

Estimation of the chemical rank for the three-way data: a principal norm vector orthogonal projection approach

Xie Hong-Ping^{a,b}, Jiang Jian-Hui^a, Shen Guo-Li^a, Yu Ru-Qin^{a,*}

^a *Laboratory for Chemometrics and Chemical Sensing Technology, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, People's Republic of China*

^b *Department of Chemistry, Sichuan Teachers' College, Nanchong 637002, People's Republic of China*

Received 1 May 2001; received in revised form 4 June 2001; accepted 25 June 2001

Abstract

A new approach for estimating the chemical rank of the three-way array called the principal norm vector orthogonal projection method has been proposed. The method is based on the fact that the chemical rank of the three-way data array is equal to one of the column space of the unfolded matrix along the spectral or chromatographic mode. A vector with maximum Frobenius norm is selected among all the column vectors of the unfolded matrix as the principal norm vector (PNV). A transformation is conducted for the column vectors with an orthogonal projection matrix formulated by PNV. The mathematical rank of the column space of the residual matrix thus obtained should decrease by one. Such orthogonal projection is carried out repeatedly till the contribution of chemical species to the signal data is all deleted. At this time the decrease of the mathematical rank would equal that of the chemical rank, and the remaining residual subspace would entirely be due to the noise contribution. The chemical rank can be estimated easily by using an *F*-test. The method has been used successfully to the simulated HPLC-DAD type three-way data array and two real excitation–emission fluorescence data sets of amino acid mixtures and dye mixtures. The simulation with added relatively high level noise shows that the method is robust in resisting the heteroscedastic noise. The proposed algorithm is simple and easy to program with quite light computational burden. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Chemical rank; Three-way array; Principal norm vector; Orthogonal projection; Excitation–emission fluorescence data

1. Introduction

With the development of the high-order hyphenated analytical instrumentation, three-way data treatment has become an extensive research subject in analytical chemistry. This kind of the data contains more information than second-order ones with an outstanding advantage of the uniqueness of the trilinear decomposition.

Based on the general concept of singular value decomposition, the decomposition models of the three-way array may be divided into three main groups, e.g. Tucker3, PARAFAC and unfolded second-order data model. The first step in establishing a right model is the estimation of the component number of the model. The physical meaning of the component number might be different in different models. In PARAFAC model the component number is the chemical rank. A three-way array can be formulated by a number of triads, each being produced by a tensor product of three vectors (Burdick, 1995). According to Kruskal (1977), the minimum number of these triads which can describe rightly

* Corresponding author. Tel./fax: +86-731-8822-782.
E-mail address: rquyu@mail.hunu.edu.cn (Y. Ru-Qin).

the model is called the rank of the three-way array. Since a three-way array contains more information than the second-order data, its chemical rank has many differences from that of the second-order data. Among others, one such difference is that the maximal rank of the three-way array may be larger than the maximal dimension of the modes of the array (Kruskal, 1977), while the rank of the second-order matrix can only be less than or equal to the minimum dimension of the columns or rows. Therefore, the method of estimating the rank for a three-way array should be different from that (Malinowski, 1991) for the second-order data.

Bro (1997) has pointed out the importance of the research on the rank estimation methods for the trilinear data. He divided these methods into three groups, e.g. methods based on the split-half experiments, examining the residual variance and utilizing the field knowledge concerning the data being modeled.

More recently, Louwerse et al. (1999) have proposed two alternative generalizations of two-way cross-validation method for Tucker3 model, i.e. expectation maximization (EM) and leave-bar-out (LBO) approaches. These methods seem to work well for the given examples, though the authors pointed out that possibly there is no minimum for the predictive residuals in the higher level of noise and the EM version which outperformed the LBO requires rather long computation time.

We have proposed an alternative approach for estimating the chemical rank of the three-way array based on the orthogonal projection with the help of the vector of the maximum Frobenius norm, which is called the principal norm vector (PNV).

The proposed method seems to belong to the second group of Bro's classification (Bro, 1997), i.e. methods based on examining the residual variance, though here the residual refers to that obtained after the orthogonal projection, rather than the residual in the ordinary sense of a model fitness measure.

