Data Analysis for Multi-Dimensional Liquid Chromatography

Pirok, B.W.J.; Rutan, S.C.; Stoll, D.R.

DOI
10.1201/9781003090557-8

Publication date
2023

Document Version
Final published version

Published in
Multi-Dimensional Liquid Chromatography

8 Data Analysis for Multi-Dimensional Liquid Chromatography

Bob W.J. Pirok, Sarah C. Rutan, and Dwight R. Stoll

CONTENTS

DOI: 10.1201/9781003090557-8

8.1 INTRODUCTION

Upon maximizing the separation power of 2D-LC, particularly in the case of comprehensive 2D-LC, after data acquisition we are often left with a large and rich set of data. It is at this moment that we realize that our gain in peak capacity comes at the cost of much greater data complexity. Indeed, some have referred to the sheer size of our LC×LC-MS/MS datasets as a "tsunami of data" [1]. Making the step from 1D-LC to 2D-LC, and perhaps even adding an extremely powerful high-resolution mass spectrometer, may offer powerful analytical resolution, but all our work would be in vain unless we find a way to find meaning in the resulting data and use it to answer our analytical questions. New analytical tools typically generate more and more complex data, from which it is increasingly difficult to deduce clear answers and useful information. In this chapter, we will address all the relevant steps related to data analysis for 2D-LC. We will learn about the format of the raw data and see how this can be visualized into a two-dimensional plot. Next, post-acquisition corrections such as baseline correction, peak alignment, and removal of undesired background signals will be discussed. We will then focus on those methods that allow for quantification, with an emphasis on techniques for peak detection and for curve resolution.

As we proceed through these steps, we will note that the datasets obtained from (comprehensive) two-dimensional liquid chromatography separations are really a collection of one-dimensional (second dimension) chromatograms. Thus, we will sometimes employ chemometric approaches that focus on analysis of one-dimensional signals. We will also see that the field of chemometrics and data analysis is highly dynamic and active, with a large number of new methods proposed in recent years. While such innovations are advancing the field, making the most of them will require a critical comparison of performance across a variety of datasets to help users understand the strengths and weaknesses of different approaches.

8.2 DATA REFORMATTING AND VISUALIZATION

To rapidly gain insight from the data, one of the first priorities is to visualize the data. As with 1D-LC, this will allow us to manually inspect the chromatogram. In addition, we will see in later sections that the choice of preprocessing technique often depends on the type of data acquired. For heartcut 2D-LC approaches, additional 1D chromatograms for each heartcut can be found in the datafile. These may be visualized by simply plotting the detected signal vs. the time vector as is done with any 1D chromatogram, and analyzed using tools already available for 1D-LC. Thus, most of the following discussion pertains to sLC×LC and LC×LC data.

8.2.1 Data Reformatting

For sLC×LC and LC×LC data, producing the 2D chromatogram requires an additional step. We have learned in Chapter 4 that LC×LC often employs a single detector that monitors the effluent of the ^{2}D separation. Consequently, raw LC×LC datafiles essentially contain single, long, one-dimensional data sequence of the detector signal, where individual ^{2}D chromatograms are serially connected. An example of such a concatenation of the ^{2}D chromatograms is shown in Figure 8.1A. The repeating pattern shows the presence of system peaks in the ^{2}D. The most prominent example of this is the dead-volume marker. Indeed, transferred fractions of ^{1}D effluent contain large quantities of solvents and buffers in the ^{1}D mobile phase. Consequently, these components may each systematically give rise to a peak in the second dimension. The length of each ^{2}D section of the long string of data is, of course, equal to the modulation time.

In order to plot the 2D chromatogram, the data must first be reformatted by dividing the long, one-dimensional data sequence into sections (i.e., vectors) using the modulation time that was used to acquire the data. These vectors are then rearranged into a $M \times N$ matrix, which is plotted as a 2D

chromatogram (Figure 8.1B). This process is also sometimes referred to as folding the data. Here, *M* equals the number of modulations and *N* the number of datapoints within each modulation (i.e., the product of the ^{2}D detector sampling frequency and the modulation time).

8.2.2 Visualization

Visualization of the matrix is generally done by plotting the two-dimensional space as a color map or contour plot. An example of the first is shown in Figure 8.1B. As is customary, the ^{1}D time is plotted on the x-axis and the ^{2}D time on the y-axis. The detector signal level is represented by the color.

At this stage the representation in Figure 8.1B is accurate, but sometimes interpreting a chromatogram in this format can be confusing. Because the sampling frequency of the ^{1}D data points (i.e., Figure 8.1D) is dictated by the modulation frequency (as discussed in Section 3.4), it can be beneficial to interpolate the ^{1}D data points to achieve a smoother profile [2]. Thus, LC×LC chromatograms are often smoothed by interpolating the datapoints, resulting in chromatograms like that in Figure 8.1C. The chromatogram shown was obtained using bilinear interpolation [3] (using interp2 function in Matlab) where the linear interpolation is first applied in one dimension and then in the second dimension. Cubic spline interpolation methods are frequently used for either one- or two-dimensional interpolation.

While visualization of these datafiles is not very complicated, it is useful to realize that there are a number of different options in the way we can visualize the data. First of all, summing all datapoints vertically (i.e., the sum of all rows in the *M*x*N* matrix) returns the full ^{1}D chromatogram as if it were recorded by a 1D-LC (Figure 8.1D), but at a data acquisition rate much lower than we would normally use in 1D-LC. Similarly, summing all datapoints horizontally (i.e., the sum of all columns in the matrix) yields the 1D-LC version of the ^{2}D separation (Figure 8.1E). However, we see immediately that the quality of description is rather poor compared to original 1D chromatograms obtained using either the ^{1}D or the ^{2}D retention mechanisms, carried out under conventional 1D-LC conditions (Figures 8.1F and 8.1G, respectively). This brings us to another realization: if there is no dedicated ^{1}D detector, all our information about the ^{1}D separation is obtained indirectly through the ^{2}D detector. Consequently, the ^{2}D separation can be regarded as the detector of the ^{1}D [4], and, as stated above, the modulation time essentially establishes the ^{1}D sampling frequency. This underlines the importance of minimizing undersampling, which results in the loss of ^{1}D resolution by sampling the ^{1}D effluent at a frequency that is too low (see Section 3.4 for more detail on this topic).

In the event that a multi-channel detector is used (e.g., MS or DAD), the raw data also contains a full spectrum for each time point allowing the assembly of a 3D data matrix.

8.2.3 Correction for Wrap-Around

When analytes do not completely elute within the modulation period set by the injection of fraction of ^{1}D effluent that contains them, they may elute entirely or partly in subsequent modulations. This is referred to as "wrap-around". While it is not very common in LC×LC (because gradient elution is frequently used in the second dimension), it is rather common in GC×GC because ^{2}D separations are nominally isothermal.

One approach to correct for this issue is to treat the 2D chromatogram as a continuous 3D cylinder, essentially connecting one modulation to another. Weusten *et al.* developed an algorithm exploiting this principle to correct for wrap-around in the analysis of urine samples using GC×GC-MS [5]. Another approach was developed by Micyus *et al.*, which specifically detects occurrences of wrap-around using an integer fraction of the original modulation period [6]. Using this approach, the algorithm determines the absolute retention times.

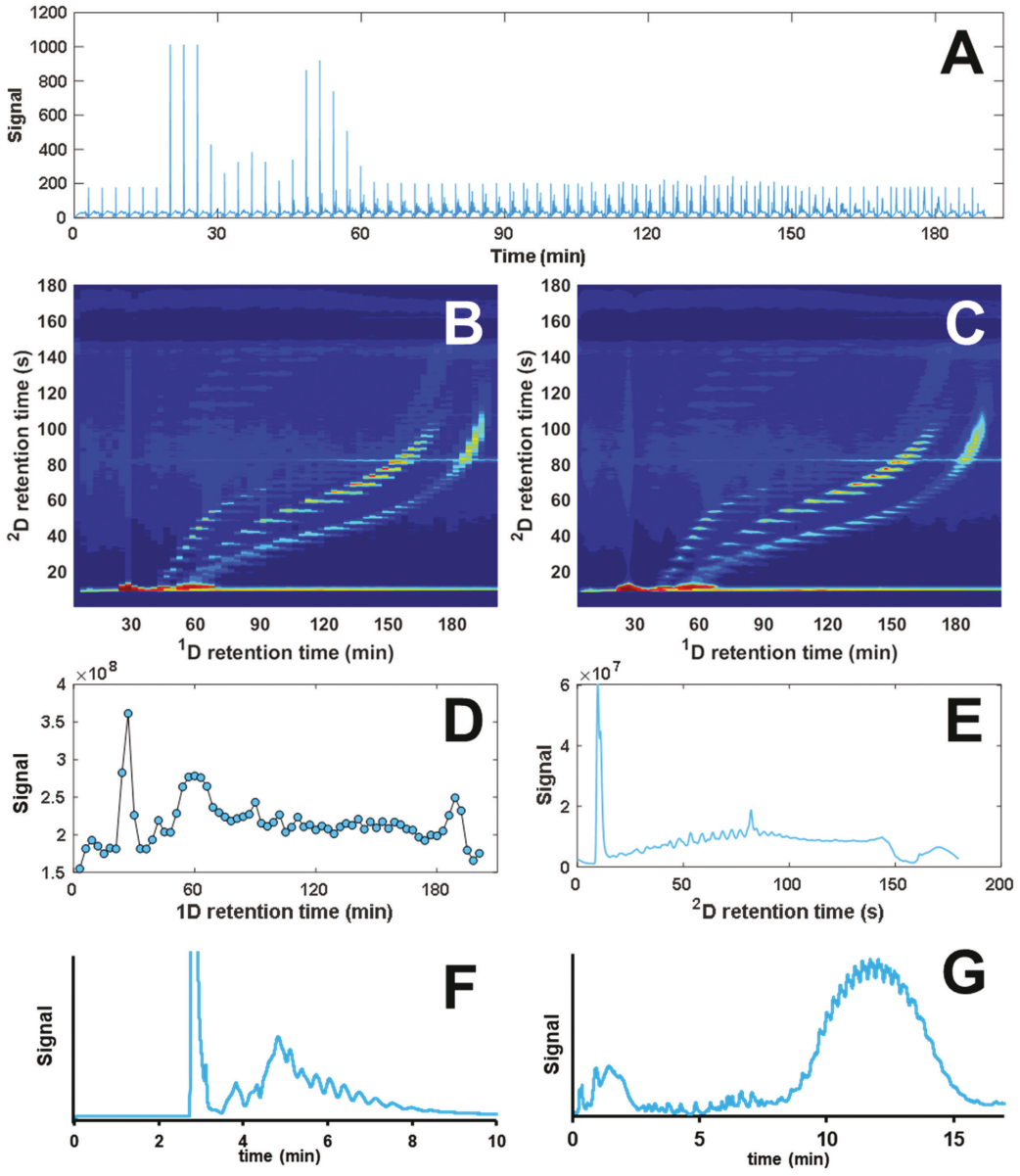


FIGURE 8.1 A) Raw LC×LC detector signal from the separation of some ionic industrial surfactants as acquired from the instrument. B) Two-dimensional LC×LC chromatogram produced from the data shown in A. C) Bilinear interpolation of LC×LC data from 2D plot shown in B. D) Reconstructed ^{1}D chromatogram (i.e., the signal at each time point is the sum of all signals within the modulation period corresponding to that time point). E) Reconstructed chromatogram that would be obtained if the sample were analyzed by 1D-LC using the conditions of the ^{2}D separation (i.e., the signal at each time point is the sum of all signals at that ^{2}D time across the entire 2D separation). F) 1D-LC chromatogram of the same sample using the ^{1}D separation conditions. G) 1D-LC chromatogram of the identical sample using a similar gradient program.

8.3 DATA PREPROCESSING – GOALS AND TECHNIQUES

The need for data preprocessing can perhaps best be understood from the quote that is popular in the field of computer science – "Garbage in, garbage out" – which is often abbreviated simply as GIGO. This phrase was coined to indicate that no matter how good data-processing methods are, they all rely on the quality of the original data. In our case, this means that our ability to distill useful information from the data depends on the quality of the original chromatogram. Here, quality is a nebulous, difficult-to-quantify term that relies heavily on the perspective of the scientist.

For example, if we look at the 1D and 2D chromatographic signals shown in Figure 8.2 from the perspective of a chromatographer, we may see well separated peaks and deem the quality of separation as good. However, from the perspective of a chemometrician, the signal contains undesired distortions (e.g., high-frequency noise and baseline drift) that often complicate data analysis, as we will see as we progress through this chapter. These distortions may impede our ability to obtain accurate, quantitative information, and even prevent us from locating all relevant regions of interest (ROI) in the chromatogram (i.e., those that contain chromatographic peaks).

While it is clear that data preprocessing is crucial to derive accurate conclusions from the data, it must also be noted that preprocessing steps do manipulate the data. Great care must be taken to prevent improper removal of useful information. Often, data preprocessing techniques rely on premises with respect to characteristics of the data and selection of the appropriate technique must be tailored to the dataset; this will also be addressed in this section.

8.3.1 Noise Reduction

As is the case with any analytical signal, a chromatogram comprises several components, each with a different frequency: (low-frequency) baseline drift, the signal of relevance, and (high-frequency) noise. This is shown in Figure 8.2. In LC, drift mainly arises from the gradient programs used, whereas noise is mainly induced by small fluctuations in flow rate, mobile-phase temperature, mobile phase composition fluctuations due to pump imprecision, but also random fluctuations in the signal induced by disturbances in the detector (e.g., shot noise, thermal noise, etc.). Removing these undesired low and high frequency components from the signal will significantly improve quality of the chromatogram, most notably the signal-to-noise ratio (S/N). In addition, the successful performance of derivative-based peak detection methods depends critically

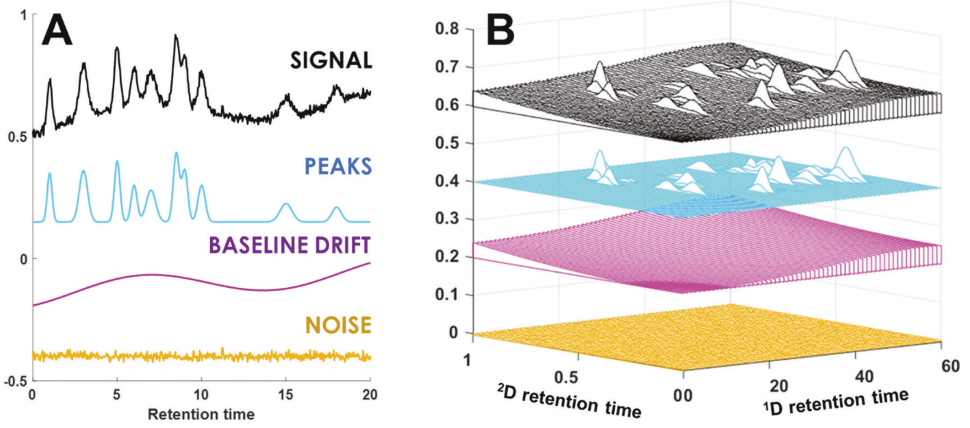


FIGURE 8.2 A chromatographic signal consists of several frequency components, including the actual chromatographic peaks plus low-frequency baseline drift and high-frequency noise. This is shown for A) 1D-LC and B) LC×LC.

on noise reduction. Unsurprisingly, research into removal of these background components goes back to the 1960s [7, 8].

Depending on the origin, noise typically features normally distributed (in terms of signal) distortions occurring at higher frequencies. Since noise may significantly impact both the precision and accuracy of quantitation and the performance of various subsequent data-processing techniques (e.g., derivative-based peak detection, and the determination of peak start and stop times), it is imperative that it must be properly reduced. However, it must be realized that noise-filtering essentially treats a symptom and not the cause. Thus, the GIGO principle applies here and a proper first approach focuses on analysis of the source of any excessive noise.

Noise filtering can be conducted both in the time domain, and the frequency domain, although filtering noise directly in the time domain is equivalent to indirect filtering in the frequency domain. Additionally, noise filters can be classified as low-pass or high-pass filters. Low-pass filters cut off frequencies above a set threshold, whilst allowing lower frequencies through. This renders them ideal for filtering baseline noise. Conversely, high-pass filters cut-off frequencies below a set threshold, making them more suitable for removal of baseline drift.

Some of the most popular filtering approaches in analytical chemistry are the moving average and polynomial filters, also known as Savitzky-Golay filters [9–14]. These filters are based on replacing the center point of a window that moves across the chromatogram with either the average, or the result of a *local* polynomial fit to the data within each window. This is achieved practically by using a weighted average of the windows points, with well-established weighting factors. This type of filter is a low-pass filter. First and higher order of the signal are easily calculated using this method.

The selection of the window width is key in the application of these filters. A filter with a window width of 3 will not be as effective in removing noise as a much wider filter that consults more neighboring points. However, filters with wider windows risk averaging out actual chromatographic peaks, potentially destroying information and thus rendering the method less accurate. This is particularly problematic for modern chromatographic separations executed at UHPLC conditions including very fast ^{2}D separations employed in comprehensive 2D-LC. An example is shown in Figure 8.3, where, regardless of the width of the filter, an increasing fraction of the area belonging to the Gaussian peaks is lost as the filter window width is increased. This example illustrates the importance of carefully selecting preprocessing methods and parameters. Another example is tuned filters, which relate to the concept of matched filtration [15]. Such filters are sometimes used in commercial packages.

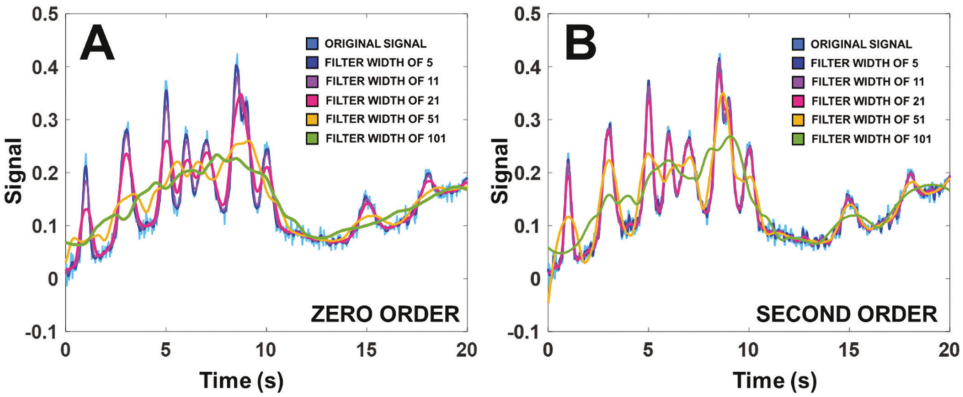


FIGURE 8.3 Two common signal filters applied to a section of a chromatographic signal. A: Zero-order moving average filter, B: Quadratic Savitzky-Golay filter.

