ELSEVIER

Chemometrics and Intelligent Laboratory Systems 52 (2000) 75–86

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

A novel trilinear decomposition algorithm for second-order linear calibration

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

College of Chemistry and Chemical Engineering, Hunan University, Changsha, 410082, People's Republic of China

Received 20 November 1999; accepted 16 May 2000

Abstract

A novel trilinear decomposition algorithm for second-order linear calibration called self-weighted alternating trilinear decomposition (SWATLD) has been designed in this paper. Experiments show SWATLD has the features of fast convergence and being insensitive to the excess factors used in calculation. Due to the unique optimizing scheme employed, SWATLD is much more efficient than the ordinary PARAFAC algorithm. In terms of the variance, the performance of SWATLD is very stable when the number of factors used in calculation varies (as long as it is no less than the actual number of factors). Such a feature will facilitate the analysis of three-way data arrays, since it is now unnecessary to spend a lot of time and effort to accurately determine the number of underlying factors in the system studied as does in PARAFAC. Furthermore, as far as the deviations of the results are concerned, experiments show SWATLD can supply acceptable results in most cases.

Keywords: Second-order linear calibration; PARAFAC; Self-weighted alternating trilinear decomposition (SWATLD)

1. Introduction

For centuries, analysts have tried to design procedures or contrive devices to quantitatively determine the content of a component of interest in a mixture, while eliminating the influence of other coexistent compounds to the greatest extent. There are three conventional approaches to eliminate the influence of interference(s), i.e. utilizing agents selective to the target species, employing wavelength-selection techniques in spectrometry and using masking agents. Regardless of which approach is applied, however, the influence of interference(s) could not be thoroughly eliminated. Furthermore, background absorption could also not be perfectly corrected just by subtracting the absorption of a blank solution. The limitation of these approaches forces analysts to resort to chemometrics — a promising tool for solving the above problems.

Due to the rapid development of chemometrics and wide application of two-way instruments such as HPLC-DAD and excitation/emission matrix spectrofluorometer, theoretically, it is now possible to solve the above problems [1,2]. The first attempt to solve these problems may be symbolized by the introduction of PARAFAC algorithm proposed by

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

0169-7439/00/$ - see front matter
PII: S0169-7439(00)00081-2

Harshman [3] and Caroll and Chang [4]. PARAFAC is actually an alternating least square algorithm which is sensitive to the number of the factors used in calculation. With the correct estimation of the number of components in mixtures, PARAFAC can accurately determine the contents of specific components regardless of the existence of interference(s) and background absorption. Such a favorable feature has attracted many researchers to engage in the studies of theoretical aspects and practical application of PARAFAC [5–8]. Consequently, people found that the implementation of PARAFC is not an easy task, since PARAFAC requires an exact determination of the number of components in mixtures, which is very difficult even for experts with much experience. Though there are many sophisticated methods for determining the number of factors in mixtures [9–15], due to the complexity of the problem none of them can guarantee the correctness of its results under all circumstances. Algorithms, which do not require an accurate estimation of the number of factors in mixtures, are therefore of utmost importance. As far as we know, little attention has been paid to construct iterative algorithms that do not require an accurate estimation of the number of factors in mixtures.

Another annoying characteristic of PARAFAC is its low convergence rate [7]. In chemical data arrays, the size of at least two models is rather large, and hence they generally show at least mild degrees of multicollinearity, which tends to cause PARAFAC to require very many iterations before convergence. The large size of the data arrays causes each iteration to be computationally expensive, which further worsens the situation. Several approaches have been proposed to remedy this demerit of PARAFAC [16–23]. Kiers and Krijnem [16] studied the problem caused by large numbers of observation units and suggested the computation of some combinations of loading matrix and data sets instead of the loading matrixes themselves in iterative procedure. Since much smaller matrixes have been stored and calculated, a great decrease in computation time can be gained. Alsberg and Kvalheim [17,18] have described in a series of papers a method to accelerate the convergence of PARAFAC by compressing high dimensional data arrays, which is essentially equivalent to the CANDELINE approach [19]. Bro and Andersson [20] have also developed an efficient method for compressing large arrays using a fast Tucher3 algorithm [21]. All the above approaches try to estimate the model parameters from much smaller data matrixes to decrease the computation time in each iteration. There are also some other methods, which speed the optimizing procedure of PARAFAC by reducing the number of iterations required. Mitchell and Burdick [22] suggested avoiding ''swamping'' due to ''degeneracy'' by stopping a run as soon as it enters a swamp area and randomly starting a new run. Paatero introduced a penalty term with decreasing impact to accelerate the optimizing procedure [23]. Recently, Kiers proposed a three-step algorithm, which can greatly reduce the number of iterations required [24]. Though these methods have their different advantages, they all suffer form the same problem as PARAFAC, i.e. requiring accurate estimation of the number of factors. Thus methods which can converge fast and do not require accurate estimation of the factors are preferable and badly in need.

In the present paper, a novel algorithm called self-weighted alternating trilinear decomposition (SWATLD) is designed, which alternatively minimizes three objective functions with intrinsic relationship rather than the objective function of PARAFAC. The proposed algorithm not only can resist the influence of the excess factors used in calculation, but also largely reduce the number of iterations required, which will be illustrated by simulated and real data arrays.

2. Nomenclature

Throughout this paper, scalars are represented by lowercase italics, vectors by bold lowercase characters, bold capitals designate two-way matrices and underlined bold capitals symbolize three-way arrays. The letters I, J, K are kept for denoting the dimensions of different modes in three-way arrays; F, for the number of actual underlining factors, and N, for the number of factors used in calculation. $\underline{\mathbf{X}}$ represents three-way data array. $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ with dimensions of $I \times F$, $J \times F$, $K \times F$ respectively are the three loading matrixes of $\underline{\mathbf{X}}$. If three-way array $\underline{\mathbf{X}}$ is assembled by several data matrixes of different samples recorded by HPLC-DAD, loading matrixes $\mathbf{A}$, $\mathbf{B}$,

