Published in final edited form as:

Anal Chim Acta. 2011 October 31; 705(0): 253–260. doi:10.1016/j.aca.2011.06.022.

Investigation of interpolation techniques for the reconstruction of the first dimension of comprehensive two-dimensional liquid chromatography-diode array detector data

Robert C. Allen and Sarah C. Rutan*

Virginia Commonwealth University, 1001 West Main Street, P.O. Box 842006, Richmond, VA 23284-2006, United States

Abstract

Simulated and experimental data were used to measure the effectiveness of common interpolation techniques during chromatographic alignment of comprehensive two-dimensional liquid chromatography-diode array detector (LC $\times$ LC-DAD) data. Interpolation was used to generate a sufficient number of data points in the sampled first chromatographic dimension to allow for alignment of retention times from different injections. Five different interpolation methods, linear interpolation followed by cross correlation, piecewise cubic Hermite interpolating polynomial, cubic spline, Fourier zero-filling, and Gaussian fitting, were investigated. The fully aligned chromatograms, in both the first and second chromatographic dimensions, were analyzed by parallel factor analysis to determine the relative area for each peak in each injection. A calibration curve was generated for the simulated data set. The standard error of prediction and percent relative standard deviation were calculated for the simulated peak for each technique. The Gaussian fitting interpolation technique resulted in the lowest standard error of prediction and average relative standard deviation for the simulated data. However, upon applying the interpolation techniques to the experimental data, most of the interpolation methods were not found to produce statistically different relative peak areas from each other. While most of the techniques were not statistically different, the performance was improved relative to the PARAFAC results obtained when analyzing the unaligned data.

Keywords

Data alignment; Interpolation; Comprehensive; Two dimensional liquid chromatography; Chemometrics; Parallel factor analysis

1. Introduction

Comprehensive two dimensional separation systems, such as LC $\times$ LC and GC $\times$ GC, produce data that possess a two-way structure [1-6]. If a multichannel detector (frequently a diode array detector (DAD) for LC $\times$ LC and a mass spectrometer (MS) for GC $\times$ GC) is used to record the system response, then the data structure becomes three-way [1,7,8]. When multiple samples are analyzed, the data structure is extended to four-way [9]. In order to process these multi-dimensional data structures, multi-dimensional algorithms, such as multivariate curve resolution with alternating least squares (MCR-ALS), generalized rank annihilation method (GRAM), and parallel factor (PARAFAC) analysis can be employed. Bailey and Rutan [10] used MCR-ALS to analyze five isolated peaks from six replicate

injections of LC $\times$ LC-DAD standard chromatograms. The resulting resolved chromatograms were manually integrated to determine the relative area of each of the peaks. While this approach resulted in relative areas with percent relative standard deviations (%RSD) ranging from 1.4 to 4.7%, the manual integration approach can become very labor intensive if a large number of peaks are present or if a large number of samples are analyzed. For chromatograms containing a larger number of peaks or larger sample sizes, a more automated quantification approach is desirable. Both GRAM and PARAFAC have been automated using the MATLAB programming environment to analyze GC $\times$ GC and LC $\times$ LC chromatograms [2,11-14]. However, in order to correctly implement either GRAM or PARAFAC, the chromatographic dimensions need to be aligned so that the peaks occur at the same retention time for each injection.

While the second chromatographic dimension (2D) is easily aligned by a simple linear shift due to the high data collection frequency [15], the alignment of the first chromatographic dimension (1D) is complicated by the sampling process that occurs between the two chromatographic columns. As noted by Fraga et al. [15], the 1D peaks typically possess a low data density (three to four data points per peak for LC $\times$ LC chromatograms), thereby requiring interpolation to insure that accurate shifting of peaks between injections is possible. The predominant interpolation strategy found within the literature for aligning the 1D of two dimensional chromatographic data consists of performing a linear interpolation on the existing data points in order to ensure that a sufficient number of points exist for alignment [2,4]. Typically a localized region of the two dimensional chromatogram containing the peaks of interest is selected for alignment and then interpolated. A reference chromatogram (typically a standard) is chosen as the basis to which the other chromatograms will be aligned. The 2D and 1D are then iteratively shifted until the criteria for selecting the optimal shift has been met. Currently, several different methods have been used for determining when the alignment criterion has been met.

