ELSEVIER

Analytica Chimica Acta 423 (2000) 187–196

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

# Novel constrained PARAFAC algorithm for second-order linear calibration

Zeng-Ping Chen, Hai-Long Wu, Yang Li, Ru-Qin Yu*

*College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, PR China*

Received 28 March 2000; received in revised form 4 July 2000; accepted 31 July 2000

**Abstract**

A novel constrained PARAFAC was designed to mitigate the influence of model deviations on the predictive accuracy in second-order calibration. It combines the two steps of decomposition and calibration of PARAFAC in second-order calibration into just one step by imposing the concentrations of the calibration samples on the model parameters in the third mode as constraints. Such a scheme not only simplifies the whole calibration procedure, but also renders some extra advantages over PARAFAC when model deviations such as Rayleigh scattering in fluorescence spectroscopy present in measurements. The results for fluorescent and HPLC data arrays have demonstrated that the relative deviations of the predicted concentrations could be significantly reduced when the constrained version of PARAFAC replaces PARAFAC to do second-order linear calibration. 

*Keywords:* Second-order linear calibration; PARAFAC; Deviations from trilinear model; Alternating least squares algorithm

## 1. Introduction

Due to the so-called second-order advantage, i.e. calibration can be performed in the presence of unknown interference(s) [1], second-order linear calibration has been one of the most active areas in analytical chemometrics. Among all the methods for second-order calibration [2–12], PARAFAC [2–4] may be the most widely used one, though it has often been criticized for its low convergence rate and requirement of a correct estimation of the underlying factors. The popularity of PARAFAC in chemometrics may lie in the model uniqueness and the optimality of its results in least square sense.

The most important prerequisite for the successful application of PARAFAC is that the data arrays should satisfy the trilinear assumption (see Section 3 for details). Such an assumption, however, may be violated in practice. Rayleigh scattering in fluorescence spectroscopy is but one instance, where slight model inadequacy can cause the estimated model parameters to be misleading. Booksh and Kowalski have cataloged five broad classes of deviations from the trilinear model that are common with chemical data [13]. Theoretically, there are two approaches to improve the reliability of the results resolved by PARAFAC in the presence of deviations. One is to do some pretreatment to the data arrays before the employment of PARAFAC. The other is to impose some reasonable constraints on the model parameters. The weighted PARAFAC of Anderson et al. [14] may be classified into the first category, for assigning weights to the data elements is essentially a kind of pretreatment of the data array. As for the second approach, one may argue that constraining the PARAFAC model is superfluous

* Corresponding author.
*E-mail address:* rqyu@mail.hunu.edu.cn (R.-Q. Yu).

0003-2670/00/$ – see front matter 
PII: S0003-2670(00)01110-7

**Nomenclature**

| | |
|---|---|
| $a_{if}$, $b_{jf}$ and $c_{kf}$ | the *if*th, *jf*th and *kf*th elements of the three underlying loading $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, respectively |
| $\boldsymbol{A}_{I\times F}$, $\boldsymbol{B}_{J\times F}$, $\boldsymbol{C}_{K\times F}$ | the three underlying loading matrices of $\underline{\boldsymbol{X}}$ with dimensions of $I \times F$, $J \times F$, $K \times F$, respectively (which will be simply represented by $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, respectively, in this paper) |
| diag($\boldsymbol{c}_k$) | diagonal matrix with elements equal to the *k*th row of $\boldsymbol{C}$ |
| $e_{ijk}$ | the *ijk*th element of the three-way residue array $\underline{\boldsymbol{E}}$ |
| $\boldsymbol{E}_{..k}$ | the *k*th frontal slice of the three-way residue array $\underline{\boldsymbol{E}}$ |
| $F$ | the number of underlying factors, i.e. the total number of detectable species, including the components of interest and interference(s) as well as background |
| $I$, $J$, $K$ | the dimensions of different modes in three-way array |
| $N$ | the number of factors used in calculation |
| $x_{ijk}$ | the *ijk*th element of three-way array $\underline{\boldsymbol{X}}$ |
| $\boldsymbol{X}_{..k} = \boldsymbol{A}_{I\times F}\text{diag}(\boldsymbol{c}_k)\boldsymbol{B}^{\mathrm{T}}_{J\times F} + \boldsymbol{E}_{..k}$ | the *k*th frontal slice of three-way array $\underline{\boldsymbol{X}}$ |

since the model in itself is unique. However, the existence of possible model inadequacy can cause the final results of PARAFAC to be misleading, which justifies the necessity of reasonable constraints. Among all the constraints used in PARAFAC model, non-negativity is the most common one [15], since the model parameters in chemistry are often with non-negative characteristic. For data arrays generated by HPLC-DAD, sometimes the unimodal property of the chromatographic profiles of chemical compounds can also be an effective constraint to enhance the quality of the final results of PARAFAC [16]. A variety of other effective constraints and corresponding algorithms have been summarized and supplied by Bro [17]. The successes of the introduction of constraints on PARAFAC model as well as other multi-way chemometric models [18,19] imply the advantages of incorporating known information into the model building.

