14

Application of Chemometrics to the Interpretation of Analytical Separations Data

James J. Harynuk, A. Paulina de la Mata and Nikolai A. Sinkov
Department of Chemistry, University of Alberta
Canada

1. Introduction

Interesting real-world samples are almost always present as mixtures containing the analyte(s) of interest and a matrix of components that are irrelevant to answering the analytical question at hand. Additionally, the compounds comprising the matrix are usually present in far greater abundance (both number and concentration) than the analytes of interest, making quantification or even detection of these analytes difficult if not impossible.

When tasked with these types of samples, analysts turn to some form of separations technique such as gas or liquid chromatography (GC or LC) or capillary electrophoresis (CE) so that individual components in each sample may be quantified. More recently, more complex analytical questions are being probed, for example profiling blood or urine to identify a disease state or ascertaining the geographic origin of a food/beverage sample. These tasks often go beyond the simple quantification of one or two analytes in a sample. For these and other similar questions, separations scientists are turning more often to chemometric tools as a means of visualizing and interpreting the rich data that they obtain from their separations systems.

Here we present a brief overview of separations approaches, with a focus on the data that are derived from different methods and on phenomena in the separations approach that lead to challenges in data interpretation. This is followed by a discussion of approaches that exist for the chemometric interpretation of separations data, specific challenges that arise in the chemometric treatment of these data, and solutions that have been implemented to deal with these challenges.

1.1 Separations techniques

Chromatography is widely used for the separation, purification, and analysis of mixtures. In general, analytes contained in either a gaseous or liquid mobile phase are flowed past a stationary phase which is usually confined within a column. Depending on the chemistries of the analytes and the conditions of the separation (mobile/stationary phase compositions, temperature, etc.) different compounds will partition between the two phases to varying degrees. The separation arises due to this differential partitioning, with analytes which associate weakly with the stationary phase passing through the column more quickly than those with a greater affinity for the stationary phase (Miller, 2005; Cazes, 2010).

There are many types of chromatography, with the most common being liquid chromatography (LC) where analytes partition between a mobile liquid phase and an immobile stationary phase, and gas chromatography (GC) where the mobile phase is a gas and the stationary phase is a solid or more often a viscous, liquid-like polymer. There are numerous modes for LC separations, including for example reverse-phase (RPLC), normal-phase (NPLC), ion (IC), size exclusion (SEC), and hydrophilic interaction (HILIC) to name a few. From a point of view of chemometric data interpretation and the discussion in this chapter, all of these LC separations generate data which are equivalent. In any chromatographic separation, the sample is delivered to the inlet of the column while the outlet is connected to a detector, which records a continuous signal. The detector response rises and then falls to baseline based on the analyte flux passing through it, ideally generating one separate peak with an approximately Gaussian shape for each individual analyte. Assuming that the conditions for repeat analyses are not changed, the peak for a given analyte will appear at the same time in every analysis, with the peak area/height being proportional to the quantity of analyte present in a sample (Poole, 2003; Miller, 2005).

Another separations technique which is popular for some samples is capillary electrophoresis (CE). Here, an electric field applied across a fused silica capillary containing a buffer induces motion of the buffer and analytes in the sample. The CE separation is dependent on differential mobilities of analytes in the solution in the presence of the electric field. This difference in mobilities is based on the fact that different analytes have different charges and sizes in solution. While the separation mechanism of CE is fundamentally different from the chromatographic mechanism, the data are a series of peaks recorded as a function of time. Consequently, the same tools can be applied to data from a CE separation, and similar concerns exist for the interpretation of these data (Poole, 2003; Miller, 2005). For ease of readability, and because chemometrics are more often applied to chromatographic data than electrophoretic data, we will often refer to a chromatogram in this chapter. This could equally be an electropherogram; when considering the application of chemometric techniques to separations data whether the origin is electrophoretic or chromatographic is largely irrelevant.

When tasked with incredibly complex samples, analysts are now turning more and more frequently to so-called comprehensive multidimensional separations (e.g.: GC×GC, LC×LC, CE×CE) (Liu & Phillips, 1991; Erni & Frei, 1978; Michels et al., 2002). In these techniques, the mixture of compounds is sequentially separated by two different separation mechanisms. In the case of GC×GC, for example, a sample might be separated first on an apolar column, followed by a polar column. The exact workings of comprehensive multidimensional separations are beyond the scope of this work, and are discussed elsewhere (Górecki et al., 2004; Cortes et al., 2009; François et al., 2009; Kivilompolo et al., 2011; Li et al., 2011). However, these techniques are gaining in popularity, and are capable of separating exceedingly complex mixtures comprising thousands of individual compounds. Due to the vastly improved separation power of these techniques, the data are much more information-rich, and without some form of chemometric treatment it is essentially impossible to do more than scratch the surface of the information contained therein.

1.2 Separations data

The detector signal from a separations experiment, when plotted vs. time, yields a series of (ideally) Gaussian peaks, each representing one compound in the sample. Acquisition speed

is one consideration for a chromatographic detector: it must be sufficient to faithfully record the profile of each compound as it passes through the detector. In order to obtain an accurate peak profile, the minimum number of acquisition points required across a peak is 10. Thus, the required speed of the detector is intrinsically linked to the nature of the separation. In separations where the base width of the peaks are on the order of 5 s, a data rate of 2 Hz would be acceptable, but when peak widths are 100-200 ms, as in GC×GC, then detector rates on the order of 50-100 Hz are required for quantitative analysis.

From a point of view of chemometric analysis of separations data, another important consideration is whether the detector is univariate or multivariate. Univariate detectors, such as the flame ionisation detector, or single-wavelength UV-visible spectrometer, record only one variable as a function of time, generating data which take the form of a vector of instrument response. Other detectors, typically mass spectrometers and multi-channel spectroscopic instruments, can be operated such that they record a multivariate response. Data from these instruments comprise an array of signal responses with each row representing a time when a response was recorded, and each column representing a variable that was recorded (e.g.: detector wavelength, ion mass-to-charge ratio). To the chemometrician, it is immediately obvious that there are numerous advantages to collecting multivariate chromatographic data; however, it is worth noting that most of this advantage has been by and large ignored by chromatographers. Typically, only the profile of a single variable vs. time would be used to selectively quantify an analyte, or the detector response across all channels at a given time used to help identify a peak.

One other aspect of raw separations data is the sheer number of variables measured for each sample. When a univariate detector is used for a 15 min separation, operating with an acquisition speed of 10 Hz, the data will be a vector of 9000 individual measurements per sample. If a multivariate detector is employed instead, for example a mass spectrometer operating over a 30-300 m/z mass range, this number increases to 2 439 000 individual variables arranged in a 9000 × 271 array per sample! In the case of GC×GC-MS analyses, which are typically 60 min in length but have a high-speed MS collecting data at rates of ~100 Hz, there are on the order of 100 million data points collected for each sample.

2. Challenges with chromatographic data

Variations in analytical separations data are, in principle, no different from those derived from any other instrument; being based on both chemical and non-chemical aspects of the analysis. All relevant information will be contained within the chemical variations and any chemometric approach to interpreting chromatographic data must be capable of identifying relevant chemical variation while minimizing the effects of irrelevant chemical and non-chemical variations. Sources of irrelevant chemical variation include matrix peaks, here defined as any chemical source of signal introduced with the sample, but having no bearing on the conclusions drawn from the data. Additionally, there is background signal which can for example derive from changes in mobile phase concentration which influence detector signals in LC or chemical "bleed" signatures from stationary phases as they degrade in GC. Non-chemical variations include, for example, baseline drift (for non-chemical reasons), retention time shifts (due to minor fluctuations in operating conditions), and electronic noise. These may easily interfere with the relevant chemical information, degrading model performance and the validity of results (de la Mata-Espinosa et al., 2011a). Figure 1 presents

an overlay of several LC chromatograms of similar samples exemplifying the challenges of baseline drift and retention time shifts. One of the major challenges in handling chromatographic data using chemometric tools is appropriate pre-processing to remove as many non-chemical and irrelevant chemical variations as possible from the data set.

