Trilinear Component Analysis in Modern Analytical Chemistry

Hai-Long WU [1†], Ru-Qin YU [1] and Koichi OGUMA [2]

[1†] *College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China (E-mail: hlwu@mail.hunu.edu.cn)*
[2] *Faculty of Engineering, Chiba University, Chiba 263-8522, Japan*

The purpose of this paper is to present an overview of the recent developments in three-mode three-way trilinear component analysis with its applications in modern analytical chemistry. Emphasis is placed on the relatively new triadic algorithms that provide new ways to decompose the three-way data arrays from HPLC/DAD or two-dimensional excitation-emission fluorescence spectra, the rank estimation of three-way data as well as the uniqueness of the presentations. It has been shown that the combination of three-way data analysis methods with fluorescence spectroscopy or HPLC/DAD is a practical way to uniquely estimate the spectra and concentration of the analyte(s) of interest in the presence of the unknown interferents.

(Received on August 9, 2001; Accepted on October 4, 2001)

Introduction

Chemometrics is an important developing forward field in analytical sciences. [1,2] In the past three decades, modern analytical instrumentation capable of generating a vast amount of response data stimulates the emergence and development of chemometrics.[3] Second-order instruments and techniques have become commonplace in analytical chemistry. Modern analytical chemistry has benefited from the development of second- or higher-order instruments, that is, ones capable of providing bi- and multi-dimensional data arrays, which usually results in a substantial improvement of the analytical capability. This is the case for the so-called hyphenated techniques such as GC-MS, HPLC-DAD, HPLC-MS, and other techniques such as excitation-emission matrix spectrofluorimetric analysis (EX-EM). It is very vital to study the methods as well as their algorithms of three-mode three-way data analysis, to develop the new chemometric methods and algorithms such as tensor calibration or tensor standard addition method in order to make a direct fast qualitative and quantitative analysis, especially in the presence of unknown interferents. In this paper, the recent developments in three-mode three-way trilinear component analysis with its applications in modern analytical chemistry are reviewed. Emphasis is placed on the relatively new triadic algorithms that provide new ways to decompose the three-way data arrays from HPLC/DAD or excitation-emission matrix fluorescence spectra, the rank estimation of three-way data and the uniqueness of the presentations.

Three-mode three-way trilinear component model

It was known intimately that most of chemometric methods are based on the theory of two-dimensional data array, matrix, or the bilinear model. They use the matrix eigenanalysis for multivariate curve resolution, estimating number of principal components as well as multivariate calibration. However, it would be difficult to obtain physical resolutions since an intrinsic drawback, that is, the flexibility of rotation, exists in matrix decomposition. In the case of multivariate calibration, an essential prerequisite that the measuring systems must be very similar to background and matrix in both calibration and prediction steps should be satisfied. If unknown interferences exist in predicted samples, then it would produce an serious prediction error for the predicted samples.[4,5] Practically, this drawback has hampered the practical applications of classical chemometric methods in analytical chemistry. For this reason, the Kowalski's Group in Washington University, U. S. A., developed successively the generalized rank annihilation method (GRAM) [6] for an analytical systems including one reference and one unknown sample, and the direct trilinear decomposition (DTLD) for more than two sample system.[7] Unfortunately, the latter sometimes produces imaginary resolution and then lacks fidelity. The possible reason is that the dimensional number in sample mode is set to two when the pseudo samples are constructed. However, by using the same experimental tools, it is easy to formulate a three-mode three-way trilinear component model:

$$x_{ijk} = \sum_{n=1}^{N} a_{in} b_{jn} c_{kn} + e_{ijk} \qquad (1)$$

$(i = 1,2, \ldots, I; \quad j = 1,2,\ldots,J; \quad k = 1,2,\ldots,K)$

Here, N denotes the number of factors, which should be considered as the total number of detectable species, containing physically meaningful component(s) of interest, and background as well as uncalibrated interferent(s). x_{ijk} is the element of the three-mode three-way data array $\underline{\mathbf{X}}(I\times J\times K)$. a_{in}, b_{jn}, c_{kn} are the corresponding elements of three profile matrices **A**, **B** and **C** dimensioned by $I\times N$, $J\times N$ and $K\times N$, respectively. e_{ijk} is the element of a three-way residual array $\underline{\mathbf{E}}$ $(I\times J\times K)$. For

example, in an excitation-emission matrix obtained in spectrofluorimetry, **A**, **B** and **C** are the relative intensity matrices of excitation and emission spectra, and the relative concentration matrix ,respectively. For HPLC-DAD, **A**, **B** and **C** are the corresponding chromatogram and spectrum (relative values) matrices, and the relative concentration matrix, respectively.

