An Algorithm for Three-way Data Analysis That Alternatively Minimizes Coupled Vector (COV) Resolution Error and PARAFAC Error

Yu-Zhen CAO,*,† Hong CHEN,* Hai-Long WU,*** and Ru-Qin YU*****

**Key Laboratory of New Packaging Materials & Technology of China National Packaging Corporation, Zhuzhou Institute of Technology, Zhuzhou, 412008, China*

***Environmental Monitoring Central Station of Guangzhou City, NO. 95, Jixiang Road of Guangzhou City, Guangzhou, 510030, China*

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

A novel algorithm, alternatively minimizing coupled vector (COV) resolution error and PARAFAC error algorithm, is proposed in this paper. This algorithm can overcome the problem of slow convergence and is insensitive to the estimation of component number, such problems are unavoidable while using the traditional parallel factors analysis (PARAFAC) algorithm. In other words, this algorithm is capable of improving the computing speed and providing accurate resolutions provided that the number of factors used in the computation is no less than that of the actual underlying ones. The characteristic performances were demonstrated with a novel fluorescence data array.

(Received June 3, 2002; Accepted October 9, 2002)

Introduction

Most of chemometric methods are based on the theory of the matrix or the bilinear model. However, it is difficult to obtain physical resolutions due to the indeterminacy of the rotation. For this reason, the methods for three-way data analysis came into being. There are two main types of methods for three-way data analysis in chemometrics. Methods of the first type are based on generalized eigenanalysis, exemplified by generalized rank annihilation method (GRAM)[1] and its extension, direct trilinear decomposition (DTLD) method.[2] The second type of methods, typically represented by the parallel factors analysis (PARAFAC) method,[3-9] seek to fit a trilinear model to the three-way data using an iteration procedure. These methods not only provide unique physical resolutions but also implement direct calibration for the components of interest with unknown components co-existing, due to the so-called “Second-order calibration advantage”. Although successful applications of the PARAFAC algorithm to many chemical problems[10,11] have been made in recent years, there still remain several obstacles limiting its further applications in many areas. Among those, the annoyingly slow convergence problem, often requiring a much too long procedure to accomplish a three-way data analysis and the problem of requiring an accurate estimate of the component number, always leading to great deviations from the model, are both aspects worthy of detailed considerations.

For the sake of solving the slow convergence problem, several methods have been reported.[3,5,12,13] A more convenient technique used to accelerate the computing speed is the combination of compression and alternative least squares. Alsberg and Kvalheim,[14] Kiers and Harshman,[15] Bro and Anderson[16] and Kiers[17] have utilized this technique to improve the computing speed. They accelerated the computing speed by reducing the size of the original data array to a smaller one with a principal component analysis procedure.

Jiang *et al*.[18] have proposed an algorithm, coupled vector resolution algorithm, for chemometric calibration with three-way data array. However, there is a requirement of a *priori* chemical information while using this method. That is to say, it can not be applied to resolve a black system. On the basis of what Jiang *et al*. have done, the present authors provided an algorithm, alternatively minimizing the coupled vector (COV) resolution error and the PARAFAC error algorithm (COV-PAR), to resolve a black system, which has 11 samples consisting of salicylic acid, gentisic acid and *p*-aminobenzic acid. In the 11 samples, the concentrations of the three drugs are linearly independent. The resolved results illustrated that this algorithm can not only overcome the annoying slow convergence problem but also the problem of requiring an accurate estimate of the number of the actual underlying factors.

Theory

The model

In second calibration, each element x_{ijk} of $\underline{\mathbf{X}}$ will be expressed as a sum of a series of products:

$$x_{ijk} = \sum_{f=1}^{F} a_{if}b_{jf}c_{kf} + e_{ijk} \qquad (1)$$
$$i = 1, 2, \cdots; I; j = 1, 2, \cdots; J; k = 1, 2, \cdots; K.$$

where $\mathbf{a}_i$, $\mathbf{b}_j$ and $\mathbf{c}_k$ are respectively, the i-th row of $\mathbf{A}$, the j-th row of $\mathbf{B}$ and the k-th row of $\mathbf{C}$ and e_{ijk} is the error of the element. In matrix notation, the trilinear model can be written as follows:

† To whom correspondence should be addressed.

$$\mathbf{X}_{\cdot\cdot k} = \mathbf{A}\mathrm{diag}(\mathbf{c}_k)\mathbf{B}^t + \mathbf{E}_{\cdot\cdot k} \quad k = 1, 2, \cdots; K \tag{2}$$

where diag($\mathbf{c}_k$) symbolizes the diagonal matrix of $F \times F$ in which the corresponding diagonal elements are those of the vector $\mathbf{c}_k$ and $\mathbf{E}_{\cdot\cdot k}$ is the residual matrix and "*t*" denotes the transpose of matrix.