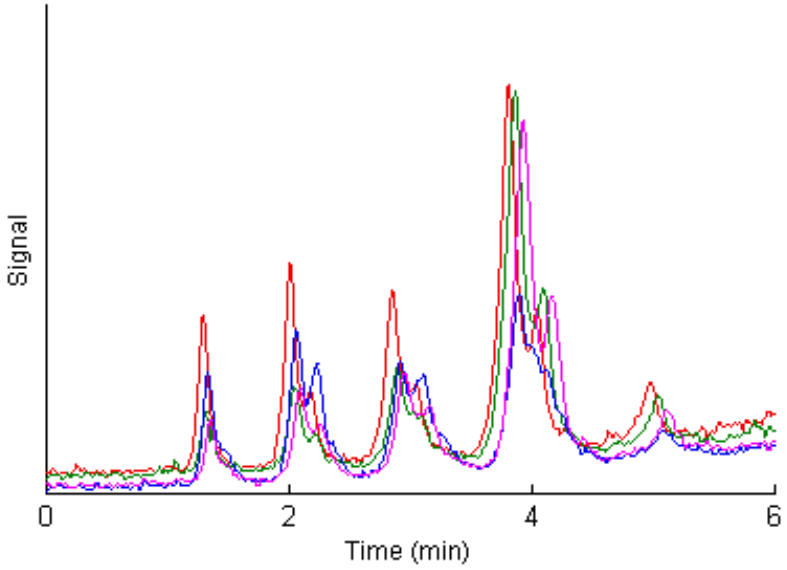


Fig. 1. LC chromatograms of edible oils showing a high degree of variation in baseline.

Initial efforts into the application of statistical and chemometric tools to chromatographic data were accomplished using data that were processed to provide a list of detected, integrated peak areas or heights (or the calibrated concentrations for known compounds). However, the trend in recent years has turned towards the direct chemometric interpretation of raw chromatographic signals (Watson et al., 2006; Johnson & Synovec, 2002). The reason for this trend is that many errors can occur during integration of raw signals (Asher et al., 2009; de la Mata-Espinosa et al., 2011b). By applying chemometric tools directly to the raw data, many of these errors can be avoided. Of course, when working with the raw data, other issues become more important, most notably retention time shifts and the population of available variables.

2.1 Baseline and noise

Baseline variations, such as noise and drift, are due to small changes in experimental conditions, for example changes in detector response due to the mobile phase gradient in LC separations or increased levels of stationary phase bleed at higher temperatures in temperature-programmed GC. Other sources of noise and drift could include changes in detector response as its components age, contamination of solvents or gases, and of course electronic noise (which is minimal in modern chromatographic systems).

Chemometric approaches to handling chromatographic data should incorporate baseline correction of some form. When raw chromatographic data are processed, the method of baseline correction and its importance are generally obvious to the analyst. In the case where integrated peak tables are used, this is often done automatically by the chromatographic software with little consideration by the analyst, even though the manner in which the baseline is calculated will significantly influence the determination of peak areas/heights.

2.2 Retention time shifts

In all separations, retention times of peaks can easily shift by a few seconds from one analysis to the next. This is not much of an issue with simple samples having only a few peaks which are then integrated prior to chemometric analysis. However, retention times of peaks are used for identifying the compounds. With complex separations, unstable retention times may result in unreliable peak identification, making comparisons from one run to the next impossible. When comparing raw data this is even more important as one must ensure that the peak for a given component is always registered in the exact same position in the data matrix so that the algorithms will recognize the signals correctly.

The causes of retention time shifts depend on the separations technique being used. In GC, peaks may shift due to degradation of the stationary phase, decreasing retention times over time; build-up of heavy matrix components which foul the column, effectively changing the chemistry of the stationary phase; minor gas leaks which alter the flow rate; or even matrix effects on the evaporation rate in the injector, affecting the rate of mass transfer to the column. In LC, peak shifts may be due to small fluctuations in mobile phase chemistry from one run to the next; temperature fluctuations which in turn affect solvent viscosity and solute diffusion coefficients, altering the kinetics as well as the thermodynamics of the separation; or degradation / fouling of the stationary phase of the column. CE is the technique most prone to drastic shifts in migration time, due to the instability of the electroosmotic flow in the capillary (Figure 2). Electroosmotic flow depends on the applied voltage, the buffer concentration and composition, and is incredibly sensitive to the surface chemistry of the capillary. The act of analyzing a sample by CE will often have a minor, possibly irreversible effect on the capillary surface, resulting in a change in the migration time of an analyte.

Shifts in retention times are minimized by proper instrument maintenance, precise control of instrumental conditions or by using approaches such as retention time locking in GC to account for variations in instrument performance (Etxebarria et al., 2009; Mommers et al., 2011) and relative retention times in CE. Even with these approaches, some retention time shifting will occur and require more advanced alignment techniques for correction prior to chemometric analysis.

2.3 Incomplete separation

Another challenge with the interpretation of chromatographic data is incomplete separation of peaks. If two or more compounds have similar retention characteristics under a given set of separation conditions, they will not be completely resolved, as evidenced by the peak clusters in Figure 1. In these cases, apportioning the signal between the different compounds

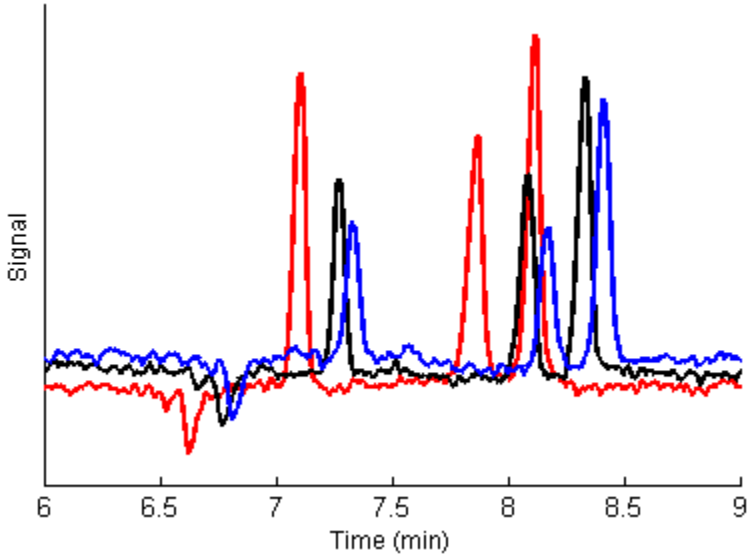


Fig. 2. CE of substituted benzenes showing extreme misalignment.

becomes a challenge, especially for univariate signals. The general approach used for these cases is one of deconvolution: decomposing the analytical signal to determine the contribution of each coeluting compound, or to determine the contribution of the compound of interest, disregarding the remaining data.

2.4 Data overload

As shown in Section 1.2, raw chromatographic signals present an overabundance of data to the analyst. This poses several challenges. From a practical point of view, attempts to construct a chemometric model using the entirety of the data set could easily exceed the capabilities of the computer system being used. More fundamentally, if the raw data are considered, the number of variables measured for each sample will vastly outnumber the number of samples available in the data set. These overdetermined systems can defeat many chemometric techniques due, for example, to collinear variables. Finally, for most chromatograms, especially multidimensional ones, only a small fraction of the data points actually contain meaningful signal. Most of the signal is due to background noise or irrelevant matrix components. Consequently, the raw data must somehow be reduced in size prior to chemometric analysis. This is typically achieved via a feature selection approach, as discussed in Section 3.3.3.

3. Pre-processing steps for chromatographic data

3.1 Baseline correction

The aim of baseline correction is to separate the analyte signal of interest from signal which arises due to changes in mobile phase composition or stationary phase bleed and signal due to electronic noise. Several baseline correction methods have been proposed in literature,

with the two most common approaches being to fit a curve to the data and subtract this value from the signal, and modeling the baseline to exclude it using factor models (Amigo et al., 2010).

Curve fitting is the classical approach used in virtually all commercial software packages provided by vendors of separations equipment. The algorithms used in this approach fit a polynomial function across segments of the chromatogram using regions where no analyte peaks elute to determine the coefficients of the polynomial and then interpolating the background signal for regions where peaks are eluting. The functions are usually first-order polynomials; however, higher-order polynomials or a series of connected first-order polynomials are also used in some situations. Having determined the equation of the background signal, the fitted line is then subtracted from the signal (Brereton, 2003; Gan et al., 2006; Kaczmarek et al., 2005; Zhang et al., 2010; Persson & Strang, 2003; Eilers, 2003). Correction of the baseline using curve fitting is demonstrated in Figure 3.

