Contents lists available at ScienceDirect



Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab



Determining the number of principal factors by eigenvector comparison of the original bi-linear data matrix and the one reconstructed from key variables

CrossMark

Shi-Ling Hao, Li-Min Shao *

Department of Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China

ARTICLE INFO

Article history: Received 9 July 2015 Received in revised form 26 October 2015 Accepted 27 October 2015 Available online 4 November 2015

Keywords: Principal factors Key variables Orthogonal projection approach (OPA) Window factor analysis (WFA)

ABSTRACT

It is an essential step in analyzing hyphenated chromatographic data of complex chemical systems to determine the number of principal factors of the bi-linear matrix. The determination is difficult due to the co-existence of non-chemical factors, such as background, noise, *etc.* A new method was proposed for the determination based on comparing eigenvectors of the original data matrix and the one reconstructed from key spectral variables that are selected with orthogonal projection approach (OPA). The proposed method is mathematically rigorous and the determination is clear. In comparison with other four indices, i.e., NPFPCA (noise perturbation in functional principal component analysis), RESO (the ratio of eigenvalues calculated by smoothed principal component analysis and those calculated by ordinary principal component analysis), DRAUG (determination of rank by augmentation) and DRMAD (determination of rank by median absolute deviation), this proposed method was proven to have good performance in both simulated GC-IR and experimental HPLC-DAD data.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Measurements from hyphenated techniques, such as GC-MS, are usually in the form of matrix. Such data matrices are bi-linear, which means columns and rows are linear combinations of pure elution profiles and spectra, respectively. Bi-linear matrix features rich chemical information and significant interferences and redundancy, and it is important to extract as much chemical information as possible for subsequent analysis.

Variables that mostly contain chemical information are referred to as principal factors. In a mathematical perspective, principal factors are linearly independent variables that span the same space as the original variables do [1]. There is a similar concept in regression, which is latent variable. Latent variables are those that build a model of regression with the best fit to calibration data [1,2]. Principal factors and latent variables are similar, and yet intrinsically different.

Principal factors play an important role in data analysis of bi-linear matrices, so it is an essential step in many multivariate analysis methods to determine the number of principal factors, also known as "chemical rank" of the bi-linear matrix [3]. In the ideal situation where each chemical component makes independent and noise-free contribution to the data matrix, the number of principal components equals the number of chemical components. However, the practical determination is

* Corresponding author. *E-mail address:* lshao@ustc.edu.cn (L-M. Shao). difficult due to the co-existence of instrumental factors and experimental noise [4,5], such as unresolved background, different types of noise, overlapping signals, *etc.* Since the number of principal factors is important information in many multivariate analysis methods, an incorrect estimation would lead to erroneous qualitative and quantitative results [3].

Several methods have been proposed to determine the number of principal factors, and they could be classified into three categories: (1) empirical, (2) mathematically rigorous and (3) statistical [6]. In this paper, a novel method, which is in the second category, was proposed to determine the number of principal factors in hyphenated chromatographic data. First, the orthogonal projection approach (OPA) [7–10] is used to select key variables, which are a set of either rows (spectra) or columns (elution profiles) of the data matrix; the contribution of each is from one dominant chemical component [4]. Second, the method of least squares (LS) [11] is applied to obtain pure elution and spectral estimates, and subsequently these estimates are used to reconstruct a purified data matrix. Finally, by comparing eigenvectors obtained from both the original and the reconstructed matrices, the number of principal factors is determined.

The proposed method, referred to as OPALS, employs singular value decomposition (SVD) to obtain eigenvectors from both the original and the reconstructed matrices, respectively [1,12]. According to the theory proposed by Lu *et al.* [6], OPALS follows the second pattern in mathematically rigorous category.

Both simulated GC-IR and experimental HPLC-DAD data were used to evaluate OPALS. In evaluations with simulated data, the results were

compared to the number of chemical components in the simulation. In the case of experimental data, the results may account not only for chemical information but also for instrumental aberrations [2], such as baseline drifts, pulsations, nonlinearity, *etc.*, if their influence is higher than the expected signal-to-noise ratio. Therefore, principal factors usually outnumber chemical components for experimental data, which requires additional measures to evaluate the obtained number of principal factors. In this paper, window factor analysis (WFA) was employed for the evaluation because only with correct number of principal factors could WFA yield meaningful results [13].

