ELSEVIER

Chemometrics and Intelligent Laboratory Systems 49 (1999) 195–213

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Weighted parallel factor analysis for calibration of HPLC-UV/Vis spectrometers in the presence of Beer's law deviations

Greger G. Andersson *, Brian K. Dable, Karl S. Booksh

Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287, USA

Received 1 April 1999; received in revised form 8 June 1999; accepted 15 June 1999

Abstract

An extension of the parallel factor analysis (PARAFAC) methodology is presented to allow accurate and reliable quantitative and qualitative analysis of nonlinear data collected from hyphenated instrumentation. The weighted PARAFAC method is applied to high-performance liquid chromatography-ultraviolet/visible (HPLC-UV/Vis) diode array spectrometry analysis. It is demonstrated that this method improves the quantitative errors when spectroscopic nonlinearities from solvent–solute interactions or detector saturation are introduced. As much as 50% improvements in the root mean squared errors of estimation are realized for test samples. This weighted PARAFAC algorithm implicitly treats nonlinear data as missing values. A method requiring no a priori information is presented, that facilitates determination of the nonlinear regions and optimal application of the weighted PARAFAC algorithm. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Factor analysis; Nonlinear; N-way calibration; Variable selection

1. Introduction

Parallel factor analysis (PARAFAC) [1–3] based calibration has been demonstrated to be a powerful method for extracting qualitative and quantitative information from collections of multiway data such as that from high-performance liquid chromatography-ultraviolet/visible (HPLC-UV/Vis) diode array spectrometers and excitation-emission matrix (EEM) fluorescence spectrometers [4–12]. Theory states that, when applied appropriately, PARAFAC exploits the structure of such data to uniquely resolve the underlying instrumental profiles and the relative concentrations of each component in the system [13]. Thus, by employing only a small number of standards, an analyte profile can be accurately extracted from a mixture containing unknown, uncalibrated interferents. This is known as the 'second order advantage' [14].

PARAFAC assumes that a trilinear model accurately describes all sources of variance in the three-way collection of data. With this model comes the implicit assumption that, for a pure component, any sample analyzed will yield a matrix of rank equal to 1. Furthermore, the observed signal of each analyte is independent of concentration and the presence of any interferents. That is to say, the final observed

* Corresponding author. Scotia Lipid Teknik, P.O. Box 6686, S-113 84, Stockholm, Sweden

0169-7439/99/$ - see front matter
PII: S0169-7439(99)00040-4

signal of any mixture is the linear combination of signals from all constituents. The theory and implementation of PARAFAC and other linear three-way methods are further covered in the tutorial by Bro [15].

It should be noted that problems do occur when directly applying the trilinear PARAFAC model/algorithm to data that deviates from the ideal trilinear model. Booksh and Kowalski [16] have catalogued five broad classes of deviations from the trilinear model that are common with chemical data. In some of these cases, methods like the extended trilinear decomposition [17], constrained Tucker models [18,19], and second order standard addition method [20] may be employed to preserve the second order advantage. However, when the observed deviations from the trilinear model result from a nonlinear detector response, no solution for accurate modeling of, and calibration with, this type of data was proposed.

This paper presents a method for maintaining the second order advantage when analyzing three-way data of the type collected with a nonlinear detector (i.e., one that deviates from Beer's law). A weighted PARAFAC algorithm is employed to eliminate the observations with the largest deviation from the trilinear model. The method is tested with Monte Carlo simulations and applied HPLC-UV/VisDAS data. It is demonstrated that this weighted PARAFAC method yields accurate and precise estimates of resolved analyte chromatograms, spectra, and concentrations even when a significant fraction of the observed data for a given sample is in the nonlinear range of the detector. It is further demonstrated that the weighted PARAFAC method is also capable of accurate quantization when other concentration dependant nonlinearities, such as solvent–solute interactions, are present in UV/Vis data.

2. Theory

PARAFAC employs the trilinear model

$$R_{ijk} = \sum_{n=1}^{N} \hat{X}_{in}\hat{Y}_{jn}\hat{Z}_{kn} + E_{ijk}. \quad (1)$$

Where the kth slice of R is the $I \times J$ matrix of data collected from the instrumental analysis of the kth sample. In the case of HPLC-UV/VisDAS, each of these matrices contain I spectra of J discretely digitized wavelengths. Thus, the N columns in the matrices $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$, and $\hat{\mathbf{Z}}$ are the PARAFAC estimates of the chromatograms, spectra, and relative concentrations of the N species that coexist in R. Only the number of factors employed in the model, N, is supplied by the analyst. E is the collection of model and random errors residual from fitting this trilinear model to R.

The parameters of Eq. (1) are found by an alternating least squares procedure where the PARAFAC algorithm begins with an initial guess of the X-way and Y-way starting profiles. The initial Z-way profiles are determined by solving

$$\mathbf{R}_{\mathrm{C}} = \mathbf{C}\mathbf{Z}^{T} \quad (2a)$$

such that $\hat{\mathbf{Z}} = \mathbf{C}^{+}\mathbf{R}_{\mathrm{C}}$ with $\mathbf{C}^{+}$ being the generalized inverse of $\mathbf{C}$ that can be calculated from the normal equations or singular value decomposition of $\mathbf{C}$. In Eq. (2a), $\mathbf{R}_{\mathrm{C}}$ is a $I * J \times K$ matrix constructed by unfolding the K slices of R in the IJ plane where $R_{C_{(j-1)I+i,k}} = R_{i,j,k}$. Similarly, $\mathbf{C}$ is a $I * J \times N$ matrix formed from the N columns of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Z}}$ where $C_{(j-i)l+i,n} = X_{i,n}Y_{j,n}$.

Updated estimates of the X-way and Y-way profiles are found by solving

$$\mathbf{R}_{\mathrm{A}} = \mathbf{A}\mathbf{X}^{T} \quad (2b)$$

such that $\hat{\mathbf{X}} = \mathbf{A}^{+}\mathbf{R}_{\mathrm{A}}$, and

$$\mathbf{R}_{\mathrm{B}} = \mathbf{B}\mathbf{Y}^{T} \quad (2c)$$

such that $\hat{\mathbf{Y}} = \mathbf{B}^{+}\mathbf{R}_{\mathrm{B}}$ where $\mathbf{R}_{\mathrm{A}}$ and $\mathbf{R}_{\mathrm{B}}$ are constructed analogously to $\mathbf{R}_{\mathrm{C}}$ by unfolding $\mathbf{R}$ in the YZ and XZ planes, respectively. This forms a $J * Z \times I$ matrix for $\mathbf{R}_{\mathrm{A}}$ and a $X * Z \times J$ matrix for $\mathbf{R}_{\mathrm{B}}$. Similarly to $\mathbf{C}$, $A_{(k-1)j+k,n} = Y_{j,n}Z_{k,n}$ and $B_{(k-1)l+k,n} = X_{i,n}Z_{k,n}$. The algorithm proceeds iteratively, cycling through Eqs. (2a), (2b) and (2c) until the convergence criterion is satisfied. At each step, the most recent estimates of $\mathbf{X}$ and $\mathbf{Y}$ are used to determine $\hat{\mathbf{Z}}$ (or $\mathbf{Y}$ and $\mathbf{Z}$ to determine $\hat{\mathbf{X}}$, or $\mathbf{X}$ and $\mathbf{Z}$ to determine $\hat{\mathbf{Y}}$, depending on the equation currently being solved). Thus, the squared residual error penalty function, $\sum_{i,jk}((R_{ijk} - \sum_{n=1}^{N}\hat{X}_{in}\hat{Y}_{jn}\hat{Z}_{kn})^2)$ is minimized by minimizing $\|\mathbf{R}_{\mathrm{A}} - \mathbf{A}\hat{\mathbf{X}}^{T}\|_F^2$,, $\|\mathbf{R}_{\mathrm{B}} - \mathbf{B}\hat{\mathbf{Y}}^{\mathrm{T}}\|_F^2$, and $\|\mathbf{R}_{\mathrm{C}} - \mathbf{C}\hat{\mathbf{Z}}^{T}\|_F^2$ at each appropriate step in the iterative cycle. Note that $\|\mathbf{Q}\|_F^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}Q_{i,j}^2$ and is the squared Frobenius, or Euclidian, norm of $\mathbf{Q}$.

However, any optimization penalty function may be employed to determine $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$, and $\hat{\mathbf{Z}}$. In this work, it is realized that some elements of $\mathbf{5}$ contain a high degree of deviation from the underlying trilinear model and are best left unmodeled. These elements, once determined, are assigned a weight of zero relative to the other elements. The weight matrix, W_{ijk} is thus constructed by determining the observed value for the onset of nonlinearities. If R_{ijk} is less than the cut-off value, W_{ijk} is assigned a weight of unity; if R_{ijk} is greater than or equal to the cut-off value, W_{ijk} is assigned a weight of one. This yields an overall penalty function of $\sum_{i,j,k}(W_{ijk}(R_{ijk}\sum_{n=1}^{N}\hat{X}_{in}\hat{Y}_{jn}\hat{Z}_{kn})^2)$ which is minimized by minimizing $\|\mathbf{W}_A \bigcirc (\mathbf{R}_A - \mathbf{A}\hat{\mathbf{X}}^T)\|_F^2$, $\|\mathbf{W}_B \bigcirc (\mathbf{R}_B - \mathbf{B}\hat{\mathbf{Y}}^T)\|_F^2$, and $\|\mathbf{W}_C \bigcirc (\mathbf{R}_C - \mathbf{C}\hat{\mathbf{Z}}^T)\|_F^2$ at each appropriate step in the iterative cycle where $\bigcirc$ is the element wise, or Hadamard, product. Here, $\mathbf{W}_A$, $\mathbf{W}_B$, and $\mathbf{W}_C$ are constructed by unfolding: equivalently to unfolding $\mathbb{R}$ to construct $\mathbf{R}_A$, $\mathbf{R}_B$, and $\mathbf{R}_C$.