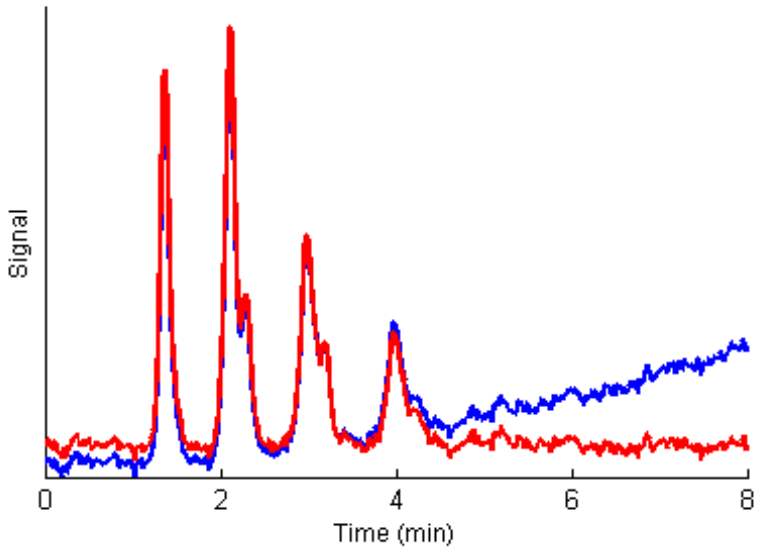


Fig. 3. An LC chromatogram before (blue) and after (red) baseline correction.

The approach of using models such as parallel factor analysis (PARAFAC) for background correction is analogous to the use of these approaches for deconvoluting coeluting peaks. As these models are more often used for this purpose than for simple background correction, they will be discussed in more detail in Section 3.3. These approaches often rely on having a multivariate signal and are applied to the chromatogram or more typically small selected regions where a single analyte elutes. The result of applying these deconvolution techniques for background correction is essentially the deconvolution of a single analyte peak, with the background noise making up the error matrix (Amigo et al., 2010). These approaches are generally more powerful and likely result in better quality analytical data, but they are not widely used in separation science. The reason for this is likely historical as these tools have

only recently become available to the separation sciences, while the classical curve fitting approach is well established, works with univariate detectors, and performs well in most practical situations.

3.2 Alignment of separations data

The retention times of analytes in separations fluctuate from one analytical run to the next and, in order for chemometric techniques to be applied to separations data, these fluctuations must be corrected during pre-processing. This ensures that the signal from each analyte in each analysis is correctly registered within the data matrix to be processed. There are essentially two approaches to this problem: integrated peak tables, or mathematical warping and alignment of the raw signal.

3.2.1 Peak tables

Integrated peak tables are the simplest way to ensure that analytical separations data are properly aligned for chemometric processing. In order to use this approach, one must be able to reliably assign a unique identifier to each peak in each sample of the data set, and ensure that the same compound is identified with the same identifier in each sample. It should be noted that while the compound name is an obvious identifier, a series of labels such as *Unknown x*, where *x* is a numerical identifier would also be acceptable in the event that compound names were unknown, so long as compounds are matched correctly. Rather than identifying peaks by retention time, one could use relative retention times or retention indices in order to adjust for slight variations in the retention times of peaks. Algorithms for aligning peak tables exist and perform well, so long as some peaks can be easily and reliably matched across all chromatograms (Lavine et al., 2001).

The challenges with this approach stem from its reliance on integrated peak tables. Thus, any integration errors due to poorly-resolved peaks or peaks that are missed due to falling outside of integration parameters in the software will impact any subsequent analysis.

3.2.2 Raw signal alignment

Alignment of raw chromatographic signals prior to chemometric processing is more complex than the alignment of peak tables. In addition to the three more popular algorithms that will be presented below, there are several others that have been developed (Yao et al., 2007; Toppo et al., 2008; Eilers, 2004; Van Nederkassel et al., 2006). In deciding which approach to use, one of the first questions to be answered is if the analysis is to be qualitative or quantitative. This is because some alignment methods can distort peaks, affecting their quantification. Some of the more common algorithms include correlation optimized warping (COW) (Nielsen et al., 1998; Tomasi et al., 2004), correlation optimized shifting (coshift) (Van den Berg, 2005), and a piecewise peak-matching algorithm (Johnson et al., 2003).

In instances where there are non-systematic peak shifts, COW is a popular algorithm. COW relies on stretching or compressing segments of a sample signal such that the correlation coefficient between it and a reference signal is maximized for each interval. Care must be taken with the selection of the input parameters to avoid significant changes in peak shapes

as this approach to the warping of the chromatogram has been shown to affect peak areas, leading to poor quantitative conclusions (Nielsen et al., 1998; Tomasi et al., 2004).

A fast and simple alignment algorithm is coshift. This algorithm is useful when data only require a single left-right shift in retention time. The entire data matrix is shifted in one direction or the other by a set amount, maximizing the correlation between a target and the data matrix that required alignment. The single shifting value for the entire data matrix is a weakness, especially for chromatographic data where peaks can shift in different directions and to different extents in a single file. To handle this, an algorithm termed icoshift (interval-correlation-shifting) has been derived from coshift. Icoshift aligns each data matrix to a target by maximizing the cross-correlation between the sample and the target within a series of user-defined intervals (Savorani et al., 2010). The use of multiple intervals permits the alignment of separations data where shifts of different magnitudes and directions occur. These alignment algorithms have been used successfully for both one-dimensional data (de la Mata-Espinosa, 2011a; Liang, 2010; Laursen, 2010) and two-dimensional data, with some modifications (Zhang, 2008). It is important to note that the shifting of chromatograms using coshift or icoshift does not lead to distortions of peak shape, and consequently does not introduce errors into quantitative results.

The piecewise peak matching approach (Johnson et al., 2003) provides another avenue for chromatographic alignment. In this approach, peaks are identified in a target signal to which all other signals will be aligned. The algorithm then identifies peaks within the sample signals located within predetermined windows of the peaks in the target. Peaks within windows are deemed to come from the same compound, and matched. The chromatograms are aligned by stretching or compressing the regions between peak apexes. A variant of this algorithm can be used when MS data are available. In this case, the mass spectrum at the apex of each peak in the target signal is compared to the mass spectrum of each peak within a set window on the sample signal and peaks are matched if their spectra have a high enough match quality (Watson et al., 2006). A general scheme for peak alignment using this approach is described in Figure 4. Depending on the number and relative positions of the peaks in chromatograms matched using this approach, peak shapes may be altered, possibly affecting quantitative results.

One of the biggest challenges for all alignment algorithms is that they depend on the data to be aligned being reasonably similar in terms of both matrix and analyte peaks. In some instances this will not be the case. In our laboratory, we have observed this when analyzing arson debris where the matrix and analytes form an incredibly complex and variable chromatogram from one sample to the next. A similar situation can be easily imagined when processing samples of biological origin. One solution to this issue is to add markers to every sample prior to the separation step in the analysis. These markers should be easily identifiable within the samples, even under conditions where they coelute with matrix components; should occur in multiple, evenly distributed locations along the chromatogram, and should not occur natively in the samples. One choice is a series of deuterated compounds which, with MS detection, are trivial to identify even in a complex mixture (Sinkov et al., 2011b). One additional benefit is that these compounds can act as internal standards if quantitative results are desired.

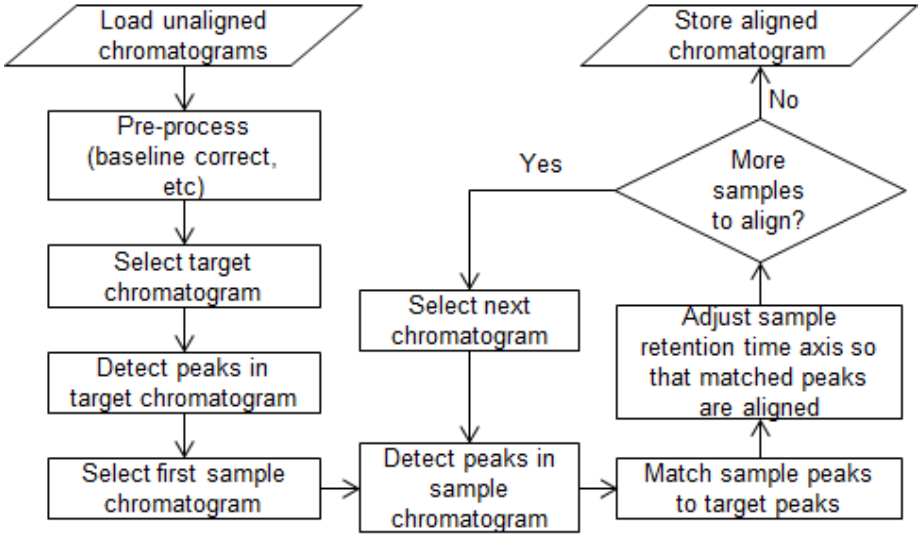


Fig. 4. Flowchart for target-based chromatographic alignment, adapted from (Johnson et al., 2003).

