the mixture of toluene and ethylbenzene from Fig. 2a was eluted with 85%

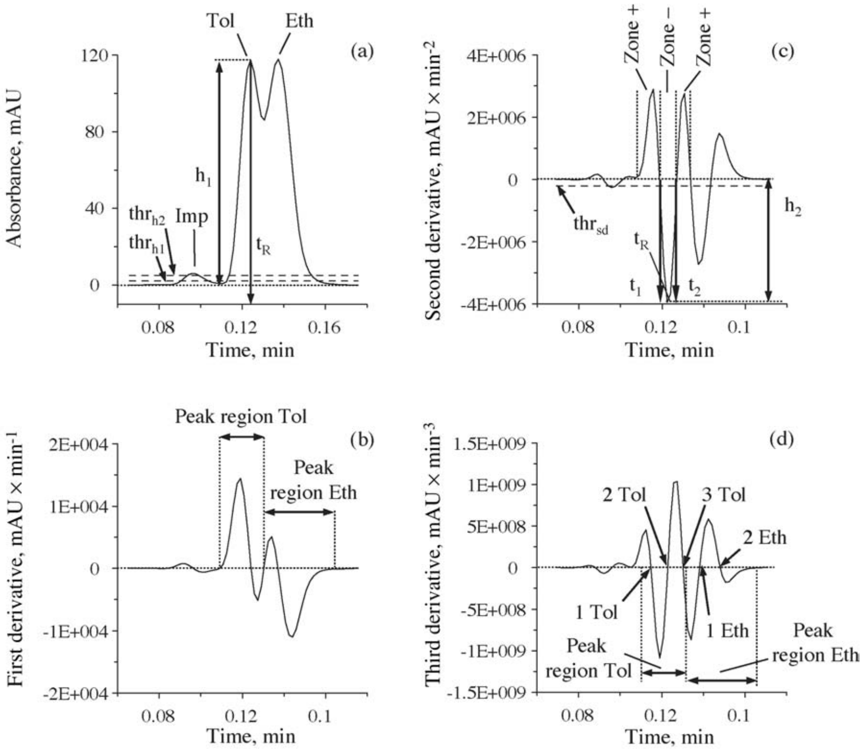

Fig. 2. Chromatogram (a) and first- (b), second- (c) and third-order (d) derivatives for toluene (Tol) and ethylbenzene (Eth). Mobile phase: 80% methanol. Both analytes have the same peak height. The same analysis plotted in Fig. 1 is depicted here for toluene. The peak “Imp” is a perturbation also found by the peak detection algorithm.

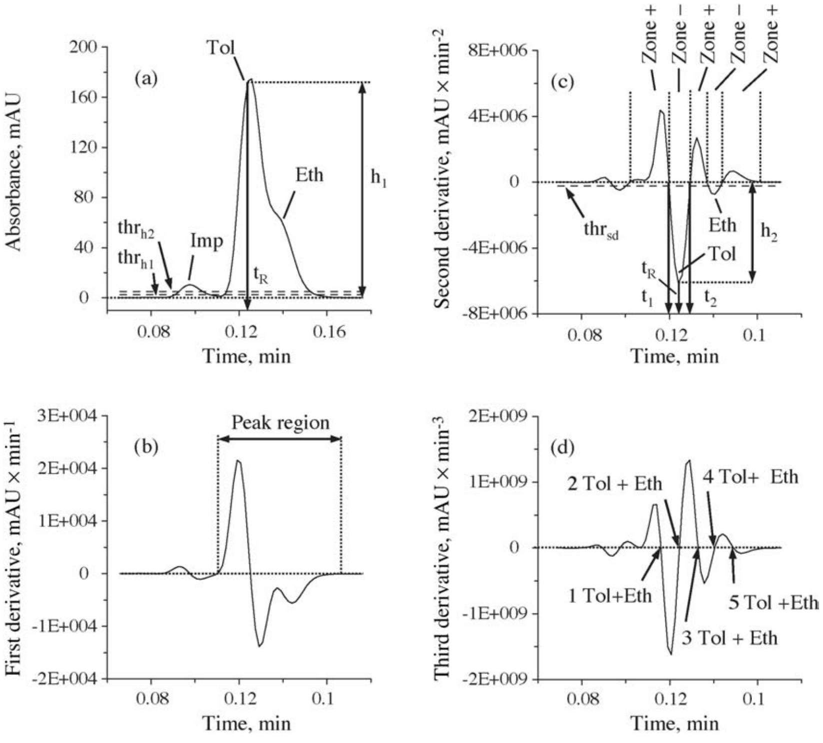

Fig. 3. Chromatogram (a) and first- (b), second- (c) and third-order (d) derivatives for toluene (Tol) and ethylbenzene (Eth). Mobile phase: 80% methanol. The smaller size of Eth with regard to Tol, makes the difference with Fig. 2. The same analysis plotted in Fig. 1 is depicted.