The algorithm

From Eq. (2), one can obtain the following equation:

$$\mathbf{X}_{\cdot\cdot k} = \sum_n c_{kn}\mathbf{a}_n\mathbf{b}_n^t + \mathbf{E}_{\cdot\cdot k} \quad k = 1, 2, \cdots; K \tag{3}$$

Assume $\mathbf{q}_m$ to be the m-th coupled vector of factors matrix *B*, which satisfies the following equation with m and n varying from 1 to N.

$$\mathbf{b}_n^t\mathbf{q}_m = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases} \tag{4}$$

Utilizing Eq. (4), one can transform the model Eq. (3) to the following equation:

$$\mathbf{X}_{\cdot\cdot k}\mathbf{q}_n = c_{kn}\mathbf{a}_n + \mathbf{e}_{k,n}^{a} \quad k = 1, 2, \cdots; K \tag{5}$$

$$L_1 = \| \mathbf{X}_{\cdot\cdot k}\mathbf{q}_n - c_{kn}\mathbf{a}_n \|_2^2 \quad k = 1, 2, \cdots; K \tag{6}$$

where $\mathbf{e}_{k,n}^{a}$ has been defined as the error vector and $\| \bullet \|_2$ defines the norm of a vector.

Similarly, assume $\mathbf{p}_n$ as the n-th coupled vector of factor matrix **A**, one can obtain:

$$L_2 = \| \mathbf{X}_{\cdot\cdot k}^{T}\mathbf{p}_n - c_{kn}\mathbf{b}_n \|^2 \quad k = 1, 2, \cdots; K \tag{7}$$

Here, L_1 and L_2 were defined as coupled vector resolution errors by Jiang.[18]

This algorithm computes factor matrices **A** and **B** with **C** fixed using the least squares fit of coupled vector resolution error L_1 and L_2 (COV procedure). Then it calculates matrix **C** with **A** and **B** fixed using least squares fit (PAR).

A simple introduction to the computation of factor matrices **A** and **B** by minimizing coupled vector resolution error with matrix **C** fixed is presented below. For further details, the readers are recommended to consult Ref. 18.

From Eqs. (6) and (7), one can obtain the following equations using the partial differential equations $\dfrac{\partial L_1}{\partial \mathbf{q}_n} = 0$, $\dfrac{\partial L_1}{\partial \mathbf{a}_n} = 0$, $\dfrac{\partial L_2}{\partial \mathbf{p}_n} = 0$ and $\dfrac{\partial L_2}{\partial \mathbf{b}_n} = 0$.

$$\mathbf{S}^{-1}\mathbf{U}^t \sum_{k,h=1}^{K} c_{kn}c_{hn}\mathbf{X}_{\cdot\cdot k}^{t}\mathbf{X}_{\cdot\cdot h}\mathbf{U}\mathbf{S}^{-1}\boldsymbol{\alpha} = \boldsymbol{\alpha} \tag{8}$$

$$\mathbf{q}_n = \frac{\sum_{k=1}^{K} c_{kn}^2}{\left\| \sum_{k=1}^{K} c_{kn}\mathbf{X}_{\cdot\cdot k}\mathbf{U}\mathbf{S}^{-1}\boldsymbol{\alpha} \right\|_2} \tag{9}$$

$$\mathbf{a}_n = \frac{\sum_{k=1}^{K} c_{kn}\mathbf{X}_{\cdot\cdot k}\mathbf{q}_n}{\sum_{k=1}^{K} c_{kn}^2} \tag{10}$$