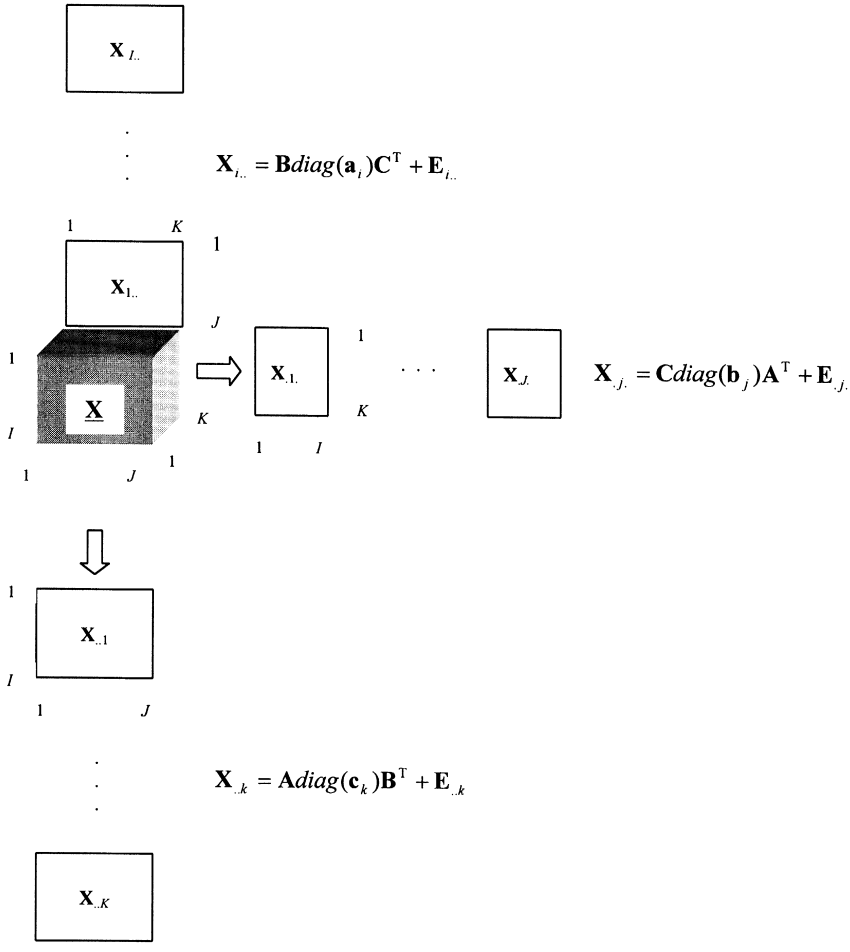

Fig. 1. The frontal, lateral and horizontal slices of a three-way array $\underline{\boldsymbol{X}}$.

$\mathbf{C}$ would be called the relative chromatogram matrix, the relative spectral matrix and the concentration matrix, respectively. As depicted in Fig. 1, $\mathbf{X}_{..k} = \mathbf{A}_{I\times F}\mathrm{diag}(\mathbf{c}_k)\mathbf{B}_{J\times F}^{\mathrm{T}} + \mathbf{E}_{..k}$ $(k = 1, 2, \ldots, K)$, $\mathbf{X}_{.j.} = \mathbf{C}_{K\times F}\mathrm{diag}(\mathbf{b}_j)\mathbf{A}_{I\times F}^{\mathrm{T}} + \mathbf{E}_{.j.}$ $(j = 1, 2, \ldots, J)$ and $\mathbf{X}_{i..} = \mathbf{B}_{J\times F}\mathrm{diag}(\mathbf{a}_i)\mathbf{C}_{K\times F}^{\mathrm{T}} + \mathbf{E}_{i..}$ $(i = 1, 2, \ldots, I)$ are the kth frontal slice, jth lateral slice and ith horizontal slice of three-way array $\underline{\mathbf{X}}$, respectively. $\mathbf{E}_{..k}$, $\mathbf{E}_{.j.}$ and $\mathbf{E}_{i..}$ are the corresponding slices of the three-way residue array $\underline{\mathbf{E}}$. Note that $\mathrm{diag}(\mathbf{c}_k)$, $\mathrm{diag}(\mathbf{b}_j)$, and are diagonal matrixes with elements equal to the kth, jth and ith rows of matrixes $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively. $\mathrm{diagm}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ is a vector with elements equal to the corresponding diagonal ones of matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. $\mathrm{cond}(\mathbf{A})$ signifies the 2-norm condition number of loading matrix $\mathbf{A}$.

3. Model and algorithm

In second-order linear calibration, the famous trilinear decomposition model proposed by Harshman [3] and Caroll and Chang [4] has been widely accepted, due to its consistence with Beer law in chemistry. The model can be expressed as the following matrix form:

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

or

$$\mathbf{X}_{.j.} = \mathbf{C}_{K\times F}\mathrm{diag}(\mathbf{b}_j)\mathbf{A}^{\mathrm{T}}_{I\times F} + \mathbf{E}_{.j},$$
$$j = 1, 2, \ldots, J \quad (2)$$

or

$$\mathbf{X}_{i..} = \mathbf{B}_{J\times F}\mathrm{diag}(\mathbf{a}_i)\mathbf{C}^{\mathrm{T}}_{K\times F} + \mathbf{E}_{i.},$$
$$i = 1, 2, \ldots, I \quad (3)$$

Along with the above model, Harshman [3] and Carroll and Chang [4] proposed an alternating least squares approach (often referred to as PARAFAC) to solve the problem by successively assuming the loading matrixes in two modes known and estimating the unknown parameters of the last mode, which minimizes the following objective function.

$$S(\mathbf{A},\mathbf{B},\mathbf{c}_1, \cdots, \mathbf{c}_K) = \sum_{k=1}^{K} \| \mathbf{X}_{..k} - \mathbf{A}\mathrm{diag}(\mathbf{c}_k)\mathbf{B}^{\mathrm{T}} \|^2_{\mathrm{F}} \quad (4)$$

where $\| \cdot \|_{\mathrm{F}}$ represents the Frobenius matrix norm.

