JOURNAL OF CHEMOMETRICS, VOL. 12, 1–26 (1998)

AN ALTERNATING TRILINEAR DECOMPOSITION ALGORITHM WITH APPLICATION TO CALIBRATION OF HPLC–DAD FOR SIMULTANEOUS DETERMINATION OF OVERLAPPED CHLORINATED AROMATIC HYDROCARBONS

HAI-LONG WU,[1,2] MASAMI SHIBUKAWA[1] AND KOICHI OGUMA[1]*

[1] *Faculty of Engineering, Chiba University, Yayoi-cho, Inage-ku, Chiba 263, Japan*
[2] *Department of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, People's Republic of China*

SUMMARY

In this paper an alternating trilinear decomposition (ATLD) algorithm that is an improvement of the traditional PARAFAC algorithm without any constraints is described as an alternative algorithm for decomposition of three-way data arrays. It is based on an alternating least squares principle and an improved iterative procedure that, in a real trilinear sense, uses the Moore–Penrose generalized inverse with singular value decomposition. Its performance is compared with that of the traditional PARAFAC algorithm by a series of Monte Carlo simulations with different noise levels. It was found that the ATLD algorithm has a capability to converge faster than the traditional PARAFAC algorithm. The ATLD-based second-order calibration retains the second-order advantage that calibration in the presence of unknown interferents can be performed to provide satisfactory concentration estimates. Both algorithms have been used for simultaneous determination of overlapped chlorinated aromatic hydrocarbons measured by means of a high-performance liquid chromatograph with a diode array detector.
© 1997 John Wiley & Sons, Ltd.

J. Chemometrics, Vol. **12,** 1–26 (1998)

KEY WORDS alternating trilinear decomposition; trilinear model; PARAFAC; second-order calibration; Moore–Penrose generalized inverse; multiwavelength chromatogram; chlorinated aromatic hydrocarbons

INTRODUCTION

With the development of modern analytical instruments that generate a second-order tensor or two-way matrix of data for each sample, it has become ever more important to develop useful methods that may be applied to these data.[1–4] Second-order calibration is one such method. Through use of a three-way data array, calibration in the presence of unknown interferents can be performed, which is called the 'second-order advantage'.[4–6] Second-order calibration is usually performed by decomposition of a three-way data array and regression of the relative concentration contributions of each component of interest in sample space against its standard concentrations. There are two main approaches to decomposition of three-way data arrays. One approach is based on generalized eigenanalysis. Several methods[7–13] based on this approach have been proposed. The generalized rank annihilation method (GRAM)[7,8] and the direct trilinear decomposition (DTLD) method[9,10,13] have been used for second-order calibration. These methods work well when the signal-to-noise ratio is high.

The other main approach utilizes alternating least squares in an iterative procedure that exploits the

* Correspondence to: K. Oguma, Faculty of Engineering, Chiba University, Yayoi-cho, Inage-ku, Chiba 263, Japan.

CCC 0886–9383/98/010001–26 $17.50
© 1998 John Wiley & Sons, Ltd.
Received 29 July 1996
Accepted 7 May 1997

conditional linearity of the trilinear model. Its iterative nature means that starting values are required, but it is guaranteed to improve the least squares fit of the model to the data at each iteration. This approach is one commonly used by psychometricians working in three-mode factor analysis.[14–17] Its prototype is the PARAFAC algorithm developed and popularized by Harshman.[14] It has also been utilized to solve chemical problems.[2, 3, 18]

A recurring problem is that DTLD tends to yield imaginary eigenvalues when significant deviations from the model occur[9] and the PARAFAC algorithm does not always converge to chemically meaningful solutions.[19, 20] Methods to overcome these weaknesses by acquiring the strong points of each may be extremely important in improving the quality of the second-order calibration. Medium-rank second-order calibration with restricted Tucker models[6] is a good attempt and has been applied to multicomponent determinations of chlorinated hydrocarbons.[21]

This paper describes an alternating trilinear decomposition (ATLD) algorithm which is mathematically equivalent to an improvement of the traditional PARAFAC algorithm without any constraints. It is based on an alternating least squares principle and replaces the iterative procedure used in the traditional PARAFAC algorithm by an improved procedure. This improved procedure contains Moore–Penrose pseudoinverse computations based on singular value decomposition (SVD) which should be theoretically more robust to similarities in spectra and time profiles. The performance of the ATLD algorithm is compared with that of the traditional PARAFAC algorithm on two sets of simulated data by a series of Monte Carlo simulations at different noise levels. Both algorithms have been used for simultaneous determination of overlapped chlorinated aromatic hydrocarbons in a chromatogram obtained by a diode array detector.

THEORY

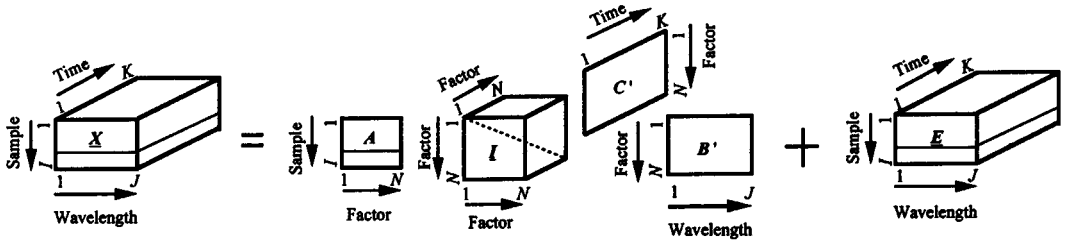
Trilinear model for second-order calibration

Second-order data are usually produced from hyphenated instruments such as a high-performance liquid chromatograph with a diode array detector (HPLC–DAD) or an excitation/emission matrix spectrofluorometer. Suppose that response values for an HPLC–DAD system are available from I samples, consisting of calibration samples and unknown mixture samples with uncalibrated interferents, on J variables (wavelengths) measured at K occasions (points of elution time). These second-order data can be collected in an $I\times J\times K$ three-way data array $\underline{\mathbf{X}}$ which can be visualized as a box of response values with frontal slices $\mathbf{X}_{\cdot\cdot k}$ (k=1, . . . , K) containing the $I\times J$ data matrices for each of the K occasions (points of elution time), with horizontal slices $\mathbf{X}_{i\cdot\cdot}$ (i=1, . . . , I) containing the $J\times K$ data matrices for each of the I samples and with lateral slices $\mathbf{X}_{\cdot j\cdot}$ (j=1, . . . , J) containing the $K\times I$ data matrices for each of the J variables (wavelengths). A trilinear model for such a three-way array $\underline{\mathbf{X}}$, as depicted in Figure 1, has the form

$$x_{ijk}=\sum_{n=1}^{N} a_{in}b_{jn}c_{kn}+e_{ijk}, \quad i=1,\ldots,I, \quad j=1,\ldots,J, \quad k=1,\ldots,K \tag{1}$$

where N denotes the number of factors, which should be considered as the total number of detectable species, containing component(s) of interest and background as well as uncalibrated interferent(s); x_{ijk} is the element (i, j, k) of $\underline{\mathbf{X}}$; a_{in} is the element (i, n) of an $I\times N$ matrix $\mathbf{A}$ with relative concentrations of the samples on the N factors; b_{jn} is the element (j, n) of a $J\times N$ matrix $\mathbf{B}$ with relative sensitivity coefficients corresponding to the wavelengths on N factors; c_{kn} is the element (k, n) of a $K\times N$ matrix $\mathbf{C}$ with elution profiles on N factors; and e_{ijk} is the i,j,kth residual element of an $I\times J\times K$ three-way residual array $\underline{\mathbf{E}}$. In subsequent discussions the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ will be called the relative

© 1998 John Wiley & Sons, Ltd.



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

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

Trilinear model for second-order calibration

Figure 1. Graphical representation of trilinear model of three-way data array $\underline{\mathbf{X}}$: **A**, relative concentration matrix of size $I \times N$; **B**, relative spectrum matrix of size $J \times N$; **C**, relative chromatogram matrix of size $K \times N$; $\underline{\mathbf{I}}$, superdiagonal core array of size $N \times N \times N$ with ones on the superdiagonal and zeros elsewhere; $\underline{\mathbf{E}}$, three-way residual data array of size $I \times J \times K$. Note that I is the number of samples including standards and unknowns, J is the number of variables (wavelengths), K is the number of occasions (points of elution time) and N is the estimated number of factors

concentration matrix, the relative spectrum matrix and the relative chromatogram matrix respectively.

Let $\mathbf{a}_i$, $\mathbf{b}_j$ and $\mathbf{c}_k$ be the ith row of **A**, the jth row of **B** and the kth row of **C** respectively. In matrix notation the trilinear model can be written as

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

$$\underset{(K\times I)}{\mathbf{X}}{}^{\cdot\,j\,\cdot} = \mathbf{C}\ \mathrm{diag}(\mathbf{b}_j)\mathbf{A}^{\mathrm{T}} + \underset{(K\times I)}{\mathbf{E}}{}^{\cdot\,j\,\cdot}, \quad j = 1, 2, \ldots, J \tag{3}$$

$$\underset{(I\times J)}{\mathbf{X}}{}^{\cdot\,\cdot\,k} = \mathbf{A}\ \mathrm{diag}(\mathbf{c}_k)\mathbf{B}^{\mathrm{T}} + \underset{(I\times J)}{\mathbf{E}}{}^{\cdot\,\cdot\,k}, \quad k = 1, 2, \ldots, K \tag{4}$$

where diag($\mathbf{a}_i$), diag($\mathbf{b}_j$) and diag($\mathbf{c}_k$) denote the diagonal matrices of order $N \times N$ in which the corresponding diagonal elements are elements of the vectors $\mathbf{a}_i$, $\mathbf{b}_j$ and $\mathbf{c}_k$ respectively. The superscript (T) denotes the transpose of a matrix.

Principle for alternating trilinear decomposition

As mentioned above, second-order calibration requires decomposition of a three-way data array and regression of relative concentration contributions of the component(s) of interest in sample space against the corresponding standard concentrations. For decomposition of three-way data arrays, several procedures have been suggested.[7–18]

In the traditional PARAFAC algorithm[17] a typical iterative procedure used for updating **A**, **B** and **C** is given by

© 1998 John Wiley & Sons, Ltd.

$$\mathbf{A}=\left(\sum_{k=1}^{K}\mathbf{X}_{..k}\mathbf{B}\mathbf{D}_k\right)\left(\sum_{k=1}^{K}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{D}_k\right)^{-1} \tag{5}$$

$$\mathbf{B}=\left(\sum_{k=1}^{K}\mathbf{X}_{..k}^{\mathrm{T}}\mathbf{A}\mathbf{D}_k\right)\left(\sum_{k=1}^{K}\mathbf{D}_k\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{D}_k\right)^{-1} \tag{6}$$

$$\mathbf{c}_k^{\mathrm{T}}=(\mathbf{A}^{\mathrm{T}}\mathbf{A}*\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}(\mathrm{diag}\mathbf{A}^{\mathrm{T}}\mathbf{X}_{..k}\mathbf{B})\mathbf{1}, \quad k=1,\ldots,K \tag{7}$$

where **1** denotes the N-vector with unit elements and $*$ denotes the Hadamard product. $\mathbf{D}_k$ is the same as $\mathrm{diag}(\mathbf{c}_k)$ in (4).

Obviously, updating **A**, **B** and **C** iteratively according to (5)–(7) does not always work well in the rank-deficient least squares problem, especially in the case of $N>N^*$, owing to experimental errors (N^* is the number of true compounds in the samples). This leads to the conclusion that (i) the traditional PARAFAC algorithm does not always converge to chemically meaningful solutions in the presence of two-factor degeneracy[19, 20] and (ii) the convergence is rather slow.

