Available online at www.sciencedirect.com

SCIENCE DIRECT®

Analytica Chimica Acta 487 (2003) 171–180

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Mitigating model deficiency in three-way data analysis by the combination of background constraining and iterative correcting techniques

Zeng-Ping Chen, Ru-Qin Yu*

College of Chemistry and Chemical Engineering, State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082, PR China

Received 19 July 2002; received in revised form 10 February 2003; accepted 12 May 2003

Abstract

PARAFAC is a popular model for trilinear data analysis in analytical chemistry. The prerequisite for the successful application of PARAFAC in analytical chemistry is that the three-way data array should follow a trilinear model, which is always violated by the presence of deviations such as Rayleigh scattering in fluorescence spectroscopy. In order to mitigate the influence of model deviations, background constraining and iterative correcting techniques are advocated in this contribution. The method established on these two techniques can nearly eliminate the effect of model deviation on the chemical loading parameters estimated. Compared with other methods for mitigating model deviations, the proposed method requires no prior knowledge about the chemical loading parameters. It is also unnecessary to assign weights to data entities as the weighted PARAFAC of Anderson does. Its implementation is comparable to PARAFAC-ALS and can be programmed to be completely automatic. Its performance has been demonstrated by fluorescent and chromatographic experiments.

Keywords: Trilinear model; Model deficiency; PARAFAC; Background constraining; Iterative correcting

1. Introduction

Due to the so-called second-order advantage [1], trilinear data analysis has been one of the most active areas in analytical chemometrics and attracted attentions of many chemometricians. Researches on the theoretical and application aspects of trilinear data analysis have being vigorously flourished [2–12]. Among the decomposition algorithms for trilinear data analysis, PARAFAC-ALS [2–4] might be one of the two most popular ones (the other is GRAM [5]), for the uniqueness and optimality of its results. The most important prerequisite for the successful application of PARAFAC model in analytical chemistry is, however, that the data arrays should follow a trilinear model, which might be violated in practice. Rayleigh scattering in fluorescence spectroscopy is but one instance. Other deviations (thereafter, "deviations" denotes the "deviations from trilinear model") in chemical data had been classified by Booksh and Kowalski [13]. Hence, measures should be taken to mitigate the influence of model deficiency in analyzing data arrays contaminated with deviations, otherwise the estimated model parameters would be misleading.

The presence of deviations justifies the necessity to impose some reasonable constraints on the trilinear

* Corresponding author. Tel.: +86-731-8822782; fax: +86-731-8822782.
E-mail address: rqyu@hnu.net.cn (R.-Q. Yu).

0003-2670/03/$ – see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0003-2670(03)00569-5

model parameters. Among all the constraints used, non-negativity is the most common one [14]. Unimodel property of chromatograms is also an effective constraint in decomposing chromatographic data arrays [15]. Furthermore, the known concentration matrix of the calibration samples was also used as constraints in second-order linear calibration, which is an important application branch of trilinear data analysis [16]. A variety of other effective constraints on model parameters have been summarized by Bro [17]. Some times, the loading parameters in one or two modes of the underlying factors designating chemical variations are known before decomposition. So, it is natural to preset them equal to the prior known values. It seems that the quality of the final results could be improved to some extent through adopting this kind of constraints, the least squares fitting procedure of PARAFAC-ALS in essence, however, would distort the loading parameters in the rest unconstrained modes with a view to account for data variations as large as possible. Therefore, the estimated loading parameters in the unconstrained mode(s) would be even worse than the corresponding ones when the loading parameters in all the three modes are unconstrained. Theoretically, an efficient way to alleviate the influence of deviations is to include extra factors other than those representing chemical variations in the trilinear model to account for background deviations. However, the sensitivity of PARAFAC-ALS to the number of factors used in calculation hinders a straightforward utilization of such technique. The goals of mitigating the influence of deviations and simultaneously stabilizing the results of PARAFAC-ALS may be attained through constraining the loading parameters of the added factors accounting for background deviations.

Besides the constraints on the model parameters, constraints on the original data entities can also be used to enhance the quality of final results under the circumstance of model deficiency. Anderson et al. suggested assigning different weights to data entities measured by fluorescent spectroscopy according to prior knowledge and then performing a weighted PARAFAC [18]. Though better results can be obtained by weighted PARAFAC, it is not an easy task to assign appropriate weights to data entities for data arrays produced by instruments other than fluorescent spectroscopy. In order to avoid the difficulty in presetting weights for data entities, an iterative correcting procedure has been suggested. Along with the aforementioned background constraining technique, it can be expected to significantly enhance the quality of the decomposing results. The power of combining background constraining and iterative correcting procedures in mitigating deviations in three-way data analysis has been demonstrated by two kinds of real data arrays, one produced by the fluorescence spectrometer, and the other by HPLC–DAD.

2. Nomenclature

Throughout this paper, scalars are represented by lower-case italics, vectors by bold-italics lower-case characters, bold capitals designate two-way matrices and underlined bold capitals symbolize three-way data arrays.