3. Experimental

3.1. Programs and analysis

The PARAFAC algorithm and all modifications were written in-house in the Matlab 5.2 (MathWorks, Natick, MA) environment. Matlab was launched on a 200 MHz Pentium (Intel) with 96MB RAM under the Windows95 operating system (Microsoft, Redmond, WA).

3.2. Simulations

Simulated HPLC-UV/VisDAS instrumental profiles of 50 spectra, each with 100 digitized wavelengths, were constructed for three components. Correlation coefficients in the HPLC-way ranged from 0.73 to 0.78 and correlation coefficients in the UV/Vis-way ranged from 0.49 to 0.88. A 3^3 experimental design was employed to create 27 simulated instrumental responses scaled to have a maximum intensity of 3000 mAU. When appropriate, normally distributed random errors with a mean of zero and standard deviation of 50 mAU were added to the 27 simulated instrumental responses.

Two types of nonlinearities were induced in the data. The first simulated a nonlinear detector response. The data remained linear up to 1500 mAU. Above 1500 mAU the instrumental response deviates significantly from Beer's law (Fig. 1, dashed line). The second nonlinearity studied simulates 5% stray light in a UV/VisDAS system [21]. This is a continuously nonlinear function where no true linear region exists (Fig. 1, dotted line).

3.3. HPLC-UV/VisDAS data

HPLC data were collected on a Shimadzu LC-10AS liquid chromatograph with a Shimadzu SPD-M10AVP diode array detector (Shimadzu, Kyoto, Japan). Separations were performed on a 15 cm × 4.6 cm Supelcosil C18 column (Supelco, Bellefonte, PA). All data were converted to ASCII flat files and imported into Matlab via a Matlab program written in-house.

Two sets of HPLC-UV/VisDAS were collected. The first set contained 33 standards and mixtures of naphthalene (Fisher, Pittsburgh, PA), toluene (Fisher), and *p*-xylene (Sigma-Aldrich, St. Louis, MO). Maximum standard sample concentrations were 0.328 ppt, 6.0 ppt, and 8.0 ppt, respectively. Nine ternary and three binary mixture samples were prepared with concentrations of 1, 3/4, 1/2 and 1/4 of the standards as appropriate. The entire experimental design was repeated with a two-fold dilution of all samples. An additional standard of 20% of the maximum concentration for each component was prepared. Each sample was run in duplicate to yield 66 total HPLC-UV/Vis spectra. The flow rate was set to 2.5 ml/min with a 95% methanol (Mallinckrodt, Phillipsburg, NJ)/5% distilled water isocratic mobile phase. Spectra were collected at 1 nm intervals from 240 nm to 300 nm. A sampling frequency of 1 UV/Vis spectrum per 0.24 s was employed. Spectra were collected from 48 s to 72 s following injection. This resulted in a $100 \times 60 \times 66$ data cube from the experiment. Spectra and chromatograms from the standards are shown in Fig. 5a and Fig. 7a, respectively.

The second set contained 34 standards and mixtures of naphthalene, toluene, *p*-xylene, and ethylbenzene (Sigma-Aldrich). Maximum standard sample concentrations were 0.328 ppt, 3.2 ppt, 4.0 ppt, and 3.2 ppt, respectively. Seven quaternary and four ternary mixture samples were prepared with concen-

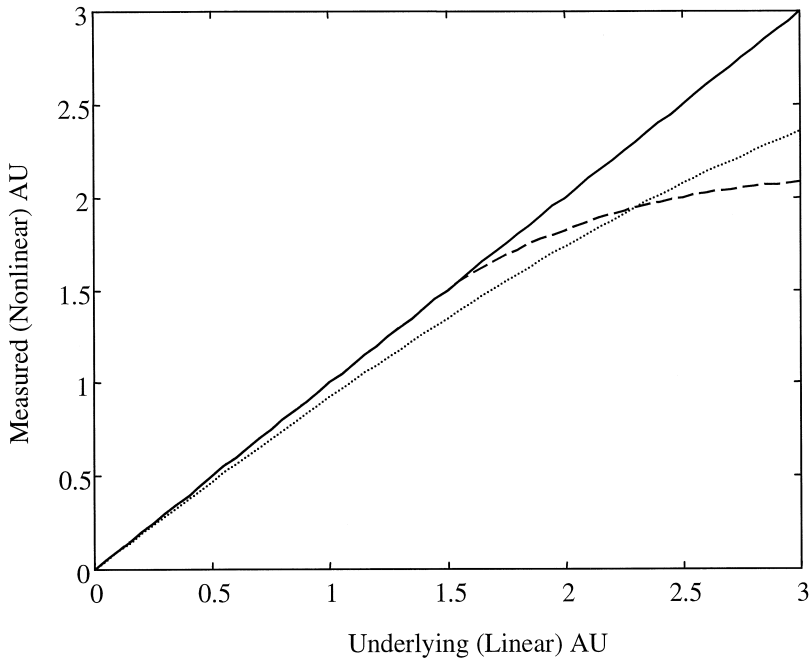

Fig. 1. Observed vs. latent response for the two types of nonlinearities investigated in the simulation. The piece-wise nonlinear function (dashed line) is given by $R - 0.5*(R - 1.5)^{1.5}$ for $R > 1.5$. The continuously nonlinear function (dotted line) simulates 5% stray light, $\log_{10}(1.05/(e^{-R} + 0.05))$.

trations of 1, 3/4, 1/2 and 1/4 of the standards as appropriate. The entire experimental design was repeated with a two-fold dilution of all samples. An additional standard of 20% of the maximum standard concentration for each component was prepared. Each sample was analyzed in duplicate to yield 68 total HPLC-UV/Vis spectra. The flow rate was set to 2.0 ml/min for a 90% methanol/10% distilled water isocratic mobile phase. Spectra were collected at 1 nm intervals from 240 nm to 300 nm. A sampling frequency of 1 UV/Vis spectrum per 0.63 s was employed as the data were collected from 72 s to 105 s following injection. This resulted in a $50 \times 60 \times 68$ data cube from the experiment. Spectra and chromatograms from the standards are shown in Fig. 5b and Fig. 7b, respectively.

4. Results and discussion

4.1. Choice of PARAFAC models

The traditional trilinear PARAFAC model was applied in analysis of all simulations and data sets. A three-factor model was employed for the simulations and first data set. A four-factor model was employed for the second data set. These models were deemed optimal based on knowledge of the number of real factors employed to construct the data/samples and visual inspection of the estimated profiles. The estimated profiles were compared, as appropriate, to the profiles employed for construction of the data set or to the profiles of pure compounds analyzed on the HPLC-UV/VisDAS. Also, estimated profiles from PARAFAC models employing one fewer and one additional factor were compared to the instrumental profiles and known concentrations of the standards to further insure that the model employed was the most accurate.

4.2. Simulations

The first computer test set simulated the HPLC-UV/VisDAS response with a detector nonlinearity above 1.5 absorbance unit (AU). The instrumental response is linear below 1.5 AU and contains a sharp nonlinearity above 1.5 AU. Fig. 2a presents the sim-