Though PARAFAC has many merits in statistic sense such as optimal unbiased estimations of the final results in least square sense, it suffers from low convergence rate and the requirement of an accurate estimation of the number of factors in mixture. The true and estimated loading matrixes **A**, **B** and **C** will coincide only when the number of factors has been correctly estimated [3,25,26]. As mentioned in the introduction section, the slow convergence is mainly resulted from the high multicollinearity in chemical data arrays, which makes the response surface of the loss function extremely flat, i.e. considerable differences in the estimated parameters induce only a slight difference in terms of loss function value. Therefore, a straightforward way to speed the optimizing procedure is to change the response behavior of the loss function by introducing some reasonable penalty terms [23]. As pointed out by Kiers [24], it requires a certain amount of tinkering to choose penalty parameters. Another possible approach is to optimize some other objective functions with favorable response surface instead of that adopted in PARAFAC. Certainly, the objective functions employed should have intrinsic relationship with the trilinear model and their solutions should be equivalent to the actual underlying loading matrixes. In the following sections, we will study the feasibility of such an idea.

For a trilinear model $\mathbf{X}_{..k} = \mathbf{A}_{I\times F}\mathrm{diag}(\mathbf{c}_k)\ \mathbf{B}^{\mathrm{T}}_{J\times F} + \mathbf{E}_{..k}$ $(k = 1, 2, \ldots, K)$, the following equations hold.

$$\mathbf{A}^{+}\mathbf{X}_{..k} = \mathrm{diag}(\mathbf{c}_k)\mathbf{B}^{\mathrm{T}} + \mathbf{A}^{+}\mathbf{E}_{..k},$$
$$\mathbf{X}_{..k}(\mathbf{B}^{+})^{\mathrm{T}} = \mathbf{A}\mathrm{diag}(\mathbf{c}_k) + \mathbf{E}_{..k}(\mathbf{B}^{+})^{\mathrm{T}},$$
$$k = 1, 2, \ldots, K, \quad (5)$$

Similarly,

$$\mathbf{B}^{+}\mathbf{X}_{i..} = \mathrm{diag}(\mathbf{a}_i)\mathbf{C}^{\mathrm{T}} + \mathbf{B}^{+}\mathbf{E}_{i..},$$
$$\mathbf{X}_{i..}(C^{+})^{\mathrm{T}} = \mathbf{B}\mathrm{diag}(\mathbf{a}_i) + \mathbf{E}_{i..}(\mathbf{C}^{+})^{\mathrm{T}},$$
$$i = 1, 2, \ldots, I, \quad (6)$$
$$\mathbf{C}^{+}\mathbf{X}_{.j.} = \mathrm{diag}(\mathbf{b}_j)\mathbf{A}^{\mathrm{T}} + \mathbf{C}^{+}\mathbf{E}_{.j},$$
$$\mathbf{X}_{.j.}(\mathbf{A}^{+})^{\mathrm{T}} = \mathbf{C}\mathrm{diag}(\mathbf{b}_j) + \mathbf{E}_{.j.}(\mathbf{A}^{+})^{\mathrm{T}},$$
$$j = 1, 2, \ldots, J, \quad (7)$$

where + symbolizes the Moore–Penrose generalized inverse.

With a view to establish an algorithm with high convergence rate, the present authors consider the possibility to solve the trilinear model by alternatively optimizing the following three objective functions rather than that utilized in PARAFAC.

$$S(\mathbf{C}) = \sum_{k=1}^{K} \Big(\| \left(\mathbf{A}^{+}\mathbf{X}_{..k} - \mathrm{diag}(\mathbf{c}_k)\mathbf{B}^{\mathrm{T}}\right)^{\mathrm{T}} \times \mathrm{diag}\left(\mathrm{sqrt}\left(\mathbf{1}./\mathrm{diagm}(\mathbf{B}^{\mathrm{T}}\mathbf{B})\right)\right) \|^2_{\mathrm{F}} + \| \left(\mathbf{X}_{..k}(\mathbf{B}^{+})^{\mathrm{T}} - \mathbf{A}\mathrm{diag}(\mathbf{c}_k)\right) \times \mathrm{diag}\left(\mathrm{sqrt}\left(\mathbf{1}./\mathrm{diagm}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\right)\right) \|^2_{\mathrm{F}} \Big) \quad (8)$$

(with **A** and **B** fixed, minimizing $S(\mathbf{C})$ to obtain **C**)

$$S(\mathbf{B}) = \sum_{j=1}^{J} \Big(\| \left(\mathbf{C}^{+}\mathbf{X}_{.j.} - \mathrm{diag}(\mathbf{b}_j)\mathbf{A}^{\mathrm{T}}\right)^{\mathrm{T}} \times \mathrm{diag}\left(\mathrm{sqrt}\left(\mathbf{1}./\mathrm{diagm}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\right)\right) \|^2_{\mathrm{F}} + \| \left(\mathbf{X}_{.j.}(\mathbf{A}^{\mathrm{T}})^{+} - \mathbf{C}\mathrm{diag}(\mathbf{b}_j)\right) \times \mathrm{diag}\left(\mathrm{sqrt}\left(\mathbf{1}./\mathrm{diagm}(\mathbf{C}^{\mathrm{T}}\mathbf{C})\right)\right) \|^2_{\mathrm{F}} \Big) \quad (9)$$

(with **A** and **C** fixed, minimizing $S(\mathbf{B})$ to renew **B**)

$$S(\mathbf{A}) = \sum_{i=1}^{I} \Big(\| \left(\mathbf{B}^{+}\mathbf{X}_{i..} - \mathrm{diag}(\mathbf{a}_i)\mathbf{C}^{\mathrm{T}}\right)^{\mathrm{T}} \times \mathrm{diag}\big(\mathrm{sqrt}\big(\mathbf{1}./\mathrm{diagm}(\mathbf{C}^{\mathrm{T}}\mathbf{C})\big)\big) \|_{\mathrm{F}}^{2} + \| \left(\mathbf{X}_{i..}(\mathbf{C}^{\mathrm{T}})^{+} - \mathbf{B}\mathrm{diag}(\mathbf{a}_i)\right) \times \mathrm{diag}\big(\mathrm{sqrt}\big(\mathbf{1}./\mathrm{diagm}(\mathbf{B}^{\mathrm{T}}\mathbf{B})\big)\big) \|_{\mathrm{F}}^{2} \Big) \quad (10)$$

(with **B** and **C** fixed, minimizing $S(\mathbf{A})$ to estimate **A**)

Here, **1** is an identity vector with appropriate size; the operator ''./'' denotes array division, for example, $\mathbf{x} = (x_i)$, $\mathbf{y} = (y_i)$, then $\mathbf{x}./\mathbf{y} = (x_i/y_i)$; sqrt(·) is a square root operator. The three empirically selected diagonal matrixes, diag(sqrt(**1**./diagm($\mathbf{A}^{\mathrm{T}}\mathbf{A}$))), diag(sqrt(**1**./diagm($\mathbf{B}^{\mathrm{T}}\mathbf{B}$))) and diag(sqrt(**1**./diagm($\mathbf{C}^{\mathrm{T}}\mathbf{C}$))) function as weight matrixes to balance two parts of each objective function.