2. Theory

Consider a two-way data matrix $X_{m \times n}$ of a multi-component chemical system, where *m* is the number of samples and *n* is the number of variables. If Beer's law is strictly obeyed, there exists a bi-linear model,

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} \tag{1}$$

where **C** and **S** are the elution and spectral matrices of pure components, respectively.

Performing singular value decomposition (SVD) on X yields

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{1} \tag{2}$$

where **U** and **V** are orthonormal matrices, and **S** is a diagonal matrix whose diagonal elements are equal to the square roots of corresponding eigenvalues. Columns of matrices **US** and **V** represent score vectors and eigenvectors, respectively.

2.1. Selecting key spectral variables

Key spectral variables were selected from row vectors of matrix **X** with orthogonal projection approach (OPA). OPA is a stepwise approach to search the most dissimilar rows of the matrix. Key spectral variables, which are the least correlated rows of the data matrix and have a large mean absorbance [9], form a matrix **Xs** in the following manner,

$$\mathbf{Xs} = \begin{bmatrix} \mathbf{x}_1^{\mathsf{T}} \ \mathbf{x}_2^{\mathsf{T}} \ \mathbf{x}_3^{\mathsf{T}} \dots \ \mathbf{x}_p^{\mathsf{T}} \end{bmatrix}$$
(3)

where \mathbf{x}_i is the *i*th key spectral variable of data matrix \mathbf{X} . The number of key spectral variables, *p*, should not be greater than the number of rows or columns of \mathbf{X} . The size of $\mathbf{X}\mathbf{s}$ is $n \times p$.

2.2. Reconstructing the new data matrix

Let S^* denote an approximation of the pure spectra matrix in Eq. (1), and then compute an approximate matrix of elution profiles with the following equation. The initialization of S^* is Xs.

$$\mathbf{C}^* = \mathbf{X}\mathbf{S}^* \left(\mathbf{S}^{*T}\mathbf{S}^*\right)^{-1} \tag{4}$$

where **C**^{*} denotes the elution profiles.

Then constraints of non-negativity and unimodality were applied to C^* [8,14] before obtaining a new approximate matrix of spectra profile S^* by least squares.

$$\mathbf{S}^* = \mathbf{X}^T \mathbf{C}^* \left(\mathbf{C}^{*T} \mathbf{C}^* \right)^{-1}$$
(5)

Finally, a new data matrix is reconstructed from C^{*} and S^{*}.

$$\mathbf{X}^* = \mathbf{C}^* \mathbf{S}^{*T} \tag{6}$$

2.3. Determining the number of principal factors

The original data matrix **X** and the reconstructed one **X**^{*} are similar in terms of chemical information and different when non-chemical information is considered. This is also true for eigenvectors obtained with SVD, which means principal eigenvectors from **X** and **X**^{*} are similar, but non-principal ones are different. Therefore, in a series of stepwise comparisons between eigenvectors, a significant difference indicates a non-principal eigenvector, and all previous ones are principal, so the number of principal factors are determined.

Let **v** and **v**^{*} denote eigenvectors from **X** and **X**^{*}, respectively. For principal factors, **v** and **v**^{*} are very similar; for non-principal factors, **v** and **v**^{*} are significantly different [15]. This fact is the foundation of OPALS to determine the number of principal factors. The consistency between two eigenvectors can be measured by congruence coefficients [16],

$$c_k = \mathbf{v}_k^{\mathsf{T}} \mathbf{v}_k^* \tag{7}$$

where c_k is the congruence coefficient for the *k*th eigenvectors, \mathbf{v}_k and \mathbf{v}_k^* . If c_k is close to 1, and c_{k+1} is much smaller than 1 or c_{k+1} decreases abruptly from c_k , then the number of principal factors is estimated as *k*. Major steps of OPALS were summarized as follows:

- 1. Select key spectral variables with OPA and compose a matrix **Xs**, as shown in Eq. (3).
- 2. Let **Xs** be an initial estimate of pure spectra matrix, **S**^{*}, then compute an approximate matrix of elution profiles, **C**^{*}, and an approximation of spectra matrix, **S**^{*}, by means of least square regression, as shown in Eqs. (4) and (5).
- 3. Reconstruct a new data matrix, $\mathbf{X}^* = \mathbf{C}^* \mathbf{S}^* \mathbf{T}$.
- 4. Perform SVD to **X** and **X**^{*} respectively to obtain two sets of eigenvectors, and compute congruence coefficients with Eq. (7).
- 5. Determine the number of principal factors from congruence coefficients.

3. Experimental

The proposed method was tested with both simulated GC-IR and experimental HPLC-DAD data. All programs were written in MATLAB 2010a (The MathWorks, Inc., Natick, MA).

3.1. Simulated data

Three-component GC-IR data were simulated with IR spectra of diethyl ether, ammonia and beta propiolactone within the region from 750 cm⁻¹ to 1250 cm⁻¹, as shown in Fig. 1a; elution profiles were generated with a Gaussian function where the width of each Gaussian peak is 2, 2 and 0.8 respectively, as shown in Fig. 1b. By altering relative heights and positions of the Gaussian peaks, the degree of minor component and overlapping level of elution profiles were simulated. Homoscedastic noise with standard deviation being 1% of the maximum absorbance was added. The size of the simulated matrix is 50 by 2075.

Bi-linear matrices were simulated with the following equation.

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} \tag{8}$$

where **C** and **S** are the elution and spectral matrices of pure components, respectively.

3.2. Experimental data

HPLC-DAD data of mixed rare-earth oxides in hydrochloric acid solvent were collected using FL 2000 HPLC Workstation (Spectra-Physics, USA) and multiple wavelengths of UV-Vis detector (Spectra-Physics, USA) within region 580 to 720 nm at 5 nm interval. The sampling



Fig. 1. (a) The spectra and (b) the elution profiles of three chemical components in simulated GC-IR data. The components are (1) diethyl ether, (2) ammonia and (3) beta propiolactone, respectively.

duration is 15 min at 0.344 s interval. Six samples were prepared and analyzed within the chromatographic system. Two two-component (Yb and Tm) data sets were recorded between 4.5 and 8.6 min and labeled as HPLC-DAD1 (high concentration) and HPLC-DAD2 (low concentration); two three-component (Yb, Tm and Er) data sets were recorded between 4.5 and 9.9 min and labeled as HPLC-DAD3 (high concentration) and HPLC-DAD4 (low concentration); and two six-component (Lu, Yb, Tm, Er, Ho and Tb) data sets were recorded between 3.9 and 12 min and labeled as HPLC-DAD5 (high concentration) and HPLC-DAD6 (low concentration).



Fig. 2. Congruence coefficient of eigenvectors versus number of factors from three-component simulated GC-IR data matrix (50 × 2075). The number of selected key spectral variables was (a) 10, (b) 20, (c) 30 and (d) 40, respectively.

Table 1

Numbers of principal factors^a determined by five indices for simulated three-component data with various degrees of elution profile overlapping^b.

Position	OPALS				NPFPCA	RESO	DRAUG	DRMAD
	10	20	30	40				
3.14	3	3	3	3	3	3	3	3
3.10	3	3	3	3	3	3	3	2
3.06	3	3	3	3	3	3	3	2
3.04	2	3	3	3	3	2	2	2

^a Incorrect results were highlighted in bold.

^b The concentration peak position of ammonia varied, while the positions of diethyl ether and beta propiolactone remained at 3 and 7. Noise level is 0.01. Concentration peak heights are all 1.

Table 2

Numbers of principal factors^a determined by five indices for simulated three-component data with various degrees of minor component^b.

Height	OPA	LS			NPFPCA	RESO	DRAUG	DRMAD
	10	20	30	40				
0.50	3	3	3	3	3	3	3	3
0.10	3	3	3	3	3	2	3	3
0.05	3	3	3	3	3	2	3	2
0.01	2	2	2	2	2	2	2	2

^a Incorrect results were highlighted in bold.