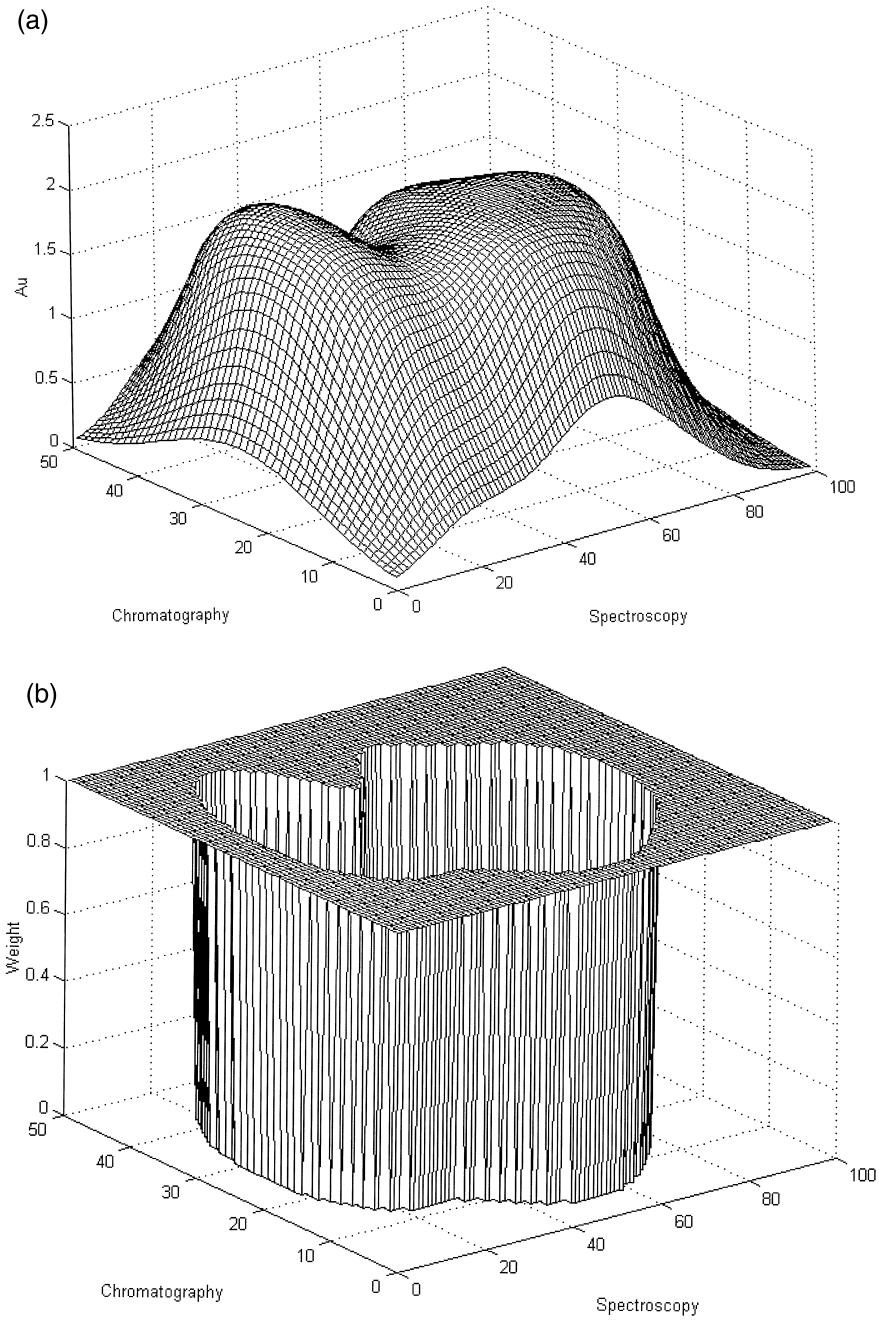

Fig. 2. (a) Simulated nonlinear HPLC-UV/VisDAS instrument response with (b) associated weight matrix at 1.5 AU cut-off.

ulated instrumental response of the most intense sample as a 3-D mesh plot. Note that the broad peak to the rear of the figure flattens as the peak's intensity increases above 1.5 AU. The weight matrix ap-

plied to this sample, for a 1.5 AU cut-off, is shown in Fig. 2b. If a greater intensity was used as cut-off, a greater percentage of the sample would receive non-zero weights. Simultaneously, if a lesser cut-off value were adapted, a greater proportion of the sample would be assigned weights of zero. Consequently, fewer of the measurements would be employed to determine the underlying instrumental and concentration profiles in the data.

Fortunately, in many cases, the data collected for PARAFAC analysis is greatly mathematically overdetermined. Therefore, accurate estimation of the model parameters can be achieved with only a small fraction of the data. The ability to accurately deconvolve the spectral profiles with weighted PARAFAC is evident in Fig. 3 (open symbols). When no random errors are added to the data, the root mean squared error (RMSE) of estimating the concentrations of the three species in all 27 samples markedly increases when a cut-off above 1.5 AU is employed. Here, only 58% of the individual measurements, 34% of the integrated signal intensity, are available for parameter (concentration) estimation. The RMSE of all three components in the most intense sample is 3.8×10^{-4}; consistent with the RMSE of the other 26 samples. Also, there is little loss in accuracy when a cut-off as low as 0.4 AU is employed. At this cut-off, only 17% of the individual measurements, or 2% of the integrated intensity, is employed for parameter estimation. A 5.0×10^{-4} RMSE of estimating the concentration of all three components in the most intense sample is still realized. Theoretically, the lower limit of the cut-off is governed by the available degrees of freedom for estimating the PARAFAC parameters.

When random, instrumental errors are present, there is a practical limit on the acceptable cut-off levels. As with the error free data, employing a cut-off above the linear region will add a bias to the estimated parameters in the model. This is evident in the rapidly increasing RMSE seen in Fig. 3 (solid symbols) as the cut-off becomes increasingly greater than 1.5 AU. At the higher cut-offs, the bias in the PARAFAC model predominates and the RMSE for

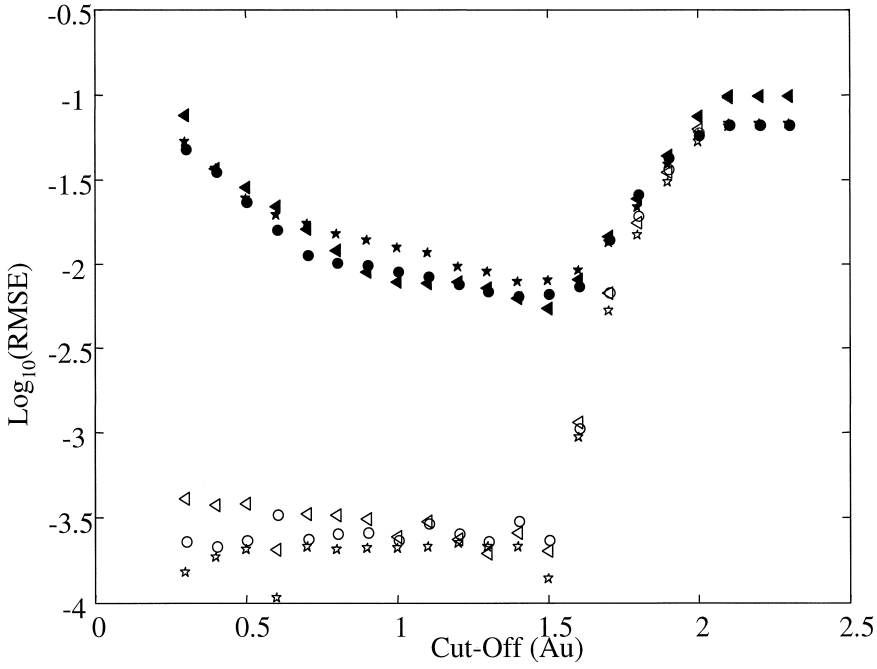

Fig. 3. Root mean squared error (RMSE) for piece-wise nonlinear three component mixture set without added random errors (open symbols) and with added random errors (solid symbols).

the noiseless and noise-added data are equivalent. At lower cut-offs, the signal averaging advantage is lost with the degrees of freedom and the RMSE rapidly increases.

A second, realistic, nonlinearity that occurs in HPLC-UV/VisDAS data is induced by stray light impinging on the detector. For the second simulation, the ability of weighted PARAFAC to mitigate the effect of stray light was investigated. Since there is no true linear region with this type of data, no explicit cut-off exists that completely constrains the model to a linear region of the data. Instead, since the data becomes increasingly nonlinear as the absorbance increases, the RMSE decreases with decreasing cut-off (Fig. 4, open symbols). However, when random, instrumental errors are present, signal averaging advantages are eventually lost as fewer and fewer individual measurements are employed to estimate the PARAFAC model parameters. Thus, there are two competing effects; systematic errors from bias increase as the cut-off level increases and random errors from precision increase as the cut-off level decreases (Fig. 4, closed symbols). Consequently, the minimum RMSE occurs with a 1 AU cut-off in this application with a cut-off range of 1.0 to 1.5 AU yielding equivalent RMSEs.

4.3. HPLC-DAS data

4.3.1. Preliminary analysis

The nonlinear response imbedded in this dataset is demonstrated in the plot of normalized (unit area) spectra of the three analytes (Fig. 5a). Here, the UV/Vis spectra of each analyte are collected throughout the chromatographic peak of the most concentrated standard. The relative intensities of the toluene 271 nm and *p*-xylene 276 nm peaks decrease compared to the toluene 263 nm and *p*-xylene 271 nm peaks, respectively. Little change is seen in the naphthalene spectrum, although a slight relative decrease is observed in the 278 nm peak. It is probable that this nonlinearity is not derived from the detector response but is a result of concentration dependent

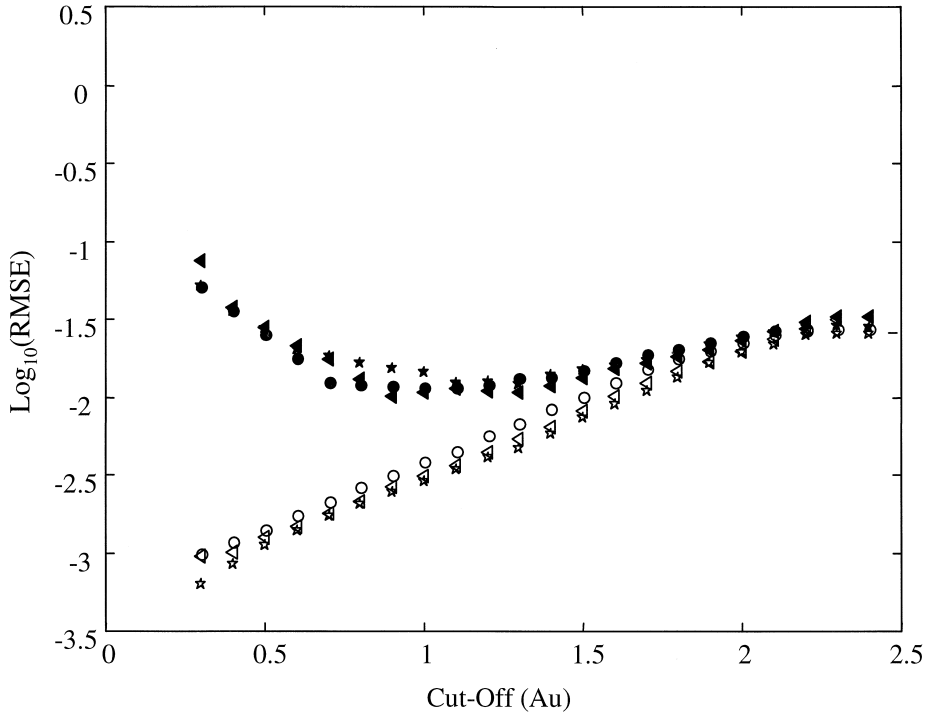

Fig. 4. RMSE for continuously nonlinear three component mixture set without added random errors (open symbols) and with added random errors (solid symbols).