In second-order calibrations, other than the non-negativity of the model parameters and unimodal property of chromatographic profiles, the concentrations of the calibration samples are also known. In PARAFAC for second-order calibration, all the model parameters (including the concentration matrix $\boldsymbol{C}$) of the trilinear model are regarded to be unknown and initialized randomly. After the model parameters being estimated by the so-called alternating least squares algorithm, the estimated concentrations are finally regressed on the known concentrations to predict the unknown samples. The known concentration matrix of the calibration samples is not utilized in the trilinear decomposition step. In our opinion, the incorporation of the concentration matrix of the calibration samples in trilinear decomposition may render some advantages when the data arrays deviate from a strictly trilinear model. Therefore, in this paper, a novel constrained PARAFAC algorithm is proposed and its performance has been demonstrated by two kinds of real data arrays, one is produced by fluorescence spectrometer, the other, by HPLC-DAD.

## 2. Nomenclature

Throughout this paper, scalars are represented by lower-case italics, vectors by bold lower-case characters, bold capitals designate two-way matrices and underlined bold capitals symbolize three-way arrays. Before reading the following parts of this paper, readers are recommended to refer the nomenclature for detailed information.

## 3. Theory

PARAFAC is actually a trilinear model:

$$x_{ijk} = \sum_{f=1}^{F} a_{if}b_{jf}c_{kf} + e_{ijk}, \quad i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K \tag{1}$$

The matrix form of this trilinear model is as follows:

$$\boldsymbol{X}_{..k} = \boldsymbol{A}\,\mathrm{diag}(c_k)\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{E}_{..k}, \quad k = 1, \ldots, K \qquad (2)$$

In chemistry, the columns of loading matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are always assigned with certain physical meanings, i.e. excitation and emission spectra in fluorescence spectrometry, and chromatographic profiles and ultraviolet-visible spectra in HPLC-DAD. The columns of loading matrix $\boldsymbol{C}$ often represent concentrations of components in mixtures.

In second-order calibration, PARAFAC first treats all the loading matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ as unknowns, these loading matrices are then estimated by alternating least squares algorithm after random initialization. The concentrations of the components in unknown samples are predicted by regressing the loading matrix $\boldsymbol{C}$ estimated by PARAFAC on the known concentrations of the calibration samples. It is evident that the known concentrations of the calibration samples have not been used in the trilinear decomposition procedure.

For data arrays which strictly follow a trilinear model, no matter whether the known concentrations being utilized in the decomposition procedure or not, the final results of both approaches will have no significant differences. Unfortunately, the deviations of data arrays from trilinear model are common in practice. For data arrays contaminated by deviations (thereafter, 'deviations' denotes the 'deviations from trilinear model', otherwise, specified), the response can be decomposed into four parts as follows:

$$\underline{\boldsymbol{X}} = \underline{\boldsymbol{X}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \underline{\boldsymbol{D}}_{\mathrm{independent}} + \underline{\boldsymbol{D}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \underline{\boldsymbol{E}} \qquad (3)$$

where $\underline{\boldsymbol{X}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ contains the linear responses of chemical species. $\underline{\boldsymbol{D}}_{\mathrm{independent}}$ represents deviations independent of the loading matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ which are associated with chemical species. Generally, it does not follow a trilinear model. $\underline{\boldsymbol{D}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ denotes deviations which have certain relationship with the three loading matrix $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$. $\underline{\boldsymbol{E}}$ signifies random noise.

If $\underline{\boldsymbol{D}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \neq 0$, PARAFAC cannot be at least directly applied to such data arrays. Some subtle pretreatment is necessary. For this kind of data arrays, there is no versatile solution to improve the reliability of the estimated model parameters. For data arrays satisfying the condition $\underline{\boldsymbol{D}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = 0$, the influence of $\underline{\boldsymbol{D}}_{\mathrm{independent}}$ on the accuracy of the estimated model parameters can be reduced by putting some constraints on the model parameters [15,16]. Like non-negativity and unimodality, reasonable constraints should be based on information known in a priori. In second-order linear calibration, the concentration matrix of the calibration samples is known, it can also be used as constraint in trilinear decomposition. Therefore, a novel constrained PARAFAC algorithm is designed (*For the convenience of presentation, the proposed constrained PARAFAC algorithm will be just simply represented by CPARAFAC, though it actually does not deserve a new acronym*).

Suppose $\underline{\boldsymbol{X}} = \{\boldsymbol{X}_{..1}, \boldsymbol{X}_{..2}, \ldots, \boldsymbol{X}_{..r}, \boldsymbol{X}_{..r+1}, \ldots, \boldsymbol{X}_{..K}\}$, $\boldsymbol{X}_{..1}, \boldsymbol{X}_{..2}, \ldots, \boldsymbol{X}_{..r}$ are response matrices of $r$ calibration samples. $\boldsymbol{X}_{..r+1}, \ldots, \boldsymbol{X}_{..K}$ are response matrices of $K - r$ samples under test. Therefore

$$\boldsymbol{X}_{..k} = [\boldsymbol{A}_{\mathrm{calibration}}, \boldsymbol{A}_{\mathrm{interf}}] \times \begin{bmatrix} \mathrm{diag}(\mathbf{c}_{k,\mathrm{calibration}}) & 0 \\ 0 & \mathrm{diag}(\mathbf{c}_{k,\mathrm{interf}}) \end{bmatrix} \times [\boldsymbol{B}_{\mathrm{calibration}}, \boldsymbol{B}_{\mathrm{interf}}]^{\mathrm{T}}, \quad k = 1, \ldots, r$$

$$\boldsymbol{X}_{..k} = [\boldsymbol{A}_{\mathrm{calibration}}, \boldsymbol{A}_{\mathrm{interf}}] \times \begin{bmatrix} \mathrm{diag}(\mathbf{c}_{k,\text{sought-for}}) & 0 \\ 0 & \mathrm{diag}(\mathbf{c}_{k,\mathrm{interf}}) \end{bmatrix} \times [\boldsymbol{B}_{\mathrm{calibration}}, \boldsymbol{B}_{\mathrm{interf}}]^{\mathrm{T}}, \quad k = r+1, \ldots, K$$