Fraga et al. [15] calculated the pseudo-rank (the minimum number of components required to explain a data matrix without noise) to determine the necessary amount of shifting between the injections to properly align GC $\times$ GC peaks. An augmented matrix is generated by stacking the sample and reference in the dimension being aligned. The pseudo-rank was determined by calculating the percent residual variance after dividing the singular values associated with noise by the sum of all the singular values obtained from singular value decomposition of the augmented matrix [16]. Alignment was achieved when the pseudo-rank was calculated to be at a minimum. Fraga and Corley [2] attempted to use this approach when aligning simple peaks in LC $\times$ LC-UV chromatograms. However, in some cases, the rank determination alignment method proved insufficient to correctly align the LC $\times$ LC-UV data. Instead, Fraga and Corley iteratively shifted the LC $\times$ LC-UV chromatograms and then performed GRAM and PARAFAC analysis on the shifted data. The best shift was determined when the smallest sum of squares was calculated after the PARAFAC analysis. However, since GRAM can only be applied to two injections at a time (typically a standard and a sample), this technique is not applicable to the alignment of large number of samples. In addition, the necessity of performing GRAM and PARAFAC analysis after every shift becomes very time consuming if large shifts are required. As an alternative approach, Pierce et al. [4] modified a one dimensional alignment algorithm [17] to account for a second chromatographic dimension. The Pearson correlation coefficient from each retention time shift was calculated between the reference and sample chromatograms. The retention time shift corresponding to the maximum correlation was then used for alignment. This approach was successfully used to align GC $\times$ GC-FID data. However, the use of linear interpolation in the 1D may prove to be insufficient to correctly align the peaks for two reasons. First, the peaks within a LC $\times$ LC chromatogram may experience non-linear shifts, resulting in different sampled 1D peak shapes, due to changes in pump performance or ambient

temperature, column degradation, or evaporation of volatile compounds from the mobile phase [18]. Second, the modulation ratio, M_R , typically used for LC $\times$ LC experiments is much less than the M_R in GC $\times$ GC experiments [4,19,20]. An alternative approach that will be explored in this article is to use common interpolation techniques to recreate the original 1D peak shape and align the reconstructed 1D peak to a common retention time.

2. Theory

2.1. Sampling of first dimension

Comprehensive two-dimensional separations are achieved when the entire effluent from the 1D column is injected into the 2D column. The injection of the 1D effluent into the 2D is accomplished through the use of a sampling device, usually a switching valve in LC $\times$ LC [21] and a thermal modulator in GC $\times$ GC [22]. An example of the injection pattern resulting from this modulation is illustrated in Fig. 1. The dashed lines in Fig. 1A represent the individual 2D chromatograms rearranged into a two dimensional structure. The length of time between each valve switch determines the length of the 2D chromatograms and the number of 2D chromatograms corresponds to the number of times the 1D effluent is sampled. Once the area of each of the sequential 2D chromatograms is summed, the pattern illustrated in Fig. 1B is obtained. Each marker represents the total amount of signal (area) present in each 2D chromatogram. The alternating $\blacktriangledown$ and $\oplus$ markers represent the alternating 1D column effluent injections onto the 2D column. The shape of the peak pattern is determined by the sampling phase. The sampling phase (ϕ) is determined by the position of the retention time of the 1D peak (t_R) relative to the center (T) of the sampling window (t_s) [23,24]. The meaning of these parameters is depicted graphically in Fig. 1C. If the retention time of the peak is in the center of the window

then the peak has a sampling phase of $\phi = 0$ and is considered to be completely “in-phase”. If the retention time of the peak exists at the edge of the sampling window then the peak has a sampling phase of $\phi = \pm\pi$ and is considered to be completely “out-of-phase”. If the retention time of the peak exists somewhere in between then the peak has a sampling phase between 0 and π . The sampling phase is calculated according to Eq. (1) [24,25]:

$$\phi = \frac{2\pi(T - t_R)}{t_s} \quad (1)$$

2.2. PARAFAC

As noted previously, the data structure from a LC $\times$ LC-DAD system is four dimensional if multiple injections are being analyzed. If initial estimates for three of the four dimensions have been determined, the PARAFAC model [26]:

$$x_{ijkl} = \sum_{n=1}^N a_{kn} b_{in} c_{jn} d_{ln} + e_{ijkl} \quad (2)$$

can be used to deconstruct the original data matrix, $\mathbf{X}$, into four matrices, $\mathbf{A}_{K \times N}$, $\mathbf{B}_{J \times N}$, $\mathbf{C}_{K \times N}$, and $\mathbf{D}_{L \times N}$, which contain the pure component profiles as a function of the sample number, the 2D elution time, the 1D elution time, and the wavelength, respectively. The error array, $\mathbf{E}$, contains the residuals. N is the number of components being determined, and I , J , K , and L are number of data points in each of the respective dimensions. The main advantage of PARAFAC is that the component solutions produced by PARAFAC are truly unique if the correct number of components is selected and the data obey the multilinearity rule [27]. If an incorrect number of components are selected, the resolution results (i.e., the final $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$,

and **D** matrices) may not be correct. Therefore, it is important to be able to determine the correct number of components. The number of components are usually determined by either the use of singular value decomposition and plotting the resulting singular values or by running PARAFAC multiple times until the amount of variance explained by the model reaches a desired threshold [9,13,14].