In this study, regularization for an iterative trilinear decomposition procedure was done. It improves the quality of the trilinear decomposition solution by using Moore–Penrose generalized inverses based on SVD. The loss function to be minimized is the sum of squares of the elements of the residual matrices, which may be written as

$$\sigma=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{ijk}-\sum_{n=1}^{N}a_{in}b_{jn}c_{kn}\right)^2 \tag{8}$$

In matrix notation, (8) may also be represented as

$$\sigma_1(\mathbf{A})=\sum_{i=1}^{I}\|\mathbf{X}_{i..}-\mathbf{B}\mathrm{diag}(\mathbf{a}_i)\mathbf{C}^{\mathrm{T}}\|_{\mathrm{F}}^2 \tag{9}$$

$$\sigma_2(\mathbf{B})=\sum_{j=1}^{J}\|\mathbf{X}_{.j.}-\mathbf{C}\mathrm{diag}(\mathbf{b}_j)\mathbf{A}^{\mathrm{T}}\|_{\mathrm{F}}^2 \tag{10}$$

$$\sigma_3(\mathbf{C})=\sum_{k=1}^{K}\|\mathbf{X}_{..k}-\mathbf{A}\mathrm{diag}(\mathbf{c}_k)\mathbf{B}^{\mathrm{T}}\|_{\mathrm{F}}^2 \tag{11}$$

where $\|.\|_{\mathrm{F}}$ denotes the Frobenius matrix norm.[22] Equations (9)–(11) can be considered equivalent to each other owing to the symmetry property of the trilinear model.

According to the above-mentioned loss functions, an alternating trilinear decomposition algorithm can be described as follows. It minimizes alternately one of the above-mentioned loss functions over **A** for fixed **B** and **C**, over **B** for fixed **A** and **C** and over **C** for fixed **A** and **B**. The updates for **A**, **B** and **C** from (9)–(11), based on a least squares principle,[23, 24] are

$$\mathbf{a}_i^{\mathrm{T}}=\mathrm{diag}(\mathbf{B}^{+}\mathbf{X}_{i..}(\mathbf{C}^{\mathrm{T}})^{+}), \quad i=1,\ldots,I \tag{12}$$

$$\mathbf{b}_j^{\mathrm{T}}=\mathrm{diag}(\mathbf{C}^{+}\mathbf{X}_{.j.}(\mathbf{A}^{\mathrm{T}})^{+}), \quad j=1,\ldots,J \tag{13}$$

© 1998 John Wiley & Sons, Ltd.

$$\mathbf{c}_k^{\mathrm{T}}=\mathrm{diag}(\mathbf{A}^{+}\mathbf{X}_{\cdot\cdot k}(\mathbf{B}^{\mathrm{T}})^{+}), \quad k=1,\ldots,K \tag{14}$$

where diag(.) denotes a column N-vector whose elements are diagonal elements of a square matrix. The superscript + denotes the Moore–Penrose generalized inverse, e.g. $\mathbf{A}^{+}=(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$, which may be automatically computed in the MATLAB environment by using a function PINV.[22, 25] For example, the PINV(**A**) computation is based on SVD(**A**) and any singular values less than a tolerance are treated as zero. The default tolerance is

$$\mathrm{tol}=\max(\mathrm{size}(\mathbf{A}))*\mathrm{norm}(\mathbf{A})*\mathrm{EPS} \tag{15}$$

where norm(**A**) is the largest singular value of **A**. EPS denotes the floating point relative accuracy: $\mathrm{EPS}=2^{-52}$, which is roughly $2{\cdot}22\times10^{-16}$. Similarly the computations of $\mathbf{B}^{+}$ and $\mathbf{C}^{+}$ in each iterative cycle can also be performed.

The numerical rank, rank$\underline{\mathbf{X}}$, of the three-way data array $\underline{\mathbf{X}}$ has been distinguished from the numerical ranks of **A**, **B** and **C** in this study, though the numerical rank of $\underline{\mathbf{X}}$ is equal to the column numbers of the latter owing to the sample number I, which is sometimes less than the number of factors, N. The rank deficiency problem, present possibly in inverse computations of **A**, **B** and **C** with N columns in each iterative cycle, may be handled automatically and preferably by using the computations of their Moore–Penrose inverses. Thus it is possible to apply it for decomposition of three-way data arrays when N is equal to or larger than the number of chemical species in order to obtain accurate concentration predictions. This may be considered one of the main advantages of the ATLD algorithm. In each iteration cycle, **B** and **C** are normalized to unit length columnwise, as is done frequently.[26] The process of updating **A**, **B** and **C** is continued until the one stopping criterion with a maximum of M iterative cycles is satisfied. In ATLD, M is usually set to be 30. This criterion is

$$\left|\frac{\sigma^{(m)}-\sigma^{(m-1)}}{\sigma^{(m-1)}}\right|\leq\varepsilon \tag{16}$$

where m denotes the iteration number in the decomposition of the three-way array and ε is some arbitrary small value (in this study, usually $\varepsilon=10^{-6}$). The above-mentioned procedures may also be called an improved PARAFAC algorithm without any constraint conditions. Their solutions also have the rotational uniqueness which has been considered a major advantage of this type of trilinear model.[20] In the following section the performance of the ATLD algorithm will be compared with that of the traditional PARAFAC algorithm.

After finishing the iterative procedure, if the columns corresponding to the components of interest in the finally obtained estimates of **A**, **B** and **C** are appropriately postprocessed according to the uniqueness property of trilinear decomposition,[20] then the physical significance of **A**, **B** and **C** can be more easily understood. The final concentration estimates in unknown samples may be obtained by a plot of relative concentration contributions of each component of interest versus it standard concentrations in the reference samples, similar to a calibration curve plot for one component, or by regression of relative concentration contributions of each component of interest against its standard concentrations.

Determination of number of factors, *N*, present in three-way data array and selection of starting values

Determination of the number of factors, N, in a three-way data matrix $\underline{\mathbf{X}}$ is an important problem that is closely related to the rank problem in three-way arrays.[20] In this study the number of factors will

© 1998 John Wiley & Sons, Ltd.

be determined as follows (see Appendix I):

$$N \geq \mathrm{rank}\underline{\mathbf{X}} = \max\{\mathrm{rank}(\mathbf{X}_{\mathrm{P}}^{I}), \mathrm{rank}(\mathbf{X}_{\mathrm{P}}^{J}), \mathrm{rank}(\mathbf{X}_{\mathrm{P}}^{K})\} \quad (17)$$

where rank$\underline{\mathbf{X}}$ is the numerical rank of the three-way data array $\underline{\mathbf{X}}$; rank$\mathbf{X}_{\mathrm{P}}^{I}$, rank$\mathbf{X}_{\mathrm{P}}^{J}$ and rank$\mathbf{X}_{\mathrm{P}}^{K}$ can be considered as the estimated numerical ranks of the relative concentration, relative spectrum and relative elution profile matrices respectively. Here RANK(.) calculates the rank of a matrix, where rank$\underline{\mathbf{X}}$ has to be calculated prior to the decomposition of the three-way data array. Let $\mathbf{X}_{\mathrm{P}}^{I}$, $\mathbf{X}_{\mathrm{P}}^{J}$ and $\mathbf{X}_{\mathrm{P}}^{K}$ be two-way partitioned matrices obtained by unfolding $\underline{\mathbf{X}}$ in the following way:

$$\mathbf{X}_{\mathrm{P}}^{I} = [\mathbf{X}_{\cdot\cdot 1}\ \mathbf{X}_{\cdot\cdot 2}\ \ldots\ \mathbf{X}_{\cdot\cdot K}] \quad (18)$$

$$\mathbf{X}_{\mathrm{P}}^{J} = [\mathbf{X}_{1\cdot\cdot}\ \mathbf{X}_{2\cdot\cdot}\ \ldots\ \mathbf{X}_{I\cdot\cdot}] \quad (19)$$

$$\mathbf{X}_{\mathrm{P}}^{K} = [\mathbf{X}_{\cdot 1\cdot}\ \mathbf{X}_{\cdot 2\cdot}\ \ldots\ \mathbf{X}_{\cdot J\cdot}] \quad (20)$$

If we calculate their SVDs respectively, then we obtain their singular value vectors. The variance can be calculated directly from the singular values. By comparing the magnitudes of each of the first singular values as well as their variances, we may determine rank$\mathbf{X}_{\mathrm{P}}^{I}$, rank$\mathbf{X}_{\mathrm{P}}^{J}$ and rank$\mathbf{X}_{\mathrm{P}}^{K}$ respectively and then consider the maximum of them as the estimated numerical rank of the three-way data array. Usually the numbers of factors for PARAFAC and ATLD are selected as N=rank$\underline{\mathbf{X}}$ and N=rank$\underline{\mathbf{X}}$ + (0–2) respectively owing to experimental errors.

In addition, the first singular vectors obtained from SVD of $\mathbf{X}_{\mathrm{P}}^{J}$ and $\mathbf{X}_{\mathrm{P}}^{K}$ can also be chosen as one of the starting values of **B** and **C** respectively for decomposition of the three-way data array. The starting values of **A** in this study may be given according to (12). This is the way in which the starting values are chosen in the Kroonenberg/de Leeuw ALS algorithm.[26] In this study a way to produce random starting values of **A**, **B** and **C** is adopted. The former converges faster than the latter. Usually the number of iterations is less than ten.

Determination of column position corresponding to each component

According to a series of Monte Carlo simulations as mentioned below, it was found that the column position corresponding to each component of interest in the estimates of **A**, **B** and **C** is variable with random starting values. Therefore it is necessary to determine the column position corresponding to each species or component prior to plotting or simple regression of relative concentration contributions for each component against its standard (known) concentrations. Once its column position is determined, the concentrations of each component of interest in the unknown samples can be given at once. Designing an appropriate concentration matrix of reference samples is helpful to determine rapidly the column position corresponding to each component of interest. The column position may also be found according to the characteristics of known spectra and time profiles of the components of interest.

Algorithms for both traditional PARAFAC and ATLD

According to the above-mentioned principles, an algorithm for ATLD was developed as given in Appendix II. In order to illustrate the advantage of ATLD over the traditional PARAFAC method, a basic algorithm for PARAFAC was also used. This algorithm is similar to that of ATLD except that three iterative substeps based on (5)–(7) were used.[17] In addition, a maximum of 200 iterative cycles in the traditional PARAFAC algorithm were used.

© 1998 John Wiley & Sons, Ltd.

EXPERIMENTAL

Simulations

Simulations were performed in the MATLAB (MathWorks Inc., Natick, MA, U.S.A.) programming environment. Two sets of simulated data were constructed: (i) one reference and one unknown sample containing one interferent (S-I) and (ii) four reference and two unknown samples containing one interferent (S-II). Both were constructed from a pool of six simulated samples, four simulated chromatographic profiles and four simulated spectral profiles. Table 1 gives the concentration values of four components in six simulated samples, designated as #1–#6. The four simulated spectra are each 40 digitized channels with complete spectral overlap. The first spectrum, $\mathbf{b}_1$, comprises two Gaussian curves centred at channels 6 and 20 with widths at half-height of four and eight channels respectively. The second spectrum, $\mathbf{b}_2$, comprises two Gaussian curves centred at channels 9 and 22 with widths at half-height of four and eight channels respectively. The third spectrum, $\mathbf{b}_3$, is the sum of two Gaussian curves centred at channels 12 and 30 with widths at half-height of four and eight channels respectively. The fourth spectrum, $\mathbf{b}_4$, is the sum of two Gaussian curves centred at channels 16 and 35 with widths at half-height of four and eight channels respectively. The relative intensities of all four spectra are 2:1. The simulated spectra are shown in Figure 2(1). The simulated chromatographic profiles, designated as $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_3$ and $\mathbf{c}_4$, are each 60-channel Gaussian curves with a width at half-height of six channels for $\mathbf{c}_1$, $\mathbf{c}_3$ and $\mathbf{c}_4$ and of ten channels for $\mathbf{c}_2$, and centred at channels 20, 22, 30 and 40 respectively (Figure 2(2)). The chromatographic profile $\mathbf{c}_1$ may be embodied in $\mathbf{c}_2$.