3. Theory and algorithm

The matrix form of trilinear model can be expressed as follows:

$$\boldsymbol{X}_{..k} = \boldsymbol{A}_{I\times F}\,\mathrm{diag}(\boldsymbol{c}_k)(\boldsymbol{B}_{J\times F})^{\mathrm{T}} + \boldsymbol{E}_{..k},\qquad k = 1, 2, \ldots, K. \quad (1)$$

In the above series of equations, $\boldsymbol{X}_{..k}$ is the kth frontal slice of three-way data array $\underline{\boldsymbol{X}}$; $\boldsymbol{E}_{..k}$ the kth frontal slice of residue array $\underline{\boldsymbol{E}}$. Elements of loading matrixes $\boldsymbol{A}_{I\times F}$, $\boldsymbol{B}_{J\times F}$ and $\boldsymbol{C}_{K\times F}$ are related to chemical species in the mixtures (in the rest part of this paper, they will be simply represented by $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$). I, J and K are the dimensions of three modes in three-way data array, respectively. F is the number of underlying factors, i.e. the total number of detectable chemical species. In general, the columns of loading matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are assigned with certain physical meanings, i.e. excitation and emission spectra in fluorescence spectrometry, and chromatograms and ultraviolet-visible spectra in HPLC–DAD, etc. Note that $\mathrm{diag}(\boldsymbol{c}_k)$ is a diagonal matrix with diagonal elements equal to the kth row of loading matrix $\boldsymbol{C}$, $\boldsymbol{c}_k$, which designates the concentration vector of chemical components in mixture k.

If data arrays follow the above trilinear model, the three underlying loading matrixes $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ can be accurately estimated by PARAFAC-ALS algorithm.

Unfortunately, the deviations of data arrays from trilinear model are common in chemistry. Instead of the above trilinear model, data arrays contaminated by deviations can be decomposed into four parts:

$$\boldsymbol{X}_{..k} = \boldsymbol{A}\,\mathrm{diag}(\boldsymbol{c}_k)\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{D}1_{..k} + \boldsymbol{D}2_{..k}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \boldsymbol{E}_{..k}, \qquad k = 1, 2, \ldots, K. \tag{2}$$

Here $\boldsymbol{D}1_{..k}$ represents deviations independent of the chemical loading matrixes $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ (such as Rayleigh scattering in fluorescent data arrays and background fluctuation in chromatographic ones). $\boldsymbol{D}2_{..k}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ denotes deviations originated in the non-linear response in certain wavelengths or retention time points, which are associated with the three chemical loading matrix $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$.

Generally, imposing reasonable constraints on the loading parameters of underlying factors representing chemical variations (for convenience of presentation, '*loading parameters of factors representing chemical variations*' will be simplified to '*chemical loading parameters*' in the following parts) was considered an effective measure to mitigate the influence of $\boldsymbol{D}1_{..k}$. Sometimes satisfactory results may be observed with the adoption of this technique in practice. However, due to the least square fitting procedure of PARAFAC-ALS, if chemical loading parameters in only one or two modes are constrained to equal to prior known values, the estimated chemical loading parameters in the rest unconstrained mode(s) would be distorted so as to explain data variations as large as possible. Therefore, they might be getting even worse than the counterparts when chemical loading parameters in all the three modes are free of constraints. A straightforward solution to this problem might be to include extra factors in trilinear model to account for $\boldsymbol{D1}_{..k}$. Since $\boldsymbol{D1}_{..k}$ does not follow a trilinear model, the employment of a trilinear model with factors larger than the number of chemical species in mixtures would cause the estimated loading parameters of PARAFAC-ALS to lose their chemical meanings. In order to retain the chemical meaning assigned to loading matrixes $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, the present authors advocate exerting constraints on the loading parameters of the extra factors in two modes. With the space spanned by the extra factors being confined, the estimated chemical loading parameters can thus be stabilized. The presence of extra factors will account for parts of the deviations $\boldsymbol{D1}_{..k}$, and hence diminishes the distortion of estimated chemical loading parameters during the least squares fitting procedure. It is obvious that the efficiency of the above approach in mitigating the influence of $\boldsymbol{D1}_{..k}$ on the chemical loading parameters is directly related to the overlapping degree between the space spanned by extra factors and that of $\boldsymbol{D1}_{..k}$. The higher the overlapping degree, the smaller the distortion of the chemical loading parameters estimated. Hence, the choice of constraints for the extra factors is of utmost importance. Generally, the space spanned by the extra factors can be constrained to be the space spanned by the response matrixes of blank solutions. A high degree of overlapping between the space spanned by extra factors and that of $\boldsymbol{D1}_{..k}$ can then be expected. The mathematical representation of this approach is as follows. Suppose the singular value decomposition of $\boldsymbol{X}_{\mathrm{blank}}$, the response matrix of the blank solution, is expressed as $\boldsymbol{X}_{\mathrm{blank}} = USV^{\mathrm{T}}$. Then, the decomposition of the original data array with the space spanned by extra factors being constrained is:

$$\boldsymbol{X}_{..k} = \boldsymbol{A}\,\mathrm{diag}(\boldsymbol{c}_k)\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{U}_m\,\mathrm{diag}(\tilde{\boldsymbol{c}}_k)\boldsymbol{V}_m^{\mathrm{T}} + \tilde{\boldsymbol{D}}_{..k} + \boldsymbol{E}_{..k}, \qquad k = 1, \ldots, K. \tag{3}$$

Here, $\boldsymbol{U}_m$ and $\boldsymbol{V}_m$ are matrixes assembled by the first m columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. The kth row of loading matrix $\tilde{\boldsymbol{C}}$, $\tilde{\boldsymbol{c}}_k$, signifies the contributions of m extra factors to mixture k. $\tilde{\boldsymbol{D}}_{..k}$ is the deviation remained which cannot be explained by the addition of excess factors. The unknown loading matrixes $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\tilde{\boldsymbol{C}}$ can be estimated through minimizing $\sum_{k=1}^{K}||\boldsymbol{X}_{..k} - (\boldsymbol{A}\,\mathrm{diag}(\boldsymbol{c}_k)\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{U}_m\,\mathrm{diag}(\tilde{\boldsymbol{c}}_k)\boldsymbol{V}_m^{\mathrm{T}})||_F^2$ by alternating least square algorithm.