When unfolding the three-way array along the mode of spectral or chromatographic profiles, the column space of the unfolded matrix is the combination of the spectral or chromatographic mode spaces of all sample matrix slices of the three-way array. All the base vectors of this mode space of the three-way array are contained in the column space of the unfolded matrix, all information concerning the co-existing chemical species is also embedded in the column space, and the column space is a rank deficient space. As the orthogonal projection transformation of the column space of the unfolded matrix is carried out with the orthogonal projection matrix formulated by the PNV as defined above, a residual matrix of the unfolded matrix could be obtained. Such a transformation makes the column space decompose into two orthogonal complement subspaces, e.g. the PNV and the column space of the residual matrix. Therefore, the mathematical rank of

the column space of the residual matrix should decrease by one. After such transformations are carried out n cycles for the column space of unfolded matrix containing n chemical species, its mathematical rank should decrease by n . At same time, its chemical rank should also decrease by n , and its residual matrix would become a noise matrix, that is, the column space of the residual matrix would become a noise subspace. One could distinguish the chemical signal vector and the noise vector using the F -test for the residuals of the PNVs, and estimate the chemical rank of the three-way array.

2. Theory

2.1. Trilinear model

The chemical rank of a three-way array is based on the trilinear model. As the three-way data for an $(I \times J \times K)$ array \mathbf{R} , a trilinear model can be expressed (Bro, 1997) as

$$\mathbf{R}_{I \times J \times K} = \sum_{n=1}^N x_n \otimes y_n \otimes z_n + \mathbf{E}_{I \times J \times K} \quad (1)$$

where x_n , y_n and z_n are the response profiles of the n th response-active component along x , y and z axis, respectively; N is the component number; \otimes is the tensor product; \mathbf{E} is the measuring error array. The matrices of the response profiles are expressed as

$$\mathbf{X}_{I \times N} = (x_1, x_2, \dots, x_N) \quad (2)$$

$$\mathbf{Y}_{J \times N} = (y_1, y_2, \dots, y_N) \quad (3)$$

$$\mathbf{Z}_{K \times N} = (z_1, z_2, \dots, z_N) \quad (4)$$

The goal of the trilinear resolution is to resolve the response profiles $\mathbf{X}_{I \times N}$, $\mathbf{Y}_{J \times N}$ and $\mathbf{Z}_{K \times N}$ for obtaining the chemical information concerning the measured processes.

Eq. (1) might be expressed as three matrices along these axes:

$$\mathbf{R}_{\cdot k} = \mathbf{X} \text{diag}(z_{(k)}) \mathbf{Y}^T + \mathbf{E}_{\cdot k} \quad (k = 1, 2, \dots, K) \quad (5)$$

$$\mathbf{R}_{\cdot j} = \mathbf{Z} \text{diag}(y_{(j)}) \mathbf{X}^T + \mathbf{E}_{\cdot j} \quad (j = 1, 2, \dots, J) \quad (6)$$

$$\mathbf{R}_{\cdot i} = \mathbf{Y} \text{diag}(x_{(i)}) \mathbf{Z}^T + \mathbf{E}_{\cdot i} \quad (i = 1, 2, \dots, I) \quad (7)$$

where $\mathbf{R}_{\cdot k}$ is the k th matrix slice of $\mathbf{R}_{I \times J \times K}$ along the z axis, $\mathbf{R}_{\cdot j}$ the j th matrix slice along the y axis, $\mathbf{R}_{\cdot i}$ the i th matrix slice along the x axis, $\text{diag}(z_{(k)})$ is the diagonal matrix whose diagonal elements are the corresponding ones of the k th row vector $Z_{(k)}$ of the response profile matrix $\mathbf{Z}_{K \times N}$, $\text{diag}(y_{(j)})$ corresponds to the j th row vector $y_{(j)}$ of $\mathbf{Y}_{J \times N}$, $\text{diag}(x_{(i)})$ corresponds to the i th row vector $x_{(i)}$ of $\mathbf{X}_{I \times N}$. The superscript T denotes the matrix transposition. Trilinear resolutions are carried out for all components at the same time or one-by-

one for each component according to Eqs. (5)–(7). The proposed principal norm vector orthogonal projection (PNVOP) approach is based on the forementioned trilinear model.

2.2. The chemical rank of the three-way array and its unfolded matrices

The six different unfolded matrices (Westerhuis et al., 1999) can be given out by unfolding a three-way array \mathbf{R} ($I \times J \times K$) along the three modes I , J and K according to Eqs. (5), (7) and (6), respectively.