Here, $\boldsymbol{A}_{\mathrm{calibration}}$ and $\boldsymbol{B}_{\mathrm{calibration}}$ are the two loading matrices of the known components in calibration samples, while $\boldsymbol{A}_{\mathrm{interf}}$ and $\boldsymbol{B}_{\mathrm{interf}}$ are those of possible unknown interference(s) in mixtures. $\boldsymbol{c}_{k,\mathrm{calibration}}$ is a vector with elements equal to the concentrations of the components in the $k$th calibration sample. $\boldsymbol{c}_{k,\text{sought-for}}$ contains the concentrations to be determined of the interested components in the $k$th unknown sample. $\boldsymbol{c}_{k,\mathrm{interf}}$ represents the concentrations of possible interference(s) in the $k$th sample. In CPARAFAC, with the $\boldsymbol{c}_{k,\mathrm{calibration}}$ ($k = 1, \ldots, r$) fixed, $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{c}_{k,\text{sought-for}}$ ($k = r+1, \ldots, K$) are obtained by alternating least squares algorithm after random initialization.

With the incorporation of the concentrations of the calibration samples into the decomposition procedure, the response modes of the sought-for species are constrained. Hence, the influence of the possible deviations on the sought-for species can be mitigated. Other than the potential capability of coping with deviations, the proposed algorithm has another feature

of combing the two-step calibration procedure into just one step. A great increase in the convergence rate has also been observed for CPARAFAC. Since this paper focuses mainly on the ability of CPARAFAC to cope with model inadequacy, its other properties such as fast convergence will be omitted.

## 4. Experimental

### 4.1. Programs

All the programs used in the 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.

### 4.2. Excitation–emission fluorescent data arrays

#### 4.2.1. Reagents and stock solutions

All reagents used were of analytical grade. Stock solutions of L-phenylalanine (0.011 mg $ml^{-1}$), L-tyrosine (0.012 mg $ml^{-1}$) and L-tryptophan (0.011 mg $ml^{-1}$) were prepared by accurately weighting correspondingly appropriate amount of reagents and dissolving them in distilled water (Table 1). A total of 10 working solutions with different concentration ratios of the three components were made by taking appropriate volumes of stock solutions and 2 ml of buffer solution ($NaOH–KH_2PO_4$, pH = 7.4) into a 25 ml volumetric flask and then making them to 25 ml with distilled water.

Table 1
Compositions of the 10 samples for fluorescent data arrays

| No. | Concentration (mg $ml^{-1}$) | | |
|---|---|---|---|
| | Tyrosine | Tryptophan | Phenylalanine |
| $1^{\#}$ | $8.80 \times 10^{-4}$ | 0 | 0 |
| $2^{\#}$ | 0 | $19.20 \times 10^{-5}$ | 0 |
| $3^{\#}$ | 0 | 0 | $4.40 \times 10^{-3}$ |
| $4^{\#}$ | $4.40 \times 10^{-4}$ | $9.60 \times 10^{-5}$ | 0 |
| $5^{\#}$ | 0 | $9.60 \times 10^{-5}$ | $2.20 \times 10^{-3}$ |
| $6^{\#}$ | $4.40 \times 10^{-4}$ | 0 | $2.20 \times 10^{-3}$ |
| $7^{\#}$ | $4.40 \times 10^{-4}$ | $9.60 \times 10^{-5}$ | $2.20 \times 10^{-3}$ |
| $8^{\#}$ | $8.80 \times 10^{-4}$ | $9.60 \times 10^{-5}$ | $2.20 \times 10^{-3}$ |
| $9^{\#}$ | $4.40 \times 10^{-4}$ | $19.20 \times 10^{-5}$ | $2.20 \times 10^{-3}$ |
| $10^{\#}$ | $4.40 \times 10^{-4}$ | $9.60 \times 10^{-5}$ | $4.40 \times 10^{-3}$ |

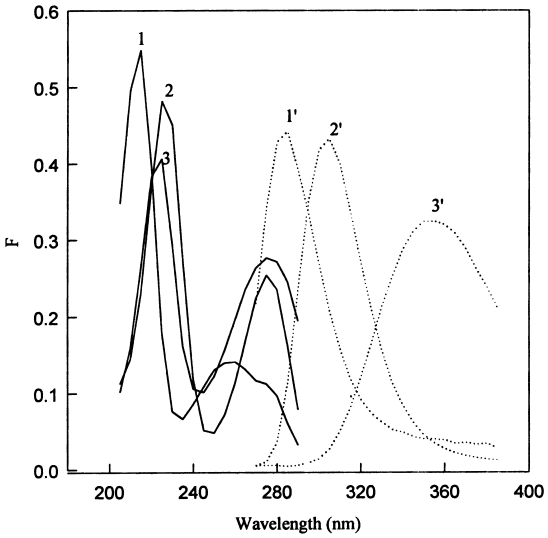


Fig. 1. The excitation (solid line) and emission (dotted line) spectra of phenylalanine (1-1′), tyrosine (2-2′) and tryptophan (3-3′) for the excitation–emission fluorescent data array.