Unlike $\boldsymbol{D}1_{..k}$, the main part of the influence of $\boldsymbol{D}2_{..k}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ on the estimation of the chemical loading parameters can only be diminished by assigning relatively small weights (even zeros) to data entities with large deviations. For fluorescent data array, assigning relatively small weights to data entities in the Rayleigh scattering regions and relatively large weights (even ones) to other entities is a reasonable scheme. However, for data arrays produced by other instruments such as HPLC–DAD and GC–MS, there is no generalized guideline to preset weights. With a view to circumvent such dilemma, the present authors advocate the following iterative correcting procedure as an alternative.

1. Decomposing the original data arrays $\boldsymbol{X}$ by PARAFAC-ALS to obtain the fitted values $\hat{\boldsymbol{X}}_{..k}$ $(k = 1, 2, \ldots, K)$, fitness value and accumulated square residue matrix $\boldsymbol{R} = \sum_{k=1}^{K}(\boldsymbol{X}_{..k} - \hat{\boldsymbol{X}}_{..k})^2$.
2. Creating a new data array $\boldsymbol{X}_{\text{new}}$ through replacing the data entities in $\boldsymbol{X}_{..k}$ $(k = 1, 2, \ldots, K)$ whose counterparts in accumulated square residue matrix $\boldsymbol{R}$ are larger than a predefined threshold r_{cut} with the corresponding fitted values in $\hat{\boldsymbol{X}}_{..k}$ $(k = 1, 2, \ldots, K)$, and retaining the rest data entities in $\boldsymbol{X}_{..k}$ $(k = 1, 2, \ldots, K)$ unchanged.
3. Decomposing the newly created data array $\boldsymbol{X}_{\text{new}}$ by PARAFAC-ALS to obtain new fitted values, fitness value and accumulated square residue matrix.
4. Repeating the steps 2–3, until the difference between fitness values of two successive decompositions reaches a predefined small value ε_1 $(1 \times 10^{-6}$, for instance).

The basic assumption, on which the iterative correcting procedure is established, is that the residues corresponding to the data entities contaminated with large deviation are larger than those of others. This assumption can be satisfied on condition that the percentage of the data entities with large deviations is small compared with those contaminated by small deviations. When fitted with PARAFAC-ALS, the data entities with large deviations will be also the ones with large residues. During the iterative correcting procedure, they are replaced by the fitted values of PARAFAC-ALS. Hence, their influence on the chemical loading parameters can be gradually mitigated. It is obvious that the preset threshold value of r_{cut} has a great impact on the quality of final results. Since in each cycle of iterative correcting procedure, the magnitude of the entities in accumulated square residue matrix $\boldsymbol{R}$ is different, employing a fixed r_{cut} for each cycle may be unreasonable. It is necessary to set different r_{cut} values for different iterations. A simple but effective solution for this problem is to connect r_{cut} with the magnitude of the entities in accumulated square residue matrix $\boldsymbol{R}$. In this contribution, the following scheme is adopted.

$$r_{\text{cut}} = \text{Min}(\boldsymbol{R}) + \frac{(1+\alpha)}{2}\left(\text{Max}(\boldsymbol{R}) - \text{Min}(\boldsymbol{R})\right), \quad 0 \le \alpha \le 1. \qquad (4)$$

Here, $\text{Min}(\boldsymbol{R})$ and $\text{Max}(\boldsymbol{R})$ signify the minimum and maximum values of accumulated square residue matrix $\boldsymbol{R}$, respectively; α is a parameter controlling the magnitude of r_{cut}. Once the value of α is set, r_{cut} will vary with $\boldsymbol{R}$ in each iteration.

The background constraining technique can be used to mitigate the influence of relatively common deviations in each data matrixes of the three-way data array. While the iterative correcting procedure aims at preventing the distortion effect of large unique deviations and non-linear response in each data matrixes. These two techniques complement each other. Combining them into one algorithm will definitely provide better performance than employing any one of them alone. The scheme of combining background constraining and iterative correcting techniques is straightforward. Detailed discussion on this subject is unnecessary. For the convenience of readers, implementing guidelines have been supplemented in Appendix A and Appendix B.

4. Experimental

4.1. Excitation–emission fluorescent data arrays

Reagents and stock solutions: All reagents used were of analytical grade. Stock solutions of 1-naphthol (0.1006 mg ml^{-1}) and 2-naphthol (1.001 mg ml^{-1}) were prepared by accurately weighting correspondingly appropriate amount of reagents and dissolving them in distilled water. In the preparation of naphthalene (0.1025 mg ml^{-1}), sufficient amount of NaOH (0.1 M) was added to enhance the solubility of naphthalene in distilled water. A total of 10 working solutions with different concentration ratios of the three components were made by taking appropriate volumes of stock solutions, 2.5 ml of C_2H_5OH and 2.5 ml of NaOH (pH = 13) into a 25 ml volumetric flask and then making them to 25 ml with distilled water.

