Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1096 (2005) 146–155
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

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

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 13 April 2005

Abstract

Several interlinked algorithms for peak deconvolution by non-linear regression are presented. These procedures, together with the peak detection methods outlined in Part I, have allowed the implementation of an automatic method able to process multi-overlapped signals, requiring little user interaction. A criterion based on the evaluation of the multivariate selectivity of the chromatographic signal is used to auto-select the most efficient deconvolution procedure for each chromatographic situation. In this way, non-optimal local solutions are avoided in cases of high overlap, and short computation times are obtained in situations of high resolution. A new algorithm, fitting both the original signal and the second derivatives is proved to avoid local optima in intermediate coelution situations. This allows achieving the global optimum without the need of background knowledge by the user. A previously reported peak model, a Gaussian with a polynomial standard deviation whose complexity can be modulated to enhance the fitting quality, was applied. However, the original formulation was modified to account baseline outside the peak region. Also, the optimal model complexity was auto-selected via error propagation theory. The method is able to process simultaneously several related chromatograms. The software was tested with both simulated and experimental chromatograms obtained with monolithic silica columns.

Keywords: Peak deconvolution; Chromatographic one-way signals; Monolithic silica; Genetic algorithms

1. Introduction

Deconvolution is a powerful mathematical tool for enhancing the selectivity offered by chemical methods. An important application is the separation of a complex chromatographic signal in its individual contributions, when partial coelution is obtained due to an insufficient separation power of the chromatographic system. As a result, compounds hidden within a peak cluster can be quantified with relatively small errors. However, these chemometric tools usually require specialised knowledge, which makes its application in routine analysis by non-expert users difficult. In this work, a set of methods is presented that automate the deconvolution process in order to facilitate the routine application of this technique.

Part I [1] was focused on the pre-treatment and analysis of the data before deconvolution. An algorithm to scan a signal and provide the number of underlying peaks, as well as estimates of the peak parameters, was developed, based on the study of the original signal and the first-, second-, and third-order derivatives. The noise of the chromatogram was removed using the Savitsky-Golay (SG) technique, and the Durbin-Watson criterion was applied to establish the optimal window size to minimise distortions by the smoothing algorithm. Also, an automatic method for peak identification was presented for the comparison of signals of the same compound injected in different samples.

This part describes different methods of deconvolution, and the tools for making the on-line decisions about which

* 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.2005.03.072

algorithm and peak model should be applied. Most parameters are obtained by non-linear regression [2]. A good collection of classical non-linear regression algorithms can be found in the literature [3]. However, most of them present the disadvantage of being local: they tend to find not the global but the closest solution to the selected candidate, not necessarily the true global optimum. Global methods are advisable to account these situations [4], since they explore the whole parameter search space instead of only the neighbourhood of a single candidate solution. Some examples of global techniques are genetic algorithms [5] and simulated annealing [6]. A hybrid method, called "locally optimised genetic algorithm" (LOGA), which combines the advantages of both the global and the local search methods, was developed recently [7]. It was demonstrated to be especially useful for the deconvolution of strong overlapping situations.

In this work, four different deconvolution algorithms with different capabilities were applied, one of them being the LOGA method. The tendency to converge into local solutions depends on the chromatographic situation. Chromatograms with slight peak overlap can usually be solved well with classical local methods, but situations with strong coelution require the selection of more powerful algorithms in order to avoid being trapped into local solutions. However, global methods are more time-consuming and, therefore, should be applied only when strictly necessary. The decision on the most appropriate deconvolution algorithm, in order to balance the difficulty of the deconvolution and computation time, is not evident and requires a quantitative estimation. This work proposes a tool for the automatic selection of the adequate deconvolution algorithm adapted to each chromatographic situation. This evaluation is based on multivariate figures of merit [8].

From a mathematical standpoint, deconvolution consists of fitting the chromatographic signal to a combination of individual peaks, each of them described with a particular peak model. Several models can be found in the literature [9], and for some of them the adaptation of their complexity to that of the experimentally obtained peak profile is possible. The selection of the proper model constitutes another problem to tackle for non-experienced users, and for this reason, this task was also automated in this work. To achieve this, a method is proposed to determine the statistical significance of each peak parameter during the deconvolution process.

To study the performance of the methods, they were first tested under controlled conditions with simulated data. In a further step, they were applied to real experimental data.

2. Theory

2.1. Peak model

Deconvolution consists of fitting an experimental chromatogram (or a set of them) to a linear combination of individual chromatographic peaks. Hence, a mathematical peak model is needed to describe each elementary peak. In this work, the selected model was the polynomially modified Gaussian (PMG) [10]:

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

where h_0 is the maximal peak height, $h(t)$ the height at time t, t_R the solute retention time, and s_0 and higher order terms are the standard deviation and distorting parameters, respectively. Eq. (1) presents the advantage of being able to describe both tailing and fronting peaks. Note that this equation represents actually a family of models, since according to the polynomial degree within the standard deviation term, several types of function may arise. The higher the degree, the more flexible the model and the better the achieved fitting to the experimental data. Theoretically, there is no limit to the polynomial degree, but in practice, parabolic or cubic functions are usually enough to describe most chromatographic signals without under- or over-fitting. In this work, the polynomial degree was selected, according to the procedure outlined in Section 2.6.

Eq. (1) has a drawback: the peak does not decay rapid enough so that the baseline tends to grow far from the peak, which is especially troublesome in situations where the individual peaks involve long baselines between peaks. These risings should not be neglected in situations such as when the minimal height is significant owing to strong asymmetry ($B/A > 2.5$). This problem was initially tackled by setting the height at each side of the peak region to the respective minimal value (Fig. 1) [11]. It should be mentioned that the parameter set on the simulation depicted in this figure was deliberately unreal to get strong raisings.

The outlined solution works when a few peaks (e.g. 2 or 3) overlap, but may lead to biased deconvolutions in some instances, especially when the accumulation of constant negligible values is translated in significant growths of the baseline owing to the large number of peaks within the cluster.