#### 4.2.2. Apparatus

The excitation–emission response matrices of all the samples plus a blank solution were recorded by a HITACHI 850 fluorescence spectrophotometer scanning at 240 nm $min^{-1}$ with excitation and emission wavelength ranging from 205 to 290 nm and from 270 to 385 nm at an interval of 5 nm, respectively. The excitation and emission spectra of phenylalanine, tyrosine and tryptophan are depicted in Fig. 1. The slit width in both excitation and emission monochromators is 10 nm.

#### 4.2.3. Data arrays

Rayleigh scattering in all response matrices was roughly corrected just by subtracting the response matrix of the blank solution. Data arrays were assembled by the Rayleigh scattering corrected response matrices.

### 4.3. 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 concentrations of three compounds in the nine samples were listed in Table 2. The corresponding nine data

Table 2
Compositions of the nine samples for HPLC data arrays

| No. | Concentration ($\mu g\,ml^{-1}$) | | |
|---|---|---|---|
| | *o*-Dichlorobenzene | *p*-Chlorotoluene | *o*-Chlorotulene |
| $1^{\#}$ | 0.0 | 75.6 | 0.0 |
| $2^{\#}$ | 0.0 | 0.0 | 91.2 |
| $3^{\#}$ | 0.0 | 50.4 | 30.4 |
| $4^{\#}$ | 0.0 | 25.2 | 60.8 |
| $5^{\#}$ | 15.2 | 12.6 | 152.0 |
| $6^{\#}$ | 152.0 | 12.6 | 15.2 |
| $7^{\#}$ | 60.8 | 25.2 | 91.2 |
| $8^{\#}$ | 91.2 | 50.4 | 30.4 |
| $9^{\#}$ | 30.4 | 75.6 | 60.8 |

sets, recorded by HPLC with diode array detector at same conditions, were used to construct data arrays. For detail information of this experiments readers please refer to [20].

## 5. Results and discussions

### 5.1. The implementation of CPARAFAC and PARAFAC

Random initialization was carried out to start the iterative optimizing procedures of both CPARAFAC and PARAFAC. The optimizing procedures are terminated when the following criterion reaches certain threshold $\varepsilon$ ($\varepsilon = 1 \times 10^{-6}$ in present paper).

$$\mathrm{SSR}^{(m)} = \sum_{k=1}^{K} \|\boldsymbol{X}_{..k} - \boldsymbol{A}^{(m)}\mathrm{diag}((\boldsymbol{c}^{(m)})_k)\boldsymbol{B}^{(m)^{\mathrm{T}}}\|_F^2$$

$$\left|\frac{\mathrm{SSR}^{(m)} - \mathrm{SSR}^{(m-1)}}{\mathrm{SSR}^{(m-1)}}\right| \leq \varepsilon$$

Where, SSR is the residual sum of squares, $m$ is the current iteration number.

As for $N$, the number of factors used in calculation, both CPARAFAC and PARAFAC take the same value for the same data array. To be specific, $N$ took value of 3 for fluorescent data arrays. For HPLC data arrays, both 3 and 4 were tried.

The constraint on the loading matrix $\boldsymbol{C}$ in CPARAFAC is set as follows:

$$\boldsymbol{C} = [\boldsymbol{C}_1 \quad \boldsymbol{C}_2 \quad \boldsymbol{C}_3] \quad \boldsymbol{C}_1 = \begin{bmatrix} \boldsymbol{C}_{\text{calibration}} \\ \boldsymbol{C}_{\text{sought-for}} \end{bmatrix}$$

$$\boldsymbol{C}_2 = \begin{bmatrix} 0 \\ \boldsymbol{C}_{\text{chemical-interference}} \end{bmatrix}$$

where $\boldsymbol{C}_{\text{calibration}}$ is the known concentration matrix of the compounds in calibration samples. $\boldsymbol{C}_{\text{sought-for}}$ contains the concentrations of the sought-for compounds in sample under test, which will be calculated by CPARAFAC. $\boldsymbol{C}_{\text{chemical-interference}}$ signifies the contents of chemical interferences in samples under test, which do not present in calibration samples. $\boldsymbol{C}_3$ denotes the degrees of the influences of other possible deviations.