Apparatus: The excitation–emission response matrices of all the samples plus four blank solutions were recorded by a HITACHI 4500 fluorescence spectrophotometer scanning at 240 nm min^{-1} with excitation wavelength in the range of 220–300 nm and emission wavelength ranging from 315 to 600 nm. The intervals for excitation and emission wavelength were 2 nm and 5 nm, respectively. The slit width in both excitation and emission monochromators was 10 nm.

4.2. HPLD–DAD data arrays

Nine mixtures of three compounds, i.e. *o*-dichlorobenzene, *p*-chlorotoluene and *o*-chlorotoluene, in different concentration ratios were prepared. The corresponding nine data sets, recorded by HPLC with diode array detector under the same conditions, were used to construct data arrays. This data array was kindly provided by Dr. H.L. Wu (for experimental details see [19]).

4.3. Programs

The program of weighted PARAFAC (WPARAFAC) was kindly supplied by Booksh and coworkers [18]. All the other programs used in this paper were written in-house in the Matlab 5.2 environment and run on a 400 MHz Pentium (Intel) with 64 MB RAM under Window 98 operating system. Random initialization was carried out to start the iterative optimizing procedures of PARAFAC-ALS, WPARAFAC and the proposed method. The optimizing procedures of all the three algorithms are terminated when any one of the following two criterions are satisfied.

$$S^{(n)} = \sum_{k=1}^{K} ||\boldsymbol{X}_{..k} - \hat{\boldsymbol{X}}_{..k}(n)||_F^2, \quad \left| \frac{S^{(n)} - S^{(n-1)}}{S^{(n-1)}} \right| \leq \varepsilon_2 \text{ or MAXIN} \geq 5000$$

Here, n is the current iteration number; ε_2 a preset small value (e.g. 1×10^{-6} in this paper), MAXIN is maximal iteration number allowed.

5. Results and discussions

For fluorescent and HPLC–DAD data arrays, comparisons have been made between PARAFAC-ALS, WPARAFAC and the proposed method. For the convenience of presentation, the proposed method will be simply referred to as MPARAFAC. The fluorescent data array input into PARAFAC-ALS, WPARAFAC and MPARAFAC employing only iterative correcting procedure is assembled by the response matrixes of 10 samples with the Rayleigh scattering being roughly corrected through subtracting the average response matrix of four blank solutions. While the fluorescent data array under the process of MPARAFAC with the two mitigating techniques embedded is composed of the original response matrixes of 10 samples and the mean response matrix of four blank solutions. Since HPLC–DAD data array consists of only the response matrixes of nine samples, and no response matrixes of blank solution are available, PARAFAC-ALS, WPARAFAC and MPARAFAC have to treat the same data array. Hence, the background constraining technique cannot be adopted. Only iterative correcting procedure will be employed in MPARAFAC to mitigate the influence of deviations in HPLC–DAD data array. In WPARAFAC, a parameter, cutoff value for assigning weights, should be preset first. Different cutoff values for both fluorescent and HPLC–DAD data arrays were tried. Only the best results were reported.

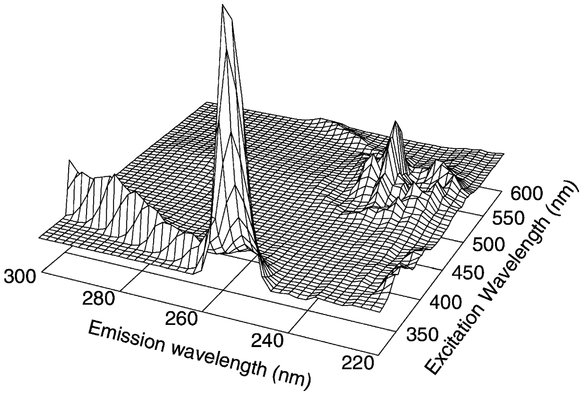

Fig. 1. Accumulated square residue matrix after fitting the excitation–emission fluorescent data arrays by PARAFAC-ALS.

Fig. 1 shows the accumulated square residue matrix after fitting the fluorescent data array by PARAFAC-ALS. It is clear that the Rayleigh scattering in fluorescent data array have not been completely corrected just by subtracting the mean response matrix of blank solutions. The Rayleigh scattering remained causes the data array to deviate from a trlinear model, and hence affect the results of PARAFAC-ALS (Table 1). Therefore, WPARAFAC, background constraining and iterative correcting techniques are employed to mitigate the influence of the embedded Rayleigh scattering. To our surprise, WPARAFAC provided almost the same results as those of PARAFAC-ALS (Table 1). The reason for the consistency between the results of WPARAFAC and PARAFAC might lie in the factor that the average

Table 1
Correlation coefficients between the real concentration profiles and those resolved by PARAFAC-ALS, WPARAFC and MPARAFAC with five extra factors and $\alpha = 5\%$ for fluorescent data array

	PARAFAC-ALS	WPARAFAC	MPARAFAC	MPARAFAC1[a]	MPARAFAC2[b]
Naphthalene	0.9963	0.9963	0.9996	0.9992	0.9956
1-Naphthol	0.9970	0.9971	0.9990	0.9982	0.9982
2-Naphthol	0.9977	0.9977	0.9994	0.9977	0.9993

[a] MPARAFAC1 represents MPARAFAC with only background constraining technique being embedded.
[b] MPARAFAC2 signifies MPARAFAC with only iterative correcting procedure being adopted.

