ELSEVIER

Talanta 44 (1997) 585–591

Talanta

Second-order data by flow injection analysis with spectrophotometric diode-array detection and incorporated gel-filtration chromatographic column

Iben Ellegaard Bechmann

Department of Chemistry, Technical University of Denmark, DK-2800 Lyngby, Denmark

Received 24 April 1996; received in revised form 29 July 1996; accepted 29 July 1996

Abstract

A flow injection analysis (FIA) system furnished with a gel-filtration chromatographic column and with photodiode-array detection was used for the generation of second-order data. The system presented is a model system in which the analytes are blue dextran, potassium hexacyanoferrate(III) and heparin. It is shown that the rank of the involved sample data matrices corresponds to the number of chemical components present in the sample. The PARAFAC (parallel factor analysis) algorithm combined with multiple linear regression and the tri-PLS (tri-linear partial least-squares regression), which allows unknown substances to be present in the sample, are implemented for FIA systems and it is illustrated how these three-way algorithms can handle spectral interferents. The prediction ability of the two methods for pure two-component samples and also the predictions ability in the presence of unknown interferents are satisfactory. However, the predictions obtained by tri-PLS are slightly better than those obtained using PARAFAC regression algorithm. © 1997 Elsevier Science B.V.

Keywords: Multivariate calibration; PARAFAC; Spectral interferents; Tri-PLS

1. Introduction

Data structures produced by analytical chemical systems in general and by flow injection analysis (FIA) systems in particular can be divided into classes reflecting the complexity of the data, ranging from scalars, vectors and matrices to higher order data structures. More than 90% of the papers published in analytical chemical journals are based on zeroth-order data structures [1]. This simple data structure arises when only one measurement per sample is recorded. A set of calibration samples thus gives a vector. The measurement can be a variety of signals, e.g. potential, current, absorbance or, as often used in FIA, the height (or area) of the sample peak. The use of zeroth-order data requires that the measured signal is absolutely selective and additionally that the signal is a known function (e.g. linear) of the concentration of the analyte. If the selectivity assumption is not fulfilled for all samples, an improvement can be achieved by using first-order data structures [1], which gives the possibility of detecting outlying samples. First-or-

PII S0039-9140(96)02066-8

der data, where each sample gives rise to a vector and a set of calibration samples hence yields a matrix, can be obtained in two ways in FIA: (1) by recording the actual measurement (e.g. the absorbance at a given wavelength) as a function of time; and (2) by recording multiple measurements at a given time (e.g. the UV/VIS spectrum at the peak maximum). The condition for employing first-order data is that the responses are linear and additive for all analytes.

The number of FIA papers published on first-order data using two-way data analytical methods for instance partial least-squares (PLS) regression is increasing [2–5]. From a multivariate chemometric point of view, second-order data structures [1], where each sample gives rise to a matrix and a set of calibration samples to a cube, are, however, much more interesting because these data make it possible not only to detect outlying samples but also to determine correct analyte concentrations even though the analytical signal is not selective. This has been called the second-order advantage [1]. Second-order data can be achieved, for instance, by scanning multiple wavelengths at multiple times. The majority of papers published concerning second-order data in FIA rely on spectrophotometric diode-array detection [6,7]. The condition for optimal use of second-order data is that the data structure is bilinear, which means that the responses are linear and additive for all analytes and that no co-elution of analytes is allowed. This bilinear data structure is found in chemical systems obeying the Lambert–Beers law and where the analytes are separated in time by a column, e.g. in chromatographic systems combined with a photodiode-array detector.