The second requirement for successful PARAFAC analysis is that the data needs to be multilinear. Fraga et al. [28] defined multilinearity by the following conditions: the detector response is the same for each component in every injection and the retention times in both the ²D and ¹D remain constant for every injection. If the data are not exactly multilinear, constraints such as non-negativity, unimodality, and selectivity [29-31] can aid in the determination of accurate **A**, **B**, **C** and **D** matrices from PARAFAC analysis. The non-negativity constraint ensures that the profiles obtained by PARAFAC contain values that are only greater than zero for the constrained components. Unimodality, which is typically used only in the chromatographic dimension, ensures that only one maximum is present in the constrained resolved profiles. Selectivity restricts either the chromatographic or spectral dimensions by ensuring that the component can not be located within certain regions of the chromatogram or spectrum. This is accomplished by setting the intensity of the corresponding component to zero within those regions.

3. Experimental

All work was conducted on a Dell® Optiplex 755, Intel(R) Core™2 Duo CPU, E6550 @ 2.33 GHz, 3.23 GB of RAM within the confines of MATLAB® software 2009a (Mathworks, Inc., Natick, MA) version 7.8.0.347.

3.1. Data sets

Two different types of data, a simulated set and an experimental set, were examined to test the five different interpolation techniques. In order to ensure that the simulated data closely modeled an experimentally obtained data set, the simulated data set was created using experimentally comparable peak widths, retention time shifts, and background signals. The ¹D chromatogram, with a time scale of 0.35–2.8 min sampled at 0.35 min intervals and a peak width defined by a σ of 0.1335 min, was constructed using an in-house sampled Gaussian function created in MATLAB® to account for the ¹D sampling, previously described by Thekkudan, Rutan, and Carr [25]. The ¹D peak area column in Table 1 shows the areas used to generate the sampled ¹D peak for both the calibration and validation samples. A normal Gaussian distribution (with an area of one) was used for the ²D chromatogram, with a time scale of 0.025–3.375 s at 0.025 s intervals and a σ of 0.2125s. After the ²D and ¹D pure component profiles were generated individually, the 2D pure component chromatograms were generated by taking the outer product of the ¹D and ²D chromatographic vectors. These 2D chromatograms are then reshaped into a one dimensional structure (a vector), where the ²D chromatograms for each sample injection are appended sequentially, and the individual sample injections appended in turn. The spectral information is then incorporated by calculating the outer product of these chromatographic vectors with a pure component spectral vector (arbitrarily taken as the spectrum of 4-methylthioamphetamine [32]), measured at every fourth wavelength from 200 nm to 700 nm. The data structure is then reshaped into a four-way structure and then added onto a pre-existing, experimental background. This approach was conducted for thirty-five injections: fifteen calibration injections and twenty validation standard injections. The fifteen calibration injections consisted of a five point calibration curve with three replicates at each calibration level interspersed throughout the data set. Four validation samples were used, with five replicates per validation sample, to gauge the effectiveness of the five interpolation methods in aligning the data set.

The experimental data set was obtained from Dr. Carr's research group at the University of Minnesota and has been previously analyzed by Bailey and Rutan [10]. This data set consisted of six replicate injections containing seven well resolved standard compounds. A localized region around five of the standard peaks was selected according to two conditions. First, the boundaries of each localized region were selected so that only a single standard peak was present. Second, each localized region was large enough to ensure that the standard peak of interest was fully present in all six replicate injections. This data set was interpolated and aligned in the 1D and then aligned in the 2D . PARAFAC decomposition was then used to determine the relative areas of five of the seven peaks. The two peaks not interpolated were affected by adjacent contaminant peaks and as the peak phase shifted, the contaminants coeluted with these peaks.

3.2. Alignment approach

The approach for the interpolation and alignment of each peak can be seen in Fig. 2. At first, a representative 1D chromatogram (denoted by the dashed line in Fig. 2A) is chosen at a wavelength where the peak of interest strongly absorbs. The representative 1D chromatogram was chosen at 216 nm to ensure that a sufficient amount of signal for each peak was present in all injections, Fig. 2B. This representative 1D chromatogram is then interpolated to provide nine data points between each sampled point, and the maximum position of each peak present is determined. Once the interpolated position of each peak is determined, each 1D chromatogram across the 2D and spectral dimension is interpolated and then shifted to the earliest retention time for that peak, Fig. 2C. The 1D chromatogram is then resampled, by taking every ninth data point starting from the first data point, in order to reduce the number of data points. This helps to ensure that the computer does not run out of memory while implementing the PARAFAC analysis. The 2D chromatograms are then aligned by determining the maximum position of each peak in the 2D chromatogram at a given 1D point, illustrated by the dashed line in Fig. 2D, and then shifting each 2D chromatogram to the earliest retention time observed for that peak.