$$\frac{\mathbf{S}^{-1}\mathbf{V}^T \sum_{k,h=1}^{K} c_{kn}c_{hn}\mathbf{X}_{\cdot\cdot k}\mathbf{X}_{\cdot\cdot h}^{T}\mathbf{V}\mathbf{S}^{-1}}{\sum_{k=1}^{K} c_{kn}^2}\boldsymbol{\beta} = \boldsymbol{\beta} \tag{11}$$

$$\mathbf{p}_n = \frac{\mathbf{V}\mathbf{S}^{-1}\boldsymbol{\beta}}{\mathbf{a}_n^t\mathbf{V}\mathbf{S}^{-1}\boldsymbol{\beta}} \tag{12}$$

$$\mathbf{b}_n = \frac{\sum_{k=1}^{K} c_{kn}\mathbf{X}_{\cdot\cdot k}^{t}\mathbf{p}_n}{\sum_{k=1}^{K} c_{kn}^2} \tag{13}$$

Here **U**, **S** and **V** can be obtained by truncating the first **N** columns of matrices, the singular value decomposition of matrices $\sum_{k=1}^{K}\mathbf{X}_{\cdot\cdot k}^{t}\mathbf{X}_{\cdot\cdot k}$ and $\sum_{k=1}^{K}\mathbf{X}_{\cdot\cdot k}\mathbf{X}_{\cdot\cdot k}^{t}$; $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the eigenvectors of Eqs. (8) and (9), whose corresponding eigenvalue is the maximal value among all eigenvalues. This algorithm calculates the whole **A** and **B** matrices with *n* changed from 1 to *N*, using Eqs. (8) – (13) with matrix **C** fixed. Then it computes **C** with **A** and **B** fixed, which is a least-squares-fit procedure and the same for the following PARAFAC procedure.

$$\mathbf{c}_k^t = [(\mathbf{A}^t\mathbf{A})^*(\mathbf{B}^t\mathbf{B})]^+ \,\mathrm{diag}(\mathbf{A}^t\mathbf{X}_{\cdot\cdot k}\mathbf{B})\mathbf{1} \tag{14}$$

where "*" defines the hadarmand product of matrices, **1** is a vector whose elements are all equal to one.

Based on the descriptions above, the algorithm could be designed as follows: Step 1, Randomly initialized **A** and **B**; Step 2, Compute **C** according to Eq. (14) with **A** and **B** fixed (PAR procedure); Step 3, Calculate **A** and **B** according to Eqs. (8) – (13) with **C** fixed (COV procedure); Step 4, Update **C**, **A** and **B** using Step 2 and Step 3 till a pre-assigned stop criterion has been fulfilled.

Experimental

Salicylic acid, gentisic acid and *p*-aminobenzoic acid, three drugs that are difficult to be simultaneously analyzed due to their closely overlapping fluorescence spectra, are taken as object analytes of simultaneouly determination in analytical samples. The steady state fluorescence of eleven samples of the aforementioned three drugs was measured in aqueous medium at pH 7 provided by the addition of phosphate buffer solution. The real concentrations of each species in eleven samples are listed in Table 1. The excitation wavelength was set from 265 to 350 nm at intervals of 5 nm and the emission wavelength varied from 305 and 550 nm with intervals of 5 nm. The scan rate was 1200 nm min^{-1}. The effect of Rayleigh's scattering on each sample was compensated by subtracting the measurement of a blank from the sample measurement. Thus a 50 × 18 × 11 data array was collected. This data array was analyzed using the COV-PAR algorithm and the ordinary PARAFAC algorithm, respectively, to recover profiles of each component.

Results and Discussion

Figure 1 shows the resolved excitation spectral and emission spectral profiles using the COV-PAR method and the PARAFAC algorithm as well with N = 3, which is the true dimensionality of this model. The concentrations are listed in Table 1. The required iteration number of the COV-PAR was 8, which is less than that (55) of the PARAFAC. These results further exemplified the point that the COV-PAR algorithm can greatly improve the computing speed without loss of accuracy.

To further explore the property of being insensitive to the estimation of component number, we decomposed the real data array with N = 4, which is greater than F, using the COV-PAR algorithm. The recovered excitation spectral, emission spectral and concentration profiles shown in Fig. 2 explain that the profiles of the actual underlying components can be perfectly resolved, provided that the number of component used in computation is no less than that of the actual underlying factors. However, it is well known that the traditional PARAFAC algorithm will provide a frustrated resolution under such circumstances.

