ELSEVIER

Analytica Chimica Acta 444 (2001) 295–307

ANALYTICA
CHIMICA
ACTA

www.elsevier.com/locate/aca

Efficient way to estimate the optimum number of factors for trilinear decomposition

Zeng-Ping Chen, Zhuo Liu, Yu-Zhen Cao, Ru-Qin Yu*

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

Received 7 December 2000; received in revised form 30 May 2001; accepted 30 May 2001

Abstract

In trilinear decomposition, one first tries to estimate the number of underlying factors in the system studied, and then employs trilinear decomposition methods such as PARAFAC to obtain the desired characteristic profiles of the underlying factors and their relative contributions. Since the results of PARAFAC are heavily dependent on the estimation of the underlying factors, either overestimation or underestimation of the underlying factors will lead the results of PARAFAC to be erroneous. Most of the existing factor-determining methods are established on the basis of factor analysis. These procedures are originally designed for two-way data sets. Only after the three-way data array was unfolded into two-way data set, could then these factor-determining methods be used. It is obvious that the trilinear character of the data array is not utilized in the factor-determining procedure. With a view to cope with non-ideal experimental conditions, such as heavy collinearity and varying backgrounds, the present authors advocated incorporating the advantages of trilinear data array into the factor-determining procedure. Hence, a novel factor-determining method has been proposed specifically for trilinear decomposition. Experiments have demonstrated that the proposed method has the features of easy implementation and excellent performance even when heavy collinearity and varying backgrounds are present. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Trilinear decomposition; Factor-determining methods; PARAFAC; ADD-ONE-UP

1. Introduction

Driven by the demands for more information to be extracted from data, the last two decades have seen a rapid development of chemometrics. It is a direct result of the wide application of sophisticated instruments such as excitation–emission matrix fluorescence spectrometer, HPLC–DAD and GC–MS, etc. Trilinear decomposition is one of the most active areas in chemometrics research, due to the so-called second-order advantage [1], i.e. the uniqueness of the decomposition results, regardless of some scaling indeterminacy. Among all the chemometric methods developed for trilinear decomposition [2–11], PARAFAC [2,3] and direct trilinear decomposition [4] are the two representatives for iterative and non-iterative types, respectively. PARAFAC is preferable in practice for its favorable statistic merits (such as optimal unbiased estimations of the final results in least square sense) comparing to non-iterative methods.

As pointed by many authors, however, PARAFAC requires an accurate estimation of the number of

* Corresponding author. Tel.: +86-731-882-4525; fax: +86-731-882-2710.
E-mail address: rqyu@mail.hunu.edu.cn (R.-Q. Yu).

PII: S0003-2670(01)01179-5

underlying factors in the system studied. Either overestimation or underestimation of the number of underlying factors will lead to erroneous results. Therefore, factor determination is of utmost importance in trilinear decomposition.

At present, trilinear data arrays are often unfolded into two-way data sets, and then factor-determining methods [12–18] for handling two-way data are used to estimate the number of underlying factors. Most of these methods are based on factor analysis, in particular SVD, which tries to separate the systematic part of the data, contributed by chemical species, from the residuals, caused by random noise. Homoscedastic noises of normal distribution are always assumed for the successful application of these methods. Generally speaking, they can not discriminate between the chemical information and background. So in practice, the presence of background may cause the estimations of the factor-analysis-based methods to be misleading. Furthermore, for the different empirical or statistical hypotheses employed, different methods may offer different estimations of the number of factors for the same data set, which will embarrass the analysts in drawing conclusion and carrying on the subsequent analysis.

In trilinear decomposition, three-way data arrays are not merely a collection of two-way data sets, but that there is actually an internal relationship between each of the matrixes. The systematic part of the three-way data array has trilinear character, which is not valid for varying backgrounds. It might not be reasonable just to unfold a trilinear data array into a two-way data set, and then estimate the number of underlying factors by factor-determining methods for two-way data as often done in literatures. The advantage of trilinear character provided by the trilinear data array is not appropriately utilized in the above factor-determining procedure. For a reliable estimation of the factors in the system studied, one should take full advantages of the information supplied by the system. The trilinear character of trilinear data arrays will be undoubtedly helpful in the process of factor determination especially when varying backgrounds are present (always the case in practice). To our knowledge, there are few true three-way factor-determining methods, which exploit the trilinear feature of trilinear data arrays. Harshman and Lundy [19] advocated for adopting split-half analysis to determine the most appropriate number of underlying factors for trilinear data arrays. It is essentially a type of jack-knife analysis, where different subsets of the data array are analyzed independently, and the same results are expected for all the subsets when the appropriate factors being used. Split-half analysis involves relatively intricate splitting skills. A poor splitting scheme will impede sound results. Recently, Louwerse et al. [20] had generalized the two-way cross-validation method to three-way array. Like its predecessors [21,22], the generalized version of cross-validation for three-way array also suffers from heavy computation burden. According to the introduction of Louwerse et al. [20], Timmerman and Kiers had proposed an alternative method, which is based on “a systematic comparison of the fit value for modes with different number of components and search for a strong decrease in added fit value by additional components”. Often, however, it is more or less arbitrary to define the degree of “strong decrease”. Bro had suggested a core-consistency-diagnostic approach for determining the appropriate number of components for PARAFAC model [23]. Though the core-consistency-diagnostic method is powerful, it seems that certain threshold value is also required for its auto-implementation. Hence, practical application needs methods with features of simple implementation, tolerable computation time and clear-cut criterion point when the most appropriate factors have been extracted.

With the above beliefs, the present authors design a relatively simple factor-determining procedure specifically for trilinear decomposition (or PARAFAC model). The performance of the proposed approach is demonstrated by both simulated and real data arrays.

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. The letters I, J, K are kept for denoting the dimensions of different modes in three-way arrays; F, for the number of actual underlying factors, and N, for the number of factors used in PARAFAC. $\underline{\boldsymbol{X}}$ represents three-way data array. $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ with dimensions of $I \times F$, $J \times F$, $K \times F$, respectively, are the three underlying loading matrixes of $\underline{\boldsymbol{X}}$. $\boldsymbol{X}_{I\times JK} = [\boldsymbol{X}_{..1}, \boldsymbol{X}_{..2}, \ldots , \boldsymbol{X}_{..K}]$ and $\boldsymbol{X}_{IK\times J} = [\boldsymbol{X}'_{..1}, \boldsymbol{X}'_{..2}, \ldots , \boldsymbol{X}'_{..K}]'$ are the two unfolded matrixes of $\underline{\boldsymbol{X}}$. $\boldsymbol{X}_{..k}$ represents the kth frontal slices of the three-way array $\underline{\boldsymbol{X}}$. The scalar c signifies the number of components retained in the trimmed data set of $\boldsymbol{X}_{I\times JK}$ or $\boldsymbol{X}_{IK\times J}$.

