Chapter 10

Pixel-Level Data Analysis Methods for Comprehensive Two-Dimensional Chromatography

Karisa M. Pierce*, Brendon A. Parsons† and Robert E. Synovec[1],†
*Department of Chemistry and Biochemistry, Seattle Pacific University, Seattle, Washington, USA
†Department of Chemistry, University of Washington, Seattle, Washington, USA
[1]Corresponding author: e-mail: synovec@chem.washington.edu

Chapter Outline

1 Introduction 427
2 Signal Preprocessing in 2D Chromatography with Pixel-Based Data 429
2.1 Baseline Correction 429
2.2 Noise Reduction 430
2.3 Normalization 431
2.4 Retention Time Alignment 432
3 Pixel-Level Methods for Deconvolution 433
3.1 Aims of Deconvolution 433
3.2 Generalized Rank Annihilation Method 433
3.3 Multivariate Curve Resolution-Alternating Least Squares 437
3.4 Parallel Factor Analysis 441
4 Fingerprinting and Pattern Recognition 447
4.1 Aim of Fingerprinting and Pattern Recognition 447
4.2 Unsupervised Pattern Recognition—PCA 447
4.3 Supervised fingerprinting and pattern recognition 453
References 457

1 INTRODUCTION

Comprehensive two-dimensional (2D) chromatography continues to evolve as an important instrumental platform to address challenges in the analysis of complex samples. The instrumentation primarily utilized is either comprehensive 2D gas or liquid chromatography (e.g., GC × GC or LC × LC) [1–10]. While other comprehensive 2D separation methods have been reported, this chapter focuses primarily on these two commonly applied instruments. Additionally, these instruments are often coupled to multichannel spectroscopic

Data Handling in Science and Technology, Vol. 29. http://dx.doi.org/10.1016/B978-0-444-63527-3.00010-2

detectors, namely time-of-flight mass spectrometry for GC × GC-TOFMS [11] and multiwavelength absorbance detection (i.e., diode array detection) for LC × LC-DAD [8]. In the analysis of complex samples, the resulting data can be very challenging to efficiently and accurately analyze. For this purpose, chemometric data analysis has emerged as the path forward to provide a successful, informative analysis. Many excellent chemometric text books and reviews are available [12–15]. Other chemometrics reviews focused primarily on comprehensive 2D separations data are also available in the literature [16–23]. Indeed, the data obtained from a single run from either GC × GC or LC × LC, when coupled with a univariate detector such as GC × GC-FID or LC × LC with single wavelength absorbance detection, produces second-order data, and third-order data when more than one sample run is simultaneously analyzed, so chemometric analysis is ideally suited to glean useful information from the rich data structure. The situation is even more urgent to apply chemometrics for the data that are inherently third order for a single sample run, i.e., using GC × GC-TOFMS or LC × LC-DAD. There are different approaches to "chemometrically" address data analysis, and in the context of this chapter, we refer to the dependence on the data-level. Data from comprehensive 2D chromatography separations can be analyzed, on the following three levels: pixel [24], peak table [25,26], or peak region [27,28] basis. While all of these three levels of data are utilized [29,30], the peak table-level and peak region-level approaches are firmly provided by the instrument manufacturers, and hence the analyst must apply such software in the commercially available state. Hence, there is generally little room for independent development of chemometric software at these two data levels, whereas at the pixel-level, there has been rapid growth in the development of chemometric tools for creative and informative analysis of comprehensive 2D chromatography separations data. Accordingly, the focus of this chapter is on the pixel-level chemometric analysis of comprehensive 2D chromatography separations data. For example, recently a novel tile-based data analysis approach with its fundamental basis at the pixel-level has been introduced for the comparative analysis of large numbers of GC × GC-TOFMS sample runs [31,32].

This chapter is organized into the following three sections: Sections 2–4. Section 2 provides an overview of the data preprocessing steps that are deemed essential for optimizing the use of pixel-level chemometric analyses of comprehensive 2D chromatography separations data. The following four generally essential preprocessing steps are covered: baseline correction, noise reduction, normalization, and retention time alignment. Additional preprocessing steps that are contingent on the particular chemometric method are discussed in due course. Since overlap of analyte peaks is commonplace in the analysis of complex samples, which are often studied using comprehensive 2D chromatography separation methods, Section 3 covers the basics of commonly applied chemometric deconvolution methods for pixel-level data. For this chapter, we due to consideration of space, we have limited the discussion primarily to generalized rank annihilation method (GRAM), multivariate curve resolution alternating least squares (MCR-ALS), and parallel factor analysis (PARAFAC), with some discussion of related methods. Finally, Section 4 provides an overview of commonly applied chemometric methods for fingerprinting and pattern recognition, in which the analyst is interested in initially comparing a series of samples on a broader scale (based upon the experimental design). In Section 4, we cover two commonly encountered experimental designs: unsupervised design using principal component analysis (PCA) and supervised design using either partial least squares (PLS) methods or Fisher ratio analysis. An overarching goal of this chapter is to provide the reader with the background and knowledge of the fundamentals of chemometric data analysis of comprehensive 2D chromatography separations data, in particular for GC × GC and LC × LC, so they can attack their own data sets at the pixel-level.

2 SIGNAL PREPROCESSING IN 2D CHROMATOGRAPHY WITH PIXEL-BASED DATA

Pixel-based analysis aims to discover the valuable chemical information in complex 2D chromatographic data with high sensitivity for real chemical signals and minimal interference from typical instrumental variations that manifest in the data. To achieve this objective, the chromatographic data must be carefully manipulated to reduce the impact of common variations, such as baseline drift and noise. These steps are performed prior to the analysis of the data, and are referred to as “preprocessing.” Preprocessing of the raw chromatographic data reduces the impact of immaterial variations in the data with the goal of improving accuracy and precision of qualitative and quantitative analyses of the chemically relevant variations that are of interest to the analyst. Preprocessing is thus critically important, and is a step that, if performed properly, allows for maximal discovery of useful information while avoiding wasted experimental, instrumental, and analytical efforts. The most common preprocessing steps which we consider here are: baseline correction, noise reduction, normalization, and retention time alignment, which are presented here in the order by which they are performed.

2.1 Baseline Correction

The first preprocessing step in most chromatographic analysis work flows is baseline correction. Baseline drift is the low-frequency signal variation that occurs in the baseline due to column stationary phase bleed, background ionization, and low-frequency variations in the detector and/or instrument-controlled parameters (such as temperature or flow). Baseline rise is the steady increase in baseline observed in temperature-programmed or gradient elution separations. Baseline correction procedures aim to mitigate the effects of baseline drift and rise, improving the detection and quantification of analyte peaks that may otherwise be impacted by baseline interference. Baseline correction methods must be carefully implemented in order to preserve the relevant chemical variations, while avoiding overfitting the chromatogram or mistaking broadened features of coeluting compounds for baseline signals. Following baseline correction, the baseline noise should be centered on zero for the length of the chromatographic separation, and in the case of multichannel detectors, for all channels, as well.

The simplest baseline correction procedure is to collect a "blank" chromatogram and to subtract the blank chromatogram from the sample chromatogram being analyzed. This method is commonly employed in one-dimensional (1D) chromatography, but is less-commonly applied to 2D chromatography, as the presence of minor run-to-run misalignment may magnify the baseline when subtracting the blank chromatogram. The next simplest baseline correction algorithm simulates a blank chromatogram by polynomial least squares fitting and subtracts it from the sample chromatogram. Baseline fitting and subtraction may also be performed on manually selected chromatographic subregions [33] or individual peaks [34–36]. Penalized least squares regression or robust orthogonal regression may also be used to model and subtract baselines [37]. Delta2D image processing software uses a powerful "rolling ball" algorithm to perform baseline subtraction for 2D chromatograms [38]. Rolling minimum methods take advantage of the dead time regions on the second separation dimension in comprehensive 2D chromatography for rapid baseline correction.

2.2 Noise Reduction

Baseline correction procedures are designed to correct low-frequency noise and offsets, but not to alter higher frequency variations. Several techniques may be applied to reduce high-frequency variations and improve signal-to-noise (S/N) ratio in chromatograms. The most commonly implemented methods are smoothing and binning. The classic Savitzky–Golay method is a running smoothing function that fits a low-order polynomial to each data pixel and its neighbors, replacing the signal of each data pixel with the value provided by the polynomial fit [39]. Wavelet smoothing methods transform the chromatogram into the frequency domain, remove the high-frequency components that are assumed to be indeterminate noise, and then perform the reverse transform to the time domain yielding the smoothed chromatogram [40]. Similarly, the low-frequency components characteristic of slowly drifting baseline variations may also be removed with the wavelet method, assisting in baseline correction. Data reduction by simple interpolation and averaging, referred to as "binning," improves S/N and additionally reduces computational load when processing large volumes of 2D separations data [41]. COMForTS, a Fourier transformation method for comprehensive

2D chromatographic separations, improves S/N and additionally reduces the separation time and data density [42].

Noise reduction techniques are critical to successfully addressing a wide variety of analytical challenges. The improvement in S/N, as well as the concurrent data reduction with some methods, improves the ability of the following pixel-based analysis to discover meaningful chemical differences between samples, while reducing false discoveries due to noise. Regardless of the technique used, noise reduction is a delicate operation that requires careful choice of parameters to avoid compromising the real chemical signals that are present in the data. Careless choice of parameters for noise removal algorithms or excessive noise removal may artificially broaden peaks, reduce resolution, introduce artifacts, and/or complicate deconvolution.

2.3 Normalization

Chromatograms are often normalized to correct between-sample variations that are unavoidably introduced during sample collection, preparation, and injection. The most commonly applied normalization technique is the internal standard method [43–46]. However, it may be difficult to find an inert standard that is completely resolved from all native components for truly unknown samples. As an alternative, isotopically labeled internal standards may be of use, taking advantage of the isotope ratio method [47] if the standard chosen happens to be natively present in the sample. However, the use of isotopic standards is limited to separations coupled to mass spectrometry. When it is not practical to use the internal standard method, many users apply the sum-normalization method, which is briefly mentioned in chromatography textbooks, sometimes in terms of area-normalization [48,49]. The sum-normalization method uses the total sum of all baseline corrected signals in a chromatogram as the signal to be normalized. The ratio of the total signal for each sample to that of the mean of all samples is sometimes used as the normalization factor to subsequently adjust all chromatographic signals in the samples. The user must assume the samples in the data set are sufficiently similar, such that equal volumes of the samples should have sufficiently equal total signals at the detector. While this assumption is rarely strictly true, it is often assumed the principal source of variation is due to injection volume. Other normalization methods involve mathematically forcing the mean signal of each chromatogram to equal 1 [50] or mathematically forcing the maximum peak signal volume to equal 1 [51].

Normalization techniques should be carefully chosen to achieve normalization of the desired source of variation. If, for example, a user desires to correct for the efficiency of an extraction, and also the variation in injection volume, a normalization method must be chosen for each. An option may be to add internal standard(s) prior to extraction, as well as a different internal standard immediately prior injection. The normalization would thus adjust samples based first on injection and then on extraction. In some cases, the additional sample handling required for internal standard normalization may introduce larger variation than that which is being normalized. Analysts should observe typical variation in their samples and proceed accordingly.

Normalization is applied to samples (every data pixel in a chromatogram is normalized using the same normalization factor), while scaling is applied to variables (each data pixel in a chromatogram is scaled by a unique factor to meet some criteria). Usually, the purpose of scaling is to reduce the influence of large signals in comparative analyses. Common scaling methods are autoscaling (where the mean of each pixel across all samples is forced to be zero and each pixel is forced to have unit standard deviation across all samples) [52], or dividing each pixel by the mean value among samples [53], or log transformations, or applying power transformations.

2.4 Retention Time Alignment

Alignment is an important preprocessing step because even minor pressure, flow and temperature fluctuations cause retention time variations that may obscure chemical information, resulting in poor performance of most chemometric methods. Alignment algorithms are designed to shift the raw signal along the time axes of the 2D separations so the peak position of a given analyte matches from one sample run to the next. The accuracy of the 2D peak signal volumes should also be preserved during alignment. Alignment algorithms can vary from very simple approaches such as application of a simple scalar shift, to a targeted peak list approach, to more sophisticated approaches involving locally and globally optimized alignment algorithms. While all of these alignment algorithm approaches have merit, for the chemometric analysis of pixel-level data, the most commonly applied preprocessing approaches implemented are the latter. The locally and globally optimized alignment algorithms can handle severe and dynamic shifting in comprehensive 2D chromatography applications, because they are robust, powerful, and essentially automated in every regard after the initial parameter selections have been made by the analyst [24,50,51,54–59].

Alignment of a "sample" chromatogram to a "target" or "reference" chromatogram at the pixel-level requires the assumption that the signal due to analyte peaks in those chromatograms are truly matched. For data sets in which samples are relatively similar, this assumption is generally realized. However, for data sets in which samples vary greatly, it is more challenging for alignment to succeed. This challenge is more confidently addressed for comprehensive 2D chromatography instruments that implement multichannel spectroscopic detection, such as GC × GC-TOFMS and LC × LC-DAD, since the alignment algorithms are designed to utilize the spectral information to correctly match the signal due to analyte peaks across samples. It should also be noted that the use of local alignment algorithms can provide successful alignment when multichannel spectral information is not available.

3 PIXEL-LEVEL METHODS FOR DECONVOLUTION

3.1 Aims of Deconvolution

Analyte peaks often coelute despite the excellent peak capacity and resolving power provided by comprehensive 2D chromatography. The analyst may need to mathematically resolve analyte peaks that are not chromatographically resolved. Peak resolution algorithms can provide the needed resolution for pixel-level chromatographic data. GRAM, MCR-ALS, and PARAFAC are some of the well-known resolution methods which will be described herein. GRAM and MCR-ALS can decompose 2D data, while PARAFAC and sometimes MCR-ALS are primarily applied to 3D data. For all of these methods, the section of either 2D or 3D data must be of suitably low rank by containing a relatively small number of analyte components. To resolve peaks in an unknown sample, GRAM requires a standard sample and is noniterative, while MCR-ALS and PARAFAC are iterative and do not require a standard sample.