This problem has been tackled in the literature in several ways [12,13]. We solved it by cropping the PMG model and substituting each outer region with an exponential decay in a given point so that the derivatives of both PMG and exponential functions coincide. This was done at both sides of the peak. The method was applied to improve the simulations of highly distorted electrophoretic signals by setting the connection at 10% peak height, since efficiency and asymmetry data involved in simulations were taken at this peak ratio. The procedure was, however, not reported in that work, but just applied [14]. For the current work, the method was improved, making the retention time where the truncation takes place depending on the peak shape. The resulting set of equations

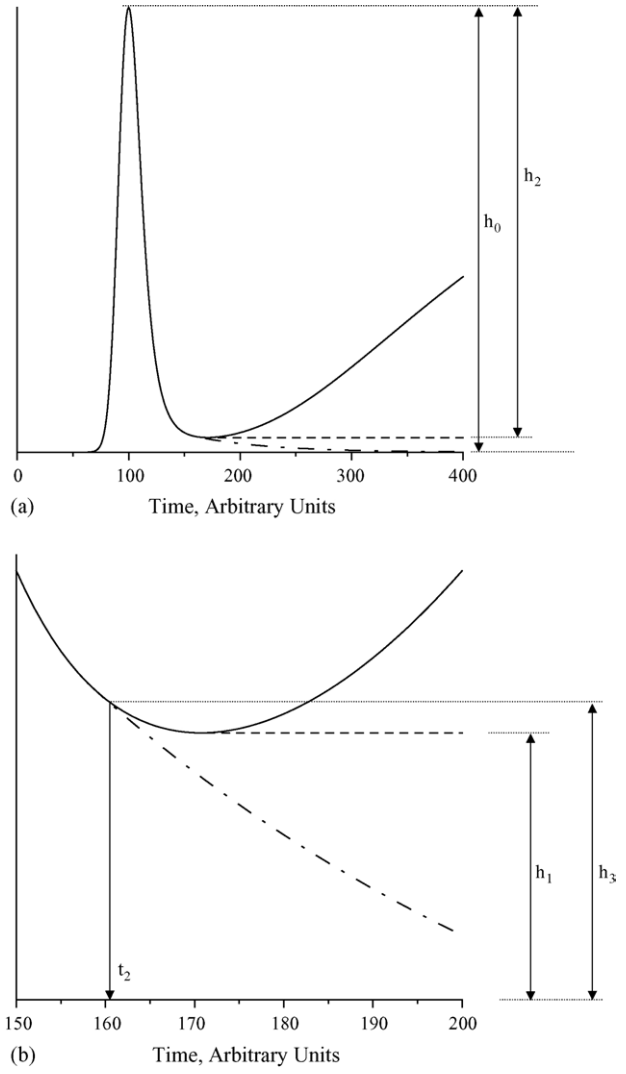

Fig. 1. Different solutions for the problem of the abnormal raisings in the original formulation of the PMG model: unaltered (solid line), setting constant values at peak distances far from both minima (dashed line), and substituting it by an exponential function for $t > t_3$ (dot-dash). See the text for the computation of h_0, h_1, h_2, h_3 and t_2. The zone of peak truncation is enlarged in (b). The simulation was done using a forced parameter set to emphasise the raisings.

is formally:

$$h(t) = \begin{cases} a_1 \ \exp[b_1(t - t_R)] & \text{if } t < t_1 \\ h_0 \ \exp\left[-\frac{1}{2}\left(\frac{t - t_R}{s_0 + s_1(t - t_R) + s_2(t - t_R)^2 \ldots}\right)^2\right] & \text{if } t_1 \le t \le t_2 \\ a_2 \ \exp[b_2(t - t_R)] & \text{if } t > t_2 \end{cases} \quad (2)$$

How to determine t_2 is shown in Fig. 1. It is calculated as the time in which the PMG function reaches the h_3 value:

$$h_3 = h_1\left(\frac{F}{h_0}h_2 + 1\right) = h_1\left(F\left(1 - \frac{h_1}{h_0}\right) + 1\right) \quad (3)$$

where h_1 is the height at the minimum of the right-side of the peak, and $h_2 = h_0 - h_1$. An identical relationship holds for t_1 (left side of the peak). Eq. (3) makes h_3 tending to zero when h_1 does. This way, when the minimum is located very close to the baseline (Fig. 1b), the exponential decay truncates the model closer to this minimum. When the minimum is not detected, the exponential function is just not used. The factor F ($0 < F \le 1$) modulates the importance of the exponential part: the higher this value, the closer the t_2 value to the retention time and the higher the importance of the exponential part. A value of $F = 0.1$ was found appropriate, and used throughout this work.

The parameters of the exponential functions (a_1, b_1, a_2 and b_2 in Eq. (2)) are computed making the values of $h(t)$ and $\partial h(t)/\partial t$ of both exponential and PMG functions equal at $h = h_3$. After solving the two-equation system at each side of the peak region, the exponential function parameters (b_i and a_i) are found to be:

$$b_i = \frac{j^2}{t_i - t_R} \times \left[-1 + j(s_1 + 2s_2(t_i - t_R) + 3s_3(t_i - t_R)^2 + \ldots)\right] \quad (4)$$

where

$$j = \pm\sqrt{-2 \ln\left(\frac{h_3}{h_0}\right)} \quad (5)$$

and

$$a_i = h_3 \exp[-b_i(t_i - t_R)] \quad (6)$$

In Eq. (5), the minus sign is used for $i = 1$ (left side of the peak) and the plus when $i = 2$ (right side of the peak). The composite peak function will be called PEMG (polynomial-exponential modified Gaussian). From this point on, the acronym PEMG0 will denote a Gaussian function, PEMG1 will include a linear standard deviation in Eq. (1), PEMG2, a parabolic standard deviation, etc.