3. Model

In trilinear data array analysis, for each sample, the response recorded by certain instrument is a matrix with its rows and columns representing some characteristic response modes, such as ultraviolet-visible spectra and chromatograms in HPLC–DAD and excitation and emission spectra in fluorescence spectrometer. For no less than two samples, the data sets recorded can be assembled into a three-way data array $\underline{\boldsymbol{X}}$. Its third mode describes the variations of the concentration ratio of compounds in samples. Under the assumption of linear response, which is always considered to be held in quantitative analysis, the three-way array $\underline{\boldsymbol{X}}$ can be decomposed into the following trilinear model:

$$\begin{aligned}\boldsymbol{X}_{..k} &= \boldsymbol{A}_{I\times N}\,\mathrm{diag}(\boldsymbol{c}_k)\boldsymbol{B}'_{J\times N} + \boldsymbol{E}_{..k}\\ k &= 1, 2, \ldots , K\end{aligned} \qquad (1)$$

where $\boldsymbol{X}_{..k}$ and $\boldsymbol{E}_{..k}$ are the kth frontal slices of three-way array $\underline{\boldsymbol{X}}$ and three-way residual array $\underline{\boldsymbol{E}}$, respectively; diag($\boldsymbol{c}_k$) is a diagonal matrix with elements equal to the kth row of loading matrix $\boldsymbol{C}$. The above representation is the famous PARAFAC model. The three underlying loading matrixes $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ can be obtained by an alternating least square algorithm, which is often referred as PARAFAC.

From Eq. (1), it is obvious that before the decomposition, the number of factors used in decomposition, N, should be determined first. Different N values would definitely lead to different version of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$. Only when N coincides with F, the number of the actual underlying factors, would $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ turn to be the actual underlying loading matrixes with physical meanings. Otherwise, their columns would just be the linear combinations of the columns of the corresponding actual underlying loading matrix, which would cause the subsequent analysis to be void. It is, therefore, of utmost importance to develop reliable methods to estimate the factors of data arrays.

4. Problems in factor determination of trilinear data array

Analogous to two-way data analysis, the factors of a trilinear data array are usually determined through unfolding the data array into a two-way data set and then applying two-way factor-determining methods on it. As stated in introduction section, many two-way factor-determining methods can only separate the systematic variance from the residuals due to homoscedastic random noise. They can hardly discriminate between the contributions of chemical species and that of varying backgrounds. Moreover, heavy collinearity is another aspect hindering their successes. In trilinear data arrays, the response of chemical species generally follows trilinear model, while that of varying backgrounds does not. This prior information should be useful in determining the number of underlying factors when varying backgrounds occur. Decomposing the trilinear data array by PARAFAC and then examining the fitness value might be the most straightforward approach to take advantage of this prior information. For a trilinear data array, it is observed that the introduction of superfluous factors in PARAFAC can only slightly improve the fitness value. Though the phenomenon does provide some hints for the determination of the appropriate factors, certain threshold for the fitness value is needed. It will raise problems, since different threshold will be required for different data arrays and there is no general guidance for the threshold setting.

5. A new approach for factors determination — ADD-ONE-UP truncating and fitting method

As pointed out in the previous section, the trilinear feature of trilinear data array would be lost, if the decision is only drawn from the analysis of the unfolded two-way data sets $\boldsymbol{X}_{I\times JK}$ and $\boldsymbol{X}_{IK\times J}$. On the other hand, though applying PARAFAC to the original three-way data array $\underline{\boldsymbol{X}}$ have taken the trilinear character into consideration, the consistent increase in the fitness value with the increase of the factors used raises the problem of threshold value setting. Such embarrassing situation drives us to investigate the feasibility of combining the information provided by the analysis of the unfolded two-way data sets and that of original trilinear data array $\underline{\boldsymbol{X}}$.

The analysis of the unfolded two-way data set $\boldsymbol{X}_{I\times JK}$ or $\boldsymbol{X}_{IK\times J}$ can supply the information of the data configuration (variation distribution). Through decomposing the unfolded two-way data set $\boldsymbol{X}_{I\times JK}$ or $\boldsymbol{X}_{IK\times J}$ by SVD into orthogonal components, one can attempt to truncate the data set on purpose of retaining the systematic variance and discarding random noise. It is difficult to determine how many components should be retained, with the employment of SVD alone. Fortunately the two-way data set $\boldsymbol{X}_{I\times JK}$ or $\boldsymbol{X}_{IK\times J}$ is unfolded from the trilinear data array $\underline{\boldsymbol{X}}$, the systematic part of variation follows a trilinear model. So one can first truncate the unfolded data set $\boldsymbol{X}_{I\times JK}$ or $\boldsymbol{X}_{IK\times J}$ into a new two-way data $\boldsymbol{X}_c$ with the first c components retained ($\boldsymbol{X}_{I\times JK} = \boldsymbol{USV}', \boldsymbol{X}_c = \boldsymbol{U}_c\boldsymbol{S}_c\boldsymbol{V}'_c$), then refold the new two-way data $\boldsymbol{X}_c$ into a new three-way data array $\underline{\boldsymbol{X}}_c$ and decompose it by PARAFAC algorithm, finally examine the residual sum of squares SSR_c. When c, the number of the retained components, coincides with the number of the actual underlying factors, $\underline{\boldsymbol{X}}_c$ ($c = F$) exactly follows a trilinear model and can be fitted perfectly by PARAFAC model. The residual sum of squares (SSR_c) is expected to reach its minimum. If some actual underlying factors are excluded ignorantly in the truncating step, part of the systematic variations will be discarded along with random noise. The discarded part of systematic variations is actually related to all the underlying factors. Hence, the trilinear feature does not hold any more for the data array $\underline{\boldsymbol{X}}_c$ ($c < F$) assembled by the trimmed two-way data set $\boldsymbol{X}_c$. The SSR_c for the new data array $\underline{\boldsymbol{X}}_c$ ($c < F$) would be much larger than that when all the underlying factors being exactly extracted. The inclusion of superfluous factors in truncating procedure ($c > F$) would also cause the increase of the residual sum of squares of PARAFAC, due to the fact that random noise does not obey the trilinear model and can not be accounted for by PARAFAC model. Moreover, for the same reason, when the retained components exceed the actual underlying factors, SSR_c would be larger than the variance introduced by the inclusion of additional factors representing noise in the truncated data set. The aforementioned reasoning is actually based on the fact that random noise does not follow the trilinear model and can not be accounted for by PARAFAC.

