Available online at www.sciencedirect.com
SCIENCE DIRECT
Talanta 68 (2006) 1371–1377
Talanta
www.elsevier.com/locate/talanta

# Vertex vector sequential projection for the resolution of three-way data

Zhi-Guo Wang, Jian-Hui Jiang, Yu-Jie Ding, Hai-Long Wu, Ru-Qin Yu *

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

Received 21 April 2005; received in revised form 28 July 2005; accepted 30 July 2005
Available online 8 September 2005

**Abstract**

Usually, the PARAFAC2 method is utilized for handling retention time shifts in resolving chromatographic three-way data. It requires all profiles shift the same amount, which, unfortunately, seems unlikely the case in the practice of chromatographic analysis of multi-component samples. The present authors deal with the problem by unfolding the three-way data array along a certain direction into one matrix and setting up a multi-bilinear model. Then, a new method called vertex vector sequential projection (VVSP) is proposed to select pure variables and then the alternating least squares (ALS) procedure is used to iteratively improve the fit of the data to the multi-bilinear model. With a good estimate that is as close as possible to the pure variables, a fast convergence can be expected. Moreover, no prerequisite on the shifting is required and the multi-bilinear model provides a plausible manner to make use of the multi-sample information. An additional advantage is that the present fitting procedure is easier to adjust when constraints such as non-negativity, unimodality, etc., are to be imposed on the loading matrix. The proposed method is evaluated with simulated and real chemical data sets. Satisfactory resolution results are obtained, which demonstrates the performance of the proposed method.
© 2005 Published by Elsevier B.V.

*Keywords:* Trilinear model; Retention time shift; PARAFAC; PARAFAC2; Vertex vector

## 1. Introduction

Parallel factor analysis (PARAFAC) is a multi-way method originated from psychometrics [1,2]. It utilizes alternating least squares in an iterative manner, exploiting the conditional linearity of a trilinear model. The PARAFAC algorithm has gained much interest in chemometrics and has been widely utilized in analytical practice due to the uniqueness and optimality of its results as well as the so-called second-order advantage [3–7].

The most important prerequisite for the successful application of PARAFAC is that the data arrays should strictly follow a trilinear model, which unfortunately might be violated in practice. There are several reasons that result in the deviation of chemical data from the trilinear model as classified by Booksh and Kowalski [8]. Chromatographic shifting in HPLC-DAD data is one instance.

High-performance liquid chromatography with diode array detection (HPLC-DAD) is a widely used analytical technique. DAD measures absorbance as a function of both time and wavelength, and provides a two-dimensional data matrix to every sample that is analyzed. In recent years, the demand for rapid HPLC analysis has increased in a number of analytical fields [9,10]. However, the chromatographic shifting could hardly be avoided because the stability of both operator and the state of the instrument could not always be guaranteed from run to run. If the shifting is severe, the trilinearity required by the PARAFAC algorithm is no longer satisfied.

Usually, the PARAFAC2 method is utilized for handling retention time shifts in resolving chromatographic data [11,12]. It requires all profiles shift the same amount, which, however, seems hardly to be the case in the practice of the chromatographic analysis of multi-component samples.

In this paper, the authors deal with the problem by unfolding the three-way data array along a certain direction into one

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

0039-9140/$ – see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.talanta.2005.07.058

matrix and setting up a multi-bilinear model. Then, a method is proposed to select pure variables and iteratively improve the fit of the data to the multi-bilinear model. Additionally, the multi-bilinear model provides a plausible manner to make use of the multi-sample information.

A two-way data matrix is obtained by unfolding the three-way data array. Thus, two-way resolution techniques can be utilized to the unfolded data matrix. In two-way resolution, the pure profiles are also referred to as pure variables. In this paper, the proposed approach is mainly based on the fact that the pure variables are all boundaries in the vector space, which means two-way data points are bracketed by the pure variables [13–15]. Submitted to normalizations, two-way data points are located on a polyhedral hyper-"spherical" surface with the pure variables on the vertices. Thus, a certain quadratic form is to be maximized by the vertex vectors, which are pure variables if they exist. A procedure for determining the pure variables called vertex vector sequential projection (VVSP) in the two-way data is proposed. To improve the resolution, the alternating least squares (ALS) procedure is adopted.

The iterative nature of the proposed algorithm means that starting variables are required, and the algorithm is guaranteed to improve the least squares fit of the data to the multi-bilinear model at each iteration. One can imagine that good starting values, which are as close as possible to the real ones, would undoubtedly speed up the convergence of the algorithm. The main idea of the VVSP is to find the variables that are most apart from each other. If the pure variables exist in the data set, VVSP could find them one by one; if not, VVSP is inclined to find the nearest point to the pure one in the vector space. So, with good estimates to the pure variables as the starting values that VVSP provides, a fast convergence could be achieved.