Simulated data set I: one reference and one unknown sample with one interferent

For simulated data set I (S-I), simulated sample #1 containing species 1 was chosen as the reference sample in which the two-dimensional responses were constructed by the formula

$$\mathbf{X}_{1\,..} = a_{11}\mathbf{b}_1\mathbf{c}_1^{\mathrm{T}} \tag{21}$$

Simulated sample #3 with one unknown interferent, species 3, was used as an unknown sample in which the response values were constructed by

$$\mathbf{X}_{3\,..} = a_{31}\mathbf{b}_1\mathbf{c}_1^{\mathrm{T}} + a_{33}\mathbf{b}_3\mathbf{c}_3^{\mathrm{T}} \tag{22}$$

Simulated data set II: four reference and two unknown samples

For simulated data set II (S-II), besides $\mathbf{X}_{1\,..}$ and $\mathbf{X}_{3\,..}$, the response values of other samples were generated by the equations

Table 1. Compositions of six simulated samples

	Composition (relative concentration)			
Samples	Species 1	Species 2	Species 3	Species 4
#1	1·000	0·000	0·000	0·000
#2	0·000	1·000	0·000	0·000
#3	1·000	0·000	1·000	0·000
#4	1·000	1·000	1·000	1·000
#5	1·000	1·000	1·000	1·000
#6	1·000	2·000	1·000	2·000

© 1998 John Wiley & Sons, Ltd.

$$\mathbf{X}_{2\,..} = a_{21}\mathbf{b}_2\mathbf{c}_2^{\mathrm{T}} \tag{23}$$

$$\mathbf{X}_{4\,..} = \sum_{n=1}^{3} a_{4n}\mathbf{b}_n\mathbf{c}_n^{\mathrm{T}} \tag{24}$$

$$\mathbf{X}_{5\,..} = \sum_{n=1}^{4} a_{5n}\mathbf{b}_n\mathbf{c}_n^{\mathrm{T}} \tag{25}$$

$$\mathbf{X}_{6\,..} = \sum_{n=1}^{4} a_{6n}\mathbf{b}_n\mathbf{c}_n^{\mathrm{T}} \tag{26}$$

The total number of samples, *I*, was six. S-II was used to simulate the simultaneous determination of multicomponents in several unknown samples containing an interferent, where species 4 was considered as an unknown interferent. The two-dimensional spectra of six simulated samples are shown in Figure 3.

Random errors

One type of random error was used to simulate the effects of random instrumental noise in the detector. After constructing second-order spectra of each mixture, normally distributed noise with a

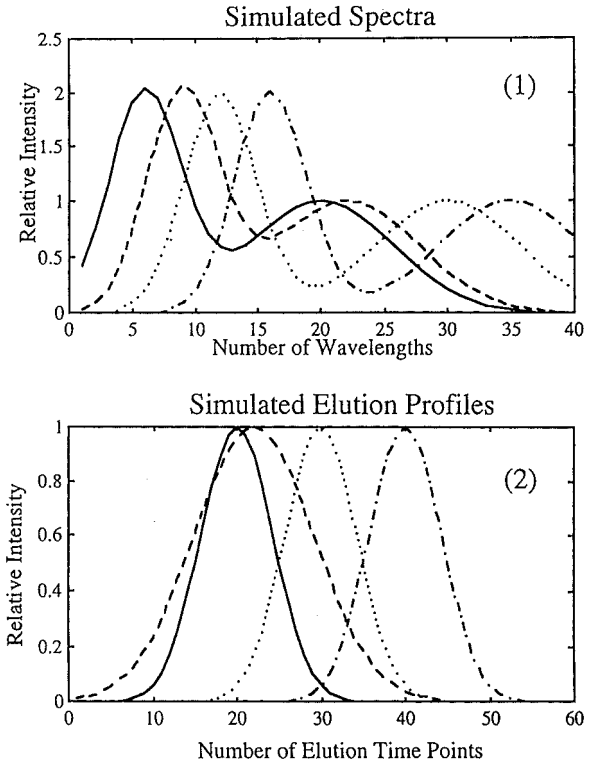


Figure 2. Simulated spectra (1) and elution profiles (2) used to generate data in two sets of simulations: (1) $\mathbf{b}_1$, ——; $\mathbf{b}_2$, – – –; $\mathbf{b}_3$, · · · ·;$\mathbf{b}_4$, – · –; (2) $\mathbf{c}_1$, ——; $\mathbf{c}_2$, – – –; $\mathbf{c}_3$, · · · ·;$\mathbf{c}_4$, – · –

© 1998 John Wiley & Sons, Ltd.

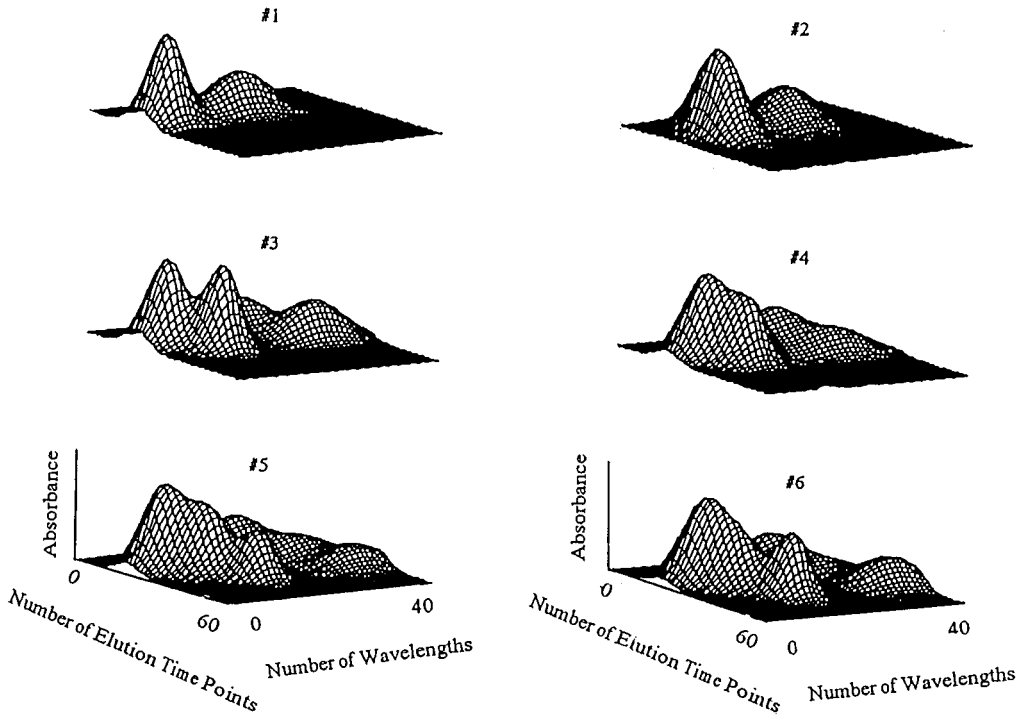


Figure 3. Two-dimensional spectra of six simulated samples

mean of zero and standard deviations of 0·5%, 1·0% and 2·0% relative to the maximum of the two-dimensional spectrum of each mixture was added. A total of 50 replicate measurements were calculated at each noise level for the Monte Carlo simulation.

HPLC–DAD data

A high-performance liquid chromatography (HPLC) system, comprising an SSC-3110T pump and an SSC-3100C2 microprocessor control unit (Senshu Scientific Co., Ltd., Japan) with an L-Column ODS ($4{\cdot}6\times 150$ mm^2, Chemical Inspection and Testing Institute, Japan) and a diode array detector (Shimadzu SPD-M10AV), was used for quantitative analysis of mixtures of chlorinated aromatic hydrocarbons such as *o*-dichlorobenzene (*o*-DCB), *p*-chlorotoluene (*p*-CT) and *o*-chlorotoluene (*o*-CT). The obtained elution time profiles and spectra for each compound are shown in Figure 4. In addition, chlorobenzene was added to these samples as an internal retention time standard in order to correct the obtained retention time of the samples. The average retention time of chlorobenzene was $3{\cdot}766\pm 0{\cdot}020$ min (n=24). The column temperature was controlled at $25{\cdot}0\pm 0{\cdot}5$ °C. A mixture of methanol and water (80:20, w/w) was used as an eluent. The flow rate of the eluent was $1{\cdot}0$ ml min^{-1}.

The two-dimensional response data were collected with a Compaq Prolinea 4/33S personal computer with CLASS-M10A program (Shimadzu). The reagents used were of analytical grade.

The response data used for second-order calibration were taken over an elution time range of 4·30–5·30 min (Δt=0·04 min, K=26) and a wavelength range of 200·0–280·0 nm ($\Delta\lambda\approx 1{\cdot}27$ nm, J=64). These data were then transferred to a Macintosh computer (PowerBook 550c) and combined into a three-way response array.

© 1998 John Wiley & Sons, Ltd.

Nine samples, designated as #1–#9, were analysed in all, in which the concentrations of each component are shown in Table 2. Three sets of experimental data were used to verify the performances of the PARAFAC and ATLD algorithms. In the first set, sample #1 containing *p*-chlorotoluene was chosen as a reference sample and #7 with two interferents, *o*-dichlorobenzene and *o*-chlorotoluene, as an unknown sample. In the second set, #1 was chosen as a reference sample and #5 and #6 as unknown samples with larger amounts of interferents. The number of samples, *I*, was three. In the third set, #1– 4 were chosen as reference samples and #7–#9 as unknown samples with an interferent, *o*-dichlorobenzene. The total number of samples was seven.