The reason for us to alternatively minimize the above three different objective functions is that they have strong intrinsic relationships, and their solutions are equal to the actual underlying loading matrixes of the error-free trilinear model. With the presence of noise, the three objective functions have different response surfaces. Thus it can be expected that alternatively minimizing the three objective functions may avoid possible ''swamp areas,'' which can hardly be circumvented by optimizing only one loss function as in PARAFAC.

Based on the above three objective functions, the present authors developed the following algorithm called self-weighted alternating trilinear decomposition (SWATLD).

1. Randomly initialize loading matrixes **A** and **B**.

2.

$$\mathbf{c}_k = \frac{1}{2}\Big(\mathrm{diagm}\big(\mathbf{B}^{+}\mathbf{X}_{..k}^{\mathrm{T}}\mathbf{A}\big)./\mathrm{diagm}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) + \mathrm{diagm}\big(\mathbf{A}^{+}\mathbf{X}_{..k}\mathbf{B}\big)./\mathrm{diagm}(\mathbf{B}^{\mathrm{T}}\mathbf{B})\Big) \quad k = 1, \ldots, K \quad (11)$$

3.

$$\mathbf{b}_j = \frac{1}{2}\Big(\mathrm{diagm}\big(\mathbf{A}^{+}\mathbf{X}_{.j.}^{\mathrm{T}}\mathbf{C}\big)./\mathrm{diagm}(\mathbf{C}^{\mathrm{T}}\mathbf{C}) + \mathrm{diagm}\big(\mathbf{C}^{+}\mathbf{X}_{.j.}\mathbf{A}\big)./\mathrm{diagm}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\Big) \quad j = 1, \ldots, J \quad (12)$$

4.

$$\mathbf{a}_i = \frac{1}{2}\Big(\mathrm{diagm}\big(\mathbf{C}^{+}\mathbf{X}_{i..}^{\mathrm{T}}\mathbf{B}\big)./\mathrm{diagm}(\mathbf{B}^{\mathrm{T}}\mathbf{B}) + \mathrm{diagm}\big(\mathbf{B}^{+}\mathbf{X}_{i..}\mathbf{C}\big)./\mathrm{diagm}(\mathbf{C}^{\mathrm{T}}\mathbf{C})\Big) \quad i = 1, \ldots, I \quad (13)$$

5. Update **C**, **B** and **A** according to steps 2–4, until certain stop criterion has been reached.

SWATLD is unique in the way that it tries to avoid possible ''swamp areas'' by alternatively minimizing three different objective functions with intrinsic relationships. It is obvious that the optimizing procedure of SWATLD may not be a monotonically decreasing one. Simulation studies, therefore, will be carried out in the following sections to investigate the performance of this new method.

4. Experimental

In this paper, several data arrays including real HPLC-DAD data arrays and simulated ones have been used to demonstrate the performance of the new algorithm proposed. Since most of the sophisticated methods mentioned in the Introduction employ some compression techniques to speed the optimizing procedure, which can also be combined into the proposed method in this paper, while the rest require much experience to choose some penalty parameters, the proposed method has, therefore, been tested by comparing it with PARAFAC only.

4.1. Simulated data arrays

4.1.1. Randomly simulated data arrays

In order to investigate the feasibility and accelerating capacity of this unique optimizing approach of SWATLD, 20 noise-free data arrays with size $20 \times 20 \times 20$ were simulated. For each data array, underlying loading matrixes **A**, **B**, **C** of order 20×3 were randomly constructed. Their elements were drawn from uniform (2,4) distribution. Loading matrixes **A**

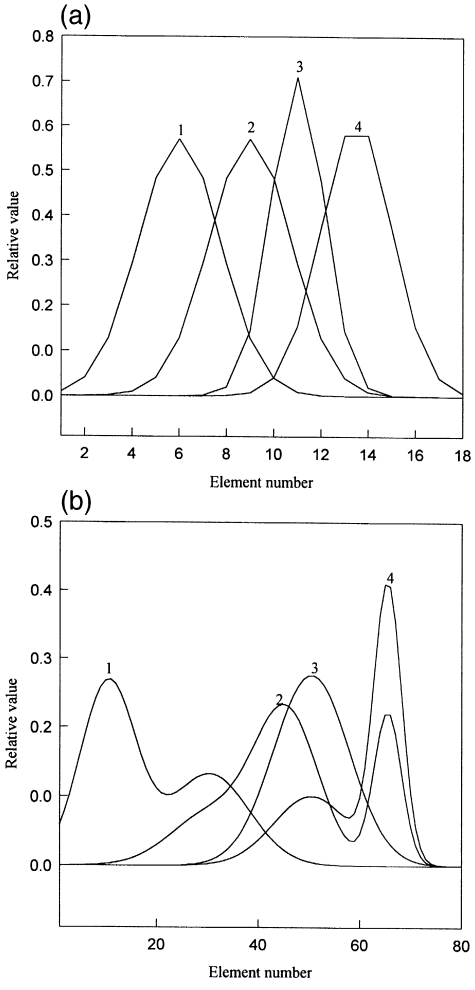

Fig. 2. The normalized chromatographic (a) and spectral (b) profiles of the simulated HPC-DAD type data array.

and **B** were column-wisely normalized to unite length. The condition numbers of **A**, **B** and **C** range from 8 to 11. Homoscedastic random noises generated with zero expectation and four different levels of standard deviations from 0.1% to 2% of the maximum absorbance were added to each data array which were constructed as follows.

$$\mathbf{X}^{p}_{..k} = \mathbf{A}\mathrm{diag}(\mathbf{c}_k)\mathbf{B}^{\mathrm{T}} \qquad k = 1, 2, \ldots, K$$

$$\mathbf{E}_{..k} = \mathrm{Max}(\mathbf{X}^{p}_{..k}) \times \mathbf{randn} \times a_{\mathrm{noise}},$$

$$k = 1, 2, \ldots, K$$

$$\mathbf{X}_{..k} = \mathbf{X}^{p}_{..k} + \mathbf{E}_{..k}, \qquad k = 1, 2, \ldots, K$$

where $\mathbf{X}^{p}_{..k}$ represents the pure data set (without noise) of kth sample, **randn** is a matrix with appropriate size and random entries chosen from a normal distribution with mean zero and variance one, a_{noise} a parameter controlling the noise level, and Max $(\mathbf{X}^{p}_{..k})$, the maximal entry of matrix $\mathbf{X}^{p}_{..k}$.