These pixel-level resolution methods rely heavily on linear algebra because chromatographic data, under ideal conditions, can be thought of as bilinear or trilinear matrices and arrays for 2D and 3D chromatography, respectively. This means that the pixel-level chromatographic data are ideally composed of unique, consistent, concentration-dependent signals from each independent and unique analyte source present, and those signals are additive, even across samples. Chromatographic instruments are designed to produce such results, though sources of uncontrollable and/or unavoidable variation do diminish the bilinearity and trilinearity of analyte signals in 2D and 3D chromatograms, e.g., various forms of retention time misalignment. Peak resolution algorithms do incorporate some methods of dealing with these sources of variation with varying degrees of success. In this text, the linear algebra notation is as follows: scalars are lower-case italics, vectors are bold lower-case text, 2D matrices are bold capitals, and 3D arrays are underlined bold capitals.

3.2 Generalized Rank Annihilation Method

GRAM can be used to mathematically resolve signals of bilinear and independent underlying factors (or components) such as overlapping analyte peaks in sections of 2D pixel-level chromatograms provided by comprehensive 2D separations [1], such as GC × GC, LC × LC, LC × CE, and so on [4,60–62]. To be bilinear, data must be composed of unique, consistent, and concentration-dependent signals from each independent and unique source present, and those signals must be additive. GRAM decomposition of such bilinear data provides the pure peak profiles as well as the concentration ratios of compounds in the sample relative to the standard. Several GRAM algorithms have been published [4,60,61]. Here, we describe the GRAM algorithm that is based on the standard eigenvalue method [61].

Consider a section of bilinear GC × GC-FID data of a standard sample represented by matrix **N** in which the column dimension represents the ${}^{2}t_{\mathrm{R}}$ dimension and the row dimension represents the ${}^{1}t_{\mathrm{R}}$ dimension. The standard might contain K components that can be expressed according to Equation (1).

$$\mathbf{N} = \sum_{k=1}^{K} \left(\mathbf{x}_k c_{k,k} \mathbf{y}_k^{\mathrm{T}}\right) = \mathbf{X}\mathbf{C}_{\mathrm{N}}\mathbf{Y}^{\mathrm{T}} \tag{1}$$

In Equation (1), **X** and **Y** are matrices whose columns $\mathbf{x}_k$ and $\mathbf{y}_k$ are vectors representing the pure component signals in the ${}^{2}t_{\mathrm{R}}$ dimension and ${}^{1}t_{\mathrm{R}}$ dimension of **N**. $\mathbf{C}_{\mathrm{N}}$ is a diagonal matrix in which the elements $c_{k,k}$ represent the concentration of the kth component in the standard sample. The superscript T denotes a transpose.

Now consider extracting the same GC × GC-FID data section from an unknown sample chromatogram, represented by matrix **M**, again in which the column dimension represents ${}^{2}t_{\mathrm{R}}$ and the row dimension represents ${}^{1}t_{\mathrm{R}}$. **M** can be expressed according to Equation (2).

$$\mathbf{M} = \mathbf{X}\mathbf{C}_{\mathrm{M}}\mathbf{Y}^{\mathrm{T}} \tag{2}$$

In Equation (2), **X**, **Y**, and superscript T have the same meaning as in Equation (1) and $\mathbf{C}_{\mathrm{M}}$ is a diagonal matrix in which the elements $c_{k,k}$ represent the concentration of the kth component in the unknown sample. Thus, given **N**, **M**, $\mathbf{C}_{\mathrm{N}}$, and K, it is possible to calculate **X**, **Y**, and $\mathbf{C}_{\mathrm{M}}$.

GRAM calculates the unique values for **X**, **Y**, and $\mathbf{C}_{\mathrm{M}}$ relative to $\mathbf{C}_{\mathrm{N}}$, given the following requirements. The unknown sample must contain relative analyte component concentrations that differ from the standard. The overlapped peaks must have some resolution on both ${}^{1}t_{\mathrm{R}}$ and ${}^{2}t_{\mathrm{R}}$. The analyte retention times and peak shapes must be equal between the unknown sample and the standard (i.e., the data must be bilinear). Figure 1 shows a standard (**N**) and sample (**M**) chromatographic data section that meets these requirements for GRAM. If retention time shifting is present, then it is necessary to align the data prior to application of GRAM [63,64].

Furthermore, the unknown sample can contain interference peaks, which are absent in the standard and vice versa. The key is that the user augments **M** and **N** into a single matrix so that all potentially present components can be modeled, even if some are absent in either **M** or **N**. For this step, GRAM requires that **M** is augmented row-wise with **N** to form a 2D bilinear augmented matrix, [**M**,**N**], where each row continues to represent a ${}^{2}t_{\mathrm{R}}$ pixel and each column continues to represent a ${}^{1}t_{\mathrm{R}}$ pixel. Next, **M** is also augmented column-wise with **N** to form another augmented matrix [**M**;**N**]. (The square brackets with comma or semi-colon between matrices denote row-wise and column-wise matrix augmentation, respectively). The GRAM algorithm then proceeds to perform singular value decomposition (SVD) on the augmented matrices, yielding Equations (3) and (4).

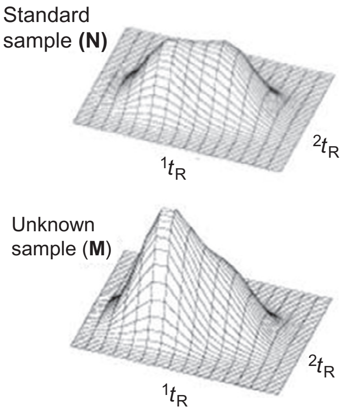

FIGURE 1 Standard (**N**) and sample (**M**) chromatographic data sections that meet the requirements for GRAM to mathematically resolve the two overlapping analyte peaks.

$$[\mathbf{M}, \mathbf{N}] = \mathbf{U}_{\mathrm{r}}\boldsymbol{\Theta}_{\mathrm{r}}\mathbf{V}_{\mathrm{r}}^{\mathrm{T}} \tag{3}$$

$$[\mathbf{M}; \mathbf{N}] = \mathbf{U}_{\mathrm{c}}\boldsymbol{\Theta}_{\mathrm{c}}\mathbf{V}_{\mathrm{c}}^{\mathrm{T}} \tag{4}$$

In Equations (3) and (4), subscript r denotes the row-wise augmented matrix, subscript c denotes the column-wise augmented matrix, the decomposed components from SVD of the augmented matrices are $\boldsymbol{\Theta}$ (a diagonal matrix of singular values) and **U** and **V** (which are orthogonal matrices of eigenvectors). The estimated pseudorank of the augmented matrices should ideally equal the expected number of analyte plus interferent components present in the subregion (K) as indicated by estimating the number of significant factors in $\boldsymbol{\Theta}_{\mathrm{r}}$ and $\boldsymbol{\Theta}_{\mathrm{c}}$. **M** and **N** are transformed to square matrices $\mathbf{M}_{\mathbf{U}_{\mathrm{r}}\mathbf{V}_{\mathrm{c}}}$ and $\mathbf{N}_{\mathbf{U}_{\mathrm{r}}\mathbf{V}_{\mathrm{c}}}$ while preserving the pseudorank of **M** and **N**. Next, the generalized eigenvalue problem is solved for the square matrices using the Moler and Stewart algorithm [65] as in Equations (5)–(7), wherein $\boldsymbol{\lambda}$ is a matrix of eigenvalues [61]. Please see the cited references for more details about the linear algebra described herein [61,65].

$$\mathbf{M}_{\mathbf{U}_{\mathrm{r}}\mathbf{V}_{\mathrm{c}}} = \mathbf{U}_{\mathrm{r}}^{\mathrm{T}}\mathbf{M}\mathbf{V}_{\mathrm{c}} = \left(\mathbf{U}_{\mathrm{r}}^{\mathrm{T}}\mathbf{X}\right)\mathbf{C}_{\mathrm{M}}\left(\mathbf{V}_{\mathrm{c}}^{\mathrm{T}}\mathbf{Y}\right)^{\mathrm{T}} \tag{5}$$

$$\mathbf{N}_{\mathbf{U}_{\mathrm{r}}\mathbf{V}_{\mathrm{c}}} = \mathbf{U}_{\mathrm{r}}^{\mathrm{T}}\mathbf{N}\mathbf{V}_{\mathrm{c}} = \left(\mathbf{U}_{\mathrm{r}}^{\mathrm{T}}\mathbf{X}\right)\mathbf{C}_{\mathrm{N}}\left(\mathbf{V}_{\mathrm{c}}^{\mathrm{T}}\mathbf{Y}\right)^{\mathrm{T}} \tag{6}$$

$$\mathbf{M}_{\mathbf{U}_{\mathrm{r}}\mathbf{V}_{\mathrm{c}}}\mathbf{V}_{\mathrm{r}} = \mathbf{N}_{\mathbf{U}_{\mathrm{r}}\mathbf{V}_{\mathrm{c}}}\mathbf{V}_{\mathrm{r}}\boldsymbol{\lambda} \tag{7}$$

The GRAM algorithm next sorts the matrices **V** and $\boldsymbol{\lambda}$ according to magnitude and applies similarity transforms ($\mathbf{T}^{\dagger}$ and $\mathbf{T}^{\ddagger}$) to **V** and $\boldsymbol{\lambda}$ to ensure only noncomplex solutions are produced as in Equations (8)–(11) wherein superscript -1 denotes the inverse matrix [61].

$$\mathbf{V}^{\dagger} = \mathbf{V}\left(\mathbf{T}^{\dagger}\right)^{-1} \tag{8}$$

$$\boldsymbol{\lambda}^{\dagger} = \mathbf{T}^{\dagger}\boldsymbol{\lambda}\left(\mathbf{T}^{\dagger}\right)^{-1} \tag{9}$$

$$\mathbf{V}^{\ddagger} = \mathbf{V}^{\dagger}\left(\mathbf{T}^{\ddagger}\right)^{-1} \tag{10}$$

$$\boldsymbol{\lambda}^{\ddagger} = \mathbf{T}^{\ddagger}\boldsymbol{\lambda}^{\dagger}\left(\mathbf{T}^{\ddagger}\right)^{-1} \tag{11}$$

The pure component profiles in the chromatographic dimensions (**X** and **Y**) are then calculated according to Equations (12) and (13) wherein the superscript + denotes the pseudoinverse.

$$\mathbf{X} = \mathbf{U}_{\mathrm{r}}\mathbf{N}_{\mathbf{U}_{\mathrm{r}}\mathbf{V}_{\mathrm{c}}}\mathbf{V}^{\ddagger} \tag{12}$$

$$\mathbf{Y} = \mathbf{V}_{\mathrm{c}}\left(\left(\mathbf{V}^{\ddagger}\right)^{+}\right)^{\mathrm{T}} \tag{13}$$

The pure analyte peak profiles provided by GRAM resolution of the overlapping peaks from Figure 1 are shown in Figure 2 using $K = 2$.

Furthermore, quantitative information can be acquired from the ratio of concentrations of the pure components (**R**) in **M** and **N** when calculated according to Equations (14)–(16).

$$\mathbf{C}_{\mathrm{M}} = \left(\mathbf{U}_{\mathrm{r}}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{M}_{\mathbf{U}_{\mathrm{r}}\mathbf{V}_{\mathrm{c}}}\left(\left(\mathbf{V}_{\mathrm{c}}^{\mathrm{T}}\mathbf{Y}\right)^{\mathrm{T}}\right)^{-1} \tag{14}$$

$$\mathbf{C}_{\mathrm{N}} = \left(\mathbf{U}_{\mathrm{r}}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{N}_{\mathbf{U}_{\mathrm{r}}\mathbf{V}_{\mathrm{c}}}\left(\left(\mathbf{V}_{\mathrm{c}}^{\mathrm{T}}\mathbf{Y}\right)^{\mathrm{T}}\right)^{-1} \tag{15}$$

$$\mathbf{R} = \mathbf{C}_{\mathrm{M}}\mathbf{C}_{\mathrm{N}}^{-1} \tag{16}$$

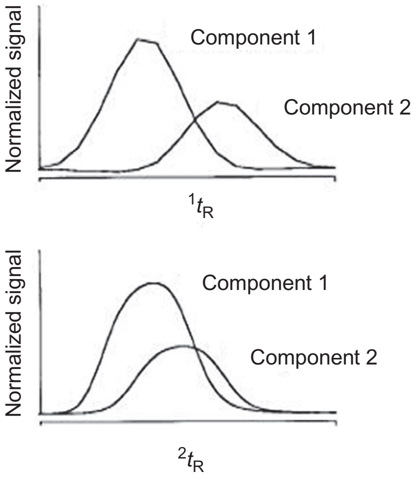

FIGURE 2 Pure peak profiles provided by the GRAM resolution of the overlapping peaks in Figure 1.