### 5.2. Excitation-emission fluorescent data arrays

In fluorescence spectroscopy the Rayleigh scattering is one of the important factors which deviate the data arrays from a strict trilinear model. Though its influence can be partly corrected by subtracting the response matrix of a blank solution from those of the samples. However, due to the ingredient difference between samples and blank solution, the influence of Rayleigh scattering can not be thoroughly eliminated, which will affect the reliability of the results of PARAFAC. Therefore, some constraints on the model parameters may render some advantages. For instance, Fig. 2 shows the excitation and emission spectra of tyrosine and tryptophan resolved by PARAFAC and CPARAFAC for data array composed of samples $1^{\#}$, $2^{\#}$, $4^{\#}$ and $8^{\#}$ with sample $1^{\#}$, $2^{\#}$ and $4^{\#}$ as calibration set. Obvious differences between the emission spectra resolved by CPARAFAC and those by PARAFAC can be observed. The emission spectrum of tryptophan resolved by PARAFAC shows a negative part in the range 270–300 nm. While the emission spectra of tryptophan and tyrosine obtained by CPARAFAC are in good consistence with the real ones. The difference between the residual sum of squares of PARAFAC ($1.568 \times 10^4$) and that of CPARAFAC ($2.0828 \times 10^4$) indicates that the relatively poor results of PARAFAC may be caused by the over-fitting of deviations. As in multivariate calibration using PCA and PLS, it is also the case in second order calibration that better fitting does not necessarily mean better predictive ability. By imposing constraint on the loading matrix $\boldsymbol{C}$, CPARAFAC tries to avoid modeling

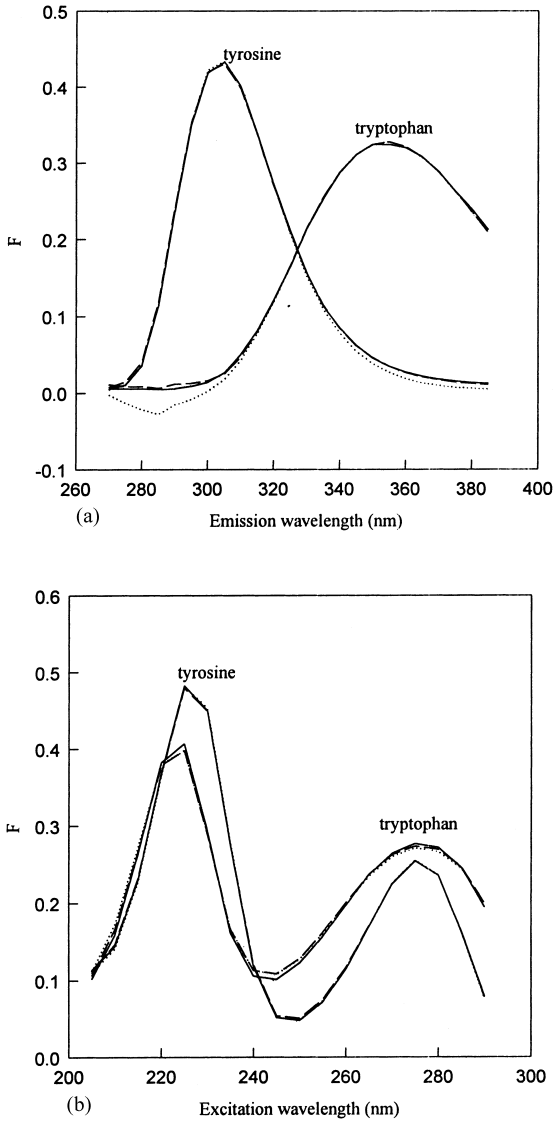


Fig. 2. The emission (a) and excitation (b) spectra of tyrosine and tryptophan for the excitation–emission fluorescent data array assembled by samples 1, 2, 4 and 8 (solid line — real, dotted line — resolved by PARAFAC, dash line — obtained by CPARAFAC with sample 8 as unknown).

deviations, and hence can improve the accuracy of prediction.

The quantitative results of CPARAFA and PARAFAC for fluorescent data arrays are listed in Tables 3–5. In Table 3, samples $1^{\#}$, $3^{\#}$ and $6^{\#}$ were taken as calibration set to predict the contents of tyrosine and phenylalanine in samples $7^{\#}$, $8^{\#}$, $9^{\#}$ and $10^{\#}$. The concentration of tyrosine in sample $8^{\#}$, predicted by PARAFAC is satisfactory with a relative deviation of 1.0%. Unfortunately, the content of phenylalanine in the same sample determined by PARAFAC shows large deviation (10.5%) from the expected one. For samples $7^{\#}$, $9^{\#}$ and $10^{\#}$, the concentrations of both tyrosine and phenylalanine determined by PARAFAC have relatively large deviations. These phenomena may be resulted from the model deviations ($\underline{\boldsymbol{D}}_{\text{independent}}$) such as Rayleigh scattering, which may be reduced by CPARAFAC. As expected, the relative large deviations are greatly reduced by exerting constraint of the known concentrations of the components in calibration samples on the loading matrix $\boldsymbol{C}$. Specifically, the large deviations of 10.5 and 10.0% for phenylalanine in sample $8^{\#}$ and sample $10^{\#}$ are reduced to 3.6 and 3.2%, respectively. Those of tyrosine in samples $7^{\#}$, $9^{\#}$ and $10^{\#}$ decreased from 7.5, 6.1 and 9.6% to 5.2, 3.6 and 5.7%, respectively. Except for tyrosine in samples $8^{\#}$, all the relative deviations of the concentrations calculated by CPARAFAC are significantly smaller than those of PARAFAC. When samples $1^{\#}$, $2^{\#}$ and $4^{\#}$ are regarded as calibration set to determine the contents of tyrosine and tryptophan in samples $7^{\#}$, $8^{\#}$ and $9^{\#}$, $10^{\#}$ (Table 4), similar results are observed. Though the relative deviations of CPARAFAC for tyrosine in $7^{\#}$, $8^{\#}$ and $9^{\#}$ increased slightly compared with those of PARAFAC, such increases are reasonable and acceptable, since relatively large decreases in the relative deviations for tryptophan are obtained (from 10.3 to 3.1%, and 4.9 to 0.1%). Moreover, the increases are so small that the qualities of the results see no obvious deterioration. The preferable feature of CPARAFAC — capable of reducing large relative deviations to acceptable range, is also observed in Table 5, which lists the results of CPARAFAC and PARAFAC for samples $4^{\#}$–$10^{\#}$ with calibration sets assembled by all samples from $1^{\#}$ to $10^{\#}$ except the sample under test.