4.1.2. Simulated HPLC-DAD type data arrays

With a view to study the results' qualities of SWATLD, a four-component system was simulated in HPLC-DAD data structure with cond(**A**) = 2.82, cond(**B**) = 4.1 and cond(**C**) = 11.90. The simulated pure chromatographic (loading matrix **A**) and spectral profiles (loading matrix **B**) of the four components are depicted in Fig. 2a and b, respectively. Six concentration levels of each of the four components were set randomly to construct six samples, which were assembled to form a three-way data array. Then homoscedastic random noises generated with zero expectation and three different levels of standard deviations from 0.5% to 2% of the maximum absorbance were added.

$$\mathbf{X}^{p}_{..k} = (\mathbf{a}_1\mathbf{a}_2\mathbf{a}_3\mathbf{a}_4)\mathrm{diag}(\mathbf{c}_k)(\mathbf{b}_1\mathbf{b}_2\mathbf{b}_3\mathbf{b}_4)^{\mathrm{T}}$$

$$k = 1, 2, \ldots, K$$

$$\mathbf{E}_{..k} = \max(\mathbf{X}^{p}_{..k}) \times \mathbf{randn} \times a_{\mathrm{noise}}$$

$$k = 1, 2, \ldots, K$$

$$\mathbf{X}_{..k} = \mathbf{X}^{p}_{..k} + \mathbf{E}_k \qquad k = 1, 2, \ldots, K$$

where, $\mathbf{a}_i$ and $\mathbf{b}_j$ are the pure chromatographic and spectral profiles of ith component, $\mathbf{c}_k$ is a row vector with elements equal to the concentration levels of the four components in the kth sample.

Table1
The concentrations of *o*-dichlorobenzen, *p*-chlorotoluene and *o*-chlorotoluene in nine real HPLC-DAD samples

Sample	Concentration ($\mu g \cdot ml^{-1}$)		
	o-Dichlorobenzene	*p*-Chlorotoluene	*o*-Chlorotoluene
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#	152.0	12.6	15.2
6#	15.2	12.6	152.0
7#	60.8	25.2	91.2
8#	91.2	50.4	30.4
9#	30.4	75.6	60.8

4.2. Real 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 are listed in Table 1. The corresponding nine data sets, recorded by HPLC with diode array detector at same conditions, were used to construct three arrays. Data array I was assembled by data sets $1^{\#}$, $2^{\#}$, $3^{\#}$, $4^{\#}$, $6^{\#}$ and $7^{\#}$; data array II, by data sets $1^{\#}$–$6^{\#}$; data array III, by data sets $1^{\#}$, $2^{\#}$, $7^{\#}$ and $9^{\#}$. The last data sets of the three data arrays were considered as unknown samples, the rest as calibration samples. In sample $7^{\#}$ of data array I and sample $6^{\#}$ of data array II, *o*-dichlorobenzene and *o*-chlorotoluene were considered as interferences, all the compounds in sample $9^{\#}$ of data array III are components of interest. For detailed information on these experiments readers are referred to ref. [27].

5. Results and discussion

5.1. 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 Windows 98 operating system.

5.2. The implementation of the algorithms

For each data array, random initialization was carried out to start the iterative optimizing procedures of both SWATLD and PARAFAC. The optimizing procedures of both SWATLD and PARAFAC are terminated when the following criterion reaches a certain threshold ε ($\varepsilon = 1 \times 10^{-6}$ in the present paper). A maximal number of 5000 iterations are adopted to avoid possible undue slow convergence.

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

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

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

5.3. Results and discussion

5.3.1. The convergence property and accelerating capacity of SWATLD

For all the 20 randomly simulated data arrays, four different noise levels were added and each noise level has been replicated five times. Therefore, there are actually a total of 400 data arrays. For each data array, 10 runs of both SWATLD and PARAFAC with random initialization were carried out. All of the 4000 runs of SWATLD have converged to satisfactory results within much fewer iterations than PARAFAC. Table 2 lists the typical performances of SWATLD and PARAFAC for a randomly simulated data array with severe multicollinearity. At low noise level $a_{\mathrm{noise}} = 0.001$, out of the 10 trials, PARAFAC has not converged within 5000 iterations for five times. With

Table2
The average number of iterations and computation time of SWATLD and PARAFAC for a randomly simulated three-component data arrays (cond(**A**) = 11.35, cond(**B**) = 9.83 and cond(**C**) = 10.58) with four different noise levels

a_{noise}	Iterations and time					
	SWATLD			PARAFAC		
	Min	Max	Average[a]	Min	Max	Average
0.001	85 (3.7s)	172 (7.6s)	116 (5.1s)	2864 (116.4s)	5000 (213.8s)[b]	4394 (179.4s)
0.005	74 (3.2s)	115 (5.4s)	89 (3.9s)	1348 (54.3s)	3257 (131.6s)	2495 (100.8s)
0.01	55 (2.4s)	92 (4.1s)	72 (3.2s)	171 (6.9s)	1900 (76.6s)	911 (36.7s)
0.02	52 (2.3s)	110 (4.8s)	70 (3.1s)	88 (3.6s)	1495 (60.4s)	430 (17.3s)

[a] The values listed are averaged over 10 trials with random initialization.
[b] Out of the 10 trials, PARAFAC has not been converged with 5000 iterations for five times.

the increase of the noise level, the average convergence rate of PARAFAC has a corresponding increase tendency, which is consistent with the observations of Kiers [[24]]. However, even if the noise level is as high as $a_{noise} = 0.02$, PARAFAC still requires many iterations to converge. In contrast, SWATLD can converge much faster. Even at low noise level $a_{noise} = 0.001$, SWATLD requires only an average number of 116 iterations (i.e. 5.1 s) to find the correct results, which is at least 37 times in terms of iterations and 35 times in terms of computation time faster than PARAFC. A similar increase tendency of the convergence rate is also observed with the increase of the noise level. However, the influence of the noise level on the convergence rate of SWALTD is not so striking compared with PARAFAC. In conclusion, either in terms of iterations or computation time, SWATLD has a faster convergence rate than PARAFAC. All these results have demonstrated that the unique optimizing scheme of SWATLD is feasible and can speed the convergence rate through alternatively minimizing three different objective functions with intrinsic relationships to avoid possible ''swamp areas.''