The filters discussed above clearly improve the S/N ratio in the time domain. A common approach to filter in the frequency domain utilizes transformation functions. While many transformation functions can be used, the Hadamard and Fourier transformation are two well-known examples. Fourier transformation (FT) exploits the fact that true chromatographic peaks and noise differ in frequency [16, 17]. The largest practical difficulty with this approach is that it is difficult to use on chromatograms with widely varying peak widths.

For chromatographic signals containing vastly different frequency components that must be retained through the filtering process, wavelets can be used which adapt to the signal characteristics using both high-pass and low-pass filtering components. Wavelets automatically apply a narrow window to find high-frequency components and a wide window for low-frequency components [18]. The orthogonal and local functions employed by wavelets render them efficient and effective for processing signals that exhibit a wide range of frequency components. Daszykowski and coworkers have used wavelets for smoothing 2D-electropherograms [19].

Of most interest to users of comprehensive 2D-LC methods, both Fourier transform [14] and wavelet transforms [14, 20] can be applied to two-dimensional data sets. However, to our knowledge neither of these methods have yet been applied to LC×LC datasets.

8.3.2 Baseline Drift and Comprehensive Background Correction Approaches

Removal of baseline drift often involves the use of a curve-fitting approach, and is often combined with noise filtering. Both methods utilize a loss function to fit a curve through the presumed background signal. The combined approach is typically referred to as "background correction".

Relative to background correction for 1D chromatographic data, background correction for LC×LC chromatograms is challenging. Some of the challenges are illustrated in Figure 8.4, which displays 3D plots from the separation of a mixture of industrial surfactants (these are the same data shown in Figure 8.1C, plotted in different ways). As we will see in the remainder of this section, the effectiveness of background correction approaches relies significantly on regional characteristics and sparsity of the data. In the case of Figure 8.4, the chromatogram exhibits very different features in different regions, at different magnitudes, due to different chromatographic phenomena. For example, the "ridge" highlighted in Figure 8.4B is due to elution of components of the ^{1}D effluent that are unretained by the ^{2}D column, and also may arise from artifacts from rapidly changing refractive indices of the mobile phase in UV-visible detection when fast gradient separations are employed [21]. With large volumes of ^{1}D effluent injected into the second dimension, the intensity of this signal is often rather large and thus the ridge feature is prominent. Correction of such a ridge requires a different approach than correction of the rather normal peaks behind it as is clearly visible from Figure 8.4C, which is the same as Figure 8.4B but rotated by 180°. If the ^{1}D column bleeds (i.e., loses stationary phase) under the conditions of the ^{1}D separation, the components of the bleed may be visible in the 2D plot as the systematic occurrence of a ridge in the middle of the chromatogram as shown in Figure 8.4D. In 1D-LC column bleed or mobile phase impurities typically appear as a single (but sometimes broad) peak or feature, in 2D-LC this turns into a significant disturbance spanning the entire 2D chromatogram. Another profound feature frequently encountered in LC×LC are the distortions typically occurring during the equilibration step at the end of each ^{2}D separation when gradient elution is used (Figure 8.4E).

The detection of analytes at trace concentrations is normally a challenge even in 1D-LC. In LC×LC this can be even more challenging because the background correction method used must perform well in both dimensions to facilitate reliable detection of real but small chromatographic features on top of the background (i.e., recognition of peaks in adjacent ^{2}D chromatograms as one 2D peak).

Background correction approaches can be classified as parametric or non-parametric. Parametric methods assume a shape of the baseline defined by a number of parameters. Examples of these types

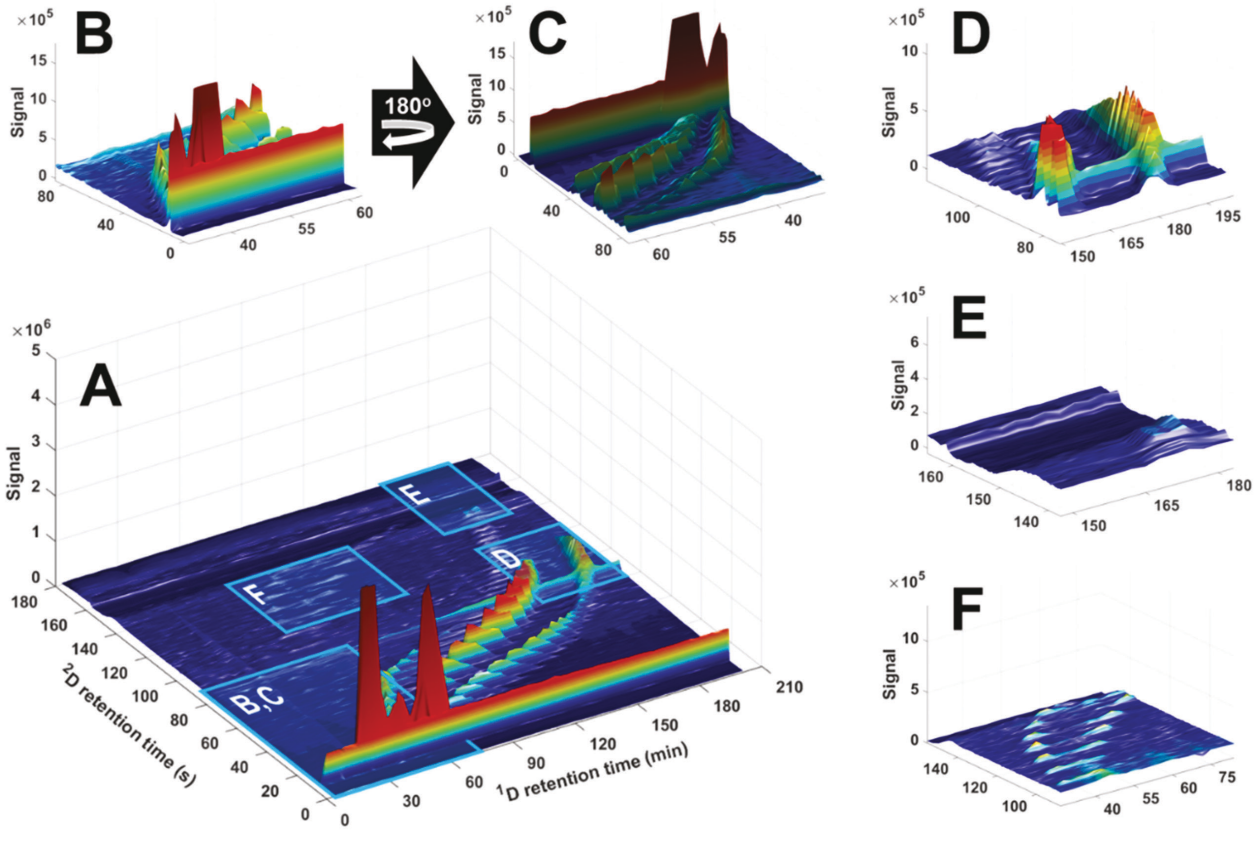


FIGURE 8.4 A) 3D plot of the LC×LC separation shown as a 2D plot in Figure. 8.2C. Various sections of the chromatogram have been selected to create a number of insets. B) The typical "ridge" resulting from the systematic elution of unretained compounds in the second dimension. C) This "ridge" requires a vastly different background correction approach than regular eluting peaks on the 2D plane. D) ^{1}D column bleed results in systematic elution of species in the 2D plots, significantly hindering data analysis. E) Distortions commonly encountered due to the equilibration step of a ^{2}D gradient. F) Separation of two homologous series of compounds.

Source: Adapted with permission from [22].

of approaches are the polynomial regression methods. Conversely, non-parametric approaches do not assume a shape of the background and the number of parameters used for the models depends solely on the data. When a large number of peaks are clustered together, fewer data points are available that describe the background. Consequently, background correction becomes increasingly more challenging with the presence of more clusters of peaks.

8.3.2.1 Penalized Least-Squares Approaches

Penalized least-squares is a smoothing method based on the Whittaker smoothing function [23], and is frequently applied for background correction. The fit of a model to the data, F, expressed as the sum of squares (SSQ), is balanced against its roughness (R) through a smoothing parameter λ. This relationship is also given by Eq. 8.1.

$$Q = F + \lambda R = \sum_{i=1}^{N} \left(x_i - z_i\right)^2 + \lambda \sum_{i=2}^{N} \left(\Delta z_i\right)^2 = \left\|\mathbf{x} - \mathbf{z}\right\|^2 + \lambda \left\|\mathbf{D}\mathbf{z}\right\|^2 \tag{8.1}$$

Here, x_i and z_i represent the data points in the signal and the model, respectively. $\mathbf{D}$ is an N x N-1 difference matrix containing values of 1, -1 and 0 such that $\mathbf{Dz} = \Delta\mathbf{z}$, where $\Delta\mathbf{z}$ is the difference vector for $\mathbf{z}$. Minimization of the cost function (Q) represented in Eq. (8.1) is an example of least-squares minimization with a regularization term to introduce an appropriate penalty (in this case against too much roughness). Solving for $\frac{\partial Q}{\partial z} = 0$ (minimization of the Q function) then returns

$$\left(\mathbf{I} + \lambda \mathbf{D}^{\mathbf{T}}\mathbf{D}\right)\mathbf{z} = \mathbf{x} \tag{8.2}$$

where the superscript T indicates a matrix transpose. For correction of the baseline, a binary matrix, $\mathbf{W}$, can be created that labels whether a datapoint belongs to a detected peak or not [24,25], as given by Eq. 8.3:

$$\left(\mathbf{W} + \lambda \mathbf{D}^{\mathbf{T}}\mathbf{D}\right)\mathbf{z} = \mathbf{W}\mathbf{x} \tag{8.3}$$

The disadvantage of this weighting method immediately becomes clear as it induces the premise that the location of peaks must be known, thus requiring a peak-detection algorithm to be executed *a priori*, which in turn may require baseline correction in order to function properly. In their asymmetrical Least Squares (asLS) method, Eilers introduced an asymmetry parameter to resolve this limitation [26]. This parameter allocates for an increase or reduction of weights imposed on positive and negative deviations of the signal from the baseline. One limitation of asLS, however, was that the asymmetry factor was determined for the entire baseline equivalently. This led to the introduction of adaptive iteratively reweighted penalized least squares (airPLS) [27]. In airPLS, a more-accurate weight vector can be obtained by iteratively solving a weighted PLS case until the difference between the signal and model is at least three orders of magnitude smaller than the signal value. The consequence is that regions of the baseline can be penalized differently.

Nevertheless, the performance of asLS and airPLS is significantly impacted by the presence of noise. Various methods have since then been introduced to improve these strategies, including asymmetrically reweighted PLS (arPLS) [28], modified adaptive iteratively reweighted penalized least squares (MairPLS) [49], and morphologically weighted penalized least squares (MPLS) [29, 30]. The latter method employs morphological analysis to determine the weighting vector more accurately. In MairPLS, the chromatogram is pre-treated before airPLS is performed.

Ultimately, all asymmetric least squares approaches rely on an accurate determination of the λ parameter to describe the baseline. Once a good value for λ has been found, it can be used for the entire dataset.

8.3.2.2 Local Minimum Values (LMV)

Baseline correction can also be carried out using local minimum values (LMVs) [31]. First, the signal is scanned for local minima, by searching all datapoints x_i which are smaller than their neighbors x_{i-1} and x_{i+1}, as is also reflected by the two conditions in Eq. 8.4. An example of these values is shown in Figure 8.5A, where each minimum value is presented as a red dot.

$$x_{i-1} > x_i \tag{8.4a}$$

$$x_i < x_{i+1} \tag{8.4b}$$

Next, the resulting vector of minima, or minimum vector, is stored. At this stage, peaks may still be present as is illustrated by Figures 8.5A and 8.5B. To remove data points corresponding to peaks, an iterative moving-window strategy is employed. All data points corresponding to S/N > 2.5 are treated as outliers and replaced by the median value within that window. This process is repeated until a convergence point is reached (Figure 8.5C).

The resulting vector of datapoints should now exclusively contain datapoints that describe the baseline. Through linear interpolation the vector is subtracted from the original data (Figure 8.5D). One disadvantage of this approach may be the required *a priori* estimation of the width of the moving window.

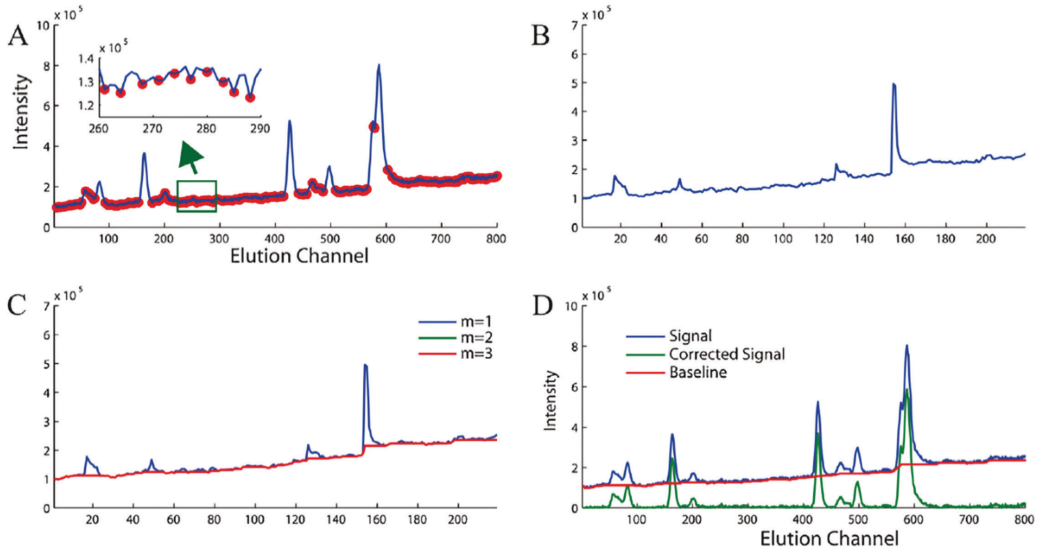


FIGURE 8.5 Example of background correction using local minimum values (LMV). A) Assignment of LMV points. B) Resulting vector of LMV, which still contains peaks. C) Removal of peak points using a moving-window strategy after *m* iterations. D) Overlay of the original signal, LMV vector and the corrected signal. Reprinted from *Journal of Chromatography, A*, 1449, H. Fu, H. Li, Y. Yu, B. Wang, P. Lu, H. Cui, P. Liu, Y. She, Simple automatic strategy for background drift correction in chromatographic data analysis, 89–99, Copyright (2016), with permission from Elsevier.

The LMV method has been compared to other approaches [32] including the moving-window-minimum-value (MWMV) [33], morphological penalized least-squares (MPLS) [34], and the orthogonal subspace projection (BD-OSP) methods [35].

For the comparison with MPLS and MWMV, simulated data were used which comprised singular peaks as well as peaks that overlapped with two to four other peaks [32]. MPLS and MWMV were used for background correction and the resulting peak areas and standard deviations were compared to values obtained after LMW combined with robust statistical analysis (LMW-RSA). The comparison was carried out for different degrees of noise present in the data. The LMV-based approach was found to yield the best accuracy for determination of peak features in all cases except for the highest noise level. MWMV was moderately less accurate, whereas MPLS yielded significantly deviating values with peak area recoveries between 53% and 74%, as opposed to the near 100% recovery by LMW-RSA. The influence of the moving-window width was found to be negligible. Because these methods use neighborhood minima in establishing the baseline, these types of approaches may be particularly useful when generalized to comprehensive 2D-LC data.

Comparison of LMW-RSA with BD-OSP was also carried out using LC-QToF-MS data [32], however the differences in performance were only evaluated on a qualitative level. The BD-OSP approach was unsuccessful in completely removing the background drift; the LMW-RSA method did remove the background drift, but also a portion of the information contained in the total ion-current chromatogram (TIC).

It should be noted that while the study found the influence of the width of the window to be negligible, chromatograms with large domains of overlapping peaks (i.e., few sections containing information regarding the baseline) will be difficult to be processed by the LMW-RSA method.

8.3.2.3 Baseline Estimation and Denoising Using Sparsity (BEADS)

In some cases, chromatograms are not necessarily expected to be completely filled with peaks. Sparsely populated chromatograms feature a large number of datapoints that describe the baseline relative to the number of datapoints describing a peak, something that is exploited by the baseline estimation and denoising using sparsity (BEADS) algorithm as developed by Ning *et al.* [36, 37].

Conceptually, BEADS does not rely on highly restrictive models to describe the frequency components of a signal. Instead, the approach is based on breaking down the chromatographic signal into its basic components:

$$\mathbf{x} = \mathbf{s} + \mathbf{w} = \mathbf{y} + \mathbf{f} + \mathbf{w} \tag{8.5}$$

Here, $\mathbf{x}$ is the original input chromatogram, comprising a trace of peaks ($\mathbf{y}$), a baseline signal ($\mathbf{f}$), and white Gaussian noise ($\mathbf{w}$). Consequently, Eq. 8.5 implies that $\mathbf{s}$ describes the noise-free input chromatogram (i.e., $\mathbf{y} + \mathbf{f}$). Thus, in the BEADS approach, a peak vector is estimated ($\hat{\mathbf{y}}$) using a regularization approach analogous to the cost function shown in Eq. 8.1. In this case, the sum of squares is subjected to a high-pass filter (low frequency residuals are rejected), and the penalties are such that the pure chromatogram signal ($\mathbf{y}$) and its derivatives are sparse (meaning there is a lot a zero baseline between the peaks), and that the peaks are positive and not negative. The baseline estimate, $\hat{\mathbf{f}}$, is then expressed by Eq. 8.6.

$$\hat{\mathbf{f}} = \mathbf{L}\left(\mathbf{x} - \hat{\mathbf{y}}\right) \tag{8.6}$$

where $\mathbf{L}$ represents a low-pass filter.