RESULTS AND DISCUSSION

Basic comparison of traditional PARAFAC algorithm and ATLD algorithm

To obtain the solution to numerical problems, it is important to develop computer programs with good performance. Speed, precision, stability and reliability are factors that have to be considered. In this

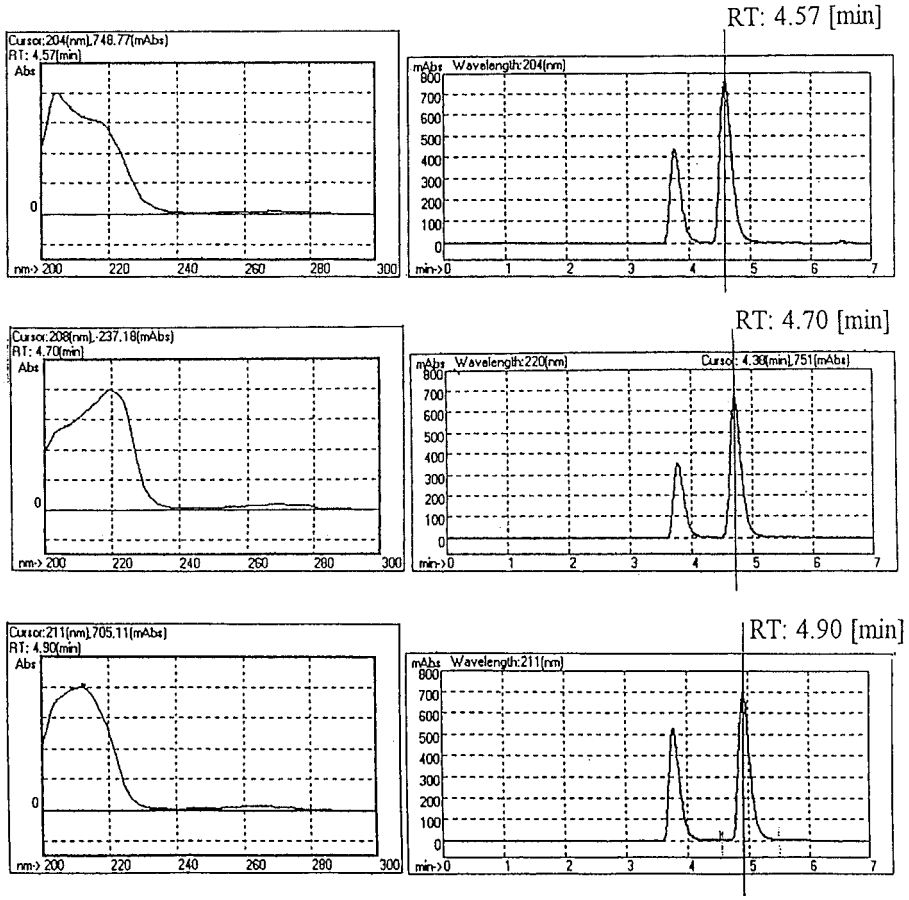


Figure 4. Standard spectra (left column) and time profiles (right column) of components to be analysed: (top), 91·2 μg ml^{-1} *o*-dichlorobenzene; middle, 75·6 μg ml^{-1} *p*-chlorotoluene; bottom, 91·2 μg ml^{-1} *o*-chlorotoluene. The first peak in the time profiles is chlorobenzene

© 1998 John Wiley & Sons, Ltd.

Table 2. Compositions of nine real samples

	Concentration (μg ml^{-1})			
Sample	*o*-Dichlorobenzene	*p*-Chlorotoluene	*o*-Chlorotoluene	Chlorobenzene[a]
#1	0·0	75·6	0·0	62·4
#2	0·0	0·0	91·2	62·4
#3	0·0	50·4	30·4	62·4
#4	0·0	25·2	60·8	62·4
#5	152·2	12·6	15·2	62·4
#6	15·2	12·6	152·0	62·4
#7	60·8	25·2	91·2	62·4
#8	91·2	50·4	30·4	62·4
#9	30·4	75·6	60·8	62·4

[a] Chlorobenzene was added to these samples as an internal elution time standard in order to keep the relative retention times obtained in a batch analysis as consistent as possible.

section we will compare the performances of the two algorithms by considering these factors.

The speed of an algorithm is related to the working efficiency when it is applied to solve a numerical problem. In the case of the PARAFAC and ATLD algorithms the speed refers to both the amount of time used by each iterative cycle and the convergence rate. Table 3 lists the measured amount of time used by each iterative cycle for both the PARAFAC and ATLD algorithms. It was observed that the amount of time used by each iterative cycle in the latter was generally shorter than that of the former by about 60%–75%. Figures 5 and 6 show different loss function curves when both algorithms are used to decompose two simulated data sets, S-I and S-II respectively, without any noise. Note that ε is about 10^{-11}. From Figure 5 it was seen that both seem to obtain identical results except for different rates of convergence for S-I. When convergence is close to being achieved, the corresponding number of iterations, M, is usually about 80 for PARAFAC and about five for ATLD. From Figure 6 it was seen that ATLD converges quickly to the minimum (M is usually less than ten), but convergence in PARAFAC cannot be achieved until the number of iterations is over 2000. Obviously, in PARAFAC, convergence is rather slow. In other words, ATLD converges faster with more satisfactory estimates than PARAFAC. In particular, in the case of complicated data this advantage of the ATLD algorithm over the PARAFAC algorithm can be clearly observed.

In addition, both the PARAFAC and ATLD alorithms were used to decompose a series of two sets

Table 3. Amount of time for each iterative cycle in traditional PARAFAC and ATLD algorithms applied to two sets of simulated data with noise level 1·0%

Simulated data set	Amount of time for each iterative cycle[a] (s)			
	PARAFAC		ATLD	
S-I	1·981 ± 0·011 (N=2)[b]	2·351 ± 0·014 (N=3)	1·364 ± 0·007 (N = 2)	1·559 ± 0·010 (N = 3)
S-II	3·783 ± 0·051 (N=4)	4·276 ± 0·043 (N = 5)	2·751 ± 0·045 (N = 4)	3·017 ± 0·038 (N = 5)

[a] Average and standard deviation of 30 iterative cycles.
[b] N is the number of factors.

© 1998 John Wiley & Sons, Ltd.

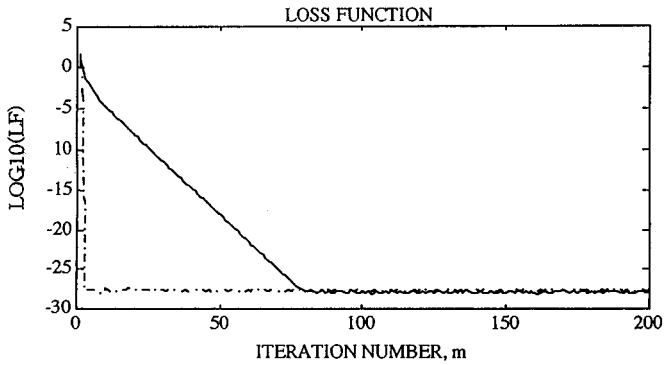


Figure 5. Loss functions of S-I without added noise using traditional PARAFAC (——) and ATLD (– · –) algorithms; N=2

of simulated data with various noise levels of 0·5%, 1·0% and 2·0%. Similar results were obtained. The PARAFAC algorithm with an appropriate number of factors can often provide an accurate solution in spite of slow convergence. In the ATLD algorithm, faster convergence can be obtained and

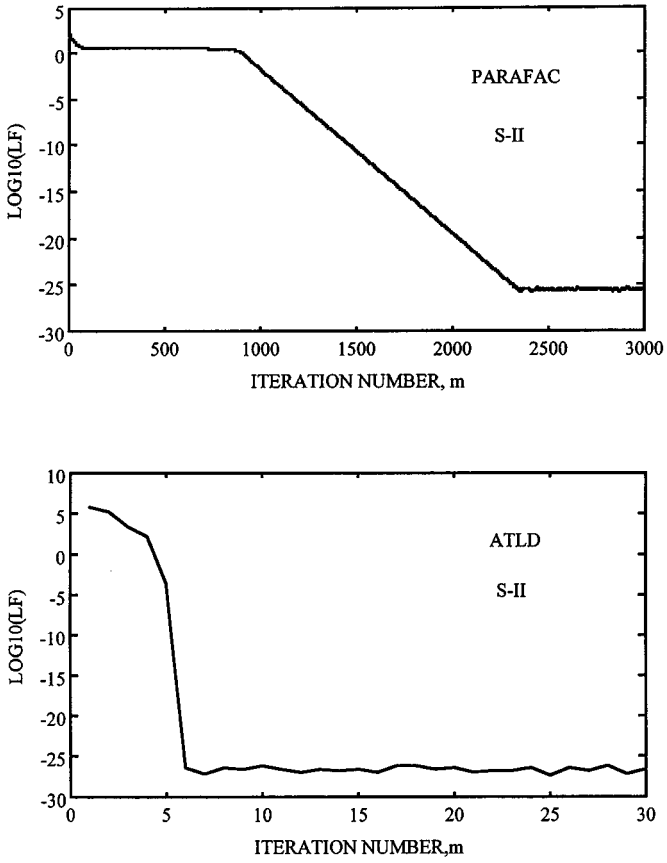


Figure 6. Loss functions of S-II without added noise using traditional PARAFAC (top) and ATLD (bottom) algorithms; N=4

© 1998 John Wiley & Sons, Ltd.

Table 4. Effect of selecting different numbers of factors on traditional PARAFAC and ATLD algorithms for decomposition of S-I with noise level of 1·0%[a]

Trilinear model	PARAFAC						ATLD					
	M	$\log_{10}(LF_M)$	Estimated A				M	$\log_{10}(LF_M)$	Estimated A			
One factor	10	2·209	14·059				12	2·209	14·060			
(N=1)			20·455						20·455			
Two factors	39	0·296	15·510	0·041			6	0·310	0·037	15·511		
(N=2)			15·540	15·202						15·199	15·538	
Three factors	200	0·288	1·179	13·918	10·103		30	0·292	0·262	15·511	0·038	
(N=3)			27·548	18·909	−3·205				−0·110	15·537	15·201	
Four factors	200	0·279	4·137	14·417	−3·168	−0·038	30	0·285	0·038	0·269	0·218	15·511
(N=4)			3·743	23·367	11·280	7·235			15·203	−0·107	−0·143	15·536
Five factors	200	0·249	16·876	1·647	19·267	0·706	−1·754	30	0·254	0·230	0·037	15·513	0·223	0·255
(N=5)			−3·025	11·505	14·042	−0·022	21·046			−0·141	15·206	15·535	−0·135	−0·106

[a] M and $\mathrm{Log}_{10}(LF_M)$ are the number of iterations and the common logarithm of the loss function respectively when the iterative procedure is finished. **A** is the relative concentration matrix of size $2 \times N$ and ε is about 10^{-11}.

a set of accurate solutions can always be given. It was also observed that the convergence by PARAFAC was not always achieved when it was used to decompose the simulated data sets with random noise.

The PARAFAC algorithm with a number of factors that is larger than the appropriate number of factors usually give unreasonable concentration estimates. As reported by other researchers,[19, 20] it does not always converge to chemically meaningful solutions, especially in the presence of two-factor degeneracies. The ATLD algorithm does not suffer from these problems. Accurate concentration estimates can always be obtained by the ATLD algorithm provided that the number of factors chosen is equal to or larger than the appropriate number of factors present in **X** (see Table 4). ATLD seems to benefit from using the Moore–Penrose generalized inverse based on SVD in each iterative cycle. The useful information present in **X** can be extracted into several factors, the number of which is equal to the chemical rank. Use of an excessive number of factors does not usually affect the estimates of the component of interest in unknown samples containing an uncalibrated interferent. This advantage over the traditional PARAFAC algorithm has been clearly confirmed in more than 100 simulations. Table 5 summarizes the performances of both algorithms for S-I and S-II. How to determine the appropriate number of factors is obviously important for trilinear decomposition, especially for the

Table 5. Performance of traditional PARAFAC and ATLD algorithms for S-I and S-II

Selected number of factors[a]	Simulated data set without noise	PARAFAC		ATLD	
		Convergence rate	Estimated **A**	Convergence rate	Estimated **A**
$N<N^*$	S-I	Slow	Not accurate	Fast	Not accurate
	S-II	Very slow	Not accurate	Fast	Not accurate
$N=N^*$	S-I	Slow	Usually accurate	Fast	Always accurate
	S-II	Very slow	Sometimes accurate	Fast	Always accurate
$N=N^*+(1-2)$	S-I	Slow	Mostly inaccurate	Fast	Always accurate
	S-II	Very slow	Mostly inaccurate	Fast	Always accurate

[a] N^* is the appropriate number of factors or chemical rank present in three-way data array.