3.3 Deconvolution of overlapping peaks

The central issue in deconvolution is depicted in Figure 5. The instrument response is represented as a black solid line which is the sum of the four dashed, coloured peaks. Ideally, the four signals should be individually quantified. This is a common problem for analytical separations, even those of relatively simple mixtures. Some of these issues may be solved by changing the experimental conditions or using characteristic features (wavelengths or ions) of the coeluting analytes and a multivariate detector to selectively detect and quantify them. However, in many cases this is insufficient and more advanced techniques must be used. The strategies used for deconvolution depend heavily on whether the detector signal is univariate or multivariate.

3.3.1 Deconvolution of univariate signals

In the case of univariate signals, one is typically limited to using univariate curve-fitting analyses where a number of Gaussian or modified Gaussian curves are determined such that the sum of these curves fits the experimentally observed cluster of peaks (Felinger, 1994). In these approaches, only a small window of chromatographic data (one peak cluster) should be processed at a time, and constraints such as fixed peak widths, shapes, unimodality, and non-negativity are often required to ensure the validity of the solution.

To solve a univariate deconvolution problem, approaches such as evolving factor analysis (EFA) (Maeder, 1987) or multivariate curve resolution (MCR) (Tauler & Barceló, 1993), among others (Vivó-Truyols et al., 2002; Sarkar et al., 1998; Kong et al. 2005) can be used. When these approaches are used with univariate data, the variables to be solved for are the number, positions, and abundances of each of the peaks that make up the signal.

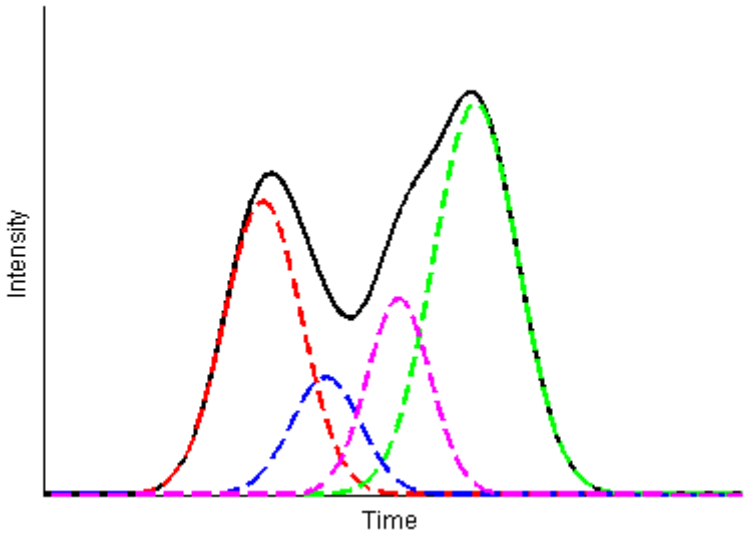


Fig. 5. Deconvolution of overlapping peaks. The black, solid trace represents the analytical signal observed at the detector, which is the sum of the four peaks represented by dashed lines.

Multivariate curve resolution is widely applicable to separations data and is one of the most common approaches (Franch-Lage et al., 2011; Marini et al., 2011, de la Mata-Espinosa et al., 2011a). The aim of this technique is to determine the number of components present in a sample and the contribution of each component to the sample. In performing MCR, the concentration and response profiles for each analyte are obtained, providing a qualitative and semi-quantitative overview of the components in an unresolved mixture without *a priori* knowledge of the mixture composition.

3.3.2 Deconvolution of multivariate signals

When multivariate detectors are used for separations, the additional dimension of information can be exploited to aid in deconvolution. MCR and EFA can also be used with multivariate data. In the case of MCR, the experimental matrix is decomposed into a matrix of concentration vs. time profiles (deconvoluted peaks) and pure spectral profiles of each compound. Knowledge of the number of components contributing to the signal in the region being deconvoluted is useful to guide the process and improve the results (de Juan & Tauler, 2006), though strictly speaking it is not required.

Parallel factor analysis (PARAFAC) (Harshman, 1970; Bro, 1997; Amigo et al., 2010) is a technique that is ideally suited for interpreting multivariate separations data. PARAFAC is a decomposition model for multivariate data which provides three matrices, **A**, **B** and **C** which contain the scores and loadings for each component. The residuals, **E**, and the number of factors, *r*, are also extracted. The PARAFAC decomposition finds the best

trilinear model that minimizes the sum squares of the residuals in the model through a procedure of alternating least squares.

The biggest advantage of using PARAFAC over other models is the uniqueness of the solution; PARAFAC is less flexible and uses fewer degrees of freedom, being a more restricted model. However, its unique solution reflects actual pure analyte profiles in both the time dimension and the spectral dimension. Thus, the results of PARAFAC analysis on a cluster of overlapping multivariate peaks provide both qualitative and quantitative data where the deconvoluted signals appear as analyte peaks. One restriction to the use of PARAFAC is that the data must be trilinear (Bro, 1997; Amigo et al., 2010). In the case of chromatographic techniques with a multivariate detector, the dimensions are retention time, detector signal, and samples. In the case of comprehensive multidimensional separations, such as GC×GC, PARAFAC considers retention in the two dimensions and the samples as the three dimensions.

3.4 Feature selection

High data acquisition rates combined with the length of time required for many separations results in a large number of data points collected for a given separation (see Section 1.2). In many situations, most of the data are collected when no analytes are eluting from the system, and represent background signal when only mobile phase is reaching the detector. In the case of spectroscopic and especially mass spectral detectors, at a given point in time, many of the recorded data in this dimension will not contain useful information, even when an analyte of interest is eluting. Furthermore, many components in the mixture can be completely irrelevant to analysis (Johnson & Synovec, 2002; Sinkov & Harynuk, 2011a). Consequently, only a small portion of separations data is potentially useful. It is also well known that any model will be heavily influenced by the specific variables that are included in its construction (Kjeldahl & Bro, 2010).

The inclusion of irrelevant data is detrimental to the model because the mathematics attempt to account for variations observed in these irrelevant variables. Consequently the model is forced to model noise, resulting in a decrease in its predictive ability. Worse yet, the model could fit the data well and provide a seemingly useful prediction, until cross-validation shows otherwise. Finally, the inclusion of extraneous variables increases the demands on the computer system being employed, making model construction slower, or in some cases outright impossible. Thus, prior reduction of separations data to a manageable size is crucial. Figure 6 depicts situations where either too few or too many variables were used to model a system.

One common manner to achieve data reduction is to use a table of integrated peaks instead of raw chromatographic data. This has the advantage of reducing the number of variables to those compounds included in the peak list, removing baseline noise and, if the analyst knows which exact peaks to use, removing signal from irrelevant compounds. Problems with this approach include the restriction to identified compounds, which may or may not include all of the information required for modeling, and integration errors that skew results. Finally, even with an error-free comprehensive peak table, the analyst must still perform feature selection since many peaks will undoubtedly be irrelevant to the analysis.

In the case of multivariate detection, it can be advantageous to monitor only one or a few channels (wavelengths, ions, etc.) as this will selectively detect only a portion of the analytes, allowing the analyst to avoid many interfering species while greatly reducing the size of the data. However, in these cases the analyst must know exactly what signals to use and runs the risk of missing important features of the data encoded in the channels that were ignored. Further, using this approach destroys much of the multivariate advantage that can be realized through using these more complex (and expensive) detection strategies.

Objective feature selection techniques generally have two steps: variable ranking, and variable selection. Objective variable ranking techniques such as analysis of variance (ANOVA) (Johnson & Synovec, 2002), the discriminating variable test (DIVA) (Rajalahti et al., 2009a, 2009b), and informative vectors (Teofilo et al., 2009) have the distinct advantage that variables are ranked based on a mathematically calculable "perceived utility" and not on subjective analyst perception. In essence, the data are given the chance to inform the user of what is relevant and what is likely noise, providing an approach that can be generalized to any set of analytical data.