$$\mathbf{RA}_{I \times JK} = [\mathbf{R}_{\cdot 1}, \mathbf{R}_{\cdot 2}, \dots, \mathbf{R}_{\cdot K}] \quad (8)$$

$$\mathbf{RB}_{J \times IK} = [\mathbf{R}_{\cdot 1}^T, \mathbf{R}_{\cdot 2}^T, \dots, \mathbf{R}_{\cdot K}^T] \quad (9)$$

$$\mathbf{RC}_{J \times KI} = [\mathbf{R}_{1 \cdot}, \mathbf{R}_{2 \cdot}, \dots, \mathbf{R}_{J \cdot}] \quad (10)$$

$$\mathbf{RD}_{K \times JI} = [\mathbf{R}_{1 \cdot}^T, \mathbf{R}_{2 \cdot}^T, \dots, \mathbf{R}_{J \cdot}^T] \quad (11)$$

$$\mathbf{RE}_{K \times IJ} = [\mathbf{R}_{\cdot 1}, \mathbf{R}_{\cdot 2}, \dots, \mathbf{R}_{\cdot J}] \quad (12)$$

$$\mathbf{RF}_{I \times KJ} = [\mathbf{R}_{\cdot 1}^T, \mathbf{R}_{\cdot 2}^T, \dots, \mathbf{R}_{\cdot J}^T] \quad (13)$$

where the row space of \mathbf{RA} is identical with that of \mathbf{RF} , the vectors of their column spaces are linear combinations of each other with a common base set. In view of the linear space theory, the two unfolded matrices have the same column and row ranks. Similarly, the matrices \mathbf{RB} and \mathbf{RD} have identical ranks as \mathbf{RC} and \mathbf{RE} , respectively.

Taking Eq. (8) as an example, one can understand the relation between the chemical rank information contained in the three-way data and that contained in its unfolded matrix.

The column space of the unfolded matrix \mathbf{RA} is constructed by the combinations of all vectors of J -mode space of \mathbf{R} . The vectors involved in each sample matrix might be independent or correlative ones. The base vector number of the column space of each sample matrix is less than or equal to J . The column space for \mathbf{RA} is also the space of J -mode and is spanned by these base vectors. The number of the independent vectors of the column space of \mathbf{RA} must be less than or equal to J . Among the vectors involved in this set (there are altogether $K \times J$ vectors) there must be some correlative ones. Consequently,

$$\text{rank}(\mathbf{RA})_{\text{col}} < K \times J \quad (14)$$

where $\text{rank}(\mathbf{RA})_{\text{col}}$ is the column rank of \mathbf{RA} . Therefore, the column space of \mathbf{RA} is a rank deficient space.

Suppose that the chemical rank of a three-way data \mathbf{R} is n , that is,

$$\text{rank}(\mathbf{R}) = n \quad (15)$$

Because the column vectors of the \mathbf{RA} are the combinations of all vectors of J -mode spaces in K sample matrices of dimension ($I \times J$), each having a chemical

rank up to n , the column space of \mathbf{RA} must contain n chemical component vectors contained in K samples, e.g.

$$\text{Rank}(\mathbf{RA})_{\text{colchem}} = n \quad (16)$$

where $\text{Rank}(\mathbf{RA})_{\text{colchem}}$ is the chemical rank of the column space for \mathbf{RA} .

From Eqs. (15) and (16), one has

$$\text{rank}(\mathbf{R}) = \text{rank}(\mathbf{RA})_{\text{colchem}} = n. \quad (17)$$

Or in words the chemical rank of \mathbf{R} is equal to the one of the column space of \mathbf{RA} , that is, the chemical species information contained in \mathbf{R} should be the same as the column space of \mathbf{RA} . Therefore, we transform the estimation of the chemical rank of the three-way array into the estimation of the chemical rank of the unfolded matrix.

2.3. The principal norm vector orthogonal projection approach

Since the information of all the chemical species is included in the matrix formulated by unfolding the three-way array along a mode of spectral or chromatographic profiles, the chemical rank of the three-way array could be found if one could determine the rank of the formulated matrix. The unfolded matrix is different from an ordinary second-order data matrix. An ordinary second-order data matrix is constructed by all vectors of spectral or chromatographic mode of a sample, and the matrix unfolded with a three-way array is formulated by juxtaposing side by side K sample matrices along spectral or chromatographic mode. Species number and concentration of the species contained in each sample may be different for each sample matrix and the species number is less than or equal to the chemical rank n for each sample. Usually there exist some samples containing the same species with different relative concentrations, and the column vectors of these sample matrices must be strongly correlated. Therefore, the column space of the unfolded matrix must contain a number of regions carrying strongly correlated information.

Suppose that \mathbf{R}_u is an unfolded matrix of the three-way array \mathbf{R} with M column vectors along spectral or chromatographic mode,

$$\mathbf{R}_u(\mathbf{u}_1, \dots, \mathbf{u}_{\text{max}}, \dots, \mathbf{u}_M) \quad (18)$$

where \mathbf{u}_{max} is PNV or the column vector with the maximum Frobenius norm in the column space. The vector must be the spectral or chromatographic vector with the most chemical components involved or the highest concentration among all samples. This vector is actually the most information-rich vector among all column ones involved in the unfolded matrix.