2.2. Deconvolution algorithms

The fitting of chromatographic profiles implies a non-linear regression. This means that the model parameters are estimated iteratively by least-squares. Several methods can be found in the literature to solve this kind of problem [3], but the risk of finding a solution not corresponding to the global least-squares minimum is always present. The lower the peak separation, the higher the probability of finding a local solution. For this reason, two families of algorithms were implemented, one of them suitable for easy deconvolutions,

and the other for more complex problems. How to select the proper algorithm is explained in Section 2.3.

Algorithms that contain a random part in their architecture have more chances of finding global optima. Examples are genetic algorithms [6] (GAs) or multi-start local search (MSLS) [15]. A disadvantage is that the user has to set some configuration parameters of the algorithm, which makes the automation difficult. The performance of these techniques deteriorates dramatically if inadequate parameters are used. For this reason, these methods require trained users able to set properly the fitting parameters.

Algorithms without random part (like the Gauss–Newton [2] or Powell [3] methods), require less user knowledge, but are less efficient. An original algorithm called the Powell-2 method is presented in this work, which solves the problem of local optima without the requirement of a random part. The algorithms can be summarised as follows:

(i) LOGA [7], which includes a random part and is appropriate for complex problems.
(ii) Powell-2, without a random part and appropriate for complex problems.
(iii) MSLS [15], with random part, appropriate for simple problems.
(iv) Powell-1 [3], without random part, appropriate for simple problems.

2.2.1. The Powell-1 algorithm

This algorithm is the well-known Powell method [3] for non-linear regression. It is the simplest and also the fastest algorithm among the four considered in this work, and is used for simple deconvolution problems. It will be called here "Powell-1" to distinguish it from the "Powell-2" algorithm, which is explained below.

2.2.2. The Powell-2 algorithm

It consists of a modification of the Powell algorithm, and is useful for more complex problems. The modified Powell fits the model parameters so that two objective functions are minimised, namely the agreement between fitted and experimental chromatograms, and the agreement between second derivatives of fitted and experimental chromatograms. Both objective functions are alternated throughout the regression.

The algorithm steps are as follows:

(i) Compute the second derivative of the experimental chromatogram, according to Ref. [1].
(ii) Obtain the fitted chromatogram by applying the Powell-1 method [3] to the second derivatives, by using a predefined number of iterations. In each iteration, the predicted signal is computed by applying Eq. (2) with the current set of parameters, and then applying SG smoothing. The SG parameters used in this step are the same as applied in (i). A numerical evaluation of Eq. (2) is preferred to compensate the deviation produced by a possible peak distortion introduced by the smoothing technique. The reason is that identical distortions will be produced in steps (i) and (ii), resulting then on no net influence introduced in the residuals between experimental and predicted second derivatives.
(iii) With the refined parameters obtained from step (ii) as initial guesses, apply the Powell-1 method by fitting this time the original signal and performing the same number of iterations as in (ii).
(iv) Determine the residuals obtained in (iii). If a significant improvement is found, return to step (ii).
(v) Apply the Powell method as in step (iii), but with a larger number of iterations for fine-tuning the solution.

2.2.3. The multi-start local search algorithm

MSLS has been proposed as an alternative to genetic algorithms for simple problems [15]. It consists of performing repeatedly local searches starting from different initial solutions. The method starts with a set of random solutions, which fall within an initial range of parameters. The Powell-1 method with a low number of iterations is applied to each candidate solution. The best of these is then fine-tuned by using the Powell-1 method with a larger number of iterations.

2.2.4. The LOGA algorithm

Several algorithms (e.g. hybrid genetic algorithms [7,16] or immune algorithms [17]) have been proposed in the literature to tackle the problem of local convergence in the deconvolution of chromatographic signals. Genetic algorithms and related tools are promising to solve these problems, but they are difficult to automate, and in some cases could require prohibitive computation times. A hybrid method called LOGA, which includes a local search as a new genetic operation [7] was applied to chromatographic deconvolution problems, with good results [11].

2.3. Selection of the deconvolution algorithm: the multivariate selectivity

The user makes the decision on the type of algorithm (i.e. with or without random part), according to his/her experience, but the complexity of the algorithm (i.e. the choice between Powell-1 and Powell-2, or between MSLS and LOGA) is selected in a automatic way. Different peaks detected in the same chromatogram may present different levels of overlap, and therefore, a single chromatogram may require tools of different complexity.

The difficulty of a deconvolution problem is estimated through the study of multivariate figures of merit of the chromatographic data, which are used to quantify the complexity of an analytical problem [18]. In previous work, first-order multivariate selectivity was demonstrated to be able to predict the quality of the deconvolution [19]. It is used here as a criterion to select the proper deconvolution algorithm. A short account of the method is given below (more details are given in [19]).

First-order multivariate selectivity for an analyte s, in the presence of compounds a and b acting as interferents, is de-

fined as follows. The peak profiles of analytes *a*, *b* and *s*, each of them containing *t* time measurements, are first outlined as three column vectors (**a**, **b** and **s**) in a *t*-dimensional space. Let $\mathbf{W}_{-s}$ be a $t \times 2$ interferents matrix that contains the peak profiles **a** and **b**. The orthogonal projection $p(\mathbf{s})$ of **s** onto $\mathbf{W}_{-s}$ is:

$$p(\mathbf{s}) = [\mathbf{W}_{-s}\mathbf{W}_{-s}^{+}]\mathbf{s} \tag{7}$$

where W_{-s}^{+} denotes the Moore-Penrose generalised inverse of $\mathbf{W}_{-s}$. The multivariate first-order selectivity for solute *s* (SEL_s) is defined as follows:

$$SEL_s = \frac{||\mathbf{s} - p(\mathbf{s})||}{||\mathbf{s}||} \tag{8}$$

It was demonstrated that this value is an estimator of the problem complexity in the deconvolution of *s* in a signal constituted by a combination of *s*, *a* and *b* [19]. The multivariate selectivity is computed for all the peaks present in a chromatogram.