3.3. Interpolation implementation

The varying requirements of each interpolation technique necessitated different approaches in applying the techniques to the raw data set. The Hermite polynomial and spline techniques were implemented using the PCHIP and spline functions available in MATLAB® [33], respectively. The inputs for these functions were the 1D chromatogram being interpolated and the new time scale (0.35–2.8 min at 0.035 min intervals for the simulated data) to which the 1D chromatograms were to be interpolated. Both the PCHIP and spline functions utilize a cubic polynomial to fit to the data. The only difference between the manner in which two functions apply the cubic polynomial is how the second derivative of the polynomial is used. In the spline function, the second derivative of the polynomial is continuous, thereby allowing the resulting interpolated peak to possess maxima and minima between the original sampled data points. In contrast, the second derivative of the polynomial used in the PCHIP function is not continuous. This lack of continuity in the second derivative ensures that the original shape of the peak is preserved with only a minimal degree of curvature existing between data points, resulting in an interpolated peak that retains the original maxima and minima of the sampled 1D peak.

The Fourier zero-filling and Gaussian fitting techniques were implemented using in-house MATLAB® functions. Prior to performing Fourier zero-filling, a time axis corresponding to the sampled 1D chromatogram and a time axis corresponding to the interpolated 1D chromatogram were generated depending on the size (number of data points) of the 1D . In the case of the simulated data set, the size of the 1D was even, necessitating truncation of the data in accordance to the requirement that the frequency domain contain an odd number of

points [34]. This truncation resulted in the sampled time axis occurring from 0.35 to 2.45 min at 0.35 min intervals and the interpolated time axis occurring from 0.35 to 2.765 min at 0.035 min intervals. If the size of the 1D was odd, then the sampled time axis would have been from 0.35 to 2.8 min at 0.35 min intervals and the interpolated time axis would have been from 0.35 to 3.115 min at 0.035 min intervals. The different interpolation time axes are necessary to account for MATLAB[®] treating the 1D data as cyclical when performing the inverse Fourier transform. After the creation of the sampled time axis and the interpolated time axis, the Fourier transform was applied to the 1D chromatogram to convert from the time domain into the frequency domain. A number of zeros equal to the difference between the original number of data points and the desired number of data points were inserted just after the median point in the frequency domain. The inverse Fourier transform was then applied to the modified 1D frequency data to convert the 1D chromatogram back into the time domain, obtaining an interpolated 1D chromatogram. The original total intensity of the data is now distributed over a larger number of data points, resulting in the 1D chromatogram possessing reduced signal intensity compared to the original 1D chromatogram. To account for this change in signal intensity, the interpolated 1D chromatogram was multiplied by the ratio between the number of data points after interpolation and before interpolation. This resulted in the interpolated 1D chromatogram having the same maximum intensity as the original data.

Gaussian fitting was accomplished through the use of the non linear least-squares function, `lsqnonlin`, available in the optimization toolbox for MATLAB[®]. In order to perform the non linear least-squares fit of the 1D peak, initial estimates for the Gaussian parameters (area, position, σ , and background level) of the peak were determined. The initial estimate for the peak area was determined by Eq. (3):

$$\text{area} = a + b + c - 1.5(d + e) \quad (3)$$

where a and c are the values of the data points to either side of the peak maximum, b is the value of the peak maximum, and d and e are the value of background levels at either end of the 1D peak, as shown in Fig. 2B. The peak area was bounded to within $\pm 50\%$ of this initial guess. In order to determine the initial estimate for the peak position, a spline curve was applied to the sampled 1D peak and the resulting maximum was chosen as the peak position. A value of 0.21 min was used as the initial estimate for σ . Furthermore, an upper and lower bound of 0.28 min and 0.14 min was used to reduce the chances of overfitting the original 1D peak if an insufficient number of points (< 4) was available for fitting. The background level was calculated by averaging the first and last data points (points d and e in Fig. 2B) in the sampled 1D chromatogram. Using the determined regression parameters for the peak area, retention time and σ , a Gaussian peak was generated using the sampled time axis (0.35–2.8 min at 0.35 min intervals for the simulated data). In order to correctly scale the interpolated 1D peak for each 2D time point, linear least squares regression was used to fit the individual 1D chromatograms at each 2D time to obtain the correct amplitude. Using these amplitudes (areas), interpolated Gaussian peaks were generated using the interpolated time axis (0.35–2.8 min at 0.035 min intervals for the simulated data), and alignment was carried out as described previously.

3.4. PARAFAC analysis

After each interpolation technique was applied and alignment subsequently carried out, the data sets were analyzed using an in-house PARAFAC function. The PARAFAC function was designed using an alternating least squares algorithm (ALS) as described by Smilde et al. [27]; however, it was expanded to account for the fourth data dimension (sample). Non-negativity, unimodality, and selectivity were implemented independently for each dimension and component using the same approach previously implemented by Bezemer and Rutan

[35]. The PARAFAC analysis was conducted with a maximum of 2000 iterations and a convergence criterion of 1×10^{-10} . The peak components were constrained by applying spectral selectivity from 440 nm to 700 nm, unimodality in both the 2D and 1D , and non-negativity was applied to all dimensions. The background components were constrained by applying non-negativity to the sample dimension, the 2D , and the 1D .