With the proposed method, it can be expected to enhance the quality of the decomposing results comparing with the direct decomposition by PARAFAC when the shifts are severe. The proposed method is evaluated with simulated examples and real HPLC-DAD data set. Satisfactory resolution results are obtained for both artificial and real chemical data sets with the proposed method.

## 2. Nomenclature

Throughout this paper, scalars are represented with lowercase italics and vectors with bold lowercase characters. Bold capitals designate two-way matrices and underlined bold capitals symbolize three-way data arrays. The letters $I$, $J$ and $K$ are kept for denoting the dimensions of different modes in three-way data arrays; $F$ is the number of actual underlying factors. $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ with dimensions of $I \times F$, $J \times F$ and $K \times F$, respectively, are the three loading matrices of $\underline{\mathbf{X}}$. If the three-way data array $\underline{\mathbf{X}}$ is gained by stacking matrices recorded by HPLC-DAD over different samples, loading matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ would be called the relative chromatographic matrix, the relative spectral matrix and the concentration matrix, respectively.

## 3. Theory and algorithm

The structural model of PARAFAC can be expressed in matrix notation as follows:

$$\mathbf{X}_{..k} = \mathbf{A}_{I\times F}\,\mathrm{diag}(\mathbf{c}'_k)\mathbf{B}^{\mathrm{T}}_{J\times F} + \mathbf{E}_{..k}, \qquad k = 1, 2, \ldots, K \quad (1)$$

where $\mathbf{X}_{..k}$ is $k$th horizontal slice of the three-way data array $\underline{\mathbf{X}}$. $\mathbf{E}_{..k}$ is the corresponding slice of the three-way residue array $\underline{\mathbf{E}}$. $\mathrm{diag}(\mathbf{c}'_k)$ is a diagonal matrix with elements equal to the $k$th row of matrix $\mathbf{C}$.

As there is assumed to be retention time shift, the relative chromatographic matrix could not be expressed by a single matrix $\mathbf{A}$. So, the three-way model for PARAFAC2 is modified as follows:

$$\mathbf{X}_{..k} = \mathbf{A}_k\,\mathrm{diag}(\mathbf{c}'_k)\mathbf{B}^{\mathrm{T}} + \mathbf{E}_{..k}, \qquad k = 1, 2, \ldots, K \quad (2)$$

where $\mathbf{A}_k$ differs from sample to sample.

Unfolding the three-way data array $\underline{\mathbf{X}}$ and arranging the unfolded matrix $\mathbf{X}$ as follows:

$$\mathbf{X} = \begin{bmatrix}\mathbf{X}^{\mathrm{T}}_{..1} & \mathbf{X}^{\mathrm{T}}_{..2} & \cdots & \mathbf{X}^{\mathrm{T}}_{..K}\end{bmatrix}^{\mathrm{T}}, \qquad k = 1, 2, \ldots, K \quad (3)$$

then one has

$$\mathbf{X}_{IK\times J} = \mathbf{D}_{IK\times F}\mathbf{B}^{\mathrm{T}}_{J\times F} + \mathbf{E}_{IK\times J} \quad (4)$$

where $\mathbf{D}$ is defined as $\mathbf{D} = [\,\mathbf{D}_1^{\mathrm{T}} \;\; \mathbf{D}_2^{\mathrm{T}} \;\; \cdots \;\; \mathbf{D}_K^{\mathrm{T}}\,]^{\mathrm{T}}$, $\mathbf{D}_k = \mathbf{A}_k\,\mathrm{diag}(\mathbf{c}_{k\prime})$ ($k = 1, 2, \ldots, K$). $\mathbf{E}$ is the unfolded residue matrix. Obviously, there are $K$ sub-matrices in the unfolded matrix $\mathbf{X}$ that follow the bilinear model, so Eq. (4) is called a multi-bilinear model and the unfolded matrix $\mathbf{X}$ is, in form, a two-way matrix. Therefore, the two-way resolution techniques can be utilized in this case. Note that $\mathbf{D}_k$ is actually the relative chromatographic matrix for the $k$th sample.