The vector \mathbf{u}_{\max} in Eq. (18) is used to construct an orthogonal projection matrix \mathbf{P} ,

$$\mathbf{P} = \mathbf{I} - \mathbf{u}_{\max} \mathbf{u}_{\max}^T \quad (19)$$

where \mathbf{I} is an identity matrix. When the column vectors of \mathbf{R}_u project along \mathbf{u}_{\max} with \mathbf{P} , the residual matrix \mathbf{R}_s formulated will be expressed as

$$\begin{aligned} \mathbf{R}_s &= \mathbf{P}\mathbf{R}_u = \mathbf{P}(\mathbf{u}_1, \dots, \mathbf{u}_{\max}, \dots, \mathbf{u}_M) \\ &= (\mathbf{P}\mathbf{u}_1, \dots, \mathbf{P}\mathbf{u}_{\max}, \dots, \mathbf{P}\mathbf{u}_M) \end{aligned} \quad (20)$$

Since the matrix \mathbf{P} in Eq. (19) is an orthogonal projection matrix constructed by \mathbf{u}_{\max} , the column space of the unfolded matrix produces the orthogonal decomposition. There are two orthogonal complement subspaces, e.g. the column space of \mathbf{R}_s is the orthogonal complement subspace of \mathbf{u}_{\max} . One has

$$\begin{aligned} \mathbf{R}_s &= (\mathbf{P}\mathbf{u}_1, \dots, \mathbf{P}\mathbf{u}_{\max-1}, 0, \mathbf{P}\mathbf{u}_{\max+1}, \dots, \mathbf{P}\mathbf{u}_M) \\ &= (\mathbf{u}'_1, \dots, \mathbf{u}'_{\max-1}, 0, \mathbf{u}'_{\max+1}, \dots, \mathbf{u}'_M) \end{aligned} \quad (21)$$

The mathematical rank of the column space of the unfolded matrix is

$$\text{rank}(\mathbf{R}_u)_{\text{col}} = \text{rank}(\mathbf{R}_s)_{\text{col}} + \text{rank}(\mathbf{u}_{\max}) \quad (22)$$

where $\text{rank}(\mathbf{R}_u)_{\text{col}}$ and $\text{rank}(\mathbf{R}_s)_{\text{col}}$ are the mathematical ranks of the column spaces of \mathbf{R}_u and \mathbf{R}_s , respectively. Since \mathbf{u}_{\max} is a vector, one has

$$\text{rank}(\mathbf{u}_{\max}) = 1 \quad (23)$$

From Eqs. (22) and (23), one obtains

$$\text{rank}(\mathbf{R}_s)_{\text{col}} = \text{rank}(\mathbf{R}_u)_{\text{col}} - 1 \quad (24)$$

Eq. (24) shows that the orthogonal projection transformation of the column space of the unfolded matrix of the three-way array along the PNV makes the mathematical rank of its residual matrix decrease by one. Taking the residual matrix \mathbf{R}_s as the current unfolded matrix \mathbf{R}_u , one can do the same as above, but the Frobenius norm of the current PNV \mathbf{u}_{\max} should be smaller than that of the PNV \mathbf{u}_{\max} of the previous cycle, though the current PNV is again the most information-rich one among all column vectors of the current unfolded matrix \mathbf{R}_u . After these kinds of the transformations being carried out n cycles for the unfolded matrix containing n chemical species, the mathematical rank of its residual matrix would decrease by n . Since these n PNVs are all most information-rich vectors in the current cycle of transformation, the column space of the residual matrix becomes a noise subspace. At this time, its chemical rank decreases by n , too. Due to the random character of the noise distribution, the decreasing rate of Frobenius norm of the noise PNV with the increase of the projection cycles would be substantially reduced. An F -test is carried out for the residuals of the PNVs of two successive cycles. When in these two successive cycles the PNVs are all contributed by the

noise, the F values would be in the rejection domain. In this way one can distinguish the PNV contributed by chemical signal and that by the noise, and the chemical rank of the three-way array can be estimated.

3. The algorithm

1. Unfold the three-way array \mathbf{R} along the mode of spectral or chromatographic profiles to formulate the unfolded matrix \mathbf{R}_u .
2. Select the column vector \mathbf{u}_{\max} of the maximum Frobenius norm or the PNV \mathbf{u}_{\max} for the column space of \mathbf{R}_u .
3. Construct an orthogonal projection matrix \mathbf{P} with \mathbf{u}_{\max} according to Eq. (19).
4. Project all the column vectors of \mathbf{R}_u along \mathbf{u}_{\max} with \mathbf{P} to obtain all the column vectors of the residual matrix \mathbf{R}_s according to Eq. (20). Record the residual sum of squares (rss) for \mathbf{u}_{\max} . Make \mathbf{R}_s equal the current \mathbf{R}_u .
5. The F -test is carried out for rss, that is,

$$F_i = \text{rss}(i) / \text{rss}(i+1) \quad (25)$$

The F_i value is compared with the presented value of threshold. The algorithm is returned to Step 2.