5.3.2. The property of SWATLD being insensitive to the excess factors used in calculation

Other than the fast convergence rate, SWATLD has another attractive feature of being insensitive to the excess factors used in calculation (Table 3). For a simulated four-component HPLC-DAD data array with noise level $a_{noise} = 0.01$, the increase of N from 4 to 6 greatly influenced the results' qualities of PARAFAC. For $N > F$, the results of PARAFAC are the linear combinations of the actual underlying factors instead of the actual underlying factors themselves. Interestingly, the increase of N seems have hardly influences on the results of SWATLD. For instance, when $N = 6$, the four columns of the loading matrixes resolved by SWATLD coincide with those of the true underlying loading matrixes (Fig. 3). The excess columns represent noise, which can be easily discriminated from the desired ones. In many other simulations, it has also been observed that provided $N \geq F$, SWATLD can always find the true solutions rather than their linear combinations (the theoretical base of this feature is out of the scope of this paper, and will be discussed in detail elsewhere). It means that SWATLD does not require an accurate estimation of the number of factor in mixtures. Only a large gross estimation can guarantee the correctness of the results, which will relieve the analysts from the troublesome determination of the number of factors.

5.3.3. The influence of noise level on the results' qualities obtained by SWATLD

Theoretically, for noise-free data array, the results obtained by SWATLD should be the same as those

Table 3
The influence of N on the final results obtained by SWATLD and PARAFAC for a simulated four-component HPLC-DAD type data array with noise level $a_{noise} = 0.01$

Mode		4		5		6	
		SWATLD	PARAFAC	SWATLD	PARAFAC	SWATLD	PARAFAC
A	$\mathbf{a}_1$	0.9999[a]	1.0000	0.9999	0.9843	0.9998	0.9696
	$\mathbf{a}_2$	0.9999	1.0000	0.9998	0.9990	0.9998	0.9827
	$\mathbf{a}_3$	0.9999	1.0000	0.9998	0.8930	0.9997	0.9823
	$\mathbf{a}_4$	0.9999	0.9999	1.0000	0.9566	1.0000	0.9779
B	$\mathbf{b}_1$	0.9996	0.9999	0.9993	0.8840	0.9993	0.9536
	$\mathbf{b}_2$	0.9997	0.9998	0.9996	0.8212	0.9996	0.8889
	$\mathbf{b}_3$	0.9997	0.9999	0.9998	0.9713	0.9997	0.9367
	$\mathbf{b}_4$	0.9998	0.9999	0.9998	0.9879	0.9998	0.8519
C	$\mathbf{c}_1$	1.0000	1.0000	1.0000	0.9362	1.0000	0.9988
	$\mathbf{c}_2$	1.0000	1.0000	1.0000	0.9970	1.0000	0.9901
	$\mathbf{c}_3$	1.0000	1.0000	1.0000	0.8487	1.0000	0.9907
	$\mathbf{c}_4$	1.0000	1.0000	1.0000	0.9759	1.0000	0.8392

[a] 0.9999 is the related coefficient between the resolved and the true profile, averaged over four runs with random initialization. For the convenience of presentation, all the related coefficients in this paper have only four significant digits after the decimal point.

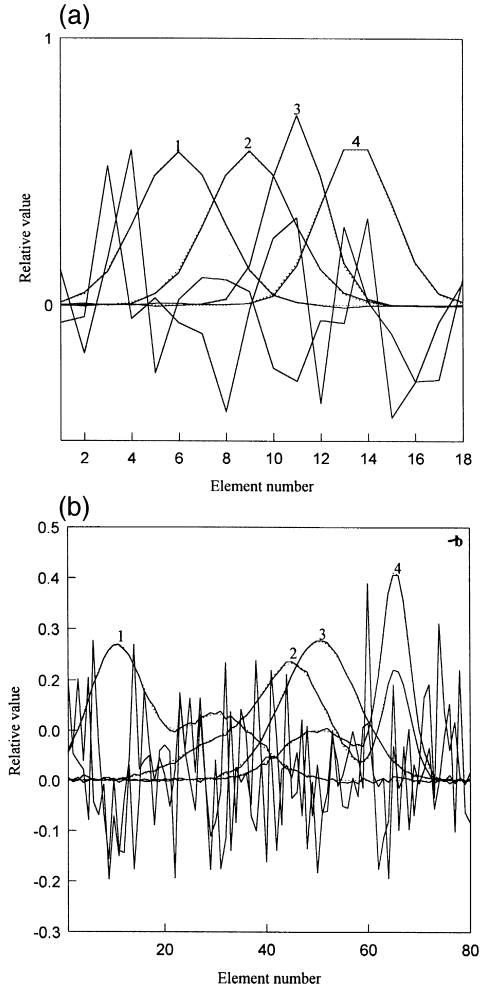

Fig. 3. The true (dotted line) and resolved (solid line) profiles (a: normalized chromatographic profiles, b: normalized spectral profiles) by SWTLD ($N = 6$) for the simulated HPLC-DAD type data array with $a_{noise} = 0.01$.

supplied by PARAFAC. In real world, noise always presents in chemical data arrays. Since different objective functions rather than that employed in PARAFAC were utilized in SWATLD, the influence of the noise level on the results' qualities should, therefore, be carefully investigated. A comparison between the results obtained by SWATLD and PARAFAC for a simulated four-component HPLC type data array with three different noise levels is shown in Table 4. It can be seen that when at relatively low noise level $a_{noise} = 0.005$, the results of SWATLD are in perfect consistence with the real underlying loading matrixes. The related coefficients between the resolved and true profiles are no less than 0.9999. When a_{noise} increases to 0.01, the results' qualities of SWATLD are still very satisfactory with the lowest related coefficient equal to 0.9996. Further increase of the noise level will deteriorate the results' qualities to some extent. However, even when the noise level a_{noise} is as high as 0.02, the results of SWATLD are still acceptable, though they are slightly inferior to those of PARAFAC. In practice, noise level seldom exceeds 2% of the maximal absorbance of the data sets, so SWATLD is suitable for practical use in chemometrics, which will be demonstrated by real HPLC-DAD data arrays in the forthcoming sectors.