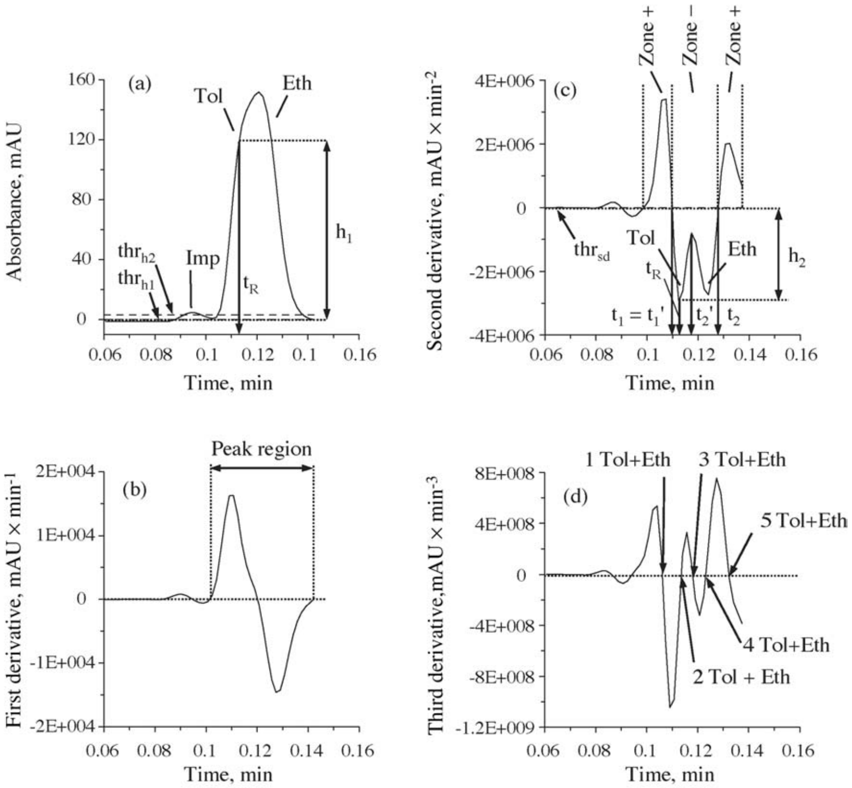

Fig. 4. Analysis of the same mixture as in Fig. 2 eluted with a mobile phase containing 85% methanol.

methanol, leading to a higher overlap. The chromatogram is shown again, together with the first, second and third derivatives. Only one significant negative region can be found in the second derivative at the peak region. However, there are five changes in sign of the third derivative in the elution region of the peak (indicated in Fig. 4d as “1Tol + Eth”, “2Tol + Eth”, “3Tol + Eth”, “4Tol + Eth” and “5Tol + Eth”). This implies that the condition of Eq. (3) is not met, since $n_2 = 1$ and $n_3 = 5$. Therefore, it can be concluded that there is strong coelution (case ii). Note that “2Tol + Eth”, “3Tol + Eth” and “4Tol + Eth” are detected within the same zone (“Zone−” of the second derivative). Since all these changes are found when the second derivative is negative, this case is classified as (ii)-a.

The general formula to determine the number of eluting compounds in a case of strong coelution –case (ii)– is the following:

$$n = \begin{cases} n_2 + \dfrac{n_3 - 2n_2 - 1}{2} & \text{if } n_3 \text{ is odd} \\ n_2 + \dfrac{n_3 - 2n_2}{2} & \text{if } n_3 \text{ is even} \end{cases} \qquad (4)$$

where n is the number of compounds, and n_2 and n_3 are defined as in Eq. (3). In the previous example, $n = 2$ and two compounds, toluene and ethylbenzene, are detected.

A (ii)-b situation is given in Fig. 5, where the mixture considered in Fig. 3 was eluted with 85% methanol. Fig. 5b and c show that Eq. (3) is not fulfilled: only one significant negative region of the second derivative is found in a single peak region, but more than three changes in sign of the third derivative are found within this peak region. This shows that there is not one, but more compounds present. In this case, “3Tol + Eth”, “4Tol + Eth” and “5Tol + Eth” are found when the second derivative is positive. Since more than one change in sign is found in a “Zone+” of the second derivative, this coelution case is classified as (ii)-b.

Additional requirements are imposed before applying Eq. (4) and deciding on the number of eluting compounds. They are slightly different for cases (ii)-a and (ii)-b. For (ii)-a, the value of the second derivative where all the secondary minima are found should be below thr$_{sd}$. This is achieved in the example of Fig. 4, in which only the retention time for toluene is depicted for clarity. When this condition is not fulfilled, a case (ii)-b is concluded. For (ii)-b, it is not the value of the second derivative but its height h_p (Fig. 5c) what should be higher than thr$_{sd}$. In both cases, the condition that the peak height should be higher than thr$_{h1}$ and thr$_{h2}$ should also be accomplished for $t = t_R$. These conditions assure that the minima yielding an additional change in sign of the third derivative are not due to noise.

The analysis of the number of changes in sign of the third derivative should be considered with care. In situations where coelution can be only detected through the third derivative (like those presented in Figs. 4 and 5), the subsequent deconvolution can yield inaccurate results, even if the correct number of compounds was determined [18]. However, in these

