Published in final edited form as:

Anal Chim Acta. 2012 April 20; 723: 7–17. doi:10.1016/j.aca.2012.02.019.

Semi-automated alignment and quantification of peaks using parallel factor analysis for comprehensive two-dimensional liquid chromatography-diode array detector data sets

Robert C. Allen and Sarah C. Rutan

Virginia Commonwealth University, 1001 West Main Street, P.O. Box 842006, Richmond, Virginia 23284-2006, (804) 828-7517

Sarah C. Rutan: srutan@vcu.edu

Abstract

Parallel factor analysis was used to quantify the relative concentrations of peaks within four-way comprehensive two dimensional liquid chromatography-diode array detector data sets. Since parallel factor analysis requires that the retention times of peaks between each injection are reproducible, a semi-automated alignment method was developed that utilizes the spectra of the compounds to independently align the peaks without the need for a reference injection. Peak alignment is achieved by shifting the optimized chromatographic component profiles from a three-way parallel factor analysis model applied to each injection. To ensure accurate shifting, components are matched up based on their spectral signature and the position of the peak in both chromatographic dimensions. The degree of shift, for each peak, is determined by calculating the distance between the median data point of the respective dimension (in either the second or first chromatographic dimension) and the maximum data point of the peak furthest from the median. All peaks that were matched to this peak are then aligned to this common retention data point. Target analyte recoveries for four simulated data sets were within 2 % of 100 % recovery in all cases. Two different experimental data sets were also evaluated. Precision of quantification of two spectrally similar and partially coeluting peaks present in urine was as good as or better than 4 %. Good results were also obtained for a challenging analysis of phenytoin in waste water effluent, where the results of the semi-automated alignment method agreed with the reference LC-LCMS/MS method within the precision of the methods.

Keywords

Comprehensive two-dimensional liquid chromatography; Diode array detector; Four-way data; Chromatographic alignment; PARAFAC

1. Introduction

Currently, there are two predominant methods, integration and multi-way analysis, for the quantification of peaks within comprehensive two dimensional gas chromatography (GC×GC) or comprehensive two dimensional liquid chromatography (LC×LC) data sets.

© 2012 Elsevier B.V. All rights reserved.

Correspondence to: Sarah C. Rutan, srutan@vcu.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The integration method is typically achieved by summing up the areas under consecutive second dimension peaks that comprise a two-dimensional peak [1–3]. However, as noted by van Mispelaar *et al.*, this method relies on complete baseline separation between adjacent peaks to produce reasonable results [4]. Bailey and Rutan developed an integration method to address the concern of van Mispelaar when analyzing LC×LC-DAD chromatograms [5]. The integration method used multivariate curve resolution by alternating least squares (MCR-ALS) to resolve coeluting peaks into two or more components, assuming that the spectra of the peaks were dissimilar enough to be resolved. However, Bailey and Rutan noted that the method was tedious, lacked automation, and somewhat subjective in regards to drawing the baseline. Bailey *et al.* also used this integration method to study the inherent difficulties in quantifying peaks in LC×LC-DAD chromatograms [6]. Bailey identified five major inherent problems found in analyzing LC×LC-DAD chromatograms: 1) data size and complexity, 2) spectral overlap, 3) chromatographic overlap, 4) retention time shifts, and 5) presence of a background signal. Alternative multi-way methods, such as the generalized rank annihilation method (GRAM) [7] or parallel factor (PARAFAC) analysis [8], address some of the problems identified by Bailey *et al.* but can only be used if no retention time shifts are present. The use of GRAM and PARAFAC to analyze four-way LC×LC-DAD data allows for a lesser degree of uncertainty in the resolved components when compared to a two-way analysis method, such as MCR-ALS, due to a decreased number of parameters requiring fitting [9].

To successfully implement multi-way methods for quantifying peaks in GC×GC or LC×LC chromatograms, two conditions must be met. First, the number of components used in the analysis needs to be correct. If the number of components is too low, chemical constituents may not be correctly resolved. If the number of components is too high, noise is included into the component profiles and/or spurious components can be introduced that have incorrect profiles (i.e., profiles that are mirror images of real components may be present in some modes). Second, the retention times of peaks within GC×GC and LC×LC chromatograms need to be reproducible for every injection. Several automated alignment methods have been developed to adjust for retention time shifts between GC×GC and LC×LC samples.

The dynamic time warping (DTW), parametric time warping (PTW), and correlation optimized warping (COW) techniques have been developed as a means of distorting the original shape of chromatograms to match with a chosen reference chromatogram for one dimensional chromatography [10–15]. Zhang *et al.* modified the COW technique to allow for a second chromatographic dimension (²D) to be aligned simultaneously with the first chromatographic dimension (¹D) [16]. Zhang *et al.* successfully tested this modified COW on GC×GC-TOFMS chromatograms and was able to resolve both linear and non-linear shifting on selective ion monitoring (SIM) and total ion counts (TIC) chromatograms.

An alternative approach to warping the sample chromatogram is to linearly shift the sample chromatogram relative to the reference chromatogram and use a defined metric to determine the point of maximum alignment [17–21]. Although these techniques successfully aligned the respective data sets, most of the data sets aligned were either single peaks or replicate injections of several samples. Several problems potentially exist if these methods are applied to data sets with more heterogeneity between samples. First, each of these techniques requires that a suitable reference injection be selected prior to alignment. Typically the chromatogram with the most peaks is selected, although this does not guarantee that these same peaks are present in every sample. Second, all of these techniques do not take into account the chemical signatures of the peaks when alignment is performed. These techniques only align the data based on the raw shape of the chromatograms. There then exists the possibility that incorrect peak alignment could occur based on the appearance of a

peak within a sample injection at approximately the same retention time as a peak in the reference chromatogram. Zhang *et al.* did propose that the use of the SIM chromatograms be used to provide an overall warping signal if this is a possibility when performing the two dimensional COW technique [16]. However, due to the large presence of background contributions, this approach is not applicable to LC×LC-DAD data.