In two-way resolution, the pure profiles are also referred to as pure variables. As the norms of spectra and the chromatograms cannot be uniquely determined, one can prescribe a certain scale constraint for the spectral or chromatographic profiles, say the spectral profiles are assumed to have unit norm. It can be deduced that subject to normalizations, the spectral points (each spectrum can be regarded as a point in a $J$-dimensional space) in the two-way data $\mathbf{X}$ are located on a polyhedral hyper-"spherical" surface. As mentioned above, the pure variables are all boundaries in the vector space, which means two-way data points are bracketed by the pure variables. Based on this fact, a procedure can be designed to identify the pure variables. In order to ascertain the pure variables, it is sufficient to identify the vertex vectors, and this constitutes the basis of the proposed algorithm for determining the pure variables in two-way data.

Suppose a number (not all) of pure variables have been found, there must exist an additional one, among the remaining pure variables to be determined, which is most apart from

the space and thus is the nearest to the complementary space spanned by the known ones. As the variables in the two-way data matrix are all normalized to unit length, the additional pure variable, when projected to the complementary space of the known ones, should have the longest length. Thus, the length of the normalized variables projecting to the complementary space of the known one(s) can be used as an index to determine a sought-for pure variable. If the additional pure variable is not present in the normalized data, the variable that is nearest to the additional pure one would be determined as a substitute, which is also required to have the farmost distance to the complementary space spanned by the known pure ones.

The normalized bilinear model is given below:

$$\mathbf{Y} = \mathbf{W}\mathbf{B}^{\mathrm{T}} \tag{5}$$

where $\mathbf{Y}$ is the two-way data matrix with each row of $\mathbf{X}$ normalized to unit norm. $\mathbf{B}$ is the pure spectral matrix that is normalized to unit norm columnwise. Comparing with Eq. (4), $\mathbf{W}$ can be called the normalized chromatographic matrix. Let $\mathbf{Z}_N = [\,\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_N\,]$ ($N < F$), and $\mathbf{G} = \mathbf{I}_J - \mathbf{Z}_N\mathbf{Z}_N^{+}$. $\mathbf{Z}_N$ collects the pure variables that have been found and normalized to unit length, $\mathbf{G}$ represents the complementary space spanned by the pure variables $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$. Note that $\mathbf{G}$ is a non-negative definite symmetric matrix and is also a projection matrix constituted by a part of pure variables and bears the property of $\mathbf{G} = \mathbf{G}\mathbf{G}$. If an arbitrary row in the normalized matrix $\mathbf{Y}$ is represented by $\mathbf{y}_i^{\mathrm{T}}$, then $(\mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}} = \mathbf{B}\mathbf{w}$, where $\mathbf{w} = (\mathbf{w}_i^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{w}_i^{\mathrm{T}}$ is the corresponding row of matrix $\mathbf{W}$. Then, the projected vector of a normalized row of the matrix $\mathbf{Y}$, i.e. the variables to be examined, to the complementary space of the known ones can be expressed by $\mathbf{G}\mathbf{B}\mathbf{w}$, and the length of the projected vector can be expressed as follows:

$$f(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{G}\mathbf{B}\mathbf{w} \tag{6}$$

Here, the property of $\mathbf{G} = \mathbf{G}\mathbf{G}$ and $\mathbf{G} = \mathbf{G}^{\mathrm{T}}$ is utilized.

Based on the above description, a procedure can be derived for locating the pure variables in two-way data. As a matter of fact, if the non-negative definite symmetric matrix $\mathbf{G}$ is obtained in such a way that the pure spectral variables already found are included in its null space, one can plot the values of the quadratic form $\mathbf{y}_i^{\mathrm{T}}\mathbf{G}(\mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}$ for all the normalized spectra $\mathbf{y}_i^{\mathrm{T}}$. Thus, the normalized spectrum having the maximal value among $\mathbf{y}_i^{\mathrm{T}}\mathbf{G}(\mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}$ gives one of the remaining pure spectral variables. Then, the remaining pure variables can also be identified successively. So, the first pure variable must be found to start the procedure. A simple way to decide the first pure variable is that let $\mathbf{r}^{\mathrm{T}}$ be the normalized spectrum, whose norm is the largest in the original two-way matrix $\mathbf{X}$. The normalized spectrum whose distance to $\mathbf{r}^{\mathrm{T}}$ is the farmost can be utilized as the first pure variable. Based on the geometry of the normalized two-way data, the way for selecting the first pure variable in such a way can be easily explained.