Other GRAM algorithms have been developed and implemented [4,60,62]. For example, another algorithm [4] uses an addition matrix ($\mathbf{M}+\mathbf{N}$) instead of an augmented matrix to model all of the components common to both the standard and the unknown sample; this addition ensures common components will be modeled even if the standard has more components than the sample and vice versa. In this particular algorithm, the addition matrix is submitted to SVD decomposition yielding the diagonal matrix $\boldsymbol{\Theta}$ and the orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ as in Equation (17).

$$(\mathbf{M}+\mathbf{N})=\mathbf{U}\boldsymbol{\Theta}\mathbf{V}^{\mathrm{T}} \tag{17}$$

Next, the SVD decomposition components are truncated according to K (the expected number of components or estimated rank of the addition matrix). Then the eigenvalue problem is solved for the truncated matrices (which are "overlined" as in Equation 18) producing a matrix of eigenvectors ($\mathbf{T}$) and a diagonal matrix of eigenvalues ($\boldsymbol{\Pi}$).

$$\left(\bar{\boldsymbol{\Theta}}^{-1}\bar{\mathbf{U}}^{\mathrm{T}}\mathbf{N}\bar{\mathbf{V}}\right)\mathbf{T}=\mathbf{T}\boldsymbol{\Pi} \tag{18}$$

The matrices $\mathbf{T}$ and $\boldsymbol{\Pi}$ are then submitted to a similarity transform to eliminate any complex solutions that arise [4,62]. The pure component profiles ($\mathbf{X}$ and $\mathbf{Y}$) that are common to both $\mathbf{M}$ and $\mathbf{N}$ are determined using the eigenvectors and decomposed components from SVD following Equations (19) and (20).

$$\mathbf{X}=\bar{\mathbf{U}}\bar{\boldsymbol{\Theta}}\mathbf{T} \tag{19}$$

$$\mathbf{Y}=\bar{\mathbf{V}}\left(\mathbf{T}^{-1}\right)^{\mathrm{T}} \tag{20}$$

Finally, the relative concentration information for the unknown sample ($\mathbf{C}_{\mathrm{M}}$) is determined from $\boldsymbol{\Pi}$ and $\mathbf{C}_{\mathrm{N}}$ as in Equation (21).

$$\mathrm{diagonal}(\boldsymbol{\Pi})=\frac{\mathbf{C}_{\mathrm{N}}}{\mathbf{C}_{\mathrm{M}}+\mathbf{C}_{\mathrm{N}}} \tag{21}$$

Ultimately, no matter which algorithm is applied, GRAM is designed to provide mathematically resolved chromatographic profiles of components that are present in both the standard and the unknown samples. Given the proper conditions, GRAM algorithms are designed to output results that are unambiguous and physically meaningful.

3.3 Multivariate Curve Resolution-Alternating Least Squares

MCR-ALS is used to mathematically resolve signals of bilinear and independent underlying factors (components) such as overlapping peaks in sections of 2D pixel-level chromatograms. It can also be applied to unfolded pixel-level LC × LC-DAD or GC × GC-TOFMS chromatograms wherein ${}^{1}t_{\mathrm{R}}$ and ${}^{2}t_{\mathrm{R}}$ are properly augmented in the same dimension or multiple sample chromatograms have been properly augmented to yield a bilinear 2D data matrix [66–70].

Consider an $(A \times B \times J)$ GC × GC-TOFMS data section containing A pixels in the ${}^2t_R$ dimension, B pixels in the ${}^1t_R$ dimension, and J pixels in the mass spectral dimension. If significant retention time shifting occurs from run-to-run along the second GC dimension within a single sample run on the instrument, then this 3D data array may not be sufficiently trilinear [61]. Similarly, for LC × LC-DAD, under some circumstances the data may not be sufficiently trilinear, for example, when the data are collected over a relatively long period of time [70]. However, Tauler, Rutan, and others have shown that the data can be readily analyzed if the data array is properly unfolded so that the ${}^1t_R$ and ${}^2t_R$ indices are augmented, producing a bilinear 2D $(A{*}B \times J)$ matrix [66–70]. For the remaining discussion of the matrix dimensions and MCR-ALS, $A{*}B$ will be replaced with I, the number of pixels in the unfolded dimension. MCR-ALS can then be readily applied to resolve peaks, and the resulting pure component profiles can be refolded into the original 3D format. For example, Figure 3 shows a section of an unfolded GC × GC-TOFMS chromatogram of a blend of biodiesel and conventional diesel wherein the mass spectral dimension (m/z 76–250) is perpendicular to the plane of the page. At least two components overlap in this region.

Figure 4 illustrates the MCR-ALS decomposition of the bilinear components in the unfolded GC × GC-TOFMS chromatographic data

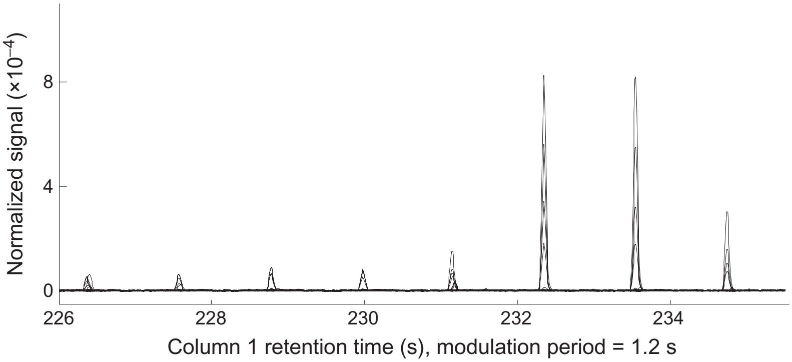

FIGURE 3 Section of an unfolded GC × GC-TOFMS chromatogram of a blend of biodiesel and conventional diesel shown as overlaid chromatograms for ions of m/z 76–250 with modulation period $(P_M) = 1.2$ s. The mass spectral dimension is perpendicular to the plane of the page.

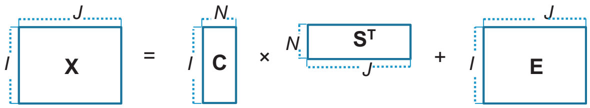

FIGURE 4 An $(A \times B \times \mathrm{J})$ section of a GC × GC-TOFMS chromatogram was unfolded into a matrix (**X**) of dimensions $(I \times J)$, wherein $A{*}B = I$, and decomposed by MCR into its N bilinear factors (pure components described by their N pure chromatographic profiles in **C** and N pure spectra in **S**) plus residual signals (**E**).

section. To be bilinear, data must be composed of unique, consistent, and concentration-dependent signals from each independent and unique source present, and those signals must be additive.

Equation (22) mathematically describes MCR-ALS applied to an unfolded GC × GC-TOFMS chromatographic data section of interest:

$$\mathbf{X} = \sum_{n=1}^{N} \mathbf{c}_n \mathbf{s}_n^{\mathrm{T}} + \mathbf{E} \tag{22}$$

In Equation (22), **X** represents the given unfolded ($I \times J$) chromatographic data section, which is equal to the sum (for $n = 1$ to N factors) of the product of the nth column ($\mathbf{c}_n$) of the ($I \times N$) loadings matrix **C** and the nth row ($\mathbf{s}_n$) of the ($N \times J$) transposed loadings matrix **S**, plus the ($I \times J$) error matrix **E**. Thus, each of the N factors present is expressed by its pure mass spectrum ($\mathbf{s}_n$) multiplied by its resolved chromatographic profile ($\mathbf{c}_n$). Ideally, each of the N columns and rows in **C** and **S**, respectively, fully describe a single analyte signal source. **E** contains all signals not accounted for in the N factors and can be used as a metric for the fit between the model and the original data, **X**.

Given initial estimates of the pure concentration profiles and pure spectra of components in a matrix such as **X**, MCR-ALS implements an alternating least-squares (ALS) method to iteratively optimize the estimates of those pure analyte peak profiles and pure spectra while using appropriate constraints to achieve convergence. The alternating least-squares initialization values for **C** and **S** can be user-provided pure spectra or pure concentration profiles if the analysis is targeted and the information is available. However, for nontarget analysis, initialization values can be provided by evolving factor analysis, orthogonal projection approach, or simple-to-use self-modelling analysis [66]. These pure variable selection methods strive to detect the most dissimilar rows or columns in a matrix and use those as initial estimates of **C** and **S** for MCR-ALS. MCR-ALS goes on to calculate **C** and **S** while obeying appropriate constraints, which may include unimodality in the unfolded ${}^{2}t_{\mathrm{R}}$ dimension and nonnegativity in both dimensions. The unimodality constraint is appropriate for chromatographic data which nominally produces Gaussian-like peaks, but if the user expects to observe a factor for noise or background contributions, then the unimodality constraint may not be appropriate. An MCR-ALS algorithm may be capable of accepting local rank/selectivity constraints wherein the user identifies regions in **X** where components are known to be present or absent [66]. The convergence criterion can be a preset number of iterations or a threshold value defining the difference in fit improvement between consecutive iterations. Metrics for fit between the model and original data may be the relative magnitude of **E**, the variance explained, or the lack-of-fit (LOF). LOF is expressed in Equation (23), where x_{ij} is an element of the data matrix **X** and e_{ij} is the corresponding element from matrix **E**.

$$\%\text{LOF} = 100\% \times \sqrt{\left(\sum_{i,j} e_{ij}^2\right) \Big/ \left(\sum_{i,j} x_{ij}^2\right)} \tag{23}$$

Ambiguous predictions may pose a challenge for MCR-ALS if and when various combinations of **C** and **S** reproduce **X** with equally good fit metrics. An example of intensity ambiguity is when $\mathbf{c}_n$ and $\mathbf{s}_n$ have signal intensity shifted from one dimension to the other such that **X** and **E** in Equation (1) are equal to **X** and **E** in Equation (24):

$$\mathbf{X} = \sum_{n=1}^{N} (\mathbf{c}_n k_i) \left(\mathbf{s}_n^{\mathrm{T}} \frac{1}{k_i}\right) + \mathbf{E} \tag{24}$$

In Equation (24), k_i is a scalar that simultaneously upscales one dimension while downscaling the other. Thus, Equations (22) and (24) result in chromatographic and spectral profiles that have the *same shape* but differ in scale, so this is known as intensity ambiguity. To treat intensity ambiguity, the chromatographic dimension is often normalized so that the relative signal intensity information is stored in the spectral dimension. Another source of ambiguity in MCR-ALS can predict sets of pure chromatographic and spectral profiles that yield equally good fit metrics, yet have *different shapes*. This is known as rotational ambiguity. Any transform matrix (**T**) that can be multiplied by **C** while its inverse ($\mathbf{T}^{-1}$) is multiplied by $\mathbf{S}^{\mathbf{T}}$ will cause this type of ambiguity in MCR-ALS results. Constraints are the main method of suppressing rotational ambiguity. In addition, some degree of chemical selectivity in both separation dimensions is required for all factors present in the 2D data section to provide diversity that reduces the number of possible k_i values and/or **T** matrices that adequately model **X** while fulfilling constraints. Quantitative information for an analyte can be obtained by summing all the elements of the outer product of the *n*th loadings vectors $\mathbf{c}_n$ and $\mathbf{s}_n$ to yield a scalar pure peak volume that should be proportional to relative concentration. When this quantification procedure is applied to a matrix composed of augmented sample chromatograms, the result yields concentration ratios between the common analyte components in the different samples.

The number of factors (value of *N*) is critical to the accuracy of the resolved profiles in **C** and **S**. If *N* is too large (over fitting), the signal from a single source is described by two or more factors in the loadings $\mathbf{c}_n$ and $\mathbf{s}_n$. If *N* is too small (under fitting), the signal from a single source might be absent in the loadings or it might be inappropriately combined with the loadings of another source. SVD may be used to estimate *N*. It is common practice to model a variety of *N* values and then the user selects the MCR-ALS model that captures the highest degree of variance and predicts pure profiles with reasonable shapes. At this point, it may be necessary to note that MCR-ALS models are not nested. This means the *N*-resolved components in a chromatographic data section are represented by *N* unique factors, which are all different than the factors produced by an MCR-ALS model for the same data section but built using $N-x$ factors, where x is any positive integer from 1 to $N-1$.

An example of implementing MCR-ALS to decompose the GC × GC-TOFMS data from Figure 3 is provided in Figure 5. MCR-ALS yielded three pure chromatographic profiles using $N=3$, nonnegativity constraints in both dimensions, and 99.02% variance captured. The pure mass spectra (inset) reveal background contributions in component 3 were filtered out simultaneously as diesel components 1 and 2 were resolved.

3.4 Parallel Factor Analysis

PARAFAC [71–73] can be used to mathematically resolve signals of independent underlying factors (components) for third-order data, such as overlapping analyte peaks, background contributions, and/or noise in sections of 3D pixel-level GC × GC-TOFMS chromatograms where the first dimension is ^1t_R, the second dimension is ^2t_R, and the third dimension is the mass spectral dimension [74]. Figure 6 (top) is an example of a section of GC × GC-TOFMS data containing three diesel components where the mass spectral dimension has been temporarily summed for viewing purposes to produce a total ion current (TIC) chromatogram. Also shown (at the bottom of Figure 6) is the top-down view of the same section which shows the diesel components are unresolved, to some extent.

The PARAFAC decomposition of the trilinear components in a GC × GC-TOFMS chromatographic section is illustrated in Figure 7. To be trilinear, data must be composed of unique, consistent, concentration-dependent signals from each independent and unique analyte source present, and those signals must be additive.

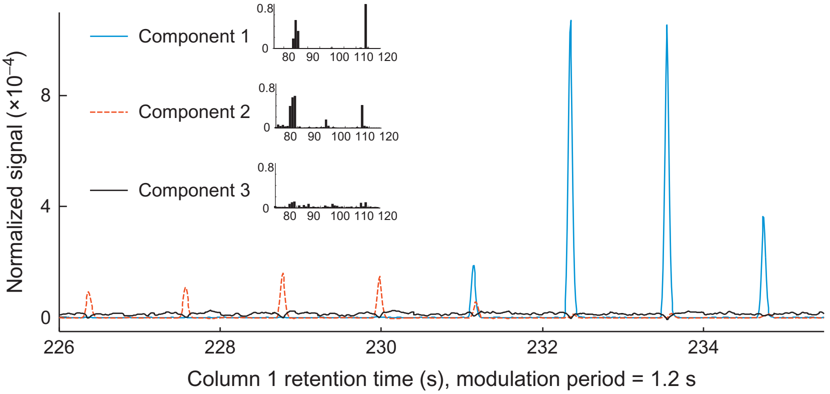

FIGURE 5 Pure chromatographic profiles and pure mass spectra (inset) provided by MCR-ALS decomposition of three components in the GC × GC-TOFMS chromatographic data section wherein $P_M=1.2$ s.