ANOVA is an effective method when the goal is to discriminate between classes of samples. ANOVA calculates the F ratio for each variable: the ratio of between-class variance to within-class variance. If the F ratio for a given variable is high, it is deemed to be more valuable for describing the difference between classes. Once the F ratio has been calculated for every data point in the chromatogram, the variables can be ranked in order of decreasing F ratio. A chemometric model is then constructed using a fraction of variables having the highest F ratio. One significant advantage of ANOVA is that the algorithm can be written with memory conservation in mind and thus is easily applied to data sets with very large numbers of samples and variables (hundreds or thousands of samples, each containing millions of variables). Consequently, it can be easily applied to a set of GC-MS chromatograms across the entire chromatogram, something that is difficult for other feature ranking approaches.

DIVA is a feature ranking technique that aids feature selection prior to chemometric analysis (Rajalahti et al., 2009a, 2009b). This approach involves the creation of a PLS-DA model using all candidate variables. Projecting this PLS-DA model onto a new single LV yields what is termed a target projected (TP) model (Rajalahti et al., 2009a). From this, the ratio of explained variance to residual variance for each variable in the TP model provides its selectivity ratio, upon which variables are ranked (Rajalahti et al., 2009a, 2009b; Kvalheim, 1990; Kvalheim & Karstang, 1989). DIVA produces a ranking that is slightly different than that produced by ANOVA, though a direct comparison on chromatographic data has not yet been performed to our knowledge.

Once variables have been ranked, those to be included in the model must be selected. This is generally achieved by constructing a model using a forward-selection or backwards elimination approach, in an attempt to maximize some metric of model quality. Model quality can be assessed based on several metrics such as mean correct classification rates (Rajalahti et al., 2009b) or the degree of separation between classes of samples in principal component (or latent variable) space, for example using either a Euclidian distance-based metric (Pierce et al., 2005) or a metric that accounts for size and shape of clusters (Sinkov & Harynuk, 2011a).

The one exception to the rank-and-select approach are genetic algorithms (Yoshida et al., 2001), though due to the sheer number of variables present in a typical separation, these are not often used on the raw separations data as arriving at the optimal number and combination of variables is computationally inefficient and uncertain.

Sometimes, several feature selection methods are used for a given analysis. For example, an analyst might reduce chromatogram to a peak table, selecting a series of candidate variables of interest and then perform further variable ranking and optimization on the integrated peak table, especially in the case of multidimensional separations where hundreds, if not thousands of compounds can be resolved (Felkel et al., 2010).

Finally, cross-validation is extremely important, especially when processing raw separations data and using a feature ranking approach such as ANOVA. As discussed previously, raw separations data contain on the order of 10^5 to 10^6 data points for each sample. In these cases of overdetermined systems it is entirely possible that some combinations of variables containing only noise will, by random chance, indicate a difference between samples. When handling raw separations data, a good approach to avoid this problem is to break the data set into three separate sets: a training set to construct the model, an optimization set to optimize data processing parameters (such as alignment and feature selection), and finally a test set to determine if the optimized model has any meaning (Brereton, 2007). Of course this does require that one collect data for a large number of samples so that a representative population of samples exists for each of the three subsets of data.

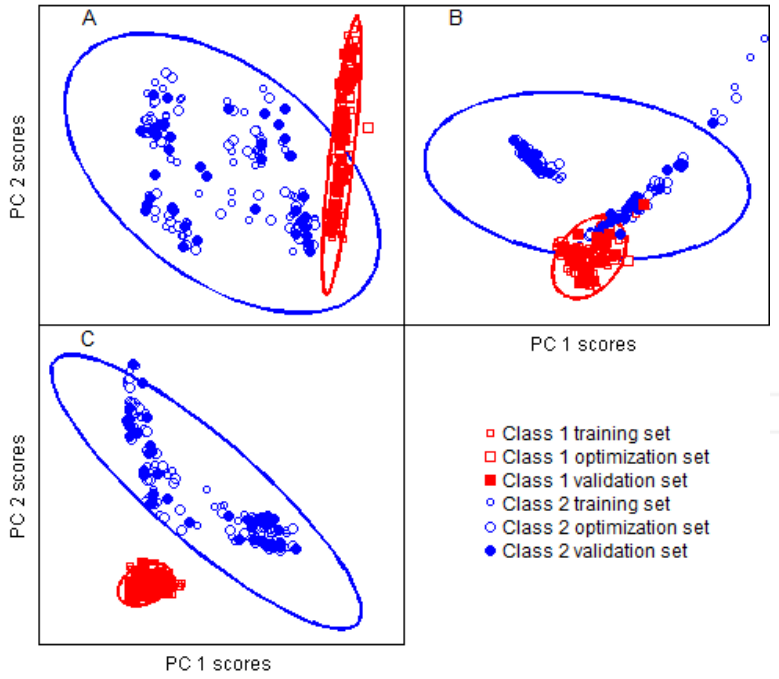


Fig. 6. Models constructed from the same data set using different numbers of top-ranked variables. (A) Too few variables; (B) Too many variables; (C) Optimal number of variables.

4. Applications and examples

After applying the appropriate pre-processing, different chemometric techniques can be applied according to the aim of the study. Pattern recognition is one of the chemometric methods most used in analytical chemistry and this is true for separations data. Pattern recognition can be generally divided into two classes: exploratory data analysis and unsupervised and supervised pattern recognition (Otto, 2007; Brereton, 2007).

Exploratory data analysis aims to extract important information, detect outliers and identify relationships between samples and its use is recommended prior to the application of other chemometric techniques. Examples of the use of exploratory data analysis tools applied to separations data include principal component analyisis (PCA) (de la Mata-Espinosa et al., 2011a; Ruiz-Samblas et al., 2011) and factor analysis (Stanimirova et al., 2011).

Unsupervised pattern recognition techniques uncover patterns within a data set without *a priori* class assignment of samples. Here, the objective is to find patterns in the data which allow grouping of similar samples using, for example, cluster analysis which has been applied to separations data by Reid et al. (2007). When supervised pattern recognition is used, the classes of samples in a training set are known and used to calibrate a model, which is then used to predict class assignments of unknown samples. Some examples of which are linear discriminant analysis (LDA), and partial least squares-discriminant analysis (PLS-DA) (de la Mata-Espinosa et al., 2011b; Zorzetti et al., 2011; Sinkov et al., 2011b). In a study performed by Sinkov et al., two alignment techniques for chromatographic data were compared. The data comprised raw GC-MS chromatograms of simulated arson debris where some samples contained different types of gasoline weathered to different extents spiked into debris samples which themselves exhibited a high degree of variability in their chemical composition. The goal was to build a PLS-DA model that could correctly classify debris samples based on whether or not they contained gasoline (Figure 7). As can be seen, the alignment algorithm used has a direct impact on the quality of the predictions. In Figure 7A, there are multiple false positives, false negatives, and ambiguous samples. In Figure 7B, all samples are classified correctly and there are no ambiguous samples.

Another example of applying chemometrics to separations data is depicted in Figures 8 and 9. Here, interval PLS (iPLS) was applied to blends of oils in order to quantify the relative concentration of olive oil in the samples (de la Mata-Espinosa et al., 2011b). iPLS divides the data into a number of intervals and then calculates a PLS model for each interval. In this example, the two peak segments which presented the lower root mean square error of cross validation (RMSECV) were used for building the final PLS model.

As mentioned in Section 3.3.2, PARAFAC is a chemometric tool for multidimensional data treatment. The scores and loadings obtained with PARAFAC can be used in two-way models for data exploration and quantitative analysis (Vosough et al., 2010). When small deviations in trilinearity exist within the data, usually due to relatively small shifts in retention time in the case of separations data, a modified version of PARAFAC called PARAFAC2 is recommended for use (Bro et al., 1999).

Like PARAFAC, PARAFAC2 decomposes raw data into loading and score matrices but without the imposition of trilinearity as in PARAFAC. Even without this constraint, the PARAFAC2 model preserves the property of uniqueness that is so advantageous with PARAFAC. Thus, analyte profiles and concentrations can be estimated by PARAFAC2 even if chromatographic alignment is not perfect (Amigo et al., 2008; Skov et al., 2009).