4. Results and discussion

4.1. Retention time prediction

The predicted retention times for the simulated data set were calculated for each of the four proposed interpolation techniques over the course of all of the injections. A plot of the predicted retention times for the four interpolation techniques of the simulated data versus the actual retention times used to generate the peaks can be seen in Fig. 3. The solid line found in Fig. 3 represents the actual retention times of the simulated peak. The squares ($\square$) show the predicted retention times obtained from the PCHIP interpolation procedure for the thirty-five injections. The PCHIP predicted retention times were found to be 1.4 min for the first sixteen injections and 1.75 min for the remaining nineteen injections. This abrupt shift from one retention time to another is due to the nature of the PCHIP interpolation. PCHIP interpolation is almost a linear interpolation between existing data points. A small amount of curvature is allowed to exist between data points as a result of the second derivative being non-continuous. Due to the shape preserving nature of the PCHIP technique, the peak maxima observed in the sampled 1D peaks remains the maxima after interpolation. The abrupt change in predicted retention time is due to a change in the 1D peak sampling phase. The sampling phase of the simulated 1D peak is 2.73 at injection fifteen, 3.07 at injection sixteen, and 2.77 at injection seventeen. Since the sampled 1D sampling phase passes through π between injections sixteen and seventeen, the predicted retention time abruptly shifts when the signal in the 1D at 1.75 min becomes larger than the signal in the 1D at 1.4 min. The linear interpolation used in the literature also produces the same predicted retention times as the PCHIP interpolation technique.

The triangles (Δ) in Fig. 3 show the predicted retention times obtained from the spline interpolation procedure for the thirty-five injections. Unlike the PCHIP interpolated peak, the predicted retention time of the spline interpolated 1D peak changes over the course of the injections. This change is due to the difference in how the spline is implemented versus PCHIP. Both techniques use a cubic polynomial. However, where the PCHIP technique only requires the first derivative to be continuous, the spline interpolation requires that both the first and second derivatives are continuous. This results in the maximum of the interpolated 1D peak shifting from the existing data points to points in between. The spline interpolated peak begins to take on the more rounded shape of the simulated peak. However, while this is an improvement over the PCHIP interpolation, the predicted retention times do not closely correspond to the actual retention times of the simulated peak.

The $\times$'s in Fig. 3 show the predicted retention times obtained from the Fourier zero-filling interpolated 1D peak. The predicted retention times were closer to the actual retention times compared to the Hermite and spline interpolation techniques. However, the predicted retention times from the Fourier zero-filling oscillated between either being too high or too low.

The circles ($\circ$) in Fig. 3 show the predicted retention times obtained from the Gaussian fitting interpolated 1D peak. Unlike the previous three interpolation techniques, the Gaussian fitting interpolation technique produces predicted retention times that are nearly identical to the actual retention times used to generate the simulated peak. Since the original sampled 1D peak was created using a Gaussian curve, the fact that the Gaussian fitting interpolation

technique produces the most reliable calculated retention times makes sense. However, the Gaussian fitting interpolation only produces an accurate depiction of the unsampled 1D peak if the fitting is properly conducted, i.e. all parameters are properly constrained. If the bounded range for the σ value, from Section 3.3, is not relatively close to the actual sigma, then the resulting Gaussian fitted interpolated 1D peak will be either too narrow or too wide. While the width of the peak does not have an impact on the predicted retention times, an erroneous σ value will significantly affect the accuracy of the results (i.e., relative peak areas) obtained from the PARAFAC analysis.

4.2. Calibration curves

In order to further measure the effectiveness of each interpolation technique, the results from the PARAFAC analysis were used to generate a calibration curve for the simulated calibration points. The Excel linest function was used to calculate the slopes and intercepts with the corresponding errors, shown in Table 2, for each of the five interpolation techniques. The relative standard deviations of the slopes were used to provide an initial estimate of how well the PARAFAC relative areas fit to the calculated trend line, shown in Table 2. The relative slope errors were calculated by dividing the error by the slope and multiplying by one hundred. The relative slope errors for each of the interpolation techniques and the unaligned calibration curve were calculated to be as follows: 17.9% for the unaligned calibration curve, 11.5% for the linear with cross correlation of the 1D chromatogram calibration curve, 4.7% for the PCHIP calibration curve, 2.9% for the spline calibration curve, 3.7% for the Fourier calibration curve and 1.9% for the Gaussian calibration curve. From the relative slope errors given in Table 2, the unaligned data set, as expected, possessed the most error in the calibration points. This is easily explained by the lack of multilinearity in the unaligned data set. The relative slope error for the linear interpolation followed by cross correlation of the 1D chromatogram was better than the unaligned relative slope error. However, in comparison to the relative slope errors of the other interpolation techniques, the relative slope error of the linear interpolation followed by cross correlation of the 1D chromatogram was much higher. This large difference in the standard errors may be due to the manner in which the cross correlation alignment approach works. Unlike the four proposed interpolation techniques, the cross correlation alignment approach may not necessarily align a peak to the same point. In addition, the shape of the interpolated peak may not be consistent across all of the injections. Both of these reasons could result in the aligned data set not being completely multilinear. The PCHIP and Fourier zero-filling interpolation techniques produced the next set of comparable relative slope errors. The PCHIP relative slope error is due to the shape preserving nature, as discussed previously, of the technique. Likewise, the Fourier zero-filling relative slope error may be explained by a similar shape preserving ability. Finally, the spline and Gaussian fitting interpolation techniques produced the lowest relative slope errors. The most probable reason for the small relative slope errors is their ability to impose a uniform shape on the interpolated 1D peak. This imposition of a uniform shape aids in forcing multilinearity for the simulated data set.