The above speculations motivate us to design the following ADD-ONE-UP truncating and fitting approach for factor determination in trilinear decomposition (from now on, it will be simply refereed as ADD-ONE-UP method). The proposed method is mainly composed of the following steps:

1. For a trilinear data array $\underline{\boldsymbol{X}}$, unfold it into a two-way data set $\boldsymbol{X}_{I\times JK}$.
2. Decompose $\boldsymbol{X}_{I\times JK}$ by SVD, $\boldsymbol{X}_{I\times JK} = \boldsymbol{USV}'$.
3. Define $\boldsymbol{X}_c = \boldsymbol{U}_c\boldsymbol{S}_c\boldsymbol{V}'_c$, where $\boldsymbol{U}_c$ and $\boldsymbol{V}_c$ consist of the first c columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively, $\boldsymbol{S}_c$ is a diagonal matrix with diagonal elements equal to the first c diagonal elements of $\boldsymbol{S}$.
4. Fold $\boldsymbol{X}_c$ into a three-way data array $\underline{\boldsymbol{X}}_c$ and decompose it by PARAFAC with $N = c$. The residual sum of squares is represented by SSR_c.
5. Repeat steps 3 and 4, for $c = 1, 2, 3, \ldots,$ until SSR_{c_1}reaches its minimum or satisfies the following inequations: $\mathrm{SSR}_{c_1} < s^2_{c_1}$ and $\mathrm{SSR}_{c_1+1} > s^2_{c_1+1}$ and $\mathrm{SSR}_{c_1+2} > s^2_{c_1+2}$ (s_i is the ith diagonal element of matrix $\boldsymbol{S}$, $s^2_{c_1}$ is the variance introduced by the inclusion of c_1th component in the truncating step).
6. Unfold $\underline{\boldsymbol{X}}$ in another mode to get $\boldsymbol{X}_{IK\times J}$, and then perform the same operations from step 2 to 5 to get c_2, which satisfies similar relationships as for c_1.
7. The number of factors needed in decomposing the trilinear data array $\underline{\boldsymbol{X}}$ should be the smaller one between c_1 and c_2, i.e. $F = \min(c_1, c_2)$.

In the above procedure, two criteria are used to draw the final conclusion. It is worth to point out that for data arrays with only medium random noise, the two criteria would be satisfied simultaneously. In practice, it was observed that the first criterion has a strong ability to cope with high random noise, but relatively weak ability to tolerate varying backgrounds. The presence of high varying backgrounds might cause SSR_c to decrease consistently. On the contrary, though the second criterion is inferior to the first criterion in the presence of high random noise, it performs much better when varying backgrounds exist. Therefore, if SSR_c has its minimum, which means varying backgrounds have little influence on the data structure, the first criterion is adopted to draw the final conclusion. If SSR_c shows a consistent decrease due to the influence of high varying backgrounds, one should draw the final conclusion based on the second criterion. In the second criterion, three inequations should be satisfied for the correct number of underlying factors. The reason to set these requirements lies in the fact that: for a great variety of data structures, there may exist cases that for certain $c < F$, SSR_c may also larger than s_c^2. However, it is practically unlikely that $\text{SSR}_c > s_c^2$ and $\text{SSR}_{c+1} > s_{c+1}^2$ hold for two successive c and $c+1$ which are $<F$. So if for certain c, $\text{SSR}_c < s_c^2$, $\text{SSR}_{c+1} > s_{c+1}^2$ and $\text{SSR}_{c+2} > s_{c+2}^2$ hold, it can be concluded with great confidence that the system studied should have c underlying factors.

For a three-way data array $\underline{\boldsymbol{X}}$ assembled by $\boldsymbol{X}_{..k} = \boldsymbol{A}_{I\times F}\text{diag}(\boldsymbol{c}_k)\boldsymbol{B}'_{J\times F}$ $(k = 1, 2, \dots, K)$ (suppose the columns of $\boldsymbol{A}_{I\times F}$ and $\boldsymbol{B}_{J\times F}$ are excitation and emission spectra of corresponding chemical components, respectively; $\boldsymbol{c}_k$ is a vector containing the concentrations of all the chemical components in kth sample), if it follows the PARAFAC model, the number of underlying factors would equal to the ranks of $\boldsymbol{A}$ and $\boldsymbol{B}$, i.e. $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{B}) = F$. The rank of $\boldsymbol{C}$ ($\boldsymbol{C} = [\boldsymbol{c}'_1, \boldsymbol{c}'_1, \dots, \boldsymbol{c}'_K]$) satisfies the following inequation: $2 \le k\_\text{rank}(\boldsymbol{C}) \le F$. Therefore, in add-one-up procedure, only the two unfolded two-way data matrixes $\boldsymbol{X}_{I\times JK}$ and $\boldsymbol{X}_{IK\times J}$ are considered, while $\boldsymbol{X}_{K\times IJ}$ is omitted.

6. Experimental

In this paper, simulated data arrays and three real chemical data arrays including two emission/excitation fluorescent data arrays and one HPLC–DAD data array have been used to demonstrate the performance of the proposed method. Comparisons between the new approach and three methods: factor indicator function (IND) [12], eigenvalue ratio (ER) [14] and VPVRS [15] had been made. All the selected methods are automatic ones; i.e. they require no human involvement. Methods requiring threshold value or confidence degree setting such as F-test [13,18] are excluded in this paper. For the convenience of readers, a brief description about how to use IND, ER and VPVRS to determine the number of components in a two-way data set is supplied in APPENDIX.