### 5.3. HPLC data arrays

Table 6 presents the results of CPARAFAC and PARAFAC for HPLC-DAD data arrays. A first glance of the results will give one an impression that the relative deviations for HPLC data arrays are much larger than those for fluorescent data arrays. This phenomenon may be partly attributed to the presence

Table 3
The concentrations and their relative deviations, predicted by PARAFAC and CPARAFAC with fluorescent samples 1#, 3#, and 6# as calibration set

| No. | Tyrosine ($10^{-4}$ mg ml$^{-1}$) | | | Phenylalanine ($10^{-3}$ mg ml$^{-1}$) | | |
|---|---|---|---|---|---|---|
| | PARAFAC | CPARAFAC | Known | PARAFAC | CPARAFAC | Known |
| 7# | 4.07 (7.5%[a]) | 4.17 (5.2%) | 4.40 | 2.05 (6.8%) | 2.18 (0.9%) | 2.20 |
| 8# | 8.89 (1.0%) | 8.69 (1.3%) | 8.80 | 1.97 (10.5%) | 2.12 (3.6%) | 2.20 |
| 9# | 4.13 (6.1%) | 4.24 (3.6%) | 4.40 | 1.98 (10.0%) | 2.13 (3.2%) | 2.20 |
| 10# | 4.82 (9.6%) | 4.65 (5.7%) | 4.40 | 4.67 (6.1%) | 4.65 (5.7%) | 4.40 |

[a] The absolute value of the relative deviation.

Table 4
The concentrations and their relative deviations, predicted by PARAFAC and CPARAFAC with fluorescent samples 1#, 2#, and 4# as calibration set

| No. | Tyrosine ($10^{-4}$ mg ml$^{-1}$) | | | Tryptophan ($10^{-5}$mg ml$^{-1}$) | | |
|---|---|---|---|---|---|---|
| | PARAFAC | CPARAFAC | Known | PARAFAC | CPARAFAC | Known |
| 7# | 4.49 (2.1%) | 4.55 (3.4%) | 4.40 | 9.65 (0.5%) | 9.63 (0.3%) | 9.60 |
| 8# | 9.00 (2.3%) | 9.07 (3.1%) | 8.80 | 10.59 (10.3%) | 9.90 (3.1%) | 9.60 |
| 9# | 4.50 (2.3%) | 4.53 (3.0%) | 4.40 | 20.14 (4.9%) | 19.22 (0.1%) | 19.20 |
| 10# | 5.28 (20.0%) | 5.26 (19.6%) | 4.40 | 9.97 (3.8%) | 9.81 (2.2%) | 9.60 |

Table 5
The concentrations and their relative deviations, predicted by PARAFAC and CPARAFAC with calibration sets assembled by fluorescent samples 1#–10# except the sample under test

| No. | Tyrosine ($10^{-4}$ mg ml$^{-1}$) | | | Tryptophan ($10^{-5}$ mg ml$^{-1}$) | | | Phenylalanine ($10^{-3}$ mg ml$^{-1}$) | | |
|---|---|---|---|---|---|---|---|---|---|
| | PARAFAC | CPARAFAC | Known | PARAFAC | CPARAFAC | Known | PARAFAC | CPARAFAC | Known |
| 4# | 4.08 (7.3%) | 4.21 (4.3%) | 4.40 | 9.82 (2.3%) | 10.3 (7.3%) | 9.60 | – | – | – |
| 5# | – | – | – | 8.70 (9.4%) | 9.13 (4.9%) | 9.60 | 2.35 (6.8%) | 2.35 (6.8%) | 2.20 |
| 6# | 4.30 (2.3%) | 4.43 (0.7%) | 4.40 | – | – | – | 1.94 (11.8%) | 2.28 (3.6%) | 2.20 |
| 7# | 4.25 (3.4%) | 4.25 (3.4%) | 4.40 | 9.25 (3.7%) | 9.46 (1.5%) | 9.60 | 2.00 (9.1%) | 2.06 (6.4%) | 2.20 |
| 8# | 8.77 (0.3%) | 8.65 (1.7%) | 8.80 | 9.64 (0.4%) | 9.54 (0.6%) | 9.60 | 2.11 (4.1%) | 2.18 (0.9%) | 2.20 |
| 9# | 4.31 (1.9%) | 4.25 (3.4%) | 4.40 | 18.60 (3.1%) | 18.23 (5.1%) | 19.20 | 2.14 (2.8%) | 2.13 (3.2%) | 2.20 |
| 10# | 5.10 (16.0%) | 5.00 (13.6%) | 4.40 | 9.41 (1.9%) | 9.18 (4.4%) | 9.60 | 4.71 (7.1%) | 4.32 (1.7%) | 4.40 |