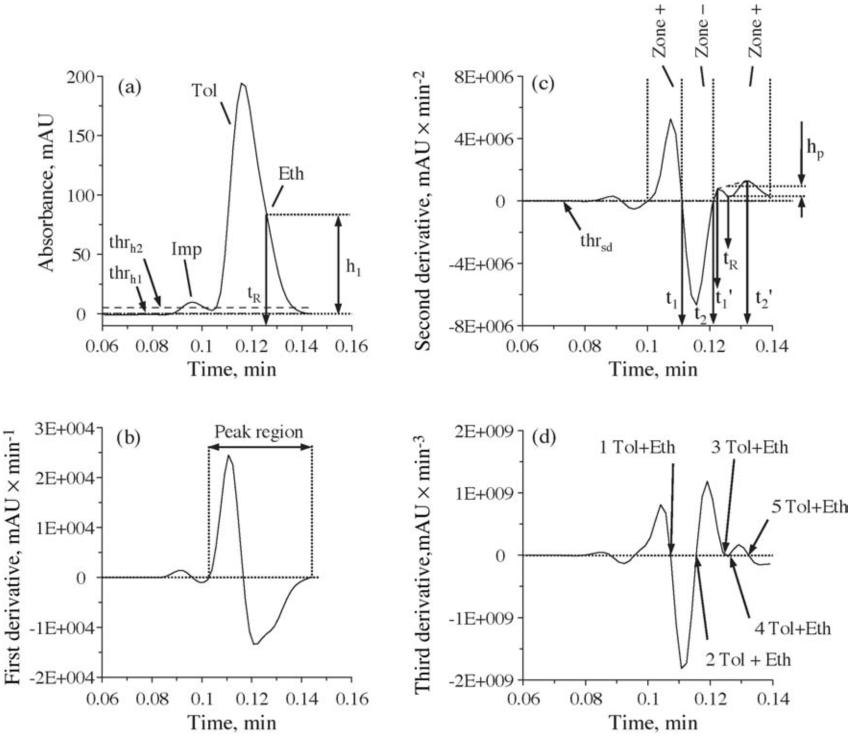

Fig. 5. Analysis of the same mixture as in Fig. 3 eluted with a mobile phase containing 85% methanol. The analysis of the figures of merit is performed for Eth.

situations of strong overlap, reasonable results can be found if the information from several injections is combined [8,19]. As a practical rule, we recommend the use of the third derivative only when several chromatograms are available.

2.3. Peak parameter assessment

2.3.1. Computation of the initial guesses and parameter boundaries

A chromatographic signal is a linear combination of individual peaks, each one described by a peak model. The deconvolution process requires initial guesses for the parameters of the solute peak model. In some instances, ranges within which the final parameter value is expected to vary must be computed, because these ranges are needed as constraints in some of the algorithms that are used in the data treatment.

The peak model used in this work is a modification of a Gaussian model [20]:

$$h(t){=}h_0\ \exp\left[-\frac{1}{2}\left(\frac{t-t_R}{s_0{+}s_1(t{-}t_R){+}s_2(t{-}t_R)^2{+}\dots}\right)^2\right] \quad (5)$$

where h_0 is the peak height, t_R the retention time, s_0 the standard deviation, and s_1, $s_{2,\dots}$ s_n are terms related to the peak tailing. Problems due to abnormal baseline raisings when this model is used for deconvolution are explained in detail in Part II of this work.

The parameters that require initial estimates to be estimated for each solute are h_0, t_R, s_0, and s_1. These values are calculated from figures of merit of the input signal and the second derivative. Cases (i) and (ii) require a different way of computing the parameters. Other parameters related to the standard deviation (s_2, s_3, etc.) that give rise to more realistic peak fittings were not used in this work.

2.3.1.1. Case (i).

Table 1 shows the calculations needed to obtain the upper and lower boundaries for each parameter in case of moderate overlap. Figs. 1–3 indicate the values of t_R, t_1, t_2, h_1 and h_2 used to build the initial parameter guesses and the lower and upper boundaries. Thus, h_1 is the value of the input signal at $t=t_R$ (i.e. the maximal peak height) (Fig. 1a). The remaining parameters are computed from the second derivative as follows (Fig. 1c): t_R is calculated by finding the minimum of the derivative, t_1 and t_2 are the times where the second derivative is zero, and h_2 is the value of the second derivative at $t=t_R$.

The computation of the s_0 and h_0 parameters is not straightforward. For a Gaussian peak, it can be deduced that:

$$s_0=\sqrt{\left\|\frac{h_1}{h_2}\right\|} \quad (6)$$

Also:

$$s_0=\left(\frac{t_2-t_1}{2}\right) \quad (7)$$

where t_1 and t_2 were defined above. It should be noted that the hypothesis of a Gaussian curve is not used in the eventual deconvolution step, but is useful in this concern because it allows deducing initial estimates for the parameters.

Table 1
Initial guesses, lower and upper boundaries for h_0, t_R, s_0 and s_1 in Eq. (5) model for case (i) (moderate overlap)

Parameter	Value[a]	Lower boundary[b,c,d]	Upper boundary[b,c,d]
Retention time (t_R)	t_R	t_1	t_2
Standard deviation (s_0)	$\frac{((t_2-t_1)/2)+\sqrt{\|\|h_1/h_2\|\|}}{2}$	$\min\left\{\begin{matrix} t_2-t_R \\ t_R-t_1 \\ \sqrt{\|\|h_1/h_2\|\|}\end{matrix}\right\}-\varepsilon$	$\max\left\{\begin{matrix} t_2-t_R \\ t_R-t_1 \\ \sqrt{\|\|h_1/h_2\|\|}\end{matrix}\right\}+\varepsilon$
Peak height (h_0)	$\frac{h_{0,\max}+h_{0,\min}}{2}$	$\min\left\{\begin{matrix} h_1-\mathrm{thr}_{h1} \\ \left(\frac{t_2-t_1}{2}\right)^2(h_2-\mathrm{thr}_{\mathrm{sd}})\end{matrix}\right\}$	$\max\left\{\begin{matrix} h_1+\mathrm{thr}_{h1} \\ \left(\frac{t_2-t_1}{2}\right)^2(h_2+\mathrm{thr}_{\mathrm{sd}})\end{matrix}\right\}$
Fronting/tailing term (s_1)[e]	$\frac{(t_2-t_R/t_1-t_R)-1}{(t_2-t_R/t_1-t_R)+1}$	–	–

The meaning of h_1, t_R, t_1, t_2 and h_2 is given in Fig. 1.

[a] $h_{0,\max}$ and $h_{0,\min}$ are the upper and lower boundaries for peak height.

[b] The ε value is given by: $\varepsilon=\varepsilon_{\mathrm{sd}}/2\sqrt{(\partial^3h/\partial t^3|_{t=t_1})^{-2}+(\partial^3h/\partial t^3|_{t=t_2})^{-2}}$ where $\varepsilon_{\mathrm{sd}}$ is the noise in the second derivative, and $\partial^3h/\partial t^3$ the third derivative of the signal.