6. The chemical rank of unfolded matrix should be less than or equal to the number of rows for the matrix. When the number of cycles is equal to the number of rows of the unfolded matrix, the calculation is terminated.

There is an alternative way to locate the point of terminating the calculation process, which can lighten the computation burden. When five or more successive cycles give F_i values below the threshold, the calculation is terminated. The cycle number before these five (or more) successive cycles is taken as the chemical rank of the three-way data array.

4. The experiments

The proposed method was tested using three data arrays, including one simulated HPLC-DAD type data set with relatively high noise added and two experimentally recorded excitation–emission fluorescence data sets.

4.1. The simulated data

The HPLC-DAD type data of ten samples consisting of four components have been simulated. The pure spectra of the four components are simulated according to the following expressions:

Table 1
The component concentrations of the amino acid mixtures (mol l⁻¹)

Sample	#1	#2	#3	#4	#5	#6	#7
Tryptophan	0	9.6×10^{-5}	0	9.6×10^{-5}	9.6×10^{-3}	1.92×10^{-4}	1.92×10^{-4}
Tyrosine	0	0	4.4×10^{-4}	4.4×10^{-4}	8.8×10^{-4}	4.4×10^{-4}	8.8×10^{-4}
Phenylalanine	4.4×10^{-3}	2.2×10^{-3}	2.2×10^{-3}	2.2×10^{-3}	2.2×10^{-3}	2.2×10^{-3}	4.4×10^{-3}

$$s_1 = 0.2 \text{ gs}(4i - 3, 30, 30) + 0.5 \text{ gs}(4i - 3, 70, 10)$$

$$(i = 1, 2, \dots, 50)$$

$$s_2 = 0.5 \text{ gs}(4i - 3, 20, 10) + 0.2 \text{ gs}(4i - 3, 70, 10)$$

$$(i = 1, 2, \dots, 50)$$

$$s_3 = 0.3 \text{ gs}(4i - 3, 40, 10) + 0.4 \text{ gs}(4i - 3, 90, 20)$$

$$(i = 1, 2, \dots, 50)$$

$$s_4 = 0.7 \text{ gs}(4i - 3, 50, 25) \quad (i = 1, 2, \dots, 50)$$

where $\text{gs}(x, m, n) = \exp[-(x - m)^2 / (2n^2)]$. The pure chromatograms of the four components are simulated as

$$c_1 = 0.5 \text{ gs}(4i - 3, 40, 5) \quad (i = 1, 2, \dots, 20)$$

$$c_2 = 0.5 \text{ gs}(4i - 3, 30, 10) \quad (i = 1, 2, \dots, 20)$$

$$c_3 = 0.5 \text{ gs}(4i - 3, 50, 10) \quad (i = 1, 2, \dots, 20)$$

$$c_4 = 0.5 \text{ gs}(4i - 3, 20, 9) \quad (i = 1, 2, \dots, 20)$$

The concentrations of the components c_j ($j = 1, \dots, 4$) are taken as random numbers of uniform distribution in the region [0,1]. Five samples contain all the four components, and the remaining five contain only three components c_1 , c_2 and c_3 . The matrices of spectral mode \mathbf{X} , the chromatographic mode \mathbf{Y} and the concentration mode \mathbf{Z} are formulated according to Eqs. (2)–(4), respectively. The noise added includes the homoscedastic one which is simulated by random numbers of normal distribution with zero mean and a standard deviation of 0.005, and the heteroscedastic noise with a relatively high intensity of 1.0% of the signal. The response metrics $\mathbf{R}_{\cdot k}$ ($k = 1, 2, \dots, 10$) of the samples are formulated according to Eq. (5) and a $(50 \times 20 \times 10)$ three-way data array is obtained.

4.2. The fluorescence excitation–emission spectra of amino acid and dye mixtures

Seven mixture samples of tyrosine, tryptophan and phenylalanine are prepared in phosphate buffer of pH 7.2 with 0.0028 M KH_2PO_4 (Table 1). The fluorescence dyes co-existing in the liquid laser, i.e. acridine, fluorescein and rhodamine B, are taken to prepare six samples (Table 2).