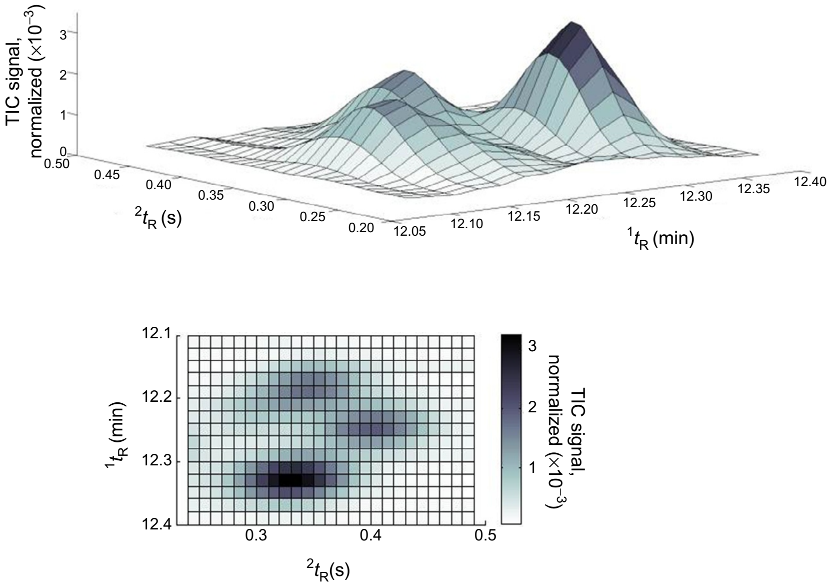

FIGURE 6 Section of a GC × GC-TOFMS chromatogram of diesel: (top) The mass spectral dimension of the section is temporarily summed and shows unresolved peaks are present. (bottom) Top-down view of the same TIC section.

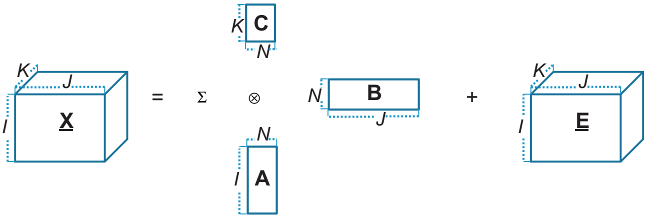

FIGURE 7 PARAFAC decomposition of an $(I \times J \times K)$ section of a GC × GC-TOFMS chromatogram ($\underline{\mathbf{X}}$) into its N trilinear factors (pure components described by N pure chromatographic profiles in **A** and **B**, as well as N pure spectra in **C**) plus residual signals ($\underline{\mathbf{E}}$).

Equation (25) mathematically describes PARAFAC applied to a GC × GC-TOFMS data section:

$$\underline{\mathbf{X}} = \sum_{n=1}^{N} \mathbf{a}_n \otimes \mathbf{b}_n \otimes \mathbf{c}_n + \underline{\mathbf{E}} \tag{25}$$

In Equation (25), $\underline{\mathbf{X}}$ represents the given $(I \times J \times K)$ chromatographic data section of interest. According to Equation (25), $\underline{\mathbf{X}}$ is equal to the sum of the outer product of each of the nth columns ($\mathbf{a}_n$, $\mathbf{b}_n$, and $\mathbf{c}_n$) of the respective $(I \times N)$, $(J \times N)$, and $(K \times N)$ loadings matrices **A**, **B**, and **C**, for $n = 1$ to N, where N is the number of factors, plus the $(I \times J \times K)$ error array $\underline{\mathbf{E}}$. Thus, $\mathbf{a}_n$, $\mathbf{b}_n$, and $\mathbf{c}_n$ essentially contain the resolved ${}^2t_\mathrm{R}$ profile, the resolved ${}^1t_\mathrm{R}$ profile, and the resolved mass spectral profile of each of the N factors present in the chromatographic subregion. Ideally, each of the N columns in $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ describe a single signal source (e.g., factors such as analytes, noise, background contribution, interferents, etc). $\underline{\mathbf{E}}$ contains all signals not accounted for in the N factors and it is a metric for the fit between the model and the original chromatographic data section.

PARAFAC can determine appropriate values for $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ using the ALS method. Traditionally, ALS iteratively seeks values of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ that minimize $\underline{\mathbf{E}}$ while optionally applying constraints. The ALS algorithm can fail to locate the global minimum (true solution) when chromatographic sections contain a large number of factors, or if there is insufficient selectivity in every dimension, or if there are sufficiently large deviations from trilinearity. In such cases, repeatedly building PARAFAC models of a single given chromatographic data section may produce a variety of solutions. This instability can often be improved by applying various constraints such as unimodality in the ${}^1t_\mathrm{R}$ and ${}^2t_\mathrm{R}$ dimensions and nonnegativity in all three dimensions. The ALS initialization values for $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ can be provided by trilinear decomposition (TLD) based on the generalized eigenvalue decomposition problem, which does not require any initial values but also does not allow constraints to be applied. Another way to initialize ALS may be to use random values and start from several different starting points of random values until the same solution is repeatedly reached. Other possible initialization values may be random orthogonalized values, or singular values. The stopping condition for ALS can be minimizing $\underline{\mathbf{E}}$, meeting a certain threshold value for sum of squares of the residuals ($\underline{\mathbf{E}}$), reaching a threshold *relative* change in $\underline{\mathbf{E}}$ between two iterations, allowing a maximum elapsed time for iterations, executing a maximum number of iterations, or meeting some other goodness-of-fit metric or convergence criterion.

The number of factors (value of N) is critical to the accuracy of the resolved peak profiles in $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. If N is too big (over fitting), the signal from a single source is described by two or more factors in the loadings $\mathbf{a}_n$, $\mathbf{b}_n$, and $\mathbf{c}_n$. If N is too small (under fitting), the signal from a single source might be absent in the loadings or it might be inappropriately combined with the loadings of another source. Some of the following methods can be used to determine the appropriate number of factors in a parsimonious manner. Some users seek the minimal value of N that still provides a PARAFAC model with a relatively high core consistency value and explains a high percentage of the variance in $\underline{\mathbf{X}}$. Another method of determining N is to examine the "noisiness" in $\underline{\mathbf{E}}$ and choose the lowest N factor model, which appears to contain a high degree of randomness and where systematic variation is not visible. Another method is known as split-half analysis wherein the appropriate number of factors is revealed by that which produces identical PARAFAC models built for independent subsamples of the given data, though this is more amenable to data wherein one of the dimensions is the sample dimension, which is not the case with the GC × GC-TOFMS chromatogram considered herein. Another method of determining the appropriate value of N for the PARAFAC model of a GC × GC-TOFMS data section is detection of "splitting" in the mass spectral loadings as N is incrementally increased. Splitting is detected in an N-factor PARAFAC model when two or more decomposed mass spectra both match a common reference spectrum above a certain match value threshold because signal from a single source was spread across two or more factors. Thus, a PARAFAC model provided by $N-1$ or fewer factors is considered appropriate. At this point, it may be necessary to note that PARAFAC models are not nested. Also of note, PARAFAC can provide a unique solution for the final values of **A**, **B**, and **C** because there is no rotational problem when the appropriate number of factors is modeled and the data are sufficiently trilinear [73]. However, the final values of **A**, **B**, and **C** may have signal intensity shifted from one dimension to another when comparing two nominally equal PARAFAC models. To treat this ambiguity, both chromatographic dimensions are often normalized so that the signal intensity information is stored in the mass spectral dimension. Quantitative information for an analyte can be obtained by summing all the elements of the outer product of the nth loadings vectors $\mathbf{a}_n$, $\mathbf{b}_n$, and $\mathbf{c}_n$ to yield a scalar pure peak volume that is proportional to relative concentration.

Figure 8A shows the pure component profiles in the chromatographic dimensions provided by PARAFAC decomposition of the chromatographic data section shown in Figure 6. In this case, the PARAFAC software that was used was from the PLS Toolbox for MATLAB (Eigenvector Research, Inc) with ALS initiated by TLD using nonnegativity constraints in all three dimensions and three factors based on core consistency and prior knowledge of the sample). Figure 8B shows the pure component mass spectra that were also provided by PARAFAC for component 1 (top), component 2 (middle), and component 3 (bottom). According to a mass spectral library search, these components are alkylated benzenes. Figure 8C shows the product of the ${}^{1}t_{\mathrm{R}}$ and ${}^{2}t_{\mathrm{R}}$ pure component chromatographic profiles multiplied by (scaled by) the sum of the full pure component mass spectrum for component 1 (top), component 2 (middle), and component 3 (bottom). Manual comparison of Figure 8C to Figure 6 reveals these mathematically resolved pure component models provided by PARAFAC can indeed be superimposed or added to closely fit the original chromatographic data section of interest where the three peaks overlapped, to some extent.

Herein, PARAFAC was described for a GC × GC-TOFMS application, but PARAFAC is also suitable for LC × LC or any comprehensive 2D separation coupled with a selective multivariate detector [8–10]. PARAFAC decomposition requires the data to have sufficient selectivity in each of the three dimensions. For example, if two overlapping components have identical spectra,

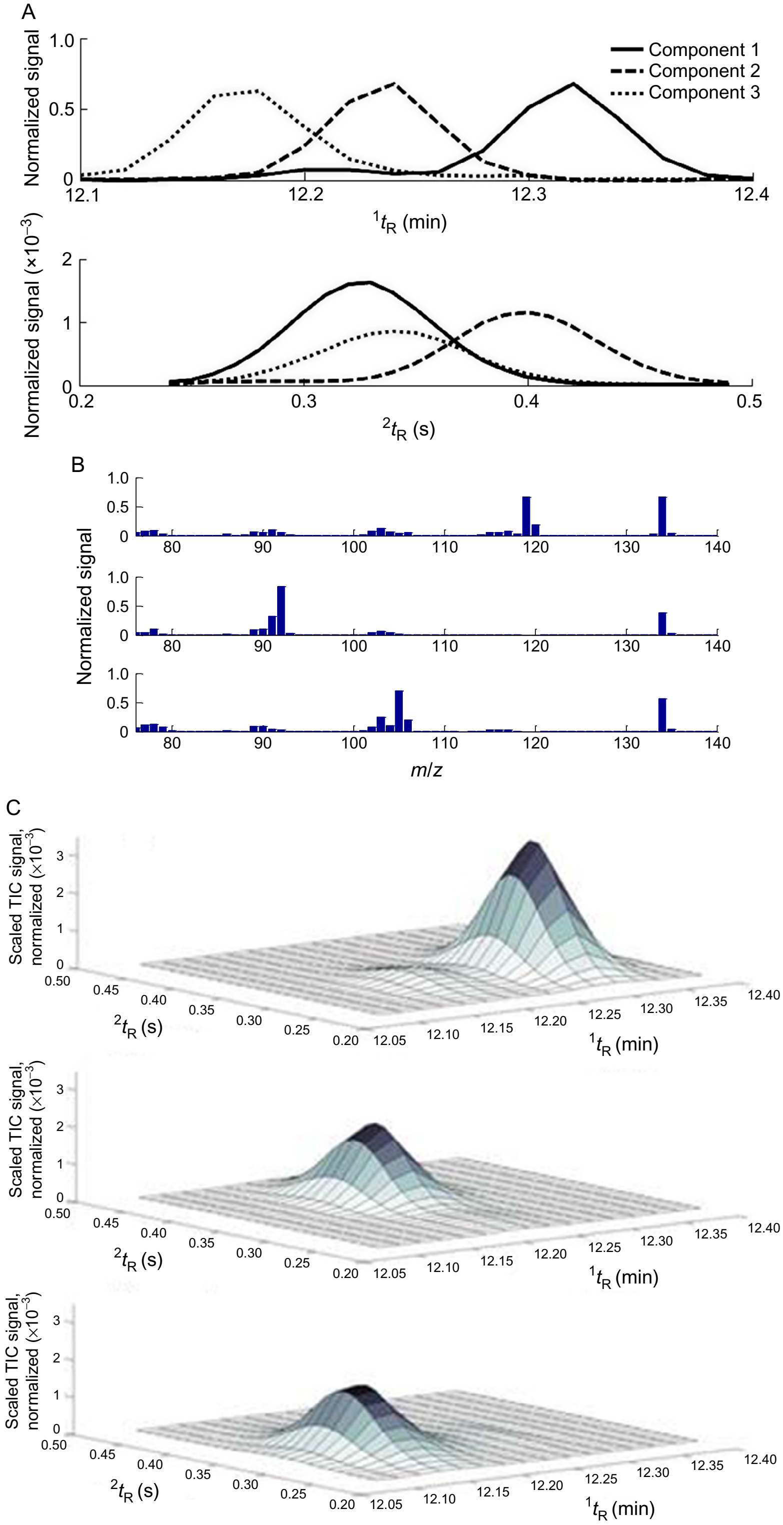

FIGURE 8 See legend on next page.

then PARAFAC will not be able to resolve them. Prudent selection of complementary columns and detector parameters for GC × GC and LC × LC separations usually help provide sufficient selectivity in all three dimensions. PARAFAC decomposition also requires the data to be additive, which is often a valid assumption for stable sample concentrations that are within the linear range of the detectors on GC × GC and LC × LC instruments. PARAFAC decomposition also requires the data to be sufficiently trilinear. Experience has shown that GC × GC-TOFMS data do not need to be completely trilinear to achieve suitable quantitative accuracy, but sufficient trilinearity should be confirmed prior to application of PARAFAC [36,53,74,75]. Indeed, GC × GC and LC × LC instruments may be prone to retention time shifting in the second separation dimension within a single sample injection onto the instrument, so the rank may vary in the ${}^{2}t_{R}$ dimension, but not in the spectral dimension, causing deviations from trilinearity. If shifting is severe, pixel-level retention time alignment may be a necessary preprocessing step for PARAFAC. In general, while achieving trilinear data in LC × LC separations is now readily achieved for short-term studies [8–10], for long-term studies the trilinearity may be compromised so MCR-ALS is preferred for LC × LC-DAD [70]. For GC × GC-TOFMS, if within-run retention time shifting in the second GC dimension is severe, implementation of MCR-ALS for such cases should be explored, as discussed earlier in this chapter [69].