In the sixties, there was a work on the three-way data analysis by L. R. Tucker[8], which is called the three-mode factor analysis, where the word "factor" can be considered as both the chemical components and the statistical principal components. Actually, the method developed by L.R. Tucker should be called three-mode principal component analysis, since an orthogonal constraints to three profile matrices has been attached. In the seventies, Carroll and Chang[9] as well as Harshman [10] suggested individually the trilinear model, which is generally called CANDECOMP/PARAFAC model. Because the chemical researchers published the work on trilinear model later than did the psychologists, the terminology of PARAFAC model, is usually used in the field of chemometrics.

In the recent years, applications of three-way data analysis to modern analytical sciences have been reported. [11, 12] Since obtaining a set of three-way measuring data array is difficult in such a science as psychology but easy using modern analytical instruments, a few psychologists started to cooperate with chemists to carry out the basic research on the three-way trilinear data analysis. The chemometricians made use of some achievements of other sciences including psychometrics and studied new applications for measuring systems based on the multidimensional response characteristics of modern analytical instrumentation. However, there exists a shortcoming that fundamental conception was not very clear, only the bilinear decompositon was used as the mathematical tool, and the advantage of multiway data decomposition was not fully utilized. Therefore, the authors [13-15] proposed a regularization of the mathematical notations in view of the characteristics of trilinear decomposition and developed a useful algorithm, alternating trilinear decomposition, ATLD, based on an alternating least-squares principle to overcome some shortcomings of the traditional PARAFAC algorithms, such as the sensitivity to the estimated component numbers and slow convergence. This ATLD algorithm utilizes an iterative procedure involving Moore-Penrose generalized inverse computations based on singular value decomposition. New algorithms of second-order calibration and standard addition methods have been developed and applied to the simultaneous determination of several organic components even in the presence of unknown interferents. In recent years, the authors [16-23] also proposed a series of other new triadic methods as well as algorithms based on different modified objective functions or alternating weighted least-squares principle. The coupled vectors resolution (COVER) [16] was developed for second-order calibration, which can yield direct and accurate solution. The alternating coupled vectors resolution (ACOVER) method [17] was proposed for trilinear analysis of a three-way data. Both resolve the three-way data in a component-wise way, while the conventional viewpoint was that in trilinear analysis all components should be resolved simultaneously. The alternating coupled matrices resolution (ACOMAR) method with some improved properties [18], the alternating slice-wise diagonalization (ASD) method [19], to find two matrices fitted to diagonal form in the least-squares sense, the self-weighted alternating trilinear decomposition (SWATLD) [20] and the modified parallel factors analysis algorithm with the penalty diagnolization error (PDE) [21] were developed in succession and improved the performances of trilinear decomposition. These methods have also been used to second-order calibration in the multicomponent analysis systems and the study on the multi-component reaction kinetic process of simultaneous degradations of chlorophyll a and b in the presence of other interferents .

Rank estimation for three-way data array

In establishing three-mode three-way component model, the traditional PARAFAC algorithm is sensitive to the estimates of principal component numbers .[14] The PARAFAC algorithm with an appropriate number of factors can often provide an accurate solution in spite of slow convergence. When a factor number that is larger than the actual number of factors is set, it usually gives unreasonable solutions, that is , it does not always converge to chemically meaningful solution, especially in the presence of two-factor degeneracies. Hence, determination of an appropriate factor number in three-way data array prior to trilinear decomposition is a very important step for the PARAFAC algorithm. The authors [14] proposes the following form:

$$\text{rank}\,\underline{X} = \max\{\text{rank}(X_P^I), \text{rank}(X_P^J), \text{rank}(X_P^K)\} \quad (2)$$

to estimate the chemical rank of the three-way data with satisfactory results. In practice, the number of factors will also be determined as follows:

$$\text{rank}\,\underline{X} = \max\{\text{rank}(\sum_{i=1}^{I} X_{i..}), \text{rank}(\sum_{j=1}^{J} X_{.j.}), \text{rank}(\sum_{k=1}^{K} X_{..k})\} \quad (3)$$

and

$$\text{rank}\,\underline{X} = \max\{\text{rank}(X_P^I (X_P^I)^T), \text{rank}(X_P^J (X_P^J)^T), \text{rank}(X_P^K (X_P^K)^T)\} \quad (4)$$