The fluorescence spectra were recorded using a Hitachi 850 fluorescence spectrophotometer with a wave-

length scan speed of 240 nm/min and wavelength intervals of 5 nm. The range of excitation wavelength for the amino acid mixtures was 205–280 nm, while that for the dye mixtures was 450–600 nm. The ranges of emission wavelength for the amino acid and dye samples were 270–385 and 480–620 nm, respectively. The effect of Rayleigh scattering was corrected by background subtraction using a blank sample. For the amino acid samples, a $(16 \times 24 \times 7)$ data array was obtained, while the dimension of the dye mixture data array was $(31 \times 29 \times 6)$.

5. Results and discussion

5.1. The simulated data array

The estimation of the chemical rank of the $(50 \times 20 \times 10)$ three-way array is first solved by estimating the chemical rank of the column space of the $I \times JK$ (50×200) unfolded matrix formulated by unfolding the original three-way array along the spectral mode. The F -test with significance level of 0.005 is shown in Fig. 1. One notices from Fig. 1 that starting from the fifth cycle all F values are in the rejection domain and their variation rate is much smaller than that for the cycles 1–4. It is evident that in the first four cycles the orthogonal projection deletes four chemical components contributing the spectral signal while the remaining cycles deal with the noise terms. Therefore, the chemical rank of the data should be 4. Fig. 2 shows the plot of the rss. One observes that the rss decreases rapidly for the first four cycles due to the successive deletion of chemical species contributed spectral information, while the decreasing rate obviously slows down after the fourth cycle.

Table 2

The component concentrations of the fluorescence dye samples (10^{-3}g l^{-1})

Sample	#1	#2	#3	#4	#5	#6
Acridine	0.00	0.00	0.00	0.00	0.24	0.12
Fluorescein	0.12	0.00	0.12	0.24	0.12	0.24
Rhodamine B	0.00	0.11	0.22	0.11	0.22	0.22

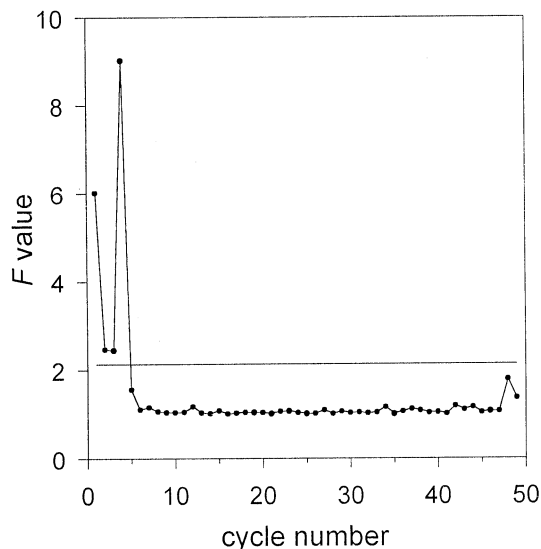


Fig. 1. F -test for the simulated system along the spectral mode.

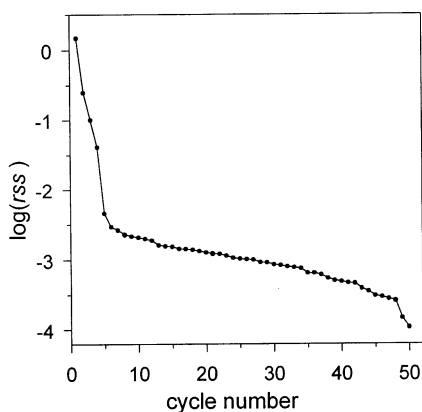


Fig. 2. The rss for the simulated system along the spectral mode.

In order to validate the conclusion obtained along the spectral mode, we have unfolded the three-way array along the chromatographic mode according to Eq. (9) to obtain a $J \times IK$ (20×500) unfolded matrix. The chemical rank can also be estimated for this matrix using the PNVOP method. The result is shown in Fig. 3, and the corresponding plot for logarithms of the rss is shown in Fig. 4. From Fig. 3 one can identify the chemical component number of 4. The variation of the F values after the fourth cycle is slightly greater as compared to Fig. 1. Comparing Fig. 4 with Fig. 2, one notices that the rss for the chromatographic mode decreases faster than that of the spectral mode for the cycles associated with noise contribution. These phe-

nomena are evidently due to the heteroscedastic noise of relatively high intensity introduced in the chromatographic mode during the simulation. These results indicate at the same time that the proposed PNVOP method is robust in resisting the effect of heteroscedastic noise. In Fig. 3 the F values obtained for the first two cycles fall into the rejection domain. The rss values for the first two cycles are close to each other. That is why in step 6 of the proposed algorithm an alternative approach is used to terminate the calculation only when five or more successive cycles give F values below the limit threshold to avoid random falling of the F into the rejection domain. This requirement might be circumvented by using a significance level higher than 0.005.