Once the chromatograms have been aligned, GRAM and PARAFAC analysis can be applied to the data sets. However, due to the way in which GRAM calculates the area of peaks, only two injections can be quantified at a time and one of the injections needs to be a standard. PARAFAC does not share this limitation and has been used to analyze more than two samples for three-way LC×LC data [20] and four-way LC×LC data [22]. As an alternative to the lengthy process necessary to correctly pre-align chromatographic data sets prior to using PARAFAC, PARAFAC2 has been developed to allow for quantification of the respective analytes present without the need to pre-align the data [23, 24]. PARAFAC2 allows peaks to shift between chromatograms by relaxing the bilinearity constraint on the dimension containing the chromatographic data. Skov *et al.* compared the performance of PARAFAC (after retention time alignment) and PARAFAC2 on GC×GC-TOFMS peaks [25]. Skov determined that PARAFAC was more robust for peaks with lower signal-to-noise ratios and lower concentrations. PARAFAC2, however, did eliminate the need for retention time alignment to analysis. This study focused on peaks that were fully resolved and did not study overlapped peaks. Van Mispelaar *et al.* also compared PARAFAC, PARAFAC2 and integration by summing directly for GC×GC-FID chromatograms [4]. Van Mispelaar determined that PARAFAC2 overestimated the concentrations when compared to PARAFAC. Also, van Mispelaar decided that differences in peak shapes were more detrimental to the PARAFAC2 analysis. While PARAFAC results were more accurate than PARAFAC2 results, the areas calculated by integration possessed better precision and accuracy when compared to the multi-way methods. In addition, the PARAFAC and PARAFAC2 methods were able to be automated, while the integration method required a great deal of user intervention. Also, to the authors' knowledge, an algorithm that allows for shifting in two dimensions does not currently exist for PARAFAC2. Therefore, as an alternative to the PARAFAC2 method, we propose the semi-automated alignment method described below to allow for accurate and precise quantification of peaks using PARAFAC analysis of LC×LC-DAD chromatograms.

2.0 Method Overview

The semi-automated alignment method is designed to align the peaks within the localized data region based on their PARAFAC resolved spectral components. The semiautomated alignment method consists of six general steps: i) selection of an appropriate region of the two-dimensional chromatogram for analysis, ii) selection of the number of components, iii) initial MCR-ALS analysis and identification of components, iv) three-way PARAFAC analysis on each injection, v) component matching between injections, vi) alignment of the ²D and ¹D, and vii) four-way PARAFAC analysis on the aligned data set. Each of these steps is discussed in further detail below.

2.1 Selection of an appropriate region of the two-dimensional chromatogram for analysis

While analyzing maize LC×LC-DAD samples, Porter *et al.* discovered that PARAFAC was not capable of analyzing the entire sample at once [22]. Instead, Porter *et al.* localized the PARAFAC analysis around a peak of interest. The localized region was selected to ensure that the peak was completely within the region in both the ²D and ¹D. Also, sufficient space around the peak was included to allow for successful accounting of background components. Bailey *et al.* further refined this approach by choosing a localized region that allowed for the

shifting of the peak in both the 2D and 1D [5]. The approach used in this method to determine the range of the localized region is the same approach used by Bailey *et al.*

2.2 Selection of the number of components

As stated in the introduction, PARAFAC requires that the correct number of components be selected for analysis. We have tried several automated methods to determine the number of components, such as cross-validation [26] or an F-test [27], however, the resulting number of components has always been overestimated. Instead, a process has been developed to provide the user with enough information to perform an educated guess for the number of components present in both the entire four-way data set and for each of the injections. To generate the necessary information, singular value decomposition (SVD) is performed on both the entire four-way data set and on each three-way data set corresponding to each injection. Once SVD is performed on the overall data set, and on each of the injections, the resulting log (base 10) of the singular values is plotted *vs.* the corresponding component number. Figure 1 is an example of the resulting scree plot. The user is prompted to choose the number of components for the both the entire raw data set and for each injection. The number of components in a scree plot is typically chosen by selecting the number of components at the point at which the slope of the plot changes suddenly, indicated by point A in Figure 1 [28]. Alternative approaches for interpretation of the scree plot have included the identification of the largest gap in the plot, point B [22, 29]. Unfortunately, due to the variability in the singular values between the entire raw data set and each injection, neither of these approaches provided a fool-proof method for determining the number of components. The criterion used in developing this method is to identify a break within the singular values, denoted by C in Figure 1, that accounts for the number of visible peaks present (by inspection of a contour plot) and the presence of background components [5].

2.3 Initial MCR-ALS analysis and identification of components

The PARAFAC algorithm used in this method allows for the implementation of non-negativity, unimodality, and selectivity constraints for each component within each mode selectively. These constraints were implemented in a similar fashion to that described by Bezemer and Rutan for MCR-ALS [30]. The non-negativity constraint is used in all four modes: 2D , 1D , sample, and spectra. Unimodality is only applied in the 2D and 1D . The unimodality constraint used in this method halves the intensity of the lesser peak over successive iterations, effectively removing the peak from the component [31]. The more common vertical and horizontal implementations of unimodality were not able to be used due to the possibility of noisy profiles in the 2D and 1D . The selectivity constraint is implemented in the 2D , the 1D , the sample dimension, and the spectral dimension.

However, to correctly construct each constraint, the components within each mode need to have been previously identified as either an analyte component or a background component. The procedure used to generate an initial estimate of each component for each mode is an approach similar to the iterative key set factor analysis (IKSFA)-ALS-ssel method developed by Bailey and Rutan [5]. The user has the choice to use either IKSFA [32, 33] or orthogonal projection approach (OPA) [34] to generate the initial guesses. The OPA used in this method has been modified to allow for the same iterative approach used in IKSFA and will be called iterative orthogonal projection approach (IOPA) to distinguish from normal OPA. Based on the number of components selected for the raw data set and each injection, IKSFA (or IOPA) produces the most dissimilar set of spectra possible for the selected number of components. These spectra are used as initial estimates for MCR-ALS to generate the chromatographic profiles of each component in both the entire raw data set and for each injection. These resolved two-way chromatographic profiles are reshaped into four-way structures and plotted as contour plots. The user is directed to identify the analyte and

background components. Once the components for the entire raw data set have been identified, the user is prompted, for each analyte component, to select the point at which selectivity will be implemented in the spectral dimension, i.e., the wavelength above which the analytes are not expected to absorb. MCR-ALS is then performed on the entire raw data set again and the resulting constrained spectra are then used as the reference spectra for the component matching step described in Section 2.5.