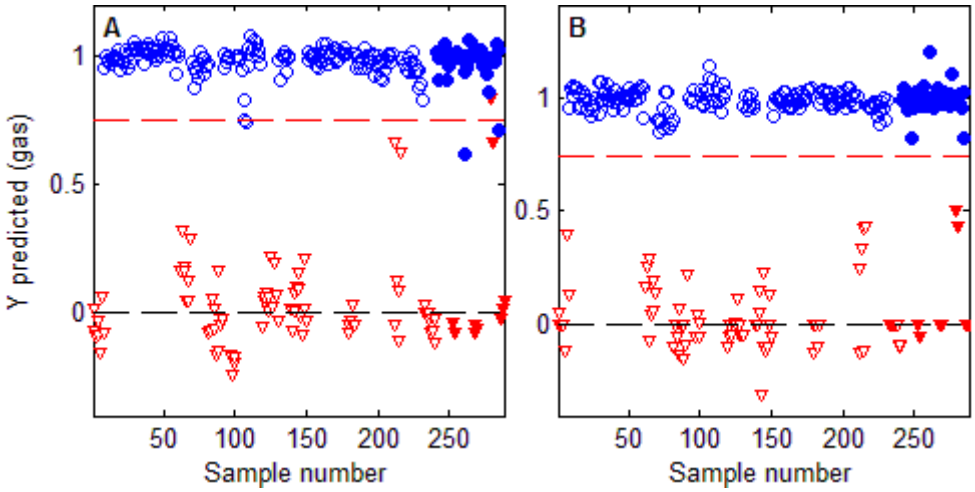


Fig. 7. PLS-DA Models for identifying gasoline in simulated arson debris derived from the same raw data, but aligned with different techniques. (A) Feature-based alignment; (B) Deuterated alkane ladder – based alignment. All other treatment and model construction algorithms were the same in both cases. Hollow markers indicate data in the training set while filled markers indicate data in the validation set. Circles represent debris containing gasoline while triangles represent gasoline-free debris. Reprinted from Sinkov et al., 2011b, with permission.

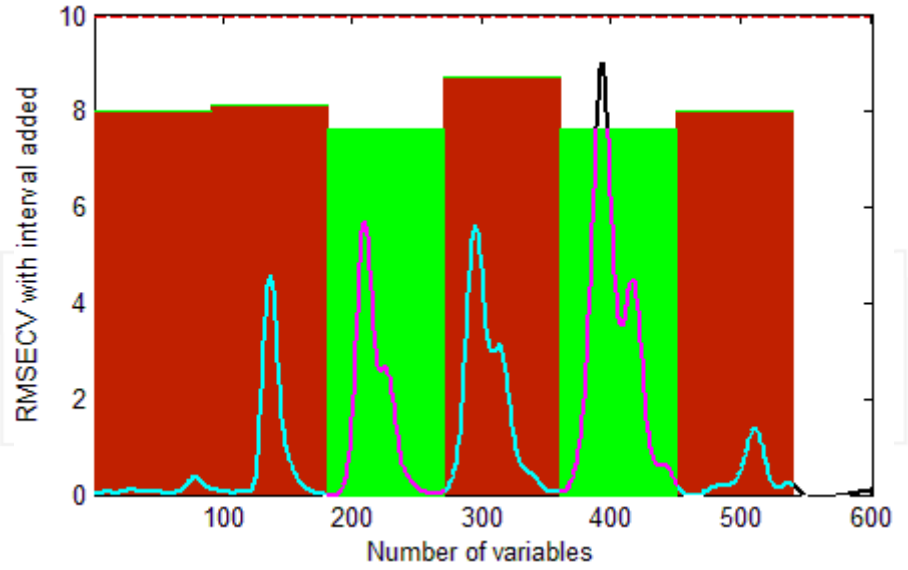


Fig. 8. Feature selection using iPLS. Segments in green showed lower RMSECV and were thus used to construct the final model. Reprinted from de la Mata-Espinosa et al., 2011b, with permission.

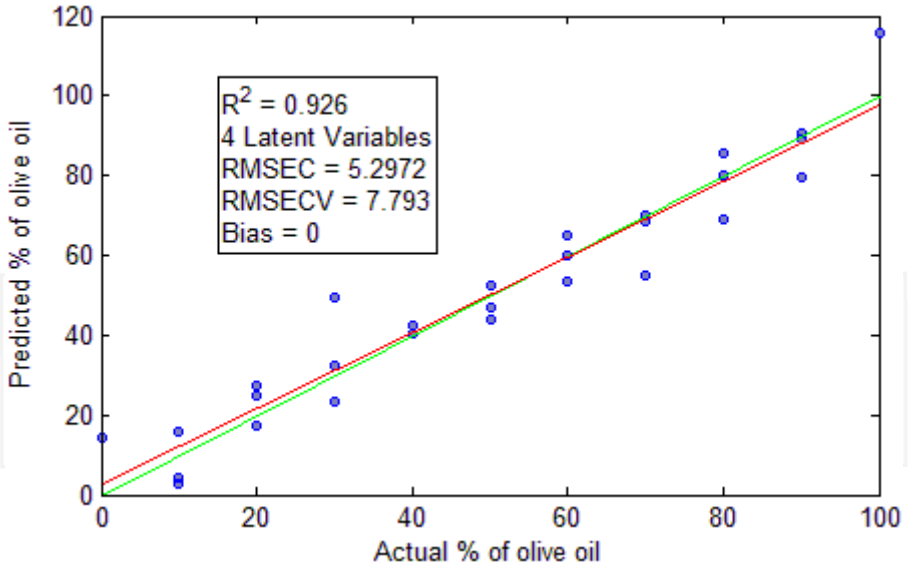


Fig. 9. Predicted vs. actual % olive oil using PLS model constructed based on results in Figure 8. Reprinted from de la Mata-Espinosa et al., 2011b, with permission.

5. Conclusions

The analyst must choose from a plethora of methods for processing separations data, a potentially daunting task. It is our hope that this review will help chromatographers entertaining thoughts of applying chemometrics to their data understand what they must consider when choosing how to prepare their data. Likewise, it is hoped that we have informed chemometricians of some of the specific challenges associated with the processing of chromatographic data and the origins of those limitations. In the development of a chemometric model for the interpretation of separations data, there are numerous opportunities for missteps that will exclude key information from the model and/or generate meaningless results. However, when due care is taken there are also many opportunities to apply chemometric techniques to transform the rich data generated by these powerful analytical tools into valuable information effectively and efficiently.

6. References

Amigo, J.M.; Skov, T.; Bro, R.; Coello, J. & Maspoch, S. (2008). Solving GC-MS problems with PARAFAC2. *Trends in Analytical Chemistry,* Vol.27, No.8, (September 2008), pp. 714-725, ISSN 0165-9936

Amigo, J.M.; Skov, T. & Bro, R. (2010). ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics. *Chemical Reviews*, Vol.110, No.8, (May 2010), pp. 4582-4605, ISSN 1520-6890

Asher, B.J.; D'Angostino, L.A.; Way, J.D.; Wong, C.S. & Harynuk, J.J. (2009). Comparison of peak integration methods for the determination of enantiomeric fraction in

environmental samples. *Chemosphere*, Vol.75, No.8, (May 2009), pp. 1042-1048, ISSN 0045-6535

Brereton, R.G. (2003). *Chemometrics Data Analysis for the Laboratory and Chemical Plant,* Wiley, ISBN 0-474-78977-8, UK

Brereton, R.G. (2007). *Applied Chemometrics for Scientists*, John Wiley & Sons Inc., ISBN 978-0-470-01686-2, Toronto, Canada

Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics Intelligent Laboratory Systems,* Vol.38, No.2, (October 1997), pp. 149-171, ISSN 0169-7439

Bro, R.; Andersson, C.A.& Kiers, H.A.L. (2009). PARAFAC-Part II. Modeling chromatographic data with retention times shifts. *Journal of Chemometrics,* Vol.13, No.3-4, (May-August 1999), pp. 295-309, ISSN 0886-9383

Casez, J. (2010). *Encyclopaedia of Chromatography*, (3rd ed.) CRC Press, ISBN 1-4200-8483, Florida, USA