The authors developed an optimization algorithm to efficiently search for the minimum. Readers interested in details associated with this concept are referred elsewhere [36]. In their concept paper, the authors compared BEADS to airPLS and backcor approaches applied to both real and simulated

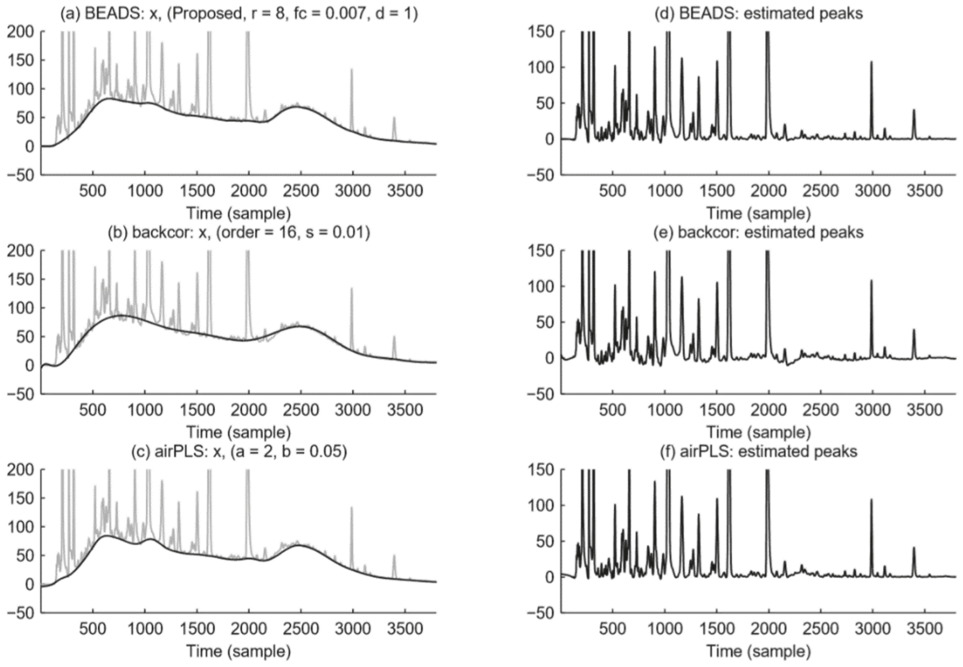


FIGURE 8.6 Comparison of background correction by BEADS (top), backcor (middle) and airPLS (bottom). The left-hand set of traces show the original signal (grey) and the modeled baseline (black), whereas the right-hand panels reflect the resulting background-corrected signal. Reprinted from *Chemometrics and Intelligent Laboratory Systems*, 139, X. Ning, I. Selesnick, L. Duval, Chromatogram baseline estimation and denoising using sparsity (BEADS), 156–167, Copyright (2014), with permission from Elsevier.

data as is also shown in Figure 8.6. Where airPLS and backcor were found to over- and underestimate the baseline, respectively, BEADS was found to be the most accurate method.

However, one disadvantage is that BEADS requires the baseline characteristics at the start to be similar to the baseline at the end. Indeed, in Figures 8.6A and D, the baseline returns to a value similar to that found at the start of the measurement. If this premise was not met, the modeled baseline was found to deviate in that its slope towards the final datapoint would be directed to the starting value of the first datapoint. While the baseline modeling does not require complicated parameters to be set, the parameters for the employed filters must be tailored (e.g., cut-off frequency and cost function parameters).

Small changes in these parameters were found to significantly impact the accuracy of the algorithm. However, while the authors admitted these flaws of the approach, they also noted that this made the algorithm conceptually flexible, as – with correct tweaking of the parameters – the algorithm could work for all types of data. Moreover, Navarro-Huerta *et al.* addressed most of these limitations in their assisted-BEADS algorithm [37], and Selesnick proposed solutions for the endpoint artifacts resulting from non-periodic signals [38].

8.3.2.4 Bayesian Approaches

We have seen for earlier methods that crowded chromatograms (i.e., large regions with co-eluting peaks) complicate background correction using most methods [39]. This is particularly true if the S/N is poor for a given peak of interest. To solve this, Lopatka *et al.* developed a method based on Bayesian statistics using a probabilistic peak-detection algorithm. Their peak-weighted (PW)

approach fits a number of different models through a set domain of datapoints using least squares. For each model, the probability of the datapoint belonging to a peak is then computed and expressed as weight vectors. The authors compared their algorithm to several approaches and demonstrated that the PW method performed particularly well for crowded chromatograms [39] as is shown by Figure 8.7.

The authors also applied their method to a comprehensive two-dimensional GC-FID chromatogram obtained from a separation of volatile components from fire debris, as shown in Figures 8.7 C–E, yet were unable to compare the performance of the PW method to other background correction methods due to the absence of accepted benchmark methods.

8.3.2.5 Background Correction Using Profile Spectra from Multi-Channel Detectors

When separations are coupled with mass spectrometric (MS) detection, the recorded spectra may be exploited to enable background correction. Erny and coworkers developed such an approach and applied this to CE-ToF-MS and UHPLC-QToF-MS data [40]. Their approach utilized full spectra rather than centroided spectra, the latter of which are known to merge overlapping peaks [41]. The authors favored this approach over other approaches as it facilitated improved acquisition of base-peak ions.

Prior to background correction, the authors first reduced the number of profiles as their typical dataset contained 141,000 profiles each containing 3,581 points. For selection, the authors removed all profiles with a certain number of non-zero values, as zero values indicate that no ion was detected at a given m/z interval. Consequently, when larger fractions of the mass spectrum contain non-zero values this is a strong indicator of background ions, and the authors used this criterion to reduce the number of profiles used for background correction to 37,000. For the actual background correction arPLS was used and no significant deviations in the total-ion chromatogram were observed, suggesting that no important information was removed.

For an elaborate approach such as this, the computational time was approximately 20 minutes for a 2.9 GB dataset. This makes clear the need for data reduction as a preprocessing step when working with MS data.

8.3.2.6 Dedicated Approaches Relevant to Comprehensive Two-Dimensional Chromatography

Several studies have focused specifically on development of background correction tools for comprehensive 2D chromatography. One approach utilized trilinear decomposition to remove the background drift from LC×LC-DAD data [42]. Using alternating trilinear decomposition (ATLD) to treat the raw dataset, the analytical signal component was separated from the background signal component. Parallel factor analysis (PARAFAC, see Section 8.6.2) and self-weighted alternating trilinear decomposition (SWATLD) have also been used for this purpose [42].

Given the fact that 2D chromatographic data are generally visualized on a plane, image-treatment software has also been applied. Reichenbach *et al*. applied such an approach to GC×GC data using various statistical and structural characteristics of the background from 2D chromatograms, including the white noise properties of noise in chromatographic signals [43]. Both the GC Image and LC Image commercial software packages employ this algorithm [44, 45]. A method by Zeng *et al*. utilizes linear least-squares curve fitting in combination with moving average smoothing to correct all one-dimensional peaks of the ^{2}D chromatograms [46].

A number of algorithms for background correction have been developed in recent years. However, unfortunately these algorithms are rarely compared and limited quantitative information about their performance has been published. As a consequence it is often difficult to discern which algorithm would be best for a particular application. A recent study thus focused on generating objective data for such numerical comparisons [47], and found that the performance of various background-correction algorithms depends largely on specific signal characteristics and are not as generally

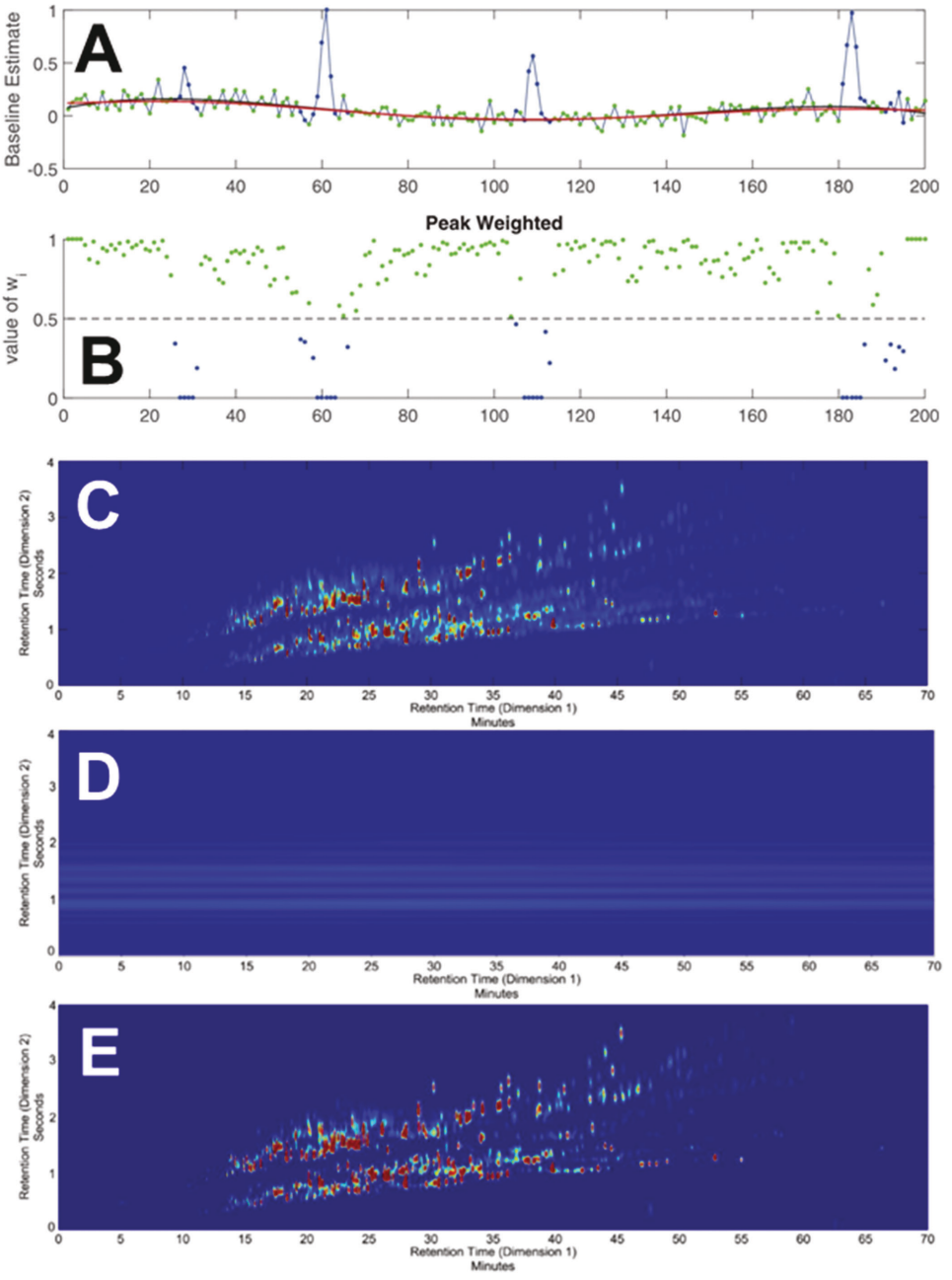


FIGURE 8.7 A) Application of the Bayesian peak-weighted approach for background correction demonstrated using a simulated chromatogram. B) Weight points indicating the magnitude with which each point contributed to the shape of the baseline; green points indicate a strong influence on baseline shape, C) Uncorrected GC×GC-FID separation of fire debris material, D) Peak-weighted estimate of the background, E) Corrected GC×GC-FID chromatogram after removing the background. Adapted from *Journal of Chromatography, A*, 1431, M. Lopatka, A. Barcaru, M. Sjerps, G. Vivó-Truyols, Leveraging probabilistic peak detection to estimate baseline drift in complex chromatographic samples, 122–130, Copyright (2016), with permission from Elsevier.

applicable as we would like; further study is needed in this area. It is thus not surprising that some approaches focus on combining the strengths of a number of tools. A case in point is the orthogonal background correction (OBGC) method developed by Filguiera *et al.* for use in 2D-LC [48]. The OBGC method exploits the fact that the ^{1}D chromatogram features a lower frequency of baseline fluctuations relative to those found in the ^{2}D chromatograms.

8.4 RETENTION-TIME ALIGNMENT

8.4.1 Introduction

When multiple chromatograms need to be compared, the next step after background correction is the alignment of retention axes. This is particularly important for two-dimensional LC where shifts in retention time are rather common. The actual alignment is generally carried out using either peak tables or the chromatograms themselves, often employing integrated peak-detection and peak-tracking algorithms.

There are two ways that retention time shifts can affect the analysis of comprehensive 2D-LC data. First, there can be shifts between the sequential ^{2}D chromatograms, such that it is difficult to determine whether two or more peaks that appear in adjacent ^{2}D separations are associated with the same compound *within* a single 2D-LC chromatogram. This can be addressed with either alignment algorithms, or directly within the peak detection method, as discussed in Section 8.5. Second, alignment can be used to address retention time shifts *between* chromatograms in order to confirm that peaks in multiple 2D chromatograms are associated with the same compound.

The complexity of algorithms that have been used for alignment varies from relatively simple local approaches including scalar-shift alignment and alignment of a selection of peaks, to global alignment, where multiple regions of the chromatogram are comprehensively aligned. In this section, we will mainly focus on the latter category, which also have found their application in forensics [49] and metabolomics [50].

In addition to the recent developments addressed below, a large number of other 2D approaches have been developed. One example is the algorithm using windowed rank minimization with interpolative stretching by Johnson *et al.* [51], which was applied to GC×GC data obtained for the analysis of naphthalene in jet fuel. Another approach by Pierce *et al.* employs indexing schemes for warping in both dimensions and was applied to GC×GC data [52]. Alignment based on images has also been extensively investigated and applied to comprehensive two-dimensional data for various applications [46, 53, 54]. While most approaches are suitable for three-way analysis, Allen and Rutan developed an algorithm for LC×LC-DAD with four-way data structures [2].

Attention has also been devoted to within-analysis retention shifts from modulation to modulation using PARAFAC in combination with PARAFAC2 [55] (see Section 8.6.2).

8.4.2 Correlation-Optimized Warping

One well-known approach for alignment is correlation-optimized warping (COW), where the chromatogram is divided into a number of local regions. Next, each section of chromatogram is compressed or stretched and compared to a reference until the correlation is maximized. The approached employs the Pearson correlation coefficient (PCC) defined as

$$PCC = \frac{\left(\boldsymbol{r}-\bar{r}\right)^{T}\left(\boldsymbol{x}-\bar{x}\right)}{\sqrt{\left(\boldsymbol{r}-\bar{r}\right)^{T}\left(\boldsymbol{r}-\bar{r}\right)\left(\boldsymbol{x}-\bar{x}\right)^{T}\left(\boldsymbol{x}-\bar{x}\right)}} \tag{8.7}$$

with r representing the reference, x the sample chromatogram, and $\bar{r}$ and $\bar{x}$ their mean values (i.e., average chromatograms), respectively. Interestingly, COW has also been expanded to support two-dimensional separations by Zhang et al. [56] and Gros et al. [57]. The latter method was recently successfully applied for alignment of GC×GC chromatograms based on high resolution mass spectrometry (HRMS) data [58].

Using a 2D chromatogram as reference for aligning regions of interests, van Mispelaar *et al.* proposed a correlation-optimized shifting through inner-product correlation for all selected regions in GC×GC chromatograms [59]. Paraster *et al.* introduced a bilinear peak-alignment method based on multivariate-curve resolution (MCR, see Section 8.6.1) and applied it for retention alignment of GC×GC data [60].

8.4.3 Automatic Time-Shift Alignment

Zheng *et al.* developed the automatic time-shift alignment (ATSA) protocol, which employs a two-stage alignment protocol [61]. After baseline correction by LMV-RSA (Section 8.3.2.2) and peak detection using multi-scale Gaussian smoothing (Section 8.5.1), the chromatogram is divided into a distinct number of segments. In the first stage of alignment, the authors opted to use the total peak correlation coefficient (TPC) as defined by

$$\mathrm{TPC} = \left(\frac{\sum_{i=1}^{I} w_i \mathrm{PCC}_i}{\sum_{i=1}^{I} w_i} \right) \frac{I}{n_{\mathrm{tot}}} \tag{8.8}$$

where w_i is the ratio between area and width (number of datapoints) of peak i, and I and n_{tot} are the total number of peaks in the sample and reference chromatogram, respectively. Segments that could not be aligned were labeled as outliers and subsequently realigned using PCC instead. Any convoluted or severed segments were corrected using a warping strategy to adjust the boundaries between segments.

The second stage of the protocol focused on precise alignment by again segmenting the preliminarily aligned chromatogram using the number of peaks. Here, boundaries between segments were located precisely in the middle between two chromatographic peaks. Next, each segment was aligned to the reference chromatogram. When no peak is available on the reference chromatogram, the algorithm was programmed to use an average time shift based on the other segments. The segments were reconnected using the principle of warping. In their study, the authors showed that ATSA was able to improve the correlation coefficient from 0.72 to 0.96 and eventually to 0.99 after the first and second stages of alignment, respectively.

ATSA and similar alignment tools require two parameters to be specified before use: the segment size and the initial time shift. While the authors showed that varying the segment size between 1 and 10 minutes had little effect, it was observed that larger sizes (particularly those above 10 minutes) would reduce consumption of computational capacity, yet resulted in erroneous time shifts. The size of the initial time shift, varied in the study between 0.1 and 1 min, was found to have little influence on the outcome.

Nevertheless, utilizing such an alignment tool is not without risk. The warping strategy applied by ATSA may influence peak areas, which may influence quantitation. While the authors did not find any evidence of changes in peak areas, application of ATSA on data obtained from a study focused on degradation of oils yielded a different conclusion [61]. Without ATSA, the data suggested that oil components were degrading, whereas after ATSA correction, the data indicated the opposite.

8.4.4 Alignment Using Mass Spectra

Similar to background correction, consulting mass spectra may also be fruitful for the purpose of retention alignment. One approach by Fu *et al.* is illustrated in Figure 8.8. After background correction by LMV, the algorithm calculates the PCC for each sample and reference peak which falls within a pre-specified time window (Figure 8.8A). In their study, the authors used 0.5 minutes as the time-shift window. Next, a correlation matrix was compiled using the resulting PCC values to establish a maximum-correlation path (Figure 8.8C). The green cells depict unaligned values, whereas the orange boxes represent corrected values. The correction is also apparent from the schematic representation shown in Figure 8.8D.

An underlying assumption that is made with most alignment algorithms is that the elution order is similar between samples. To account for dissimilar elution orders, the authors programmed the algorithm to specify landmark peaks, which are peaks with a correlation coefficient larger than 0.99. The time shifts of the found landmark peaks were then collected in a vector and outliers removed. Based on this vector, time shifts between two landmark peaks are linearly interpolated to calculate an expected time shift. The resulting value is compared to the earlier calculated time shift, and the peak is realigned in the event the difference is significantly larger.