In light of the principles above, the procedure for the ascertainment of the pure spectral variables, vertex vector sequential projection, is developed as follows:

- *Step 1*: Select a feasible set that is composed of the spectra having norms larger than a specified positive constant $c$.
- *Step 2*: Normalize each spectrum in the feasible set to unit norm.
- *Step 3*: Determine the first pure variable $\mathbf{b}_1$ in such a way—select the spectrum with the largest norm in the original data matrix $\mathbf{X}$ as $\mathbf{r}^{\mathrm{T}}$ and normalize $\mathbf{r}^{\mathrm{T}}$, find the spectrum $\mathbf{y}_i^{\mathrm{T}}$, in the feasible set that maximizes $\left\|\mathbf{r}^{\mathrm{T}} - \mathbf{y}_i^{\mathrm{T}}\right\|_2$ as $\mathbf{b}_1$.
- *Step 4*: Let $\mathbf{G} = \mathbf{I}_J - \mathbf{Z}_N\mathbf{Z}_N^{+}$ and $\mathbf{Z}_N = [\,\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_N\,]$ ($N = 1, 2, \ldots, F-1$), and find the spectrum, $\mathbf{b}_{N+1}$, in the feasible set that maximizes the quadratic form $\mathbf{y}_i^{\mathrm{T}}\mathbf{G}(\mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}$.
- *Step 5*: Repeat Step 4 until $F$ pure spectral variables have been identified.

Note that in the procedure, a step to select a feasible set for the pure variables is adopted to exclude those spectra with very poor signal-to-noise ratios. This step is essential to maintain that, even in the presence of measurement errors, the normalized spectra are still approximately contained in the polyhedral hyper-"spherical" surface. In addition, it ensures that the pure variables to be determined have a sufficient signal-to-noise ratio. The gradually improved non-negatively definite matrix $\mathbf{G}$ in Step 4 is to make the identified pure variables have zero values for the quadratic form such that the pure variables already found can be excluded in subsequent searching for unidentified ones.

If some pure spectral variables are absent or the noise in the data is too heavy, the resolved "pure" variables by the proposed procedure may lead to an irrational solution for the other set of pure variables gained via the least square method. To improve the resolution, an optimization procedure, the alternating least squares procedure is adopted. Combining with some constraints such as non-negativity, unimodality, etc., ALS is expected to improve the results iteratively.

Note that an additional advantage of the proposed method is that the present fitting procedure is easier to adjust when constraints such as non-negativity, unimodality, etc., are to be imposed on the loading matrix. As in the present study, the applications involve the data from hyphenated chromatography; the unimodality rectification step is included in the algorithm. This is implemented in the same way as orthogonal projection analysis [16].

After the resolution finished, the relative concentrations in all the samples can be easily obtained by the integral of the resolved relative chromatographic matrices $\mathbf{D}_k$, which is the basis of chromatographic analysis, i.e. the integral calculus of a chromatogram is proportional to the concentration of an analyte. Thus, a three-way resolution is achieved by the multi-bilinear model and the proposed method.

The proposed method was evaluated with simulated and real chemical data sets in the following section. Satisfactory

resolution results were obtained, which demonstrates the performance of the proposed method.

## 4. Experimental

### 4.1. Simulated samples

The proposed method was performed on an artificial three-way data array of a three-component system. The data were simulating HPLC-DAD profiles of 10 samples with dimension of $70 \times 40 \times 10$ (retention time × wavelength × sample). Spectra and chromatograms were generated as sums of Gaussian peaks and depicted in Fig. 1. For simplicity, only the first chromatogram shifts from sample to sample and the other two chromatograms remain steady to the retention time. The noise for the data sets is created using a three-way array of random numbers that are normally distributed with zero mean and standard deviation of 0.002. The cross-products of relative chromatographic matrices of the samples 1 and 10 are shown as below:

$$\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_1 = \begin{bmatrix} 1.0000 & 0.0030 & 0.0000 \\ 0.0030 & 1.0000 & 0.0449 \\ 0.0000 & 0.0449 & 1.0000 \end{bmatrix} \quad \text{and}$$

$$\mathbf{A}_{10}^{\mathrm{T}}\mathbf{A}_{10} = \begin{bmatrix} 1.0000 & 0.1502 & 0.0001 \\ 0.1502 & 1.0000 & 0.0449 \\ 0.0001 & 0.0449 & 1.0000 \end{bmatrix}$$

The cross-products of relative chromatographic matrices are required to be constant over $k$ by the PARAFAC2 model. The cross-products shown above have been scaled by making the first value 1 for the convenience of comparison. It is readily seen that these matrices are not identical and hence the requirement for the PARAFAC2 model to hold is not valid here. So, the comparison is mainly carried out between PARAFAC and the proposed method.