6.1. Randomly simulated data arrays

A total of 20 data arrays with size of 18 × 18 × 9 were simulated. For each data array, underlying loading matrixes $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ of order 18×4, 18×4 and 9×4, respectively, were randomly constructed. Their elements were drawn from uniform (0, 1) distribution. Loading matrixes $\boldsymbol{A}$ and $\boldsymbol{B}$ were normalized column-wise to unit length. Homoscedastic and proportional heteroscedastic random noises were considered. For data set $\boldsymbol{X}_{..k}$ of sample k, homoscedastic and heteroscedastic noises added were simulated according to the following schemes, respectively:

$$\text{homo\_noise} = a_{\text{homo}} \times \text{RANDN} \times \max(\boldsymbol{X}_{..k}),$$
$$\text{heter\_noise} = a_{\text{heter}} \times \text{RANDN} \circ \boldsymbol{X}_{..k}$$

where a_{homo} and a_{heter} are two parameters controlling the homoscedastic and heteroscedastic noise level, respectively. RANDN is a matrix with appropriate size and random entries chosen from a normal distribution with zero mean and unit variance; $\max(\boldsymbol{X}_{..k})$ is the maximal entry of matrix $\boldsymbol{X}_{..k}$. The symbol '∘' represents Hadamard product (if $\boldsymbol{Z} = \boldsymbol{X} \circ \boldsymbol{Y}$, then $z_{ij} = x_{ij}y_{ij}$).

6.2. Simulated HPLC–DAD type data arrays

With a view to investigate the influence of varying backgrounds and collinearity on the performance of these methods, an HPLC–DAD type data array (24 × 41 × 9) of four components has been simulated. The varying backgrounds used were created by measuring the HPLC–DAD response matrixes of nine samples of o-dichlorobenzene, p-chlorotoluene and o-chlorotoluene in different concentration ratios, subtracting the contributions of the three chemical components and then retaining the residual matrixes as backgrounds. Before adding to the simulated HPLC–DAD data array, each background matrix should be divided by its maximal element, and then multiply a parameter $a_{\text{background}}$ and the maximal element of the corresponding simulated HPLC–DAD data set. As for collinearity, it was controlled by replacing the spectrum of the fourth component with spectral profile produced by the equation: $\mathrm{sp} = \mathrm{norm}(\mathrm{sp}_4 + a_{\mathrm{collineary}} \times \mathrm{sp}_3)$. Here, norm is a function normalizing vectors to unit length; sp_3 and sp_4 symbolize the spectra of components 3 and 4, respectively; $a_{\mathrm{collinearity}}$ is a parameter regulating the degree of collinearity.

6.3. Excitation–emission fluorescent data array I

6.3.1. Reagents and stock solutions

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

6.3.2. Apparatus

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

6.3.3. Data array

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

6.4. Excitation–emission fluorescent data array II

Fluorescent data array II has been used in elsewhere [24]. It is a collection of fluorescent excitation–emission response matrixes of 11 mixtures of tyrosine, tryptophan and phenylalanine with different concentration ratios. All the response matrixes were recorded by a HITACHI-850 fluorescence spectrophotometer with excitation and emission wavelength ranging from 205 to 290 nm and from 270 to 385 nm at an interval of 5 nm, respectively.

6.5. HPLC–DAD data array

This data array with size of 24 × 51 × 9 is assembled by data sets of nine mixtures of three compounds, i.e. *o*-dichlorobenzene, *p*-chlorotoluene and *o*-chlorotoluene, in different concentration ratios. The corresponding nine data sets were recorded by HPLC with diode array detector at same conditions. In order to illustrate the ability of ADD-ONE-UP dealing with heavy collinearity, only part of the data array, composed of the 35 slices from the seventh to the 51 horizontal slices of the original data arrays, was used. For detail information of this experiments readers please refer to [25].

6.6. 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.

7. Results and discussions

7.1. General remarks on the factor-determining procedures for trilinear data arrays

For two-way data set, the row and column ranks are always identical for mathematical reasons. As far as the three-way data arrays are concerned, it does not hold any more. The pseudo rank in a specific mode shows how many linear independent chemical variations are present in the given model. For chemical data array, the underlying factors often involve chemical components and sometimes background or some other interference(s). So only two modes of the data array are concerned in factor-determining procedure. The sampling mode is often omitted. In this paper, the smaller one among the two pseudo ranks of the two

Table 1
The general behavior of ADD-ONE-UP method for a randomly simulated four-component data array with homoscedastic noise level $a_{\mathrm{homo}} = 0.01$

i	s_i^2	SSR_i
1	23.48	0.3703
2	0.4526	0.2598
3	0.2162	8.337×10^{-2}
4	0.1081	2.392×10^{-3}
5	1.391×10^{-3}	3.630×10^{-3}
6	1.293×10^{-3}	4.796×10^{-3}

Table 2
The performances of four factor-determining methods for a randomly simulated four-component data array[a] with homoscedastic noise level $a_{\mathrm{homo}} = 0.01$ ranging from 0.01 to 0.07

a_{homo}	ADD-ONE-UP	IND	ER	VPVRS
0.01	4	4	4	4
0.03	4	4	4	4
0.05	4	4	4	4
0.07	4	2	4	4

[a] A total of 20 randomly simulated data arrays were tested with quite similar performance. The results of one randomly selected array are shown here.

unfolded matrixes: $\boldsymbol{X}_{I\times JK}$ and $\boldsymbol{X}_{IK\times J}$ is taken as the number of underlying factors.

7.2. Testing the general behavior of ADD-ONE-UP method

Table 1 lists the common performance of ADD-ONE-UP for data arrays with random noise. As expected in the design stage, SSR_c attains its minimum when the number of retained factors equal to 4, which is exactly the number of actual underlying factors. Significant increases do emerge with the inclusion of superfluous factors representing random noise. It is also confirmed that the variances (s_i^2) brought in by the addition of noise factors in truncating procedure are smaller than the corresponding residuals sum of square of PARAFAC with N equaling to the number of factors retained. The two criteria employed in ADD-ONE-UP method have been simultaneously satisfied. All these phenomena are in perfect agreement with our former conjectures. It demonstrates the feasibility of ADD-ONE-UP in factor-determining tasks. Further investigation of the performance of ADD-ONE-UP for a variety of situations will be pursued in the following sections.