^b The concentration peak height of ammonia varied, while the heights of diethyl ether and beta propiolactone both remained at 1. Noise level is 0.01. Concentration peak positions are 3, 4.283 and 7, respectively.

Table 3

Numbers of principal factors^a determined by five indices for simulated three-component data with various levels of homoscedastic noise^b.

Noise level	OPALS				NPFPCA	RESO	DRAUG	DRMAD
	10	20	30	40				
0.01	3	3	3	3	3	3	3	3
0.05	3	3	3	3	3	3	3	2
0.08	3	3	3	3	3	2	2	1
0.10	2	3	3	3	3	2	2	1

^a Incorrect results were highlighted in bold.

^b Concentration peak heights of diethyl ether, ammonia and beta propiolactone are all 1, and peak positions are 3, 4.283 and 7, respectively.

4. Results and discussion

4.1. Three-component simulated GC-IR data

From the simulated data matrix, 10 key spectral variables were selected by OPA. Following the procedure outlined in Section 2.3, a series of congruence coefficients were obtained, and plotted versus numbers of factors, as shown in Fig. 2(a). The figure shows a sharp decline of congruence coefficients when the number of factors changes from 3 to 4. Therefore, the number of principal factors was determined to be 3, which is consistent with the number of components in the simulated data.

4.1.1. Effect of the number of key spectral variables

In order to investigate possible effects of the number of key spectral variables on OPALS, 20, 30 and 40 key spectral variables were used in determining the number of principal factors. The results were shown in Fig. 2b to d. The congruence coefficient profile in Fig. 2b is sufficiently clear to indicate 3 principal factors despite an increase at c_4 . When the number of key spectral variables increases to 30 or 40, there is a gradual decline from c_3 to c_6 , causing the turning point for distinguishing the principal factors from the secondary factors less prominent. This is due to noise accumulation in the reconstructed matrix when more key spectral variables were used. Nonetheless, the congruence coefficients in Fig. 2c and d are all close to 1 for the first 3 factors, making them readily distinguished from others, and the correct number of principal factors could still be obtained.

4.1.2. Effect of three interfering factors

Effects of three interfering factors, i.e., elution profile overlapping, minor component and homoscedastic noise, on OPALS were investigated and results are listed in Tables 1–3. For comparison, other four indices were used to process the same data sets, which are NPFPCA (noise perturbation in functional principal component analysis) [16], RESO (the ratio of eigenvalues calculated by smooth principal component analysis and those calculated by ordinary principal component analysis) [3], DRAUG (determination of rank by augmentation) [2] and DRMAD (determination of rank by median absolute deviation) [17]. The tables show that the three interfering factors



Fig. 3. A 3D plot of experimental HPLC-DAD data set 4.

Table 4

Numbers of principal factors^a determined with various number of selected key spectral variables for six HPLC-DAD data sets.

Data	Size	Number of key spectral variables				
		10	15	20	25	
HPLC-DAD1	717 imes 29	7	7	7	7	
HPLC-DAD2	717 imes 29	5	5	5	5	
HPLC-DAD3	932 imes 29	8	9	9	9	
HPLC-DAD4	932 imes 29	5	5	5	5	
HPLC-DAD5	1600 imes 25	6	6	6	6	
HPLC-DAD6	1600×25	6	6	6	6	

^a Inconsistent results are highlighted in bold.

have similar effect on these methods except DRMAD, which performed relatively poor in those cases. OPALS performs similarly to NPFPCA and better than others.

In some extreme cases when the degree of elution profile overlapping is serious, minor component is very weak, or the noise level is quite high, OPALS of using10 key spectral variables yielded 2 principal factors, one fewer than the true value. It is probably ascribable to the insufficiency of 10 key spectral variables, which resulted in some loss of chemical information in the reconstructed matrix. Nonetheless, correct results were obtained once more key spectral variables were used. Therefore, it is desirable to compare the results of using different number of key spectral variables in order to reduce effects of various interferences in original data matrix.