Conclusion

A method that alternatively minimizes coupled vector resolution error and PARAFAC error for three-way data resolution was developed in this paper. This algorithm reduced the three-step computation in each iteration in the ordinary PARAFAC algorithm to two steps. It can not only overcome the slow convergence problem but also is insensitive to the estimation of component number, which is of great importance for second-calibration.

Table 1 Recovered concentration using the COV-PAR and the PARAFAC with N = 3 (ppm)

Sample	Real concentration			COV-PAR			PARAFAC		
	A	B	C	A	B	C	A	B	C
1	560	640	0	572	640	–3	572	639	–7
2	720	960	0	729	963	–3	729	962	–12
3	560	0	800	542	1	812	542	1	805
4	720	640	800	733	640	815	733	638	801
5	0	960	800	4	963	809	4	965	826
6	720	0	1200	6	3	1186	706	2	1170
7	0	640	1200	6	618	1186	8	619	1201
8	560	960	1200	567	963	1196	567	962	1188
9	0	0	800	1	2	809	0	5	825
10	0	960	0	9	965	–2	–10	967	19
11	720	0	0	710	2	–6	709	1	16
MAE				9.5	4	8.3	9.6	4.5	14

All values have been multiplied by 10^4.
MAE is defined as the mean absolute error between the expected profiles and the actual profiles.
A, Salicylic acid; B, gentisic acid; C, p-aminobenzoic acid.

Acknowledgements

This work was financially supported by Natural Science Foundation of Hunan (Grant No. 01JJY3006) and the Key Laboratory of New Packaging Materials and Technology of China National Packaging Corporation.

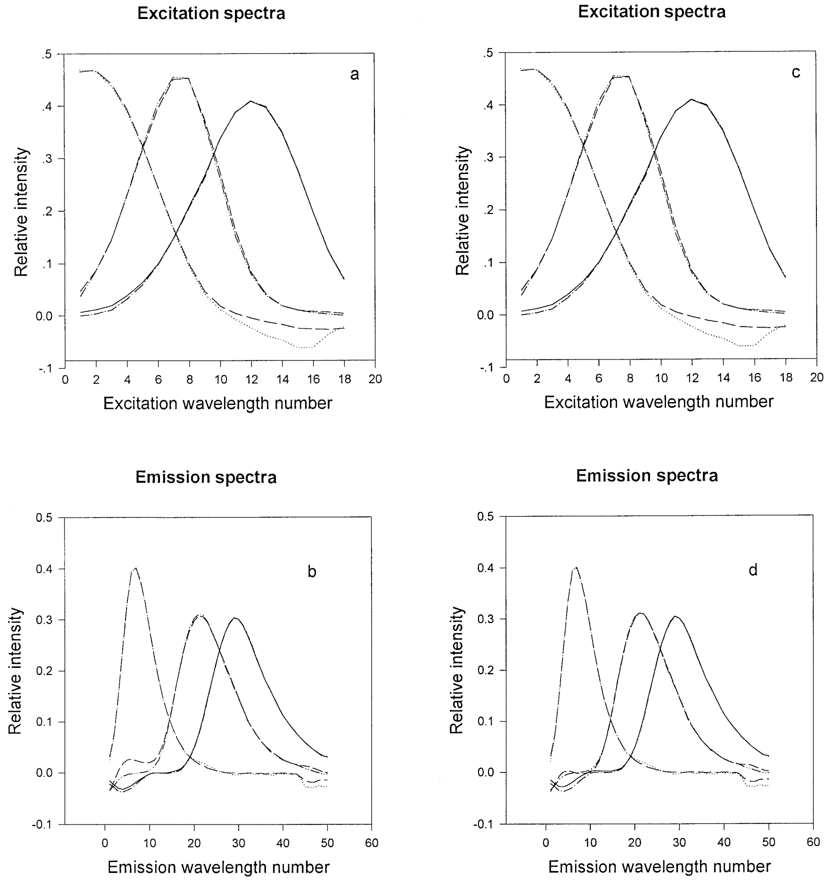

Fig. 1 Resolved profiles (dashed lines) of COV-PAR and PARAFAC and true profiles (solid lines) in real data array with N = 3. (a), (b) COV-PAR against true and (c), (d) PARAFAC against true.