7.3. Randomly simulated data arrays

The typical performances of four factor-determining methods for all the 20 randomly simulated data arrays are showed in Tables 2 and 3. With the presence of only homoscedastic noises, all the four methods can separate the systematic variations from the random ones, even when noise level a_{homo} is as high as 0.05. Generally speaking, IND is slightly inferior to the other three methods, since it can perceive the existence of only two components for a four-component system with a higher noise level $a_{\mathrm{homo}} = 0.07$, while ADD-ONE-UP, ER and VPVRS can still give the right results (Table 2). The capabilities of the tested methods to cope with heteroscedastic noise have a similar sequence (Table 3). Once again, IND is broken down first, followed by ER and ADD-ONE-UP. VPVRS has the strongest ability to deal with heteroscedasticity. The failure of IND originates in the violation of the assumption it established on that heteroscedastic noise does not follow a spherically symmetric distribution. The relatively stronger abilities of ER and especially VPVRS are due to the fact that they employ certain ratio forms of eigenvalues of the covariance data set rather than eigenvalues themselves. As for ADD-ONE-UP, it puts no extra assumption on the noise form, therefore, theoretically mild heteroscedastic noise has little influence on its performance as the experiments show. For $a_{\mathrm{heter}} = 0.2$, it seems that ADD-ONE-UP and ER failed to give the right answers. Actually, the heteroscedastic

Table 3
The performances of four factor-determining methods for a randomly simulated four-component data array[a] with $a_{\mathrm{homo}} = 0.01$ and a_{heter} equaling to 0.1, 0.15 and 0.2

a_{heter}	ADD-ONE-UP	IND	ER	VPVRS
0.1	4	4	4	4
0.15	4	3	3	4
0.2	3	2	3	4

[a] A total of 20 randomly simulated data arrays were tested with quite similar performance. The results of one randomly selected array are shown here.

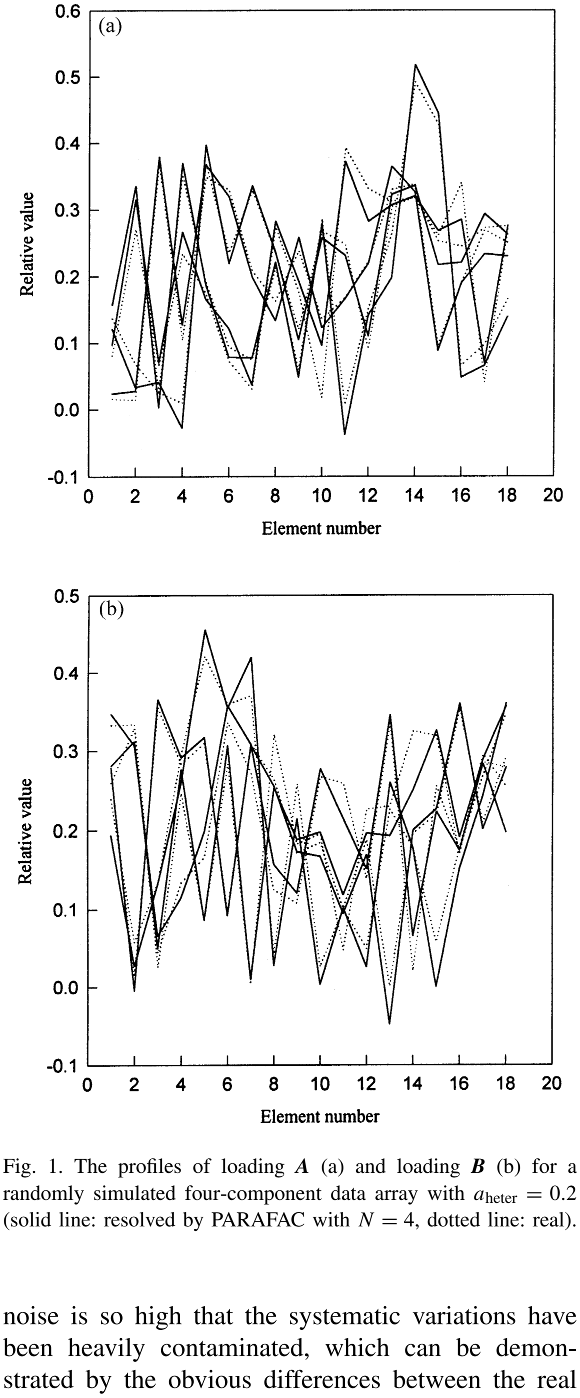

Fig. 1. The profiles of loading ***A*** (a) and loading ***B*** (b) for a randomly simulated four-component data array with $a_{\mathrm{heter}} = 0.2$ (solid line: resolved by PARAFAC with $N = 4$, dotted line: real).

noise is so high that the systematic variations have been heavily contaminated, which can be demonstrated by the obvious differences between the real loadings and those resolved by PARAFAC ($N = 4$) for the third system in Table 3 (Fig. 1). In analytical practice, noise level will not be so high. Otherwise, even when the actual underlying factors have been accurately estimated, the severely distorted results resolved would have little use. In this sense, the ability of ADD-ONE-UP withstanding heteroscedastic noise is enough for the purpose of trilinear data analysis.

Table 4
The performances of four factor-determining methods for a simulated four-component HPLC–DAD type data array with $a_{\mathrm{homo}} = 0.01$ and $a_{\mathrm{colinearity}}$ equaling to 0.9, 1.2 and 1.4

$a_{\mathrm{colinearity}}$	ADD-ONE-UP	IND	ER	VPVRS
0.9	4	4	3	4
1.2	4	4	3	3
1.4	4	4	3	3

7.4. Simulated HPLC–DAD data arrays

All the above comparisons are based on the data arrays with mild collinearity. For data arrays with heavy collinearity, the sequence may be different. It can be deduced theoretically that ER and VPVRS will obviously inferior to ADD-ONE-UP and IND, for their eigenvalue-ratio characters require the eigenvalues representing systematic variations to be comparable. Large differences among the systematic eigenvalues caused by high collinearity will bring disaster to them. The results listed in Table 4 verified such a conjecture. Without any additional assumption on the systematic eigenvalues, ADD-ONE-UP and IND can handle situations of high collinearity under condition of mild random noise.

Under comparatively ideal conditions, ADD-ONE-UP already shows its advantages over the other three methods. However, in practice, non-ideal situations such as varying backgrounds and other deviations are always encountered. A competent method should also suit for such cases other than the above comparatively ideal situations.