In this work, a flow injection system with an incorporated gel-filtration chromatographic column and a UV/VIS photodiode-array detector was used for the generation of second-order data. In an earlier paper [8], the application of a gel-filtration chromatographic column in a FIA system for enzymatic determination of formaldehyde in aqueous fish extracts was reported. In that work, the purpose of providing the FIA system with a gel-filtration chromatographic column was on-line removal of the protein fraction of the extract prior to the enzymatic analysis. The purpose of this work was partly to improve the possibilities of the application of gel-filtration chromatographic columns in FIA and partly to illustrate the general advantages of using second-order FIA data, especially in cases where unknown spectral interferents are present. In that context, attention is drawn to an early paper [9] where the resemblance of flow injection analysis with chromatographic systems in general is discussed further. The system presented here is a model system where the analytes blue dextran, potassium hexacynoferrate(III) and heparin were chosen to illustrate the benefit of the methodology used. The data structure produced by the FIA system is a bilinear data structure, where each sample produces a 76 (times) × 188 (wavelength) matrix. Among the most common methods used for second-order calibration are unfolding methods, where the calibration cube is unfolded to an ordinary matrix which is then treated by standard two-way methods (e.g. PLS). In the case of response matrices where each chemical component gives rise to a matrix of rank one, the rank annihilation factor analysis (RAFA) method [10] or the general rank annihilation method (GRAM) can be used [10,11] for the prediction of correct analyte concentration in presence of spectral interferents.

In this paper, the tri-linear parallel factor analysis (PARAFAC) algorithm [12–14] and the newly developed tri-linear partial least-squares regression (tri-PLS) [15] are implemented for FIA systems. PARAFAC is a generalization of PCA to higher order data arrays, but there are some important differences between the two methods. Tri-PLS is a generalization of the ordinary two-way PLS methods. A short description of the two methods is given in Section 3. Compared with the unfolding methods, the multi-linear models are much simpler because they use fewer parameters.

2. Experimental

2.1. Apparatus

2.1.1. FIA system

The FIA system used in the experiments was a simple one-line FIA system. All tubing employed was PTFE (0.5 mm i.d.) and the carrier stream was propelled by an Ismatec MS-4 Reglo peristaltic pump. The sample (50 μl) was aspirated into the valve by the same pump. The carrier stream was pumped through a gel-filtration chromatographic column (HiTrap Desalting; Pharmacia, Uppsala, Sweden). The detector was a TIDAS (Zeiss J and M Analytische Mess- und Regeltechnik, Germany) photodiode array spectrophotometer furnished with a 8 μl flow cell. A 12-port injection valve (made in this laboratory) was used for injection of the sample. The photodiode array starts scanning 20 s after injection and continues for a further 75 s at 1.0 s intervals. The wavelength range recorded in each scan is 220–594 nm (every 2 nm). Injection of one sample thus, generates a data matrix containing 76 × 188 absorbances. An example of the data structure produced by injection of 50 μl sample is depicted in Fig. 1.

2.1.2. *Gel-filtration column*

The column used in this work was a HiTrap desalting column (Pharmacia) of dimensions 25 × 16 mm i.d. The column is filled with Sephadex G-25 Superfine and has a fractionation range between 1000 and 5000 Da. This column shows a significantly better separation of low and high molecular weight substances than the Sephadex G-25 column used in the previous work [8], which was packed in this laboratory and consisted of a piece of PVC into with was machined a tubular cavity of 52 × 3.3 mm i.d. When not in use, the column reactor was stored filled with a 20% (v/v) ethanol solution in a refrigerator at 5°C to prevent microbiological growth.

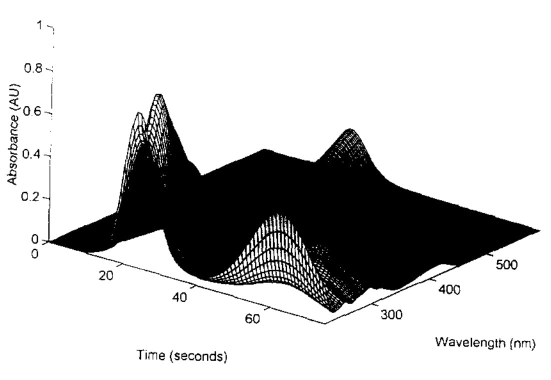

Fig. 1. Three-dimensional plot of the data matrix corresponding to injection of a mixture of blue dextran and $K_3Fe(CN)_6$.