[c] $\mathrm{thr}_{\mathrm{h1}}$ is the threshold in peak height defined in Section 2.2.

[d] $\mathrm{thr}_{\mathrm{sd}}$ is the threshold in the second derivative defined in Section 2.2.

[e] Lower and upper boundaries of s_1 are not computed.

Eqs. (6) and (7) are both taken into account in the computation of s_0 and its boundaries (Table 1). This assures that a good estimate of the value of the standard deviation is obtained. In case of coelution, Eq. (7) tends to underestimate the true value of s_0. This bias is corrected by using Eq. (6), since it is less sensitive to deviations in s_0 introduced by overlapping peaks.

The error ε in s_0 is deduced by applying error propagation theory to Eq. (7):

$$\varepsilon=\sqrt{\left(\frac{\partial s_0}{\partial t_1}\varepsilon_{t1}\right)^2+\left(\frac{\partial s_0}{\partial t_2}\varepsilon_{t2}\right)^2} \tag{8}$$

where ε_{t1} and ε_{t2} are the errors (measured as standard deviation) associated to the determination of t_1 and t_2, and the derivatives of s_0 with respect to t_1 and t_2 are calculated from Eq. (7). Taking into account that t_1 and t_2 are obtained from the second derivative, and applying again error propagation theory, ε_{t1} can be approximated to:

$$\varepsilon_{t1}=\left(\left.\frac{\partial^3 h}{\partial t^3}\right|_{t=t1}\right)^{-1}\varepsilon_{\mathrm{sd}} \tag{9}$$

where $\varepsilon_{\mathrm{sd}}$ is the noise in the second derivative, computed as explained in Section 2.1, and $(\partial^3h/\partial t^3)|_{t=t1}$ is the third derivative evaluated at $t=t_1$. A similar definition is found for $t=t_2$. By combining Eqs. (7)–(9), the expression of ε in Table 1 is obtained.

The boundaries of the h_0 parameter are also computed by taking into account Eq. (6), expressing h_0 as a function of h_2 and s_0, and replacing the latter by its value, according to Eq. (7):

$$h_0=\left(\frac{t_2-t_1}{2}\right)^2||h_2|| \tag{10}$$

Note that this expression, which appears in the computation of h_0 limits, is based only on the second derivative. This corrects the overestimation of the peak height measured directly from the input signal (h_1) in cases of moderate coelution. In such cases, the value of h_1 is greater than expected because the tail or front of the interferents make the whole signal higher at $t=t_R$.

2.3.1.2. Case (ii). Table 2 gives the initial values and lower and upper boundaries for each parameter when the solutes coelute strongly. The main modification is the inclusion of the approximation $(t_2-t_1)/2n_4$, which implies that all the underlying compounds contribute equally to the standard deviation.

The computation of s_0 requires the use of other parameters: t'_1 and t'_2, which depends on the classification of the coelution case as (ii)-a or (ii)–b. Fig. 4c (case (ii)-a) includes the values of t'_1 and t'_2 for toluene, which are those times where the second derivative is maximal at each side of $t=t_R$ within the corresponding "Zone−". In (ii)-a, the value of the second derivative at $t=t'_1$ and $t=t'_2$ should be negative. If this condition does not hold, the value of t_1 or t_2 is substituted by t'_1 or t'_2. In (ii)–b, the latter condition is not applied (see the values of t'_1 and t'_2 for ethylbenzene in Fig. 5c), since the peak is found in a "Zone+" of the second derivative.

2.3.2. Correction of the peak height

Once computed the initial estimates of the peak parameters according to Tables 1 and 2, a better h_0 estimate is obtained for all the peaks through linear regression. The predicted chromatogram is built using the peak model (Eq. (5)) with the initial parameters obtained according to Section 2.3.1. The linear regression can be written as follows [16]:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{11}$$

Table 2
Initial guesses, lower and upper boundaries for h_0, t_R, s_0 and s_1 in Eq. (5) model for case (ii) (strong overlap)

Parameter	Value[a]	Lower boundary[b,c,d,e]	Upper boundary[b,c,d,e]
Retention time (t_R)	t_R	t'_1	t'_2
Standard deviation (s_0)	$\frac{(t'_2 - t'_1)/2 + \sqrt{\Vert h_1/h_2 \Vert} + (t_2 - t_1)/2n_4}{3}$	$\min \left\{ \begin{array}{c} t'_2 - t_R \\ t_R - t'_1 \\ \sqrt{\Vert h_1/h_2 \Vert} \\ \frac{t_2 - t_1}{2n_4} \end{array} \right\} - \varepsilon$	$\max \left\{ \begin{array}{c} t'_2 - t_R \\ t_R - t'_1 \\ \sqrt{\Vert h_1/h_2 \Vert} \\ \frac{t_2 - t_1}{2n_4} \end{array} \right\} + \varepsilon$
Peak height (h_0)	$\frac{h_{0,\max} + h_{0,\min}}{2}$	$\min \left\{ \begin{array}{l} h_1 - thr_{h1} \\ thr_{h1} \end{array} \right\}$	thr_{h1}
Fronting/tailing term (s_1)[f]	0	–	–

The meaning of h_1, t_R, t_1 and t_2 is given in Fig. 1, and an example of the definition of t'_1 and t'_2 values for toluene is plotted in Fig. 3.

[a] $h_{0,\max}$ and $h_{0,\min}$ are the upper and lower boundaries for peak height.