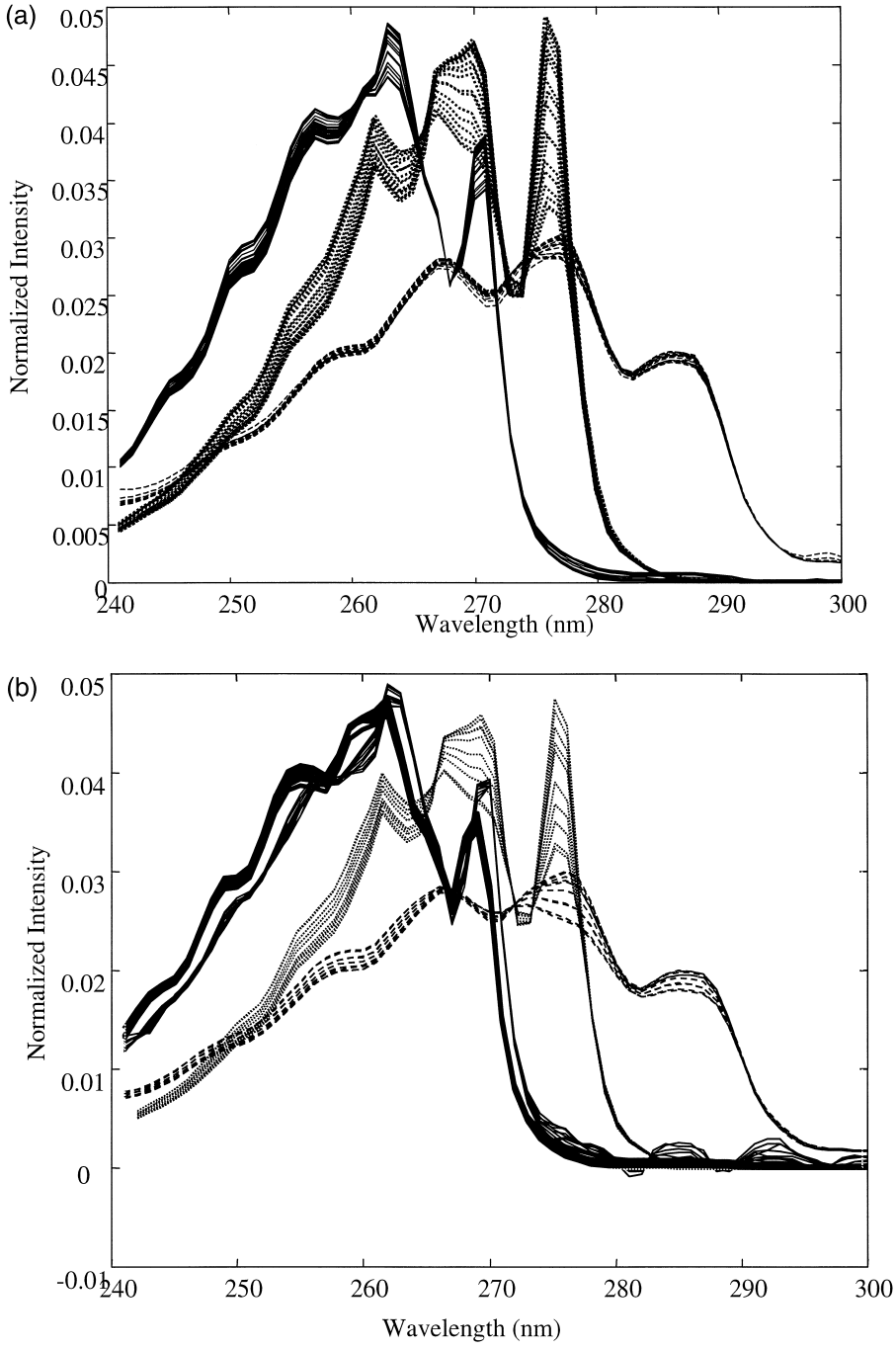

Fig. 5. Unit area normalized spectra from (a) three component data set and from (b) four component data set of toluene (solid lines), naphthalene (dashed lines), *p*-xylene (dotted lines) and ethylbenzene (bold lines).

changes in the toluene and *p*-xylene UV/Vis spectra. This assertion stems from the fact that the maximum naphthalene and *p*-xylene signals are 1.60 AU and 1.62 AU, respectively, and the toluene standard's

maximum absorbance is 1.30 AU. Yet naphthalene shows little nonlinearity compared to the less intense toluene. Furthermore, when the chromatographic profiles are plotted for each wavelength, toluene and *p*-xylene show significant flattening at the top of the chromatographic peaks at 271 nm and 276 nm, respectively. The second data, collected at lesser average concentrations, demonstrated no significant nonlinearity in the toluene spectra (1.59 AU maximum) and increased nonlinearity in the naphthalene 278 nm peak (Fig. 5b). The nonlinearity observed in *p*-xylene (1.49 AU maximum) is comparable to the nonlinearity observed with *p*-xylene in the first data set; although a plot of absorbance at 276 nm vs. absorbance at 271 nm (or 263 nm) shows that the second data set demonstrates a slightly greater degree of nonlinearity. The increased nonlinearity could be a result of increased solvatochromatic effects in the 90:10 methanol:water mobile phase compared to the 95:5 methanol water mobile phase. No appreciable nonlinearity is observed with the ethylbenzene (0.9 AU maximum).

4.3.2. Three-analyte data set (high concentrations)

Employing the weighted PARAFAC significantly improves the RMSE of prediction for all three analytes in the first data set. The RMSE of naphthalene is reduced 24.4% from 19.9 ppm to 14.9 ppm. Toluene prediction error is reduced 38.9% from 537.1 ppm to 328 ppm and the RMSE for *p*-xylene decreases 31.6% from 384 ppm to 263 ppm. Therefore, the overall accuracy of analysis improves from approximately 10% of maximum concentration to only 3% to 4% for these analytes. The correlation between cut-off employed for the weighted PARAFAC and the RMSE is presented in Fig. 6. Toluene and *p*-xylene experience a minimum in the RMSE when a cut-off of 1.1 AU is employed; naphthalene is minimized with a cut-off of 1.3 AU. The minima of these curves are quite broad, ± 0.2 AU, so exact determination of the optimal cut-off is not critical. Choosing an intermediate cut-off value of 1.2 AU does not drastically degrade the predictive accuracy. The RMSE for *p*-xylene increases by 1.6 ppm while toluene and naphthalene increase by 0.2 and 0.02 ppm, respectively.

4.3.3. Four-analyte data set (low concentrations)

Investigation of optimal cut-off values for the four component, lower concentration, data set yields additional insights for the application of weighted PARAFAC to nonlinear three-way data.

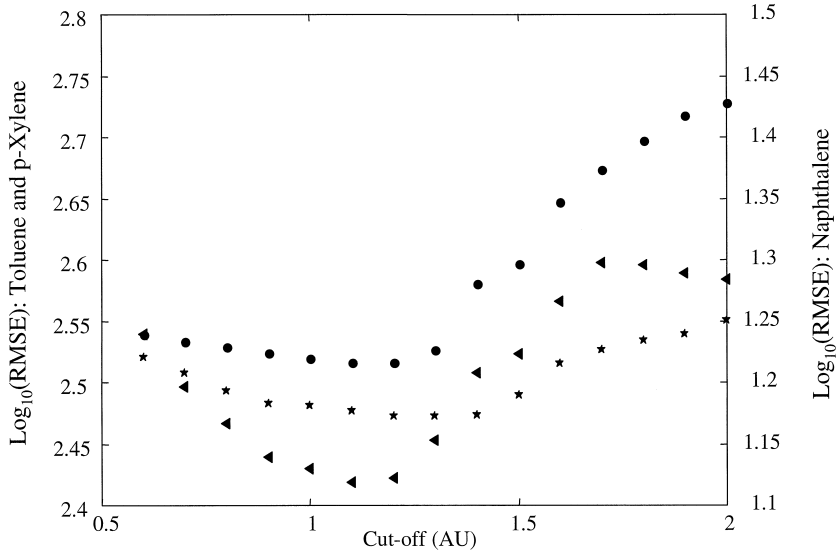

Fig. 6. RMSE as a function of applied cut-off calculated from the three component data set for toluene (circles), naphthalene (stars) and *p*-xylene (triangles).