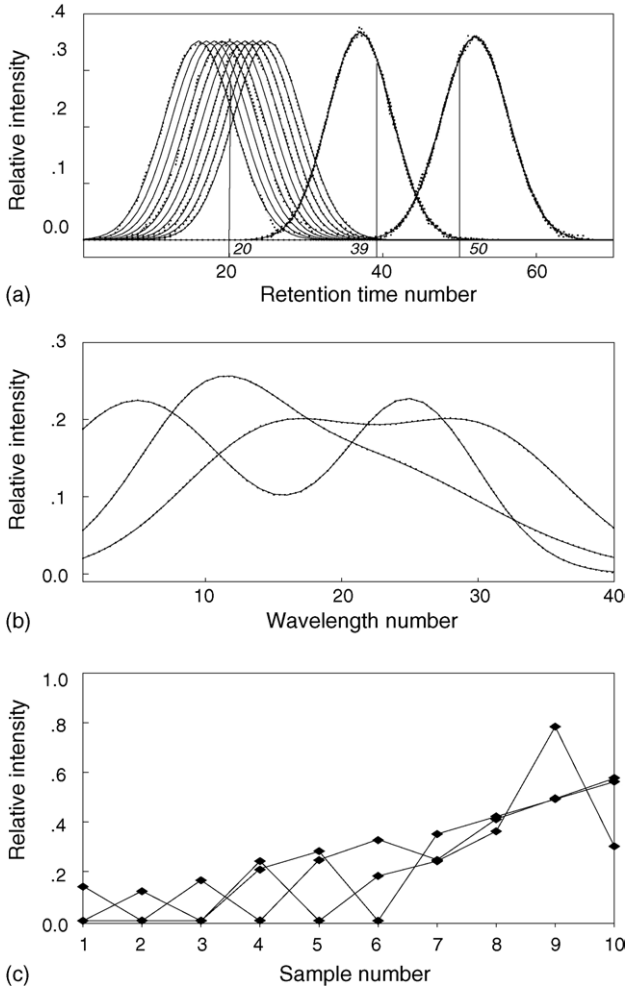


Fig. 1. The resolved profiles (dotted line for chromatograms (a) and spectra (b) and diamonds for concentrations (c)) by VVSP–ALS compared with real profiles (solid line) for the three components in the simulated samples. The numbers in italic style in (a) indicate the corresponding retention time numbers that give the pure variables identified by VVSP.

### 4.2. HPLD-DAD data array

#### 4.2.1. Reagents and stock solutions

All reagents used were of analytical grade. Stock solutions of Vitamins $B_1$, $B_2$, $B_3$ and $B_6$ were prepared by accurately weighting appropriate amount of reagents and dissolving Vitamins $B_1$, $B_3$ and $B_6$ in eluant solution and Vitamin $B_2$ in NaOH solution (0.01 mol $L^{-1}$). The stock solutions were all dissolved in 100 mL volumetric flasks and the eluant is used as diluting agent. In the preparation of stock solutions, an appropriate amount of hydrochloric acid was added to Vitamin $B_2$ solution to make its pH value less than 7. A total of 11 working solutions with different concentrations of the 4 components were made by taking appropriate volumes of stock solutions into 10 mL volumetric flasks and then making them to 10 mL with the eluant.

#### 4.2.2. Apparatus

The HPLC-DAD response matrices of all the samples as well as six blank solutions were recorded by Agilent-1100 with Eclipse ADB-C8 ($4.6 \times 250$, 5 μm) as the separating column. The wavelength range is 248–318 nm with a fixed interval of 2 nm and the monitoring time is from 59.7 s to 107.7 s with a fixed interval of 0.4 s. The solution of methanol and water (volume rate 70:30) was used as eluant. The flowing rate is 1.0 mL $min^{-1}$, column temperature is 25 °C and injection volume is 20.0 μL.

## 5. Results and discussion

The proposed method needs an input of the component number, which is vital in two-way data resolution. The component number can be determined by analysis of the so-called rank map. As it is not an emphasis discussed in the present

study, for simplicity, it is assumed that the component number is known for the simulated and real HPLC-DAD data sets.

### 5.1. Simulated samples

With the component number set to 3, VVSP sequentially found the pure spectral variables, which are the 330th, the 249th and the 440th spectra in the unfolded matrix **X**, respectively. As the number of retention time is 70, the corresponding spectra are the 50th spectrum in the fifth sample, the 39th spectrum in the fourth sample and the 20th spectrum in the seventh sample, which are shown in Fig. 1a. It can be seen that the pure retention time points identified are all dominated by a single component, which indicates that VVSP is capable of determining the pure variables, when they actually exist.