For validation, the alignment tool was applied to GC-MS data comprising the characterization of plant samples [62]. The algorithm was assessed for its performance to correct the 15 most co-eluting peaks across a series of 30 samples. Figure 8.9A displays an overlay of the resulting 30 chromatograms. The extent of misalignment without correction is shown for the peaks within the highlighted box in Figure 8.9C. Here, all 30 chromatograms are displayed in series (y-axis) with the intensity represented by color. Retention-time alignment using the developed approach yielded aligned peaks as shown in Figure 8.9D (fully corrected chromatogram shown in Figure 8.9B).

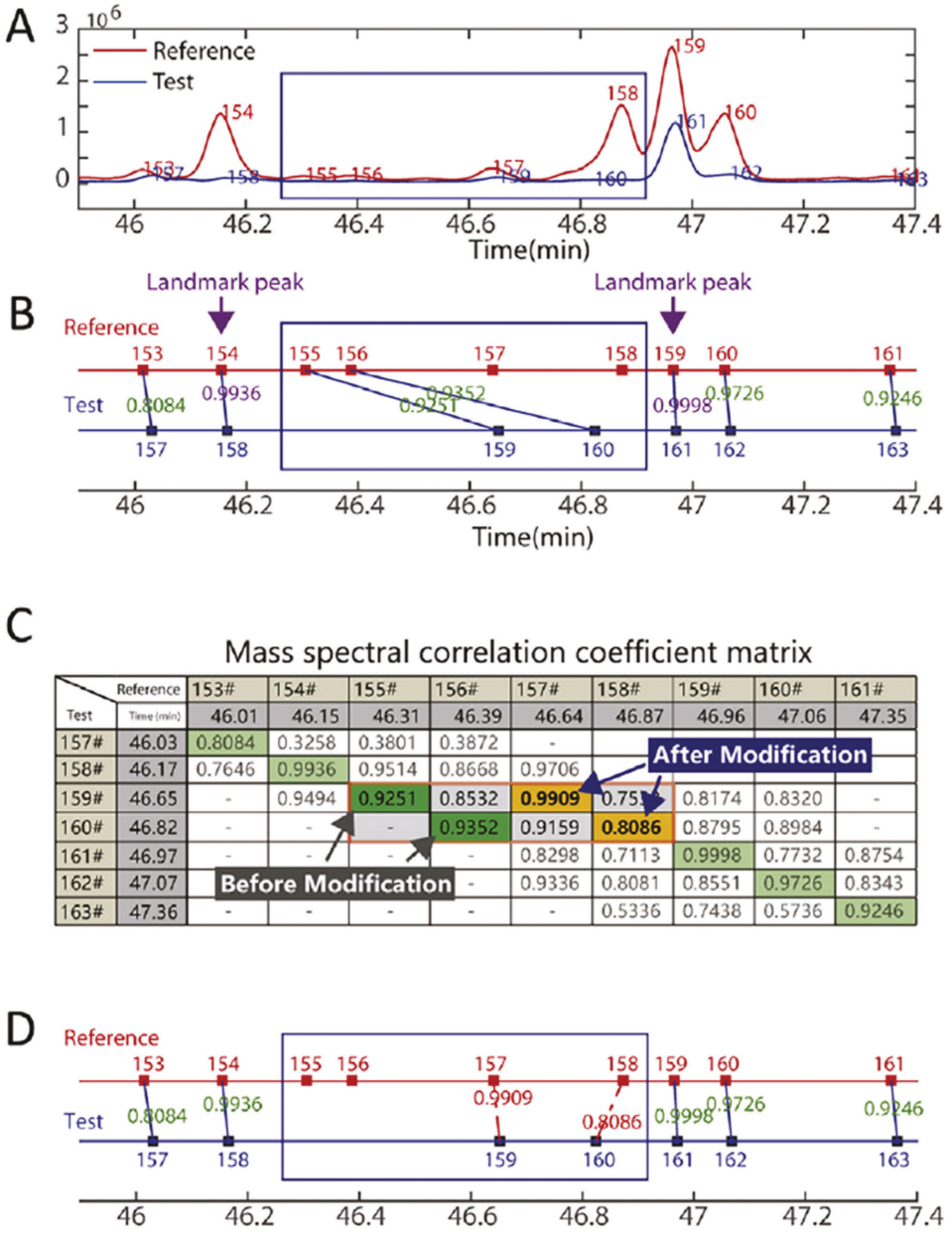
When the elution order is not expected to change, another method of interest establishes the chromatogram with the largest number of peaks as the reference chromatogram. After background correction and peak detection using automated peak detection and baseline correction (ACPD-BCP, see Section 8.5.3), a rough alignment was carried out in a way similar to the COW approach [63]. In this case, however, a cosine correlation was computed instead of the PCC. The cosine correlation is a measure of the similarity between two vectors of a product space, by calculating the cosine between the two vectors. Such a metric is often used in pattern recognition algorithms [64].

After preliminary alignment, the next stage utilized the relative distances between a particular peak found in a sample chromatogram compared to the reference chromatogram, as well as the cosine values and absolute distances. The resulting differences were collected in an alignment table. However, no actual information on robustness of the algorithm was shown, nor was the algorithm compared to alternative approaches.

8.5 PEAK DETECTION

After completing initial data preprocessing of 2D-LC data, we can proceed to locate all true chromatographic peaks. In signal processing, this process is referred to as peak detection. Similar to data preprocessing, peak detection of higher-order data, such as our two-dimensional chromatograms, relies on lower-order data processing techniques. While direct two-dimensional peak detection is possible, the limited number of data points in the first dimension typically seriously limit this possibility, leaving chemometricians with no choice but to use one-dimensional peak detection for all ^{2}D separations, after which the resulting peaks are clustered into 2D peaks. Thus, again, most of the following paragraphs will concern one-dimensional approaches to peak detection.

Once a peak has been located, determination of a number of elementary characteristics of the peak may be useful. Examples include the peak area, retention time and asymmetry. These and



Test \ Reference	Time (min)	153#	154#	155#	156#	157#	158#	159#	160#	161#
		46.01	46.15	46.31	46.39	46.64	46.87	46.96	47.06	47.35
157#	46.03	0.8084	0.3258	0.3801	0.3872	-				
158#	46.17	0.7646	0.9936	0.9514	0.8668	0.9706				
159#	46.65	-	0.9494	0.9251	0.8532	**0.9909**	0.751	0.8174	0.8320	-
160#	46.82	-	-	-	0.9352	0.9159	**0.8086**	0.8795	0.8984	-
161#	46.97	-	-	-	-	0.8298	0.7113	0.9998	0.7732	0.8754
162#	47.07	-	-	-	-	0.9336	0.8081	0.8551	0.9726	0.8343
163#	47.36	-	-	-	-	-	0.5336	0.7438	0.5736	0.9246

FIGURE 8.8 Peak alignment using the maximum-correlation path and landmark peaks. A) Selected segment from chromatogram, B) misalignment resulting from exclusively consulting mass spectra, C) locations of misaligned peaks in the maximum correlation coefficient path. D) Schematic representation of reference chromatogram and aligned peaks of a test chromatogram after correction, where the x-axis is time in minutes, and the points are peaks associated with peak numbers. Reprinted from *Journal of Chromatography, A*, 1513, H. Fu, Y. Zhang, L. Zhang, J. Song, P. Lu, Q. Zheng, P. Liu, Q. Chen, B. Wang, X. Wang, L. Han, Y. Yu, Mass-spectra-based peak alignment for automatic nontargeted metabolic profiling analysis for biomarker screening in plant samples, 201–209, Copyright (2017), with permission from Elsevier.

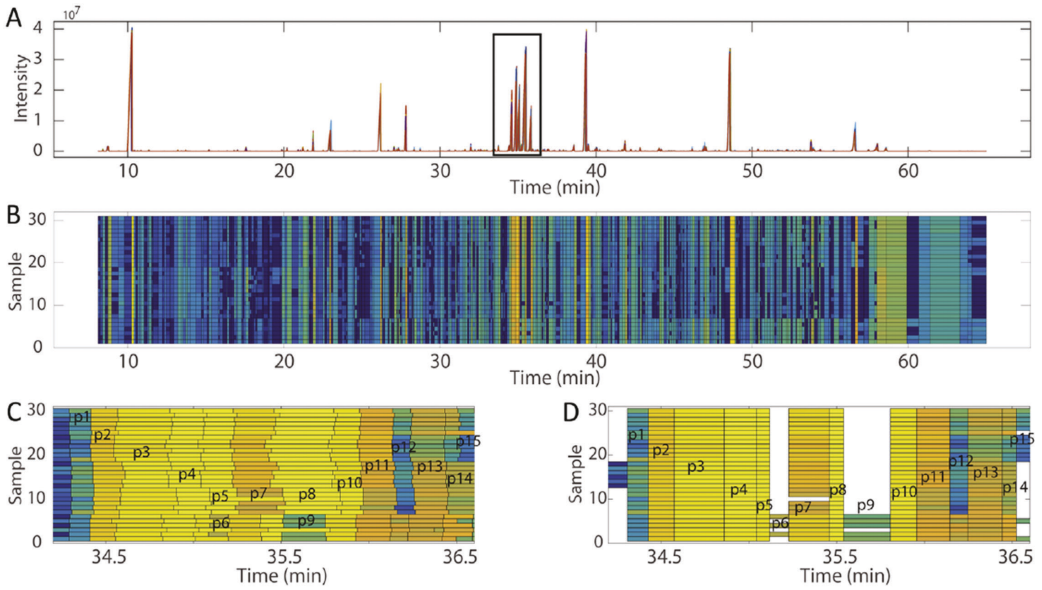


FIGURE 8.9 Illustration of peak alignment using mass spectra. A) Original chromatogram – highlighted box depicts dense region of peaks. B) Alignment of all chromatograms (time, x-axis) for all 30 samples (y-axis). Color depicts signal intensity. Panels C and D show the peak structure for the dense region with C) original chromatogram, D) corrected chromatograms. Adapted from *Journal of Chromatography, A*, 1513, H. Fu, Y. Zhang, L. Zhang, J. Song, P. Lu, Q. Zheng, P. Liu, Q. Chen, B. Wang, X. Wang, L. Han, Y. Yu, Mass-spectra-based peak alignment for automatic nontargeted metabolic profiling analysis for biomarker screening in plant samples, 201–209, Copyright (2017), with permission from Elsevier.

other characteristics are often automatically calculated by the software supplied with the instrument. Generally, the algorithms define the peak start and end points as boundaries, although the user can typically adjust these graphically. A default approach in many instrument software packages is the perpendicular drop method for integrating overlapped peaks. In the case of even moderate coelution, this rather simplistic approach leads to erroneous results, as is illustrated schematically in Figure 8.10.

For a more accurate determination of the various properties of a well-separated peak, the statistical moments may be used [66] (Table 8.1). In Eqs. 8.9–8.13 x_i are time points in the chromatogram, $f(x_i)$ is the signal value at time x_i, and the index i runs from the start to the stop points of the peak. When using curve fitting for peak detection, the function $f(t)$ can be replaced by the model used for fitting. When no model is available, the solution can be numerically determined. The accuracy of the statistical moments relies heavily on the preprocessing and sampling frequency of the detector [67–69].

8.5.1 Classical Peak Detection

Peak detection is generally performed using either a derivative-based approach or a curve-fitting approach, although other methods such as tuned or matched filters exist [15]. When peaks are well-resolved, the noise is minimal or adequately removed by preprocessing, and the background is reasonably constant, derivative methods, which are usually incorporated into chromatographic data system software, work well. However, when overlapped peaks are present, derivative detection methods are often inadequate. Figure 8.11 illustrates some of these challenges [65]. In Figure 8.11

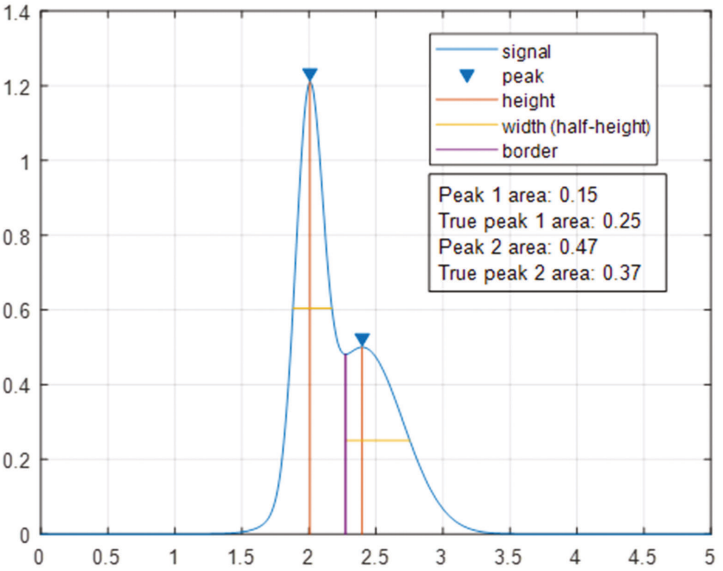


FIGURE 8.10 Illustration of an outcome of simple, automated peak integration as encountered with standard data-analysis software. Inset provides a comparison of the results with the true values for a case with overlapping peaks 1 (left) and 2 (right).

Source: Reproduced from [65].

TABLE 8.1
Overview of Statistical Moments

Moment ordinal	Property	Formula	Eq.
0 (m_0)	Area	$\sum_{i=1}^{N} f(x_i)$	(8.9)
1 (m_1)	Retention time	$\frac{\sum_{i=1}^{N} x_i f(x_i)}{m_0}$	(8.10)
2(μ_2)	Variance (σ^2)	$\sum_{i=1}^{N} (x_i - m_1)^2 f(x_i)$	(8.11)
3 ($\widetilde{\mu_3}$)	Skewness	$\frac{\sum_{i=1}^{N} (x_i - m_1)^3 f(x_i)}{\mu_2{}^3}$	(8.12)
4 ($\widetilde{\mu_4}$)	Kurtosis	$\frac{\sum_{i=1}^{N} (x_i - m_1)^4 f(x_i)}{\mu_2{}^2}$	(8.13)

Source: [70].

A, a signal containing two overlapped peaks with no noise or baseline drift is shown. If we take the second derivative of this signal, we obtain Figure 8.11C. The new signal shows maxima that reflect the presence of inflection points in the chromatogram. More importantly, the minima indicate the presence of peak apexes in the original chromatogram.

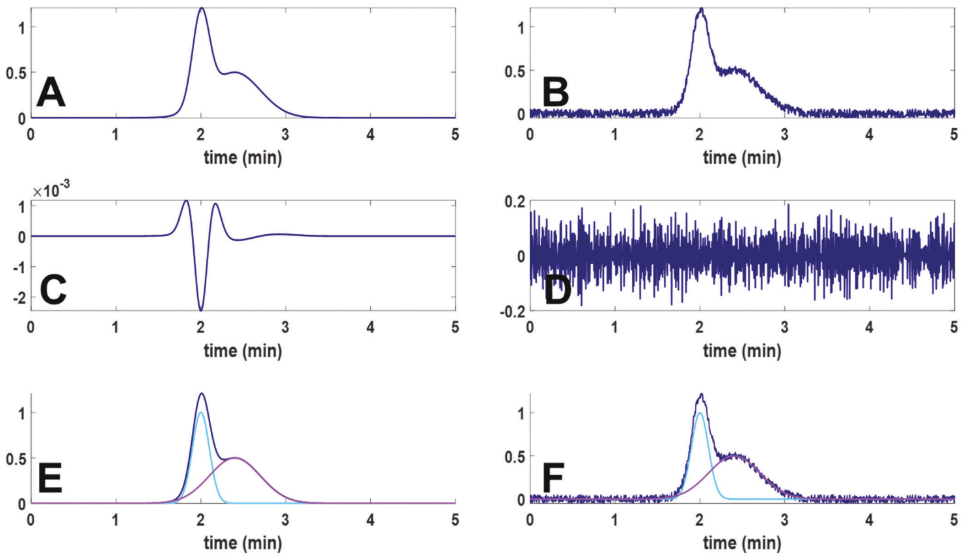


FIGURE 8.11 A) Ideal (noise-free) signal containing two overlapped peaks. B) Equivalent of signal in (A) but superimposed with high frequency noise. C) Second derivative of (A) where a minimum indicates the apex of a peak, D) Second derivative of (B) where the extreme number of peaks induced by the noise, yields an equal number of minima, again yielding noise. E) and F): two curves fitted to (A) and (B), respectively. Source: Reproduced from [71].

Now we consider the noisy chromatogram shown in Figure 8.11B. While the signal to noise ratio and resolution between the two peaks is sufficient to observe two peaks, the second derivative only amplifies the high-frequency noise. To make things worse, our two peaks have now completely disappeared from the derivative signal. From this simple comparison, we can immediately draw a number of conclusions. First, derivative-based peak detection approaches require significant data preprocessing to remove most of the noise. From Section 8.3.1, we also know that too extensive noise filtering is likely to remove useful information. Second, we see that the peaks must be sufficiently resolved, as the resolution in Figure 8.11A, and thus 8.11C, is just sufficient to detect the second peak.

The curve-fitting approach to peak detection suffers much less from the challenges discussed above. Here, a section of the chromatogram, such as the ones in Figures 8.11A/B, is taken and x number of curves as defined by the selected model are fit to the chromatogram. Classically, a Gaussian function is used for fitting, but that has the disadvantage that the algorithm may completely miss non-Gaussian shaped peaks. Indeed, peaks in LC often comprise a tailing component in the distribution function. Consequently, it is often difficult to find a distribution function which describes the peak accurately.

Arguably the biggest challenge in peak detection for 2D-LC, however, is that the algorithm must determine the number of peaks present in the overlapped section of the chromatogram consistently across all modulations in the LC×LC space. Suppose the signal shown in Figure 8.11A actually contains a third, smaller peak buried underneath the other two. For curve-fitting approach to pick this up, it must know *a priori* that three peaks are present. The exercise of determining the correct number of peaks is paradoxically the whole aim of the curve-fitting approach, rendering curve fitting very challenging for chromatograms containing regions with a high degree of peak overlap. Nevertheless, as is shown in Figures 8.11E and F, once the correct number of peaks is known, curve fitting is very robust, even for moderately noisy data.

Most recent developments in peak detection for 1D chromatography have focused on improving either of these two generic peak-detection strategies. Due to the susceptibility of the algorithms to noise, many of these approaches utilize integrated preprocessing algorithms that remove some of the noise prior to the peak detection step.