4.3.3.1. Ethylbenzene. In general, ethylbenzene is not precisely predicted at any cut-off level. The best RMSE of ethylbenzene is 354 ppm (11% error relative to maximum concentration) in this data set compared to 80 ppm for the other analytes in the same concentration range and does not systematically

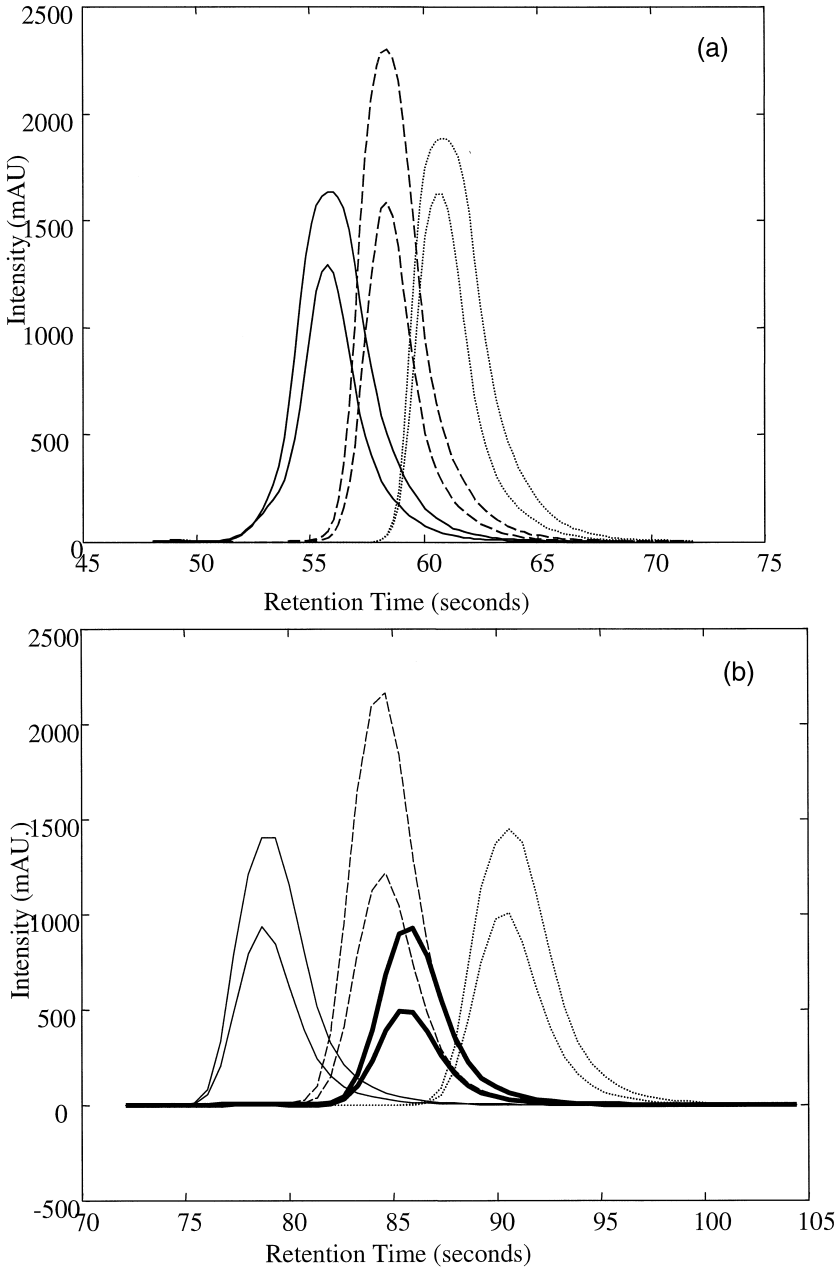

Fig. 7. Pure component chromatograms from (a) three component data set and from (b) four component data set of toluene (solid lines), naphthalene (dashed lines), *p*-xylene (dotted lines) and ethylbenzene (bold lines). Each analyte is shown at maximum concentration and 1/2 maximum concentration.

change with varying the cutoff level. Precise resolution of the spectral, chromatographic, and concentration profiles are hindered in part due to the large spectral similarity between ethylbenzene and toluene (Fig. 5b) and large chromatographic similarity with naphthalene (Fig. 7b). The signal overlap between components, theoretically, does not degrade the performance of PARAFAC to a great extent. However, the chromatographic irreproducibility between samples is significant compared to the difference in retention times for naphthalene and ethylbenzene. Retention time reproducibility has been demonstrated to be a limiting factor in successful application of three-way calibration models [4,22,23]. Weighted PARAFAC does not attempt to alleviate modeling problems associated with chromatographic shifts. The effects of retention time shifts are also seen in the errors of model fit for each sample. The majority of model errors display strong time dependent derivative features that are indicative of chromatographic shifting among the samples.

4.3.3.2. p-Xylene. Application of weighted PARAFAC gains a 34% improvement on the RMSE. As in the first data set, *p*-xylene demonstrates a sharp dip in the plot of RMSE vs. cut-off (Figs. 6 and 8). However, in the second data set, the minimum RMSE occurs at 0.9 AU while the minimum for the first data set occurs at 1.1 AU. This effect could be partially due to the increased *p*-xylene nonlinearity in the second data set such that additional signal should be excluded. While it is unlikely that the decrease in average *p*-xylene concentration in the second data set compared to the first data set resulted in the shift in optimal cut-off, confounding effects from variance in chromatographic overlap and retention time reproducibility between the two data sets make it difficult to reliably attribute this change in optimal cut-off to any particular cause.

4.3.3.3. Toluene. Analysis for toluene yields only a 2.3% improvement in RMSE. That the decrease in RMSE for toluene with application of weighted PARAFAC is much less than the improvement realized with *p*-xylene is expected since the nonlinearity associated with toluene is less than the nonlinearity associated with *p*-xylene; the toluene spectra are modeled well with high cut-off values even though *p*-xylene spectra are not well-modeled with the four-factor PARAFAC model.

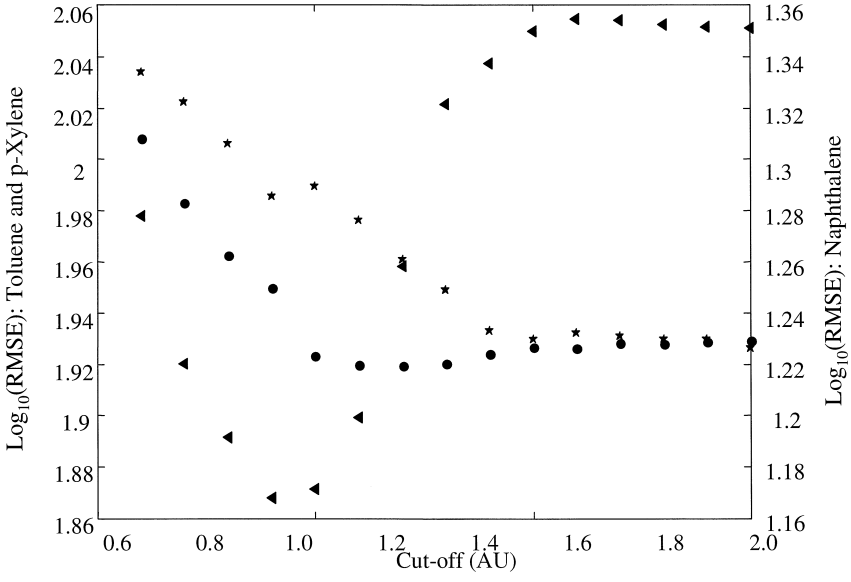

Fig. 8. RMSE as a function of applied cut-off calculated from the four component data set for toluene (circles), naphthalene (stars) and *p*-xylene (triangles).

Table 1
Correlation coefficients for RMSE of each analyte as a function of interfering analytes

Interferent \ analyte (cut-off)	Naphthalene (none)	Ethylbenzene (none)	Toluene (1.2 AU)	*p*-Xylene (0.9 AU)
Naphthalene	0.00	0.20	−0.03	0.15
Ethylbenzene	−0.41	0.00	−0.11	0.46
Toluene	−0.23	−0.31	0.00	0.40
p-Xylene	−0.21	0.01	−0.11	0.00

4.3.3.4. Naphthalene. Naphthalene showed no improvement in RMSE with decreasing cut-off in application of weighted PARAFAC. In fact, a steady increase in RMSE was observed with the most rapid ascent at cut-offs less than 1.3 AU. The lack of improvement in naphthalene prediction is attributed to the chromatographic instability coupled with the chromatographic collinearity with ethylbenzene. Table 1 presents the correlation coefficients between the concentration each analyte with the prediction error for each analyte. The largest correlation coefficients were with ethylbenzene as an interferent overlapping *p*-xylene and naphthalene. Eliminating the largest measurements primarily eliminates data in the naphthalene/ethylbenzene and ethylbenzene/toluene measurement areas. This would account for the earlier, and more rapid, increase in RMSE with decreasing cut-off for naphthalene and toluene.

4.3.3.5. Split data set. By splitting this data set in half (high concentration samples and low concentration samples), it can be seen that the observed behavior of the weighted PARAFAC model with this data set agrees with the observed behavior of the weighted PARAFAC model in the Monte Carlo Simulation. Fig. 9 presents a plot of RMSE vs. cut-off for the high and low concentration halves of the four component data set. When just the higher concentration samples are analyzed for *p*-xylene (solid triangles), a minimum in the RMSE vs. cut-off is observed at 0.9 AU.

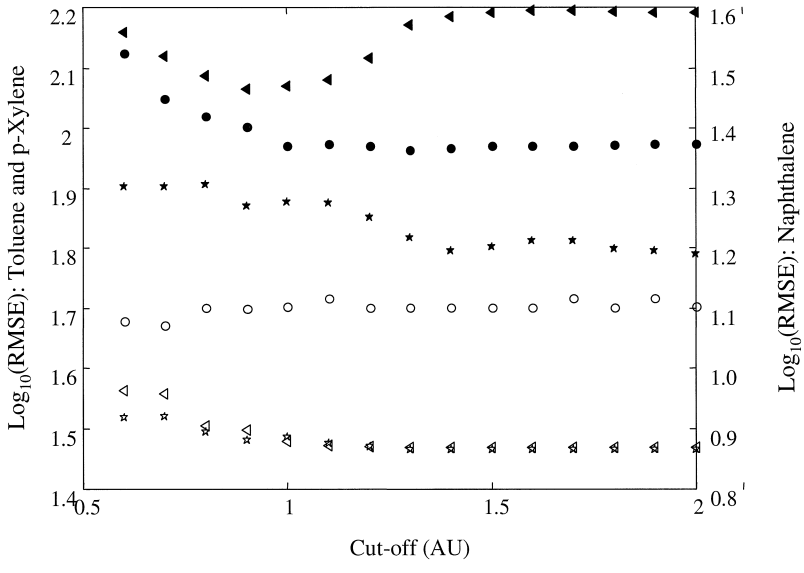

Fig. 9. RMSE as a function of applied cut-off calculated from the higher concentration half (solid symbols) and the lower concentration half (open symbols) of the four component data set for toluene (circles), naphthalene (stars) and *p*-xylene (triangles).