2.4 Three-way PARAFAC analysis on each injection

To allow for each peak within the data set to be aligned separately, three-way PARAFAC is performed on each injection to obtain the component profiles in the 2D , 1D , and spectral mode. Two benefits accrue from applying three-way PARAFAC to each injection separately. First, the likelihood of overfitting the data due to non-existent components present in other injections is minimized. Second, applying three-way PARAFAC to each injection allows for each injection to be constrained by trilinearity. To ensure that each injection is trilinear prior to performing three-way PARAFAC analysis, three options are given to allow the user to better constrain and correct the analyte components.

The first option is whether the user wishes to split an analyte component into two or more components in the 2D , the 1D , or both. This option is included in the method due to the possibility that more than one peak may be present in an analyte component, because two compounds have highly similar spectra. If this is the case, the use of unimodality during the three-way PARAFAC analysis would result in the smaller of the two peaks not being included in the analyte component. The approach used to split the analyte component is a modification of the approach used by Tistaert *et al.* [35]. An example of this process is shown in Figure 2 where a single PARAFAC component containing two peaks is being split into two analyte components. In Figure 2A, the 1D is selected as the dimension where the component is being split, and a point is selected where the split will occur, denoted by the dashed white line. Next, an identical duplicate of the analyte component in Figure 2A is generated as a new component. Since the MCR-ALS components will be recombined prior to the three-way PARAFAC analysis, the sections of the analyte components outside of the specified region need to be set to zero. In the example in Figure 2, the region to the right of the dashed line is set to zero, Figure 2B, and the intensities at the point at which the component was split, the dashed line, are then multiplied by 0.5. Likewise, the region of the newly generated component, Figure 2C, to the left of the dashed line is then set to zero and halved at the dashed line. To ensure that the split components do not recombine during three-way PARAFAC analysis, a chromatographic selectivity constraint, in this case in the 1D , is used so that the region where the peak is not present is set to zero for each component. Finally, the number of components for this injection is increased by the number of additional components created by splitting.

To provide the requisite trilinearity for three-way PARAFAC analysis, the user next has the option to correct for within injection retention time shifts. If a peak does not possess the same retention time in each 2D chromatogram, then that analyte component does not possess trilinearity. The process to correct for within injection retention time shifting is shown in Figure 3. In Figure 3A, the analyte component has an asymmetric contour, indicating that the retention times between the 2D chromatograms are not identical. The user selects the 1D time points where the peak is present, 7.86 min, 7.89 min, 7.92 min and 8.00 min in Figure 3A, and the 2D chromatograms corresponding to those 1D data points are plotted, Figure 3B. The user is prompted to select the maximum positions of the 2D chromatograms, which give the numbers shown in Figure 3B. From these maximum positions, each of the 2D chromatograms is shifted so that the maxima all occur at the earliest selected point, Figure 3C. To accommodate change in the length of the 2D chromatogram, each 2D chromatogram is extended by using the intensity value at the latest point so that all of the 2D

chromatograms remain the same size. When the aligned analyte component is plotted as a contour plot, the resulting shape of the peak is more symmetrical, as shown in Figure 3D.

The third option is to constrain the ²D or ¹D with selectivity. Four choices are available to the user. The user can decide to constrain each analyte component in either the ²D or ¹D, in both the ¹D and ²D, or not constrain the analyte component at all. However, if the analyte was previously split in either the ²D or ¹D, the option to implement selectivity on that component in the split dimension is not available because the selectivity constraint has been implemented previously.

After all three options have been either implemented or rejected, the modified MCR-ALS components, both analyte and background, are reconstructed into a pseudo-data set as follows

$$\mathbf{X}_{reconstructed} = \mathbf{CS}^T \quad (1)$$

where the matrix **C** contains the resolved chromatographic profiles and the matrix **S** contains the resolved spectra. The residuals are not added since, in theory, the residuals should just be random noise. At this point, three-way PARAFAC is performed on the reconstructed data for each individual injection.

2.5 Component matching between injections

After three-way PARAFAC is performed on each injection separately, the possibility exists that the component profiles may not been in the same order for each injection. A matching scheme was created to allow for analytes to be identified in comparison to the reference spectra generated in the MCR-ALS step as described in section 2.3. In addition, the general trend of the position of the analyte peaks in both the ²D and ¹D is taken into consideration to prevent outliers with the same spectra from being aligned incorrectly.

The general procedure for matching the spectral component profiles between injections is illustrated in Figure 4. A table, containing the Pearson correlation coefficients between all possible pairs, is constructed with the reference components occupying the rows and the injection components occupying the columns, as shown in Figure 4A. At the start of the matching, the user is prompted to select a minimum threshold for confirming a spectral match; 0.9 has proven more than sufficient to ensure accurate matching for the data sets examined in this article. Since only analyte components are aligned, the background components in the reference data set, identified as components one and three in this example, are removed from consideration, the grayed out boxes in Figure 4B. The removal of background components is then repeated for those components which had been identified as background in the injection of interest, Figure 4C. Once the background components have been removed from consideration, starting with the first available correlation coefficient, each coefficient is compared to the threshold. If the correlation coefficient is larger than the threshold and larger than any previous coefficients then that coefficient is selected as a match. This process is repeated until all components are matched or all remaining coefficients are less than the threshold, Figure 4D. A slight modification exists if a matched component was previously split within an injection. In this case, the generated component is not directly compared to the reference spectra. Instead, only after a match has been determined between one of the reference spectra and the pre-split component, the split component is automatically matched to the same reference spectrum. Once the spectral matching is complete for all injections, a trail of component matches from the injections for each reference component has been generated and will be used as a template for matching the ²D components between each injection.

After all of the components for each injection have been matched to the reference spectra, incorrect spectral matches are eliminated based on the peak positions in the 2D . Unlike the spectral matching procedure, verifying component matches between injections in the 2D does not use correlation coefficients. This change in strategy is due to the inconsistency in the position of the peak within the 2D and the lack of a set reference 2D chromatograms. Instead, the 2D matching scheme uses the maximum positions of the analyte components from the three-way PARAFAC analysis. The 2D chromatograms are smoothed using a Whitaker smoothing using a weighting parameter of nine [36, 37] prior to determining the maximum position to reduce the impact of noise on both the verification of components and the alignment of the 2D in Section 2.6. Using the spectral component trail for each reference data set component identified as an analyte, linear regression is used to calculate a slope and intercept from the maximum positions, y , and the corresponding injection number, x . The residuals are then calculated for every injection. Starting with the largest residual, the residual is then compared to a user defined threshold, the absolute maximum amount of deviation from the line allowed for each of the positions. If the residual is greater than the threshold, the corresponding injection and component is eliminated as a match, and linear regression is performed again. The comparison of the residuals to the threshold is continued until all residuals are less than or equal to the threshold. These injections are used as the starting variables for the matching within the 1D .