Table 5 shows the results of the four methods for a simulated four-component HPLC–DAD data array with varying backgrounds taken from real experiments. For all the four levels of the backgrounds, $a_{\mathrm{background}}$ varying from 0.01 to 0.07, IND failed to provide estimations of the underlying factors. All its 18 values investigated show a consistent decrease, no minimum occurs. The ER and VPVRS can work only at low level of backgrounds. Slightly higher level of

Table 5
The performances of four factor-determining methods for a simulated four component HPLC–DAD type data array with $a_{\text{background}}$ varying from 0.01 to 0.07

$a_{\text{background}}$	ADD-ONE-UP	IND	ER	VPVRS
0.01	4	–[a]	4	4
0.03	4	–	3	3
0.05	4	–	3	3
0.07	5	–	3	3

[a] The 18 values of IND surveyed show a consistent decrease and no minimum occurs, therefore, the number of underlying factors can not be determined.

backgrounds, such as $a_{\text{background}} = 0.03$, make ER and VPVRS give wrong results. It is interesting that intuitively, high backgrounds would introduce additional underlying factors. Nevertheless, instead of finding five or more factors, ER and VPVRS even neglect the existence of an original component and only find three underlying factors. To some extent, it may indicate that the differences among the values of some ratio forms of eigenvalues as adopted in ER and VPVRS, may not be appropriate indexes to discriminate the systematic part of data arrays from varying backgrounds.

As expected in design, ADD-ONE-UP has an outstanding advantage to cope with non-ideal experimental conditions. The increase of background level $a_{\text{background}}$ from 0.01 to 0.05 exerts no influence on the correctness of its estimations. Further increase of the background level ($a_{\text{background}} = 0.07$) will introduce additional factors other than the original four components. ADD-ONE-UP can detect one additional factor creeping into the systematic part of the data array. So it announces the presence of five underlying factors. The correctness of it estimation is strongly supported by the results of PARAFAC with $N = 4$ and 5. The obvious distortions of the loading profiles resolved by PRAFAC with $N = 4$ (Fig. 2) suggest there may be other factors being neglected and four-component model is not sufficient for the specific data array. With an extra factor being included, the results of PARAFAC (Fig. 3) show some improvements. Though the consistency between the real and resolved profiles are still not perfect. Taking the presence of varying backgrounds into consideration, which do not abide by a triliear model, the results seem reasonable and acceptable. Further increase of the factors used in PARAFAC sees a rapid deterioration of the results resolved (not shown), which means a five-component model is most appropriate for the specific data array, just as predicted by ADD-ONE-UP.

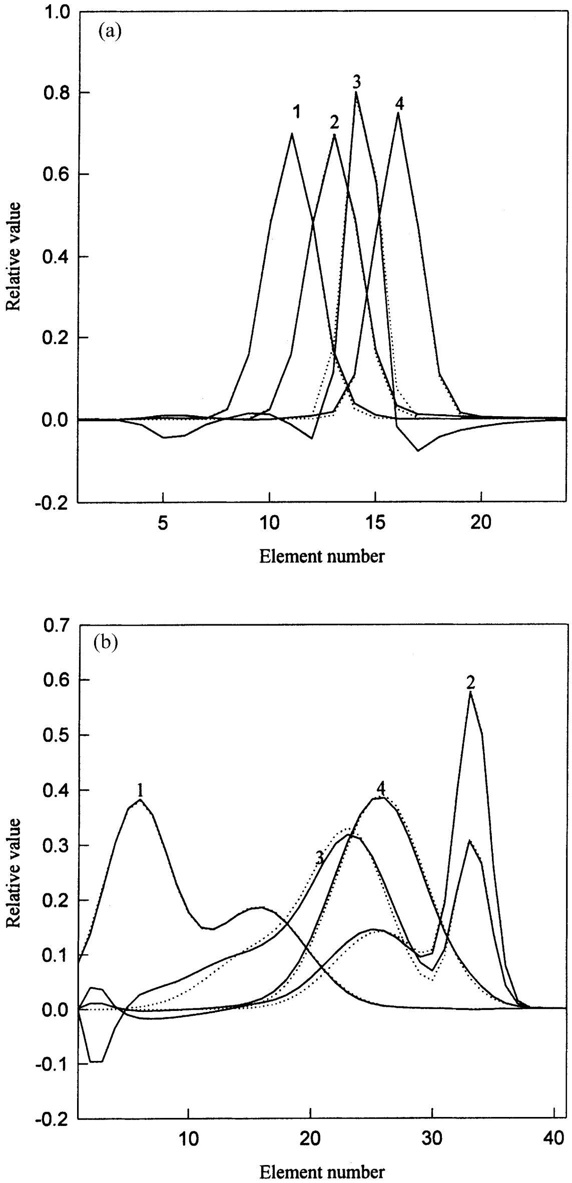

Fig. 2. The chromatographic (a) and spectral (b) profiles of the four components for the simulated HPLC data array with $a_{\text{background}} = 0.07$ (solid line: resolved by PARAFAC with $N = 4$, dotted line: real).

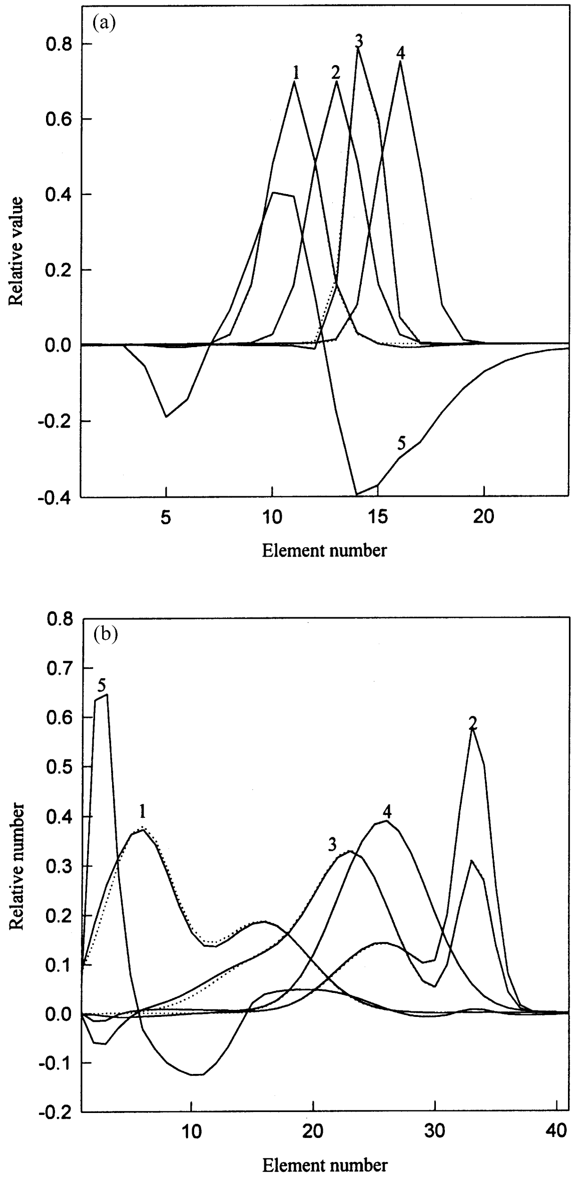

Fig. 3. The chromatographic (a) and spectral (b) profiles of the four components for the simulated HPLC data array with $a_{\mathrm{background}} = 0.07$ (solid line: resolved by PARAFAC with $N = 5$, dotted line: real). The extra profiles numbered with 5 are introduced by varying backgrounds.