As with the *p*-xylene analysis, the RMSE for toluene at higher concentrations (Fig. 8 circles and Fig. 9 solid circles) begins to increase when a cut-off less than 1.0 AU is employed. At this cut-off level, the negative effect of the decreasing signal to noise in the fitted data begins to surpass the positive effect of applying the PARAFAC model to the increasingly linear region of the data.

It is interesting to note that no improvement in RMSE is observed when weighted PARAFAC is applied the lower concentration half of the data set (Fig. 9, open symbols). The lack of improvement is understood by realizing that only a small fraction of the data have an absorbance value greater than 1.0 AU. At a cut-off of 0.9 AU, the most intense (and theoretically informative) 2.9% of the data are excluded from analysis of the full data set with 5.6% of the high half of the data set being excluded and only 0.6% of the low half of the data set being excluded. Considering that 60.5% of the data are less than 10 mAU and 76.8% of the data are less than 100 mAU in the full data set, a cut-off of 0.9 AU excludes between 7% and 25% of the best quality data from the full data set. In terms of spectral intensity, a cut-off of 0.9 AU excludes 31.0%, 43.0%, 9% of the total signal from the whole, high concentration half, and low concentration half of the data set, respectively.

4.4. Determination of optimal cut-off

One method for determining the optimal cut-off is to observe the RMSE of estimation. This is fine when the analyte, or analytes, concentration is known in many of the samples investigated and these known samples exceed the linear range of signal intensities. However, this is not always the case and possessing and alternative, corroborating, means to assess the proper cut-off is useful.

4.4.1. Simulations (with linear region)

Empirically, the optimal cut-off can be also be determined by analyzing a plot of modeled data vs. measured data. The modeled data is formed by reconstructing one or more samples of the data set from the estimated PARAFAC parameters. In the case of data with an explicit linear range, this plot will be linear up to the optimal cut-off value whenever a cut-off less than or equal to the optimal cut-off is chosen. This is shown in Fig. 10a for the most intense sample in the simulation. If a cut-off greater than the optimal cut-off is chosen, the plot will be slightly nonlinear up to the optimal cut-off; larger deviations from linearity will occur at greater intensities than the optimal cut-off (Fig. 10a). In fact, when a cut-off in the linear region is chosen, the plot of modeled vs. measured data passes through the point ⟨cut-off, cut-off⟩. The greatest value that the plot passes through a point with the same ordinate and abscissa value is the maximum cut-off that includes only linear data. The appropriateness of choosing cut-offs in the linear region to model the data can be seen in Fig. 10b. At cut-offs below 1.5 AU, the reconstructed data from the PARAFAC model extrapolates to the true underlying data before any nonlinearities were introduced. Concurrently, when data in the nonlinear region are included in determining the model parameters, deviations occur both above and below the maximum linear region.

4.4.2. Simulations (with no linear region)

When no explicit linear region exists in the data, the choice of optimal cut-off is more difficult. Fig. 11a presents a plot of model errors vs. modeled intensity for four distinct cut-off levels applied to the simulated stray light data with no added random errors. The modeled data from the 2.0 AU cut-off demonstrates a significant lack of fit to the measured data both above and below the cut-off value. This cut-off value does not result in accurate estimation of analyte concentrations (Fig. 4). Concurrently, although the degree of fit is very poor above the cut-off value, the 0.5 AU cut-off fits the experimental data accurately up to the employed cut-off. This cut-off also yielded the lowest RMSE of estimation for analyte concentration. Comparing Fig. 11a to Fig. 4, the strong correlation between fit of the model to the data below the cut-off value to RMSE of prediction is evident. When the model fit is excellent below the cut-off, the RMSE of prediction over the whole data set is low. However, when the model fit below the cut-off is poor, the RMSE of prediction is high throughout the whole data set.

When random errors are added to the continuously nonlinear data, a strong correlation between model fit below the applied cut-off and RMSE is observed. As with the errorless data, the 2.0 AU cut-off

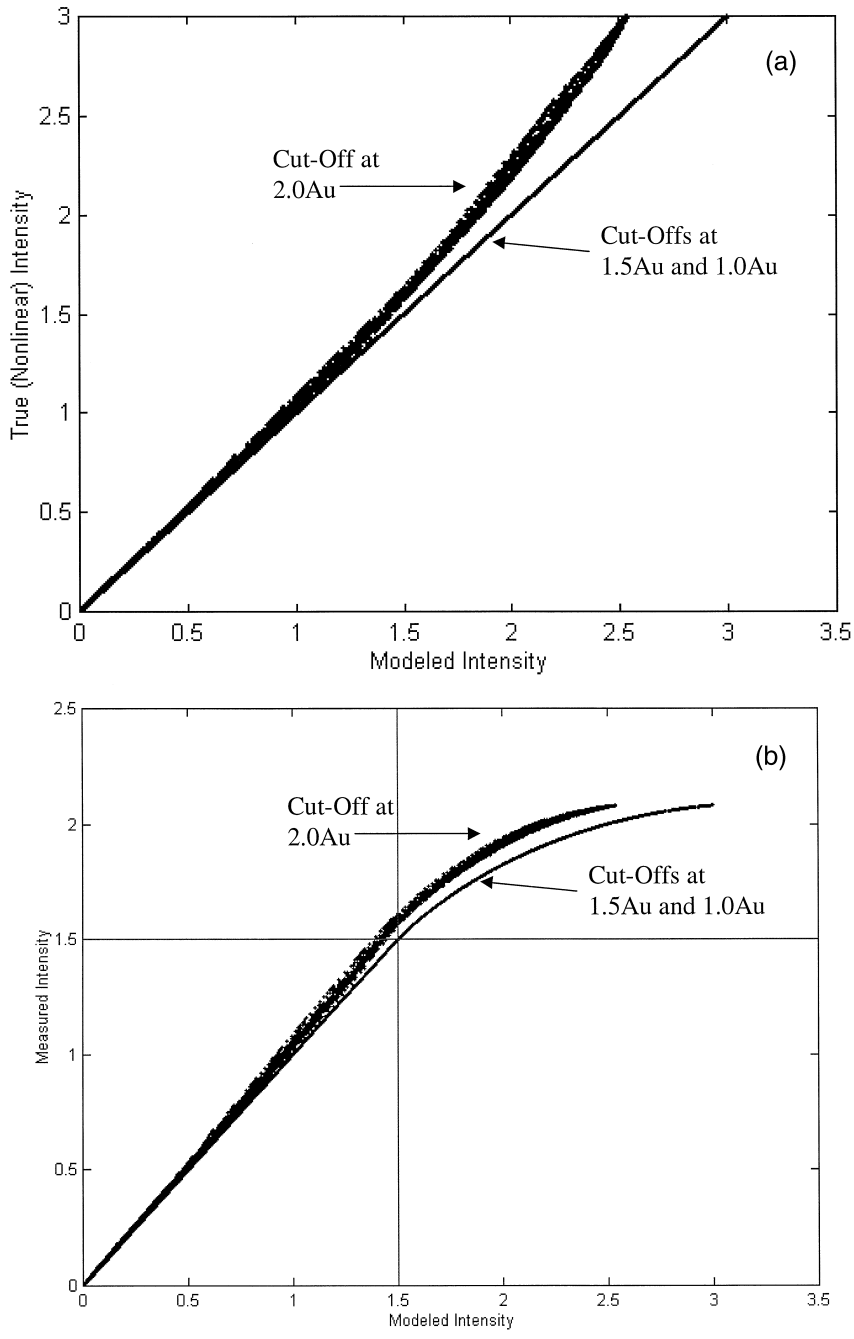

Fig. 10. (a) Modeled vs. true latent linear spectral intensities as a function of three separate cut-offs applied with weighted PARAFAC to errorless simulated HPLC-UV/Vis data. The optimal cut-off is 1.5 AU or less. (b) Modeled vs. measured spectral intensities as a function of three separate cut-offs applied with weighted PARAFAC to errorless simulated HPLC-UV/Vis data. The optimal cut-off is 1.5 AU or less.