response of four blank solutions has been subtracted from the original data sets of samples. Assigning weights of zero to the pretreated data entities in the area with intensity of the blank above a predetermined cutoff value might not be as effective as the original data sets being used. Such a speculation has at least partly supported by the results of WPARAFAC for the original data sets without any pretreatment. The correlation coefficients between the real concentration profiles and those resolved by WPARAFAC are 0.9323, 0.9783 and 0.9964, respectively. Comparing with those of PARAFAC-ALS for the original data sets (0.9310, 0.9770 and 0.9963, respectively), some improvements over PARAFAC-ALS can be seen. But the results are still not satisfying, which indicates the necessity to subtract the average response from the original data sets of samples. From Table 1, it is obvious that employing any one of the two techniques, background constraining and iterative correcting, alone can enhance the results to some extent. Background constraining technique enhanced the correlation coefficients between the resolved and real concentration vectors of naphthalene and 1-naphthol from 0.9971 and 0.9970 to 0.9992 and 0.9982, respectively, and retained that of 2-naphthol unchanged. While iterative correcting procedure favors 2-naphthol and 1-naphthol. These results at least partly support our assumption that the two techniques can mitigate different types of deviations and they supplement each other. The effectiveness of combining the background constraining and iterative correcting techniques in MPARAFAC can be also demonstrated by the results listed in Table 1. The number of extra factors used in MPARAFAC equals to 5. The value of α is set to 5%. The three correlation coefficients between the estimated concentration profiles by MPARAFAC and the corresponding real ones are 0.9996, 0.9990 and 0.9994, respectively, which are significantly superior to the counterparts of PARAFAC-ALS, i.e. 0.9971, 0.9970 and 0.9977, respectively. It should be noted that such improvements are attained without using any prior knowledge such as non-negativity and unimodelity of chemical loading parameters, and also requiring no extra experiments. It is even unnecessary to assign weights to data entities as the weighted PARAFAC does.

Similar conclusion can also be drawn from the results of HPLC–DAD data array. The chromatograms resolved by PARAFAC-ALS shows negative parts during the region between the 4th and 10th retention time points (Fig. 2a). The difference between resolved spectra and the actual ones are also perceptible (Fig. 2b). Along with the above evidences, the accumulated square residue matrix (Fig. 3) after fitting the HPLC–DAD data array by PARAFAC-ALS suggests the existence of model deviation. Since only small parts of the whole data entities are contaminated by large deviations, the iterative correcting technique might be effective in reducing the influence of model deviations. As expected, the chromatograms and spectra estimated by MPARAFAC are in perfect consistency with the actual ones (Fig. 2a and b). The correlation coefficients between the three concentration vectors obtained by MPARAFAC and the real ones are 0.9992, 0.9992 and 0.9991, respectively (Table 2). Comparisons between the results of PARAFAC-ALS demonstrate the capability of MPARAFAC in treating data arrays with deviations. A similar improvement is also obtained by WPARAFAC. The results of WPARAFAC for the HPLC–DAD data sets are obviously superior to those of PARAFAC-ALS which manifesting it ability to cope with non-linearity. Though both MPARAFAC and WPARAFAC can effectively mitigate the influence of non-linearity,

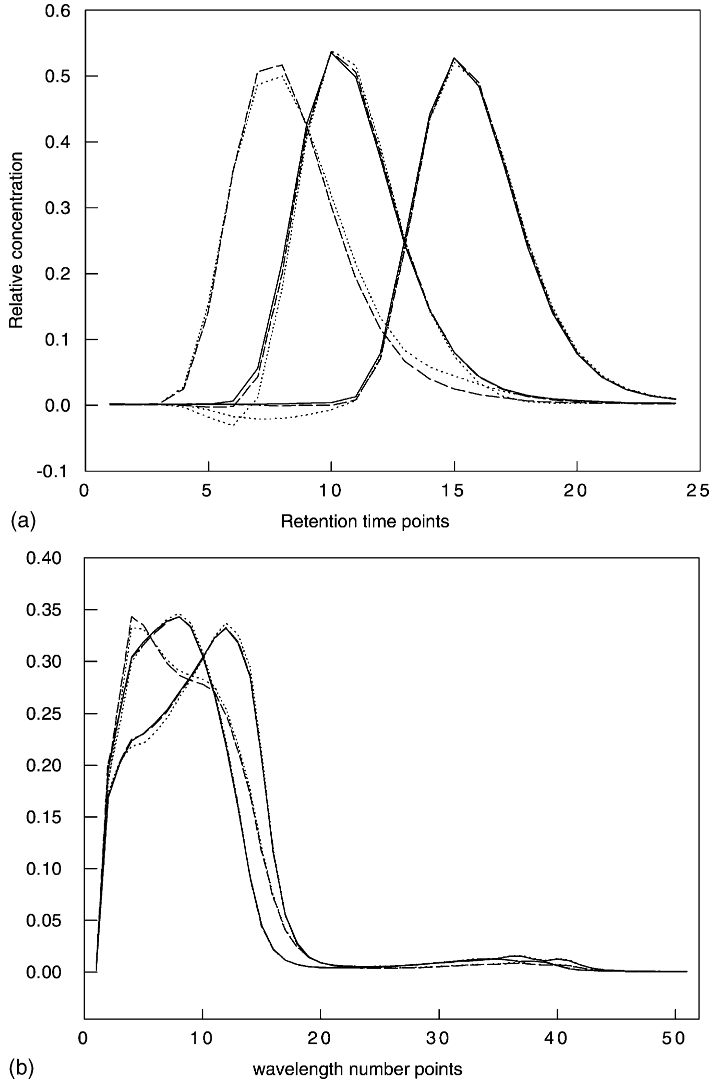

Fig. 2. (a) The chromatograms and (b) ultraviolet visible spectra for HPLC–DAD data array (solid line, real; dotted line, resolved by PARAFAC-ALS; dash line, obtained by MPARAFAC1 with $\alpha = 5\%$).