In addition to calculating the relative slope error, the calibration curves were used to determine the calculated concentration of the validation samples. The calculated validation sample concentrations were used to determine the percent standard error of prediction (%SEP), a measure of the accuracy of the approach, and the average percent relative standard deviation (%RSD), a measure of precision. Table 3 shows the calculated %SEP and %RSD values for the unaligned data and for the different interpolation techniques. As expected when the data was not aligned prior to analysis by PARAFAC, the resulting relative peak areas were the most inaccurate and imprecise. Of the five interpolation techniques examined, the linear followed by cross correlation of the 1D chromatogram was

the most inaccurate but the imprecision of the technique was on par with most of the other interpolation techniques, except for the Gaussian fitting. Of the remaining four interpolation techniques, the Gaussian fitting calculated concentrations were the most accurate and precise. This observation is not surprising, as Gaussian peak shapes were used to generate the simulated data.

4.3. Experimental data

From the results of the simulated data analysis, the Gaussian fitting interpolation technique was expected to be the most reliable technique for aligning the experimental sampled ^1D peaks. Each of the five peaks in the experimental data set were interpolated, aligned, and analyzed using PARAFAC. Since a calibration curve was not available for the experimental data, only the precision measurement, %RSD, is shown in Table 4. Since this data set has been previously analyzed by Bailey and Rutan [10], a column titled MCR-ALS, where the results from this MCR-ALS analysis are reported, was included for comparison purposes. None of the four interpolation techniques were able to match the degree of reproducibility obtained by Bailey and Rutan [10]. This MCR-ALS method, which requires manual integration, while somewhat tedious and less automated than PARAFAC, does not require that the data have a multilinear structure.

In addition to visually comparing the %RSDs for the five techniques versus the unaligned %RSD, Levene's test was used to determine if there was a statistical difference between the different normalized relative peak areas [36]. Levene's test is a robust method that works under the assumption that the underlying data possesses a normal distribution. Since the number of data points for each technique is small, the Brown and Forsythe test could not be used to determine if the data deviates from a normal distribution [36]. The critical F -value for Levene's test is 2.53 obtained from an α of 0.05 with 5 and 30 degrees of freedom respectively. Levene's test revealed that the normalized relative peak areas for Peaks 1, 2, and 3 possessed statistically significant differences. The reason for the statistical significant difference for normalized relative peak areas for Peaks 1 and 2 is due to the large unaligned standard deviation, seen in Fig. 4A and 4B. The standard deviations for the unaligned normalized relative peak areas are a clear outlier in comparison to the standard deviations from the other techniques. The unaligned standard deviation was much larger due to the degree of peak shifting for Peaks 1 and 2 relative to the other peaks. In both cases, the normalized relative peak areas for injections three and four were higher due to the presence of the peak existing in the center of the data window for these injections. The reason for the statistical significant difference for the normalized relative peaks areas for Peak 3 is not as easily explained, Fig. 4C. In comparison to the retention time shifting of Peaks 1 and 2, Peak 3 experiences less retention time shifting over the course of the experimental run. An examination of the normalized relative peak areas reveals that for the PCHIP, Fourier zero-filling, and Gaussian fitting techniques, the normalized relative peak area for injection six of the experimental data set is less than the other techniques. An identifiable cause for this decreased value has not been determined.

Even though the Gaussian fitting was found to produce the most accurate and precise calculated concentrations for the simulated data set, the Gaussian fitting for the experimental peaks did not produce better results than the other techniques. One possible reason for this deviation from the simulated data is the inability of the non-linear least squares fitting to correctly account for the background signal within the sampled ^1D peak. The background used

in the simulated data set was not the same background present in the experimental data set. As a result, the resulting fitting parameters calculated for the experimental data set were not as consistent as the fitting parameters calculated for the simulated data set. This deviation in

the fitting parameters results in the interpolated Gaussian peak showing a greater degree of inconsistency, in comparison to the simulated interpolated Gaussian peak, across each of the replicate injections. A second consideration is that the actual experimental data may not follow the Gaussian model exactly, such that for the Gaussian method specifically may not give improved results for the experimental data.