© 1998 John Wiley & Sons, Ltd.

traditional PARAFAC algorithm. In the following section we discuss the problem of choosing the appropriate number of factors used to carry out trilinear decomposition.

Determining the number of factors in both sets of simulated data with various noise levels

S-I is a two sample problem that is usually solved by the GRAM method. The data array to be decomposed is of size $2\times40\times60$. S-II is a six-sample calibration problem, which is used to demonstrate a multicomponent calibration problem composed of several reference and several unknown samples in the presence of an uncalibrated interferent, that is usually solved by the TLD method. The array to be decomposed is of size $6\times40\times60$. By using the above-mentioned procedure, the numbers of factors present in these three-way data arrays may be decided (see Tables 6 and 7). By comparing the magnitudes of each first singular value as well as their percentage variance criterion, we may determine the numerical ranks, rank**A**, rank**B** and rank**C** respectively, and then consider the maximum of them as the numerical rank of the three-way data array. For S-I the estimated numbers

Table 6. Summary of first singular values obtained from $\mathbf{X}_{\mathrm{P}}^{I}$, $\mathbf{X}_{\mathrm{P}}^{J}$ and $\mathbf{X}_{\mathrm{P}}^{K}$ and determination of number of factors in S-I[a]

Factor no.	First singular value (percentage variance (%))		
	$\mathbf{X}_{\mathrm{P}}^{I}$	$\mathbf{X}_{\mathrm{P}}^{J}$	$\mathbf{X}_{\mathrm{P}}^{K}$
	Noise level 0·0%		
1	26·43 (89·90)	25·71 (85·07)	25·23 (81·97)
2	8·83 (10·10)	10·77 (14·93)	11·83 (18·03)
3		0·00 (0·00)	0·00 (0·00)
4		0·00 (0·00)	0·00 (0·00)
	*N**=2 for both PARAFAC and ATLD		
	Noise level 0·5%		
1	26·42 (89·97)	25·68 (85·02)	25·21 (81·91)
2	8·82 (10·03)	10·76 (14·91)	11·83 (18·03)
3		0·18 (0·00)	0·16 (0·00)
4		0·17 (0·00)	0·16 (0·00)
	N=2 for PARAFAC; *N*=2 or 3 for ATLD		
	Noise level 1·0%		
1	26·48 (89·90)	25·73 (84·85)	25·26 (81·77)
2	8·88 (10·10)	10·79 (14·91)	11·85 (17·99)
3		0·33 (0·01)	0·32 (0·01)
4		0·32 (0·01)	0·31 (0·01)
	N=2 for PARAFAC; *N*=2 or 3 for ATLD		
	Noise level 2·0%		
1	26·51 (89·69)	25·71 (84·31)	25·24 (81·33)
2	8·99 (10·31)	10·82 (14·72)	11·77 (17·69)
3		0·66 (0·06)	0·65 (0·05)
4		0·66 (0·06)	0·63 (0·05)
	N=2 for PARAFAC; *N*=2 or 3 for ATLD		

[a] Estimated number of components present in three-dimensional data array given by rank $\underline{\mathbf{X}}$=max-{rank($\mathbf{X}_{\mathrm{P}}^{I}$), rank($\mathbf{X}_{\mathrm{P}}^{J}$), rank($\mathbf{X}_{\mathrm{P}}^{K}$)}. For PARAFAC rank $\underline{\mathbf{X}}\rightarrow N$ and for ATLD rank $\underline{\mathbf{X}}$ +(0–2)$\rightarrow N$ owing to experimental errors.

© 1998 John Wiley & Sons, Ltd.

of factors are two for PARAFAC and around two or three for ATLD when the standard deviations of the added random noise are 0·5%, 1·0% and 2·0% relative to the maximum of the second-order spectrum of each mixture. For S-II the estimated numbers of factors are four for PARAFAC and around four or five for ATLD. Note that the unknown samples #5 and #6 contain one interferent, species 4.

With regard to the development of further methods for deducing the number of factors present in a three-way data array, Malinowski's work[27] should be considered as an important foundation.

Table 7. Summary of first singular values obtained from $\mathbf{X}_p^I$, $\mathbf{X}_p^J$ and $\mathbf{X}_p^K$ and determination of number of factors in S-II[a]

Factor no.	First singular value (percentage variance (%)) $\mathbf{X}_p^I$	$\mathbf{X}_p^J$	$\mathbf{X}_p^K$
	Noise level 0·0%		
1	96·90 (94·87)	93·90 (87·95)	92·67 (86·76)
2	18·58 (3·49)	30·51 (9·40)	31·80 (10·21)
3	9·69 (0·95)	15·24 (2·35)	16·96 (2·91)
4	8·29 (0·70)	5·44 (0·30)	3·41 (0·12)
5	0·00 (0·00)	0·00 (0·00)	0·00 (0·00)
6	0·00 (0·00)	0·00 (0·00)	0·00 (0·00)
	N*=4 for both PARAFAC and ATLD		
	Noise level 0·5%		
1	96·90 (94·84)	93·30 (87·92)	92·66 (86·73)
2	18·62 (3·50)	30·50 (9·40)	31·79 (10·21)
3	9·70 (0·95)	15·25 (2·35)	16·97 (2·91)
4	8·31 (0·70)	5·43 (0·30)	3·42 (0·12)
5	0·94 (0·01)	0·44 (0·00)	0·41 (0·00)
6	0·66 (0·00)	0·42 (0·00)	0·40 (0·00)
	N=4 for PARAFAC; N=4 or 5 for ATLD		
	Noise level 1·0%		
1	96·88 (94·74)	93·25 (87·77)	92·62 (86·59)
2	18·72 (3·54)	30·53 (9·41)	31·80 (10·21)
3	9·76 (0·96)	15·31 (2·37)	17·04 (2·93)
4	8·39 (0·71)	5·51 (0·31)	3·45 (0·12)
5	1·92 (0·04)	0·88 (0·01)	0·80 (0·01)
6	1·28 (0·02)	0·86 (0·01)	0·80 (0·01)
	N=4 for PARAFAC; N=4 or 5 for ATLD		
	Noise level 2·0%		
1	96·91 (94·43)	93·36 (87·44)	92·65 (86·21)
2	19·04 (3·64)	30·65 (9·30)	31·78 (10·14)
3	9·90 (0·98)	15·39 (2·37)	17·06 (2·92)
4	8·55 (0·73)	5·62 (0·31)	3·54 (0·13)
5	3·76 (0·14)	1·75 (0·03)	1·63 (0·03)
6	2·62 (0·07)	1·69 (0·03)	1·60 (0·03)
	N=4 for PARAFAC; N=4 or 5 for ATLD		

[a] See footnote to Table 6.

© 1998 John Wiley & Sons, Ltd.

Predicting concentrations of component(s) of interest in unknown samples with uncalibrated interferent(s) by using both PARAFAC and ATLD algorithms

Both algorithms with the above-estimated numbers of factors were applied to the decomposition of two sets of simulated data with different noise levels (S-I and S-II). Some of the results obtained using ATLD are shown in Figures 7 and 8. In the case of S-I the column position corresponding to the component of interest, species 1, in the relative concentration matrix is judged by comparing the characteristics of the obtained relative spectra **B** and relative chromatograms **C** with the known spectrum and elution profile of the component of interest. The final concentration of species 1 was calculated by using its relative concentration ratio. Table 8 shows the effects of random noise with different levels on the predictive performance of both the PARAFAC and ATLD algorithms for S-I. Both algorithms with an appropriate number of factors obtained identical concentration estimates. The magnitude of the bias increases linearly with the noise level.

Table 9 lists the recoveries of three components of interest in S-II using the ATLD algorithm. The results of 50 Monte Carlo simulations using the traditional PARAFAC algorithm could not be obtained owing to very slow convergence and poor reproducibility in convergence resolution.

It was also observed that even in the case of large condition number for one or two of the estimated **A**, **B** and **C** (e.g. cond(**B**) in S-I is sometimes larger than 100), ATLD still produced satisfactory concentration predictions in a batch analysis of unknown samples even in the presence of interferents.

By comparing the results obtained from these simulated data sets, it was found that simultaneous determination of several components of interest and qualitative analysis of all components present in two or more samples, which generally contain reference and unknown samples, are possible in the presence of one interferent or more. As the noise level increases, correct concentration estimates and correct profiles can still be obtained, but the standard deviations of the estimated concentrations will increase.

It is also shown that both PARAFAC with $N=N^*$ and ATLD with $N \geq N^*$ may be considered as

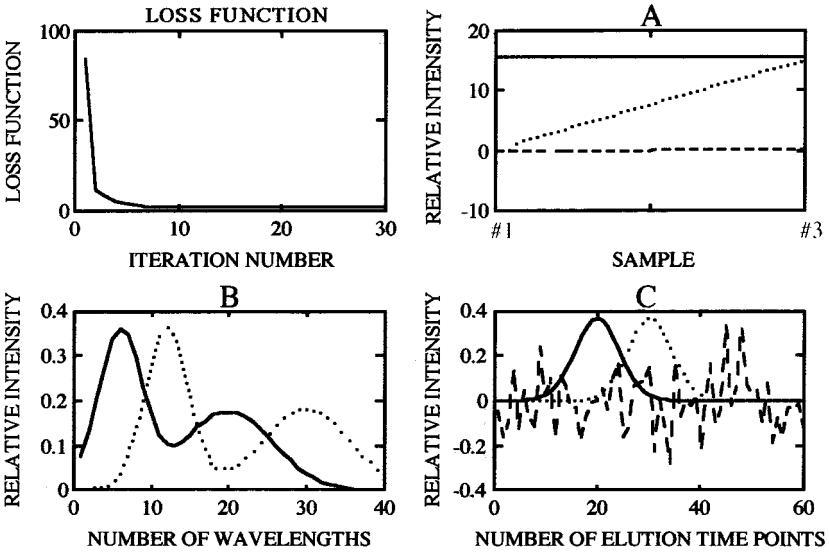


Figure 7. One set of results for S-I using ATLD algorithm: A, relative concentrations; B, relative spectra; C, relative elution profiles. The noise level is 1·0% and N=3. The predicted concentration of species 1 (——) is $15{\cdot}61 \times 1{\cdot}000/15{\cdot}58 = 1{\cdot}002$ and its recovery is $1{\cdot}002/1{\cdot}000 \times 100 = 100{\cdot}2\%$. M=30

© 1998 John Wiley & Sons, Ltd.