Table 2
Correlation coefficients between the real concentration profiles and those resolved by PARAFAC-ALS, WPARAFAC and MPARAFAC2 with $\alpha = 5\%$ for HPLC data array

	PARAFAC-ALS	WPARAFAC	MPARAFAC2
o-Dichlorobenzene	0.9900	0.9962	0.9992
p-Chlorotoluene	0.9968	0.9994	0.9992
o-Chlorotulene	0.9977	0.9991	0.9991

MPARAFAC has an advantage of easy parameter setting (which will be discussed in the following sector).

In the implementation of MPARAFAC, there are two parameters to be preset first. One is *m*, the number of extra factor used, the other is α. Theoretically, in order to guarantee a high degree of overlapping between the space spanned by extra factors and that of $\boldsymbol{D}1_{..k}$, a large *m* is preferred. Actually, when *m* is larger than certain value, further increase of *m* will not bring further improvement on the decomposition results.

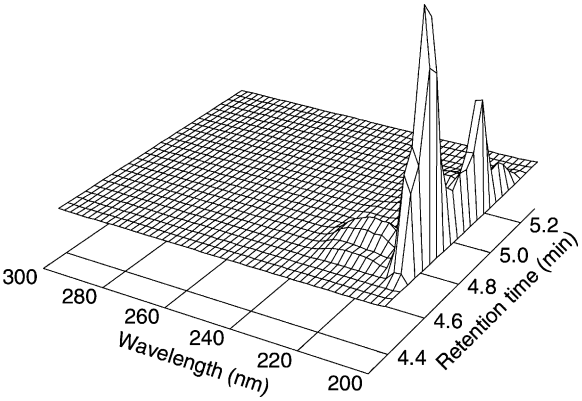

Fig. 3. Accumulated square residue matrix after fitting the HPLC–DAD data arrays by PARAFAC-ALS.

Therefore, a moderate m with which the most part of the response of blank solution (say 95% of the total variance) can be accounted by $\boldsymbol{U}_m\boldsymbol{S}_m\boldsymbol{V}_m^{\mathrm{T}}$ is recommended for most cases. For the fluorescent data array, the results of MPARAFAC show no significant differences for m larger than 3 (Table 3). Hence, satisfactory results can be expected by choosing a moderate m such as 5. As far as α is concerned, the variation of α from 0 to 50% hardly has impact on the final results of both fluorescent and HPLC–DAD data arrays (Tables 4 and 5). This phenomenon results from the iterative property of iterative correcting procedure. It is, therefore, preferable to select a large α (say 50%). Due to the robustness of MPARAFAC to the variation of α, it is recommended to set $\alpha = 50\%$ for general applications. The simplicity of the implementation of MPARAFAC can thus be comparable to PARAFAC-ALS. However, better results might be expected from MPARAFAC.

Table 3
Correlation coefficients between the real concentration profiles and those resolved by MPARAFAC with $\alpha = 5\%$ and different extra factors (m) for fluorescent data array

m	Naphthalene	1-Naphthol	2-Naphthol
2	0.9967	0.9919	0.9991
3	0.9996	0.9987	0.9994
4	0.9997	0.9988	0.9994
5	0.9996	0.9990	0.9994
6	0.9994	0.9989	0.9994
7	0.9995	0.9989	0.9993

Table 4
Correlation coefficients between the real concentration profiles and those resolved by MPARAFAC with five extra factors and different α for fluorescent data array

α (%)	Naphthalene	1-Naphthol	2-Naphthol
50	0.9993	0.9989	0.9998
40	0.9994	0.9989	0.9998
30	0.9995	0.9990	0.9997
20	0.9995	0.9990	0.9995
10	0.9996	0.9990	0.9994
5	0.9996	0.9990	0.9994
0	0.9996	0.9990	0.9993

Table 5
Correlation coefficients between the real concentration profiles and those resolved by MPARAFAC1 with different α for HPLC data array

α (%)	*o*-Dichlorobenzene	*p*-Chlorotoluene	*o*-Chlorotulene
50	0.9993	0.9993	0.9996
40	0.9993	0.9993	0.9996
30	0.9993	0.9992	0.9993
20	0.9992	0.9992	0.9993
10	0.9992	0.9992	0.9990
5	0.9992	0.9992	0.9991
0	0.9991	0.9992	0.9990

6. Conclusions

The background constraining and iterative correcting techniques developed in this contribution are effective in alleviating the influence of model deviations on the chemical loading parameters in trilinear data analysis. Since the two techniques are designed to mitigate different types of deviations, they supplement each other. Their combination, MPARAFAC, can provide further advantages over any one of them. Comparisons of MPARAFAC with PARAFAC-ALS and WPARAFAC for the treatment of fluorescent and HPLC–DAD data arrays demonstrated that the influence of model deviation on the chemical loading parameters could be significantly abated by MPARAFAC. It should be noted that such improvements are attained without using any prior knowledge such as non-negativity and unimodelity of chemical loading parameters, and requiring also no extra experiments. It is even unnecessary to assign weights to data entities as the weighted PARAFAC does. The cost is only a small increase in computation time.