5.2. The fluorescence data of the amino acid mixtures

Since the three-way fluorescence data set has two spectral modes, that is, the excitation and the emission ones, the chemical ranks of the unfolded matrices formulated by unfolding the three-way array along these

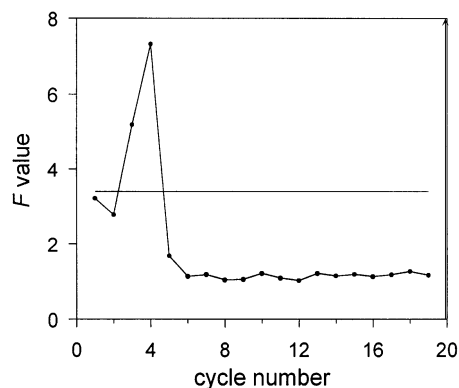


Fig. 3. F -test for the simulated system along the chromatographic mode.

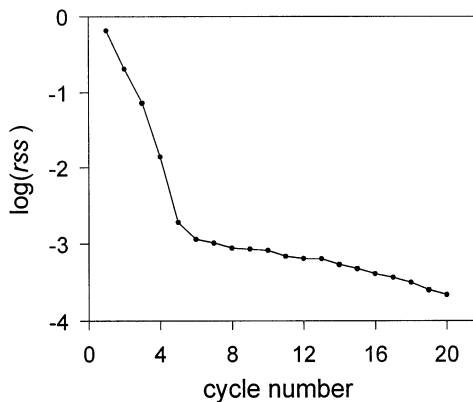


Fig. 4. The rss for the simulated system along the chromatographic mode.

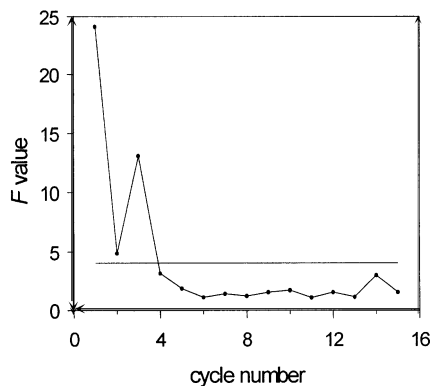


Fig. 5. *F*-test for the amino acid system along the excitation spectral mode.

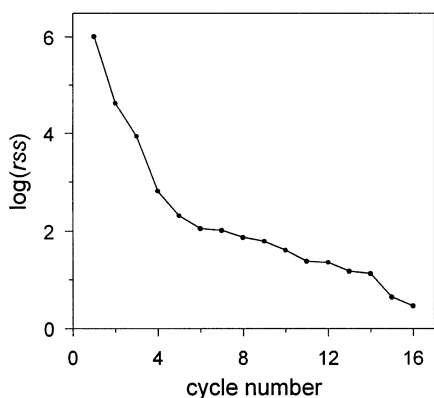


Fig. 6. The rss for the amino acid system along the excitation spectral mode.

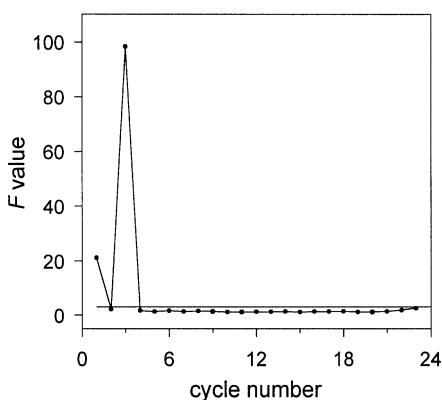


Fig. 7. *F*-test for the amino acid system along the emission spectral mode.

two modes must be equal to each other, and they represent the number of the chemical species involved.

A (16×168) unfolded matrix was obtained by un-

folding the $(16 \times 24 \times 7)$ three-way array of amino acid mixtures along the excitation spectral mode according to Eq. (8). The chemical rank has been estimated using the PNVOP method with the results shown in Figs. 5

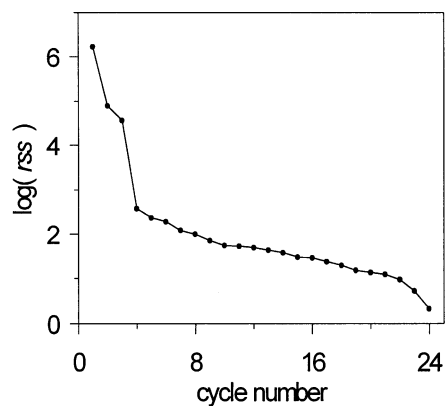


Fig. 8. The rss for the amino acid system along the emission spectral mode.