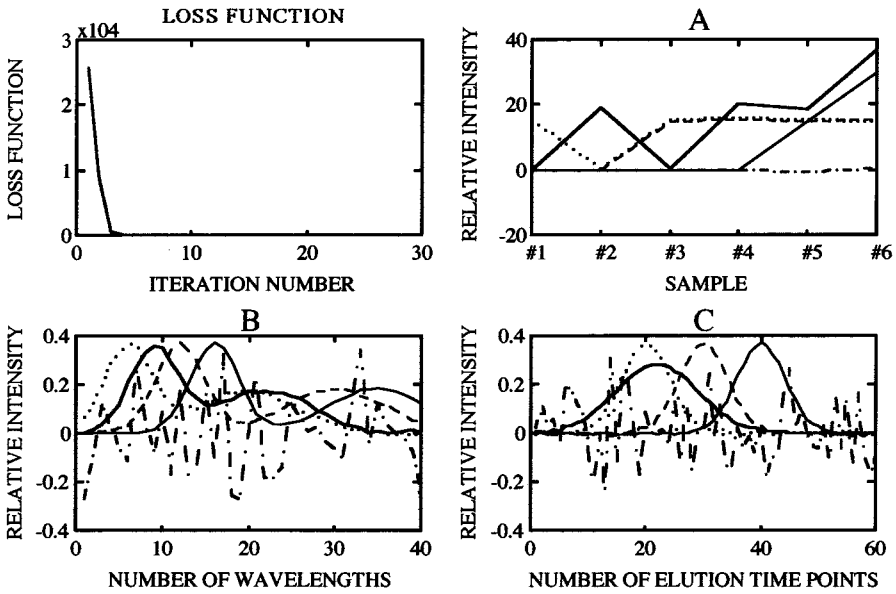


Figure 8. One set of results for S-II using ATLD algorithm: A, relative concentrations; B, relative spectra; C, relative elution profiles. The noise level is 1·0% and N=5. See Table 9 for predicted concentrations of species 1 (· · · ·), species 2 (——) and species 3 (– – –). M=30

trilinear component analysis methods or trilinear factor analysis methods rather than trilinear principal component analysis (PCA) or trilinear principal factor analysis despite similar forms.[26, 27] Note that N^* is the true physical or chemical rank. In a sample measured by HPLC–DAD, for example, N^* is the total number of detectable chemical species. In a three-way trilinear data array $\underline{\mathbf{X}}$ of size $I \times J \times K$ and with rank $\underline{\mathbf{X}} = N^*$, if and only if one of I, J and K and the product of the other two are equal to or larger than N^*, then rank$\underline{\mathbf{X}}$ equals N^*, i.e. the true physical or chemical rank N^* can be obtained using (17). The difference between trilinear component analysis and trilinear PCA is that the former is performed without any constraints and the latter imposes orthogonality conditions.[19, 20] In this sense, trilinear

Table 8. Effect of random noise with different levels on predictive performance of traditional PARAFAC and ATLD algorithms for S-I

	Predicted concentration of species 1[a]			
	PARAFAC (M=200)[b]		ATLD (M=30)	
Random noise level (%)	N[c]	$x \pm s$	N	$x \pm s$
0·0	2	1·0000 ± 0·0000	2	1·0000 ± 0·0000
0·5	2	0·9999 ± 0·0011	3	0·9999 ± 0·0014
1·0	2	0·9994 ± 0·0025	3	0·9999 ± 0·0023
2·0	2	0·9998 ± 0·0053	3	1·0006 ± 0·0051

[a] A total of 50 replicate measurements were calculated at each noise level for each simulation. The actual concentration of species 1 in simulated sample #3 is 1·0000.
[b] M is the maximum number of iterative cycles.
[c] N is the number of factors.

© 1998 John Wiley & Sons, Ltd.

Table 9. Recoveries of three components of interest in S-II using ATLD algorithm[a]

Random noise level (%)	Recovery (mean ± standard deviation) (%)		
	Species 1	Species 2	Species 3
	For unknown sample #5		
0·0	100·0±0·0	100·0±0·0	100·0±0·0
0·5	99·6±1·3	99·3±1·3	99·8±1·7
1·0	97·0±2·9	96·5±4·8	99·3±3·0
2·0	96·9±5·1	89·5±8·3	95·7±4·6
	For unknown sample #6		
0·0	100·0±0·0	100·0±0·0	100·0±0·0
0·5	99·0±2·0	99·8±1·0	98·9±0·9
1·0	98·4±3·6	93·5±4·0	97·2±3·3
2·0	97·2±6·9	87·4±6·6	96·7±5·3

[a] A total of 50 replicate measurements were calculated at each noise level.

decomposition or PARAFAC should not be viewed as a generalization of PCA to second-order data arrays. In this study the trilinear model has been considered the basic structure present in three-way data arrays rather than a generalization of PCA or singular value decomposition of three-way data arrays. Therefore it should be appropriate to refer to the above-mentioned method which decomposes three-way data arrays as ATLD. Usually, *N*-factor trilinear component analysis or PARAFAC ($1<N<N^*$) is suitable for qualitative analysis or quasi-quantitative analysis, while trilinear component analysis with $N \geq N^*$ can be applied for accurate quantitative analysis. In the field of analytical chemistry, theoretically, only *N* -factor trilinear component analysis ($N \geq N^*$) is suitable for second-order calibration.

In addition, it was observed that the final loss function values of the ATLD algorithm are often larger than those of the PARAFAC algorithm when model residuals exist, but the obtained estimates of **A**, **B** and **C** of the former are closer to the theoretical values than those of the latter (see Table 4). In fact, this is related to the more fundamental problem of whether a basic trilinear structure exists in three-model three-way data arrays. Obviously a basic trilinear structure exists in HPLC–DAD data sets because they obey the generalized Lambert–Beer law. The ATLD algorithm decomposes, in a true trilinear sense (see (9)–(14)), the three-way data array into **A**, **B** and **C**. The PARAFAC algorithm is based on the only frontal slice matrices $\mathbf{X}_{\cdot\cdot k}$ (k=1, 2, . . . , K) according to Cattell's idea of parallel proportional profiles.[17] This may be considered as the main reason why the two iterative procedures produce different results. In ATLD, Moore–Penrose generalized inverses are directly computed based on SVD, so not only are the least squares properties maintained under a trilinear structure but also the introduction of large errors is avoided, especially in the case of rank deficiency. In PARAFAC, three procedures for computing inverse matrices are included in each iterative cycle that could easily introduce large errors. The third possible reason is that in ATLD one row of a matrix is estimated based on the other two estimated matrices and the procedure for obtaining the diagonals of a matrix is based on a trilinear structure, while in PARAFAC four addition procedures using a series of matrices are contained in each iterative cycle. The addition of matrices containing positive and negative elements may lead to rank deficiency. Therefore ATLD is numerically more efficient than PARAFAC.

As pointed out by Sanchez and Kowalski,[9] chemistry is perhaps more suitable for trilinear components analysis than many other branches of science owing to the abundance of instruments that

© 1998 John Wiley & Sons, Ltd.

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,[2] the subject is never static; there is still much to be learned about the theory of three-way data arrays similar to matrix-based chemometrics methods.

Application of traditional PARAFAC and ATLD algorithms to HPLC–DAD data

As mentioned in the 'Experimental' section, three sets of real experimental data were used to verify the performance of both the traditional PARAFAC algorithm and the ATLD algorithm. Their data arrays had the dimensions $2\times64\times26$, $3\times64\times26$ and $7\times64\times26$ respectively. The estimated numbers of components in these three-way data arrays are given in Table 10.

Figures 9 and 10 show the results obtained from the first set of HPLC–DAD experimental data using the traditional PARAFAC and ATLD algorithms respectively. Although the loss functions decrease monotonically, the three-factor PARAFAC solution fails to produce an accurate concentration prediction. In the case of ATLD the correct spectrum and elution profile as well as the correct predicted concentration corresponding to *p*-chlorotoluene in sample #7 were obtained even in the presence of two interferents, *o*-dichlorobenzene and *o*-chlorotoluene. The predicted concentration of *p*-chlorotoluene in sample #7 was 25·45 μg ml^{-1} and its recovery was 101·0%.

Table 10. Determination of number of components in HPLC–DAD data[a]

Component no.	First singular value (percentage variance (%))		
	$\mathbf{X}_p^I$	$\mathbf{X}_p^J$	$\mathbf{X}_p^K$
	#1 as reference and #7 as unknown sample		
1	8152·1 (86·55)	8632·4 (907·04)	8155·2 (86·61)
2	3214·3 (13·45)	1383·2 (2·49)	3154·2 (12·96)
3		585·2 (0·45)	560·1 (10·41)
4		115·1 (0·02)	127·8 (0·02)
5		13·4 (0·00)	36·3 (0·00)
6		6·4 (0·00)	17·4 (0·00)
	N=3 for PARAFAC; *N*=4 for ATLD		
	#1 as reference and #5, #6 as unknown samples		
1	10 364 (67·10)	12 528 (98·04)	10 381 (67·32)
2	6504 (26·43)	1566 (1·51)	6514 (26·51)
3	3219 (6·47)	706 (0·31)	3116 (6·06)
4		420 (0·11)	355 (0·08)
5		65 (0·00)	223 (0·03)
6		17 (0·00)	65 (0·00)
	N=4 for PARAFAC; *N*=4 or 5 for ATLD		
	#1–#4 as reference and #7–#9 as unknown samples		
1	14 574 (83·93)	15 724 (97·71)	14 528 (83·41)
2	5548 (12·16)	2237 (1·98)	5680 (12·75)
3	3119 (3·84)	823 (0·27)	3093 (3·78)
4	338 (0·05)	344 (0·05)	255 (0·03)
5	200 (0·02)	45 (0·00)	196 (0·02)
6	9 (0·00)	14 (0·00)	153 (0·01)
	N=4 for PARAFAC; *N*=4 or 5 for ATLD		

[a] See footnote to Table 6.

© 1998 John Wiley & Sons, Ltd.

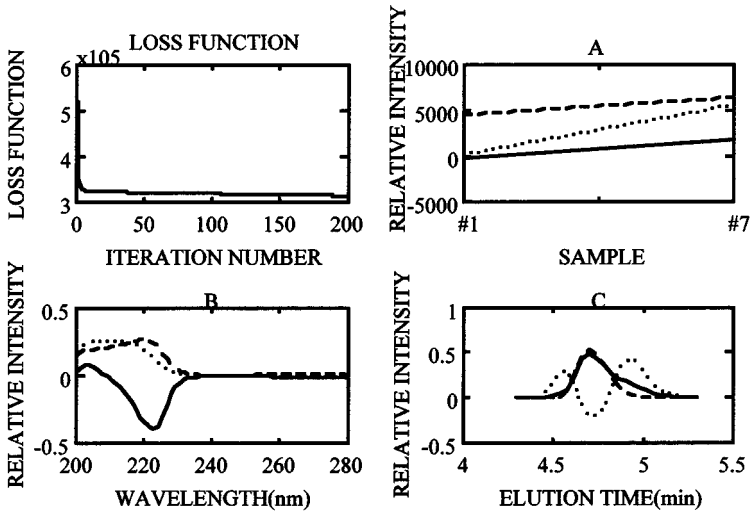


Figure 9. Results obtained for first set of HPLC–DAD experimental data using traditional PARAFAC algorithm; N=3. The predicted concentration of *p*-chlorotoluene (– – –) in unknown sample #7 is $6411 \times 75{\cdot}6/4643 = 104{\cdot}4\,\mu g\ ml^{-1}$ and its recovery is $104{\cdot}4/25{\cdot}2 = 414\%$. M=200

For the second data set the concentrations of two interferents, *o*-dichlorobenzene and *o*-chlorotoluene, in the unknown samples #5 and #6 are about ten times greater than the concentrations of the component of interest, *p*-chlorotoluene. The traditional PARAFAC algorithm mostly did not give satisfactory concentration predictions (see Figure 11). Using ATLD, the concentrations of *p*-chlorotoluene predicted for samples #5 and #6 were 11·32 and 15·62 $\mu g\ ml^{-1}$ and the recoveries were

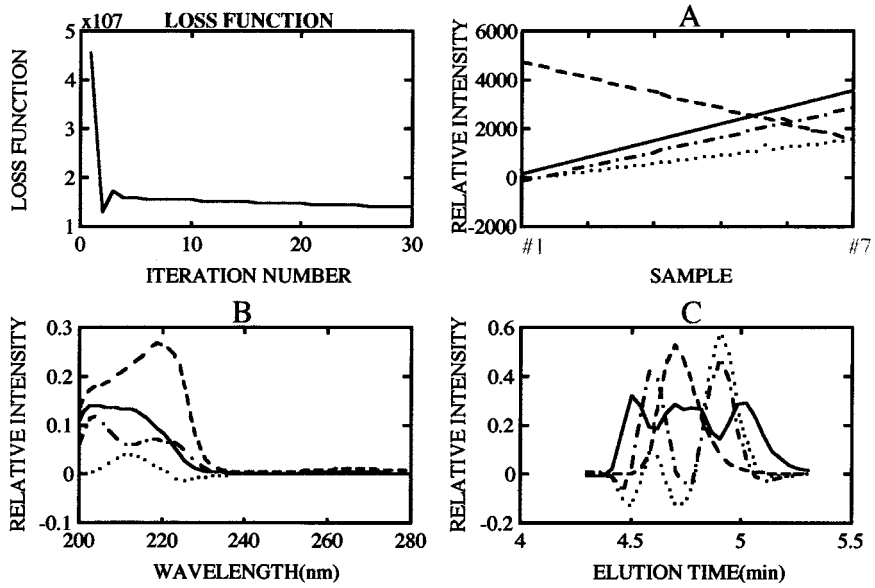


Figure 10. Results obtained for first set of HPLC–DAD experimental data using ATLD algorithm; N=4. The predicted concentration of *p*-chlorotoluene (– – –) in unknown sample #7 is $1590{\cdot}9 \times 75{\cdot}6/4725{\cdot}9 = 25{\cdot}45\,\mu g\ ml^{-1}$ and its recovery is $25{\cdot}45/25{\cdot}2 = 101{\cdot}0\%$. M=30

© 1998 John Wiley & Sons, Ltd.