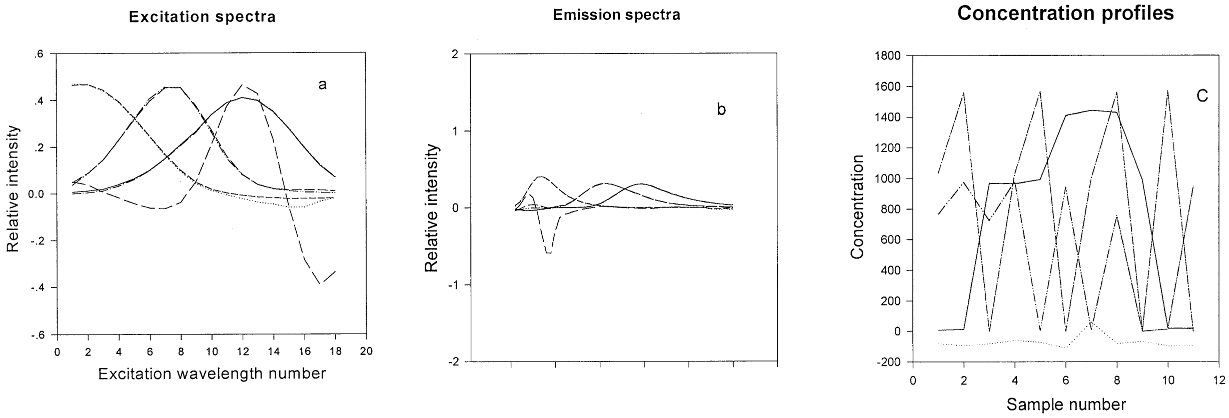

Fig. 2 Resolved profiles (solid lines) of COV-PAR and true profiles (dashed lines) in real data array with $N = 4$. (a) Emission profiles against True, (b) excitation profiles against true and (c) concentration profiles against True.

References

1. E. Sanchez and B. R. Kowalski, *Anal. Chem.*, **1986**, *58*, 499.
2. E. Sanchez and B. R. Kowalski, *J. Chemometrics*, **1990**, *4*, 29.
3. P. Paatero, *Chemometrics Intell. Lab. Syst.*, **1997**, *38*, 223.
4. J. D. Carroll and J. J. Chang, *Psychometrica*, **1970**, *35*, 283.
5. R. A. Harshman, *UCLA Working Papers Phonet.*, **1970**, *16*, 1.
6. H. A. L. Kiers and W. P. Krijnen, *Psychometrica*, **1991**, *56*, 147.
7. W. P. Krijnen, "*The Analysis of Three-way Arrays by Constrained PARAFAC Methods*", **1993**, DSWO Press, Leiden.
8. R. Bro, *Chemometrics Intell. Lab. Syst.*, **1997**, *38*, 149.
9. W. S. Rayens and B. C. Mitchell, *Chemometrics Intell. Lab. Syst.*, **1997**, *38*, 173.
10. P. Geladi, *Chem. Intell. Lab. Syst.*, **1980**, *7*, 11.
11. R. Bro, *Chem. Intell. Lab. Syst.*, **1997**, *38*, 149.
12. B. C. Mitchell and D. S. Burdic, *J. Chemometrics*, **1994**, *8*, 155.
13. R. Bro and Sijmen De Jong, *J. Chemometrics*, **1997**, *11*, 393.
14. B. K. Alsberg and O. M. Kvalheim, *Chem. Intell. Lab. Syst.*, **1994**, *24*, 43.
15. H. A. L. Kiers and R. A. Harshman, *Chem. Intell. Lab. Syst.*, **1997**, *36*, 31.
16. R. Bro and C. A. Anderson, *Chem. Intell. Lab. Syst.*, **1998**, *42*, 105.
17. H. A. L. Kiers, *J. Chemometrics*, **1998**, *12*, 155.
18. J.-H. Jiang, H.-L. Wu, Z.-P. Chen, and R.-Q. Yu, *Anal. Chem.*, **1999**, *72*, 4254.