of the deviations ($\underline{\boldsymbol{D}}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$) related to the chemical species — the possible run-to-run variability in chromatographic peak shape and position. Nevertheless, a significant improvement in the accuracy of the concentration prediction for this kind of data arrays is also observed by the introduction of constraints on the model parameters. With $N$ taking 3, out of the 12 concentrations of *o*-dichlorobenzene, *p*-chlorotoluene and *o*-chlorotoluene predicted by PARAFAC, eight estimated concentrations are of poor quality with relative deviations larger than 10.0%. The relative deviation for *o*-chlorotoluene in sample 6# is even as large as 55.3%, which is by no means acceptable. In comparison, when CPARAFAC in stead of PARAFAC is employed to analyze the HPLC data arrays, the number of concentrations estimated with relative deviation larger than 10.0% is reduced to three. The advantages brought by the utilization of constraint on model parameter are further verified by the prominent differences between the chromatographic profiles resolved by CPARAFAC and PARAFAC (Fig. 3a). Both the two chromatographic profiles of *p*-chlorotoluene

Table 6
The concentrations and their relative deviations, predicted by PARAFAC and CPARAFAC with calibration sets assembled by HPLC-DAD samples $1^{\#}$–$9^{\#}$ except the sample under test

| No. | $N$ | *o*-Dichlorobenzene ($\mu g\,ml^{-1}$) | | | *p*-Chlorotoluene ($\mu g\,ml^{-1}$) | | | *o*-Chlorotoluene ($\mu g\,ml^{-1}$) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | PARAFAC | CPARAFAC | Known | PARAFAC | CPARAFAC | Known | PARAFAC | CPARAFAC | Known |
| $6^{\#}$ | 3 | 127.88 (16.0%) | 135.73 (10.8%) | 152.20 | 8.83 (29.9%) | 13.65 (8.3%) | 12.60 | 6.80 (55.3%) | 14.25 (6.3%) | 15.20 |
| | 4 | 126.67 (16.8%) | 135.99 (10.7%) | 152.20 | 13.73 (9.0%) | 13.77 (9.3%) | 12.60 | 11.36 (25.3%) | 13.97 (8.0%) | 15.20 |
| $7^{\#}$ | 3 | 72.54 (19.3%) | 71.61 (17.8%) | 60.80 | 19.43 (22.9%) | 20.91 (17.0%) | 25.20 | 94.35 (3.5%) | 96.80 (6.1%) | 91.20 |
| | 4 | 70.28 (15.6%) | 61.52 (1.2%) | 60.80 | 19.76 (21.6%) | 21.90 (13.1%) | 25.20 | 89.74 (1.6%) | 96.47 (5.8%) | 91.20 |
| $8^{\#}$ | 3 | 96.50 (5.8%) | 95.98 (5.2%) | 91.20 | 48.75 (3.3%) | 51.55 (2.3%) | 50.40 | 25.77 (15.2%) | 27.79 (8.6%) | 30.40 |
| | 4 | 96.52 (5.8%) | 92.74 (1.7%) | 91.20 | 52.25 (3.7%) | 52.31 (3.8%) | 50.40 | 27.06 (11.0%) | 27.39 (9.9%) | 30.40 |
| $9^{\#}$ | 3 | 37.82 (24.4%) | 31.61 (4.0%) | 30.40 | 73.22 (3.2%) | 74.60 (1.3%) | 75.60 | 67.29 (10.7%) | 66.24 (9.0%) | 60.80 |
| | 4 | 36.37 (19.6%) | 29.12 (4.2%) | 30.40 | 75.20 (0.5%) | 75.11 (0.7%) | 75.60 | 66.66 (9.6%) | 66.17 (8.8%) | 60.80 |

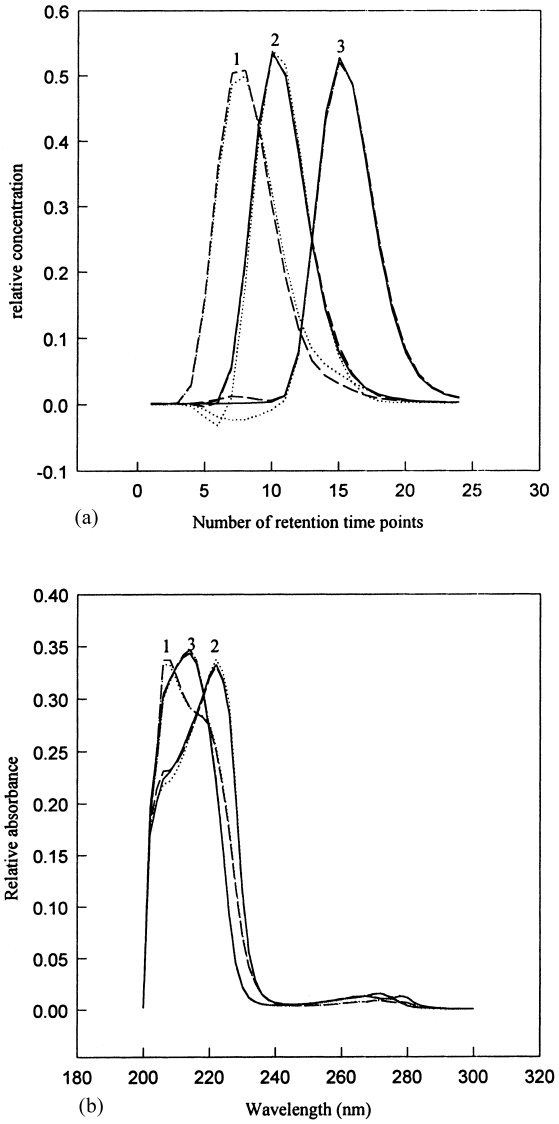


Fig. 3. The chromatographic (a) and spectral (b) profiles of *o*-dichlorobenzene (1), *p*-chlorotoluene (2) and *o*-chlorotoluene (3) for the HPLC data array composed of all the nine samples (solid line — real, dotted line — resolved by PARAFAC, dash line — obtained by CPARAFAC with sample 5 as unknown).