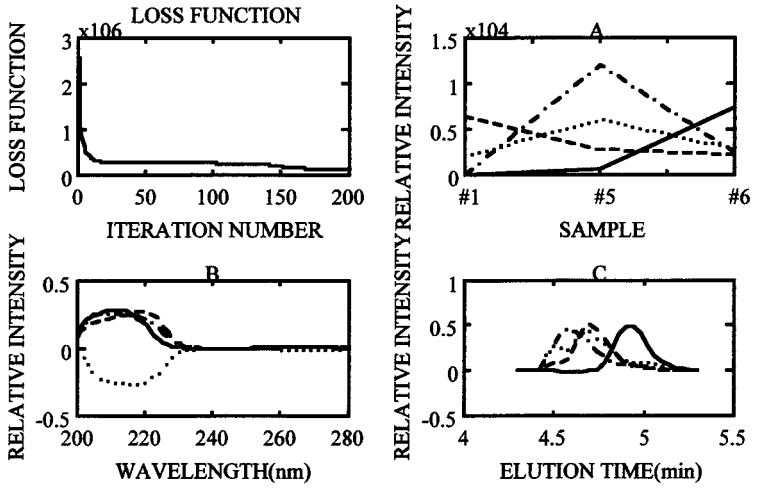


Figure 11. Results obtained for second set of HPLC–DAD experimental data using traditional PARAFAC algorithm; N=4. The predicted concentrations of *p*-chlorotoluene (– – –) in samples #5 and #6 are $2814\times75{\cdot}6/6434=33{\cdot}06\,\mu g\,ml^{-1}$ and $2268\times75{\cdot}6/6434=26{\cdot}65\,\mu g\,ml^{-1}$ and their recoveries are 33·06/12·6=262% and 26·65/12·6=211% respectively.M=200

89·7% and 123·9% respectively (see Figure 12).

For the third data set, *p*-chlorotoluene and *o*-chlorotoluene were chosen as analytes and *o*-dichlorobenzene as the unknown interferent. A total of 20 repeated calculations based on random starting values for each method were done. In the case of four-factor PARAFAC, four calculations

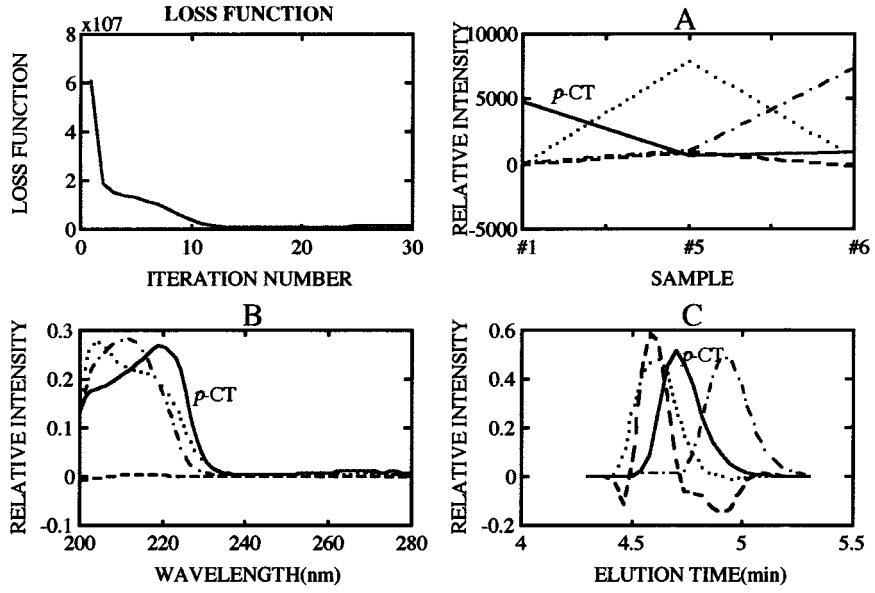


Figure 12. Results obtained for second set of HPLC–DAD experimental data using ATLD algorithm; N=4. The predicted concentrations of *p*-chlorotoluene (——) in samples #5 nd #6 are $709{\cdot}3\times75{\cdot}6/4736{\cdot}0=11{\cdot}32\,\mu g\,ml^{-1}$ and $978{\cdot}3\times75{\cdot}6/4736{\cdot}0=15{\cdot}62\,\mu g\,ml^{-1}$ and their recoveries are 11·32/12·6=89·7% and 15·62/12·6=123·9% respectively. M=30

© 1998 John Wiley & Sons, Ltd.

Table 11. Estimated concentrations and recoveries of *p*-chlorotoluene and *o*-chlorotoluene in third set of HPLC–DAD data in presence of interferent, *o*-dichlorobenzene, using PARAFAC and ATLD algorithms[a]

	p-Chlorotoluene (μg ml^{-1})					*o*-Chlorotoluene (μg ml^{-1})				
		Four-factor PARAFAC		Four-factor ATLD			Four-factor PARAFAC		Four-factor ATLD	
Unknown sample	Known	Predicted	Recovery (%)	Predicted	Recovery (%)	Known	Predicted	Recovery (%)	Predicted	Recovery (%)
#7	25·2	24·2±3·9	95·8±15·5	25·2±0·2	100·0±0·7	91·2	88·1±18·2	108·7±19·9	90·8±0·9	99·6±1·0
#8	50·4	54·9±2·8	109·3±5·5	53·7±0·5	106·6±1·0	30·4	30·7±12·0	101·1±39·5	29·5±0·8	97·2±2·7
#9	75·6	75·3±6·1	99·6±8·1	78·2±0·1	103·4±0·1	60·8	67·7± 9·9	111·3±16·3	64·7±0·9	106·4±1·4

[a] A total of 20 repeated calculations were done with different random starting values. M=200 for PARAFAC and M=30 for ATLD. In the case of PARAFAC the averages and standard deviations of only 16 repeated calculations are shown. The other four calculations were deleted owing to unreasonable estimates.

produced unreasonable predictions and the other 16 calculations are summarized in Table 11. Twenty predictions using ATLD are also summarized in Table 11. It was found that the ATLD algorithm was clearly superior to the traditional PARAFAC algorithm.

CONCLUSIONS

The alternating trilinear decomposition (ATLD) algorithm has been proposed as an alternative algorithm for decomposition of three-way data arrays. It is based on an alternating least squares principle and an iterative procedure that uses Moore–Penrose generalized inverse computations based on singular value decomposition, which should theoretically be more robust to similarities in spectra and time profiles. Its performance was compared with that of the traditional PARAFAC algorithm by a series of Monte Carlo simulations and by practical application to HPLC–DAD data. It was found that the ATLD algorithm has a capability to converge faster than the traditional PARAFAC algorithm. The ATLD-based second-order calibration retains the second-order advantage that calibration in the presence of unknown interferents can be performed to provide more satisfactory concentration predictions. In addition, a procedure for reducing the numerical rank of three-way trilinear data arrays has been proposed.

ACKNOWLEDGEMENTS

This work was supported by the Sasakawa Scientific Research Grant from The Japan Science Society. We wish to thank Professor Paul J. Gemperline, Assistant Professor Rasmus Bro and an anonymous reviewer who offered many very helpful suggestions and for their encouragement. We admire Professor Bro's capacity for having implemented correctly the ATLD method.

APPENDIX I: DEDUCING THE RANK OF A THREE-MODEL THREE-WAY TRILINEAR DATA ARRAY

Consider an analytical subject with an intrinsic physical or chemical rank N^*.

Deducing the rank of a matrix

Suppose the responses obtained from an analytical subject are arranged in a matrix $\mathbf{X}$ of size $I \times J$, e.g. a data matrix involving the absorbances of I different mixtures of N^* different absorbing components

© 1998 John Wiley & Sons, Ltd.

measured at j different wavelengths. Then the numerical rank, rank$\mathbf{X}$, of this bilinear matrix $\mathbf{X}$($\mathbf{X}=\mathbf{AB}^{\mathrm{T}}$, where $\mathbf{A}$ is a concentration matrix of size $I\times N^*$ and $\mathbf{B}$ is an extinction coefficient matrix of size $J\times N^*$) can be represented by

$$\mathrm{rank}\mathbf{X}=\max\{\mathrm{rank}(\mathbf{X}),\ \mathrm{rank}(\mathbf{X}^{\mathrm{T}})\}=\max\{\min(I,\ \min(J,\ N^*)),\ \min(J,\ \min(I,\ N^*))\} \tag{27}$$

Here RANK(.) calculates the rank of a matrix in the MATLAB environment based on singular value decomposition. Then we have the following.

1. If $I\geq N^*$ and $J\geq N^*$, then rank$\mathbf{X}=N^*$ and rank($\mathbf{A}$)=rank($\mathbf{B}$)=N^*.
2. If only one of I and J is equal to or larger than N^*, then rank$\mathbf{X}<N^*$ and max{rank($\mathbf{A}$), rank($\mathbf{B}$)}$<N^*$.
3. If $I<N^*$ and $J<N^*$, then rank$\mathbf{X}<N^*$.

In other words, if and only if both the number of different rows and the number of different columns of a matrix obtained from an analytical subject are equal to or larger than the intrinsic physical or chemical rank, then the numerical rank, rank$\mathbf{X}$, obtained by (27) can be considered equal to the intrinsic physical or chemical rank.