Creative and extremely useful variations and automations of PARAFAC have been described by experts. For example, software for automated nontargeted PARAFAC decomposition of *entire* GC × GC-TOFMS chromatograms was developed. The software automatically determines the optimal *N*-factor PARAFAC model for each component in a chromatographic subregion and then it continues to comprehensively analyze the entire chromatogram [75]. In another advancement, PARAFAC2 was shown to provide accurate solutions for GC × GC-TOFMS data exhibiting deviations from trilinearity caused by within-run retention time shifting in ${}^{2}t_{R}$ [76,77]. PARAFAC2 relaxes the strict trilinearity requirement by allowing the estimated loading matrix for the ${}^{2}t_{R}$ dimension to vary for each factor in the ${}^{1}t_{R}$ loadings matrix. As is the case for PARAFAC, the PARAFAC2 method provides a unique solution that cannot be rotated without a loss of fit and the solution provides pure peak profiles and pure mass spectra when the proper number of factors is used. Another advancement of note was PARAFAC applied to quantify analytes in comprehensive three-dimensional gas chromatography (GC^{3}) chromatograms [78].

FIGURE 8 (A) Pure component profiles in the ${}^{1}t_{R}$ dimension (top) and the ${}^{2}t_{R}$ dimension (bottom) as provided by PARAFAC decomposition of the chromatographic subregion shown in Figure 1. (B) Pure component mass spectra provided by PARAFAC for component 1 (top), component 2 (middle), and component 3 (bottom). (C) Product of the ${}^{1}t_{R}$ and ${}^{2}t_{R}$ pure component profiles multiplied by (scaled by) the sum of the full pure component mass spectrum for component 1 (top), component 2 (middle), and component 3 (bottom).

4 FINGERPRINTING AND PATTERN RECOGNITION

4.1 Aim of Fingerprinting and Pattern Recognition

Chromatograms are composed of hundreds, if not thousands, of variables (retention time units) for which signals proportional to chemical concentration are recorded. Consider comparing many chromatograms of various samples of interest. Some of the chromatographic variables will reveal important chemical similarities (correlations) and important chemical differences (variations) among the samples. Manually examining chromatograms to discover interesting patterns is subjective, and large volumes of data may make manual observations unfeasible. Instead, an unsupervised pattern recognition algorithm can automatically explore a given data set and objectively reveal interesting patterns, variations, correlations, covariations, similarities, and differences among complex samples. Additionally, supervised pattern recognition algorithms use given quantitative and/or qualitative information to model a given data set and meet various goals such as discovering statistically significant similarities and differences among complex samples, or building a multivariate calibration model to predict a quantitative or qualitative property of an unknown sample. A major purpose of unsupervised and supervised pattern recognition algorithms is to reduce the data set dimensionality while preserving the original information, which allows more efficient manipulation of the data and perhaps a better understanding of underlying phenomena related to the samples.

For pattern recognition algorithms to be successful, sources of correlation and variation in the data must be true chemical correlations, manifested as relevant variations in the detector response. But chemical correlations and variations can be obscured by other sources of variation such as retention time shifting, uncontrollable instrumental fluctuations, injection volume discrepancies, and scaling variations among series of measurements with wide ranging magnitudes. Preprocessing techniques such as baseline correction, normalization, alignment or peak matching, mean centering, and scaling are often applied to data prior to pattern recognition analysis to diminish variations in the data that may obscure true chemical variations among samples.

Theoretically, if peak-level data are comprehensive, accurate, and precise, and if pixel-level data are bilinear and of suitable dimensions, then pattern recognition algorithms should return results that are equally accurate for both pixel-level and peak-level data. However, these requirements are often challenging to achieve for these two data types. The following sections explore the implementation of pattern recognition studies with pixel-level 2D separations data.

4.2 Unsupervised Pattern Recognition—PCA

PCA is used to find linear combinations of factors (latent variables (LVs) or principal components, PCs) in a bilinear 2D data matrix that model the reproducible variations and correlations (e.g., covariations, patterns) in that data matrix more succinctly. To be bilinear, the 2D data matrix must be composed of unique, consistent, concentration-dependent signals from each independent and unique source present, and those signals must be additive. The data matrix can be composed of many kinds of data including pixel-level data or peak-level data. PCA is certainly applicable to any kind of chromatography, but this text will use examples from pixel-level GC × GC-TOFMS data to explain PCA.

The first step in preparing pixel-level GC × GC-TOFMS chromatograms for PCA is to properly unfold and augment chromatograms from various samples of interest into a 2D matrix such that the first dimension is the sample dimension. Assume I samples are included so the first dimension contains I elements. The second dimension of the 2D matrix is the properly unfolded retention time dimension. For pixel-level analysis of GC × GC with a univariate detector (e.g., FID or TIC), the second dimension will contain J elements, where $J = (\text{pixels in } {}^{2}t_R) \times (\text{pixels in } {}^{1}t_R)$. For pixel-level analysis of GC × GC with a multivariate detector, the second dimension will contain $J = (\text{pixels in } {}^{2}t_R) \times (\text{pixels in } {}^{1}t_R) \times (\text{pixels in detector dimension})$ elements, so in this case, J is often extremely large. Without specialized computer equipment, analysts will encounter limits to the maximum size of the 2D matrix that can be successfully submitted to PCA algorithms. Thus, analysts may need to reduce data density prior to PCA using data binning techniques that were discussed earlier in this chapter. Assume the properly unfolded and bilinear chromatograms from many samples are augmented into matrix $\mathbf{X}$ of size $(I \times J)$ where the rows represent I samples and the columns represent J variables.

The goal of PCA is to produce a decomposition model of $\mathbf{X}$ which can be written in matrix notation as $\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E}$, wherein $\mathbf{T}$ is an $(I \times F)$ scores matrix, $\mathbf{P}^{\mathrm{T}}$ is a transposed $(F \times J)$ loadings matrix, $\mathbf{E}$ is an $(I \times J)$ residuals matrix, and F is the number of PCs necessary to model 100% of the variation in $\mathbf{X}$. The PCA decomposition model of $\mathbf{X}$ is illustrated in Figure 9. The scores contain information about how the samples (rows in $\mathbf{X}$) relate to each other. The loadings contain information about how the variables (columns in $\mathbf{X}$) relate to each other. Conceptually, if two samples have perfectly matching chromatograms, they should ideally have equal scores.

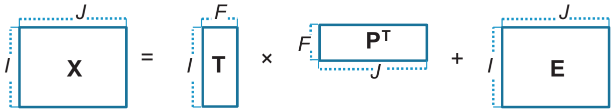

FIGURE 9 Chromatograms composed of J pixels from comprehensive 2D chromatography of I samples were properly unfolded and augmented into a data matrix ($\mathbf{X}$) of dimensions $(I \times J)$. PCA decomposes $\mathbf{X}$ into an $(I \times F)$ scores matrix ($\mathbf{T}$), a transposed $(F \times J)$ loadings matrix ($\mathbf{P}^{\mathrm{T}}$), and an $(I \times J)$ residuals matrix ($\mathbf{E}$). F is less than or equal to the minimum of I and J and F is the number of PCs necessary to model 100% of the variation in $\mathbf{X}$.

The PCA model of **X** can also be written as the sum of the F outer products of a score vector ($\mathbf{t}_f$) and its corresponding loadings vector ($\mathbf{p}_f$) plus residuals **E** as in Equation (26) [12,14,79,80].

$$\mathbf{X} = \sum_{f=1}^{F} \mathbf{t}_f \mathbf{p}_f^{\mathrm{T}} + \mathbf{E} \tag{26}$$

In Equation (26), F is less than or equal to the minimum of I and J, and **p** and **t** ideally preserve the important variations in **X**. To preserve and summarize the important variations in **X** while finding **t** and **p**, the PCA algorithm converts this model into a standard eigenvalue decomposition problem as in Equation (27).

$$\mathrm{cov}(\mathbf{X})\mathbf{p}_f = \boldsymbol{\lambda}_f \mathbf{p}_f \quad \text{wherein } \mathrm{cov}(\mathbf{X}) = \frac{(\mathbf{X}^{\mathrm{T}}\mathbf{X})}{(I-1)} \text{ for } I > J$$

or

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}\mathbf{X}^{\mathrm{T}}}{(I-1)} \text{ for } I \le J \tag{27}$$

In Equation (27), cov(**X**) is defined as the covariance matrix of **X** when the columns of **X** have been mean centered. Mean centering is a preprocessing technique wherein the original mean value of each column of **X** is subtracted from each variable in the column to force the average of each column to be zero. If the columns of **X** have been autoscaled, then Equation (27) substitutes the correlation matrix of **X** (corr(**X**)) for cov(**X**). Autoscaling is a preprocessing technique wherein the columns of **X** are mean centered and then the variance in each column is normalized to unity by dividing each column by its original standard deviation. PCA loadings are a function of both signal variation and signal magnitude so autoscaling helps reduce the effect of signal magnitude on the loadings.

Equation (27) has a standard linear algebra solution that uses SVD to decompose cov(**X**) into the product of three terms: an orthogonal matrix **U**, a diagonal singular values matrix **S**, and a transposed orthogonal matrix **V**. The fth squared singular value (s_f) is the same as the fth eigenvalue ($\boldsymbol{\lambda}_f$) in Equation (27) and so the unit normalized $\mathbf{p}_f$ (standardized eigenvectors) can be determined. The $\mathbf{p}_f$ loadings vectors are orthonormal to each other $\left(\mathbf{p}_i^{\mathrm{T}}\mathbf{p}_{\mathrm{j}} = 0 \text{ for } i \neq j,\ \ \mathbf{p}_i^{\mathrm{T}}\mathbf{p}_j = 1 \text{ for } i = j\right)$. The $\boldsymbol{\lambda}_f$ reveal the percent variation captured in each factor. Finally, each score vector $\mathbf{t}_f$ is supposed to be the linear combination of the original **X** variables defined by $\mathbf{p}_f$, so using the $\mathbf{X}\mathbf{p}_f = \mathbf{t}_f$ projection reveals the $\mathbf{t}_f$ scores vectors which are orthogonal to each other ($\mathbf{t}_i{}^{\mathrm{T}}\mathbf{t}_j = 0$ for $i \neq j$). Ambiguity can be a problem for PCA because any scalar value can simultaneously upscale the **t** or **p** vector while downscaling the other vector and still solve Equations (26) and (27). Thus **p** is normalized to unit length in Equation (27) to treat this scaling ambiguity.

The $\mathbf{t}_f$, $\mathbf{p}_f$ pairs are sorted in descending order of variation based on $\boldsymbol{\lambda}_f$ values. The linear combinations of the pairs ($\mathbf{t}_f\,\mathbf{p}_f^{\mathrm{T}}$) are known as PCs (specifically PC 1 through PC F for $f=1$ to F). PC 1 most closely models $\mathbf{X}$ in Equation (26) with residuals as minimal as possible, considering PC 1 is a linear factor trying to capture maximum variation in $\mathbf{X}$. Each successive PC models diminishing amounts of variation in $\mathbf{X}$. The PCs are nested, meaning a model retaining $F+1$ PCs has PC 1 through PC F which are identical to a model retaining F PCs, for any integer value of F. Therefore, the user can truncate $\mathbf{T}$ and $\mathbf{P}^{\mathrm{T}}$ to F columns and rows, respectively, to keep only the initial PCs that capture sufficient variation to adequately model $\mathbf{X}$. Choosing the number of PCs to retain is a subjective matter and there is not a single best method for making this decision. Some of the methods that exist involve inspecting the ordered $\boldsymbol{\lambda}_f$ and retaining only the corresponding PCs which captured a relatively high percentage of variation compared to that of the final PCs which are assumed to represent noise.

The analyst can plot and inspect the model residuals to glean information about variables that were not modeled. This may reveal weaknesses or strengths in assuming the data set is suitable for PCA modeling. The analyst can examine their data set using various metrics related to PCA modeling, particularly outlier detection metrics. Hotelling's T^2 measures the Mahalanobis distance of a score from the PCA model origin and accounts for different variation in different PCs. Samples with large Hotelling's T^2 might be outliers, assuming the samples are supposed to represent a multivariate normal distribution. The metric sum square error (Q) measures the squared distance of a score from a plane of scores in a PCA model, and samples with large Q might be considered outliers. When using such metrics the analyst should be conservative and always carefully review all assumptions about normal distributions and variations in the data set when deciding whether the metrics are appropriate for labeling a sample as an outlier.

In addition to the above description of PCA, the following geometric conceptualization of PCA might be helpful. Consider a data set of 10 sample chromatograms, each composed of 20,000 pixels. The first conceptual step in applying PCA to this data set is to imagine each preprocessed and unfolded chromatogram as a 1D data vector plotted in independent variable-space. This means each chromatogram is geometrically represented by a single vector endpoint located in the independent variable space that was defined by 20,000 orthogonal axes ($J=20{,}000$). Then, among those 10 vector endpoints (representing the $I=10$ sample chromatograms), PCA calculates a vector (PC 1) which is fit so that it captures the greatest variation in the locations of those 10 vector endpoints. A second principal component (PC 2) that is orthogonal to PC 1 is calculated and fit such that it captures the next greatest variation in those 10 vector endpoints. This is repeated until up to 10 orthogonal PCs are fit and ordered based on captured percent variation. The 10 PCs will cumulatively have modeled 100% of the variation in the locations of those original 10 vector endpoints. The 10 vector endpoints can now be thought of as being in PC space relative to (at most) 10 PC axes, rather than 10 points relative to the original 20,000 independent variable-axes. Since the PCs are ordered and nested, the analyst can truncate the later PCs and project the 10 vector endpoints onto only the first few primary PCs. The distance in PC space between each vector endpoint and each PC axis is the score for each sample. Samples that are similar to each other will have similar score values because the vector endpoints of similar samples cluster together in PC space. Samples that are different from each other will have different score values because the vector endpoints of different samples may be farther apart in PC space. When the 10 sample scores on PC 1 are plotted, this "scores plot" reveals chemical similarities (or differences) among the 10 samples (similar samples may have similar scores and different samples may have different scores).