7.5. Real data arrays

For the two excitation–emission fluorescent data arrays, data array I and data array II, the three methods ADD-ONE-UP, ER and VPVRS all offer correct results (Table 6). The reason for the success of ER and VPVRS for this two fluorescent data arrays may lies in the mild collinearity in emission and excitation spectra (the 2-norm condition numbers are in the range of (2.25, 7.5)) and the well-known good trilinear character of fluorescent data arrays. Background and other deviation levels are rather low. Though they are too low to affect ER and VPVRS, they are still high enough to deceive IND and make it exaggeratedly overestimated the underlying factors for the two fluorescent data arrays (9 for data array I, and 8 for data array II). Therefore, in practice, IND should be used with great care. Otherwise the results will be rather puzzling.

Table 6
The performances of four factor-determining methods for the two fluorescent data arrays

	ADD-ONE-UP	IND	ER	VPVRS
Data array I	3	9	3	3
Data array II	3	8	3	3

Things are much complicated for the HPLC–DAD data array. When the data array is unfolded in the retention time direction, ER reaches its maximum at the second value implying only two components detected (Table 7). VPVRS suggests that there are five

Table 7
Results of IND, ER and VPVRS for HPLC–DAD data array unfolded in chromatographic direction

i	IND ($\times 10^5$)	ER	VPVRS
1	219.9	17.14	16.64
2	42.34	33.04	33.99
3	13.17	17.42	2.6.12
4	8.614	2.692	1.967
5	5.562	7.155	143.2
6	5.224	1.045	0.1699
7	4.676	1.359	0.8204
8	4.099	1.779	2.017
9	3.729	1.629	1.668
10	3.485	1.605	2.178
11	3.353	1.385	0.6097
12	3.264	2.713	8.686
13	3.436	1.246	2.313
14	3.657	1.119	0.7965
15	3.916	1.175	5.356
16	4.233	1.034	1.025

Table 8
Results of IND, ER and VPVRS for HPLC–DAD data array unfolded in spectral direction

i	IND ($\times 10^5$)	ER	VPVRS
1	885.1	6.108	9.002
2	544.8	2.312	1.317
3	51.35	252.9	429.9
4	38.58	2.415	2.334
5	30.33	2.541	5.488
6	26.22	1.390	1.007
7	20.47	1.633	0.8418
8	12.49	4.032	3.682
9	8.656	5.661	7.385
10	8.406	2.711	12.72
11	9.143	1.155	0.3587
12	10.02	1.764	5.517
13	11.58	1.161	2.084
14	13.65	1.083	2.008
15	16.42	1.043	0.2877
16	20.16	1.178	3.188

underlying factors (Table 7). For data set unfolded in wavelength direction, both ER and VPVRS support the conclusion of three components (Table 8). The failure of ER to detect the existence of another component in spectral mode lies in its weak ability to cope with severe collinearity in the spectra with the 2-norm condition number equaling to 19.5. The very reason also lead to the puzzling results of VPVRS. It is well known that in HPLC–DAD experiments, the run-to-run variations in retention times and chromatogram shapes may introduce extra underlying factors in chromatographic mode. Therefore, unless some chemical components elute in and out simultaneously, the underlying factors in chromatographic mode are often one or more factors larger than that of the spectral mode. Contrarily, VPVRS find more factors in spectral mode than in chromatographic mode. Though according to the rule set in the former section that the smaller one of the two estimations for their corresponding modes should be taken as the number of the underlying factors for the data array, VPVRS seems to offer a correct estimation. Due to the obvious contradiction of its intermediate results with the common belief, it might be reasonable to own its specious success for this specific data array to lottery rather than its competence.

As a comparison, the results of ADD-ONE-UP are in fine accordance with the prior knowledge. It recommends three components for the spectral mode under the second criterion (Table 9). Though SSR$_i$ shows no minimum, which indicates the presence of unfavorable interference(s) or backgrounds, they are still in level of harmlessness to ADD-ONE-UP procedure. For the chromatographic mode (Table 10), the SSR attains its minimum when a three-component model being considered. So a three-component model is adopted for the HPLC–DAD data array. Further scrutinizing the values of s_i^2 and SSR$_i$, one can observe that the variance s_4^2 introduced by the inclusion of the fourth component is slightly larger than the corresponding residual SSR$_4$. It may suggest the existence of some model deviations, which may be brought in by the possible retention time shifts or run-to-run variation in chromatogram shapes. Since the difference is relatively small, one can imagine that though the model deviations would have certain degree of influence on the decomposition results of the data array, they might not be so great to significantly distort the loading profiles. Such conjecture has been at least partly supported by the results obtained by PARAFAC with $N = 3$ (Fig. 4). The slight differences between the real profiles and the resolved ones imply the presence of run-to-run variations. Nevertheless, the distortions are small; the results of a three-component mode as recommended by ADD-ONE-UP are acceptable.

Table 9
Results of ADD-ONE-UP for HPLC–DAD data array unfolded in chromatographic direction

I	s_i^2	SSR$_i$
1	7.872×10^5	1.201×10^5
2	4.593×10^4	5.136×10^4
3	1389	382.5
4	79.80	317.6
5	29.65	162.7

Table 10
Results of ADD-ONE-UP for HPLC–DAD data array unfolded in spectral direction

i	s_i^2	SSR$_i$
1	6.758×10^5	7670
2	1.106×10^5	5172
3	4.786×10^4	165.0
4	189.2	165.6
5	78.33	294.4
6	30.83	178.2