where rank$\underline{\mathbf{X}}$ is the numerical rank of the three-way data array $\underline{\mathbf{X}}$; rank (.) denotes the numerical rank estimates of a matrix based on a singular value decomposition procedure with a default tolerance. $\mathbf{X}_{i..}$, $\mathbf{X}_{j.,}$, $\mathbf{X}_{..k}$ denote the i-th horizontal slice of size $J\times K$, the j-th lateral slice of size $K\times I$, and the k-th frontal slice of size $I\times J$. The definition of X_P^I , X_P^J and X_P^K are given elsewhere [14].

Uniqueness of trilinear decomposition

If one considers the extension of a matrix to a three-way data similar to the extension of a bilinear model to a trilinear model, the uniqueness property can easily be understood. [10] In the case of a matrix, the expression formula have only two forms, a matrix and its transposition. However, for a three-way data array, since the mode number is three, it can lead to six expressions. Fortunately, owing to the symmetry and the cyclic property of the trilinear component model, the regularity of their changes can easily be found.

After finishing an iterative process of trilinear decomposition, if the columns corresponding to the components of interest in the finally obtained estimates **A**, **B** and **C** are appropriately post-processed, then the physical significance of **A**, **B** and **C** can easily be understood. The so-called uniqueness property is that the only alternative solutions involve changing the order of columns and/or stepping all the entries in a given column up or down by a constant multiplier. To express this algebraically, let

an alternative solution be represented as

$$X_{..k} = \dot{A}\, diag\ (\dot{c}_k)\dot{B}^T + E_{..k}, k = 1,..., K \qquad (5)$$

where $\mathbf{A}^*$, $\mathbf{B}^*$ and $\mathbf{C}^*$ are the alternative versions of the original solutions **A**, **B** and **C**, respectively. The uniqueness or intrinsic axis property insures that $\mathbf{A}^*$ can only differ from A by a rearrangement of columns (which can be represented by post-multiplicating **A** by a permutation matrix P) and /or a multiplicative rescaling of its columns (which can be represented by post-multiplication of **A** by a diagonal matrix $\mathbf{D}_a$). This is compensated for by an inverse rescaling of the columns of **B** and **C**. Similar relationships hold for the **B** and **C** matrices. Thus the alternative $\mathbf{A}^*$, $\mathbf{B}^*$ and $\mathbf{C}^*$ matrices must be related to the corresponding " original " versions of these matrices as follows :

$$A^* = A \cdot Da \cdot P, B^* = B \cdot D_b \cdot P, C^* = C \cdot D_c \cdot P. \qquad (6)$$

Here the **D** must conform to the requisite that

$$\mathbf{D}_a \ \underline{\mathbf{I}}\ \mathbf{D}^T_c \mathbf{D}^T_b = \underline{\mathbf{I}} \qquad (7)$$

since the effects of any internal rescalings (described by the D matrices) must "cancel out". Here $\underline{\mathbf{I}}$ is a superdiagonal core array of size $N \times N \times N$ with ones on the superdiagonal and zeros elsewhere.

These indeterminacies can be considered trivial because they do not affect the interpretation of the solutions. Geometrically, the effect of the permutation matrix P is simply to " renumber " the axes without moving them, and the effect of the rescaling D_a and so forth is simply to stretch or contract the axis in one space with compensatory contraction or stretch of the corresponding axis in the space of another mode. None of these changes affect the orientation of the axes in any of the three spaces. [10, 24] In other words, the trilinear decomposition of a three-way data array has an uniqueness property. When they are used to second-order calibration or resolution of a kinetic process, one should keep away from orthogonal constraints on the profile matrices, otherwise the resolution structure and the physical significance would be destroyed.

Future Prospects

The trilinear component model is a physically meaningful mathematical one that exists in practical measuring data obtained by means of modern analytical instrumentation. It is different from the three-mode principal component model, since an orthogonal constraint has been added to the latter but not to the former. The numerical rank of three-way trilinear data is estimable and the three formula described in this paper can basically meet this requirement. In spite of a variety of trilinear decomposition algorithms, the physically meaningful solutions may be obtained provided that (1) the selected component number is not less than the objective component number; (2) the rank-defending computations such as SVD is used in an iterative way based on an alternating least-squares principle. The proper post-process of the obtained resolutions is helpful to understand their physical and chemical meaning. In short, three-way trilinear component analysis is a useful exploratory data analysis method. It is also a promising chemometric method to the three-way data obtained in modern analytical chemistry. It can be used to substitute mathematical resolution for chemical separation. When combined with modern analytical instruments, it can be applied to simultaneous determination of several sought-for components even in the presence of unknown interferents.