As for a conceptual description of the loadings in PCA, imagine the 10 PCs in PC space are 10 axes that are projected onto the original 20,000 independent variable-axes. Each PC axis has a calculable angle between it and each of the original 20,000 independent variable-axes. If an original variable had the largest variations across samples, then it would have a large influence on the calculation of PC 1 and the angle at the intersection of the two axes is small. If an original variable had little variation across samples, then it would have a small influence on PC 1 and the angle at the intersection of the two axes is large. As that angle of intersection decreases from $\Pi/2$ to 0°, the cosine of the angle goes from 0 to 1. Summarily, the loadings values are the cosines of the angles between each PC axis and each original independent variable axis. For mean centered chromatograms, the loadings values range from -1 to 1. Original independent variables (retention times) that have the most influence on the clustering of vector endpoints in PC space will have the greatest absolute loadings. Retention times with the most substantial amounts of signal variation across all the samples will be the most highly loaded variables in the primary PCs. When loadings on PC 1 are plotted as a function of the original variables, the variables with greatest absolute magnitude in this "loadings plot" had the most influence on any clustering observed in the PC 1 scores plot.

Consider an example of PCA applied to a data set of normalized, properly unfolded, and mean centered GC × GC TIC chromatograms of seven samples (two pure conventional diesels and five blends of various biodiesels with various conventional diesels). In this case, the PCA software that was used was from the PLS Toolbox for MATLAB (Eigenvector Research, Inc.). A chromatogram of one of the blends is shown in Figure 10A. PCA is not given this information, but it is useful to know that the peaks eluting before 47 min are primarily conventional diesel components and the peaks eluting after 47 min are primarily biodiesel components, specifically fatty acid methyl esters (FAMEs). Figure 10B shows some results of PCA: the scores on PC 1 captured 60.78% of the variation in the seven chromatograms. It is apparent that samples 1 and 2 have negative scores, samples 3, 4, and 5 have moderately positive scores, and samples 6 and 7 have the highest positive scores in this data set. Conceptually, this means that samples 1 and 2 may be similar to each other, samples 3, 4, and 5 may be similar to each other, and samples 6 and 7 may be similar to each other, if some subjectivity is acceptable in deciding that three classes are present rather than two classes. Figure 10C

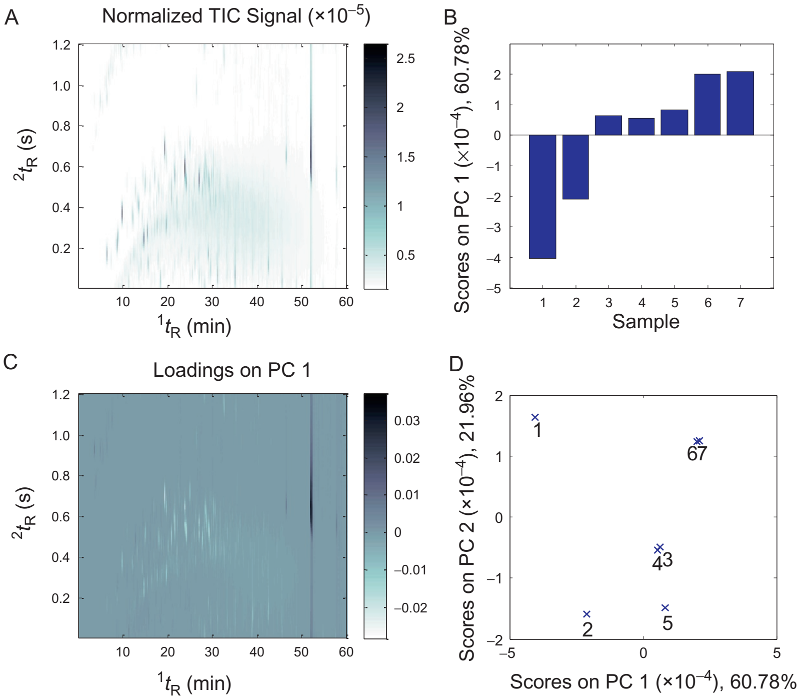

FIGURE 10 GC × GC TIC chromatograms of seven samples* composed of various biodiesels and various conventional diesels were submitted to PCA. (A) The chromatogram of Sample 3. Peaks eluting before 47 min are primarily conventional diesel components. Peaks eluting after 47 min are primarily biodiesel components. (B) Scores on PC 1 which captured 60.78% of the variation. (C) Loadings plot for PC 1. The most positively loaded variables are black, while the most negatively loaded variables are white, and variables with moderate-to-zero absolute loadings values are gray. (D) Scores plot for PC 1 versus PC 2. PC 2 captured 21.96% variation.
*Sample 1 = 100% (v/v) Conventional Diesel K;
*Sample 2 = 100% (v/v) Conventional Diesel L;
*Sample 3 = 12.5% (v/v) Biodiesel P with 87.5% (v/v) Conventional Diesel L;
*Sample 4 = 12.5% (v/v) Biodiesel Q with 87.5% (v/v) Conventional Diesel L;
*Sample 5 = 12.5% (v/v) Biodiesel R with 87.5% (v/v) Conventional Diesel L;
*Sample 6 = 20% (v/v) Biodiesel P with 80% (v/v) Conventional Diesel M;
*Sample 7 = 20% (v/v) Biodiesel Q with 80% (v/v) Conventional Diesel N.

shows the loadings plot for PC 1. The most positively loaded chromatographic variables are black and it appears that these are primarily the FAMEs. The most negatively loaded variables are white and it appears that these are primarily the conventional diesel components. The variables with little to zero loadings (both positive and negative) are gray. Combining the basic conclusions from the scores plot with the basic conclusions from the loadings plot could lead an analyst to surmise that PCA reveals samples 1 and 2 have similar conventional diesel components, while samples 3, 4, 5, 6, and 7 have relatively similar FAME components. The analyst might further conclude that PCA reveals samples 3, 4, and 5 have biodiesel components similar to each other, while samples 6 and 7 are in a different category of biodiesel composition. In fact, the samples can be loosely classified into three categories: (category 1) samples 1 and 2 are pure conventional diesels, (category 2) samples 3, 4, and 5, are composed of 12.5% (v/v) biodiesel blended with conventional diesel, and (category 3) samples 6 and 7 are composed of 20% (v/v) biodiesel blended with conventional diesel. Figure 10D shows a plot of scores on PC 1 versus scores on PC 2. PC 2 captured 21.96% of the variation in the seven chromatograms and it is plotted on a slightly different scale than the PC 1 axis which captured the 60.78% variation mentioned earlier. In this scores plot, samples 6 and 7 cluster together, while samples 3 and 4 cluster together; this matches the tentative categorizations gleaned from inspecting the clustering in Figure 10B. However, contrary to what might be expected, samples 1 and 2 do not cluster together and instead sample 2 is closest to sample 5. This is because PC 2 captures variations that are orthogonal to PC 1. PC 1 primarily focused on the presence or absence of biodiesel FAMEs, but the variables highly loaded in PC 2 must be indicative of other variations, perhaps only variations in the conventional diesels; the analyst would have to look at the loadings on PC 2 to find out. Since the sample 1 score on PC 2 is positive while the sample 2 score on PC 2 is negative, an analyst might conclude there are chemical differences between these two samples of pure conventional diesel, so perhaps there really are more than three categories of samples present. Finally, sample compositions are revealed in the caption of Figure 10, and the reader can observe the similarities and differences among the samples. Perhaps, as Figure 10D indicated, there is some similarity between sample 2 and sample 5, after all. The reader can evaluate how well PCA, an unsupervised pattern recognition method, preserved and modeled important variations in the data set.

4.3 Supervised fingerprinting and pattern recognition

4.3.1 PLS Analysis and Multilinear Partial Least Squares for N*-way analysis*

PLS analysis is a multivariate calibration method for modeling a relationship of maximum covariance between given 1D sample data (independent variables) and corresponding quantitative data (dependent variables) [81–83]. Once the relationship between the independent variables and the dependent variables is modeled, the values of the dependent variable for unknown samples can be predicted through regression.

Multilinear partial least squares for N-way analysis (NPLS) generalizes PLS from 1D data to higher order ND data (where N is a small positive integer) [84–86]. NPLS can be applied to data sets of many 2D or 3D chromatograms that have been augmented into a single matrix where one dimension is the sample dimension. The augmented data matrix must be composed of trilinear data. To be trilinear, data must be composed of unique, consistent, concentration-dependent signals and those signals must be independently additive. NPLS is applicable to both pixel-level and peak-level data; theoretically, both levels of data may return equally accurate results as long as both are comprehensive and peaks are properly aligned across samples.

Consider a 3D data array ($\underline{\mathbf{X}}$) composed of augmented samples of pixel-level GC × GC-FID chromatograms (or chromatograms from any comprehensive 2D chromatography with a univariate detector). The dimensions of $\underline{\mathbf{X}}$ are (I = number of samples × J = data points on $^{1}\mathrm{t}_{\mathrm{R}}$ × K = data points on $^{2}\mathrm{t}_{\mathrm{R}}$). Also consider a single column vector ($\mathbf{y}$) of given quantitative information for each sample. NPLS decomposes $\underline{\mathbf{X}}$ into F factors (e.g., LVs or linear combinations of the original variables) defined by F score vectors ($\mathbf{t}$ ($I \times 1$)), F weight vectors in the $^{1}\mathrm{t}_{\mathrm{R}}$ dimension ($^{J}\mathbf{w}$ ($J \times 1$)), and F weight vectors in the $^{2}t_{\mathrm{R}}$ dimension ($^{K}\mathbf{w}$ ($K \times 1$)). These produce a model of $\underline{\mathbf{X}}$ that can be represented by Equation (28) and a model of $\mathbf{y}$ with regression coefficients ($\mathbf{b}$) that can be represented by Equation (29):

$$x_{ijk} = t_i {}^{J}w_j {}^{K}w_k \tag{28}$$

$$y_i = t_i b_i \tag{29}$$

NPLS finds $\mathbf{t}_i$, $^{J}\mathbf{w}$, and $^{K}\mathbf{w}$ that fit the model in Equation (28) while maximizing the covariance of $\underline{\mathbf{X}}$ with $\mathbf{y}$.

The following steps describe an algorithm to build an NPLS model of $\underline{\mathbf{X}}$ [84].

Step 1. Mean center the given $\underline{\mathbf{X}}$ and $\mathbf{y}$ and save the mean data so it can be later used during model regression and prediction. Do the same for scaling the data, if scaling is appropriate [87]. Initially, use $f = 1$ to calculate the first LV.

Step 2. Calculate the product of the cube $\underline{\mathbf{X}}$ and the vector $\mathbf{y}$ to yield the matrix $\mathbf{Z}$ ($J \times K$).

$$z_{jk} = \sum_{i=1}^{I} x_{ijk} y_i$$

Step 3. Calculate the weight vectors $^{J}\mathbf{w}$ ($J \times 1$) and $^{K}\mathbf{w}$ ($K \times 1$), each normalized to unit length, which yield a solution to the model in Equation (28) by using the first set of normalized vectors from SVD of $\mathbf{Z}$. Some texts interchange the terms “weights” and “loadings.”

$$\max_{{}^J\mathbf{w}{}^K\mathbf{w}}\left[\left({}^J\mathbf{w}\right)^{\mathrm{T}}\mathbf{Z}{}^K\mathbf{w}\right] \approx \left({}^J\mathbf{w}{}^K\mathbf{w}\right) = \mathrm{SVD}(\mathbf{Z})$$

Step 4. Calculate the $\mathbf{t}$ that has maximal covariance with $\mathbf{y}$. When the $\mathbf{w}$ vectors are known and of length one, the solution to the problem of finding $\mathbf{t}$ is found by projection of $\underline{\mathbf{X}}$ onto the $\mathbf{w}$ vectors. This can be described in summation notation and combined with the prior expression for z_{jk} to yield [84]:

$$\max_{{}^J\mathbf{w}{}^K\mathbf{w}}\left[\sum_{i=1}^{I} t_i y_i | t_i = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}{}^J w_j{}^K w_k\right] = \max_{{}^J\mathbf{w}{}^K\mathbf{w}}\left[\sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}{}^J w_j{}^K w_k\right]$$

An alternative format for this calculation of each element of $\mathbf{t}$ (t_i) under the constraint of maximizing covariance is [88,89]:

$$t_i = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}{}^J w_j{}^K w_k$$

(As an aside, at this point the fth LV is definable by $\mathbf{t}$, ${}^J\mathbf{w}$, and ${}^K\mathbf{w}$, and $\hat{x}_{ijk} = \mathbf{t}_i{}^J\mathbf{w}_j\left({}^K\mathbf{w}_k\right)^{\mathrm{T}}$ is the prediction of x_{ijk} that represents the best least squares model of $\underline{\mathbf{X}}$. An alternative format of expressing the model of $\underline{\mathbf{X}}$ is $\hat{x}_{ijk} = \sum_{f=1}^{F} t_i = {}^J w_j{}^K w_k$).

Step 5. Use $\mathbf{t}$ to calculate regression coefficients: $\mathbf{b} = (\mathbf{t}^{\mathrm{T}}\mathbf{t})^{-1}\mathbf{t}^{\mathrm{T}}\mathbf{y}$

(Note that at this point the prediction of $\mathbf{y}$ is $\hat{\mathbf{y}} = \mathbf{tb}$.)

Step 6. Calculate the residuals for the chromatograms: $\mathbf{X}_{i,\mathrm{res}} = \mathbf{X}_i – t_i\,{}^J\mathbf{w}\,({}^K\mathbf{w})^{\mathrm{T}}$.

Step 7. Calculate the residuals for the dependent variable values: $\mathbf{y}_{\mathrm{res}} = \mathbf{y} – \mathbf{tb}$.