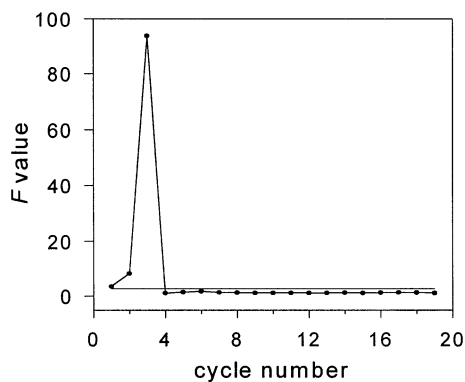


Fig. 9. *F*-test for the fluorescence dye system along the emission spectral mode.

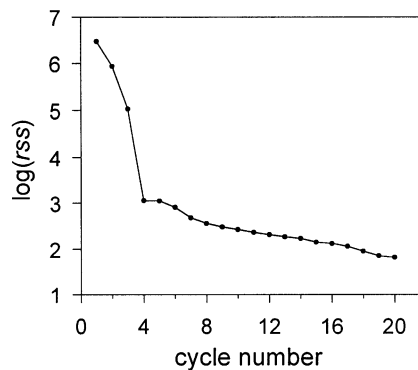


Fig. 10. The rss for the fluorescence dye system along the emission spectral mode.

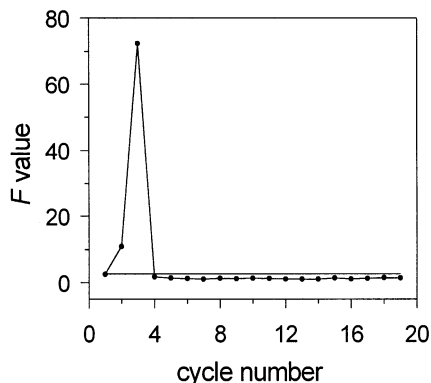


Fig. 11. *F*-test for the fluorescence dye system along the excitation spectral mode.

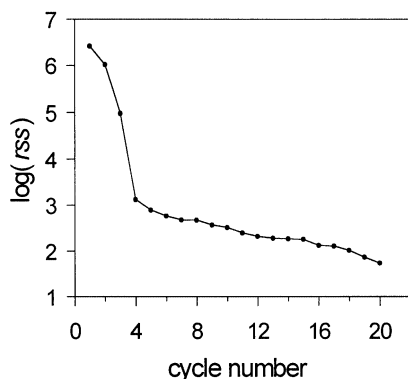


Fig. 12. The rss for the fluorescence dye system along the excitation spectral mode.

and 6. One can quite easily estimate the chemical rank of three for this system. The results calculated for the emission spectral mode are shown in Figs. 7 and 8. A comparison of Figs. 6 and 8 indicates that for the excitation spectral model a relatively high heteroscedastic noise is existing, though this does not affect the application of the proposed method for obtaining the correct results.

5.3. The fluorescence data for the dye mixtures

Unfolding the $(31 \times 29 \times 6)$ three-way data array of dye mixtures along the emission spectral mode gives a

(29×186) matrix and along the excitation spectral mode a (31×174) matrix.

The PNVOP treatment results are shown in Figs. 9–12. The correct chemical rank of three is obtained from these figures.

6. Conclusions

The column space of the unfolded matrix formulated by unfolding the three-way array along its spectral or chromatographic mode is a rank deficient one, and the chemical rank of this column space is equal to that of the three-way array. The orthogonal projection of the column vectors of the unfolded matrix performed with the aid of PNV would decompose the column space into two orthogonal complement subspaces. The mathematical rank of the residual subspace would decrease by one. By using the proposed algorithm the orthogonal projection transformation is repeated until all the information contributed by the chemical species is deleted. An *F*-test can easily locate this point when the decrease of the mathematical rank by the projection is equal to that of chemical rank. An outstanding feature of the proposed method is the robustness toward the heteroscedastic noise which might cause problems when one uses the traditional cross-validation type methodology. The algorithm can easily be programmed, which runs very fast.

Acknowledgements

This research was supported by grants from the National Natural Science Foundation of China (Grant Nos. 29735150 and 20075006).

References

- Burdick, D.S., 1995. Chemom. Intell. Lab. Syst. 28, 229–237.
- Kruskal, J.B., 1977. Linear Algebra Appl. 18, 95–138.
- Malinowski, E.R., 1991. Factor Analysis. Wiley–Interscience, New York.
- Bro, R., 1997. Chemom. Intell. Lab. Syst. 38, 149–171.
- Louwerse, D.J., Smilde, A.K., Kiers, H.A.L., 1999. J. Chemom. 13, 491–510.
- Westerhuis, J.A., Kourti, T., Macgregor, J.F., 1999. J. Chemom. 13, 379–413.