5.3.4. The results of HPLC-DAD data arrays

The results of both SWATLD and PARAFAC for the three HPLC-DAD data arrays are shown in Tables 5, 6 and 7, respectively. From all the three tables, one can see that with N taking a value of 3, both algorithms converged comparatively fast. However, the increase of N put heavy computation burden on PARAFAC. When N exceeds certain value, PARAFAC can not converge within 5000 iterations. For all the data arrays, SWATLD generally converged within several seconds. The longest time needed by SWATLD is 16.7 s. Furthermore, the value of N just slightly affects the computation time required. In terms of either computation time or iterations, SWATLD converged obviously much faster than PARAFAC, and its performance with respect to N is much more stable.

Another advantage of SWATLD over PARAFAC is the property of being insensitive to the excess factors used in calculation. This conclusion is drawn from simulation studies, and corroborated by the three real HPLC-DAD arrays. For data array I, though there are only three chemical species, the sharp decreases of the relative deviations of the two algorithms with the increase of N from 3 to 4 may suggest data array I is a four-factor system. The extra factor may denote the presence of background or other possible interference. The content of *p*-chlorotoluene predicted by SWATLD with N taking values of 4, 5 and 6, are satisfactory and stable. The predicted contents with relative deviations ranging from 0.4% to 2.8% are virtually equivalent. While PARAFAC can obtain an

Table 4
The influence of noise level a_{noise} on the final results obtained by SWATLD and PARAFAC with $N = 4$ for a simulated four-component HPLC-DAD type data array

Mode		0.005		0.01		0.02	
		SWATLD	PARAFAC	SWATLD	PARAFAC	SWATLD	PARAFAC
A	$\mathbf{a}_1$	1.0000	1.0000	0.9999	1.0000	0.9997	0.9999
	$\mathbf{a}_2$	1.0000	1.0000	0.9999	1.0000	0.9996	0.9999
	$\mathbf{a}_3$	1.0000	1.0000	0.9999	1.0000	0.9995	0.9997
	$\mathbf{a}_4$	1.0000	1.0000	0.9999	0.9999	0.9997	0.9999
B	$\mathbf{b}_1$	0.9999	1.0000	0.9996	0.9999	0.9981	0.9996
	$\mathbf{b}_2$	0.9999	1.0000	0.9997	0.9999	0.9987	0.9993
	$\mathbf{b}_3$	0.9999	1.0000	0.9997	0.9999	0.9988	0.9996
	$\mathbf{b}_4$	1.0000	1.0000	0.9998	0.9999	0.9992	0.9996
C	$\mathbf{c}_1$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	$\mathbf{c}_2$	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	$\mathbf{c}_3$	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997
	$\mathbf{c}_4$	1.0000	1.0000	1.0000	1.0000	0.9998	0.9998

acceptable result only when N is equal to 4, the exact number of the underlying factors. With excess factors used, PARAFAC converged to erroneous results (not shown in Table 5). The perfect agreement among the concentrations predicted by SWATLD with different N values of 3, 4 and 5 for data array III further verified that SWATLD is immune from the influence of the excess factors used in calculation.

Table 5
Concentration ($\mu g \cdot ml^{-1}$) of *p*-chlorotoluene in sample $7^{\#}$ of data array I estimated by SWATLD and PARAFAC with N taking 3, 4, 5 and 6

Algorithm	N	Known	Predicted	Iterations and time
SWATLD	3	25.2	22.1 (12.3%)[a]	19 (0.8s)
PARAFAC	3	25.2	20.4 (19.0%)	311 (7.2s)
SWATLD	4	25.2	25.6 (1.6%)	55 (2.8s)
PARAFAC	4	25.2	23.4 (7.1%)	2327 (58.3s)
SWATLD	5	25.2	25.1 (0.4%)	141 (7.5s)
PARAFAC	5	25.2	–[b]	5000 (143.5s)[c]
SWATLD	6	25.2	24.5 (2.8%)	137 (7.7s)
PARAFAC	6	25.2	–	5000 (152.1s)

[a] For real HPLC-DAD data arrays, all the values listed are averaged over five runs with random initialization. The value in the bracket is the absolute value of the relative deviation.

[b] The bar ''–'' means when the algorithm stopped within 5000 iterations, the loading matrixes obtained have no physical meanings. Therefore the contents of the interested compounds cannot be determined.

[c] The underline means the algorithm did not converge within 5000 iterations.

As far as the accuracy of the predicted concentration is concerned, the performance of SWATLD is also very satisfactory. For data array I, the relative deviation of the concentration of *p*-chlorotoluene calculated by SWATLD with N taking 4 is 1.6% while that of PARAFAC is 7.1%. An even better result with relative deviation of 0.4% was found by SWATLD when $N = 5$. Coincidentally, SWATLD with N equal to 5 gave the best result with relative deviation of 0.8% for data array II. The results of SWATLD for data array III are more than acceptable. All the contents of the three components are accurately determined. Surprisingly, for all the three data arrays, the results of PARAFAC are inferior to those of SWATLD, though it can not thus to conclude that SWATLD is superior to PARAFAC in term

Table 6
Concentration ($\mu g \cdot ml^{-1}$) of *p*-chlorotoluene in sample $6^{\#}$ of data array II estimated by SWATLD and PARAFAC with N taking 3, 4, 5 and 6

Algorithm	N	Known	Predicted	Iterations and time
SWATLD	3	12.6	15.9 (26.2%)	30 (1.4s)
PARAFAC	3	12.6	14.6 (15.9%)	220 (5.1s)
SWATLD	4	12.6	14.3 (13.5%)	118 (5.9s)
PARAFAC	4	12.6	14.8 (17.1%)	1171 (29.3s)
SWATLD	5	12.6	12.7 (0.8%)	314 (16.7s)
PARAFAC	5	12.6	–	5000 (144.1s)
SWATLD	6	12.6	13.6 (7.9%)	181 (10.2s)
PARAFAC	6	12.6	–	5000 (156.5s)