Peak position matching for the 1D follows the same procedure except that the 1D component profiles are interpolated prior to matching. Interpolation is performed to alleviate the sampling effects on the shapes of the 1D chromatograms and the apparent retention times. The interpolation used in this method is either Gaussian fitting or exponentially modified Gaussian (EMG) fitting depending on the user's choice. Gaussian fitting should be sufficient in most cases except where significant tailing occurs in the 1D . Gaussian fitting was chosen as the interpolation technique based on results obtained from comparing five different interpolation strategies in a previous study [38]. Also, the flat baseline obtained when implementing Gaussian fitting allows for peak shifting while keeping the number of the data points in the resolved profile equal to the number of data points in the original profile, which is required for implementing the data reconstruction step described in section 2.7. The matching results from the 1D are then used to initialize the alignment of the 2D and 1D .

2.6 Alignment of the 2D and the 1D

The alignment for each reference component is conducted separately from other components, taking into consideration components that were previously split within certain injections. This approach allows for each peak to be aligned individually instead of relying on an overall alignment of the localized window. This approach was also designed to prevent peaks that exist on the edge of the localized window from being distorted during alignment. Within the 2D and 1D , the positions of the peak closest to the beginning and end of the localized window are determined based on the component trail, Figure 5. The distance between each point and the median of the window is calculated. Three possible scenarios arise depending on the distance. The first scenario, Figure 5A, occurs when the greater distance between the positions to the median is positive. This results in the peak within this component being shifted to a later retention time within the localized window. The second scenario, Figure 5B, occurs when the distance between the two points to the center of the window is equal. The peaks within this component are then aligned so that the retention time of the peak is set to the center of the localized region. The third scenario, Figure 5C, occurs when the greater distance between the positions to the center is negative. The peaks are then shifted to an earlier retention time within the localized region. In each case, after being shifted, the lengths of the 2D or 1D chromatograms are different between each injection. To compensate for this, the baseline of the Gaussian fitting is extended to ensure that the

chromatogram lengths are equal for all injections. Likewise, the baseline of the ²D chromatogram is also extended to ensure that the number of points in the ²D chromatograms remains the same for each injection. Depending on the direction of the shift, the baseline is extended in the opposite direction.

2.7 Four-way PARAFAC analysis on the aligned data set

After the ²D and ¹D chromatograms are aligned, the three different modes for each injection are recombined according to

$$\mathbf{X}_{ijkl} = \sum_{n=1}^N a_i b_j c_k d_l \quad (2)$$

To maintain the size of the aligned data set compared to the raw data set, the ¹D is resampled to remove the interpolated data points. The reconstructed four-way data are then reshaped into a two-way structure. The IKSFA (or IOPA) results from Section 2.3 are then used to reinitialize a constrained (using the same constraints as used previously) MCR-ALS of this reconstructed two-way data set. The resolved chromatographic and spectral profiles are then used as initial estimates for four-way PARAFAC analysis. The resulting sample component profiles provide the relative concentrations of the analytes within each injection.

3.0 Experimental

The development and validation of the method was conducted on a Dell® Optiplex 755, Intel(R) Core™2 Duo CPU, E6550 @ 2.33 GHz, 3.23 GB of RAM. The software used to implement the method was MATLAB® software 2009a (Mathworks, Inc. Natick, MA) version 7.8.0.347. The non-linear least squares Gaussian and EMG fitting required the use of the MATLAB optimization toolbox. Both the semi-automated alignment method script functions and the simulated data sets can be obtained by contacting the corresponding author.

3.1 Data Sets

The method was developed and validated using four simulated and two experimental data sets. The simulated data sets were constructed using the same methodology as previously reported by Allen and Rutan [38]. The simulated data sets consisted of four separate peaks; A, B, C, and D. Peak A was the primary peak of interest in all four simulated data sets and consisted of a five-level calibration, with each level in triplicate, and four “unknown” samples, with each unknown in quintuplicate. Peak B was an interferent peak present in each of the “unknown” samples in the second simulation. Peak C was a secondary peak with the same style of calibration curve and “unknown” samples as Peak A in the third simulated data set. Peak D was an interferent present in some calibration samples and “unknown” samples for the fourth simulated data set. Peak D also possessed an identical spectra to peak A.

The first experimental data set was obtained from the Carr research group at the University of Minnesota and consisted of fourteen replicate injections (with a ²D run time of 21 s and a ¹D run time of 30 min) of a urine sample [39]. A small section, 2.15 s in the ²D and 2.45 min in the ¹D, was selected for analysis, shown in Figure 6A. The resulting size of the urine data set was 87 data points in the ²D, 8 data points in the ¹D, 14 data points in the sample dimension, and 126 data points in the spectral dimension. The spectral range of the data set was 200 to 700 nm with an interval between collected wavelengths of 4 nm. This section contained four peaks, the two primary peaks of interest (Peak 11 and Peak 12, respectively,

as identified by Bailey and Rutan [6] (shown in Figure 6 of that reference) located approximately in the center of the section and two coeluting peaks in the upper left corner of the section. The two primary peaks were chosen for analysis due to both peaks possessing the same spectra [6], Figure 6B.

The second experimental data set was obtained from the Stoll research group at Gustavus Adolphus College, MN and consisted of two sets of calibration samples [40, 41]. The first set of calibration samples was a series of duplicate spikes (0, 25, 50, 75, 150 ppb) of phenytoin in distilled water. The second calibration curve was a standard addition (0, 25, 50, 75, 150 ppb) analysis of a 1000x concentrated waste water extract sample containing phenytoin. Like the distilled water calibration samples, each of the levels in the standard addition experiment was duplicated. In addition to the phenytoin peak, the phenytoin data set also contained an unknown interferent peak that was severely overlapped with the phenytoin peak. A localized region (8.75 to 20.0 s in the ²D and 7.82 to 8.02 min in the ¹D) surrounding the phenytoin and interferent peak was selected to minimize extraneous influences (i.e., other interferents or background fluctuations) on the resulting relative areas, shown in Figure 7A. The resulting size of the phenytoin data set was 225 data points in the ²D, 6 data points in the ¹D, 20 data points in the sample dimension, and 101 data points in the spectral dimension. The spectral range of the data set was 200 to 600 nm with an interval between collected wavelengths of 4 nm.