4.2. Experimental data

Experimental data contain both chemical and non-chemical information that includes noise, instrumental fluctuations [18]. Nonchemical information, e.g., the oscillations in baseline in Fig. 3, which are from pulses of the pump, is not necessarily separable from chemical information and principal factors ought to account for it [2]. As a result, the number of principal factors of experimental data is usually larger than the number of chemical components. The numbers of principal factors for experimental data sets of HPLC-DAD1, HPLC-DAD2, HPLC-DAD3 and HPLC-DAD4 were confirmed to be 7, 5, 9 and 5 with window factor analysis (WFA), and these values are references to results of OPALS. For more complicated data sets HPLC-DAD5 and HPLC-DAD6, WFA failed to yield the referential numbers of principal factors, so we used the number of chemical components instead.

4.2.1. Results of using different numbers of key spectral variables

As found in Table 4, the number of key spectral variables did not affect the results of data sets HPLC-DAD1, HPLC-DAD2, HPLC-DAD4, HPLC-DAD5 and HPLC-DAD6, and the determined numbers of principal factors are 7, 5, 5, 6 and 6, respectively. Those values are consistent with the references. For data set HPLC-DAD3, the result of using 10 key spectral variables is 8, while the results of using more key spectral variables are unanimously 9, as shown in Fig.4a–d. Therefore, the final result should be 9, and it is also consistent with the reference.

For all six HPLC-DAD data sets, plots of congruence coefficient versus number of factors are presented in Fig. 5a–f, and the numbers



Fig. 4. Congruence coefficient of eigenvector (*c_k*) versus number of factors for data HPLC-DAD3. The number of selected key spectral variables was (a) 10, (b) 15, (c) 20 and (d) 25, respectively.



Fig. 5. Congruence coefficients of eigenvector (ck) versus number of factors for data sets from HPLC-DAD1 to HPLC-DAD6 (a-f).

of principal factors were readily determined to be 7, 5, 9, 5, 6 and 6, respectively. Congruence coefficients were obtained with 20 key spectral variables.

4.2.2. Comparison with other methods

The experimental HPLC-DAD data were also processed with four popular methods, and the determined numbers of principal factors were listed in Table 5. It was found OPALS yielded similar results to those of NPFPCA, DRAUG and DRMAD. Moreover, OPALS and NPFPCA achieved highest accuracy. Results of RESO are generally smaller than those of other methods and incorrect for data HPLC_DAD5 and

Table 5						
Numbers of p	orincipal factor	s ^a determin	ed by five	indices for s	six HPLC-DA	D data sets.
Data	OPALS	NPFPCA	RESO	DRAUG	DRMAD	Referenceb

Data	OPALS	NPFPCA	RESO	DRAUG	DRMAD	Reference ^b
HPLC-DAD1	7	6	3	7	8	7 (2)
HPLC-DAD2	5	5	2	8	9	5(2)
HPLC-DAD3	9	8	4	7	6	9(3)
HPLC-DAD4	5	5	3	8	7	5 (3)
HPLC-DAD5	6	6	3	7	8	6(6)
HPLC-DAD6	6	6	5	7	8	6 (6)

^a Incorrect results are highlighted in bold.

^b Parenthesized values are numbers of chemical components.

HPLC_DAD6 since the number of principal factors should not be smaller than the number of chemical components.

5. Conclusion

A new method called OPALS was proposed in this paper to determine the number of principal factors of data matrix. The core idea of OPALS is to compare eigenvectors obtained from the original data matrix and the one reconstructed from key spectral variables. OPALS is mathematically rigorous, and the determination was clear. OPALS yielded accurate results for both simulated and experimental data, as compared with four popular methods NPFPCA, RESO, DRAUG and DRMAD. The performance of OPALS is hardly affected by the number of key spectral variables that are used to reconstruct a new data matrix. However, for some complicated data, it is advisable to run OPALS with different numbers of key spectral variables and choose the stable result. It is also noteworthy that key spectral variables should not outnumber either the rows or the columns of the data matrix.