With the identified pure spectral variables as the starting estimate, VVSP–ALS gives the resolution in 89 iterations. The recovered profiles are perfectly consistent with the true ones (Fig. 1a and b, solid line for true ones and dotted line for resolved ones), showing that the proposed method is able to resolve the spectra as well as the true chromatograms for each sample when there exist retention time shifts. After the resolution is finished, the relative concentrations are obtained simply by the integral calculus of each $\mathbf{D}_k$, which are depicted (diamonds) in Fig. 1c together with the true concentrations (solid line). One can see that they are matching very well, which further testifies the performance of the proposed method.

We also carried out the PARAFAC algorithm on the simulated data set. As the PARAFAC procedure uses the same chromatographic matrix **A** for all the samples, the resolved chromatogram for the first eluting component, which shifts in different samples in the simulated data, should be an average of the true shifting chromatograms. It is found that the resolved chromatographic profile for the first component is more close to that of sample seven, comparing other samples. Due to the influence of shifting, the resolved spectra and the concentrations of the first component and its overlapping component (the second eluting component) deviate slightly from the true ones, while the third eluting component is nearly free of influence since it overlaps only a very small part with the shifting component. This is supported by the correlation coefficients between the resolved and true profiles for the three components, which, in orders, are 0.9986, 0.9999 and 1.0000 for spectra, and 0.9982, 0.9992 and 0.9998 for concentrations. Comparing these results, the values for VVSP–ALS are actually 1.0000. This sustains that if the shift in the chromatography is severe, PARAFAC bears a more serious influence, while the proposed method is robust to the shifts, since it is designed for solving the problem.

### 5.2. HPLD-DAD data array

The real HPLD-DAD data array comprises 120 retention times, 35 wavelengths and 11 samples. The pure spectral variables successively ascertained by VVSP are the 292nd, 804th, 209th and 29th spectra in the unfolded matrix **X**, respectively. Similarly, they are the 52nd spectrum of the third sample, the 84th spectrum of sample seven, the 89th spectrum of the second sample and the 29th spectrum of the first sample. With the identified pure spectral variables as the starting estimate, the resolution of the unfolded real HPLA-DAD data matrix by VVSP–ALS is achieved in 35 iterations. The relative concentrations were also obtained by the integral calculus of all $\mathbf{D}_k$. The resolved chromatographic profiles together with the location numbers of the pure spectral variables are shown in Fig. 2a and the resolved spectra (dotted line) and relative concentration (diamonds) profiles as well as the true ones (solid line) are depicted in Fig. 2b and c, respectively. Note that the peak with an arrow in Fig. 2a belongs to the first eluting component according to the resolution results. It heavily deviates from the peaks of the first group.

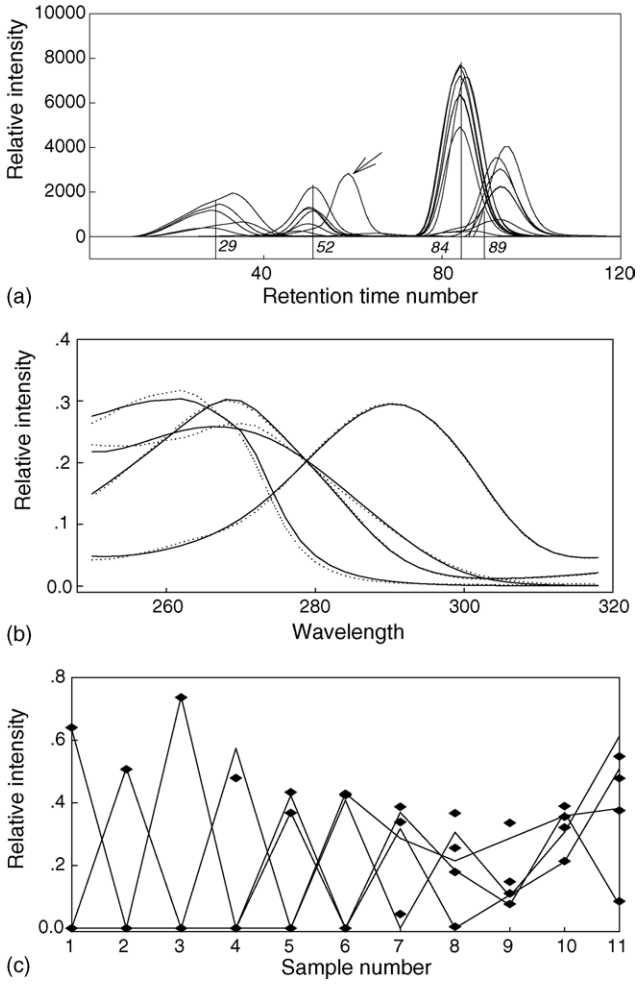


Fig. 2. The resolved profiles (solid line for chromatograms (a), dotted line for spectra (b) and diamonds for concentrations (c)) by VVSP–ALS compared with real spectral (b) and concentration (c) profiles (solid line) for the four components in the real HPLC-DAD data array. The numbers in italic style in (a) indicate the corresponding retention time numbers that give the pure variables identified by VVSP. The peak with an arrow in (a) belongs to the first eluting component.