3.2 Data analysis scheme

The resulting relative areas from the four-way PARAFAC analysis and the methods developed by Tistaert *et al.* [35] and Bailey *et al.* [6] were normalized by dividing each relative area by the maximum relative area for each component. This normalization scheme was carried out for the four simulated data sets and the two experimental data sets. The normalization was conducted to allow for a better comparison between each of the methods.

The first, second, and third simulated data sets were analyzed using the method described in this paper, applying four-way PARAFAC to the unaligned data set, and applying MCR-ALS to each data set. The fourth simulated data set was analyzed using the method described in this paper, applying four-way PARAFAC to the unaligned data set, and applying the MCR-ALS with unimodality method [35]. Where possible, the constraints for the analysis of the four data sets were identical within the limitations of the methods. The component containing the two peaks with identical spectra in simulation four was split at the same ¹D for each appropriate injection for both the semiautomated alignment method and IKSFA-ALS-ssel with unimodality.

The resulting relative areas obtained by four-way PARAFAC analysis and summed chromatographic components from MCR-ALS were used to calculate the accuracy of the calculated values for the “unknown” samples for peaks A and C using the following formula:

$$\% \text{recovery} = 100 + \frac{(\text{actual} - \text{expected})}{\text{expected}} \times 100 \quad (3)$$

where actual is the value obtained from using the slope and intercept obtained from linear regression and expected is the value originally used as the area of the sampled ¹D peak. The average % recovery between the four “unknown” samples was then calculated. The corresponding errors for the % recoveries were determined by calculating the percent relative standard deviations (% RSD) from the normalized four-way PARAFAC relative areas. Appropriate error propagation was used to generate the error for the average % recovery of the four “unknown” samples.

Both of the experimental data sets were analyzed using the method described in this paper. The MCR-ALS with unimodality method [35] was used to analyze the urine data set and the % RSDs of the summed resulting peak chromatograms were compared to the relative areas % RSDs from the present semi-automated alignment method. The component containing peaks 11 and 12 was split at the same 1D point for both methods. All other constraints used were applied in the same fashion as previously described for the simulated data sets.

The IKSFA-ALS-ssel method [6] was used with the phenytoin data set to provide a comparison to results obtained from the semi-automated alignment method. The unknown phenytoin sample concentrations were calculated using both the direct and standard addition calibration curves. The reported errors for the phenytoin concentrations were determined using a 95 % confidence interval. Instead of manually integrating the peaks as was done by Tistaert *et al.* [35] and Bailey *et al.* [6], the resulting chromatographic components were summed to generate the relative peak areas. This approach was used to provide a more direct comparison to the relative areas produced by PARAFAC.

4.0 Results and discussion

4.1 Simulated data sets

The % recoveries and % RSDs for the four simulated data sets are shown in Table 1. The semi-automated alignment method produced the % recoveries closest to 100 % for each of the four simulated data sets. The summed MCR-ALS-ssel components produced the next set of % recoveries closest to 100 %. The large % recovery for simulation 3C (348 %) was due to the inability of MCR-ALS-ssel to adequately remove the background contributions from the peak C component. As a result of this lack of background removal, the calibration curve for the simulation 3C MCR-ALS-ssel component had a negative slope compared to the slopes of the semi-automated alignment method. Unlike, simulation 2, two peaks were present in both the calibration and unknown samples. The presence of both peaks led the algorithm to generate a computational form of cross contamination and shifted some of the relative areas for both peaks A and C into background components.

4.2 Urine data set

Peaks 11 and 12 have been previously analyzed by Bailey *et al.* [6] and Tistaert *et al.* [35] using IKSFA-ALS-ssel and IKSFA-ALS-ssel with unimodality, respectively. The resulting % RSDs for each technique, calculated by manual integration of the resolved component, were found to be 1.04 and 3.58 by Bailey *et al.* [6] and 2.33 and 3.68 by Tistaert *et al.* [35]. By comparison, the % RSDs obtained using the present semiautomated alignment method were 4.01 and 2.67, as shown in Table 2. A more direct comparison can be made between the proposed method and the IKSFA-ALS-ssel method of Tistaert by summing up the resulting chromatographic profiles obtained. Because the present semi-automated alignment method used a localized region of the urine data smaller than the region previously analyzed by Tistaert *et al.*, this smaller localized region has been re-analyzed using the IKSFA-ALS-ssel with unimodality method and the calculated % RSDs are reported in Table 2. The resulting % RSDs from the summation method were 11.0 and 10.3 for peaks 11 and 12, respectively, vs. 4.01 and 2.67 for the present method (values mentioned previously).

Even though a different window size was used in each of the previous papers, the resulting % RSDs were comparable taking into account the various limitations of each method. The IKSFA-ALS-ssel method used by Bailey *et al.* did not allow the separation of the two peaks into two components. Instead, Bailey *et al.* had to make a best guess of how to manually integrate the peaks leading to a favored integration of peak 11 in their analysis. The IKSFA-ALS-ssel with unimodality method used by Tistaert *et al.* did allow for the splitting of the two peaks into two different components. However, as noted by Tistaert *et al.*, the split was

not optimal for every injection, resulting in some cross contamination between the two components.

While the semi-automated alignment method produced slightly better precision for peak 12 in comparison to the IKSFA-ALS-ssel and IKSFA-ALS-ssel with unimodality, the % RSD for peak 11 was slightly higher. A possible reason for the slightly higher % RSD was that peak 11 was aligned so that the entire peak was not present in the data window. This is due to the location of peak 11 in the fourteenth injection occurring at the far left side of the window. While this allowed peak 11 and peak 12 to be further resolved during the alignment, the resulting four-way PARAFAC component for peak 11 did not return to zero in the ¹D, shown in Figure 6B. The decision to limit the width of the data window to 2.45 min, instead of 2.8 min, in the ¹D was due to the previous analysis of Bailey *et al.* [6]. Bailey *et al.* identified approximately four additional peaks to the immediate left of the data window that would have been included in the analysis had the window been expanded. An attempt was made at analyzing this larger window but the resulting spectra obtained by IKSFA and IOPA did not allow MCR-ALS to generate reliable chromatographic initial guesses for the three-way PARAFAC step of the semi-automated alignment method.