Table 1
Concentrations of blue dextran and $K_3Fe(CN)_6$ for nine calibration samples

Sample	Blue dextran (mg ml^{-1})	$K_3Fe(CN)_6$ (mg ml^{-1})
C1	0.00	0.00
C2	2.65	0.00
C3	5.30	0.00
C4	0.00	0.80
C5	2.65	0.80
C6	5.30	0.80
C7	0.00	1.60
C8	2.65	1.60
C9	5.30	1.60

2.2. *Reagents*

The samples use in the model system were made from stock solutions of a high molecular weight substance (Blue dextran, Mw $\approx 10^6$ g mol^{-1}; Sigma) and a low molecular weight substance (potassium hexacyanoferate(III), Mw = 329.2 g mol^{-1}; Merck) in degassed, distilled Millipore-filtered water containing 0.9% NaCl. The stock solutions, which contained 26.5 g l^{-1} blue of dextran and 1.62 g l^{-1} of potassium hexacyanoferrate(III), were diluted with 0.9% NaCl solution for the preparation of samples. A solution of heparin (Tinzaparine, MW $\approx$ 4000 g mol^{-1}) containing 10 g l^{-1} was used in some of the test samples as an interfering component.

2.3. *Programs*

Data collection and control of the pump and valve were performed by a software package from J&M Analytische Mess- und Regeltechnik (Germany). All calculations were preformed in Matlab ver. 4.0 for Windows (Math Works), partly by use of PARAFAC and multi-linear PLS algorithms available from the Internet at http://newton.foodsci.kvl.dk/foodtech.html (R. Bro and C.A. Andersson).

3. Theory

3.1. Data structure

The data structure produced by the FIA system is a bilinear data structure, where each sample produces a J (times) × K (wavelength) matrix. If all responses are linear and no noise is present, the rank of the measured response matrix will be equal to the number of chemical substances present in the sample. However, the true rank of the sample matrix is often not determinable because of real data contain non-linearities and noise. For this reason, it is necessary to determine a pseudo-rank instead. In the present work, principal component analysis (PCA) [10] is used for this purpose. In this context the rank equals the number of principal components which is necessary to explain 99% of the variance in the response matrix. The dimension of the three-way calibration data array, **X**, is $I \times J \times K$, where I is the number of samples in the calibration set. The pre-processing of this kind of raw data is more complicated than in the two-way case. In this work no centering and no scaling were used in the PARAFAC models. The data arrays used for tri-PLS are centered by unfolding the data array to a $I \times JK$ matrix and then mean center the unfolded matrix. No scaling was used.

3.2. PARAFAC

PARAFAC performs a tri-linear decomposition of the data array, **X**, and can be regarded as a generalization of the bilinear PCA. In PARAFAC the tri-linear model of **X** is found to minimize the sum of squares

$$\left(\sum e_{ijk}^2\right)$$

in the model

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}$$

where x_{ijk} is an element in **X** ($i = 1, ..., I$, $j = 1, ..., J$, and $k = 1, ..., K$), a are the scores, b and c are loadings in the model and F is the number of factors. In contrast to PCA, the number of factors to be included in the PARAFAC model must be determined before the modelling.

A very obvious advantage of PARAFAC over PCA is the uniqueness of the solution. The loadings in a spectral bilinear PCA model reflect the pure spectra of the analytes measured, but it is not possible without external information actually to find the pure spectra. In the data used for the PARAFAC model represent a second-order data structure, the true analyte spectra can often be found directly, if only the concentrations vary independently pairwise and no spectra are linearly dependent on any of the others. In this work, PARAFAC is used for prediction purposes by