Table 1
Concentrations of 11 mixtures in HPLC-DAD data array

| Samples | Concentration ($\mu g\,mL^{-1}$) | | | |
|---|---|---|---|---|
| | Vitamin $B_1$ | Vitamin $B_6$ | Vitamin $B_3$ | Vitamin $B_2$ |
| 1 | 0.1500 | 0.0000 | 0.0000 | 0.0000 |
| 2 | 0.0000 | 0.0000 | 0.0000 | 0.0500 |
| 3 | 0.0000 | 0.0800 | 0.0000 | 0.0000 |
| 4 | 0.0000 | 0.0000 | 0.6000 | 0.0000 |
| 5 | 0.1000 | 0.0400 | 0.0000 | 0.0000 |
| 6 | 0.0000 | 0.0000 | 0.4500 | 0.0400 |
| 7 | 0.0750 | 0.0400 | 0.3000 | 0.0000 |
| 8 | 0.0000 | 0.0200 | 0.2250 | 0.0300 |
| 9 | 0.0250 | 0.0080 | 0.3000 | 0.0100 |
| 10 | 0.0500 | 0.0400 | 0.3750 | 0.0300 |
| 11 | 0.1200 | 0.0080 | 0.4000 | 0.0600 |

It can be seen from Fig. 2a that the pure retention time points identified for the components 1–3 are dominated by a single component, while for the component 4, the retention time point (time 89) seems not a point that gives pure variable. But by the observation of the true concentrations (Table 1), one can find that the third eluting component is not present in sample 2. So, time 89 in the second sample, which is identified by VVSP as a pure spectral point, is actually providing a pure spectral estimate for component 4. This can be regarded as an advantage of the proposed method, i.e. it could make use of different presentation and separation of components in the different samples to search pure variables. On the other hand, one can carry out analyses on just one sample using different separating conditions. For a complicated system, different separating conditions of chromatography might result in different eluting overlaps of chemical components, which might contribute pure variables for different components. By simply arranging the obtained matrices as Eq. (3), one can get the similar multi-bilinear model for one sample as Eq. (4), since the spectral matrix holds unchanged in each analysis. The only difference between the multi-bilinear model for one sample and that of multi-samples is that the concentrations are a vector for one sample and hold unconverted in each run. For a complicated system, the search for conditions of a thorough separation of all components is often exhausting and time-consuming. The proposed method provides an alternative for such a case. One can choose several separating conditions that are believed to be more effective to at least a part of components, and then use the present method to achieve the resolutions.

As there exist severe retention time shifts in the real HPLC-DAD data array, the results of PARAFAC could be expectably not very satisfactory. The correlation coefficients between the resolved and true profiles for the four components are given in Table 2 for spectra and Table 3 for concentrations. It can be seen that the results of PARAFAC for components 1 and 4 are not so good, since more serious shifts can be observed from Fig. 2a. One can also find that the results of the proposed method are superior to that of PARAFAC as a whole, especially for the first component, whose shift is the most severe one; a more precise estimate to the concentrations is obtained by the VVSP–ALS method. This upholds that the proposed method is sure to give a better resolution for the heavily shifting chromatograms comparing with PARAFAC and in the meantime, this is also propitious to the second-order calibration.

Table 2
Correlation coefficients between the real and resolved spectral profiles for the real HPLC data array

| | Vitamin $B_1$ | Vitamin $B_6$ | Vitamin $B_3$ | Vitamin $B_2$ |
|---|---|---|---|---|
| VVSP–ALS (11 samples) | 0.9994 | 0.9999 | 0.9990 | 0.9999 |
| PARAFAC (11 samples) | 0.9892 | 0.9999 | 0.9992 | 0.9868 |
| PARAFAC (10 samples) | 1.0000 | 1.0000 | 0.9995 | 0.9999 |

From the results that VVSP–ALS provides, it is found that the chromatogram of the first component in sample 11 has the most severe deviation. If sample 11 is excluded from the data set, the remaining 10-sample system is approximately following the trilinear model. So, the present authors redo the decomposition with PARAFAC by removing sample 11 from the data set. The correlation coefficient results of the 10-sample data set are shown in Table 2 for the spectra and Table 3 for the concentrations. Comparing with the results of 11-sample data set, the decomposing results are remarkably improved, since the deviation from the trilinear model for the system is alleviated. This validates that PARAFAC is capable of managing data sets when they are slightly deviated from the trilinear model with satisfactory results. However, when the deviation is beyond a certain extent and the trilinear model is no longer a valid hypothesis, PARAFAC would be unsuitable for interpreting the data set. Thus, alternative procedures should be adopted.