As mentioned above, a more direct comparison between the % RSDs obtained by the semi-automated alignment method can be made by comparing the summed areas from the IKSFA-ALS-ssel with unimodality method. Both methods were constrained as closely as possible to one another (the chromatographic selectivity constraints were not possible due to the limitations of the implementation of the IKSFA-ALS-ssel with unimodality method) and the components were split at the same points in every injection. The % RSD from the simple summation method using the IKSFA-ALS-ssel method with unimodality was calculated to be 11.0 and 10.3 for Peaks 11 and 12 respectively, Table 2. The large difference between the simple summation method and PARAFAC can be attributed to two factors. First, due to the data reconstruction used in the semi-automated alignment method, much more noise is eliminated using this method. This can be seen in the fit errors of the two methods; 2.51% for the IKSFA-ALS-ssel-unimod method and 11.9 % for the semi-automated alignment method. The most likely reason for this difference in the fit errors is due to the discarding of the error matrices from the MCR-ALS and three-way PARAFAC analysis for each individual injection. The second factor was the ability of the semi-automated alignment method to implement an additional constraint beyond what was possible in the IKSFA-ALS-ssel with unimodality method. The unimodality constraint, as implemented by the IKSFA-ALS-ssel method with unimodality, simply splits components into two or more components, thereby preserving each individual analyte. However, the unimodality constraint used by the semiautomated alignment method is the more traditional version where only one maximum is permitted in each component. In its current form the IKSFA-ALS-ssel method with unimodality is unable to implement this version of unimodality due to the augmented nature of the chromatographic dimension. This limitation of the IKSFA-ALS-ssel with unimodality method results in small contributions of background remaining within the resolved analyte component.

4.3 Phenytoin data set

The phenytoin data set analyzed in this paper has been previously analyzed by Groskreutz *et al.* [41] and discussed in more detail by Bailey *et al.* [42] with reported concentrations for the average phenytoin concentration in waste water samples of 42 ± 1 ppb from the standard addition curve, 36 ± 1 ppb from the direct calibration curve (both calculated by manually integrating MCR-ALS-ssel components) and 42 ± 19 ppb by LC-LC MS/MS (all error ranges are given as 95 % confidence intervals). These calculated concentrations were similar to those obtained by directly summing the MCR-ALS-ssel components after careful application of the chromatographic selectivity constraint. The summed MCR-ALS

components resulted in 32 ± 2 ppb and 37 ± 7 ppb for the direct and standard addition calibration curves, as shown in Table 3. The results obtained by the semi-automated alignment method differed depending on the type of fitting used, Gaussian or EMG. If the 1D was fitted using a Gaussian curve then the calculated concentrations of the unknown were 23 ± 5 ppb and 26 ± 3 ppb from the direct calibration and standard addition curves, respectively, Table 3. If the 1D was fitted using an EMG curve then the calculated concentrations of the unknown were 28 ± 5 ppb and 31 ± 3 ppb from the direct calibration and standard addition curves respectively, Table 3. A possible reason for the increase in calculated concentrations from the EMG fitting compared to the Gaussian fitting is an approximate average factor of 1.23 increase in the fitted 1D areas from the EMG fitting compared to the Gaussian fitting. While not an overly large increase in fitted areas, the 1.23 factor increase in fitted areas correlates to an approximate increase in the calculated concentrations for both the direct calibration and standard addition curves. Regardless of which 1D peak model is used, the resulting calculated concentrations of phenytoin are still lower than the reported literature concentrations. The lower concentrations may be due to several challenges that were encountered during the analysis of the phenytoin data set.

The first challenge was the irregular shifting of the phenytoin peak in both the 2D and 1D . The irregular shifting in the 2D was compounded by the presence of noise in the three-way PARAFAC 1D component for the one of the unknown replicate samples. The irregular shifting in the 1D resulted in different apparent peak shapes which were then fitted to either a Gaussian or EMG curve. Neither the Gaussian nor the EMG curves were able to completely reproduce the exact structure of the 1D peak. The introduction of small discrepancies between replicate injections, during the fitting step of the semiautomated alignment method, did not produce calibration curves as linear as those reported by Groskreutz *et al.* ($R^2 = 0.988$ for the standard addition and $R^2 = 0.998$ for the direct calibration) [41], as shown in Table 3. In addition to the irregular shifting of retention times between injections in the 2D and 1D , retention time shifting was observed within some of the injections. While an attempt was made to correct for these shifts, some information loss most likely occurred during the three-way PARAFAC step due to the enforcement of trilinearity. The third challenge was the presence of the interferent which coeluted very closely with the phenytoin peak in the spiked waste water treatment samples, Figure 7A. After the semi-automated alignment method analysis was concluded, the resolved spectral signature of the phenytoin component and interferent component were compared. The resulting correlation coefficient of 0.9687 indicated why the semi-automated alignment method had difficulty separating the two peaks into different components, Figure 7B. Once again, partial compensation for this challenge was achieved by selective use of a chromatographic selectivity constraint in the 2D and 1D , Figures 7C and 7D. Given the low intensity of the phenytoin peak in the unknown samples, as well as the noisy nature of the resolved component, it is entirely possible that some of the phenytoin component was included in the interferent component. While the calculated concentrations from using the semi-automated alignment method were lower than those obtained using MCR-ALS-ssel, the semi-automated alignment method did produce better calculated concentrations in comparison to the application of four-way PARAFAC analysis to the unaligned data set.

Conclusion

The semi-automated alignment method was successfully used to align and subsequently quantify peaks within four simulated and two experimental data sets. The semi-automated alignment method consistently provided % recoveries closest to 100 % compared to summing MCR-ALS-ssel components. While the % RSD for peak 11 from the urine data set was not as low as previous manual integration results from Bailey *et al.* and Tistaert *et al.*, the % RSD for peak 12 was slightly lower than the previously reported literature values.