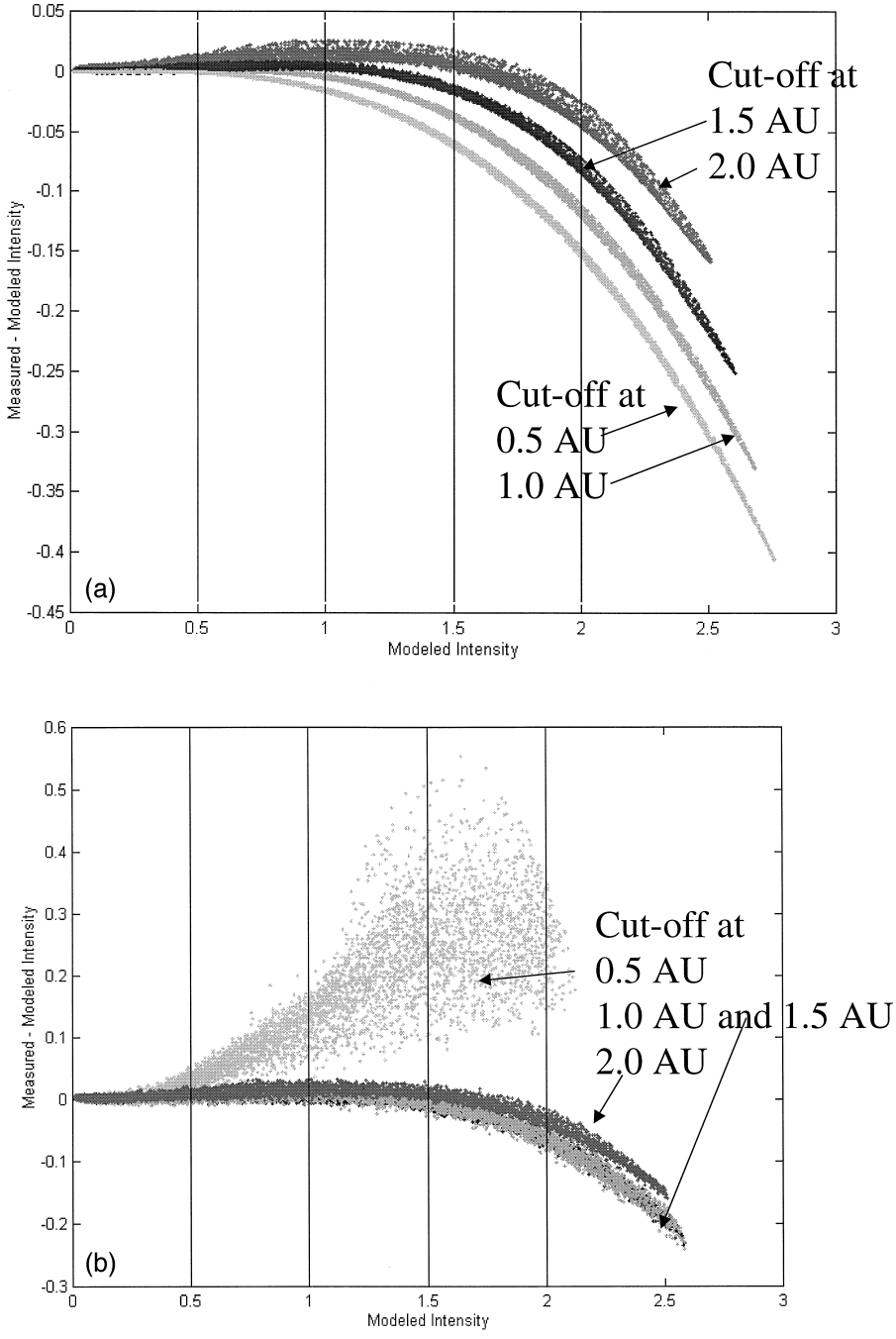

Fig. 11. Model error vs. modeled intensity for four different cut-offs applied to continuously nonlinear simulated (5% stray light) HPLC-UV/Vis data. (a) No random errors added, (b) random errors added.

model neither fits the data below the cut-off (Fig. 11b), nor yields a low RMSE of estimation (Fig. 4). Unlike the errorless data, the PARAFAC model based on the 0.5 au cut-off yields large deviations from the

experimental data and a relatively large RMSE of estimation compared to the minimum RMSE of fit observed for the data in Fig. 4. The reversal in weighted PARAFAC's ability to quantitatively model the data with this low cut-off is understandable since, in this application, the signal-to-noise ratio of the data employed to estimate the model parameters is at most 10 for any given datum and usually much less. Interestingly, the 1.0 AU and 1.5 AU cut-off models fit the data almost identically over the whole range of measured intensities. Based on the fact that the two models fit the data in the range of 0 to 1.0 AU equivalently, and that the 1.5 AU cut-off based model attempts to model an additional set of slightly nonlinear data, it is understandable that the two models (and those models in the continuum between them) yield roughly equivalent RMSE of estimation with the model derived from the most linear region of data yielding the lowest RMSE. However, the 1.0 AU cut-off data yield lower errors of fit over its range of intensities employed to calculate the model parameters (0 to 1.0 AU) that does the 1.5 AU cut-off based model (0 to 1.5 AU) and consequently, yields the lowest RMSE of estimation.

4.4.3. HPLC-UV / Vis data

The utility of estimating the optimal cut-off by plotting the error of model fit error vs. the measured

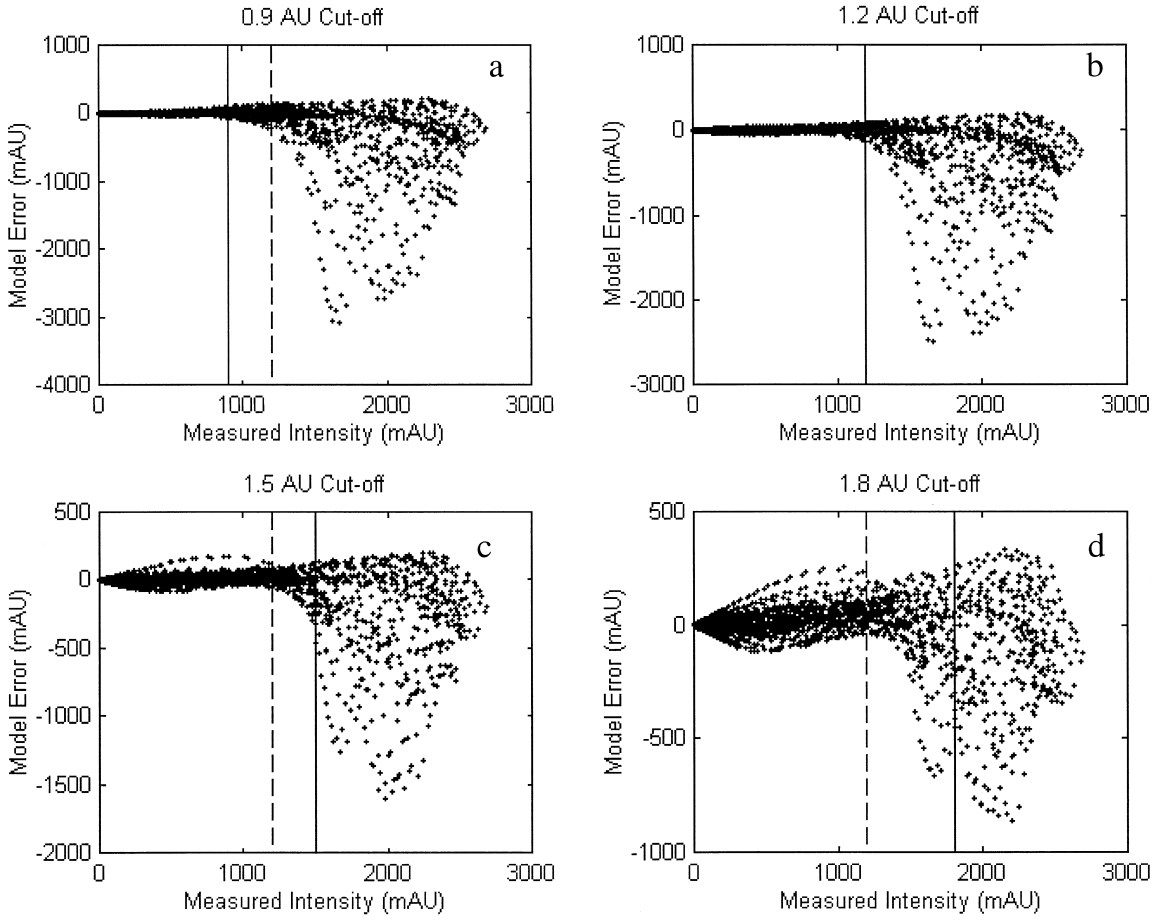

Fig. 12. Model error vs. measured intensity for four different cut-off values applied with weighted PARAFAC to the three component HPLC-UV/Vis data. The solid line represents the value of the applied cut-off; the dashed line is the optimal cut-off as determined from Fig. 6.

(or modeled) data seen in the analysis of Figs. 12 and 13. In figure set 12, the fit error plot is constructed for one of the more concentrated samples (4000 ppm toluene, 427 ppm naphthalene, 6000 ppm *p*-xylene). For this data set, cut-off between 1.1 AU and 1.3 AU were found to yield the minimum RMSE of estimation. When either a cut-off of 0.9 AU (Fig. 12a) or 1.2 AU (Fig. 12b) are employed, the fit error plots do not show significant model error at measured intensities below 1.1 to 1.3 AU. In both cases, the model deviations do not occur at measured intensities less than the chosen cut-off. This is consistent with having employed a cut-off that is equal to or less than the appropriate cut-off value. When higher cut-offs are tested (Fig. 12c and d), model errors are observed at measured intensities both above and below the chosen cutoff and above and below the optimal cut-off. This error structure is consistent with choosing a cut-off greater than the optimal cut-off value.