In MPARAFAC, there are two controlling parameters, i.e. the number of extra factor (m) used in background constraining and α in iterative correcting procedure. Fortunately, the performance of MPARAFAC is robust to the variations of these two parameters. Hence, they cause no trouble in implementing MPARAFAC. For general applications, it is recommended to employ a moderate m with which most part of the response of blank solution can be accounted by $\boldsymbol{U}_m\boldsymbol{S}_m\boldsymbol{V}_m^{\mathrm{T}}$ and set $\alpha = 50\%$. The simplicity of the implementation of MPARAFAC is thus comparable to PARAFAC-ALS.

Acknowledgements

The authors would like to thank the National Nature Science Foundation of China for financial support (Grant Nos. 29735150 and 20075006).

Appendix A

Implementation guideline for combining background constraining and iterative correcting techniques in mitigating model deviations

1. Determining the number of chemical species (F) in the three-way data array $\boldsymbol{X}$.
2. Decomposing the response matrix of blank solution, $\boldsymbol{X}_{\text{blank}}$ by SVD: $\boldsymbol{X}_{\text{blank}} = USV^{\mathrm{T}}$, and then constraining the loading parameters of m extra factors in two modes of the PARAFAC model to be $\boldsymbol{U}_m$ and $\boldsymbol{V}_m$ (consisting of the first m columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively).
3. Decomposing the original data array $\boldsymbol{X}$ through minimizing the $\sum_{k=1}^{K}||\boldsymbol{X}_{..k} - (\boldsymbol{A}\,\mathrm{diag}(\boldsymbol{c}_k)\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{U}_m\,\mathrm{diag}(\tilde{\boldsymbol{c}}_k)\boldsymbol{V}_m^{\mathrm{T}})||_F^2$ by constrained PARAFAC-ALS with $F + m$ loading factors to obtain the fitted values $\hat{\boldsymbol{X}}_{..k}$ $(k = 1, 2, \ldots, K)$, fitness value and accumulated square residue matrix $\boldsymbol{R} = \sum_{k=1}^{K}(\boldsymbol{X}_{..k} - \hat{\boldsymbol{X}}_{..k})^2$ (see Appendix B).
4. Creating a new data array $\boldsymbol{X}_{\text{new}}$ through replacing the data entities in $\boldsymbol{X}_{..k}$ $(k = 1, 2, \ldots, K)$ whose counterparts in accumulated square residue matrix $\boldsymbol{R}$ are larger than a predefined threshold r_{cut} with the corresponding fitted values in $\hat{\boldsymbol{X}}_{..k}$ $(k = 1, 2, \ldots, K)$, and retaining the rest data entities in $\boldsymbol{X}_{..k}$ $(k = 1, 2, \ldots, K)$ unchanged.
5. Decomposing the newly created data array $\boldsymbol{X}_{\text{new}}$ by constrained PARAFAC-ALS with $F + m$ loading factors to obtain new fitted values, fitness value and accumulated square residue matrix.
6. Repeating the steps 4–5, until the difference between fitness values of two successive decompositions reaches a predefined small value ε_1 (1×10^{-6}, for instance).

Appendix B

Implementation guideline for decomposing data array $\boldsymbol{X}$ through minimizing the $\sum_{k=1}^{K}||\boldsymbol{X}_{..k} - (\boldsymbol{A}\,\mathrm{diag}(\boldsymbol{c}_k)\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{U}_m\,\mathrm{diag}(\tilde{\boldsymbol{c}}_k)\boldsymbol{V}_m^{\mathrm{T}})||_F^2$ by constrained PARAFAC-ALS

1. $\boldsymbol{X}_A = [\boldsymbol{X}_{..1}|\boldsymbol{X}_{..2}|\cdots|\boldsymbol{X}_{..K}]$, $\boldsymbol{X}_B = [\boldsymbol{X}_{..1}^{\mathrm{T}}|\boldsymbol{X}_{..2}^{\mathrm{T}}|\cdots|\boldsymbol{X}_{..K}^{\mathrm{T}}]$;
2. Randomly initialize loading matrixes $\boldsymbol{A}$ and $\boldsymbol{B}$;
3. $\boldsymbol{A}_{\text{comb}} = [\boldsymbol{A}|\boldsymbol{U}_m]$, $\boldsymbol{B}_{\text{comb}} = [\boldsymbol{B}|\boldsymbol{V}_m]$;
4. $\boldsymbol{Z} = ((\boldsymbol{A}_{\text{comb}}^{\mathrm{T}}\boldsymbol{A}_{\text{comb}}) \circ (\boldsymbol{B}_{\text{comb}}^{\mathrm{T}}\boldsymbol{B}_{\text{comb}}))^{+}$ (where '$\circ$' is element wise, or Hadamard product); $\boldsymbol{c}_{\text{comb},k} = [\boldsymbol{c}_k|\tilde{\boldsymbol{c}}_k] = (\boldsymbol{Z}\,\mathrm{diag}(\boldsymbol{A}_{\text{comb}}^{\mathrm{T}}\boldsymbol{X}_{..k}\boldsymbol{B}_{\text{comb}}))^{\mathrm{T}}$, $k = 1, \ldots, K$;
5. $\boldsymbol{Y}_A = [\mathrm{diag}(\boldsymbol{c}_{\text{comb},1})\boldsymbol{B}_{\text{comb}}^{\mathrm{T}}|\mathrm{diag}(\boldsymbol{c}_{\text{comb},2})\boldsymbol{B}_{\text{comb}}^{\mathrm{T}}|\cdots|\mathrm{diag}(\boldsymbol{c}_{\text{comb},K})\boldsymbol{B}_{\text{comb}}^{\mathrm{T}}]$;