Cortes, H.J.; Winniford, B.; Luong, J. & Pursch, M. (2009). Comprehensive two dimensional gas chromatography review. *Journal of Separation Science,* Vol.32, No.5-6, (March 2009), pp. 883-904, ISSN 1615-9306

de Juan, A. & Tauler, R. (2006). Multivariate Curve Resolution (MCR) from 2000: Progress in concepts and applications. *Critical Reviews in Analytical Chemistry,* Vol.36, No.3-4, (2006) pp. 163-176, ISSN 1040-8347

de la Mata-Espinosa, P.; Bosque-Sendra, J.M.; Bro, R. & Cuadros-Rodríguez, L. (2011a). Discriminating olive and non-olive oils using HPLC-CAD and chemometrics. *Analytical and Bioanalytcial Chemistry,* Vol.399, No.6, (February, 2011), pp. 2083-2092, ISSN 1618-2650

de la Mata-Espinosa, P.; Bosque-Sendra, J.M.; Bro, R. & Cuadros-Rodríguez, L. (2011b). Olive oil quantification of edible vegetable oil blends using triacylglycerols chromatographic fingerprints and chemometric tools. *Talanta,* Vol.85, No.1, (July 2011), pp. 183-196, ISSN 0039-9140

Eilers, P.H.C. (2003). A perfect Smoother. *Analytical Chemistry,* Vol.75, No.14, (July 2003) pp. 3631-3636, ISSN 0003-2700

Eilers, P.H.C. (2004). Parametric Time Warping. *Analytical Chemistry,* Vol.76, No.2, (January 2004), pp. 404-411, ISSN 0003-2700

Erni, F. & Frei, R.W. (1978). 2-Dimensional column liquid-chromatographic technique for resolution of complex mixtures. *Journal of Chromatography,* Vol.149, (February 1978), pp. 561-569 ISSN 0021-9673

Etxebarria, N.; Zuloaga, O.; Olivares, M.; Bartolomé. L.J. & Navarro, P. (2009). Retention-time locked methods in gas chromatography. *Journal of Chromatography A*, Vol.1216, No.10, (March 2009), pp. 1624-1629 ISSN 0021-9673

Felinger, A. (1994). Deconvolution of overlapping skewed peaks. *Analytical Chemistry,* Vol.66, No.19, (October 1994), pp. 3066-3072, ISSN 0003-2700

Felkel, Y.; Dorr, N.; Glatz, F. & Varmuza, K. (2010). Determination of the total acid number (TAN) of used gas engine oils by IR and chemometrics applying a combined strategy for variable selection. *Chemometrics and Intelligent Laboratory Systems,* Vol. 101, No. 1, (March, 2010), pp. 14-22 ISSN 0169-7439

Franch-Lage, F.; Amigo, J.M.; Skibsted, E.; Maspoch, S. & Coello, J. (2011). Fast assessment of the surface distribution of API and excipients in tablets using NIR-hyperspectral

imaging. *International Journal of Pharmaceutics,* Vol.441, No.1-2, (June 2011), pp. 27-35, ISSN 0378-5173

François, I.; Sandra, K. & Sandra, P. (2009). Comprehensive liquid chromatography: Fundamental aspects and practical considerations—A review. *Analytica Chimica Acta,* Vol.641, No.1-2, (May 2009), pp. 14-31, ISSN 0003-2670

Gan, F.; Ruan, G. & Mo, J. (2006). Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems,* Vol.82, No.1 (May 2006), pp. 59-65, ISSN 0169-7439

Górecki, T.; Harynuk, J. & Panić, O. (2004). The evolution of comprehensive two-dimensional gas chromatography, *Journal of Separation Science,* Vol.27 (2004) pp. 359-379, ISSN 1615-9306

Harshman, R.A. (1970). Foundations of the PARAFAC procedure: models and conditions for an 'exploratory' multimodal factor analysis. *UCLA Working Papers Phonet.* Vol 16, (1970), pp. 1-84

Johnson, K.J. & Synovec, R.E. (2002). Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, Vol.60, No.1-2, (January 2002), pp. 225-237, ISSN 0169-7439

Johnson, K.J.; Wright, B.W.; Jarman, K.H. & Synovec, R.E. (2003). High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *Journal of Chromatography A*, Vol.996, No.1-2, (May 2003), pp. 141-155, ISSN 0021-9673

Kaczmarek, K.; Walczak, B.; de Jong, S. & Vandeginste, B.G.M. (2005). Baseline reduction in two dimensional gel electrophoresis images. *Acta Chromatographica,* Vol.15 (2005), pp. 82-96, ISSN 1233-2356

Kivilompolo, M.; Pol, J. & Hyotylainen, T. (2011). Comprehensive two-dimensional liquid chormatography (LC×LC): A review. *LC GC Europe*, Vol.24, No 5 (May 2011), pp. 232-+, ISSN 1471-6577

Kjeldahl, K. & Bro, R. (2010). Some common misunderstandings in chemometrics. *Journal of Chemometrics*, Vol.24, No.7-8, (July-August, 2011), pp. 558-564, ISSN 0886-9383

Kong, K.; Ye, F.; Guo, L.; Tian, J. & Xu, G. (2005). Deconvolution of overlapped peaks based on the exponentially modified Gaussian model in comprehensive two-dimensional gas chromatography, *Journal of Chromatography A,* Vol.1086, No.1-2 (September 2005) pp. 160-164, ISSN 0021-9673

Kvalheim, O.M. & Karstang, T.V. (1989). Interpretation of latent-variable regression models. *Chemometrics and Intelligent Laboratory Systems*, Vol.7, No.1-2, (December 1989), pp. 39-51, ISSN 0169-7439

Kvalheim, O.M. (1990). Latent-variable regression models with higher-order terms: An extension of response modelling by orthogonal design and multiple linear regression. *Chemometrics and Intelligent Laboratory Systems*, Vol.8, No.1, (May 1990), pp. 59-67, ISSN 0169-7439

Lavine, B.K.; Brzozowski, D.; Moores, A.J.; Davidson, C.E. & Mayfield, H.T. (2001). Genetic algorithm for fuel spill identification. *Analytica Chimica Acta*, Vol.437, No.2, (June 2001), pp. 233-246, ISSN 0003-2670

Laursen, K.; Frederiksen, S.S.; Leuenhagen, C. & Bro, R. (2010). Chemometric quality contrl of chromatographic purity. *Journal of Chromatography A,* Vol.1217, No.42 (October 2010), pp. 6503-6510, ISSN 0021-9673

Li, Y.H.; Wojcik, R & Dovichi, N.J. (2011). A replaceable microreactor for on-line protein digestion in a two dimensional capillary electrophoresis system with tandem mass spectrometry detection. *Journal of Chromatography A*, Vol.1218, No.15 (April 2011), pp. 2007-2011, ISSN 0021-9673

Liang, Y.; Xie, P. & Chau, F. (2010). Chromatographic fingerprinting and related chemometric techniques for quality control of traditional Chinese medicines. *Journal of Separation Science,* Vol.33, No.3 (February 2010), pp. 410-421, ISSN 1615-9314

Liu, Z. & Phillips, J.B. (1991). Comprehensive 2-dimensional gas-chromatography using a modulator interface. *Journal of Chromatographic Science*, Vol.29, No.6 (June 1991), pp. 227-231, ISSN 0021-9665

Maeder, M. (1987). Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Analytical Chemistry,* Vol.59, No.3, (February 1987), pp 527-530, ISSN 0003-2700

Marini, F.; D'Aloise, A.; Bucci, R.; Buiarelli, F.; Magri, A.L. & Magri, D. (2011), Fast analysis of 4 phenolic acids in olive oil by HPLC-DAD and chemometrics, *Chemometrics and Intelligent laboratory systems,* Vol.106, No.1, (March 2011), pp. 142-149, ISSN 0169-7439

Michels, D.A.; Hu, S.; Schoenherr, R.M.; Eggertson, M.J. & Dovichi, N.J. (2002), Fully automated two-dimensional capillary electrophoresis for high sensitivity protein analysis, *Molecular & Cellular Proteomics,* Vol.1, No.1, (January 2002), pp. 69-74, ISSN 1535-9476

Miller, J.M. (2005). *Chromatography: concepts and contrasts*, (2nd ed.) Wiley, ISBN 0471472077, Hoboken, USA

Mommers, J.; Knooren, J.; Mengerink, Y.; Wilbers, A.; Vreuls, R. & van der Wal, S. (2011). Retention time locking procedure for comprehensive two-dimensional gas chromatography. *Journal of Chromatography A*, Vol.1218, No.21 (May, 2011), pp. 3159-3165 ISSN 0021-9673

Nielsen, N-P.; Cartensen, J.M. & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A,* Vol.805, No.1-2 (May 1998), pp. 17-35 ISSN 0021-9673

Otto, M. (2007). *Chemometrics*, Wiley-VCH, ISBN 978-3-527-31418-8, Weinheim, Germany