[b] The ε value is given by: $\varepsilon = \varepsilon_{sd}/2\sqrt{(\partial^3 h/\partial t^3|_{t=t_1})^{-2} + (\partial^3 h/\partial t^3|_{t=t_2})^{-2}}$ where ε_{sd} is the noise in the second derivative, and $\partial^3 h/\partial t^3$, the third derivative.

[c] thr_{h1} is the threshold in peak height defined in Section 2.2.

[d] thr_{sd} is the threshold in the second derivative defined in Section 2.2.

[e] n_4 is given by: $n_4 = ((n_c - 1)/2) + 1$, where n_c is the number of changes in sign of the third derivative in those time ranges where the second derivative is negative.

[f] Lower and upper boundaries of s_1 are not computed.

where **y** is the column vector containing the experimental signal (i.e. the input chromatogram), **X** is a matrix whose columns describe the contribution of the signal due to the individual peaks (built with Eq. (5), using the peak parameters of Table 1 or 2), **β** is a column vector containing the regression coefficients, and **ε** stores the residuals. The regression coefficients are obtained as follows:

$$\boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad (12)$$

where $\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}^{-1}$ denote the transpose and inverse **X**, respectively. Since h_0 is a factor that multiplies each individual peak profile, and since this is also what **β** does, it can be written as:

$$\boldsymbol{\eta} = \mathbf{h}\boldsymbol{\beta}^{\mathrm{T}} \qquad (13)$$

where **h** is the column vector with the initial h_0 values found in Section 2.3.1.1. The values contained in **η** are used as updated initial h_0 guesses for the deconvolution.

The regression step removes those peaks that are not found significant to explain the whole chromatogram. Peaks for which the updated value of h_0 is negative are removed, and the procedure is performed again with the corrected number of peaks. This step is performed until all the diagonal elements of **η** are positive.

2.4. Splitting the chromatograms

A chromatogram usually exhibits peaks or peak clusters that can be isolated from the rest of the signal. This fact allows splitting the chromatogram in convenient smaller blocks, which can be processed independently. This speeds up the computation and increases the accuracy, since the deconvolution is performed only in those time domains, where a change in a parameter has a significant impact on the value of the sum of squared residuals.

The proposed algorithm includes a routine able to detect these elution zones without user interaction. To perform this, the depth of the points comprised between two consecutive peaks was monitored as the ratio between the valley point height to the interpolated peak height. When this ratio becomes smaller than 0.01, the chromatogram is divided in two blocks. Also, large baseline regions are discarded to speed up the computation.

3. Experimental

A high-performance liquid-chromatographic system, equipped with an L–7100 pump, L–7612 solvent degasser, L–7250 autosampler, L–7400 UV detector and a D–7000 interface from Merck-Hitachi (Tokyo, Japan) was used for the study. The detection wavelength was 254 nm, and the sampling frequency was kept to 600 points/min, in order to get at least more than 20 points per peak for the faster compounds, that were eluted within 0.04 min. The injection volume and flow-rate were 5 μl and 9 ml/min, respectively. The column was submerged in a water bath whose temperature was kept constant at 30 °C with a Protherm pt 5000 thermostat.

A monolithic SpeedROD RP–18e (50 mm × 4.6 mm) HPLC column from Merck (Darmstadt, Germany) was used. The test mixtures contained amylbenzene (Amy) (Sigma-Aldrich, Steinheim, Germany), butylbenzene (But), ethylbenzene (Eth), *o*-terphenyl (Tph), triphenylene, (Trp) (Fluka, Buchs, Switzerland), and toluene (Tol) (Merck), in methanol/water (80/20 m/m). Mobile phases were prepared using methanol from Hipersolv for HPLC (BDH Laboratory

Supplies, Poole, England), and ultrapure water, obtained with the Milli-Q water purification system (Millipore, Molsheim, France).

The HPLC system was operated with the LaChrom D–7000 HPLC Manager Software (Merck-Hitachi). The computation was carried out with a Pentium IV/2400 MHz computer. Home-made routines were written in Matlab 6.5 (Natick, MA, USA).

4. Results and discussion

4.1. Selection of the proper window size in SG filtering

Fig. 6 depicts an example of the application of the DW test to select the most adequate window size. A peak cluster containing two highly overlapped compounds was generated using Eq. (5), and including blank noise of 0.01 standard