One example of an integrated approach is the smoothing-based peak-detection method, such as the multi-scale Gaussian-smoothing algorithm developed by Fu *et al.* [72], which operates in three steps. The first and second steps involve removal of background drift and subsequent detection of all local maxima. The key step involves application of a smoothing filter to the signal with various window sizes of the filter. Working on the assumption that true peaks retain a constant location of their maxima after smoothing, noise peaks are automatically removed. By varying the width of the filter, the intensity of the peak can be assessed. This renders the algorithm more robust against noise and baseline drift than classical derivative-based peak detection which tends to be more sensitive to such factors although this highly depends on the specific applications.

8.5.2 Continuous Wavelet Transformation (CWT)

The critical dependence of curve-fitting approaches on the determination of the number of overlapped components (see Section 8.5.1) has also received attention. One example is the development of wavelet-transform based peak detection, where Peters *et al.* applied cross validation to estimate the number of components [73]. Another challenge for curve-fitting approaches is robustness against the presence of a large variation in characteristics between neighboring peaks, and a number of different wavelet morphologies have been suggested [74], including the continuous-wavelet-transform (CWT) approach. CWT is more sensitive to peak characteristics such as symmetry, yielding fewer false positives than classical derivative-based approaches [75, 76] and the multi-scale Gaussian-smoothing approach [77]. CWT has also been incorporated with ridge-detection algorithms [78], which locate peaks using local maxima [74].

Nevertheless, the CWT method is not without flaws. This method produces chromatograms analogous to second derivative analysis, so the focus is more on peak detection rather than profile determination, although some studies have shown that areas can be extracted and calibration curves can be obtained [79, 80]. The CWT needs to be optimized in terms of the selection of the appropriate wavelet function as well as the scale factor. These weaknesses have been addressed using a heuristic and recursive approach that resulted in improved peak detection, as well as determination of peak characteristics, such as area [81]. Despite developments for both CWT and Gaussian-smoothing, both approaches struggle with cases where coelution is severe [72, 82].

An approach completely different from the above-mentioned methods employs Bayesian statistics. Unlike the previous techniques, which yield a binary answer (true or false) for the detection of a peak at a given datapoint, Bayesian methods employ probabilities. While initial implementations struggled with overlapping peaks [83], statistical-overlap theory [84] was incorporated to improve this [85]. This is a characteristic advantage of Bayesian statistics, in that these methods can incorporate prior knowledge. Woldebriel *et al.* further developed a probabilistic model to allow untargeted peak detection of LC-MS without requiring any preprocessing [86], and thus reducing the risk of accidentally destroying information.

8.5.3 Automatic Peak Detection and Background Drift Correction

Another example of background drift correction combined with peak detection is the algorithm developed by Yu *et al.* [76]. This automatic peak detection and background drift correction (ACPD-BDC) approach focuses on start and end points of peaks. A datapoint x_i was considered a starting point if its value was lower than the next three points x_{i+1} to x_{i+3}. Similarly, the end point x_j must be larger than the next three points, as is also reflected by Eqs. 8.14a/b.

$$x_i < x_{i+1} < x_{i+2} < x_{i+3} \tag{8.14a}$$

$$x_j > x_{j+1} > x_{j+2} > x_{j+3} \tag{8.14b}$$

The list of start and end points was stored in two vectors; the linear combination of these vectors represents the list of peak elution ranges. Next, all detected peak regions were subtracted from the chromatogram ($\mathbf{x}$), yielding the background ($\mathbf{x}_{\text{filtered}}$) as a result. By taking the derivative of this background signal ($d\mathbf{x}_{\text{filtered}}$), outliers can be detected and removed using Eq. 8.15.

$$\frac{\left| dx_{\text{filtered},i} - \overline{d\mathbf{x}_{\text{filtered}}} \right|}{\sigma} > 3 \tag{8.15}$$

Here, σ is the standard deviation of the $d\mathbf{x}_{\text{filtered}}$ vector. By iteratively removing outliers, the noise level – which is the first-order derivative of $d\mathbf{x}_{\text{filtered}}$ – is fine-tuned.

The removed regions that contain peaks in $\mathbf{x}_{filtered}$ are then linearly interpolated, thus constructing $\mathbf{x}_{background}$. This signal is then filtered again using a moving-average filter with a width of 3 points. Meanwhile, the first- and second-order derivatives of the original signal (x) are compared. In this approach, peaks are only considered a true peak if 1) $|dx_i|$ is five times larger than 3σ, and 2) the second-order derivative crosses zero fewer than eight times. Finally, the background, $\mathbf{x}_{background}$, is subtracted from $\mathbf{x}$, yielding a background-corrected chromatogram with peaks detected.

The authors of this elaborate approach compared their ACPD-BDC algorithm with airPLS and MairPLS, and applied the three methods to: 1) simulated data; 2) a GC separation of a plant-based flavor extract; and 3) a LC separation of pharmaceuticals in water. MairPLS and ACPD-BDC were found to perform better than airPLS (Figure 8.12) [76].

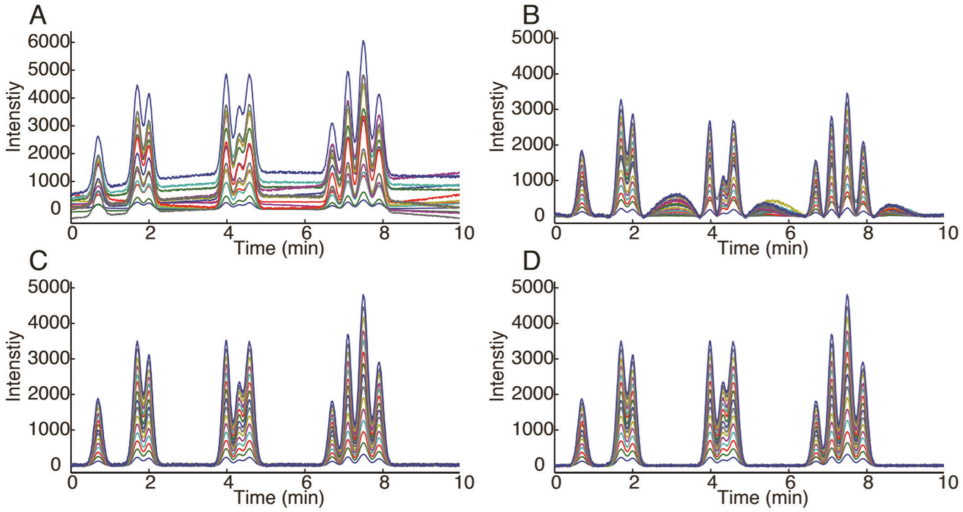


FIGURE 8.12 Comparison of three background correction and peak detection approaches. A) Raw data, B) airPLS, C) MairPLS, and D) ACPD-BDC. Reprinted from *Journal of Chromatography, A*, 1359, Y. Yu, Q. Xia, S. Wang, B. Wang, F. Xie, X. Zhang, Y. Ma, H. Wu, Chemometric strategy for automatic chromatographic peak detection and background drift correction in chromatographic data, 262–270, Copyright (2014), with permission from Elsevier.

8.5.4 Comprehensive Two-Dimensional Approaches

For single-channel detectors, two categories of approaches have been used for peak detection in comprehensive 2D chromatography. The first category detects peaks on the 1D signal and subsequently clusters them into 2D peaks [87]. The second category employs the image-based watershed algorithm [45].

Difficulties in peak detection mainly arise from the variability in the second dimension times from one modulation to another in LC×LC, data but most importantly undersampling in the first dimension (Chapter 3). A robust algorithm must be able to handle situations where the degree of undersampling varies within a single LC×LC chromatogram. Shifts in the retention of individual compounds in adjacent ^{2}D separations are also challenging. These shifts can result from either natural variation in retention (i.e., retention precision is not perfect), or deliberate changes in ^{2}D conditions, as in the case when using shifting gradient programs (see Section 4.5).

In addition, multi-way methodologies have been developed for peak detection, specifically for datasets recorded using multi-channel detectors. These are discussed in Section 8.6.

8.5.4.1 Two-Step Peak Detection Using Peak Clustering

One method of peak detection was introduced by Peters *et al.* and utilizes two steps [87]. First, 1D peak detection algorithms, such as described in the previous sections, are used to for peak detection across the entire string of concatenated ^{2}D chromatograms. In their study, Peters *et al.* used the Savitzky-Golay method to detect the 1D peaks based on their derivatives (see Section 8.3.1). Indeed, this strategy allows the determination of various peak properties, such as the peak height, as well as the start and end points.

In the second step, an algorithm is applied to merge all 1D peaks across different modulations belonging to the same compound [87]. This process is generally referred to as peak clustering or peak merging, and remains a difficult challenge in data analysis for two-dimensional LC. The clustering algorithm associates the peaks belonging to the same compound together using retention-time alignment (see Section 8.4), but since shifting gradients deliberately induce substantial retention time shifts, merging peaks properly when shifting gradients are used is particularly challenging.

For the actual merging, the algorithm employs overlap and unimodality criteria to determine whether the 1D peaks belong together. The overlap criterion assesses the degree of overlap between two regions (a and b) in which the peak elutes in adjacent modulations (Figure 8.13). The ratio of overlap is computed by Eq. 8.16 and a threshold must be set to determine which peaks will be merged. The unimodality criterion is used to determine whether ^{2}D maxima belong to the same peak and investigates the maxima of the peak profile in the first dimension.

$$OV = \frac{b}{a} \cdot 100\% \tag{8.16}$$

For determination of the peak area, Peters *et al.* used a trapezoidal method, essentially summing the areas of peaks in adjacent ^{2}D separations that are associated with elution of the same compound [87]. This method has been applied to LC×LC data [88, 89]. In order to allow the method to profit from four-way data, which was demonstrated to be needed by Bailey and Rutan [90], Vivó-Truyols developed a Bayesian two-step approach [83].

An approach that is analogous to the approach described by Peters *et al.* is the well-known msPeak algorithm that utilizes the normal-exponential-Bernoulli (NEB) model to describe peaks [91]. This method includes a means for resolving overlapped peaks in the second dimension. The approach combines preprocessing with a scan for regions with co-eluting peaks. This approach was recently improved with the development of the normal-gamma-Bernoulli (NGB) model, which, unlike the NEB model, does not have an analytical solution [92]. The authors demonstrated, however, that their

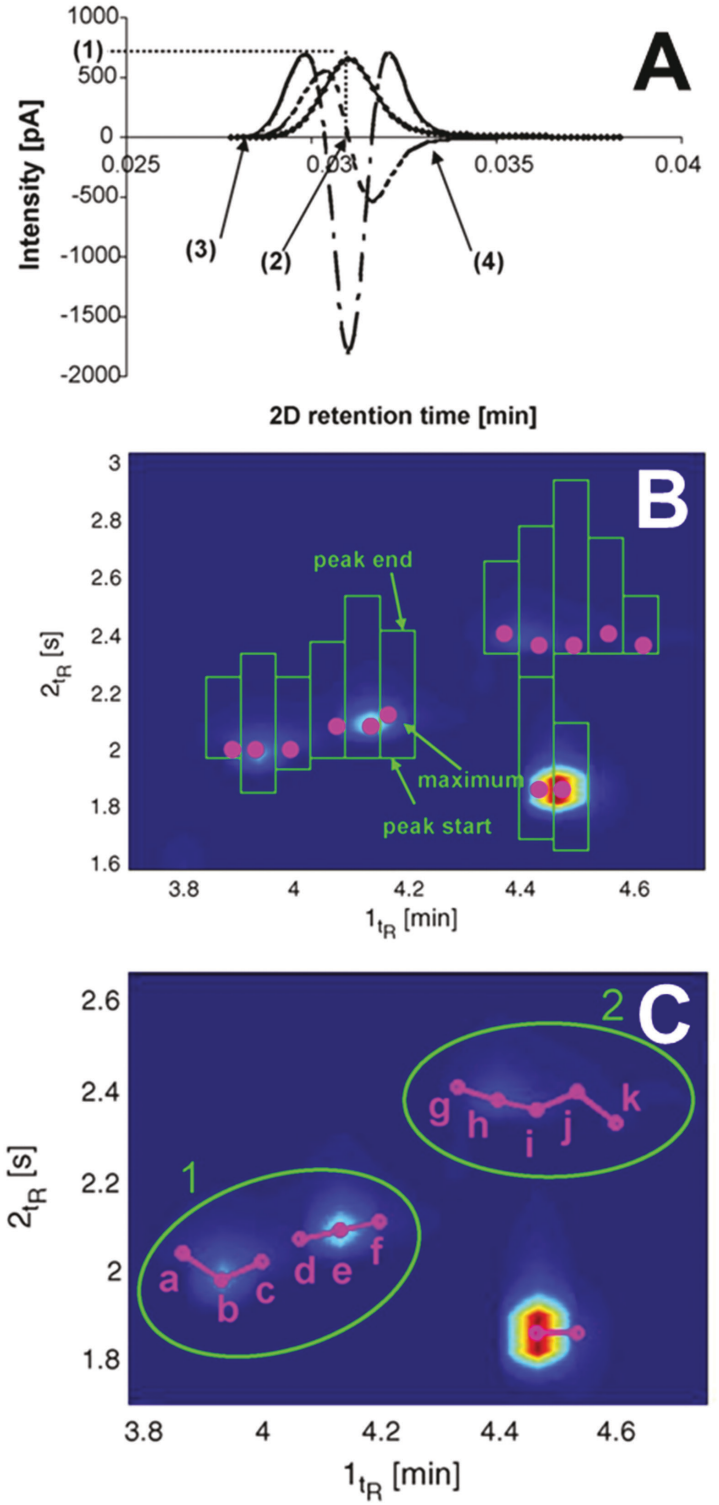


FIGURE 8.13 Illustration of the peak detection in LC×LC by first employing 1D peak detection methods on 1D data, and then clustering related ^{2}D peaks to form 2D peaks. A) Peak detection using derivative signal processing on 1D data – solid line, original signal; dashed line, first derivatives; dashed-dot line, second

FIGURE 8.13 (Continued)
derivatives. Indicated characteristics are (1) peak maximum; (2) ^{2}D retention time; (3) peak start; and (4) peak end. B) Locations of detected peaks in 2D chromatograms. The purple points in series belong to the same peak and peak-clustering algorithms aim to connect these correctly for each analyte. C) Clustered peaks as depicted by the connected purple dots. Adapted from *Journal of Chromatography, A*, 1156, S. Peters, G. Vivó-Truyols, P. Marriott, P. Schoenmakers, Development of an algorithm for peak detection in comprehensive two-dimensional chromatography, 14–24, Copyright (2007), with permission from Elsevier.

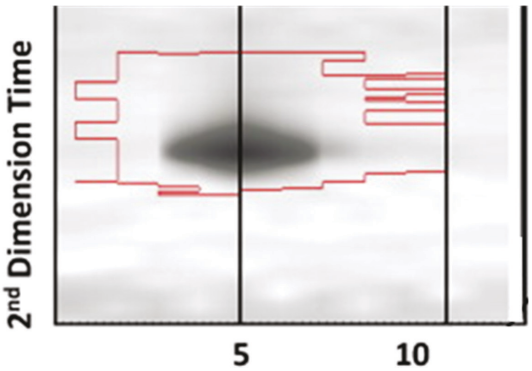


FIGURE 8.14 Contour plot of a peak after background correction using LC Image software [90]. The (red) line is the peak boundary as determined by the LC Image software via the watershed algorithm. Adapted from *Journal of Chromatography, A*, 1218, H. Bailey, S. Rutan, P. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, 8411–8422, Copyright (2011), with permission from Elsevier.

NGB model yielded better fits of the peaks and an improved true positive rate for detection of chromatographic peaks with low total ion currents. Both models were applied to GC×GC-ToF datasets, where the NGB model was shown to find more true positives.

8.5.4.2 Watershed Algorithm

A completely different approach relies on processing the two-dimensional chromatogram as a surface. The inverted watershed algorithm leverages this concept and defines the boundaries of peaks using the topology of the surface [93]. This can be imagined as viewing the chromatogram as a mountain landscape, turning it upside down and filling it with water until the different peak maxima can no longer be distinguished. The algorithm will continue this process until it reaches the background signal, thus rendering it vulnerable to noise or artifacts. Figure 8.14 shows an LC×LC peak, with the boundaries as identified by the watershed algorithm. While the watershed algorithm is routinely applied (e.g., [94]), and also is available in commercial applications (e.g., LC Image software), it has been shown to be prone to erroneous results when the modulations are not correctly aligned and the noise levels are high [95]. Figure 8.14 illustrates this point, where the peaks for the outer modulations tend to appear and disappear in a single ^{2}D chromatogram. This issue was addressed by Latha *et al.* in a study which applied skew correction to improve the watershed algorithm [96]. In a comparison, the authors concluded that the improved watershed algorithm outperformed the two-step algorithm. In any case, both methods are known to be sensitive to large degrees of coelution and noise levels, and improved methods are needed.

Another commercially available program for peak detection in comprehensive 2D-LC, ChromSquare, is available and can be used in conjunction with a Shimadzu data system. However,

details related to the algorithm(s) used in this software have not been published to the best of our knowledge.

8.6 MULTI-WAY APPROACHES

Most of the methods discussed above treat single-channel data (e.g., flame ionization detection in GC) or one channel in the case of multi-channel detectors (e.g., total ion chromatogram in MS detection). In contrast, multi-way analysis leverages all of the available data produced by multi-channel detectors. In general, these methods rely on modeling the data as a sum of linearly independent (i.e., not correlated) components, where ideally these components correspond directly to the chromatographic and spectroscopic signatures of real chemical species. These methods often treat the background contributions as additional components, which removes the need for an independent background subtraction algorithm. Examples include multivariate curve resolution-alternating least squares (MCR-ALS) and parallel factor analysis (PARAFAC) and related methods.

8.6.1 Multivariate Curve Resolution-Alternating Least Squares

Multivariate curve resolution-alternating least squares (MCR-ALS) is a tool for resolving overlapped signals (i.e., overlapped peaks, or peaks overlapped with background signals) resulting from a wide range of analytical measurements, and numerous chromatographic applications have been reported [97,98]. In this section we introduce the basics of MCR-ALS for analysis of single 1D chromatograms with full spectrum detection (i.e., diode array or mass spectrometry), and then extend the concept to show how single or multiple 2D-LC chromatograms can be analyzed. Some specific advantages of this approach are that often extensive preprocessing of the data is not required, and that when multiple chromatograms are analyzed simultaneously, quantitative results can be obtained as a direct outcome of the algorithm.