Deducing the rank of a three-model three-way trilinear data array

Similarly, suppose the responses obtained from an analytical subject are arranged in a three-way trilinear data array $\underline{\mathbf{X}}$ of size $I\times J\times K$ with an intrinsic physical or chemical rank N^*, e.g. an HPLC–DAD data array involving the absorbances of I different mixtures of N^* different absorbing components measured at J different wavelengths at K different time points. Then the numerical rank, rank$\underline{\mathbf{X}}$, of the three-way trilinear data array $\underline{\mathbf{X}}$ ($\underline{\mathbf{X}}=\mathbf{A}\,\underline{\mathbf{I}}\,\mathbf{B}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}$, where $\underline{\mathbf{I}}$ is a superdiagonal core array of size $N^*\times N^*\times N^*$ with ones on the superdiagonal and zeros elsewhere, $\mathbf{A}$ is a (relative) concentration matrix of size $I\times N^*$, $\mathbf{B}$ is a (relative) extinction coefficient matrix of size $J\times N^*$ and $\mathbf{C}$ is a (relative) chromatogram matrix of size $K\times N^*$) can be represented by

$$\begin{aligned}\mathrm{rank}\underline{\mathbf{X}}&=\max\{\mathrm{rank}(\mathbf{X}_{\mathrm{P}}^{I}),\ \mathrm{rank}(\mathbf{X}_{\mathrm{P}}^{J}),\ \mathrm{rank}(\mathbf{X}_{\mathrm{P}}^{K})\}\\&=\max\{\min(I,\ min(J\times K,\ N^*)),\ \min(J,\ min(K\times I,\ N^*)),\ \min(K,\ \min(I\times J,\ N^*))\}\end{aligned} \tag{28}$$

Here $\mathbf{X}_{\mathrm{P}}^{I}$, $\mathbf{X}_{\mathrm{P}}^{J}$ and $\mathbf{X}_{\mathrm{P}}^{K}$ are the same as in (18)–(20) in the text. Then we have the following:

1. If all I, J and K are equal to or larger than N^*, then rank$\mathbf{X}=N^*$.
2. If any two of I, J and K are equal to or larger than N^*, then rank$\mathbf{X}=N^*$. When e.g. K=1, see the numerical rank of a matrix.
3. If one of I, J and K and the product of the other two are equal to or larger than N^*, then rank$\underline{\mathbf{X}}=N^*$.
4. If one of I, J and K is equal to or larger than N^* but the product of the other two is less than N^*, then rank$\underline{\mathbf{X}}<N^*$.
5. If $I<N^*$, $J<N^*$ and $K<N^*$, then rank$\underline{\mathbf{X}}<N^*$.

In other words, provided that an obtained three-way trilinear data array $\underline{\mathbf{X}}$ satisfies the condition that one of I, J and K and the product of the other two are equal to or larger than the intrinsic physical or chemical rank N^*, then the numerical rank, rank$\underline{\mathbf{X}}$, obtained by (28) can be considered equal to the intrinsic physical or chemical rank.

Strictly speaking, the intrinsic physical or chemical rank N^* means that the N^* columns of the matrices $\mathbf{A}$ and $\mathbf{B}$ obtained from a bilinear matrix $\mathbf{X}$ as well as the N^* columns of the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ obtained from a three-way trilinear data array $\underline{\mathbf{X}}$ are independent. It also hides such a

© 1998 John Wiley & Sons, Ltd.

prerequisite that the selected I rows and J columns in the bilinear matrix $\mathbf{X}$ as well as the I rows and J columns and the K frontal slices in the three-way trilinear data array $\mathbf{X}$ need to be satisfied with a condition. Let us take the I rows as an example and similarly for the J columns and the K frontal slices. This condition is that (a) if $I<N^*$, then the I rows must be independent, while (b) if $I\geq N^*$, then at least if necessary N^* rows of the selected I rows are required to be independent. How to choose these rows and columns as well as slices in practical applications is related to optimal experimental design. Non-linearity and experimental errors in these response data may cause added complexity in deducing the rank of a three-way trilinear data array $\underline{\mathbf{X}}$.

APPENDIX II: ALGORITHM FOR ALTERNATING TRILINEAR DECOMPOSITION (ATLD)

Step 0

Estimate the number of factors, N, in the three-way data array $\underline{\mathbf{X}}$ of order $I\times J\times N$;

$$N\geq \text{rank}\underline{\mathbf{X}}=\max\{\text{rank}(\mathbf{X}_{\text{P}}^{I}),\ \text{rank}(\mathbf{X}_{\text{P}}^{J}),\ \text{rank}(\mathbf{X}_{\text{P}}^{K})\}$$

where $\text{rank}(\mathbf{X}_{\text{P}}^{I})$, $\text{rank}(\mathbf{X}_{\text{P}}^{J})$, $\text{rank}(\mathbf{X}_{\text{P}}^{K})$ are the numerical ranks of $\mathbf{X}_{\text{P}}^{I}$, $\mathbf{X}_{\text{P}}^{J}$ and $\mathbf{X}_{\text{P}}^{K}$. They may be obtained by using SVD of $\mathbf{X}_{\text{P}}^{I}$, $\mathbf{X}_{\text{P}}^{J}$ and $\mathbf{X}_{\text{P}}^{K}$ as

$$\begin{aligned}\mathbf{X}_{\text{P}}^{I}&=[\mathbf{X}_{\cdot\cdot 1}\ \mathbf{X}_{\cdot\cdot 2}\ \ldots\ \mathbf{X}_{\cdot\cdot K}]\\ \mathbf{X}_{\text{P}}^{J}&=[\mathbf{X}_{1\cdot\cdot}\ \mathbf{X}_{2\cdot\cdot}\ \ldots\ \mathbf{X}_{I\cdot\cdot}]\\ \mathbf{X}_{\text{P}}^{K}&=[\mathbf{X}_{\cdot 1\cdot}\ \mathbf{X}_{\cdot 2\cdot}\ \ldots\ \mathbf{X}_{\cdot J\cdot}]\end{aligned}$$

and then comparing the percentage variances of the first singular values respectively.

Step 1

Initialize the matrices $\mathbf{B}$ and $\mathbf{C}$ of orders $J\times N$ and $K\times N$ respectively by producing them randomly or by choosing the first singular vectors from Step 0 as one of the starting values for $\mathbf{B}$ and $\mathbf{C}$ respectively.

Step 2

Compute the starting values for $\mathbf{A}$ of order $I\times N$ according to the procedure

$$\text{P}\tilde{\mathbf{B}}=\text{pinv}(\mathbf{B}),\qquad \text{P}\tilde{\mathbf{C}}=\text{pinv}(\mathbf{C}),\qquad \mathbf{a}_i^{\text{T}}=\text{diag}(\text{P}\tilde{\mathbf{B}}\,\mathbf{X}_{i\cdot\cdot}\,\text{P}\tilde{\mathbf{C}}^{\text{T}}),\quad i=1,\ldots,I$$

where pinv(.) produces the Moore–Penrose pseudoinverse. The computation is based on SVD and any singular value less than a tolerance is treated as zero.

Step 3

Iteratively update $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{A}$ and evaluate the loss function

$$\sigma=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{ijk}-\sum_{n=1}^{N}a_{in}b_{jn}c_{kn}\right)^2$$

according to the following steps.

© 1998 John Wiley & Sons, Ltd.

(a) Compute **B**:

$$P\tilde{A} = \text{pinv}(\mathbf{A}), \qquad \mathbf{b}_j^T = \text{diag}(P\tilde{C}\,\mathbf{X}_{\cdot j \cdot} P\tilde{A}^T), \quad j=1, \ldots, J$$

(b) Compute **C**:

$$P\tilde{B} = \text{pinv}(\mathbf{B}), \qquad \mathbf{c}_k^T = \text{diag}(P\tilde{A}\,\mathbf{X}_{\cdot \cdot k}\, P\tilde{B}^T), \quad k=1, \ldots, K$$

(c) Normalize **B** and **C** to unit length columnwise respectively.

(d) Compute **A**:

$$P\tilde{C} = \text{pinv}(\mathbf{C}), \qquad P\tilde{B} = \text{pinv}(\mathbf{B}), \qquad \mathbf{a}_j^T = \text{diag}(P\tilde{B}\,\mathbf{X}_{i \cdot \cdot} P\tilde{C}^T), \quad i=1, \ldots, I$$

(e) Evaluate σ. If the one stopping criterion

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

where ε is some arbitrary small value, with a maximum of M iterative cycles is satisfied, then go to Step 4, otherwise repeat Step 3.

Step 4

Postprocess **A**, **B** and **C** according to the uniqueness or intrinsic axis property of trilinear decomposition.

Step 5

Determine the column position of each component of interest in **A** and then obtain its estimated concentration(s) in the unknown sample(s) by a plot or regression of the relative contribution(s) corresponding to each component in **A** against its standard concentration(s).

The ATLD algorithm has been implemented in MATLAB and is available from the authors.

REFERENCES

1. T. Hirschfeld, *Anal. Chem.* **52**, 297A (1980).
2. P. Geladi, *Chemometrics Intell. Lab. Syst.* **7**, 11 (1989).
3. A. K. Smilde, *Chemometrics Intell. Lab.Syst.* **15**, 143 (1992).
4. K. S. Booksh and B. R. Kowalski, *Anal. Chem.* **66**, 782A (1994).
5. Y. D. Wang, O. S. Borgen and B. R. Kowalski, *J. Chemometrics*, **7**, 439 (1993).
6. A. K. Smilde, Y. Wang and B. R. Kowalski, *J. Chemometrics*, **8**, 21 (1994).
7. E. Sanchez and B. R. Kowalski, *Anal. Chem.* **58**, 496 (1986).
8. B. Wilson, E. Sanchez and B. R. Kowalski, *J. Chemometrics*, **3**, 493 (1989).
9. E. Sanchez and B. R. Kowalski, *J. Chemometrics*, **4**, 29 (1990).
10. K. S. Booksh, Z. Lin, Z. Wang and B. R. Kowalski, *Anal. Chem.* **66**, 2561 (1994).
11. D. S. Burdick, X. M. Tu, L. B. McGown and D. W. Millican, *J. Chemometrics*, **4**, 15 (1990).
12. S. Leurgans and R. T. Ross, *Stat. Sci.* **7**, 289 (1992).
13. M. Gui, S. C. Rutan and A. Agbodjan, *Anal. Chem.* **67**, 3293 (1995).
14. R. A. Harshman, *UCLA Working Papers in Phonet.* **16**, 1 (1970).
15. J. D. Carroll and J. Chang, *Psychometrika*, **35**, 283 (1970).
16. H. Law, C. Snyder Jr., J. Hattie and R. McDonald (eds), *Research Methods for Multimode Data Analysis*, Praeger, New York (1984).
17. W. P. Krijnen, *The Analysis of Three-Way Arrays by Constrained PARAFAC Methods*, DSWO Press, Leiden (1993).

© 1998 John Wiley & Sons, Ltd.

18. A. K. Smilde and D. A. Doornbos, *J. Chemometrics*, **6**, 11 (1992).
19. B. C. Mitchell and D. S. Burdick, *J. Chemometrics*, **8**, 155 (1994).
20. J. B. Kruskal, in *Multiway Data Analysis*, ed. by R. Coppi and S. Bolasco, pp. 7–18, Elsevier, Amsterdam (1989).
21. A. K. Smilde, R. Tauler, J. M. Henshaw, L. W. Burgess and B. R. Kowalski, *Anal. Chem.* **66**, 3345 (1994).
22. G. Golub and C. Van Loan, *Matrix Computations*, 2nd edn, Johns Hopkins University Press, Baltimore, MD (1989).
23. R. Penrose, *Proc. Camb. Philos. Soc.* **52**, 17 (1956).
24. J. M. F. Ten Berge, *Least Squares Optimization in Multivariate Analysis*, DSWO Press, Leiden (1993).
25. H.-L. Wu, K. Oguma and R.-Q. Yu, *Anal. Sci.* **10**, 875 (1994).
26. P. M. Kroonenberg, *Three-Mode Principal Component Analysis*, DSWO Press, Leiden (1983).
27. E. R. Malinowski, *Factor Analysis in Chemistry*, 2nd edn, Wiley, New York (1991).

© 1998 John Wiley & Sons, Ltd.