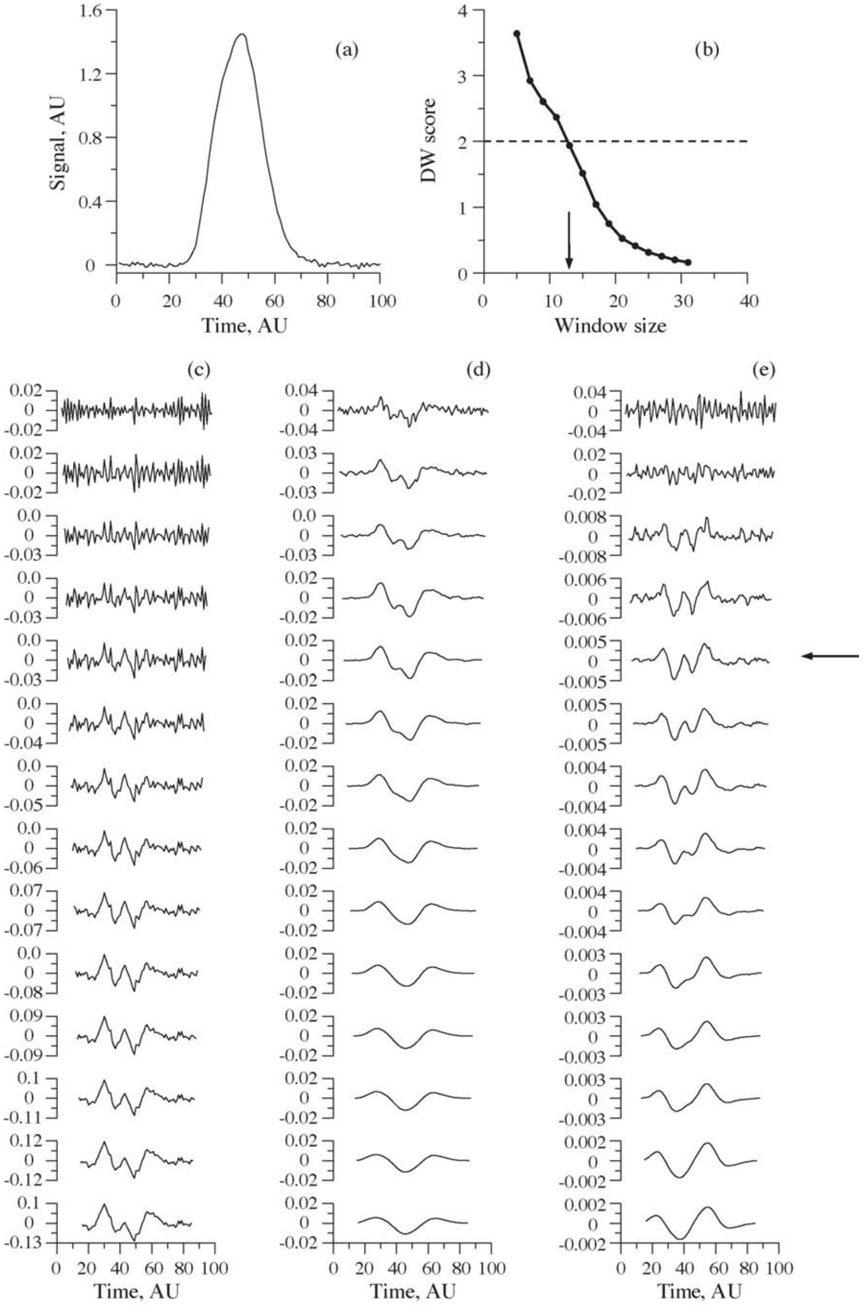

Fig. 6. Performance of the DW test as a criterion to select the adequate window size in SG smoothing. A chromatogram (a) with two underlying peaks was smoothed with SG, using a third-degree polynomial and several window sizes. Residuals (c), second- (d) and third-order derivatives (e), are also given. Each row of c, d and e corresponds to a different window size between 5 and 31 (from top to bottom, increasing the number of points symmetrically in both extremes of the window in one unit each time). The DW statistic of the residuals is plotted in part b as a function of different window sizes. The critical value of DW = 2 is depicted in (b) as a dashed line. Arrows point the optimal window size.

deviation units. The chromatogram is depicted in Fig. 6a. A third-order polynomial was used to smooth the signal by SG, considering windows from 5 to 31 points. As usual, only odd numbers of points, giving rise to symmetrical windows, were considered. From the fitted polynomial parameters in each point, the first-, second- and third-order derivatives were computed. Fig. 6b shows the DW scores plotted versus the window size according to Eq. (2). The residuals, second- and third-order derivatives are plotted in Figs. 6c–e, respectively, for different window sizes. Each row in the plot corresponds to a different value in the *X*-axis of Fig. 6b, starting from a window size of 5 (top) to 31 (bottom).

As can be seen, the selection of the proper window size is critical. Windows with too few points (first rows of columns c–e in Fig. 6) yield noisy derivatives, and no peak was reliably detected under these conditions. On the other hand, too large windows (last rows) yield highly correlated residuals, which retain only partially the information about the derivatives. As a consequence, the second and third derivatives lack of the expected details (columns d and e in Fig. 6), and the chromatographic situation is erroneously assessed as a single-peak case. In these circumstances, only one negative domain was detected in the second derivative, and no change in sign of the third derivative was found in this range. The right window size as derived from the DW criterion yielding a value closest to 2.0, is depicted by an arrow in Fig. 6b. This corresponds to a 13-point window size. With this window, the peak cluster is correctly assessed, since a clear negative zone and two clear minima were found in the second derivative plot, and a sign inversion in the third derivative is detected within this time range.

4.2. Self-adapting window size

When a chromatogram comprises large variations in peak width, a single window size for the SG smoothing cannot remove the noise in the whole chromatogram without distorting the peaks. This problem can be overcome by applying independently the DW criterion to each elution region.