As described in Part I [1], the method splits each chromatogram to be deconvolved in several elution zones. Each of these zones is then treated independently from the others. When SEL_s for any of the peaks involved within an elution region falls below a certain threshold value, the deconvolution of this region is considered complex (see Section 4.1 for the selection of the threshold value). The algorithms detailed in Section 2.2 are then selected, according to the complexity of the elution regions.

The evaluation of SEL_s is performed before the deconvolution itself. Therefore, the initial estimates of the peak parameters in the deconvolution algorithms (see Part I [1]) are used to calculate the theoretical peak profiles of *s*, *a* and *b*. Accordingly, a more or less biased value of SEL_s will be obtained, since not the actual but only an approximation of the parameters is taken. Therefore, a higher threshold value is recommended to account possible under-estimations of the SEL_s value due to biased initial parameters.

2.4. Identifying compounds from different chromatograms

In previous work [7,11], the deconvolution errors were demonstrated to notably decrease when first-order chromatograms from different batches are treated altogether, which was called multi-batch deconvolution. Peak profiles of those compounds which are present in more than one injection are better retrieved, due to a decrease in the ambiguity of the mathematical solution. This strategy concerns not only the multi-batch treatment of related samples, but also the inclusion of injections of all or some of the standards of the target compounds.

In this work, this methodology is automated in order to get maximal benefits from multi-batch deconvolution for non-experienced users. The program processes more than one chromatogram simultaneously, and the peak detection algorithm is applied to all of them to obtain a rough estimation of the peak model parameters. Then, a peak-identification step follows, which computes how dissimilar are all possible peak pairs in order to identify if the same compound is present in more than one chromatogram. This measurement should be independent of the concentration of each peak, and able to compare peaks from different chromatograms, perhaps involving peaks acquired at different sampling frequencies. For this reason, the rough estimation of the model parameters (see Section 2.3 of Part I of this work [1]) was selected to build a modelled peak. This allows building synthetic chromatograms sharing a common time vector as independent variable, which solves the problem of different sampling frequencies. The dissimilarity ($Dis_{i,j}$) between two peaks *i* and *j* in two different chromatograms is measured as the sinus of the angle between both vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ arranged in columns containing the peak profile estimates:

$$Dis_{i,j} = \frac{||\mathbf{y}_i - (\mathbf{y}_j\mathbf{y}_j^{T}/||\mathbf{y}_j||^2)\mathbf{y}_i||}{||\mathbf{y}_i||} \tag{9}$$

where $\mathbf{y}_j^{T}$ is the transpose of $\mathbf{y}_j$. The dissimilarity coincides with the multivariate selectivity for first-order data in the way defined in Ref. [8], but with only one compound as interferent. Since the dissimilarity is a sinus, it varies between 0 (complete overlap) and 1 (peaks fully resolved). Those peak pairs belonging to different chromatograms whose dissimilarity falls below a certain threshold are assumed be originated by the same compound.

The threshold value of dissimilarity should be an assessment of the confidence of the chromatographer on the reliability and difficulty of his/her results: higher values mean that peak shapes are not reproducible (two peaks originated by the same compound can be rather dissimilar), whereas low values restricts the assessment of the same compound to highly similar peaks. A default value of 0.45 is recommended for non-problematic chromatograms.

In the next steps of the program, the parameters defining the peak shape of the same compound will be made equal, although the retention times and peak heights can vary from batch to batch. This is due to the assumption that the peak shape is maintained constant among injections—usually true in most chromatographic systems—, but the retention times are not strictly reproducible due to irregularities in the flow-rate or other sources.

2.5. Sorting out automatically the deconvolution process in multi-batch treatment

There are several situations where the multi-batch treatment outlined in the previous section can be applied. A straightforward example of this approach corresponds to the consideration of standard injections of individual compounds together with the chromatogram of a mixture. In this case, the peak shape of each compound is first retrieved by fitting the chromatogram of the standards. This information is then applied to deconvolve the chromatogram

of the mixture in a second step, forcing each peak shape to be equal to the previously found with the standards. This procedure cannot be classified, however, as a true multi-batch treatment. The reinforcement of information can be also achieved even when no standard injection is available, but in this case, chromatograms corresponding to related mixtures are required. This second case corresponds to a true multi-batch deconvolution, since all chromatograms are treated not sequentially but at the same time. Enhanced results are obtained with this strategy, since the parameters defining the peak shape are less ambiguous [7,11].

Strictly speaking, the difference between these two approaches relies on the availability or not of chromatograms including baseline-resolved peaks to be used as if were standards. If the same compound is baseline-resolved in one chromatogram but poorly separated in another, the deconvolution can be performed not simultaneously but sequentially. This speeds up the computation time without loosing precision, since the peak shape can be well established from the baseline-resolved peak. The multivariate selectivity, computed as described in Section 2.3, was used here as an estimator of the resolution. The same threshold used to select the type of algorithm was adopted in this case to decide whether the chromatograms should be deconvolved at a time, or on the contrary, if a sequential treatment will report more benefits.

2.6. Selection of the polynomial degree in the PEMG model

The use of Eq. (2) as a model to deconvolve chromatographic peaks obliges to an appropriate selection of the polynomial degree (i.e. PEMG1, PEMG2, etc.). This should be addressed with care: whereas a too simple polynomial can introduce systematic errors (underfitting), a too complex one may lead to overfitting (two replicates with slightly different noise can lead to completely different solutions). In this case, the analysis of residuals is not a valid tool to assess the proper polynomial degree. The reason is that the predicted signal will fit always better the experimental part when more complex polynomials are selected, and the highest—although not necessarily correct— polynomial degree will be systematically chosen as the best.

An analysis of the uncertainty of the s_i parameters—the values describing the standard deviation in Eq. (2)—was performed to establish the proper model complexity. The 95% confidence interval of each s_i parameter was calculated, according to Ref. [2].