MCR decomposes a data matrix (**X**) into a set of chromatographic components (**C**), a corresponding set of spectral components (**S**) and the error (**E**) which ideally only contains the noise.

$$\mathbf{X} = \mathbf{C} \cdot \mathbf{S}^T + \mathbf{E} \tag{8.19}$$

The matrix **X** (*R* x *S*) consists of *R* rows, corresponding to *R* chromatographic time points and *S* columns, corresponding to *S* wavelengths or mass channels and is an LC-DAD or LC-MS chromatogram generally containing contributions from multiple chemical species. Upon application of the algorithm, the matrix **C** (*R* x *N*) contains estimates for the "pure" component chromatograms corresponding to *N* components. Each of these *N* components may be directly associated with a chemical compound or may be associated with instrumental contributions to the signal such as background (e.g., solvent impurities). The recovered matrix **S** (*S* x *N*) consists of *N* spectra of these same components.

The implementation of the algorithm requires that initial estimates for each component (either the chromatograms or spectra; here the illustration is for initial estimates for the spectra, **S**) are made. These can be obtained from the data itself using a method such as principal components analysis (PCA) [99, 100], and/or *a priori* knowledge of the component spectra (e.g., from a DAD or MS spectral library). Several other approaches have been developed to obtain the initial estimates, including key-set factor analysis [101], self-modeling [102], and orthogonal projection [103].

Using ALS, equation (8.19) is then iteratively optimized and solved as

$$\mathbf{C} = \mathbf{X}\mathbf{S} \cdot \left(\mathbf{S}^T\mathbf{S}\right)^{-1} \tag{8.20a}$$

$$\mathbf{S}^T = \left(\mathbf{C}^T\mathbf{C}\right)^{-1} \cdot \mathbf{C}^T\mathbf{X} \tag{8.20b}$$

The -1 superscript indicates a matrix inverse; Eqs. 8.20a/b are least squares solutions for **C** and **S**, respectively.

A key aspect of MCR-ALS that makes it powerful is the possibility to apply chemically meaningful constraints during the optimization process. These constraints can include non-negativity (i.e., analyte concentrations should not be negative), unimodality (i.e., well-behaved chromatographic peaks are singlets), and/or predefined elution profiles and spectra [104]. These constraints can be applied to the chromatograms and/or spectral profiles, and to one or more of the individual components, and make it more likely that the algorithm converges to a chemically reasonable solution.

The MCR-ALS method can be extended to resolve overlapped peaks in comprehensive 2D-LC as well. Figure 8.15A shows the format of the matrix **X** for a single LC-DAD (or LC-MS) chromatogram. Figure 8.15B shows the format of the array when the multiple modulations of subsequent chromatograms, i.e., ^{2}D are stacked together, and subsequently unfolded, as discussed above. This unfolded form of the **X** matrix is analyzed as described above, recognizing that the resulting

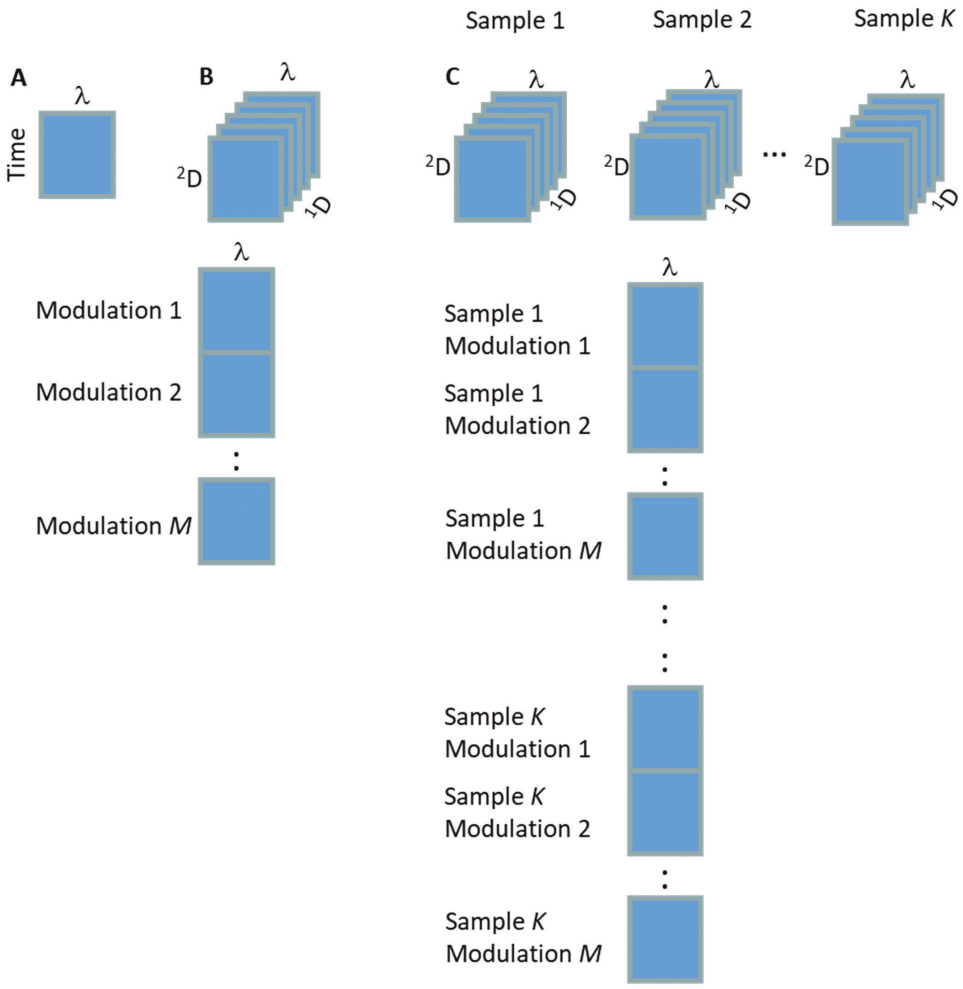


FIGURE 8.15 Schematic illustrating extension of the MCR-ALS approach to LC×LC separations and multiple samples. A) Data structure for LC-DAD or LC-MS data; B) Data structure for LC×LC-DAD or LC ×LC-MS data; C) Data structure for LC×LC-DAD or LC×LC-MS data for multiple samples.

resolved chromatograms contained in **C** will also be in this unfolded format. Finally, it is advantageous to analyze multiple samples simultaneously, as shown in Figure 8.15C. In this case the **X** matrix now contains all the ^{2}D chromatograms for all samples appended end to end; again, the resulting **C** matrix will have this same structure. A schematic of the curve resolution results from the data structure shown in Figure 8.15C is provided in Figure 8.16.

Inherent in the MCR-ALS algorithm is the requirement that the pure component spectra present in the matrix **S** must be consistent across all modulations and samples; this requirement is referred to in mathematical terms as bilinearity. This offers a degree of smoothing because there are multiple instances of the spectra across the data set. Additionally, if the *K* samples treated simultaneously consist of both standards and unknown samples, then areas under the resolved chromatograms can be used to construct "pseudo-univariate" calibration curves and provide quantitative results for the unknown samples, as discussed by Olivieri [105].

One challenge with LC×LC-DAD data is that the sample is continuously diluted during the analysis procedure, such that by the time the analytes enter the detector upon exiting the ^{2}D column, significant dilution has occurred. Cook *et al.* proposed incorporating a DAD column at the exit of the ^{1}D, when the sample has not been as diluted, and, hence, the S/N is much larger, and combining these data with the data from the ^{2}D detector [106]. MCR-ALS was then used to resolve the overlapped peaks.

A few comments on the implementation of MCR-ALS are in order here. First, as commonly implemented, the chromatogram is typically analyzed in small segments containing less than 10 components, making this approach somewhat tedious, although some progress has been made in automating this procedure [107, 108]. Second, the correct number of components must be chosen to achieve successful resolution; this is not always straightforward, and has proven to be difficult to automate in a way that is, according to the authors, robust. Additionally, the manual intervention required has so far prevented widespread implementation in chromatographic data system software, which presents a large barrier to use of these approaches by researchers who are not expert chemometricians. However, there are available toolboxes that function within the Matlab computing environment from Tauler *et al.* [109, 110] and Olivieri *et al.* [111], as well as a commercially available toolbox (PLS Toolbox) from Eigenvector [112].

While the above discussion focused primarily on DAD detector data as the basis for the curve resolution, MS detection can be advantageous because of the improved selectivity and structural information available from mass spectral data. The emergence of hyphenated LC-MS has been accompanied by a significant increase in dataset size for a single experiment. Depending on whether MS is carried out in tandem (i.e., MS/MS), at high resolution (e.g., TOF, ICR), and the chromatographic analysis time, datasets can easily reach up to 80 GB per analysis. One aim in data analysis has been the compression of the data to a manageable size. A conventional data reduction approach is binning, where the m/z axis is separated into segments, usually to unit resolution. Numerous applications of MCR-ALS to these types of LC-MS data have been reported [113–115]. Of course, this reduction is also accompanied by a loss of mass resolution, and this can result in multiple compounds being represented within a single mass data point.

Another method that has been proposed to address the large volumes of data resulting from high resolution LC-MS experiments is the region-of-interest (ROI) strategy [116]. Here, data regions with a high information density are selected using criteria such as signal intensity and the fact that peaks are typically adjacent to "data voids". This strategy allowed for peak detection without loss of mass resolution [117].

The ROI approach has been combined with MCR-ALS to provide for resolution of overlapped peaks [118]. A protocol has been reported to carry out this procedure, which includes all steps from the export of the data from the chromatographic data system to complete resolution of a set of metabolomic data [119]. The ROI-MCR-ALS method has been applied to LC×LC-HRMS data of the rice metabolome by Navarro-Reigh *et al.* [120]. The authors used ROIs that were selected using S/N, mass accuracy, and the minimum number of subsequent occurrences of the same m/z. In addition, the authors applied wavelet compression [121, 122] to further compress the data up to

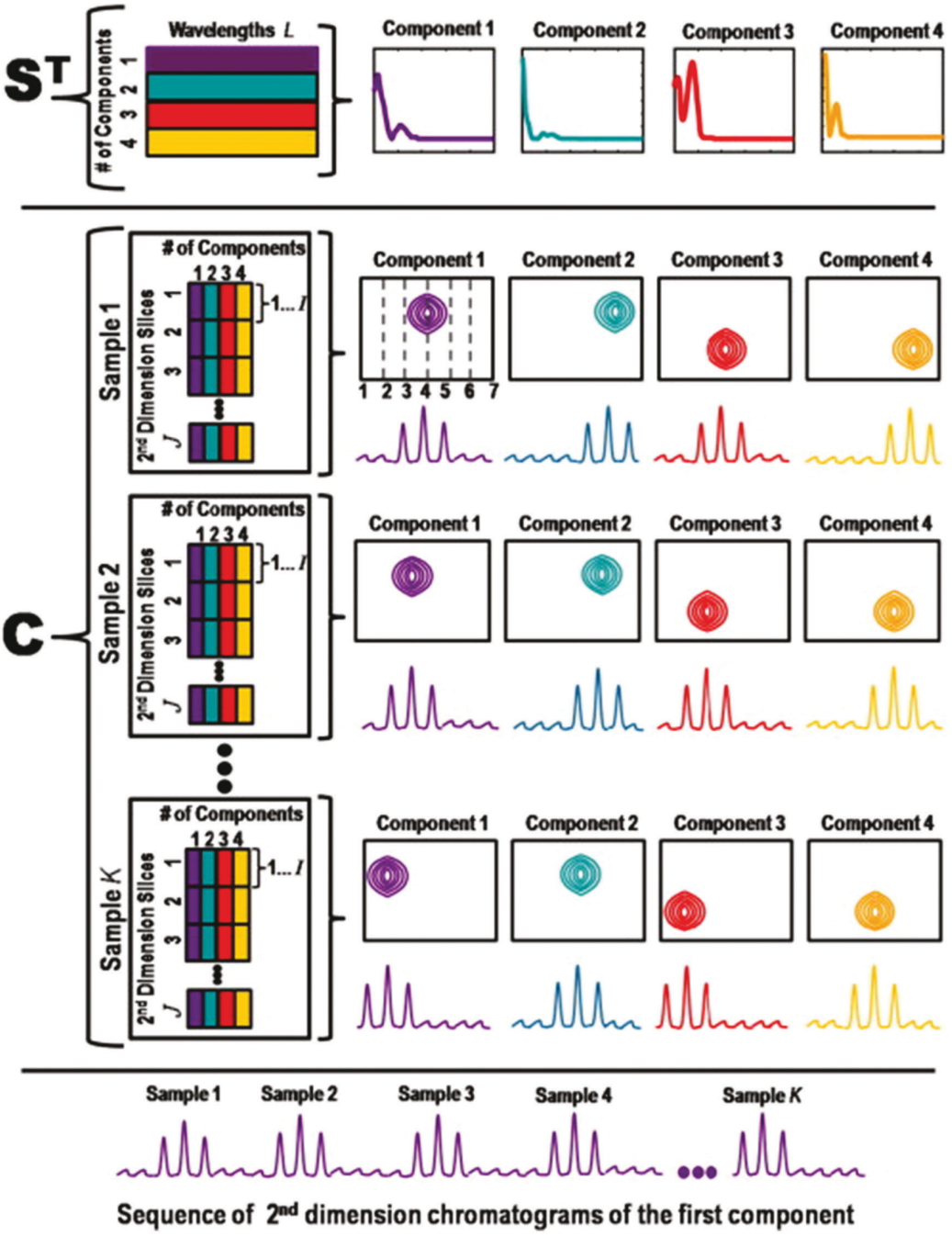


FIGURE 8.16 Example of curve resolution results from MCR-ALS for the analysis of multiple LC×LC-DAD chromatograms. The top panel shows the resolved spectra ($\mathbf{S}^T$) for the hypothetical four peaks found in the chromatogram. The next panel shows the resolved chromatograms (C) for these four peaks. On the left side, the concatenated ^{2}D chromatograms are shown in matrix format appended end to end, while on the right, the contour plots for each of the four resolved peaks are shown, with the sequence of individual ^{2}D peaks shown below the contour plot. The third panel shows the corresponding sequence of ^{2}D chromatograms for component 1 for all samples (1–K) as an example. Reprinted from *Chemometrics and Intelligent Laboratory Systems*, 106, H. Bailey, S. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, 131–141, Copyright (2011), with permission from Elsevier.

a total reduction of the data by a factor of 50. The reduced data were processed by MCR-ALS and 154 metabolites were detected. Although the protocol for this strategy is fairly complex, a detailed step-by-step procedure has been provided to aid in its implementation [119].

8.6.2 PARAFAC AND PARAFAC2

Another chemometric method that can also be used for quantification is parallel factor analysis (PARAFAC) [123, 124]. The PARAFAC method relies on an extension of the bilinearity concept described above. In the case of the analysis of a single LC×LC-DAD or LC×LC-MS experiment, as depicted in Figure 8.15 B, it is assumed that the data are trilinear, meaning that the ^{2}D chromatograms for individual pure components are perfectly reproducible across all modulations (as well as the spectra, as above), with no retention time shifts. The difference between trilinear and non-trilinear data is shown in Figure 8.17; note the differences in retention times. When multiple samples are treated simultaneously, as depicted in Figure 8.15 C, the data are assumed to have a quadrilinear structure, meaning that both the ^{2}D and ^{1}D chromatograms of the pure components are reproducible from sample to sample. In the case of either trilinear or quadrilinear data structure, the PARAFAC solution is achieved using an alternating least squares method. PARAFAC algorithms have also been developed which allow for constraints [125] such as non-negativity to be applied, but generally the implementation of constraints is much less flexible than for MCR-ALS. Algorithms for PARAFAC are available from Bro [125–127] and in Eigenvector's PLS Toolbox [112].

As LC×LC data is acquired as a series of 1D chromatograms, which are combined into a two-dimensional plane, the occurrence of retention-time shifts from one modulation to another and/or one sample to another is not uncommon. This means that relatively few applications of PARAFAC

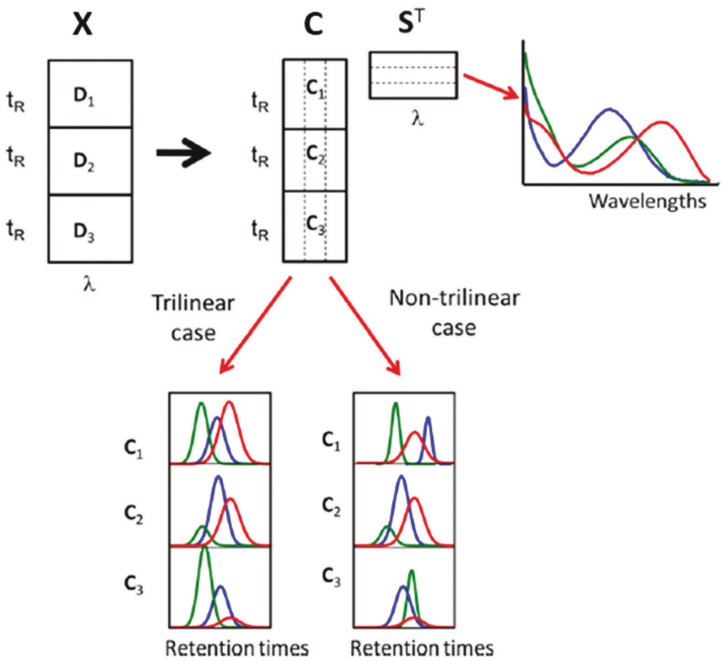


FIGURE 8.17 Schematic showing the difference between trilinear and non-trilinear data. Reprinted with permission from M. Bauza, G. Ibanez, R. Tauler, A. Olivieri, Sensitivity equation for quantitative analysis with multivariate curve resolution-Alternating Least-Squares: Theoretical and experimental approach, *Analytical Chemistry* 84 (2012), 8697–8706. Copyright 2012 American Chemical Society.

to raw LC×LC data have been reported [128]. Allen and Rutan have described a semi-automated method that incorporates initialization with MCR-ALS, along with appropriate alignment to correct for retention time shifts in both the ^{1}D and the ^{2}D to enable the use of PARAFAC to analyze both simulated data and an application for quantification of phenytoin in waste water by LC×LC-DAD with quadrilinear PARAFAC [129]. Recoveries were adequate, and in the case of the phenytoin analysis, quantitation accuracy and precision matched that of the reference LC-LC-MS/MS method.