Also, the % RSDs obtained were reproducible and because the user was not required to manually draw a baseline under the peaks. The semi-automated alignment method was also successfully demonstrated to be able to divide a single component into two spectrally similar components without a significant degradation of the results. The semi-automated alignment method, however, was found to experience difficulty when analyzing data sets containing coeluting peaks with highly similar spectra. This difficulty was able to be overcome with careful use of the ²D and spectral selectivity constraints. Because the peaks were able to be resolved into individual components, the calculated concentrations obtained from the semi-automated alignment method were within the error range of the LC-LC MS/MS results. The results obtained from the analysis of the phenytoin data set by the semi-automated alignment method were also found to be reproducible assuming that the selected peak positions when using the within sample alignment were reproducible. Small differences in the selection of the peak positions resulted in only small changes to the final concentrations (less than 0.5 ppb).

Acknowledgments

The authors would like to thank the Carr research group at the University of Minnesota and the Stoll research group at Gustavus Adolphus College for the experimental data sets. The authors also acknowledge financial support from NIH-GM-54585-13 and NSF CHE-0911330.

References

1. Beens J, Boelens H, Tijssen R, Blomberg J. J High Resol Chromatogr. 1998; 21:47–54.
2. Mondello L, Herrero M, Kumm T, Dugo P, Cortes H, Dugo G. Anal Chem. 2008; 80:5418–5424. [PubMed: 18517223]
3. Reichenbach SE. Anal Chem. 2008; 81:5099–5101. [PubMed: 19445478]
4. van Mispelaar VG, Tas AC, Smilde AK, Schoenmakers PJ, van Asten AC. J Chromatogr A. 2003; 1019:15–29. [PubMed: 14650601]
5. Bailey HP, Rutan SC. Chemom Intell Lab Syst. 2011; 106:131–141.
6. Bailey HP, Carr PW, Rutan SC. J Chromatogr A. 2011; 1217:4313–4327.
7. Wilson BE, Sanchez E, Kowalski BR. J Chemometrics. 1989; 3:493–498.
8. Bro R. Chemom Intell Lab Syst. 1997; 38:149–171.
9. Liu X, Sidiropoulos ND, Swami A. MILCOM. 2001; 2:1340–1344.
10. Wang CP, Isenhour TL. Anal Chem. 1987; 59:649–654.
11. Eilers PHC. Anal Chem. 2004; 76:404–411. [PubMed: 14719890]
12. Nielsen N-PV, Cartensen JM, Smedsgaard J. J Chromatogr A. 1998; 805:17–35.
13. Pravdova V, Walczak B, Massart DL. Anal Chim Acta. 2002; 456:77–92.
14. Tomasi G, van den Berg F, Andersson C. J Chemom. 2004; 18:231–241.
15. van Nederkassel AM, Daszykowski M, Eilers PHC, Heyden YV. J Chromatogr A. 2006; 1118:199–210. [PubMed: 16643929]
16. Zhang D, Regnier HXFE, Zhang M. Anal Chem. 2008; 80:2664–2671. [PubMed: 18351753]
17. Johnson KJ, Wright BW, Jarman KH, Synovec RE. J Chromatogr A. 2003; 996:141–155. [PubMed: 12830915]
18. Bortolato SA, Arancibia JA, Escandar GM, Olivieri AC. Chemom Intell Lab Syst. 2010; 101:30–37.
19. Fraga GG, Prazen BJ, Synovec RE. Anal Chem. 2001; 73:5833–5840. [PubMed: 11791551]
20. Fraga CG, Corley CA. J Chromatogr A. 2005; 1096:40–49. [PubMed: 16301068]
21. Pierce KM, Wood LF, Wright BW, Synovec RE. Anal Chem. 2005; 77:7735–7743. [PubMed: 16316183]
22. Porter SEG, Stoll DR, Rutan SC, Carr PW, Cohen JD. Anal Chem. 2006; 78:5559–5569. [PubMed: 16878896]

23. Kiers HAL, Berge JMFT, Bro R. *J Chemometrics*. 1999; 13:275–294.
24. Bro R, Andersson CA, Kiers HAL. *J Chemometrics*. 1999; 13:295–309.
25. Skov T, Hoggard JC, Bro R, Synovec RE. *J Chromatogr A*. 2009; 1216:4020–4029. [PubMed: 19269647]
26. Bro R, Kjeldahl K, Smilde AK, Kiers HAL. *Anal Bioanal Chem*. 2008; 390:1241–1251. [PubMed: 18214448]
27. Wasim M, Brereton RG. *Chemom Intell Lab Sys*. 2004; 72:133–151.
28. Otto, M. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*. Wiley-VCH, Federal Republic of Germany; 2007.
29. Zhu M, Ghodsi A. *Computation Stat*. 2006; 51:918–930.
30. Bezemer E, Rutan SC. *Chemom Intell Lab Syst*. 2006; 81:82–93.
31. de Juan A, Heyden YV, Tauler R, Massart DL. *Anal Chim Acta*. 1997; 346:307–318.
32. Schostack KJ, Mallinowski ER. *Chemom Intell Lab Syst*. 1989; 6:21–29.
33. Malinowski, ER. *Factor Analysis in Chemistry*. John Wiley & Sons, Inc; Hoboken, New Jersey: 1991.
34. Sanchez FC, Toft B, van den Bogaert B, Massart DL. *Anal Chem*. 1996; 68:79–85. [PubMed: 21619222]
35. Tistaert C, Bailey HP, Allen RC, Rutan SC. *Chemom Intell Lab Sys*. 2011 submitted.
36. Eilers PHC. *Anal Chem*. 2003; 75:3631–3636. [PubMed: 14570219]
37. Thekkudan, DF.; Rutan, SC. Denoising and Signal-to-Noise Ratio Enhancement: Classical Filtering. In: Brown, S.; Tauler, R.; Walczak, R., editors. *Comprehensive Chemometrics*. Elsevier; Oxford: 2009. p. 9-24.
38. Allen RC, Rutan SC. *Anal Chim Acta*. 2011; 705:253–260. [PubMed: 21962368]
39. Reichenbach SE, Carr PW, Stoll DR, Tao Q. *J Chromatogr A*. 2009; 1216:3458–3466. [PubMed: 18848329]
40. Groskreutz SR, Swenson MM, Secor LB, Stoll DR. *J Chromatogr A*. 2011 in press. 10.1016/j.chroma.2011.06.035
41. Groskreutz SR, Swenson MM, Secor LB, Stoll DR. *J Chromatogr A*. 2011 in press. 10.1016/j.chroma.2011.06.038
42. Bailey HP, Rutan SC, Stoll DR. *J Sep Sci*. 2012 submitted.