Step 8. Iterate to the next factor ($f = f + 1$) and return to Step 2 with the following deviations: (1) use the residuals from Steps 6 and 7 to replace $\mathbf{X}_i$ and $\mathbf{y}$; (2) during Steps 5 and 7 replace $\mathbf{t}$ with the ($I \times F$) scores matrix ($\mathbf{T}$) whose columns are the augmented $\mathbf{t}$ vectors for $f = 1$ to f [80,84]. Continue reiterating Steps 2–8 until convergence when a proper prediction of $\mathbf{y}$ is achieved.

To predict the dependent variable for unknown samples of interest using the NPLS model, use the saved mean data from Step 1 to center the unknown sample chromatogram. Do the same to scale the data, if appropriate. Then regress the preprocessed unknown sample chromatogram ($\mathbf{X}_{\mathrm{unk}}$) onto the model ($\mathbf{y} = \mathbf{bX}$) and predict the value of the dependent variable for the unknown sample of interest.

Generally, the model is built using an appropriate training set and the optimal number of LV (F) may be chosen by a cross-validation procedure that allows the user to appropriately minimize the root mean square error of cross-validation of predicted y_i values to avoid over fitting or under fitting the model. Then the model is evaluated using an independent test set.

The NPLS algorithm described here is limited to only one dependent variable (single column vector $\mathbf{y}$), but another version of this algorithm exists for calculating the NPLS model using several dependent variables for each sample (when **Y** is a matrix) [84]. The algorithm described here is also limited to only a univariate detector, but for comprehensive 2D chromatography with a multivariate detector the ND data array where N equals any small positive integer can be modeled by NPLS using more detailed versions of the algorithm [84,86]. If it is appropriate, 2D chromatograms could theoretically be unfolded into 1D and analyzed by PLS, but expect the results to differ because PLS calculates an additional set of weight vectors to maintain orthogonal scores while NPLS does *not* provide orthogonal scores and weights vectors [89].

Conventional figures-of-merit that are commonly reported with univariate calibration methods need multivariate propagated error techniques to be completely defined for multivariate calibration methods. In fact, instead of quantifying a single limit of detection (LOD) for a model, it is more appropriate to report an interval of LODs for PLS calibration models [90]. The standard error of prediction metric for multiway PLS regression is also available [91].

Examples of NPLS applied to comprehensive 2D chromatography include modeling and predicting the aromatic and naphthene content in naphtha samples and modeling and predicting the percent composition of blends of biodiesel and conventional diesel [92,93].

4.3.2 Supervised Pattern Recognition—Feature Selection

Supervised variable selection techniques are used to discover statistically significant chemical differences or similarities among known classes of complex samples. Since class membership must be known, experimental design is particularly important. Furthermore, eliminating uninformative variables and reducing voluminous data down to selective variables can improve predictive models. Two common techniques are uninformative variable elimination (UVE) [94] and Fisher ratios (F-ratio) feature selection [11,31,32] methods.

UVE is designed to inspect a given data set and eliminate the independent variables that contain no more information than that of random variables [94]. The original UVE algorithm is designed to analyze the regression coefficients ($\mathbf{b}$ $(1 \times K)$) of the PLS multivariate calibration model for a data set of independent variables ($\mathbf{X}$ $(J \times K)$) related to dependent variables ($\mathbf{y}$ $(J \times 1)$) with residuals ($\mathbf{e}$). The PLS model is represented by $\mathbf{y} = \mathbf{b} \times \mathbf{X} + \mathbf{e}$. UVE looks at the "reliability" ($\mathbf{c}$ $(1 \times K)$) of the regression coefficients for mean centered data, where each element of **c** is defined as each element of **b** divided by the standard deviation of **b** coefficients obtained by leave-one-out modeling. Independent variables that have the smallest **c** values are considered uninformative and are eliminated from the data set. The UVE algorithm defines the threshold for **c** as the value of **c** computed for random variables that are temporarily added to the data set. Variations of the algorithm exist to make it robust against outliers, to reduce influence of the maximum random **c** value on the final threshold value, and to calculate **c** using the absolute value of **b** coefficients from autoscaled data.

F-ratio feature selection methods are designed to inspect a data set and select variables that contain statistically significant differences among given sample classes [11,31,32,95]. The *F*-ratio of each independent variable across a given variety of sample classes is defined as the class-to-class variation of that independent variable divided by the sum of its within-class variations, which may be defined as follows when applied to chromatography [95].

Class-to-class variation $= \sigma_{\mathrm{cl}}^2 = \sum \frac{(\overline{x_i} - \bar{x})^2 n_i}{(k-1)}$ (where n_i is the number of chromatograms in the ith class, $\bar{x}_i$ is the mean of the ith class, $\bar{x}$ is the overall mean, and k is the number of classes).

Within-class variation $= \sigma_{\mathrm{err}}^2 = \frac{\sum\left(\sum\left(\overline{x_{ij}} - \bar{x}\right)^2\right) - \left(\sum(\overline{x_i} - \bar{x})^2 n_i\right)}{(N-k)}$ (where x_{ij} is the ith measurement of the jth class, and N is the total number of chromatograms).

Finally, the *F*-ratio $= \frac{\sigma_{\mathrm{cl}}^2}{\sigma_{\mathrm{err}}^2}$.

Independent variables with large *F*-ratios are likely to be features of interest that differentiate sample classes, whereas independent variables with small *F*-ratios are likely to be representative of noise or features that do not differentiate sample classes. A recently introduced *F*-ratio algorithm called the tile-based Fisher-ratio software [31,32] is designed to comprehensively analyze 4D data sets of pixel-level GC × GC-TOFMS chromatograms to discover chemical features that differentiate given sample classes. The tile-based *F*-ratio software uses a novel indexing scheme that benefits from the advantages offered by pixel-level data analysis (namely, comprehensive nontargeted information is quickly gleaned) as well as the advantages offered by peak-level data analysis (namely, the algorithm is robust against retention time variations). The independent variables are sorted in descending order of *F*-ratio value which corresponds with the analyst's level of interest in each independent variable. *F*-ratio analysis is a robust, comprehensive, nontargeted analysis method that is useful for a variety of studies, particularly forensics experiments in which cause and effect chemical signatures can be cleverly evaluated by the experimental design. It is also useful for metabolomics experiments wherein the analyst often seeks to discover interesting patterns in the chemical fingerprints that characterize cellular processes.

REFERENCES

[1] Giddings JC. Two-dimensional separations: concept and promise. Anal Chem 1984;56:1258–70. http://dx.doi.org/10.1021/ac00276a003.

[2] Liu Z, Phillips JB. Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface. J Chromatogr Sci 1991;29:227–31. http://dx.doi.org/10.1093/chromsci/29.6.227.

[3] Bushey MM, Jorgenson JW. Automated instrumentation for comprehensive two-dimensional high-performance liquid chromatography of proteins. Anal Chem 1990;62:161–7. http://dx.doi.org/10.1021/ac00201a015.

[4] Bruckner CA, Prazen BJ, Synovec RE. Comprehensive two-dimensional high-speed gas chromatography with chemometric analysis. Anal Chem 1998;70:2796–804. http://dx.doi.org/10.1021/ac980164m.

[5] Harynuk J, Marriott PJ. Fast GC×GC with short primary columns. Anal Chem 2006;78:2028–34. http://dx.doi.org/10.1021/ac0519413.

[6] Seeley JV, Micyus NJ, Bandurski SV, Seeley SK, McCurry JD. Microfluidic deans switch for comprehensive two-dimensional gas chromatography. Anal Chem 2007;79:1840–7. http://dx.doi.org/10.1021/ac061881g.

[7] Siegler WC, Fitz BD, Hoggard JC, Synovec RE. Experimental study of the quantitative precision for valve-based comprehensive two-dimensional gas chromatography. Anal Chem 2011;83:5190–6. http://dx.doi.org/10.1021/ac200302b.

[8] Porter SEG, Stoll DR, Rutan SC, Carr PW, Cohen JD. Analysis of four-way two-dimensional liquid chromatography-diode array data: application to metabolomics. Anal Chem 2006;78:5559–69. http://dx.doi.org/10.1021/ac0606195.

[9] Stoll DR, Cohen JD, Carr PW. Fast, comprehensive online two-dimensional high performance liquid chromatography through the use of high temperature ultra-fast gradient elution reversed-phase liquid chromatography. J Chromatogr A 2006;1122:123–37. http://dx.doi.org/10.1016/j.chroma.2006.04.058.

[10] Stoll DR, Li X, Wang X, Carr PW, Porter SEG, Rutan SC. Fast, comprehensive two-dimensional liquid chromatography. J Chromatogr A 2007;1168:3–43. http://dx.doi.org/10.1016/j.chroma.2007.08.054.

[11] Pierce KM, Hoggard JC, Hope JL, Rainey PM, Hoofnagle AN, Jack RM, et al. Fisher ratio method applied to third-order separation data to identify significant chemical components of metabolite extracts. Anal Chem 2006;78:5068–75. http://dx.doi.org/10.1021/ac0602625.

[12] Brereton RG. Chemometrics: data analysis for the laboratory and chemical plant. New York: Wiley; 2003.

[13] Massart DL. Chemometrics: a textbook. New York: Elsevier Sciences Ltd.; 1988

[14] Beebe KR, Pell RJ, Seasholtz MB. Chemometrics: a practical guide. New York: Wiley-Interscience; 1998.

[15] Amigo JM, Skov T, Bro R. ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics. Chem Rev 2010;110:4582–605. http://dx.doi.org/10.1021/cr900394n.

[16] Cortes HJ, Winniford B, Luong J, Pursch M. Comprehensive two dimensional gas chromatography review. J Sep Sci 2009;32:883–904. http://dx.doi.org/10.1002/jssc.200800654.

[17] Mondello L, Tranchida PQ, Dugo P, Dugo G. Comprehensive two-dimensional gas chromatography-mass spectrometry: a review. Mass Spectrom Rev 2008;27:101–24. http://dx.doi.org/10.1002/mas.20158.

[18] François I, Sandra K, Sandra P. Comprehensive liquid chromatography: fundamental aspects and practical considerations—a review. Anal Chim Acta 2009;641:14–31. http://dx.doi.org/10.1016/j.aca.2009.03.041.

[19] Reichenbach SE, Tian X, Tao Q, Stoll DR, Carr PW. Comprehensive feature analysis for sample classification with comprehensive two-dimensional LC. J Sep Sci 2010;33:1365–74. http://dx.doi.org/10.1002/jssc.200900859.

[20] Pierce KM, Kehimkar B, Marney LC, Hoggard JC, Synovec RE. Review of chemometric analysis techniques for comprehensive two dimensional separations data. J Chromatogr A 2012;1255:3–11. http://dx.doi.org/10.1016/j.chroma.2012.05.050.

[21] Cook DW, Rutan SC. Chemometrics for the analysis of chromatographic data in metabolomics investigations. J Chemom 2014;28:681–7. http://dx.doi.org/10.1002/cem.2624.

[22] Ruckebusch C, Blanchet L. Multivariate curve resolution: a review of advanced and tailored applications and challenges. Anal Chim Acta 2013;765:28–36. http://dx.doi.org/10.1016/j.aca.2012.12.028.

[23] Furbo S, Hansen AB, Skov T, Christensen JH. Pixel-based analysis of comprehensive two-dimensional gas chromatograms (color plots) of petroleum: a tutorial. Anal Chem 2014;86:7160–70. http://dx.doi.org/10.1021/ac403650d.

[24] Gröger T, Schäffer M, Pütz M, Ahrens B, Drew K, Eschner M, 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 2008;1200:8–16. http://dx.doi.org/10.1016/j.chroma.2008.05.028.

[25] Almstetter MF, Appel IJ, Dettmer K, Gruber MA, Oefner PJ. Comparison of two algorithmic data processing strategies for metabolic fingerprinting by comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry. J Chromatogr A 2011;1218:7031–8. http://dx.doi.org/10.1016/j.chroma.2011.08.006.

[26] Welke JE, Manfroi V, Zanus M, Lazzarotto M, Alcaraz Zini C. Differentiation of wines according to grape variety using multivariate analysis of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection data. Food Chem 2013;141:3897–905. http://dx.doi.org/10.1016/j.foodchem.2013.06.100.

[27] Harvey PM, Shellie RA. Data reduction in comprehensive two-dimensional gas chromatography for rapid and repeatable automated data analysis. Anal Chem 2012;84:6501–7. http://dx.doi.org/10.1021/ac300664h.

[28] Reichenbach SE, Tian X, Boateng AA, Mullen CA, Cordero C, Tao Q. Reliable peak selection for multisample analysis with comprehensive two-dimensional chromatography. Anal Chem 2013;85:4974–81. http://dx.doi.org/10.1021/ac303773v.

[29] Reichenbach SE, Tian X, Cordero C, Tao Q. Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography. J Chromatogr A 2012;1226:140–8. http://dx.doi.org/10.1016/j.chroma.2011.07.046.

[30] Brokl M, Bishop L, Wright CG, Liu C, McAdam K, Focant J-F. Multivariate analysis of mainstream tobacco smoke particulate phase by headspace solid-phase micro extraction coupled with comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry. J Chromatogr A 2014;1370:216–29. http://dx.doi.org/10.1016/j.chroma.2014.10.057.

[31] Marney LC, Siegler WC, Parsons BA, Hoggard JC, Wright BW, Synovec RE. Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry data. Talanta 2013;115:887–95. http://dx.doi.org/10.1016/j.talanta.2013.06.038.

[32] Parsons BA, Marney LC, Siegler WC, Hoggard JC, Wright BW, Synovec RE. 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 2015;87:3812–9. http://dx.doi.org/10.1021/ac504472s.

[33] Bailey HP, Rutan SC, Carr PW. Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis. J Chromatogr A 2011;1218:8411–22. http://dx.doi.org/10.1016/j.chroma.2011.09.057.

[34] Mondello L, Herrero M, Kumm T, Dugo P, Cortes H, Dugo G. Quantification in comprehensive two-dimensional liquid chromatography. Anal Chem 2008;80:5418–24. http://dx.doi.org/10.1021/ac800484y.