A parameter is considered non-significant when its confidence interval brackets zero. Based on this concept, the correct polynomial degree for each peak is selected as follows. In a first step, a PEMG1 model is fitted to all peaks. At that moment, the correct polynomial degree is still unknown. Then, the polynomial degree for those compounds whose s_1 confidence interval does not contain the zero value is increased to PEMG2, whereas the PEMG1 model is kept for the remaining peaks (this means that the minimal polynomial degree will be at least one). A second fitting of the chromatogram is then performed, using now the previously fitted parameters as initial estimates, and zero values for the s_2 parameters for those peaks whose model complexity was increased. Once got the convergence, the same study on the confidence intervals is performed, decreasing the model complexity for those compounds for which the higher term in the polynomial was found non-significant, and increasing the polynomial degree for the others. If incidentally, the treatment decreases the polynomial degree of a given compound, it will be kept and no more analysis of the uncertainty of the parameters referred to this compound will be performed. The process continues by decreasing or increasing the polynomial degrees until the degree of all compounds has been identified.

3. Experimental

The reagents, apparatus and experimental procedure were described in Part I [1].

4. Results and discussion

4.1. Comparison of the different algorithms

In order to test the performance of the proposed algorithms and the adequacy of the multivariate selectivity as an estimator of the problem complexity, synthetic experiments were carried out, consisting of several peak arrangements involving two peaks at different overlapping degree. The signals were built from two PEMG1 peaks. Parameter values were (Eq. (4)) $h_0 = 1$, $s_0 = 6$ and $s_1 = 0.1$ for both peaks. The retention time of one of them was kept constant at $t_R = 40$ (arbitrary time units), whereas the other was varied from 50 to 70 (stepped in 0.5 units from 50 to 60 and in 1 unit from 60 to 70). Blank noise of 0.01 standard deviation units was added to each chromatogram. The four algorithms were applied to each experiment.

Usually, the quality of the fitting is established through a straightforward comparison between the predicted and experimental composite signals, quantifying the discrepancies with a maximum likelihood estimator, such as the sum of squared residuals (SSR). However, the result of the deconvolution is not the composite signal but the individual ones, so that not necessarily a low error in SSR is indicative of a right deconvolution. To overcome this, SSR was calculated by comparing not the experimental and predicted composite signals, but each predicted individual profile with the corresponding theoretical one. This value, called SSR_i (individual sum of squared residuals), has been proved to be a good estimator of the deconvolution error [19]. The square root of this value tends to the standard deviation of the blank noise with complete resolved mixtures, in the absence of lack of fit.

The mean of the two SSR_i values (one by compound) was computed first at each deconvolution experiment. In order to get a more precise result, the process was applied 10-fold in each situation, using different seeds for noise generation, calculating a mean value for each overlapping case. The squared root of the sum of these values is depicted vs. the peak distance in Fig. 2a. In this figure, the value of multivariate selectivity (right axis) is overlaid. Fig. 2b plots the mean computation time of the 10 experiments as a function of the peak distance. As can be seen, the four algorithms have the same behaviour up to a certain resolution threshold (around 20 peak distance units), which corresponds to a value of selectivity of ca. 0.99. This means that a simple algorithm will get good results in the resolution of the mixture. It is noteworthy that the two complex algorithms (LOGA and Powell-2) need around 1 min of computation time with any deconvolution problem, whereas the two simplest algorithm took significantly less time (10 and 2.5 s for MSLS and Powell-1 in the considered problem, respectively). This means that using the complex algorithms for problems of resolution higher than 0.98 is a waste of time. Naturally, the computation time is increased with the number of compounds and chromatograms.

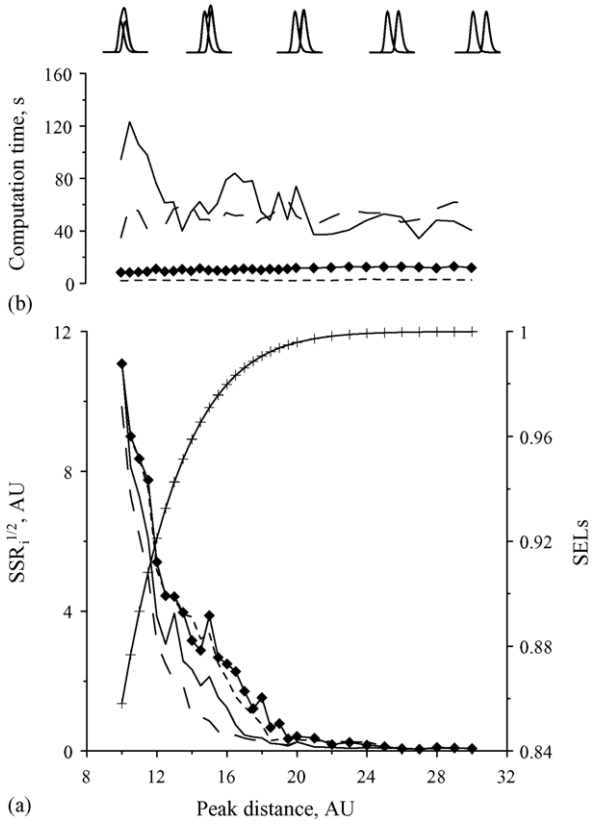

Fig. 2. Performance of the different algorithms explained in Section 2.2 for a two-peak deconvolution problem with different separation degrees. The mean error (left *Y*-axis, part a) in deconvolution (expressed as $SSR_i^{1/2}$, Section 4.1) for a 10-folded experiment is plotted vs. peak distance (*X*-axis). Four algorithms were considered: LOGA (long dash), Powell-2 (solid line), Powell-1 (short dash) and MSLS (solid line with (♦)). The multivariate selectivity is also given for one of the two compounds (right *Y*-axis, solid line with (+)). Computation times for each algorithm are plotted in part b vs. peak distance. The top diagram illustrates the chromatographic situations at 10, 15, 20, 25 and 30 arbitrary units of peak distance.