PARAFAC2 is a variant of PARAFAC that relaxes the rigid multilinearity constraint in one of the dimensions [127, 130]. This is achieved by using a mathematical constraint that the matrix cross product of the matrix containing the shifting dimension is a constant. Effectively, this constraint allows for retention time shifts, as long as the shape of the profile is not significantly changed. PARAFAC2 has been applied to GC×GC datasets [55].

Navarro-Reig *et al.* compared MCR-ALS, PARAFAC and PARAFAC2 for the analysis of corn oil samples by LC×LC-MS [131]. As others have found, the data did not rigorously follow the trilinearity assumption, so that PARAFAC2 and MCR-ALS yielded improved results as compared to PARAFAC. The authors particularly commented on the flexibility of the MCR-ALS in this regard. These authors also noted that it is possible to implement a flexible trilinearity constraint within the MCR-ALS algorithm, which can also accommodate retention time shifts.

8.7 CLASSIFICATION

Often the results from chromatographic analysis are used for pattern recognition; either supervised or unsupervised methods can be used. Once peak tables are obtained from the analysis or compounds identified and quantified, numerous, well-established methods can be used [132–134], and their application is not different from the case of data obtained from 1D chromatographic methods. In this section, we will only discuss selected methods that have been applied within the framework of the 2D-LC data analysis workflow.

Pierce *et al.* have developed a method for detecting components in GC×GC-MS chromatograms that help to classify samples based on differences in concentration using a Fisher ratio method. This method carries out the comparison on a point-by-point basis, requiring that there is no misalignment between chromatograms. Subsequently, Marney *et al.* [135] and Parsons *et al.* [136] developed a tile-based Fisher ratio method that was less sensitive to misalignment, and resulted in fewer false positives (i.e., identification regions showing significant differences in concentrations, when such differences were not really present).

Bailey *et al.* have compared the Fisher ratio method [137] to the similarity index method of Windig [133] for the screening and classification of wine samples using LC×LC-DAD data [137]. The similarity index method was originally developed for 1D LC-MS data and is based on correlation coefficients. Both simulated data and wine analysis data were evaluated. Both methods were successfully able to identify peaks that showed significant concentration differences, which could be subsequently targeted for a more thorough analysis.

Reichenbach *et al.* developed an application to distinguish two classes of patients using LC×LC-DAD data of urine samples [138]. This method was based on creating a template based on the peaks or regions of the chromatograms that were common across multiple samples. Like the tile-based Fisher ratio method described above, this method does not require precise peak alignment to successfully identify those regions of the chromatogram showing significant concentration differences. Subsequently, two conventional pattern recognition techniques, K-nearest neighbor and support vector machines, were used for classification.

8.8 SUMMARY

Alongside advances in instrumentation for 2D-LC in recent years, there have been many advances in data analysis approaches to address the “data tsunami” from 2D-LC experiments. We have

summarized many of them in this chapter. While many significant and important advances have been made, there is a desperate need for more improvements in some key areas. First, the background contributions to signals in LC×LC data are generally larger and more variable than in conventional 1D-LC experiments. Additionally, many creative methods have been applied to 1D-LC experimental data, but many of these could be better tailored to LC×LC data, for example, the very short rapid gradients in ^{2}D chromatograms lead to pronounced background features (especially for DAD detection).

There is also room for significant improvement in peak detection. On the one hand, the watershed algorithm treats the 2D data holistically, but it requires very high quality data with clean preprocessing. On the other hand, two-step methods with peak clustering can better handle less ideal data, but require a complex series of supervised steps as part of the process.

Multi-way data methods show great promise, especially when full spectrum DAD or MS detectors are employed with LC×LC separations. These methods can typically resolve background contributions to the signal quite well, and can lead directly to quantitative results. However, the limitations of these methods include the fact that the chromatogram usually needs to be segmented to process small portions of the dataset at one time, and full automation of the process across the entire chromatogram has yet to be achieved.

Because of the numerous data analysis strategies already developed, and the increasing number of new algorithms being reported, the field would benefit from a dataset repository where reference datasets could be used to compare different data analysis methods, whether background subtraction, peak alignment, peak detection, and/or quantitation. Thus, data analysis in comprehensive 2D-LC continues to be an area ripe for creative and novel developments.

REFERENCES

[1] S.J. Qin, Process data analytics in the era of big data, AIChE J. 60 (2014) 3092–3100. doi:10.1002/aic.14523

[2] R.C. Allen, S.C. Rutan, Investigation of interpolation techniques for the reconstruction of the first dimension of comprehensive two-dimensional liquid chromatography – Diode array detector data, Anal. Chim. Acta. 705 (2011) 253–260. doi:10.1016/j.aca.2011.06.022

[3] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C: The Art of scientific computing, 2nd ed., Cambridge University Press, New York, 1992.

[4] D.R. Stoll, X. Li, X. Wang, P.W. Carr, S.E.G. Porter, S.C. Rutan, Fast, comprehensive two-dimensional liquid chromatography, J. Chromatogr. A. 1168 (2007) 3–43. doi:10.1016/j.chroma.2007.08.054

[5] J.J.A.M. Weusten, E.P.P.A. Derks, J.H.M. Mommers, S. van der Wal, Alignment and clustering strategies for GC×GC–MS features using a cylindrical mapping, Anal. Chim. Acta. 726 (2012) 9–21. doi:10.1016/j.aca.2012.03.009

[6] N.J. Micyus, S.K. Seeley, J. V. Seeley, Method for reducing the ambiguity of comprehensive two-dimensional chromatography retention times, J. Chromatogr. A. 1086 (2005) 171–174. doi:10.1016/j.chroma.2005.06.016

[7] J.D. Wilson, C.A.J. McInnes, The elimination of errors due to baseline drift in the measurement of peak areas in gas chromatography, J. Chromatogr. A. 19 (1965) 486–494. doi:10.1016/s0021-9673(01)99489-0

[8] G.A. Pearson, A general baseline-recognition and baseline-flattening algorithm, J. Magn. Reson. 27 (1977) 265–272. doi:10.1016/0022-2364(77)90076-2

[9] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639. doi:10.1021/ac60214a047

[10] M.U.A. Bromba, H. Ziegler, Application hints for Savitzky-Golay digital smoothing filters, Anal. Chem. 53 (1981) 1583–1586. doi:10.1021/ac00234a011

[11] C.G. Enke, T.A. Nieman, Signal-to-noise ratio enhancement by least-squares polynomial smoothing, Anal. Chem. 48 (1976) 705A-712A. doi:10.1021/ac50002a769

[12] D.F. Thekkudan, S.C. Rutan, Denoising and Signal-to-Noise Ratio Enhancement: Classical Filtering, in: Compr. Chemom ., Elsevier, Amsterdam, 2009: pp. 9–24. doi:10.1016/B978-044452701-1.00098-3

[13] P.D. Wentzell, C.D. Brown, Signal processing in analytical chemistry, in: Encycl. Anal. Chem., John Wiley & Sons, Ltd, Chichester, UK, 2000: pp. 9764–9800. doi:10.1002/9780470027318.a5207
[14] F. Vogt, Data filtering in instrumental analyses with applications to optical spectroscopy and chemical imaging, J. Chem. Educ. 88 (2011) 1672–1683. doi:10.1021/ed100984c
[15] A. Felinger, Peak Detection, in: Data Anal. Signal Process. Chromatogr., Elsevier, Amsterdam, 1998: pp. 183–190.
[16] R. Bracewell, The Fourier Transform and Its Applications, 3rd edition, McGraw-Hill Science, New York, 1999.
[17] A. Felinger, T.L. Pap, J. Inczédy, Improvement of the signal-to-noise ratio of chromatographic peaks by Fourier transform, Anal. Chim. Acta. 248 (1991) 441–446. doi:10.1016/S0003-2670(00)84661-9
[18] B. Walczak, Wavelets in Chemistry, 1st edition, Elsevier, Amsterdam, 2000.
[19] M. Daszykowski, I. Stanimirova, A. Bodzon-Kulakowska, J. Silberring, G. Lubec, B. Walczak, Start-to-end processing of two-dimensional gel electrophoretic images, J. Chromatogr. A. 1158 (2007) 306–317. doi:10.1016/j.chroma.2007.02.009
[20] M. Li Vigni, J.M. Prats-Montalban, A. Ferrer, M. Cocchi, Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA), J. Chemom. 32 (2018) e2970. doi:10.1002/cem.2970
[21] R.C. Allen, M.G. John, S.C. Rutan, M.R. Filgueira, P.W. Carr, Effect of background correction on peak detection and quantification in online comprehensive two-dimensional liquid chromatography using diode array detection, J. Chromatogr. A. 1254 (2012) 51–61. doi:10.1016/j.chroma.2012.07.034
[22] B.W.J. Pirok, J.A. Westerhuis, Challenges in obtaining relevant information from one- and two-dimensional LC experiments, LC-GC North Am. 6 (2020) 8–14. www.chromatographyonline.com/view/challenges-obtaining-relevant-information-one-and-two-dimensional-lc-experiments.
[23] E.T. Whittaker, On a new method of graduation, Proc. Edinburgh Math. Soc. 41 (1922) 63–75. doi:10.1017/S001309150000359X
[24] J. Carlos Cobas, M.A. Bernstein, M. Martín-Pastor, P.G. Tahoces, A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data, J. Magn. Reson. 183 (2006) 145–151. doi:10.1016/j.jmr.2006.07.013
[25] Z.-M. Zhang, S. Chen, Y.-Z. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding et al., An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy, J. Raman Spectrosc. 41 (2009) 659–669. doi:10.1002/jrs.2500
[26] P.H.C.C. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631–3636. doi:10.1021/ac034173t
[27] Z.-M.M. Zhang, S. Chen, Y.-Z.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst. 135 (2010) 1138. doi:10.1039/b922045c
[28] S.-J. Baek, A. Park, Y.-J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, Analyst. 140 (2015) 250–257. doi:10.1039/C4AN01061B
[29] R. Perez-Pueyo, M.J. Soneira, S. Ruiz-Moreno, Morphology-based automated baseline removal for raman spectra of artistic pigments, Appl. Spectrosc. 64 (2010) 595–600. doi:10.1366/000370210791414281
[30] Z. Li, D.-J. Zhan, J.-J. Wang, J. Huang, Q.-S. Xu, Z.-M. Zhang et al., Morphological weighted penalized least squares for background correction, Analyst. 138 (2013) 4483. doi:10.1039/c3an00743j
[31] H.-Y. Fu, H.-D. Li, Y.-J. Yu, B. Wang, P. Lu, H.-P. Cui, et al. Simple automatic strategy for background drift correction in chromatographic data analysis, J. Chromatogr. A. 1449 (2016) 89–99. doi:10.1016/j.chroma.2016.04.054
[32] H.-Y.Y. Fu, H.-D.D. Li, Y.-J.J. Yu, B. Wang, P. Lu, H.-P.P. Cui et al., Simple automatic strategy for background drift correction in chromatographic data analysis, J. Chromatogr. A. 1449 (2016) 89–99. doi:10.1016/j.chroma.2016.04.054
[33] P. Yaroshchyk, J.E. Eberhardt, Automatic correction of continuum background in Laser-induced Breakdown Spectroscopy using a model-free algorithm, Spectrochim. Acta – Part B At. Spectrosc. 99 (2014) 138–149. doi:10.1016/j.sab.2014.06.020
[34] Z. Li, D.-J. Zhan, J.-J. Wang, J. Huang, Q.-S. Xu, Z.-M. Zhang *et al.*, Morphological weighted penalized least squares for background correction, Analyst. 138 (2013) 4483. doi:10.1039/c3an00743j
[35] Y.-J. Yu, H.-L. Wu, H.-Y. Fu, J. Zhao, Y.-N. Li, S.-F. Li *et al.*, Chromatographic background drift correction coupled with parallel factor analysis to resolve coelution problems in three-dimensional chromatographic data: Quantification of eleven antibiotics in tap water samples by high-performance liquid chromatography, J. Chromatogr. A. 1302 (2013) 72–80. doi:10.1016/j.chroma.2013.06.009

[36] X. Ning, I.W. Selesnick, L. Duval, Chromatogram baseline estimation and denoising using sparsity (BEADS), Chemom. Intell. Lab. Syst. 139 (2014) 156–167. doi:10.1016/j.chemolab.2014.09.014
[37] J.A. Navarro-Huerta, J.R. Torres-Lapasió, S. López-Ureña, M.C. García-Alvarez-Coque, Assisted baseline subtraction in complex chromatograms using the BEADS algorithm, J. Chromatogr. A. 1507 (2017) 1–10. doi:10.1016/j.chroma.2017.05.057
[38] I. Selesnick, Sparsity-assisted signal smoothing (revisited), in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc., 2017: pp. 4546–4550. doi:10.1109/ICASSP.2017.7953017
[39] M. Lopatka, A. Barcaru, M.J. Sjerps, G. Vivó-Truyols, Leveraging probabilistic peak detection to estimate baseline drift in complex chromatographic samples, J. Chromatogr. A. 1431 (2016) 122–130. doi:10.1016/j.chroma.2015.12.063
[40] G.L. Erny, T. Acunha, C. Simó, A. Cifuentes, A. Alves, Background correction in separation techniques hyphenated to high-resolution mass spectrometry – Thorough correction with mass spectrometry scans recorded as profile spectra, J. Chromatogr. A. 1492 (2017) 98–105. doi:10.1016/j.chroma.2017.02.052
[41] A. Kaufmann, P. Butcher, Strategies to avoid false negative findings in residue analysis using liquid chromatography coupled to time-of-flight mass spectrometry, Rapid Commun. Mass Spectrom. 20 (2006) 3566–3572. doi:10.1002/rcm.2762
[42] Y. Zhang, H.-L. Wu, A.-L. Xia, L.-H. Hu, H.-F. Zou, R.-Q. Yu, Trilinear decomposition method applied to removal of three-dimensional background drift in comprehensive two-dimensional separation data, J. Chromatogr. A. 1167 (2007) 178–183. doi:10.1016/j.chroma.2007.08.055
[43] S.E. Reichenbach, M. Ni, D. Zhang, E.B. Ledford, Image background removal in comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 985 (2003) 47–56. doi:10.1016/S0021-9673(02)01498-X
[44] S.E. Reichenbach, P.W. Carr, D.R. Stoll, Q. Tao, Smart templates for peak pattern matching with comprehensive two-dimensional liquid chromatography, J. Chromatogr. A. 1216 (2009) 3458–3466. doi:10.1016/j.chroma.2008.09.058
[45] S.E. Reichenbach, V. Kottapalli, M. Ni, A. Visvanathan, Computer language for identifying chemicals with comprehensive two-dimensional gas chromatography and mass spectrometry, J. Chromatogr. A. 1071 (2005) 263–269. doi:10.1016/j.chroma.2004.08.125
[46] Z.-D. Zeng, S.-T. Chin, H.M. Hugel, P.J. Marriott, Simultaneous deconvolution and re-construction of primary and secondary overlapping peak clusters in comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1218 (2011) 2301–2310. doi:10.1016/j.chroma.2011.02.028
[47] L.E. Niezen, P.J. Schoenmakers, B.W.J. Pirok, Critical comparison of background correction methodology used in chromatography and spectroscopy, Anal. Chim. Acta. (2022) submitted.
[48] M.R. Filgueira, C.B. Castells, P.W. Carr, A simple, robust orthogonal background correction method for two-dimensional liquid chromatography, Anal. Chem. 84 (2012) 6747–6752. doi:10.1021/ac301248
[49] T. Gröger, M. Schäffer, M. Pütz, B. Ahrens, K. Drew, M. Eschner, et al., Application of two-dimensional gas chromatography combined with pixel-based chemometric processing for the chemical profiling of illicit drug samples, J. Chromatogr. A. 1200 (2008) 8–16. doi:10.1016/j.chroma.2008.05.028
[50] T. Gröger, R. Zimmermann, Application of parallel computing to speed up chemometrics for GC×GC-TOFMS based metabolic fingerprinting, Talanta. 83 (2011) 1289–1294. doi:10.1016/j.talanta.2010.09.015
[51] K.J. Johnson, B.J. Prazen, D.C. Young, R.E. Synovec, Quantification of naphthalenes in jet fuel with GC×GC/Tri-PLS and windowed rank minimization retention time alignment, J. Sep. Sci. 27 (2004) 410–416. doi:10.1002/jssc.200301640
[52] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data, Anal. Chem. 77 (2005) 7735–7743. doi:10.1021/ac0511142
[53] R.K. Nelson, B.M. Kile, D.L. Plata, S.P. Sylva, L. Xu, C.M. Reddy, et al., Tracking the weathering of an oil spill with comprehensive two-dimensional gas chromatography, Environ. Forensics. 7 (2006) 33–44. doi:10.1080/15275920500506758
[54] C. Cordero, E. Liberto, C. Bicchi, P. Rubiolo, S.E. Reichenbach, X. Tian et al., Targeted and non-targeted approaches for complex natural sample profiling by GCxGC-qMS, J. Chromatogr. Sci. 48 (2010) 251–261. doi:10.1093/chromsci/48.4.251