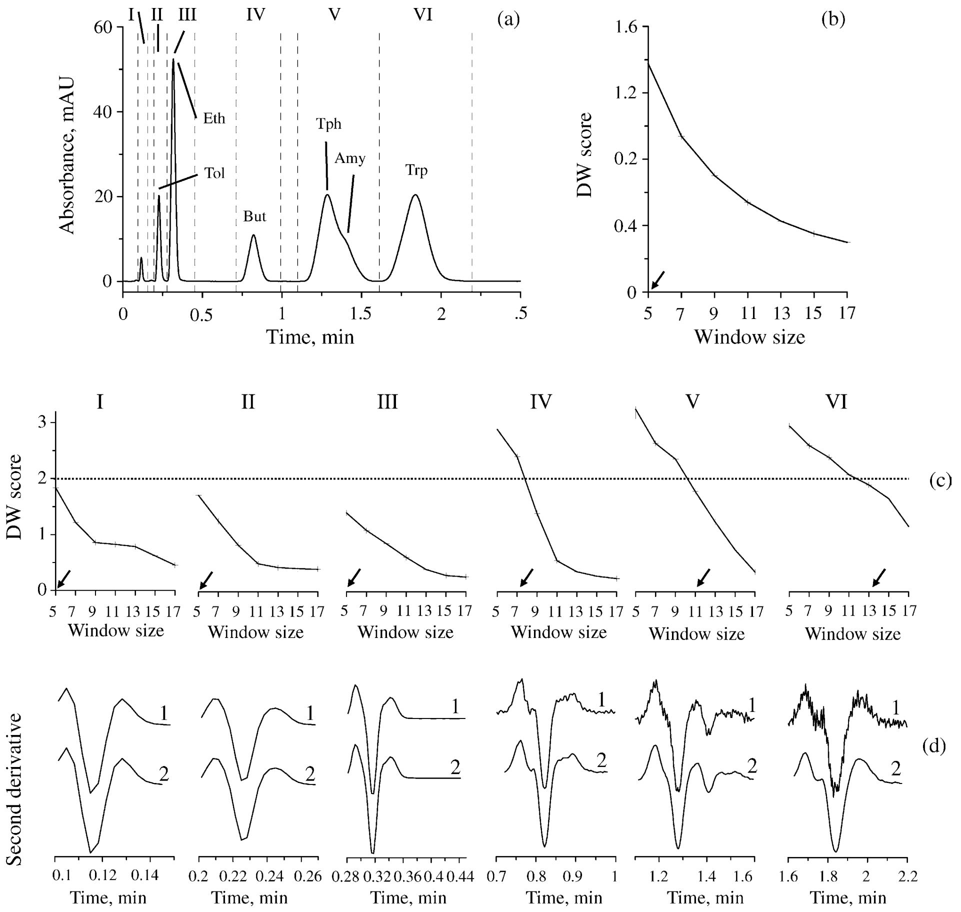

Fig. 7. Application of the self-adaptive window-size selection using the DW criterion to the chromatogram (a) of a mixture of six aromatic compounds. Mobile phase: 60% methanol. Six different elution regions (roman numerals) are separated by dashed lines (the limits were set according to Section 2.4). The injection peak appears in region I. The DW statistic computed with several window sizes over the whole signal is plotted in (b). The same test applied at each particular elution region is depicted in (c)—roman numerals corresponding to each elution region are included above for clarity. Arrows indicate the optimal window size. The second derivatives obtained from the SG smoothing are depicted in (d). Label "1" indicates that the same window size was used for the whole chromatogram, whereas label "2" indicates that the DW test was applied independently in each particular elution region.

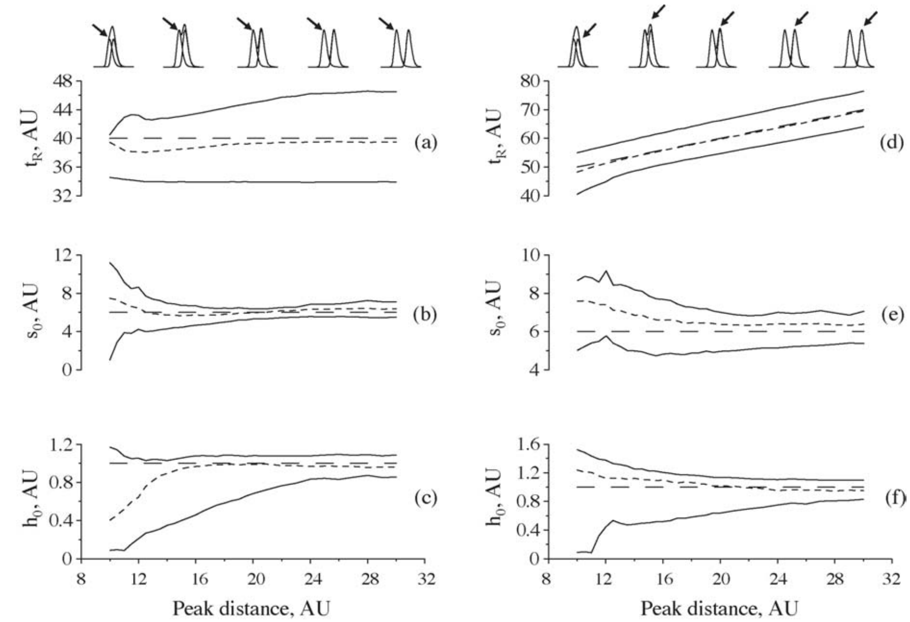

Fig. 8. Peak parameter estimation in a two-peak deconvolution problem with different separation degrees. Upper and lower boundaries (solid line), initial guesses (short dashed line) and true value (long dashed line) are plotted for t_R (a, d), s_0 (b, e) and h_0 (c, f) vs. peak distance (X-axis). The different chromatographic situations were simulated by adding the signal of two peaks. Plots a–c correspond to the first, and d–e to the last eluting peak (pointed with an arrow). Some chromatograms representing several selected situations are shown on the top.

The performance of this approach is illustrated in Fig. 7, where the results for a mixture of six aromatic compounds injected in 60% methanol (Fig. 7a) are plotted. The dashed lines represent the different elution regions (roman numerals), at which the chromatogram was split, according to Section 2.4. The DW scores computed using the whole chromatogram versus the window size are plotted in Fig. 7b. The optimal window, yielding the DW score closest to 2.0, includes five points. The second derivative of the chromatogram using this window size is plotted in Fig. 7d (line 1), for each of the six elution regions. As can be seen, the five-point window size yields good results for regions I–III, but too noisy derivatives are obtained for regions IV–VI. This can lead to a wrong identification of a peak eluting within these regions, since the noise will be misinterpreted as the elution of non-existing compounds. This effect is a consequence of the strong differences in peak width in the different elution regions.