There is not a universal solution to the problem of determining the number of principal factors, otherwise there would not have been so many methods, such as NPFPCA, RESO, DRAUG and DRMAD cited in this paper. A practically effective strategy is to use more than one method on the same data, and compare their results in order to yield a consensus. As a mathematically rigorous method, OPALS could provide yet another reliable result to refer to for the consensus.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 21175123), the Program for New Century Excellent Talents in University (No. NCET-11-0878), the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) and the Fundamental Research Funds for the Central Universities (wk2060190040).

References

- E.R. Malinowski, Determination of rank by augmentation (DRAUG), J. Chemom. 25 (2011) 323–328.
- [2] E.R. Malinowski, Adaptation of the Vogt–Mizaikoff F-test to determine the number of principal factors responsible for a data matrix and comparison with other popular methods, J. Chemom. 18 (2004) 387–392.
- [3] Z.P. Chen, Y.Z. Liang, J.H. Jiang, Y. Li, J.Y. Qian, R.Q. Yu, Determination of the number of components in mixtures using a new approach incorporating chemical information, J. Chemom. 13 (1999) 15–30.
- [4] H.L. Shen, Y.Z. Liang, O.M. Kvalheim, R. Manne, Determination of chemical rank of two-way data from mixtures using subspace comparisons, Chemom. Intell. Lab. Syst. 51 (2000) 49–59.
- [5] M. Wasim, R.G. Brereton, Determination of the number of significant components in liquid chromatography nuclear magnetic resonance spectroscopy, Chemom. Intell. Lab. Syst. 72 (2004) 133–151.
- [6] W. Lu, L.M. Shao, Methods to estimate the number of principal factors of data matrices of complex chemical systems: a review and a categorization, Sci. China Chem. 10 (2014) 1499–1507.
- [7] K. De Braekeleer, D.L. Massart, Evaluation of the orthogonal projection approach (OPA) and the simplisma approach on the Windig standard spectral data sets, Chemom. Intell. Lab. Syst. 39 (1997) 127–141.
- [8] S. Gourvénec, X. Capron, D.L. Massart, Genetic algorithms (GA) applied to the orthogonal projection approach (OPA) for variable selection, Anal. Chim. Acta 519 (2004) 11–21.
- [9] S. Gourvénec, D.L. Massart, D.N. Rutledge, Determination of the number of components during mixture analysis using the Durbin–Watson criterion in the orthogonal projection approach and in the simple-to-use interactive self-modelling mixture analysis approach, Chemom. Intell. Lab. Syst. 61 (2002) 51–61.
- [10] F. Cuesta Sánchez, J. Toft, B. van den Bogaert, D.L. Massart, Orthogonal projection approach applied to peak purity assessment, Anal. Chem. 68 (1996) 79–85.
- [11] D.M. Haaland, E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, Anal. Chem. 60 (1988) 1193–1202.
- [12] D. Kalman, A singularly valuable decomposition: the SVD of a matrix, Coll. Math. J. 27 (1996) 2–23.
- [13] E.R. Malinowski, Window factor analysis: theoretical derivation and application to flow injection analysis data, J. Chemom. 6 (1992) 29–40.
- [14] A. de Juan, Y. Vander Heyden, R. Tauler, D.L. Massart, Assessment of new constraints applied to the alternating least squares method, Anal. Chim. Acta 346 (1997) 307–318.
- [15] Y.J. Yu, H.Y. Fu, H.W. Gu, Y. Li, C. Kang, Y.P. Wang, H.L. Wu, Chemical rank estimation for second-order calibration by discrete Fourier transform coupled with robust statistical analysis, Chemom. Intell. Lab. Syst. 141 (2015) 47–57.
- [16] C.J. Xu, Y.Z. Liang, Y. Li, Y.P. Du, Chemical rank estimation by noise perturbation in functional principal component analysis, Analyst 128 (2003) 75–81.
- [17] E.R. Malinowski, Determination of rank by median absolute deviation (DRMAD): a simple method for determining the number of principal factors responsible for a data matrix, J. Chemom. 23 (2009) 1–6.
- [18] W. Lu, L.M. Shao, Estimation of primary factors based on residual matrix variance ratio, J. Univ. Sci. Technol. China 11 (2014) 881–886.