As the resolution step of the proposed method is a two-way resolution method in nature, the present authors compared the results of a conventional two-way resolution method, window factor analysis (WFA) [17], with that of the proposed method for the 10th sample of the real HPLD-DAD data array. It is found that the correlation coefficients between two correspondingly resolved profiles (chromatograms and spectra) of WFA and VVSP–ALS are all larger than 0.99. This validates that the performance of VVSP–ALS is comparable to that of unique resolution method. Moreover, it does not involve the determination of feature regions such as selective regions and zero-concentration regions, which are usually required

Table 3
Correlation coefficients between the real and resolved concentrations for the real HPLC data array

| | Vitamin $B_1$ | Vitamin $B_6$ | Vitamin $B_3$ | Vitamin $B_2$ |
|---|---|---|---|---|
| VVSP–ALS (11 samples) | 0.9993 | 0.9996 | 0.9916 | 0.9947 |
| PARAFAC (11 samples) | 0.8602 | 0.9973 | 0.9932 | 0.9906 |
| PARAFAC (10 samples) | 0.9990 | 0.9990 | 0.9940 | 0.9993 |

by unique resolution methods. Additionally, the proposed method is apter to be programmed, as it does not require interventions of a person.

## 6. Conclusions

In this paper, the multi-sample HPLC-DAD data array with different extent of retention time shifts is unfolded into a two-way matrix and handled using a multi-bilinear model. Then, a procedure (VVSP) for the identification of pure variables, which are as close as possible to the true ones, in the two-way matrix is proposed. This provides a good starting estimate for the following refining process (ALS). The results obtained with both simulated and real chemical data sets have demonstrated that the proposed method is capable of finding pure variables provided they exist and gives more precise estimates to the model parameters. The comparison of VVSP–ALS with PARAFAC shows that PARAFAC can bear only very small model deviation, while VVSP–ALS is more robust when the chromatographic shifts are very severe. Moreover, the multi-bilinear model provides a plausible manner to make use of the multi-sample information. An additional advantage is that the present fitting procedure is easier to adjust when constraints such as non-negativity, unimodality, etc., are to be imposed on the loading matrix. The proposed method is approved to be a competent tool for the resolution of three-way data with the inspection to both simulated and real chemical data sets.

## Acknowledgements

The work is financially supported by the National Natural Science Foundation of China (Grant Nos. 20435010, 20375012, 20205005 and 20475014).

## References

[1] R.A. Harshman, UCLA Working Pap. Phonetics 16 (1970) 1.
[2] J.D. Carrol, J. Chang, Psychometrika 35 (1970) 283.
[3] P. Geladi, Chemom. Intell. Lab. Syst. 7 (1989) 11.
[4] A.K. Smilde, Chemom. Intell. Lab. Syst. 5 (1992) 143.
[5] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149.
[6] H.A.L. Kiers, J. Chemom. 12 (1998) 149.
[7] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) 782A.
[8] K.S. Booksh, B.R. Kowalski, Anal. Chim. Acta 348 (1997) 1.
[9] M.R. Brunetto, M.A. Obando, M. Gallignani, O.M. Alarcón, E. Nieto, R. Salinas, J.L. Burguera, M. Burguera, Talanta 64 (2004) 1364.
[10] K. Amarnath, V. Amarnath, K. Amarnath, H.L. Valentine, W.M. Valentine, Talanta 60 (2003) 1229.
[11] H.A.L. Kiers, J.M.F. Ten Berge, R. Bro, J. Chemom. 13 (1999) 275.
[12] R. Bro, C.A. Andersson, H.A.L. Kiers, J. Chemom. 13 (1999) 295.
[13] F.J. Knorr, J.H. Futrell, Anal. Chem. 51 (1979) 1236–1241.
[14] E.R. Malinowski, Anal. Chim. Acta 134 (1982) 129–137.
[15] O.S. Borgen, B.R. Kowalski, Anal. Chim. Acta 174 (1985) 1–26.
[16] F. Cuesta-Sanchez, B. van den Bogaert, S.C. Rutan, D.L. Massart, Chemom. Intell. Lab. Syst. 34 (1996) 139–171.
[17] E.R. Malinowski, J. Chemom. 6 (1992) 29.