As expected, the difference in performance for the algorithms becomes evident in situations of increased overlap. Both errors obtained with MSLS and Powell-1 are higher than those given by the complex algorithms. One should note also that LOGA (Fig. 2, long dashed line) performs the best in all situations. As commented, its parameters (e.g. population size, mutation rate, etc.) are however, difficult to establish a priori. In this case, the parameters used in the algorithm configuration were selected, according to Ref. [7]. Although the Powell-2 algorithm performs worse than LOGA, it yields clearly better results than MSLS and Powell-1, and can be considered as a quite robust solution for non-experienced users to resolve highly overlapped situations. Note also that deconvolution errors are inversely correlated to the multivariate selectivity, which justifies the use of this figure of merit as a correct estimator of the complexity of the deconvolution problem, and, consequently, as a guide to select the proper algorithm.

4.2. Deconvolution of real mixtures of several aromatic compounds

In order to test the method performance, samples containing six aromatic compounds (toluene, ethylbenzene, butylbenzene, *o*-terphenyl, amylbenzene and triphenylene) were chromatographed with aqueous–organic mobile phases containing 70, 75, 80 and 85% (m/m) methanol. Injections of standards, together with five mixtures containing different concentration ratios of the test compounds were injected twice within each experimental condition.

The deconvolution studies were performed using three different methods: (i) processing each sample chromatogram separately (single-batch deconvolution), (ii) deconvolving each sample chromatogram together with the standards (sequential multi-batch deconvolution), and (iii) deconvolving simultaneously all sample chromatograms without the standards (multi-batch deconvolution). In methods (ii) and (iii), the assignment of the peaks to a given compound among chromatograms was performed using the method described in Section 2.4, with a threshold of 0.45 for dissimilarity. All the peaks were correctly assessed except the case of mixture 3 at 85% methanol, for which a threshold value of 0.55 was introduced. Only algorithms without random part (Powell-1 and Powell-2) were used. The deconvolution program was applied without any user supervision.

Fig. 3 plots the chromatograms obtained with all the mixtures eluted with four different mobile phases (solid lines), together with the deconvolved profiles with method (i) (dashed lines)—only one of the duplicated injections is plotted to simplify the figure. As can be seen, a virtual baseline separation was obtained at 70% methanol. This allowed eliminating the error introduced in the preparation of the mixtures: the nominal concentration of each compound was corrected con-

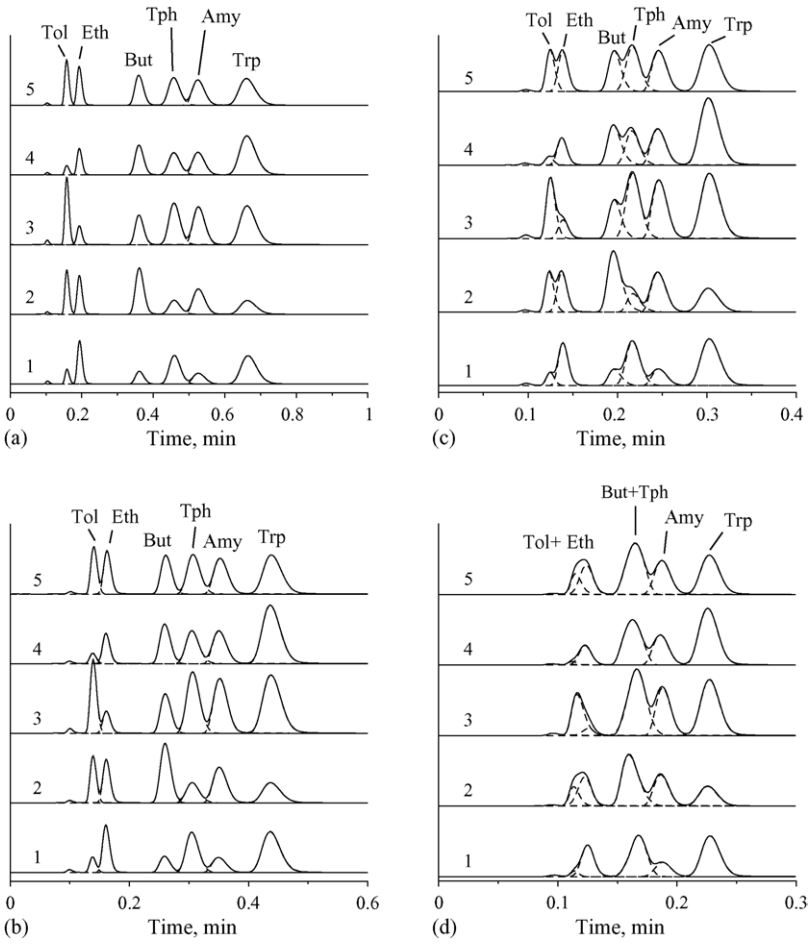

Fig. 3. Experimental chromatograms (solid line) obtained from the injection of five mixtures eluted with: (a) 70, (b) 75, (c) 80 and (d) 85% (m/m) methanol. The deconvolved individual profiles (dashed line) using single-batch deconvolution (method (i)) are overlaid. Compound identities: toluene (Tol), ethylbenzene (Eth), butylbenzene (But), *o*-terphenyl (Tph), amylbenzene (Amy), and triphenylene (Trp).

sidering the deconvolved peak areas obtained from the chromatograms with this mobile phase. For a given compound, the ratio between the peak areas of the respective standard and the mixture injection was computed at 70% methanol, and compared with those obtained for other available mobile phases. Finally, the relative error in concentration was calculated for each mixture, and the mean over the two values obtained with the duplicated injections, computed. Results are presented in Tables 1–3. Each table corresponds to the three different deconvolution methods (i)–(iii).