and *o*-chlorotoluene resolved by PARAFAC possess negative parts. Though the chromatographic profile of *o*-chlorotoluene obtained by CPARAFAC also shows some aberration from the real one, the aberration is much smaller in comparison with that of PARAFAC. Moreover, the chromatographic profile of *p*-chlorotoluene calculated by CPARAFAC coincides with the actual one. Similar results also appear in the spectra resolved by CPARAFAC and PARAFAC (Fig. 3b).

The increase of $N$ from 3 to 4 has improved the qualities of the results obtained by PARAFAC to some extent. The improvement, however, is still not enough to make the results of PARAFAC to be comparable to those of CPARAFAC, even when $N$ in CPARAFAC is equal to 3. Further increase of $N$ for PARAFAC will cause sharp deterioration in the results' qualities, due to the well-known property of PARAFAC — being sensitive to excess factors used in calculation. Therefore, the poor qualities of the results obtained by PARAFAC are caused by deviations, which cannot be decomposed into a trilinear form, rather than by insufficient number of factors used in calculation. The results' qualities of CPARAFAC with $N = 4$ also see some improvements. For example, the relative deviation of *o*-dichlorobenzene in sample $7^{\#}$ was decreased from 17.8 to 1.2%, and that of *p*-chlorotoluene in the same sample, from 17.0 to 13.1%. Generally, the results of CPARAFAC with $N = 4$ are in consistence with those of CPARAFAC with $N = 3$. Again, it verified that imposing constraints on model parameters do raise some advantages over common PARAFAC.

It is worth to point out that CPARAFAC will sometimes be trapped in local optimum. It can be relieved just by taking the results of PARAFAC as initial estimations, and then run CPARAFAC to refine the results. Non-random initial estimations based on some sophisticated chemometric methods such as EFA [21] and OPA [22] are also effective to avoid the problem of local optimum, and hence recommendable.

## 6. Conclusions

The constrained PARAFAC algorithm proposed in this paper can reduce the influence of deviations on the predictive accuracy in second-order calibration. By imposing a constraint (the concentration matrix of the calibration samples) on the loading matrix $\boldsymbol{C}$, the two steps of decomposition and calibration of PARAFAC in second-order calibration has been combined into just one step in CPARAFAC. The calibration procedure can thus be simplified with some

extra advantages over PARAFAC when deviations such as Rayleigh scattering in fluorescence spectroscopy present in data arrays. A comparison of the proposed procedure with PARAFAC for the treatment of fluorescent and HPLC data arrays demonstrated that the relative deviations of the predicted concentrations could be significantly reduced by CPARAFAC. Though CPARAFAC will sometimes be stuck in local optimum, a simple remedy is to take the results of PARAFAC as initial estimations and then run CPARAFAC to further enhance the qualities of the results.

It should be noted that CPARAFAC is just an alternative for second-order calibration, it should not be regarded as a substitute for PARAFAC, since the two-step scheme of PARAFAC allows non-linear regression of estimated loading matrix $\boldsymbol{C}$ on the concentrations of calibration samples, while CPARAFAC does not. It is, therefore, recommended that only when the results of PARAFAC show apparent abnormal characteristics such as negative parts in spectra or chromatographic profiles, should the CPARAFAC be tried.

## Acknowledgements

The authors would like to thank the National Nature Science Foundation of China for financial support (Grant No. 29735150).

## References

[1] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) 782A.
[2] R.A. Harshman, UCLA Working Papers in Phonetics 16 (1970) 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] A.K. Smilde, R. Tauler, J.M. Henshaw, L.W. Burgess, B.R. Kowalski, Anal. Chem. 66 (1994) 3345.
[11] A.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] G.G. Anderson, B.K. Dable, K.S. Booksh, Chemom. Intell. Lab. Syst. 49 (1999) 195.
[15] R. Bro, S.D. Jong, J. Chemom. 11 (1997) 393.
[16] R. Bro, N.D. Sidiropoulos, J. Chemom. 12 (1998) 223.
[17] R. Bro, Multi-way analysis in the food industry, Ph.D. Thesis.
[18] R. Tauler, Chemom. Intell. Lab. Syst. 30 (1995) 133.
[19] H.A.L. Kiers, A.K. Smilde, J. Chemom. 12 (1998) 125.
[20] H.L. Wu, M. Shibukawa, K. Oguma, J. Chemom. 12 (1998) 1.
[21] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbuhler, Talanta 32 (1985) 1133.
[22] F.C. Sanchez, S.C. Rutan, M.D. Gil Garcia, D.L. Massart, Chemom. Intell. Lab. Syst. 36 (1997) 153.