$$\boldsymbol{Y}_{A,(F+1):(F+m)} = \begin{bmatrix} \boldsymbol{y}_{A,F+1} \\ \boldsymbol{y}_{A,F+2} \\ \vdots \\ \boldsymbol{y}_{A,F+m} \end{bmatrix},$$

$$\boldsymbol{Y}_{A,1:F} = \begin{bmatrix} \boldsymbol{y}_{A,1} \\ \boldsymbol{y}_{A,2} \\ \vdots \\ \boldsymbol{y}_{A,F} \end{bmatrix};$$

($\boldsymbol{y}_{A,f}$, $(f = 1, 2, \cdots, F+m)$ is the fth row of $\boldsymbol{Y}_A$);
$\boldsymbol{A} = (\boldsymbol{X}_A - \boldsymbol{U}_m\boldsymbol{Y}_{A,(F+1):(F+m)})\boldsymbol{Y}_{A,1:F}^{\mathrm{T}}(\boldsymbol{Y}_{A,1:F}\boldsymbol{Y}_{A,1:F}^{\mathrm{T}})^{+}$;

6. $\boldsymbol{Y}_B = [\mathrm{diag}(\boldsymbol{c}_{\text{comb},1})\boldsymbol{A}_{\text{comb}}^{\mathrm{T}}|\mathrm{diag}(\boldsymbol{c}_{\text{comb},2})\boldsymbol{A}_{\text{comb}}^{\mathrm{T}}|\cdots|\mathrm{diag}(\boldsymbol{c}_{\text{comb},K})\boldsymbol{A}_{\text{comb}}^{\mathrm{T}}]$;

$$\boldsymbol{Y}_{B,(F+1):(F+m)} = \begin{bmatrix} \boldsymbol{y}_{B,F+1} \\ \boldsymbol{y}_{B,F+2} \\ \vdots \\ \boldsymbol{y}_{B,F+m} \end{bmatrix},$$

$$\boldsymbol{Y}_{B,1:F} = \begin{bmatrix} \boldsymbol{y}_{B,1} \\ \boldsymbol{y}_{B,2} \\ \vdots \\ \boldsymbol{y}_{B,F} \end{bmatrix},$$

($\boldsymbol{y}_{B,f}$, $(f = 1, 2, \cdots, F+m)$ is the *f*th row of $\boldsymbol{Y}_B$);
$\boldsymbol{B} = (\boldsymbol{X}_B - \boldsymbol{V}_m \boldsymbol{Y}_{B,(F+1):(F+m)})\boldsymbol{Y}^{\mathrm{T}}_{B,1:F}(\boldsymbol{Y}_{B,1:F}\boldsymbol{Y}^{\mathrm{T}}_{B,1:F})^{+}$;

7. Updating $\boldsymbol{c}_{\mathrm{comb},k}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ according to steps 3–6, until certain stop criterion has been reached.

References

[1] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) 782A.

[2] R.A. Harshman, UCLA Working Papers in Phonetics, vol. 16, 1970, p. 1.

[3] J.D. Carrol, J. Chang, Psychometrika 35 (1970) 283.

[4] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149.

[5] E. Sanchez, B.R. Kowalski, J. Chemom. 4 (1990) 29.

[6] S. Wold, P. Geladi, K. Esbensen, J. Öhman, J. Chemom. 1 (1987) 41.

[7] R. Bro, J. Chemom. 10 (1996) 367.

[8] M. Linder, R. Sundberg, Chemom. Intell. Lab. Syst. 42 (1998) 159.

[9] K.S. Booksh, Z. Lin, Z. wang, B.R. Kowalski, Anal. Chem. 66 (1994) 2561.

[10] K. Smilde, R. Tauler, J.M. Henshaw, L.W. Burgess, B.R. Kowalski, Anal. Chem. 66 (1994) 3345.

[11] K. Smilde, Y. Wang, B.R. Kowalski, J. Chemom. 8 (1994) 21.

[12] K.S. Booksh, J.M. Henshaw, L.W. Burgess, B.R. Kolwaski, J. Chemom. 9 (1995) 263.

[13] K.S. Booksh, B.R. Kowalski, Anal. Chim. Acta 348 (1997) 1.

[14] R. Bro, S.D. Jong, J. Chemom. 11 (1997) 393.

[15] R. Bro, N.D. Sidiropoulos, J. Chemom. 12 (1998) 223.

[16] Z.P. Chen, H.L. Wu, Y. Li, R.Q. Yu, Anal. Chim. Acta 423 (2000) 187.

[17] R. Bro, Multi-way analysis in the food industry, Ph.D. Thesis.

[18] G.G. Anderson, B.K. Dable, K.S. Booksh, Chemom. Intell. Lab. Syst. 49 (1999) 195.

[19] H.L. Wu, M. Shibukawa, K. Oguma, J. Chemom. 12 (1998) 1.