Persson, P.O. & Strang, G. (2003). Smoothing by Savitzky-Golay and Legendre filters, In: *Mathematical Systems Theory in Biology, Communications, Computation and Finance,* Rosenthal J. Gilliam D.S., pp. 301-315, IMA Vol. Math. Appl., 134, Springer, ISBN 978-0387-40319-9, New York, USA

Pierce K.M.; Hope J.L.; Johnson K.J.; Wright B.W. & Synovec R.E. (2005). Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. *Journal of Chromatography A*, Vol.1096, No.1-2, (November 2005), pp. 101-110, ISSN 0021-9673

Poole, C.F. (2003). *The Essence of Chromatography*, (1st ed.), Elsevier, ISBN 0444501983, Amsterdam, The Netherlands

Rajalahti, T.; Arneberg, R.; Berven, F.S.; Myhr, K.M.; Ulvik, R.J. & Kvalheim, O.M. (2009a). Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and Intelligent Laboratory Systems*, Vol. 95, No. 1, (January 2009), pp. 35-48, ISSN 0169-7439

Rajalahti, T.; Arneberg, R.; Kroksveen, A.C.; Berle, M.; Myhr, K.M. & Kvalheim, O.M. (2009b). Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles. *Analytical Chemistry*, Vol. 81, No. 7, (April 2009), pp. 2581-2590, ISSN 0169-7439

Reid, R.G.; Durham, D.G.; Boyle S.P.; Low, A.S. & Wangboonskul, J. (2007). Differentiation of opium and poppy straw using capillary electrophoresis and pattern recognition techniques. *Analytica Chimica Acta*, Vol.605, No. 1, (December 2007), pp. 20-27, ISSN 0003-2670

Ruiz-Samblas, C.; Cuadros-Rodriguez, L.; Gonzalez-Casado, A.; Rodriguez Garcia, F.D.P; de la Mata-Espinosa, P.; Bosque-Sendra, J.M. (2011). Multivariate analysis of HT/GC-(IT)MS chromatographic profiles of triacylglycerols for classification of olive oil varieties, *Analytical and Bionalytical Chemistry*, Vol.399, No.6 (February 2011), pp. 2093-2103, ISSN 1618-2642

Sarkar, S.; Dutta, P.K. & Roy, N.C. (1998). A blind-deconvolution approach for chromatographic and spectroscopic peak restoration, *IEEE transactions on instrumentation and measurement,* Vol.47, No.4 (August 1998), pp. 941-947, ISSN 0018-9456

Savorani, F.; Tomasi, G. & Engelsen, S.B. (2010). Icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance,* Vol.202, No.2, (February 2010), pp. 190-202 ISSN 1090-7807

Sinkov, N.A. & Harynuk, J.J. (2011a). Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta*, Vol.83, No.4, (January 2011), pp. 1079-1087, ISSN 0039-9140

Sinkov, N.A.; Johnston, B.M.; Sandercock, P.M.L. & Harynuk, J.J. (2011b). Automated optimization and construction of chemometric models based on highly variable raw chromatographic data. *Analytica Chimica Acta*, Vol.697, No.1-2, (July 2011), pp. 8-15, ISSN 1873-4324

Skov, T.; Hoggard, J.C.; Bro, R. & Synovec, R.E. (2009). Handling within run retention time shifts in two-dimensional chromatography data using shift correction and modeling. *Journal of Chromatography A,* Vol.1216, No.18, (May 2009), pp. 4020-4029, ISSN 0021-9673

Stanimirova, I.; Boucon, C. & Walczak, B. (2011). Relating gas chromatographic profiles to sensory measurements describing the end products of the Maillard reaction. *Talanta,* Vol.83, No 4, (January 2011), pp. 1239-1246, ISSN 0039-9140

Tauler, R. & Barceló, D. (1993). Multivariate curve resolution applied to liquid chromatography-diode array detection. *Trends in Analytical Chemistry,* Vol.12, No.8, (1993), pp. 319-327, ISSN 0165-9936

Teofilo, R.F.; Martins, J.P.A. & Ferreira, M.M.C. (2009). Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression.

Journal of Chemometrics, Vol.23, No.1-2, (January-February 2009), pp. 32-48, ISSN 0886-9383

Tomasi, G.; Van den Berg, F. & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics,* Vol.18, No.5, (May 2004), pp. 231-241, ISSN 0886-9383

Toppo, S.; Roveri, A.; Vitale, M.P.; Zaccarin, M.; Serain, E.; Apostolidis, E.; Gion, M., Mariorino, M. & Ursini, F. (2008). MPA: A multiple peak alignment algorithm to perform multiple comparisons of liquid-phase proteomic profiles. *Proteomics,* Vol.8, No.2, (January 2008), pp. 250-253 ISSN 1615-9861

Van den Berg, F.; Tomasi, G. & Viereck, N. (2005). Warping: investigation of NMR pre-processing and correction, In: *Magnetic Resonance in Food Science: The Multivariate Challenge,* Engelsen, S.B., Belton, P.S., Jakobsen, H.J., pp. 131-138, Royal Society of Chemistry, ISBN 0854046488, Cambridge, UK

Van Nederkassel, A.M.; Dazykowski, M.; Eilers, P.H.C. & Vander Heyden, Y. (2006). A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A,* Vol.118, No.2 (June 2006), pp. 199-210 ISSN 0021-9673

Vivó-Truyols, G.; Torres-Lapasió, J.R.; Caballero R.D. & García-Alvarez-Coque, M.C. (2002). Peak deconvolution in one-dimensional chromatography using a two-way data approach. *Journal of Chromatography A,* Vol.958, No.1-2, (June, 2002), pp. 35-49, ISSN 0021-9673

Vosough, M.; Bayat, M. & Salemi, A. (2010). Matrix-free analysis of aflatoxins in pistachio nuts using parallel factor modeling of liquid chromatography diode-array detection data. *Analytica Chimica Acta*, Vol.663, No.1, (March 2010), pp. 11-18. ISSN 0003-2670

Watson, N.E.; VanWingerden, M.M.; Pierce, K.M.; Wright, B.W. & Synovec, R.E. (2006). Classification of high-speed gas chromatography-mass spectrometry data by principal component analysis coupled with piecewise alignment and feature selection. *Journal of Chromatography A*, Vol.1129, No.1, (September, 2006), pp. 111-118, ISSN 0021-9673

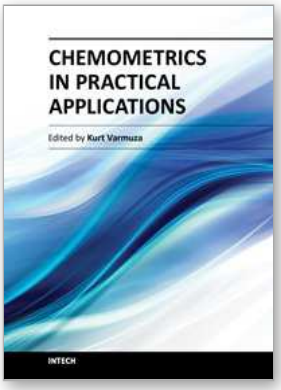
Yao, W., Yin, X. & Hu Y. (2007). A new algorithm of piecewise automated beam search for peak alignment of chromatographic fingerprints. *Journal of Chromatography A,* Vol. 1160, No.1-2, (August 2007), pp. 254-262. ISSN 0021-9673

Yoshida H.; Leardi R.; Funatsu K. & Varmuza K. (2001) Feature selection by genetic algorithms for mass spectral classifiers. *Analytica Chimica Acta*, 446, 1-2, (November 2001), pp. 485-494, ISSN 0003-2670

Zhang D.; Huang, X.; Regnier, F.E. & Zhang, M. (2008). Two-dimensional correlation optimized warping algorithm for aligning GC×GC-MS data. *Analytical Chemistry,* Vol.80, No.8 (April 2008), pp. 2664-2671, ISSN 0003-2700

Zhang, Z.M.; Chen, S. & Liang, Y.Z. (2010). Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst,* Vol.5 (February 2010), pp. 1138-1146, ISSN 0003-2654

Zorzetti, B.M.; Shaver, J.M. & Harynuk, J.J. (2011). Estimation of the age of a weathered mixture of volatile organic compounds. *Analytica Chimica Acta*, Vol.694, No.1-2, (May 2011), pp. 31-37, ISSN 0003-2670



Chemometrics in Practical Applications

Edited by Dr. Kurt Varmuza

ISBN 978-953-51-0438-4
Hard cover, 326 pages
Publisher InTech
Published online 23, March, 2012
Published in print edition March, 2012

In the book "Chemometrics in practical applications", various practical applications of chemometric methods in chemistry, biochemistry and chemical technology are presented, and selected chemometric methods are described in tutorial style. The book contains 14 independent chapters and is devoted to filling the gap between textbooks on multivariate data analysis and research journals on chemometrics and chemoinformatics.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821