Highlights

We describe a spectrally based alignment method for LC×LC-DAD data

Alignment between injections is not dependent on selection of a reference injection

The alignment process allows for non-linear peak shifting between injections

Simulated data sets aligned by this method produced % recoveries close to 100%

The method was able to align and quantify experimental data with overlapping peaks

Figure 1.

A scree plot (from the reshaped four-way phenytoin data set) is generated by plotting the log of the singular value by the component number. The bend, indicated by A, is the normal criteria used to determine the number of components [28]. The gap, indicated by B, is the previous method used by Porter *et al.* [22] and Zhu *et al.* [29] for determining the number of components. The gap, indicated by C, is the method used by Bailey *et al.* [5] and is used by the semi-automated alignment method.

Figure 2.

A) Contour plot of the MCR-ALS component from the twelfth replicate injection of the urine sample showing the two peaks of interest within the urine data set. The dashed white line indicates the ¹D point at which the component will be split into two components. B) The resulting MCR-ALS component after splitting the original MCR-ALS component into two components. C) The newly generated MCR-ALS component created when the component shown in Figure 2A was split into two components.

Figure 3.

A) A contour plot of the MCR-ALS component containing the phenytoin peak (50 ppb) from the sixth injection of the phenytoin data set. B) A plot of the 2^{D} chromatograms from the second to fifth 1^{D} data points in 3A. The numbers (in s) indicate the position of the peak maximum for the corresponding color 2^{D} chromatogram. C) The 2^{D} chromatograms shown in 3B after shifting all of the 2^{D} chromatograms to the earliest maximum (12.60 s). D) The resulting contour plot of the MCR-ALS component after aligning the 2^{D} chromatograms.

Figure 4.

Schematic of the spectral component matching system. The reference components have been labeled in the following order: 1) background, 2) analyte, 3) background, 4) analyte. The injection components have been labeled in the following order: 1) analyte, 2) background, 3) background, 4) analyte. A) Matrix generated from all possible Pearson correlation coefficient pairs. B) Removal of the raw background components (1 and 3) from consideration. C) Removal of the injection background components (2 and 3) from consideration. D) The resulting analyte component matches between the raw reference spectra and the three-way PARAFAC spectra for this injection.

Figure 5.

Schematic of how the direction of the alignment is determined. The white and black triangles represent the earliest and latest eluting peaks. The arrow indicates the direction of alignment. A) The distance between the black triangle and the median is greater than the distance between the white triangle and median resulting in the peak being aligned to the right. B) The distance between the white and black triangles and the median are equal resulting in the peak being aligned to the middle of the data window. C) The distance between the white triangle and the median is greater than the distance between the black triangle and the median resulting in the peak being aligned to the left.

Figure 6.

A) A contour plot of the seventh injection from the urine data set. Shown in the middle of the contour plot are the two selected peaks with the same spectra. The absorbance bar on the right of the contour plot shows the intensities of the peaks present. B) The resolved spectra of the two peaks with the calculated Pearson correlation coefficient shown. C) Contour plot of the reconstructed peak shapes from the resolved four-way PARAFAC components for the two peaks of interest. The positions of the two peaks are different from Figure 6A due to having been aligned prior to reconstruction.

Figure 7.

A) A contour plot of the thirteenth injection from the phenytoin data set (25 ppb spiked waste water treatment). The color bar on the right shows the relative absorbance (in mAU) of each of the peaks present. B) The resolved spectra of the phenytoin and interferent after four-way PARAFAC analysis. The solid line is the phenytoin spectra and the dashed line is the interferent spectra. The calculated Pearson correlation coefficient is also shown. C) The resolved phenytoin peak after reconstruction from the components obtained from four-way PARAFAC analysis. D) The resolved interferent peak after reconstruction from the components obtained from four-way PARAFAC analysis. The distortion of the upper shape of the peak is caused by the use of selectivity in the 2^D .

Table 1

% recoveries for the four simulated data sets after analysis by MCR-ALS and PARAFAC

% Recovery of peaks A and C		
Simulation	MCR-ALS-ssel	Aligned PARAFAC
1A	111.9±5.5	98.0±1.9
2A	107.6±4.7	98.6±1.3
3A	109.5±7.8	98.3±2.4
3C	348±22	100.0±3.0
4A	111.7±6.1 ^a	99.3±1.9

^a calculated from % recoveries shown in Table 2 from reference [34]

Table 2

Percent relative standard deviations for Peaks 11 and 12 from the urine data set

	Semi-automated alignment	IKSFA-ALS-ssel ^a	IKSFA-ALS-ssel with unimodality ^b	IKSFA-ALS-ssel with unimodality ^c
Peak 11	4.0	1.04	2.33	11.0
Peak 12	2.7	3.58	3.68	10.3

^a from reference[6]^b from reference[34]^c component areas determined by the simple summation method

Table 3Direct and standard calibration curves for phenytoin^a

Direct calibration	Linear regression parameters	R ²	Standard error (s _y)	Concentration (ppb)
Semi-automated alignment with Gaussian fitting	Slope = 0.0060(0.0003) Intercept = 0.00(0.03)	0.976	0.055	23±5
Semi-automated alignment with EMG fitting	Slope = 0.0060(0.0003) Intercept = -0.00(0.02)	0.979	0.051	28±5
IKSFA-ALS-ssel ^b	Slope = 0.0022(0.0003) Intercept = 0.07(0.03)	0.994	0.022	32±2
Standard addition	Linear regression parameters	R ²	Standard error (s _y)	Concentration (ppb)
Semi-automated alignment with Gaussian fitting	Slope = 0.0056(0.0002) Intercept = 0.14(0.02)	0.989	0.030	26±3
Semi-automated alignment with EMG fitting	Slope = 0.0054(0.0002) Intercept = 0.17(0.02)	0.989	0.030	31±3
IKSFA-ALS-ssel ^b	Slope = 0.0027(.0007) Intercept = 0.44(0.06)	0.940	0.070	37±7

^a calculated concentrations were compared to the LC-LC MS/MS concentration of 42±19 ppb obtained by Groskreutz *et al.* [40]^b calculated using the simple summation method