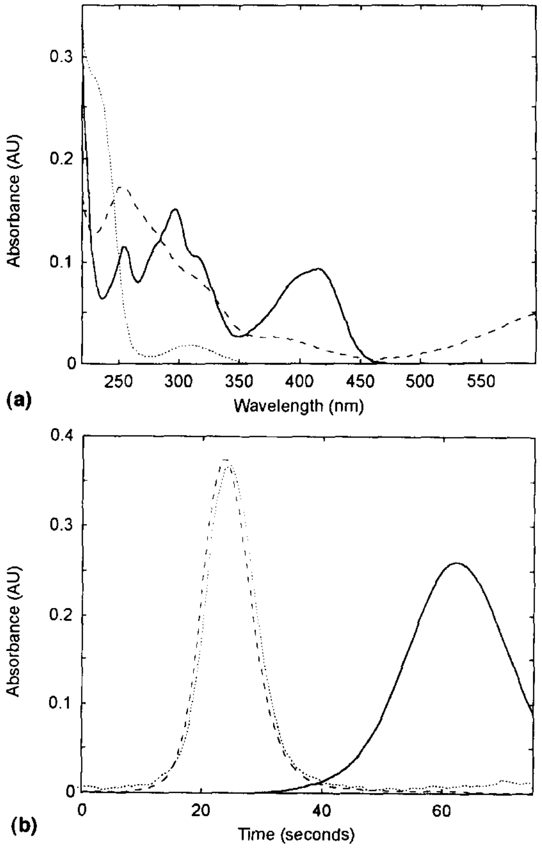

Fig. 2. (a) Normalized pure spectra of the three components. Dash line, blue dextan; solid line, $K_3Fe(CN)_6$; Dotted line, heparin. (b) Normalized concentration profiles calculated by alternating regression on each of the pure sample matrices. Lines as in (a).

following the strategy of principal component regression (PCR) [10]. The scores from the PARAFAC model are used for the predictions using multiple linear regression.

3.3. Tri-PLS

The tri-linear PLS is a natural extension of the ordinary bi-linear PLS method. By the tri-linear decomposition in tri-PLS the calibration cube **X** is decomposed into a set of rank-one cubes describing **X** in some optimal sense. This does not mean, however, that the part of **X** relevant for describing the dependent variables (the analyte concentrations) has to be of rank one. One advantage of tri-PLS over PARAFAC is the incorporation of the dependent variables (the analyte concentrations) in the decomposition of the measured calibration cube, which might stabilize the predictive model. The theory of multi-linear PLS can be found elsewhere [15].

4. Results and discussion

The samples used for calibration consisted of mixtures of the two analytes, blue dextran and potassium hexacyanoferrate(III). The samples were prepared as a two-factor, three-level factorial design, i.e. the calibration set consisted of nine samples. The compositions of the calibration samples are given in Table 1.

An independent test set of 12 different test samples was prepared. To illustrate how the PARAFAC and the tri-PLS algorithms handle spectral interferents, some of the test samples in addition contained heparin.

The pure spectra and the concentration profiles of blue dextran, potassium hexacyanoferrate(III) and heparin are shown in Fig. 2(a) and (b), respectively.

The spectra and time profiles are found by resolution of each of the three data matrices measured on pure standards using the alternating least squares (ALS) algorithm [10]. From Fig. 2(b), it is seen that the time profiles of the dextran and the $K_3Fe(CN)_6$ differ owing to different resolution times on the gel-filtration column. The resolution times of heparin and dextran are almost the same owing to their molecular weights of close to 4000 Da and higher than 5000 Da, respectively. In the current application, the exclusion limit of the column is apparently about 4000 Da and not 5000 Da as specified by Pharmacia. The spectral characteristics of the three components, however, are different. The data structure obtained by injection of a pure mixture of blue dextran and $K_3Fe(CN)_6$ is thus, expected to be a bilinear second-order data structure.