[55] T. Skov, J.C. Hoggard, R. Bro, R.E. Synovec, Handling within run retention time shifts in two-dimensional chromatography data using shift correction and modeling, J. Chromatogr. A. 1216 (2009) 4020–4029. doi:10.1016/j.chroma.2009.02.049
[56] D. Zhang, X. Huang, F.E. Regnier, M. Zhang, Two-dimensional correlation optimized warping algorithm for aligning GC×GC−MS data, Anal. Chem. 80 (2008) 2664–2671. doi:10.1021/ac7024317
[57] J. Gros, D. Nabi, P. Dimitriou-Christidis, R. Rutler, J.S. Arey, Robust algorithm for aligning two-dimensional chromatograms, Anal. Chem. 84 (2012) 9033–9040. doi:10.1021/ac301367s
[58] Y. Zushi, J. Gros, Q. Tao, S.E. Reichenbach, S. Hashimoto, J.S. Arey, Pixel-by-pixel correction of retention time shifts in chromatograms from comprehensive two-dimensional gas chromatography coupled to high resolution time-of-flight mass spectrometry, J. Chromatogr. A. 1508 (2017) 121–129. doi:10.1016/j.chroma.2017.05.065
[59] V.G. van Mispelaar, A.C. Tas, A.K. Smilde, P.J. Schoenmakers, A.C. van Asten, Quantitative analysis of target components by comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1019 (2003) 15–29. doi:10.1016/j.chroma.2003.08.101
[60] H. Parastar, M. Jalali-Heravi, R. Tauler, Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution, Chemom. Intell. Lab. Syst. 117 (2012) 80–91. doi:10.1016/j.chemolab.2012.02.003
[61] Q.-X. Zheng, H.-Y. Fu, H.-D. Li, B. Wang, C.-H. Peng, S. Wang, et al., Automatic time-shift alignment method for chromatographic data analysis, Sci. Rep. 7 (2017) 256. doi:10.1038/s41598-017-00390-7
[62] H.Y. Fu, O. Hu, Y.M. Zhang, L. Zhang, J.J. Song, P. Lu, et al., Mass-spectra-based peak alignment for automatic nontargeted metabolic profiling analysis for biomarker screening in plant samples, J. Chromatogr. A. 1513 (2017) 201–209. doi:10.1016/j.chroma.2017.07.044
[63] Y.-J. Yu, H.-Y. Fu, L. Zhang, X.-Y. Wang, P.-J. Sun, X.-B. Zhang, et al., A chemometric-assisted method based on gas chromatography – Mass spectrometry for metabolic profiling analysis, J. Chromatogr. A. 1399 (2015) 65–73. doi:10.1016/j.chroma.2015.04.029
[64] P. Xia, L. Zhang, F. Li, Learning similarity with cosine similarity ensemble, Inf. Sci. (Ny). 307 (2015) 39–52. doi:10.1016/j.ins.2015.02.024
[65] T.S. Bos, W.C. Knol, S.R.A. Molenaar, L.E. Niezen, P.J. Schoenmakers, G.W. Somsen et al., Recent applications of chemometrics in two-dimensional chromatography, J. Sep. Sci. (2020) submitted.
[66] E. Grushka, M.N. Myers, P.D. Schettler, J.C. Giddings, Computer characterization of chromatographic peaks by plate height and higher central moments, Anal. Chem. 41 (1969) 889–892. doi:10.1021/ac60276a014
[67] S.B. Howerton, C. Lee, V.L. McGuffin, Additivity of statistical moments in the exponentially modified Gaussian model of chromatography, Anal. Chim. Acta. 478 (2003) 99–110. doi:10.1016/S0003-2670(02)01472-1
[68] Y. Vanderheyden, K. Broeckhoven, G. Desmet, Comparison and optimization of different peak integration methods to determine the variance of unretained and extra-column peaks, J. Chromatogr. A. 1364 (2014) 140–150. doi:10.1016/j.chroma.2014.08.066
[69] P.G. Stevenson, X.A. Conlan, N.W. Barnett, Evaluation of the asymmetric least squares baseline algorithm through the accuracy of statistical peak moments, J. Chromatogr. A. 1284 (2013) 107–111. doi:10.1016/j.chroma.2013.02.012
[70] D.W. Morton, C.L. Young, Analysis of peak profiles using statistical moments, J. Chromatogr. Sci. 33 (1995) 514–524. doi:10.1093/chromsci/33.9.514
[71] T.S. Bos, W.C. Knol, S.R.A. Molenaar, L.E. Niezen, P.J. Schoenmakers, G.W. Somsen et al., Recent applications of chemometrics in one- and two-dimensional chromatography, J. Sep. Sci. 43 (2020) 1678–1727. doi:10.1002/jssc.202000011
[72] H.-Y. Fu, J.-W. Guo, Y.-J. Yu, H.-D. Li, H.-P. Cui, P.-P. Liu, et al., A simple multi-scale Gaussian smoothing-based strategy for automatic chromatographic peak extraction, J. Chromatogr. A. 1452 (2016) 1–9. doi:10.1016/j.chroma.2016.05.018
[73] S. Peters, H.-G. Janssen, G. Vivó-Truyols, A new method for the automated selection of the number of components for deconvolving overlapping chromatographic peaks, Anal. Chim. Acta. 799 (2013) 29–35. doi:10.1016/j.aca.2013.08.041
[74] Z.-M. Zhang, X. Tong, Y. Peng, P. Ma, M.-J. Zhang, H.-M. Lu et al., Multiscale peak detection in wavelet space, Analyst. 140 (2015) 7955–7964. doi:10.1039/C5AN01816A

[75] J. Lu, M.J. Trnka, S.-H. Roh, P.J.J. Robinson, C. Shiau, D.G. Fujimori et al., Improved peak detection and deconvolution of native electrospray mass spectra from large protein complexes, J. Am. Soc. Mass Spectrom. 26 (2015) 2141–2151. doi:10.1007/s13361-015-1235-6

[76] Y.-J. Yu, Q.-L. Xia, S. Wang, B. Wang, F.-W. Xie, X.-B. Zhang, et al., Chemometric strategy for automatic chromatographic peak detection and background drift correction in chromatographic data, J. Chromatogr. A. 1359 (2014) 262–270. doi:10.1016/j.chroma.2014.07.053

[77] V.P. Andreev, T. Rejtar, H.-S. Chen, E. V. Moskovets, A.R. Ivanov, B.L. Karger, A universal denoising and peak picking algorithm for LC–MS based on matched filtration in the chromatographic time domain, Anal. Chem. 75 (2003) 6314–6326. doi:10.1021/ac0301806

[78] R.A.R.A. Carmona, W.L.W.L. Hwang, B. Torresani, B. Torrésani, Multiridge detection and time-frequency reconstruction, IEEE Trans. Signal Process. 47 (1999) 480–492. doi:10.1109/78.740131

[79] E. Dinç, E. Büker, A new application of continuous wavelet transform to overlapping chromatograms for the quantitative analysis of amiloride hydrochloride and hydrochlorothiazide in tablets by ultra-performance liquid chromatography, J. AOAC Int. 95 (2012) 751–756. doi:10.5740/jaoacint.SGE_Dinc

[80] X. Shao, L. Sun, An application of the continuous wavelet transform to resolution of multicomponent overlapping analytical signals, Anal. Lett. 34 (2001) 267–280. doi:10.1081/AL-100001578

[81] X. Tong, Z. Zhang, F. Zeng, C. Fu, P. Ma, Y. Peng, et al., Recursive wavelet peak detection of analytical signals, Chromatographia. 79 (2016) 1247–1255. doi:10.1007/s10337-016-3155-4

[82] Y.-J. Yu, Q.-X. Zheng, Y.-M. Zhang, Q. Zhang, Y.-Y. Zhang, P.-P. Liu et al., Automatic data analysis workflow for ultra-high performance liquid chromatography-high resolution mass spectrometry-based metabolomics, J. Chromatogr. A. 1585 (2019) 172–181. doi:10.1016/j.chroma.2018.11.070

[83] G. Vivó-Truyols, Bayesian approach for peak detection in two-dimensional chromatography, Anal. Chem. 84 (2012) 2622–2630. doi:10.1021/ac202124t

[84] J.M. Davis, J.C. Giddings, Statistical theory of component overlap in multicomponent chromatograms, Anal. Chem. 55 (1983) 418–424. doi:10.1021/ac00254a003

[85] M. Lopatka, G. Vivó-Truyols, M.J. Sjerps, Probabilistic peak detection for first-order chromatographic data, Anal. Chim. Acta. 817 (2014) 9–16. doi:10.1016/j.aca.2014.02.015

[86] M. Woldegebriel, G. Vivó-Truyols, Probabilistic model for untargeted peak detection in lc–ms using bayesian statistics, Anal. Chem. 87 (2015) 7345–7355. doi:10.1021/acs.analchem.5b01521

[87] S. Peters, G. Vivó-Truyols, P.J. Marriott, P.J. Schoenmakers, Development of an algorithm for peak detection in comprehensive two-dimensional chromatography, J. Chromatogr. A. 1156 (2007) 14–24. doi:10.1016/j.chroma.2006.10.066

[88] J. Pól, B. Hohnová, M. Jussila, T. Hyötyläinen, Comprehensive two-dimensional liquid chromatography–time-of-flight mass spectrometry in the analysis of acidic compounds in atmospheric aerosols, J. Chromatogr. A. 1130 (2006) 64–71. doi:10.1016/j.chroma.2006.04.050

[89] M. Kivilompolo, T. Hyötyläinen, Comprehensive two-dimensional liquid chromatography in analysis of Lamiaceae herbs: Characterisation and quantification of antioxidant phenolic acids, J. Chromatogr. A. 1145 (2007) 155–164. doi:10.1016/j.chroma.2007.01.090

[90] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemom. Intell. Lab. Syst. 106 (2011) 131–141. doi:10.1016/j.chemolab.2010.07.008

[91] S. Kim, M. Ouyang, J. Jeong, C. Shen, X. Zhang, A new method of peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data, Ann. Appl. Stat. 8 (2014) 1209–1231. doi:10.1214/14-AOAS731

[92] S. Kim, H. Jang, I. Koo, J. Lee, X. Zhang, Normal–Gamma–Bernoulli peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data, Comput. Stat. Data Anal. 105 (2017) 96–111. doi:10.1016/j.csda.2016.07.015

[93] S.E. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, Information technologies for comprehensive two-dimensional gas chromatography, Chemom. Intell. Lab. Syst. 71 (2004) 107–120. doi:10.1016/j.chemolab.2003.12.009

[94] B. Li, S.E. Reichenbach, Q. Tao, R. Zhu, A streak detection approach for comprehensive two-dimensional gas chromatography based on image analysis, Neural Comput. Appl. (2018). doi:10.1007/s00521-018-3917-z

[95] G. Vivó-Truyols, H.-G. Janssen, Probability of failure of the watershed algorithm for peak detection in comprehensive two-dimensional chromatography, J. Chromatogr. A. 1217 (2010) 1375–1385. doi:10.1016/j.chroma.2009.12.063
[96] I. Latha, S.E. Reichenbach, Q. Tao, Comparative analysis of peak-detection techniques for comprehensive two-dimensional chromatography, J. Chromatogr. A. 1218 (2011) 6792–6798. doi:10.1016/j.chroma.2011.07.052
[97] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemom. 9 (1995) 31–58. doi:10.1002/cem.1180090105
[98] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in concepts and applications, Crit. Rev. Anal. Chem. 36 (2006) 163–176. doi:10.1080/10408340600970005
[99] R.W. Hendler, R.I. Shrager, Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners, J. Biochem. Biophys. Methods. 28 (1994) 1–33. doi:10.1016/0165-022X(94)90061-2
[100] Y. Nagai, W.Y. Sohn, K. Katayama, An initial estimation method using cosine similarity for multivariate curve resolution: application to NMR spectra of chemical mixtures, Analyst. 144 (2019) 5986–5995. doi:10.1039/C9AN01416K
[101] E.R. Malinowski, Obtaining the key set of typical vectors by factor analysis and subsequent isolation of component spectra, Anal. Chim. Acta. 134 (1982) 129–137. doi:10.1016/S0003-2670(01)84184-2
[102] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425–1432. doi:10.1021/ac00014a016
[103] F. Cuesta Sánchez, B. Van Den Bogaert, S.C. Rutan, D.L. Massart, Multivariate peak purity approaches, Chemom. Intell. Lab. Syst. 34 (1996) 139–171. doi:10.1016/0169-7439(96)00020-2
[104] A. De Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods. 6 (2014) 4964–4976. doi:10.1039/C4AY00571F
[105] M.C. Bauza, G.A. Ibañez, R. Tauler, A.C. Olivieri, Sensitivity equation for quantitative analysis with multivariate curve resolution-alternating least-squares: Theoretical and experimental approach, Anal. Chem. 84 (2012) 8697–8706. doi:10.1021/ac3019284
[106] D.W. Cook, S.C. Rutan, D.R. Stoll, P.W. Carr, Two dimensional assisted liquid chromatography – A chemometric approach to improve accuracy and precision of quantitation in liquid chromatography using 2D separation, dual detectors, and multivariate curve resolution, Anal. Chim. Acta. 859 (2015) 87–95. doi:10.1016/j.aca.2014.12.009
[107] X. Domingo-Almenara, A. Perera, J. Brezmes, Avoiding hard chromatographic segmentation: A moving window approach for the automated resolution of gas chromatography – Mass spectrometry-based metabolomics signals by multivariate methods, J. Chromatogr. A. 1474 (2016) 145–151. doi:10.1016/j.chroma.2016.10.066
[108] R. Wehrens, E. Carvalho, D. Masuero, A. de Juan, S. Martens, High-throughput carotenoid profiling using multivariate curve resolution, Anal. Bioanal. Chem. 405 (2013) 5075–5086. doi:10.1007/s00216-012-6555-9
[109] R. Tauler, A. de Juan, J. Jaumot, MCR-ALS toolbox, www.mcrals.info/. Accessed May 30, 2022.
[110] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: A new tool for multivariate curve resolution in MATLAB, Chemom. Intell. Lab. Syst. 76 (2005) 101–110. doi:10.1016/j.chemolab.2004.12.007
[111] A.C. Olivieri, H.-L. Wu, R.-Q. Yu, MVC2: A Matlab graphical interface toolbox for second-order multivariate calibration, Chemom. Intell. Lab. Syst. 96 (2009) 246–251. doi:10.1016/j.chemolab.2009.02.005
[112] Eigenvector PLS Toolbox, https://eigenvector.com/software/pls-toolbox/. Accessed May 30, 2022.
[113] G. Vivó-Truyols, J.R. Torres-Lapasió, M.C. García-Alvarez-Coque, P.J. Schoenmakers, Towards unsupervised analysis of second-order chromatographic data: Automated selection of number of components in multivariate curve-resolution methods, J. Chromatogr. A. 1158 (2007) 258–272. doi:10.1016/j.chroma.2007.03.005
[114] M. Navarro-Reig, J. Jaumot, A. García-Reiriz, R. Tauler, Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies, Anal. Bioanal. Chem. 407 (2015) 8835–8847. doi:10.1007/s00216-015-9042-2

[115] M. Farrés, B. Piña, R. Tauler, Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC–MS, Metabolomics. 11 (2015) 210–224. doi:10.1007/s11306-014-0689-z
[116] R. Stolt, R.J.O. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen et al., Second-order peak detection for multicomponent high-resolution LC/MS data, Anal. Chem. 78 (2006) 975–983. doi:10.1021/ac050980b
[117] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, MSroi: A pre-processing tool for mass spectrometry-based studies, Chemom. Intell. Lab. Syst. 215 (2021) 104333. doi:10.1016/j.chemolab.2021.104333
[118] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: A powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinformatics. 20 (2019). doi:10.1186/s12859-019-2848-8
[119] R. Tauler, E. Gorrochategui, J. Jaumot, R. Tauler, A protocol for LC-MS metabolomic data processing using chemometric tools, Protoc. Exch. (2015) 1–46. doi:10.1038/protex.2015.102
[120] M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler, Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice metabolome using multivariate curve resolution, Anal. Chem. 89 (2017) 7675–7683. doi:10.1021/acs.analchem.7b01648
[121] I. Daubechies, Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, 1992. doi:10.1137/1.9781611970104
[122] J. Trygg, N. Kettaneh-Wold, L. Wallbäcks, 2D wavelet analysis and compression of on-line industrial process data, J. Chemom. 15 (2001) 299–319. doi:10.1002/cem.681
[123] D.W. Cook, S.C. Rutan, Chemometrics for the analysis of chromatographic data in metabolomics investigations, J. Chemom. 28 (2014) 681–687. doi:10.1002/cem.2624
[124] A. Smilde, R. Bro, P. Geladi, Multi-Way Analysis with Applications in the Chemical Sciences, John Wiley & Sons, Ltd, Chichester, UK, 2004. doi:10.1002/0470012110
[125] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171. doi:10.1016/S0169-7439(97)00032-4
[126] C.A. Andersson, R. Bro, The N-way Toolbox for MATLAB, Chemom. Intell. Lab. Syst. 52 (2000) 1–4. doi:10.1016/S0169-7439(00)00071-X
[127] H.A.L. Kiers, J.M.F. ten Berge, R. Bro, PARAFAC2 – Part I. A direct fitting algorithm for the PARAFAC2 model, J. Chemom. 13 (1999) 275–294. doi:10.1002/(SICI)1099-128X(199905/08)13:3/4<275::AID-CEM543>3.3.CO;2-2
[128] S.E.G. Porter, D.R. Stoll, S.C. Rutan, P.W. Carr, J.D. Cohen, Analysis of Four-Way Two-Dimensional Liquid Chromatography-Diode Array Data: Application to Metabolomics, Anal. Chem. 78 (2006) 5559–5569. doi:10.1021/ac0606195
[129] R.C. Allen, S.C. Rutan, Semi-automated alignment and quantification of peaks using parallel factor analysis for comprehensive two-dimensional liquid chromatography – Diode array detector data sets, Anal. Chim. Acta. 723 (2012) 7–17. doi:10.1016/j.aca.2012.02.019
[130] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2 – Part II. Modeling chromatographic data with retention time shifts, J. Chemom. 13 (1999) 295–309. doi:10.1002/(SICI)1099-128X(199905/08)13:3/4<295::AID-CEM547>3.0.CO;2-Y
[131] M. Navarro-Reig, J. Jaumot, T.A. van Beek, G. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oil, Talanta. 160 (2016) 624–635. doi:10.1016/j.talanta.2016.08.005
[132] L.C. Lee, C.-Y. Liong, A.A. Jemain, Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps, Analyst. 143 (2018) 3526–3539. doi:10.1039/C8AN00599K
[133] W. Windig, W.F. Smith, W.F. Nichols, Fast interpretation of complex LC/MS data using chemometrics, Anal. Chim. Acta. 446 (2001) 465–474. doi:10.1016/S0003-2670(01)01276-4
[134] K. Vanden Branden, M. Hubert, Robust classification in high dimensions based on the SIMCA Method, Chemom. Intell. Lab. Syst. 79 (2005) 10–21. doi:10.1016/j.chemolab.2005.03.002
[135] L.C. Marney, W. Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography – Time-of-flight mass spectrometry data, Talanta. 115 (2013) 887–895. doi:10.1016/j.talanta.2013.06.038

[136] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC–TOFMS) data using a null distribution approach, Anal. Chem. 87 (2015) 3812–3819. doi:10.1021/ac504472s

[137] H.P. Bailey, S.C. Rutan, Comparison of chemometric methods for the screening of comprehensive two-dimensional liquid chromatographic analysis of wine, Anal. Chim. Acta. 770 (2013) 18–28. doi:10.1016/j.aca.2013.01.062

[138] S.E. Reichenbach, X. Tian, Q. Tao, D.R. Stoll, P.W. Carr, Comprehensive feature analysis for sample classification with comprehensive two-dimensional LC, J. Sep. Sci. 33 (2010) 1365–1374. doi:10.1002/jssc.200900859