5. Conclusions

The five interpolation techniques were successfully applied to LC $\times$ LC–DAD data during the alignment process. From the simulated data set, the Gaussian interpolation technique was found to produce the best %SEP and %RSD in comparison to the other interpolation techniques. This was expected due to the Gaussian nature of the simulated chromatographic peaks and was expected to carry over into the experimental data set. However, when the techniques were applied to the experimental data, the Gaussian fitting was found to not be statistically better or worse than the other interpolation techniques. The results from MCR-ALS reported previously [10] still provide the best precision for the experimental data set. This approach has the advantage that it is not affected by retention time shifts in the 1D and 2D separations, but is more tedious and less automated than PARAFAC, especially with respect to the final quantification step which requires manual assignment of the chromatogram baseline. To reduce the potential impact of the background on the Gaussian fitting interpolation technique, an approach consisting of performing PARAFAC on each injection individually is being developed. This method will enable alignment of more complex chromatograms; because the peaks being aligned will be constrained to have the same spectral features. The new approach will isolate the sampled 1D peak signature from the background and hopefully reduce the variations in the Gaussian fitting parameters between injections.

Acknowledgments

The authors of this paper would like to thank Dr. Dwight Stoll at Gustavus Adolphus College and Dr. Peter Carr at the University of Minnesota for the use of their data in the development of this approach. The authors also would like thank Dr. Jason Merrick at Virginia Commonwealth University for consultation regarding the appropriate implementation of Levene's test. The authors acknowledge financial support from NIH-GM-54585-13 and NSF CHE-0911330.

References

1. Marriott P, Shellie R. *Anal Chem.* 2002; 21:573.
2. Fraga CG, Corley CA. *J Chromatogr A.* 2005; 1096:40. [PubMed: 16301068]
3. Stoll DR, Carr PW. *J Am Chem Soc.* 2005; 127:5034. [PubMed: 15810834]
4. Pierce KM, Wood LF, Wright BW, Synovec RE. *Anal Chem.* 2005; 77:7735. [PubMed: 16316183]
5. Stoll DR, Cohen JD, Carr PW. *J Chromatogr A.* 2006; 1122:123. [PubMed: 16720027]
6. Francois I, Sandra K, Sandra P. *Anal Chim Acta.* 2009; 641:14. [PubMed: 19393362]
7. Opiteck GJ, Jorgenson JW. *Anal Chem.* 1997; 69:2283. [PubMed: 9212702]
8. Sinha AE, Hope JL, Prazen BJ, Fraga CG, Nilsson EJ, Synovec RE. *J Chromatogr A.* 2004; 1056:145. [PubMed: 15595544]
9. Porter SEG, Stoll DR, Rutan SC, Carr PW, Cohen JD. *Anal Chem.* 2006; 78:5559. [PubMed: 16878896]
10. Bailey HP, Rutan SC. *Chemom Intell Lab Syst.* 2011; 106:131.
11. Bruckner CA, Prazen BJ, Synovec RE. *Anal Chem.* 1998; 70:2796.
12. van Mispelaar VG, Tas AC, Smilde AK, Schoenmakers PJ, van Asten AC. *J Chromatogr A.* 2003; 1019:15. [PubMed: 14650601]
13. Hoggard JC, Synovec RE. *Anal Chem.* 2007; 79:1611. [PubMed: 17297963]

14. Hoggard JC, Synovec RE. *Anal Chem.* 2008; 80:6677. [PubMed: 18672889]
15. Fraga GG, Prazen BJ, Synovec RE. *Anal Chem.* 2001; 73:5833. [PubMed: 11791551]
16. Prazen BJ, Synovec RE, Kowalski BR. *Anal Chem.* 1998; 1998:218.
17. Pierce KM, Hope JL, Johnson KJ, Wright BW, Synovec RE. *J Chromatogr A.* 2005; 1096:101. [PubMed: 16301073]
18. Christin C, Smilde AK, Hoefsloot HCJ, Suits F, Bischoff R, Horvatovich PL. *Anal Chem.* 2008; 80:7012. [PubMed: 18715018]
19. Ong RCY, Marriott PJ. *J Chromatogr Sci.* 2002; 40:276. [PubMed: 12049157]
20. Xie L, Marriott PJ, Adams M. *Anal Chim Acta.* 2003; 500:211.
21. Fairchild JN, Horvath K, Guiochon G. *J Chromatogr A.* 2009; 1216:1363. [PubMed: 19150720]
22. Blumberg LM. *J Sep Sci.* 2008; 31:3358. [PubMed: 18798219]
23. Murphy RE, Schure MR, Foley JP. *Anal Chem.* 1998; 70:1585.
24. Seeley JV. *J Chromatogr A.* 2002; 962:21. [PubMed: 12198965]
25. Thekkudan D, Rutan SC. *J Chromatogr A.* 2010; 1217:4313. [PubMed: 20462587]
26. Bro R. *Chemom Intell Lab Syst.* 1997; 38:149.
27. Smilde AK, Bro R, Geladi P. *Multi-way Analysis with Applications in the Chemical Sciences.* John Wiley & Sons Ltd, West Sussex, England. 2004
28. Fraga CG, Prazen BJ, Synovec RE. *J High Resolut Chromatogr.* 2000; 23:215.
29. de Juan A, Heyden YV, Tauler R, Massart DL. *Anal Chim Acta.* 1997; 346:307.
30. Bro R, De Jong S. *J Chemom.* 1997; 11:393.
31. Bro R, Sidiropoulos ND. *J Chemom.* 1998; 12:223.
32. Pragst F, Herzler M, Herre S, Erxleben BT, Rothe M. *UV Spectra of Toxic Compounds.* Verlag Dr Dieter Helm, Heppenheim. 2001:627.
33. Matlab: The Language of Technical Computing. The MathWorks Inc; 2009.
34. Bracewell, RN. *The Fourier Transform and its Applications.* McGraw-Hill; New York: 1979.
35. Bezemer E, Rutan SC. *Chemom Intell Lab Syst.* 2006; 81:82.
36. Brown MB, Forsythe AB. *J Am Stat Assoc.* 1974; 69:364.