Table 7
Concentrations ($\mu g \cdot ml^{-1}$) of *o*-dichlorobenzene (ODB), *p*-chlorotoluene (PCT) and *o*-chlorotoluene (OCT) in sample 9# of data array III estimated by SWATLD and PARAFAC with N taking 3, 4 and 5

Compound	Known	Predicted by SWATLD			Predicted by PARAFAC		
		3	4	5	3	4	5
ODB	30.4	32.2 (5.9%)	30.7 (1.0%)	30.9 (1.5%)	32.1 (5.6%)	33.5 (10.2%)[a]	–
PCT	75.6	77.7 (2.8%)	77.6 (2.7%)	76.8 (1.6%)	76.2 (0.8%)	75.8 (0.3%)	–
OCT	60.8	61.7 (1.5%)	62.0 (2.0%)	62.7 (3.2%)	60.8 (0.0%)	60.3 (0.8%)	–
Iterations		23	39	71	184	703	5000
Time (s)		(1.0s)	(1.8s)	(3.5s)	(3.2s)	(13.3s)	(112.5s)

[a] Out of the five runs, only three have been converged with 5000 iterations, the values listed are based on the three trials.

of predictive ability, it at least demonstrates that SWATLD is suitable for real data arrays.

It was observed in experiments that there is a symmetry assumption on data array analyzed by SWATLD, i.e. all the ranks of the loading matrixes **A**, **B** and **C** should be no less than F. The intrinsic reasons introducing such symmetric constraint is unclear now. In quantitative analysis, the loading matrixes **A** and **B** are chromatographic or spectral profiles of chemical compounds, they generally satisfy the above constraint. In order to make loading matrix **C** also satisfy the requirement, one should prepare more calibration samples (no less than the number of components in calibration samples).

6. Conclusions

A novel optimizing scheme for trilinear decomposition is proposed. It is unique in that it alternatively minimizes three different objective functions related to the trilinear model. Based on the proposed scheme, a new algorithm called self-weighted alternating trilinear decomposition (SWATLD) was therefore contrived for second-order linear calibration. Simulations have been carried out to investigate the performance of SWATLD. Experiments demonstrated that SWATLD has the properties of fast convergence and being insensitive to the excess factors used in calculation. To be specific, SWATLD can generally converge within 100 iterations. Compared with PARAFAC that always requires several hundreds and thousands of iterations to converge, SWATLD is much more efficient either in terms of iterations or computation time. For large data arrays, compression techniques can be combined with SWATLD to further accelerate the optimizing procedure. The property of being insensitive to excess factors used in calculation means that SWATLD can decompose data arrays into the actual underlying loading matrixes as long as $N \geq F$. It is, therefore, unnecessary to spend a lot of effort to accurately determine the number of underlying factors in the system studied as it is in PARAFAC. Only a large rough estimation can guarantee SWATLD to find the correct results. As far as the deviations of the results are concerned, SWATLD can offer satisfactory results under moderate noise level. This conclusion drawn from simulation studies was corroborated by three real HPLC-DAD data arrays.

Since almost all the properties of SWALD were drawn from experiments, our future work will pursue to provide some more or less rigorous mathematical explanations on these properties as well as the conditions required by SWATLD in analyzing three-way data arrays.

Acknowledgements

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

References

[1] R.Q. Yu, Introduction to Chemometrics, Hunan Education Publishing House, Changsha, 1991.
[2] Y.Z. Liang, O.M. Kvalheim, R. Manne, Chemom. Intell. Lab. Syst. 18 (1993) 235.
[3] R.A. Harshman, UCLA Working Papers in Phonetics 16 (1970) 1.
[4] J.D. Carrol, J. Chang, Psychometrika 35 (1970) 283.

[5] P. Geladi, Chemom. Intell. Lab. Syst. 7 (1989) 11.
[6] A.K. Smilde, Chemom. Intell. Lab. Syst. 5 (1992) 143.
[7] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149.
[8] H.A.L. Kiers, J. Chemom. 12 (1998) 155.
[9] E.R. Malinowski, in: B. Kowalski (Ed.), Chemometrics: Theory and Application, ACS Symp. Ser. 52 American Chemical Society, Washington, DC, 1977, Chap. 3.
[10] E.R. Malinowski, D.G. Howery, Factor Analysis in Chemistry, Wiley-Interscience, New York, 1980.
[11] H.B. Woodruff, P.C. Tway, L.J. Cline Love, Anal. Chem. 53 (1981) 81.
[12] S. Wold, Technometrics 20 (1987).
[13] D.P. Herman, M.F. Gonnord, G. Guiochon, Anal. Chem. 56 (1984) 995.
[14] Z.P. Chen, Y.Z. Liang, J.H. Jiang, Y. Li, J.Y. Qian, R.Q. Yu, J. Chemom. 13 (1999) 15.
[15] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom., in press.
[16] H.A.L. Kiers, W.P. Krijnem, Psychometrica 56 (1991) 147.
[17] B.K. Alsberg, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 24 (1994) 31.
[18] B.K. Alsberg, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 24 (1994) 43.
[19] H.A.L. Kiers, R.A. Harshman, Chemom. Intell. Lab. Syst. 39 (1997) 31.
[20] R. Bro, C.A. Andersson, Chemom. Intell. Lab. Syst. 42 (1998) 105.
[21] C.A. Andersson, R. Bro, Chemom. Intell. Lab. Syst. 42 (1998) 93.
[22] B.C. Mitchell, D.S. Burdick, Chemom. Intell. Lab. Syst. 34 (1996) 85.
[23] P. Paatro, Chemom. Intell. Lab. Syst. 38 (1997) 223.
[24] H.A.L. Kiers, J. Chemom. 12 (1998) 155.
[25] J.B. Kruskal, Linear Algebra Appl. 18 (1977) 95.
[26] J.B. Kruskal, Psychometrika 41 (1976) 281.
[27] H.L. Wu, M. Shibukawa, K. Oguma, J. Chemom. 12 (1998) 1.