As can be seen, ∼1% mean relative error was obtained in the deconvolution of the mixture injected at 75% methanol with the three methods (Fig. 3b). This low error is not surprising, taking into account the separation achieved with this mobile phase. This allowed establishing the threshold error associated to the deconvolution: the remaining residual mean error (due both to lack of fit and peak integration) will be ca. 1%. At decreasing resolution (80% methanol), deconvolution errors became more important, particularly with the single-batch treatment—method (i), Table 1. With this method, a mean error of 3% was obtained. Note that the introduction of standards allowed a significant decrease in the deconvolution error (compare Tables 1 and 2), yielding figures similar to those obtained at 75% methanol (around 2%). On the contrary, the use of multi-batch treatment without standards (deconvolving the mixtures altogether: method (iii)) does not improve significantly the results when compared with the single-batch method at this resolution (compare Tables 1 and 3).

The deconvolution of the mixtures eluted at 85% methanol constitutes a more difficult problem. In fact, only at this mobile phase, the multivariate selectivity fell below the above mentioned threshold of 0.98 for toluene and ethylbenzene, and consequently, the Powell-2 method (instead of Powell-1, which was automatically selected for the other mobile phases) was applied to deconvolve the mixture. For this mobile phase, butylbenzene and *o*-terphenyl coeluted too strongly, and the second- and third-order derivatives were unable to distinguish them within the peak cluster (for this reason, no results are presented in the table). Also, no evidence (i.e. a shoulder) can be observed in the original chromatogram indicating the presence of two compounds (Fig. 3d). With this mobile phase, the use of the multi-batch method was mandatory to decrease the errors from 19–50% with the single-batch treatment (Table 1)

Table 1
Relative errors (%) in concentration in the deconvolution of a set of mixtures injected in different mobile phases using single-batch deconvolution—method (i)

Mobile phase[a]	Mixture	Compound[b]					
		Tol	Eth	But	Tph	Amy	Trp
75	1	2.0	1.8	1.1	0.6	0.1	1.7
	2	0.2	0.0	0.5	−1.3	−0.5	−0.6
	3	0.4	1.2	0.7	−0.5	−1.1	0.8
	4	−1.5	−0.8	−1.1	−3.0	−3.8	−1.9
	5	1.9	2.1	1.8	0.6	0.7	2.3
80	1	8.9	1.3	0.8	1.0	−2.2	2.2
	2	1.2	−3.6	−9.8	16.6	−3.2	−1.9
	3	1.8	4.6	2.1	1.2	−0.8	3.3
	4	6.6	−0.5	−4.2	1.6	−4.9	−0.6
	5	4.1	1.7	1.3	1.3	0.1	2.9
85	1	48.8	−11.9	–[c]	–[c]	−7.6	2.9
	2	37.2	−26.3	–[c]	–[c]	5.1	5.2
	3	−18.8	60.9	–[c]	–[c]	−1.0	2.2
	4	49.6	−9.7	–[c]	–[c]	−6.4	2.6
	5	30.7	−21.3	–[c]	–[c]	−4.3	4.5

[a] Methanol, % (m/m).
[b] See Fig. 3 for compound identities.
[c] Unresolved.

to 0.4–5.2% (inclusion of standards, Table 2), or 0.8–49% (multi-batch treatment with the chromatograms of all mixtures but no standards) for toluene and ethylbenzene.

The results of the deconvolution for amylbenzene deserve a special comment. For this compound, the relative error was significantly increased at 85% methanol, even when the resolution was not dramatically low (Fig. 3d). This is particularly striking for the multi-batch deconvolution without standards (Table 3). The explanation of this effect is the wrong assignment of the peaks of butylbenzene and *o*-terphenyl as a single peak case. As these peaks are unresolved, a certain lack of fit remains always in the peak cluster, which is partially compensated biasing the peak shape of amylbenzene. In the multi-batch deconvolution (cases ii and iii), the wrongly assigned as single-peak butylbenzene + *o*-terphenyl is forced to be constant among injections. This is specially troublesome since the relative peak heights of both compounds are not constant, so neither the peak shape of the cluster. The effect is less important when standards are included, since the peak shape of amylbenzene is fixed in a first step by fitting the individual injections of this compound. In case of the single-batch treatment, the lack of fit introduced by the wrong assignment of butylbenzene and *o*-terphenyl as a single peak is better processed, since the peak shape of this single peak can vary from batch to batch.

Table 2
Relative errors (%) in concentration in the deconvolution of a set of mixtures injected in different mobile phases using multi-batch deconvolution with the inclusion of standards—method (ii)

Mobile phase[a]	Mixture	Compound[b]					
		Tol	Eth	But	Tph	Amy	Trp
75	1	−0.04	2.2	1.0	0.01	2.0	2.1
	2	−1.2	0.6	0.5	−3.1	0.4	0.2
	3	0.1	1.0	0.2	−1.6	2.8	1.1
	4	−3.2	−0.8	−1.2	−3.7	−0.1	−0.8
	5	1.2	2.7	1.7	0.3	4.0	2.8
80	1	0.3	4.4	−0.7	1.8	−1.0	3.5
	2	0.3	1.9	−0.9	−3.1	−3.2	0.0
	3	3.2	5.6	0.9	2.0	1.8	3.7
	4	−2.5	2.2	−0.7	−2.1	−1.5	1.5
	5	3.1	5.2	0.8	2.2	1.3	3.5
85	1	4.6	1.3	–[c]	–[c]	−6.9	4.4
	2	0.4	5.0	–[c]	–[c]	−2.8	4.0
	3	−5.2	17.3	–[c]	–[c]	−1.1	3.5
	4	−0.6	−0.3	–[c]	–[c]	−7.7	2.4
	5	2.1	3.4	–[c]	–[c]	−3.1	4.3

[a] Methanol, % (m/m).
[b] See Fig. 3 for compound identities.
[c] Unresolved.