[35] Hoggard JC, Synovec RE. Parallel factor analysis (PARAFAC) of target analytes in GC × GC−TOFMS data: automated selection of a model with an appropriate number of factors. Anal Chem 2007;79:1611–9. http://dx.doi.org/10.1021/ac061710b.

[36] Tobias HJ, Sacks GL, Zhang Y, Brenna JT. Comprehensive two-dimensional gas chromatography combustion isotope ratio mass spectrometry. Anal Chem 2008;80:8613–21. http://dx.doi.org/10.1021/ac801511d.

[37] Daszykowski M, Wróbel MS, Bierczynska-Krzysik A, Silberring J, Lubec G, Walczak B. Automatic preprocessing of electrophoretic images. Chemom Intell Lab Syst 2009;97:132–40. http://dx.doi.org/10.1016/j.chemolab.2009.03.002.

[38] Schmarr H-G, Bernhardt J. Profiling analysis of volatile compounds from fruits using comprehensive two-dimensional gas chromatography and image processing techniques. J Chromatogr A 2010;1217:565–74. http://dx.doi.org/10.1016/j.chroma.2009.11.063.

[39] Stevenson PG, Mnatsakanyan M, Guiochon G, Shalliker RA. Peak picking and the assessment of separation performance in two-dimensional high performance liquid chromatography. Analyst 2010;135:1541–50. http://dx.doi.org/10.1039/B922759H.

[40] Soggiu A, Marullo O, Roncada P, Capobianco E. Empowering spot detection in 2DE images by wavelet denoising. In Silico Biol 2009;9:125–33. http://dx.doi.org/10.3233/ISB-2009-0393.

[41] Nadeau JS, Wilson RB, Hoggard JC, Wright BW, Synovec RE. Study of the interdependency of the data sampling ratio with retention time alignment and principal component analysis for gas chromatography. J Chromatogr A 2011;1218:9091–101. http://dx.doi.org/10.1016/j.chroma.2011.10.031.

[42] Trudgett MJE, Guiochon G, Shalliker RA. Theoretical description of a new analytical technique: comprehensive online multidimensional fast Fourier transform separations. J Chromatogr A 2011;1218:3545–54. http://dx.doi.org/10.1016/j.chroma.2011.03.061.

[43] van der Klift EJC, Vivó-Truyols G, Claassen FW, van Holthoon FL, van Beek TA. Comprehensive two-dimensional liquid chromatography with ultraviolet, evaporative light scattering and mass spectrometric detection of triacylglycerols in corn oil. J Chromatogr A 2008;1178:43–55. http://dx.doi.org/10.1016/j.chroma.2007.11.039.

[44] Castillo S, Mattila I, Miettinen J, Orešič M, Hyötyläinen T. Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry. Anal Chem 2011;83:3058–67. http://dx.doi.org/10.1021/ac103308x.

[45] Amador-Muñoz O, Marriott PJ. Quantification in comprehensive two-dimensional gas chromatography and a model of quantification based on selected summed modulated peaks. J Chromatogr A 2008;1184:323–40. http://dx.doi.org/10.1016/j.chroma.2007.10.041.

[46] Kallio M, Kivilompolo M, Varjo S, Jussila M, Hyötyläinen T. Data analysis programs for comprehensive two-dimensional chromatography. J Chromatogr A 2009;1216:2923–7. http://dx.doi.org/10.1016/j.chroma.2008.11.037.

[47] Humston EM, Hoggard JC, Synovec RE. Utilizing the third order advantage with isotope dilution mass spectrometry. Anal Chem 2010;82:41–3. http://dx.doi.org/10.1021/ac902184b.

[48] Robards K, Hadad PR, Jackson PE. Principles and practice of modern chromatographic methods. New York: Academic Press; 1994.

[49] Lee ML, Yang FJ, Bartle KD. Open tubular column gas chromatography. New York: John Wiley & Sons; 1984.

[50] Vial J, Pezous B, Thiébaut D, Sassiat P, Teillet B, Cahours X, et al. The discriminant pixel approach: a new tool for the rational interpretation of GC×GC-MS chromatograms. Talanta 2011;83:1295–301. http://dx.doi.org/10.1016/j.talanta.2010.07.059.

[51] Mohler RE, Tu BP, Dombek KM, Hoggard JC, Young ET, Synovec RE. Identification and evaluation of cycling yeast metabolites in two-dimensional comprehensive gas chromatography–time-of-flight-mass spectrometry data. J Chromatogr A 2008;1186:401–11. http://dx.doi.org/10.1016/j.chroma.2007.10.063.

[52] Carvalho PC, Hewel J, Barbosa VC, Yates III JR. Identifying differences in protein expression levels by spectral counting and feature selection. Genet Mol Res 2008;7:342–56. http://dx.doi.org/10.4238/vol7-2gmr426.

[53] Hoggard JC, Wahl JH, Synovec RE, Mong GM, Fraga CG. Impurity profiling of a chemical weapon precursor for possible forensic signatures by comprehensive two-dimensional gas chromatography/mass spectrometry and chemometrics. Anal Chem 2010;82:689–98. http://dx.doi.org/10.1021/ac902247x.

[54] Allen RC, Rutan SC. Investigation of interpolation techniques for the reconstruction of the first dimension of comprehensive two-dimensional liquid chromatography–diode array detector data. Anal Chim Acta 2011;705:253–60. http://dx.doi.org/10.1016/j.aca.2011.06.022.

[55] Johnson KJ, Wright BW, Jarman KH, Synovec RE. High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. J Chromatogr A 2003;996:141–55. http://dx.doi.org/10.1016/S0021-9673(03)00616-2.

[56] Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. J Chemom 2004;18:231–41. http://dx.doi.org/10.1002/cem.859.

[57] Gröger T, Zimmermann R. Application of parallel computing to speed up chemometrics for GC × GC–TOFMS based metabolic fingerprinting. Talanta 2011;83:1289–94. http://dx.doi.org/10.1016/j.talanta.2010.09.015.

[58] Zhang D, Huang X, Regnier FE, Zhang M. Two-dimensional correlation optimized warping algorithm for aligning GC×GC−MS data. Anal Chem 2008;80:2664–71. http://dx.doi.org/10.1021/ac7024317.

[59] Vial J, Noçairi H, Sassiat P, Mallipatu S, Cognon G, Thiébaut D, et al. Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: application to plant extracts. J Chromatogr A 2009;1216:2866–72. http://dx.doi.org/10.1016/j.chroma.2008.09.027.

[60] Faber NM, Buydens LMC, Kateman G. Generalized rank annihilation method. I. Derivation of eigenvalue problems. J Chemom 1994;8:147–54. http://dx.doi.org/10.1002/cem.1180080206.

[61] Li S, Hamilton JC, Gemperline PJ. Generalized rank annihilation method using similarity transformations. Anal Chem 1992;64:599–607. http://dx.doi.org/10.1021/ac00030a007.

[62] Sanchez E, Kowalski BR. Generalized rank annihilation factor analysis. Anal Chem 1986;58:496–9. http://dx.doi.org/10.1021/ac00293a054.

[63] Prazen BJ, Synovec RE, Kowalski BR. Standardization of second-order chromatographic/spectroscopic data for optimum chemical analysis. Anal Chem 1998;70:218–25. http://dx.doi.org/10.1021/ac9706335.

[64] Fraga CG, Prazen BJ, Synovec RE. Objective data alignment and chemometric analysis of comprehensive two-dimensional separations with run-to-run peak shifting on both dimensions. Anal Chem 2001;73:5833–40. http://dx.doi.org/10.1021/ac010656q.

[65] Moler C, Stewart G. An algorithm for generalized matrix eigenvalue problems. SIAM J Numer Anal 1973;10:241–56. http://dx.doi.org/10.1137/0710024.

[66] de Juan A, Jaumot J, Tauler R. Multivariate curve resolution (MCR). Solving the mixture analysis problem. Anal Methods 2014;6:4964–76. http://dx.doi.org/10.1039/C4AY00571F.

[67] Parastar H, Radović JR, Jalali-Heravi M, Diez S, Bayona JM, Tauler R. Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC × GC-TOFMS combined to multivariate curve resolution. Anal Chem 2011;83:9289–97. http://dx.doi.org/10.1021/ac201799r.

[68] Kuligowski J, Quintás G, Tauler R, Lendl B, de la Guardia M. Background correction and multivariate curve resolution of online liquid chromatography with infrared spectrometric detection. Anal Chem 2011;83:4855–62. http://dx.doi.org/10.1021/ac2004407.

[69] Parastar H, Tauler R. Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: a new insight to address current chromatographic challenges. Anal Chem 2014;86:286–97. http://dx.doi.org/10.1021/ac402377d.

[70] Bailey HP, Rutan SC. Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine. Chemom Intell Lab Syst 2011;106:131–41. http://dx.doi.org/10.1016/j.chemolab.2010.07.008.

[71] Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis. UCLA working papers in phonetics, 16: Ann Arbor, Michigan: University Microfilms; 1970. p. 1–84, No. 10,085.

[72] Carroll JD, Chang J-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. Psychometrika 1970;35:283–319. http://dx.doi.org/10.1007/BF02310791.

[73] Bro R. PARAFAC. Tutorial and applications. Chemom Intell Lab Syst 1997;38:149–71. http://dx.doi.org/10.1016/S0169-7439(97)00032-4.

[74] Hoggard JC, Synovec RE. Automated resolution of nontarget analyte signals in GC × GC-TOFMS data using parallel factor analysis. Anal Chem 2008;80:6677–88. http://dx.doi.org/10.1021/ac800624e.

[75] Hoggard JC, Siegler WC, Synovec RE. Toward automated peak resolution in complete GC × GC–TOFMS chromatograms by PARAFAC. J Chemom 2009;23:421–31. http://dx.doi.org/10.1002/cem.1239.

[76] Bro R, Andersson CA, Kiers HAL. PARAFAC2—Part II. Modeling chromatographic data with retention time shifts. J Chemom 1999;13:295–309. http://dx.doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4<295::AID-CEM547>3.0.CO;2-Y.

[77] Skov T, Hoggard JC, Bro R, Synovec RE. Handling within run retention time shifts in two-dimensional chromatography data using shift correction and modeling. J Chromatogr A 2009;1216:4020–9. http://dx.doi.org/10.1016/j.chroma.2009.02.049.

[78] Watson NE, Siegler WC, Hoggard JC, Synovec RE. Comprehensive three-dimensional gas chromatography with parallel factor analysis. Anal Chem 2007;79:8270–80. http://dx.doi.org/10.1021/ac070829x.

[79] Wise BM, Gallagher NB, Bro R, Shaver JM, Windig W, Koch SR. PLS Toolbox 3.5 for use with Matlab™. Eigenvector Research, Inc. ISBN: 0-97611840-8.

[80] Bro R, Smilde AK. Principal component analysis. Anal Methods 2014;6:2812–31. http://dx.doi.org/10.1039/C3AY41907J.

[81] Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta 1986;185:1–17. http://dx.doi.org/10.1016/0003-2670(86)80028-9.

[82] Haaland DM, Thomas EV. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Anal Chem 1988;60:1193–202. http://dx.doi.org/10.1021/ac00162a020.

[83] de Jong S. SIMPLS: an alternative approach to partial least squares regression. Chemom Intell Lab Syst 1993;18:251–63. http://dx.doi.org/10.1016/0169-7439(93)85002-X.

[84] Bro R. Multiway calibration. Multilinear PLS. J Chemom 1996;10:47–61. http://dx.doi.org/10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C.

[85] Andersson CA, Bro R. The N-way Toolbox for MATLAB. Chemom Intell Lab Syst 2000;52:1–4. http://dx.doi.org/10.1016/S0169-7439(00)00071-X.

[86] Olivieri AC, Wu H-L, Yu R-Q. MVC3: a MATLAB graphical interface toolbox for third-order multivariate calibration. Chemom Intell Lab Syst 2012;116:9–16. http://dx.doi.org/10.1016/j.chemolab.2012.03.018.

[87] Gurden SP, Westerhuis JA, Bro R, Smilde AK. A comparison of multiway regression and scaling methods. Chemom Intell Lab Syst 2001;59:121–36. http://dx.doi.org/10.1016/S0169-7439(01)00168-X.

[88] Bro R, Heimdal H. Enzymatic browning of vegetables. Calibration and analysis of variance by multiway methods. Chemom Intell Lab Syst 1996;34:85–102. http://dx.doi.org/10.1016/0169-7439(96)00019-6.

[89] Smilde AK. Comments on multilinear PLS. J Chemom 1997;11:367–77. http://dx.doi.org/10.1002/(SICI)1099-128X(199709/10)11:5<367::AID-CEM481>3.0.CO;2-I.

[90] Allegrini F, Olivieri AC. IUPAC-consistent approach to the limit of detection in partial least-squares calibration. Anal Chem 2014;86:7858–66. http://dx.doi.org/10.1021/ac501786u.

[91] Faber N (Klaas) M, Bro R. Standard error of prediction for multiway PLS: 1. Background and a simulation study. Chemom Intell Lab Syst 2002;61:133–49. http://dx.doi.org/10.1016/S0169-7439(01)00204-0.

[92] Prazen BJ, Johnson KJ, Weber A, Synovec RE. Two-dimensional gas chromatography and trilinear partial least squares for the quantitative analysis of aromatic and naphthene content in naphtha. Anal Chem 2001;73:5677–82. http://dx.doi.org/10.1021/ac010637g.

[93] Pierce KM, Hoggard JC. Chromatographic data analysis. Part 3.3.4: handling hyphenated data in chromatography. Anal Methods 2014;6:645–53. http://dx.doi.org/10.1039/C3AY40965A.

[94] Centner V, Massart D-L, de Noord OE, de Jong S, Vandeginste BM, Sterna C. Elimination of uninformative variables for multivariate calibration. Anal Chem 1996;68:3851–8. http://dx.doi.org/10.1021/ac960321m.

[95] Johnson KJ, Synovec RE. Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis. Chemom Intell Lab Syst 2002;60:225–37. http://dx.doi.org/10.1016/S0169-7439(01)00198-8.