Fig. 1.

(A) An illustration of an LC $\times$ LC chromatogram. The dashed lines indicate the time points where the sampling device injects the ${}^1\text{D}$ eluent into the ${}^2\text{D}$ column. (B) The ${}^1\text{D}$ peak pattern from (A) generated by summing all values along the dashed lines. The alternating ▼ and ● markers represent either the use of two loops or two ${}^2\text{D}$ columns in LC $\times$ LC. (C) Illustration of the parameters that determine the sampling phase, ϕ .

Fig. 2.

(A) A contour plot of a representative LC $\times$ LC chromatogram at 216 nm. The dashed line indicates the 2^{D} time point chosen to generate the representative 1^{D} chromatogram for interpolation purposes. (B) The representative 1^{D} chromatogram generated in 1A. The solid and dashed lines represent two different injections of the same peak at two different retention times. The letters *a-e* refer to the points indicated in Eq. (3) in the text used to estimate the peak area. (C) 1^{D} chromatograms after the cubic spline interpolation technique has been applied to the sampled 1^{D} chromatograms from 1B. The vertical dashed line indicates the point to which the dashed peak is being aligned. (D) The resulting resampled 1^{D} peak after alignment and resampling of the interpolated 1^{D} chromatogram due to computer memory issues. The vertical dashed line indicates the 1^{D} time point chosen to determine the 2^{D} time points for the 2^{D} alignment.

Fig. 3.

A plot of the retention times after reconstruction of the 1D for the simulated data set versus the actual retention times used for the generation of the simulated data. The solid line is the actual retention times plotted against the actual retention times. The four interpolation techniques, PCHIP ($\square$), spline polynomial (Δ), Fourier zero-filling ($\times$), and Gaussian fitting ($\circ$), are compared against this line.

Fig. 4. Plots of the standard deviations for the normalized relative peak areas obtained from PARAFAC for (A) Peak 1, (B) Peak 2, and (C) Peak 3. The solid black line is the average standard deviation for all techniques for each peak.

Table 1

Peak areas used to generate the sampled ^1D peak for the simulated data.

^1D peak area	
Calibration level ^a	
1	26.00
2	34.00
3	42.00
4	10.00
5	18.00
Validation level ^b	
1	39.25
2	12.98
3	14.57
4	26.92

^a Each calibration level consisted of three replicates.

^b Each validation level consisted of five replicates

Table 2

Calculated slope and intercept (with associated errors) for the simulated data set from the relative areas obtained from PARAFAC.

Interpolation method	Slope	Intercept	Standard error of the regression (s_y)	Relative standard deviation of slope (%)
Unaligned	1.2 (0.2)	-1 (5)	9.2	17.9
Linear	1.0 (0.1)	-3 (3)	4.9	11.5
PCHIP	1.21 (0.06)	0 (2)	2.5	4.8
Spline	1.39 (0.04)	0 (1)	1.8	2.9
Fourier	1.70 (0.06)	2 (2)	2.8	3.5
Gaussian	1.73 (0.03)	0 (1)	1.4	1.7

Table 3

Average percent standard error of prediction and percent relative standard deviation for the validation samples within the simulated data.

Interpolation method	%SEP	Average %RSD
Unaligned	23.8	4.0
Linear	14.2	2.4
PCHIP	9.3	2.5
Spline	8.0	2.2
Fourier	8.7	2.5
Gaussian	4.1	1.1

Table 4

Percent relative standard deviations of the relative peak areas for the five experimentally generated peaks.

Peak	Unaligned	MCR-ALS [10]	Linear	PCHIP	Spline	Fourier	Gaussian
1	13.0	1.6	4.29	6.27	5.24	5.05	4.39
2	16.8	2.2	6.76	4.33	3.77	4.60	5.84
3	7.45	4.7	13.9	19.6	8.41	17.0	16.8
4	11.4	3.5	12.8	12.9	12.1	11.2	11.8
5	17.0	1.4	15.9	11.8	13.6	10.1	10.4
Average	13.1	2.7	10.7	11.0	8.61	9.58	9.83