Table 3
Relative errors (%) in concentration in the deconvolution of a set of mixtures injected in different mobile phases, processing those mixtures with the same mobile phase altogether, and using multi-batch deconvolution without the inclusion of standards–method (iii)

Mobile phase[a]	Mixture	Compound[b]					
		Tol	Eth	But	Tph	Amy	Trp
75	1	1.0	2.4	1.0	0.6	1.4	1.8
	2	−0.8	0.5	0.1	−2.5	0.3	0.0
	3	0.3	1.0	0.8	−0.4	1.1	1.3
	4	−2.5	−0.8	−1.3	−3.1	−1.2	−1.0
	5	1.3	2.5	1.9	0.6	2.0	2.3
80	1	8.0	2.3	−1.1	1.8	0.1	2.5
	2	6.6	−3.2	−4.7	3.9	−2.1	−1.5
	3	9.2	−0.4	−0.6	3.0	2.2	3.4
	4	5.4	−0.3	−4.2	0.7	−1.8	0.1
	5	8.7	0.4	−1.6	3.1	1.2	2.6
85	1	10.2	0.8	–[c]	–[c]	−26.5	3.0
	2	−8.4	13.8	–[c]	–[c]	−17.1	1.8
	3	−15.6	48.7	–[c]	–[c]	−20.1	2.2
	4	8.8	−1.4	–[c]	–[c]	−23.7	1.6
	5	−6.8	12.2	–[c]	–[c]	−22.3	2.0

Mixtures are deconvoluted together in a multi-batch deconvolution without the inclusion of standards—method (iii).
[a] Methanol, % (m/m).
[b] See Fig. 3 for compound identities.
[c] Unresolved.

5. Conclusions

The polynomial-exponential modified Gaussian model presented in this work is useful to deconvolve chromatographic peaks. It constitutes an attractive alternative to solve the problem of baseline raisings present in the original formulation of the polynomially modified Gaussian function (PMG) previously described [10]. The PEMG model has the same advantages as the PMG, among which its high stability and easy convergence in fittings constitute features of primordial importance when this model is applied with deconvolution purposes, in an automatic program (with no user supervision).

As an aside, increasing or decreasing the polynomial degree of the standard deviation can modulate the complexity of the model. This avoids both underfitting and overfitting when this model is used for deconvolution, even when applied to chromatograms presenting peaks with different asymmetry levels. An on-line test assessing the significance of each parameter of the polynomial allows an automatic tuning of the model complexity, according to each situation. This test is applied by computing the parameter uncertainties via error propagation theory.

The practical application of deconvolution introduces different levels of complexity. The peaks usually overlap in a diverse extent, not only when comparing different chromatograms, but also within the same chromatogram. This makes the use of different deconvolution algorithms convenient, to both guarantee an accuracy level good enough in situations of high overlap, and avoid high computation times in cases of larger resolution. The multivariate selectivity computed, according to the Lorber's definition [8] is a good estimator of the problem complexity, and has allowed the development of an automatic procedure for the selection of the most adequate algorithm.

Approaches containing random part, such as conventional or hybrid genetic algorithms, are difficult to automate and require a certain background by the user to configure them appropriately. On the other hand, the classical optimisation methods (such as the Gauss–Newton or the Powell methods) require less user experience, but can be trapped into local solutions. The new algorithm presented in this work, which alternates the fitting of the original signal and the second derivatives, avoids partially the problem of local convergence, and presents the advantage of having no random part. This latter feature simplifies its implementation in an automatic program, because it requires less user experience to configure its parameters and also there are fewer opportunities to make the wrong decision.

When a high overlap is detected, treating the available information from different injections as one data set can notably decrease the errors of deconvolution. This multi-batch treatment includes different samples sharing common compounds, or the injection of the standards together with the mixture. The software presented here develops all the steps for both single- and multi-batch deconvolution in a fully automated way.

Acknowledgements

This work was supported by projects 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 pre-doctoral stay in Brussels, and for an FPI grant.

References

[1] G. Vivó-Truyols, J.R. Torres-Lapasió, A.M. van Nederkassel, Y. Vander Heyden, D.L. Massart, J. Chromatogr. A 1096 (2005) 133.
[2] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam, 1998.
[3] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, second ed., Cambridge University Press, Cambridge, 1992.
[4] D. Corne, M. Dorigo, F. Glover (Eds.), New Ideas in Optimisation, McGraw-Hill, London, 1999.
[5] D.B. Hibbert, Chemom. Intell. Lab. Syst. 19 (1993) 277.
[6] R.E. Schaffer, G.W. Small, Anal. Chem. 69 (1997) 236A.
[7] G. Vivó-Truyols, J.R. Torres-Lapasió, A. Garrido-Frenich, M.C. García-Alvarez-Coque, Chemom. Intell. Lab. Syst. 59 (2001) 107.
[8] A. Lorber, Anal. Chem. 58 (1986) 1167.
[9] V. Di Marco, G.G. Bombi, J. Chromatogr. A 931 (2001) 1.
[10] J.R. Torres-Lapasió, J.J. Baeza-Baeza, M.C. García-Alvarez-Coque, Anal. Chem. 69 (1997) 3822.
[11] G. Vivó-Truyols, J.R. Torres-Lapasió, R.D. Caballero-Farabello, M.C. García-Alvarez-Coque, J. Chromatogr. A 958 (2002) 35.
[12] R.D. Caballero, M.C. García-Alvarez-Coque, J.J. Baeza-Baeza, J. Chromatogr. A 954 (2002) 59.
[13] P. Nikitas, A. Pappa-Louisi, A. Papageorgiou, J. Chromatogr. A 912 (2001) 13.
[14] E. Fuguet, C. Ràfols, J.R. Torres-Lapasió, M.C. García-Alvarez-Coque, E. Bosch, M. Rosés, Anal. Chem. 74 (2002) 4447.
[15] W.E. Hart, Adaptive Global Optimization with Local Search, Ph.D. Thesis, University of California, San Diego, 1994.
[16] A.P. De Weijer, C.B. Lucasius, L. Buydens, G. Kateman, Anal. Chem. 66 (1994) 23.
[17] X.G. Shao, Z.H. Chen, X.Q. Lin, Chemom. Intell. Lab. Syst. 50 (2000) 91.
[18] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) 782A.
[19] G. Vivó-Truyols, J.R. Torres-Lapasió, M.C. García-Alvarez-Coque, J. Chromatogr. A 991 (2003) 47.