Concurrently, the choice of a 0.9 AU cut-off for the optimization of *p*-xylene prediction can be seen in figure set 13. Regardless of the applied cut-off, large model deviations begin at measured intensities greater than 0.9 AU. It is interesting to note that, these errors are associated only with the *p*-xylene region of the chromatogram-UV/Vis data. From this, it is more

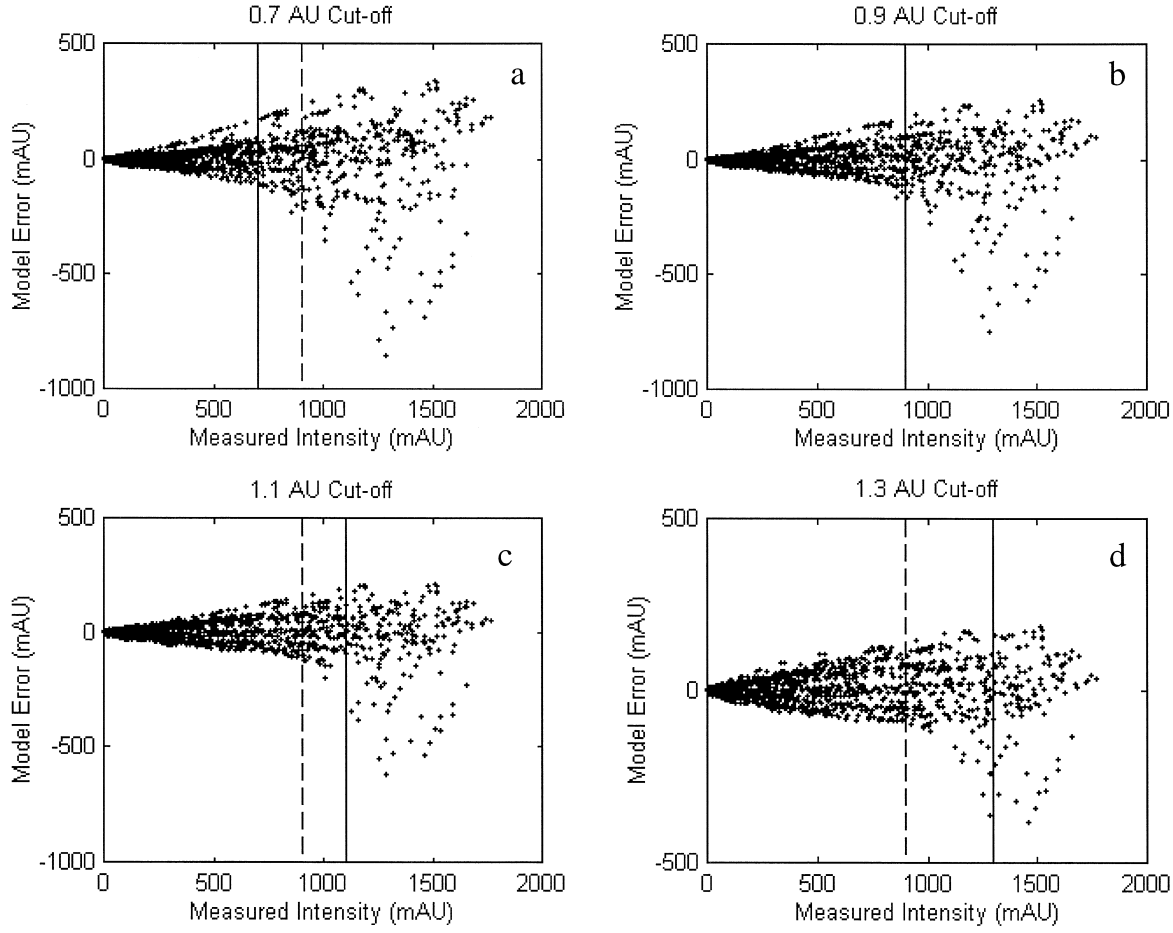

Fig. 13. Model error vs. measured intensity for four different cut-off values applied with weighted PARAFAC to the four component HPLC-UV/Vis data. The solid line represents the value of the applied cut-off; the dashed line is the optimal cut-off as determined from Fig. 9.

understandable that weighted PARAFAC presented little improvement for the prediction of well-modeled naphthalene and toluene.

5. Conclusions

Weighted PARAFAC enables accurate quantitative and qualitative analysis of nonlinear three-way data. This method works by finding the most nonlinear regions of the data set and eliminating these regions from consideration during model construction. Presented here is a limited view of the application and utility of weighted PARAFAC. In this article, the nonlinear regions have been treated as missing data and assigned a weight of zero. Also, only nonlinearities that increase with increasing signal were considered. However, it is also possible to use employ weight along a continuous scale and to distribute these weights with any number of different criteria. From this starting point, weighted PARAFAC should prove useful in extending the dynamic range of hyphenated instrumentation and increasing the applicability of three-way analysis methods to more nonlinear problems.

It should be noted that there is room for refinement in determining the exact the weight matrix from the assumed cut-off. As performed here, there will be a small number of measurements that will be assigned to the wrong side of the cut-off due to random instrumental errors. While this misassignment would have a negligible, in most cases, where the majority of the data lies in the linear region, a degradation of precision and accuracy may be observed in extreme cases when only a small fraction of the data is 'linear.' Options to avoid potential problems associated with such spurious rejection or acceptance of data include basing the rejection on the average of neighboring values [24], and iterative reweighting techniques [25,26].

References

[1] R.A. Harshman, Foundations of the PARAFAC Procedure, UCLA Working Paper in Phonetics, 1970, 16, 1–84.
[2] J.D. Carrol, J. Chang, Analysis of individual differences in multidimensional scaling via a N-way generalization of an Eckart–Young decomposition, Psychometrica 35 (1970) 283.
[3] P.M. Kroonenberg, Three-Mode Principal Component Analysis. Theory and Applications, DSWO Press, Leiden, 1983.
[4] R. Poe, S. Rutan, Effects of resolution, peak ratio, and sampling frequency in diode array fluorescence detection in liquid , Anal. Chim. Acta 283 (1993) 245–253.
[5] S. Li, P.J. Gemperline, Generalized rank annihilation method using similarity transformations, Anal. Chem. 64 (1992) 599–607.
[6] A.K. Smilde, D.A. Doornbos, 3-Way methods for the calibration of chromatographic systems — comparing PARAFAC and 3-way PLS, J. Chemom. 5 (1991) 345–360.
[7] H.L. Wu, M. Shibukawa, K. Oguma, An alternating trilinear decomposition algorithm with application to calibration of HPLC-DAD for simultaneous determination of chlorinated aromatic hydrocarbons, J. Chemom. 12 (1998) 1–26.
[8] D.S. Burdick, X.M. Tu, L.B. McGown, D.W. Millican, Resolution of multicomponent fluorescent mixtures by analysis of the excitation-emission frequency arrays, J. Chemom. 4 (1990) 15–28.
[9] M.M.C. Ferreira, M.L. Brandes, I.M.C. Ferreira, K.S. Booksh, W.C. Dolowy, M. Gouterman, B.R. Kowalski, A chemometric study of fluorescence of dental calculus by trilinear decomposition, Appl. Spectrosc. 49 (1995) 1317–1325.
[10] K.S. Booksh, A.R. Muroski, M.L. Myrick, Single-measurement EEM spectrofluorometer for determination of hydrocarbons in ocean water: 2. Calibration and quantization of naphthalene and styrene, Anal. Chem. 68 (1996) 3539–3544.
[11] R. Bro, C.A. Andersson, Improving the speed of multiway algorithms: Part II. Compression, Chemom. Intell. Lab. Syst. 42 (1998) 105–113.
[12] R.D. JiJi, G.C. Cooper, K.S. Booksh, Excitation-emission matrix fluorescence based determination of carbamate pesticides and polycyclic aromatic hydrocarbons, Anal. Chim. Acta 1999, in press.
[13] J.D. Kruskal, Rank, Decomposition, and Uniqueness for 3-way and N-way Arrays, in: R. Coppi, S. Bolasco (Eds.), Multiway Data Analysis, Elsevier, Amsterdam, 1989.
[14] K.S. Booksh, B.R. Kowalski, Theory of analytical chemistry, Anal. Chem. 66 (1994) 782A–791A.
[15] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–172.
[16] K.S. Booksh, B.R. Kowalski, Calibration method choice by comparison of model basis functions to the theoretical instrument response function, Anal. Chim. Acta 348 (1997) 1–9.
[17] K.S. Booksh, Z. Lin, Z. Wang, B.R. Kowalski, Extension of trilinear decomposition method with an application to the flow probe sensor, Anal. Chem. 66 (1994) 2561–2569.
[18] A.K. Smilde, R. Tauler, J.M. Henshaw, L.W. Burgess, B.R. Kowalski, Multicomponent determination of chlorinated hydrocarbons using a reaction based sensor: 3. Medium-rank second order calibration with restricted Tucker models, Anal. Chem. 66 (1994) 3345–3351.
[19] A.K. Smilde, Y. Wang, B.R. Kowalski, Theory of medium-rank second order calibration with restricted Tucker models, J. Chemom. 8 (1994) 21–36.
[20] K.S. Booksh, J.M. Henshaw, L.W. Burgess, B.R. Kowalski, The second order standard addition method with an applica-

tion to in situ environmental monitors, J. Chemom. 9 (1995) 263–282.
[21] J.D. Ingles, S.R. Crouch, Spectrochemical Analysis, Prentice Hall, Englewood Cliffs, NJ, 1988.
[22] K.S. Booksh, B.R. Kowalski, Error analysis of the generalized rank annihilation method, J. Chemom. 8 (1994) 45–64.
[23] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Standardization of second order chromatographic/spectroscopic data for optimum chemical analysis, Anal. Chem. 70 (1998) 218–225.
[24] P. Paatero, Least squares formulation of robust, non-negative factor analysis, Chemom. Intell. Lab. Syst. 37 (1997) 23–35.
[25] M.I. Griep, I.N. Wakeling, P. Vankeerberghen, D.L. Massart, Comparison of semirobust and robust partial least squares procedures, Chemom. Intell. Lab. Syst. 29 (1995) 37–50.
[26] G.R. Phillip, M.E. Eyring, Comparison of conventional and robust regression in analysis of chemical data, Anal. Chem. 55 (1983) 1134–1138.