As pointed out by Sanchez and Kowalski [7], chemistry is perhaps more suitable for trilinear components analysis than many other branches of science owing to the abundance of instruments that can automatically collect precise three-way data arrays in a short period of time. Since the related theories about three-way data arrays are still in their infancy, it is never static; there is still much to be learned about the fundamentals of three-way trilinear data arrays similar to the matrix-based chemometric methods.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of the P. R. of China （Grant No. 29735150 and 29975007） and the Hunan Province as well as a Grant from the Ministry of Education of China.

References

1. The Committee of the National Natural Science Foundation of The P. R. of China, *Analytical Chemistry* (A Report*),* Beijing : Science Press, **1993**.
2. E. K. Wang, ed., *Analytical Chemistry in The 21th Century*, Beijing : Science Press, **1999**, p1-22.
3. H. L. Wu, Y. Z. Liang and R. Q. Yu, *Fenxi Shiyanshi*, **1999**, *18*, 94.
4. R. Q. Yu, *Introduction to chemometrics*, Changsha, Hunan Education House, **1991**, 67.
5. H. L. Wu, K. Oguma and R. Q. Yu, *Anal. Sci.*,**1994**,*10*, 875.
6. E. Sanchez and B. R. Kowalski, *Anal. Chem.,* **1986**, *58*, 496.
7. E. Sanchez and B. R. Kowalski, *Journal of Chemometrics*, **1990**, *4*, 29.
8. L. R. Tucker, *Psychomitrika*, **1963**, *31*, 279.
9. J. D. Carroll and J. J. Chang, *Psychometrika*, **1970**, *35,* 283.
10. R. A. Harshman and M. F. Lundy, in: *Research methods for multiway data analysis*, H. G. Law, C. W. Snyder, J. A. Hattie and R. P. McDonld (Eds.), **1984**, New York, Praeger, p122-215.
11. P. Geladi, *Chemometrics Intell. Lab. Syst.*, **1980**, *7*, 11.
12. R. Bro. *Chemometrics Intell. Lab. Syst.*, **1997**, *38,* 149.
13. H. L. Wu, Y. Q. Feng, M. Shibukawa and K. Oguma, *Anal. Sci.*, **1997**, *13,* 99.
14. H. L. Wu, M. Shibukawa and K. Oguma, *J.Chemometrics*, **1998**, *12,* 1.
15. H. L. Wu, R. Q. Yu, M. Shibukawa, and K. Oguma, *Anal. Sci.*, **2000**, *16*, 217.
16. J. H. Jiang, H. L. Wu, Z. P. Chen, and R. Q. Yu, *Anal. Chem.y*, **1999**, *71*, 4254.
17. J. H. Jiang, H. L. Wu, Y. Li, and R. Q. Yu, *J. Chemometrics*, **1999**, *13,* 557.
18. Y. Li, J. H. Jiang, H. L. Wu, Z. P. Chen, and R. Q. Yu, *Chemometrics and Intell. Lab. Syst.* **2000**, *52,* 33.
19. J. H. Jiang, H. L. Wu, Y. Li and R. Q. Yu, *J. Chemometrics*, **2000**, *14*, 15.
20. Z. P. Chen, H. L. Wu, J. H. Jiang, Y. Li and R. Q. Yu, *Chemometrics Intell. Lab.Syst.*, **2000**, *52,* 75.
21. Y. Z. Cao, Z. P. Chen, C. Y. Mo, H. L. Wu, and R. Q. Yu, *Analyst*, **2000**, *125*, 2303.

22. Z. P. Chen, H. L. Wu, Y. Li and R. Q. Yu, *Anal. Chim. Acta*, **2000**, *423* , 187.
23. Y. X. Tan, J. H. Jiang, H. L. Wu, H. Cui and R. Q. Yu, *Anal. Chim. Acta*, **2000**, *412,* 195.
24. R. A. Harshman and M. E. Lundy, *Computational Statistics & Data Analysis*, **1994**, *18,* 39.
25. W. P. Krijnen, *The Analysis of Three-way Arrays by Constrained PARAFAC Methods*, DSWO, Press, **1993**.