The sharpness of the first three peaks requires small window sizes because larger windows would distort the peak shape. The wider peaks at higher retention times require a larger window for efficient noise removal. This is depicted in Fig. 7c, where the DW scores versus the window size are plotted for each peak of the I–VI regions. The optimal window size is different in each region and evolves from 5 (for regions I–III) to 7, 11 and 13 (for regions IV, V and VI). Fig. 7d (line 2) shows the second derivative obtained with the SG smoothing using the optimally-adapted window sizes. In comparison to the results obtained using the same optimal window for the whole chromatogram (Fig. 7d, line 1), the improvement in signal-to-noise ratio for the last eluting bands is evident. The larger window sizes for regions IV–VI allow removing the noise properly.

4.3. Peak parameter assessment

Different situations of coelution were simulated, in order to test the adequacy of the selected initial parameter estimates and the respective boundaries explained in Section 2.3. The chromatograms were generated by adding the individual signals of two peaks using Eq. (5). The parameters were: $t_R = 40$, $h_0 = 1$, $s_0 = 6$, and $s_1 = 0.1$ for the preceding peak, and $t_R = 50$ to 70, $h_0 = 1$, $s_0 = 6$, and $s_1 = 0.1$ for the following peak. Blank noise with 0.01 standard deviation units was added to the chromatograms. A 10-fold experiment, using a different seed for generating the noise, was produced within each situation. Then, the mean value of the initial guesses, and the upper and lower boundaries, were computed. Fig. 8 depicts the initial guesses and the boundaries for the parameters obtained by applying the peak detection algorithm explained in this work. As can be seen, the initial guesses of t_R, s_0 and h_0 are generally close to the true values. This is slightly more evident when the distance amidst peaks is high, since the evaluation of the peak parameters yields less bias. The true values are always within the boundaries and these are narrower as peak overlap decreases.

5. Conclusions

The first step in the deconvolution of chromatographic signals is the detection of peaks. This is particularly necessary when peak overlapping is detected as slight shoulders in a peak cluster. In these situations, peak detection algorithms that use only the first derivative are not powerful enough to evaluate properly the number of peaks. Only the second- (and in some cases, the third-) order derivatives provide the necessary information.

A smoothing technique is also needed when derivatives are calculated. The SG method is adequate for this purpose, allowing the computation of the smoothed signal and the derivative in a single step. Two critical parameters must be selected in SG smoothing, namely the window size and the polynomial degree. The window size is particularly critical, since the smoother should remove the noise, preserving nevertheless the chemical information. If this is not fulfilled, the number of peaks can be incorrectly inferred.

The DW criterion was demonstrated a valuable test to optimise the window size in SG smoothing. In some cases, the different elution regions of a chromatogram require a different window size and smoothing technique. The application of the DW test to each region allows the selection of a locally-adapted set of SG parameters. Initial estimates of the peak parameters and their ranges can be also obtained from the study of the high-order derivatives of the signal. The application of the error propagation theory and the study of several peak parameters were shown to yield the necessary information to establish the initial guesses and the peak boundaries of the parameters in different situations of overlap.

Acknowledgements

This work was supported by Project CTQ2004–02760/BQU (Ministerio de Educación y Ciencia of Spain) and Groups Grant 04/16 (Generalitat Valenciana). JRTL thanks the MCYT for a Ramón y Cajal position. GVT thanks the Generalitat Valenciana for providing funds to perform a predoctoral stay and for an FPI grant.

References

[1] J.H. Jiang, Y. Ozaki, Appl. Spectrosc. Rev. 37 (2002) 321.
[2] A.G. Stromberg, S.V. Romanenko, E.S. Romanenko, J. Anal. Chem. 55 (2000) 615.
[3] P. Nikitas, A. Pappa-Louisi, A. Papageorgiou, J. Chromatogr. A 912 (2001) 13.
[4] K. Lan, W. Jorgenson, J. Chromatogr. A 915 (2001) 1.
[5] T.L. Pap, Zs. Papai, J. Chromatogr. A 930 (2001) 53.
[6] V. Di Marco, G.G. Bombi, J. Chromatogr. A 931 (2001) 1.
[7] J. Li, J. Chromatogr. A 952 (2002) 63.
[8] G. Vivó-Truyols, J.R. Torres-Lapasió, R.D. Caballero-Farabello, M.C. García-Alvarez-Coque, J. Chromatogr. A 958 (2002) 35.
[9] J.L. Excoffier, G. Guiochon, Chromatographia 15 (1982) 543.
[10] E. Ziegler, G. Schomberg, Anal. Chim. Acta 147 (1983) 91.
[11] P.J.P. Cardot, P. Trolliard, S. Tembely, J. Pharm. Biomed. Anal. 8 (1990) 755.
[12] A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627.
[13] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics (Part B), Elsevier, Amsterdam, 1998.
[14] E.J. Karjalainen, U.P. Karjalainen, Data Analysis for Hyphenated Techniques, Elsevier, Amsterdam, 1996.
[15] J. Durbin, G.S. Watson, Biometrika 37 (1950) 409.
[16] N.R. Draper, Applied Regression Analysis, Wiley, New York, 1998.
[17] D.N. Rutledge, A.S. Barros, Anal. Chim. Acta 454 (2002) 277.
[18] B.G.M. Vandeginste, L. de Galan, Anal. Chem. 47 (1975) 2124.
[19] G. Vivó-Truyols, J.R. Torres-Lapasió, A. Garrido-Frenich, M.C. García-Alvarez-Coque, Chemom. Intell. Lab. Syst. 59 (2001) 107.
[20] J.R. Torres-Lapasió, J.J. Baeza-Baeza, M.C. García-Alvarez-Coque, Anal. Chem. 69 (1997) 3822.