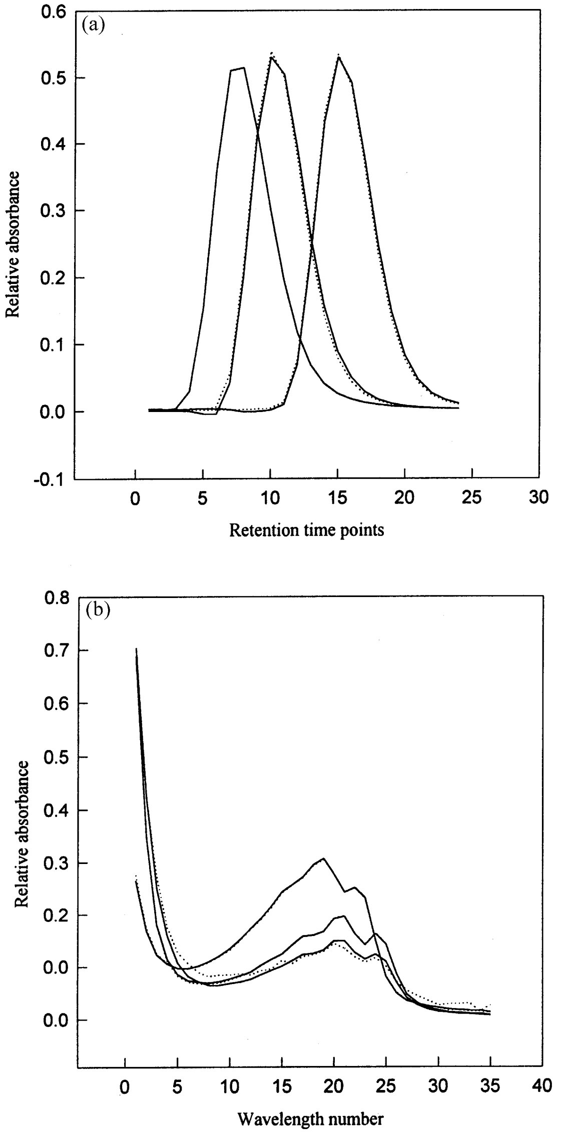

Fig. 4. The chromatographic (a) and spectral (b) profiles of the three components for the real HPLC data array (solid line: resolved by PARAFAC with $N = 3$, dotted line: real).

8. Conclusion

A new method called ADD-ONE-UP is contrived mainly for factor determination of trilinear data arrays. ADD-ONE-UP method involves the following three main steps: (1) truncating the unfolded two-way data matrix, (2) refolding the trimmed two-way data matrix into a new three-way data array, and (3) fitting the newly constructed three-way data array by PARAFAC model and then examining the residual sum of squares. The proposed method distinguishes itself from other conventional factor-determining methods originally designed for two-way data sets in the way that it takes the advantages of both the eigenvalues of factor analysis and the residuals of trilinear decomposition. The factor-determining procedure of ADD-ONE-UP is simple and can be automatically implemented. Unlike other methods such as *F*-test, no threshold value or confidence degree is required. ADD-ONE-UP has a strong ability to cope with heteroscedastic noise, heavy collinearity and varying backgrounds. Moreover, it can supply more information about the system studied rather than just an arbitrary decision on the number of underlying factors as many other methods do. Experimental results show the all-around performance of ADD-ONE-UP is superior to many other factor-determining methods, hence, it is recommended as a promising alternative for factor estimation in trilinear decomposition.

Acknowledgements

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

Appendix A

The procedure to determine the number of components in two-way data sets using IND, ER and VPVRS.

1. Using SVD algorithm to find the eigenvalues λ_i ($i = 1, 2, \ldots, n$; assume $n \leq m$) of the covariance matrix $\boldsymbol{X}'\boldsymbol{X}$, and do the following calculations:

$$\mathrm{IND}_i = \left(\sum_{j=i+k}^{n} \lambda_j\right)^{1/2} m^{-1/2}(n-i)^{-5/2}, \quad i = 1, 2, \ldots, n \qquad \text{(A.1)}$$

$$\mathrm{ER}_i = \frac{\lambda_i}{\lambda_{i+1}}, \quad i = 1, 2, \ldots, n-1 \qquad \text{(A.2)}$$

$$\mathrm{VPVRS}_i = \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_{i+2}}, \quad i = 1, 2, \dots, n-2 \tag{A.3}$$

Here, m and n are the number of rows and columns of data matrix $\boldsymbol{X}$, respectively.

2. The number of components can then be determined through examining the values of the above indexes. Suppose IND_i reaches its minimum with i equaling to c, then c is considered as the number of components in data set $\boldsymbol{X}$. As for ER and VPVRS, the number of components in mixture is defined by the value of i corresponding to their respective maximums.

References

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

[2] R.A. Harshman, UCLA Working Papers Phonetics 16 (1970) 1.

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

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

[5] S. Li, J.C. Hamilton, P.J. Gemperline, Anal. Chem. 64 (1992) 599.

[6] Y.S. Zeng, P.K. Hopke, J. Chemomet. 6 (1992) 65.

[7] S.E. Leurgans, R.T. Ross, R.B. Abel, SIAM, J. Matrix Anal. Appl. 14 (1993) 1064.

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

[9] H.A.L. Kiers, W.P. Krijnen, Psychometrica 56 (1991) 147.

[10] H.A.L. Kiers, J.M.F. Ten Berge, R. Bro, J. Chemomet. 13 (1999) 275.

[11] R. Bro, C.A. Andersson, H.A.L. Kiers, J. Chemomet. 13 (1999) 295.

[12] E.R. Malinowski, Factor Analysis in Chemistry, 2nd Edition, Wiley/Interscience, New York, 1991.

[13] E.R. Malinowski, J. Chemomet. 3 (1988) 49.

[14] X.W. He, H. Li, H.M. Shi, Anal. Chem. (China) 14 (1) (1986) 34.

[15] S. Alex, R. Savoie, Can. J. Spectros. 34 (2) (1989) 27.

[16] Z.P. Chen, Y.Z. Liang, J.H. Jiang, Y. Li, J.Y. Qian, R.Q. Yu, J. Chemomet. 13 (1999) 15.

[17] R.C. Henry, E.S. Park, C.H. Spiegelman, Chemom. Intell. Lab. Syst. 48 (1999) 91.

[18] K. Faber, B.R. Kowalski, Anal. Chim. Acta 337 (1997) 57.

[19] R.A. Harshman, M.E. Lundy, Research Methods for Multimode Data Analysis, Praeger, New York, 1984.

[20] D.J. Louwerse, A.K. Smilde, H.A.L. Kiers, J. Chemomet. 13 (1999) 491.

[21] S. Wold, Technometrics 20 (1978) 397.

[22] H.T. Eastment, W.J. Krzanowski, Technometrics 24 (1982) 73.

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

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

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