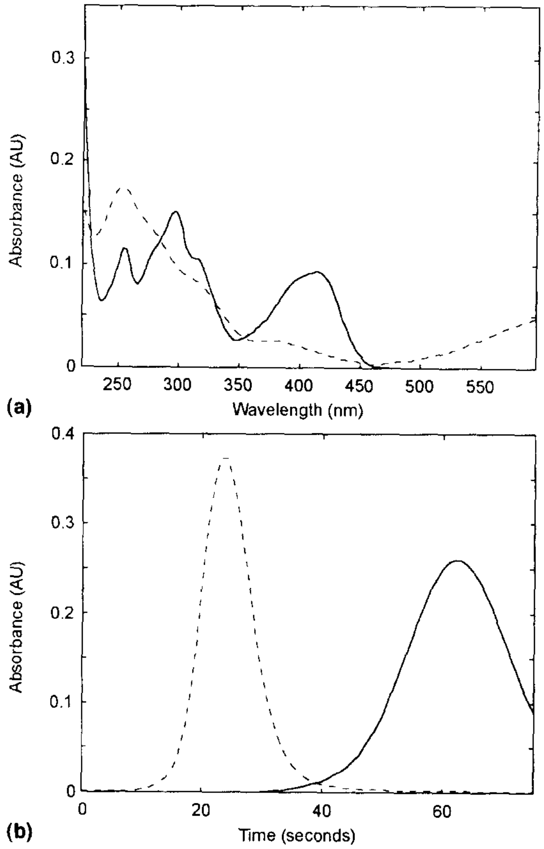

Fig. 3. (a) Spectra found by the PARAFAC algorithm (normalized loading vectors of second mode) for the analytes used for calibration. Dashed line blue dextran; Solid line $K_3Fe(CN)_6$. (b) Time profiles found by the PARAFAC algorithm (normalized loading vectors of third mode) for the analytes used for calibration. Lines as in (a).

A PARAFAC model with two components should give the right solution since there are two

Table 2
True and predicted concentrations (in mg ml^{-1}) obtained by the PARAFAC algorithm and tri-PLS algorithm for the 12 test samples

Sample	True			Predicted by PARAFAC		Predicted by tri-PLS		Rank[a]
	Dextran	$K_3Fe(CN)_6$	Heparin	Dextran	$K_3Fe(CN)_6$	Dextran	$K_3Fe(CN)_6$	
T1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	—
T2	2.25	0.00	0.00	2.26	0.00	2.25	0.00	1
T3	4.50	0.00	0.00	4.51	0.00	4.50	0.00	1
T4	0.00	0.68	0.00	0.01	0.68	0.00	0.68	1
T5	2.25	0.68	0.00	2.26	0.68	2.25	0.68	2
T6	4.50	0.68	0.00	4.51	0.68	4.50	0.68	2
T7	0.00	1.36	0.00	0.00	1.36	0.00	1.36	1
T8	2.25	1.36	0.00	2.26	1.36	2.25	1.36	2
T9	4.50	1.36	0.00	4.51	1.36	4.50	1.36	2
T10	3.15	0.00	2.00	3.20	0.00	3.18	0.00	2
T11	0.00	0.80	2.00	0.00	0.80	0.00	0.80	2
T12	3.15	0.80	2.00	3.26	0.82	3.24	0.81	3

[a] The 'Rank' column indicate the pseudo-rank of the sample matrix determined by PCA.

analytes in the calibration set. The spectral loadings of a two-component PARAFAC model are shown in Fig. 3(a) and the corresponding time loadings are shown in Fig. 3(b). From comparison with Fig. 2(a) and (b), it is seen that the model estimates precisely both the pure spectra and the time profiles of the two analytes.

The PARAFAC model was validated by test set validation with the nine samples in a calibration set and 12 other samples in a test set. For the predictions of analyte concentrations in the test samples, a regression model was made from the scores of the two-component PARAFAC model. For comparison, the results of using a tri-PLS model with two PLS components were also calculated. The true concentrations and the concentrations predicted by PARAFAC for the 12 test samples are given in Table 2 together with the concentrations predicted by tri-PLS. The pseudo-rank determined by PCA of the measured data matrices for the 12 test samples is also given in Table 2.

The rank of a sample matrix containing only one substance equals one and the rank of a sample matrix containing two substances equals two, and analogously with three substances. This indicates that an FIA system furnished with a gel-filtration column acts as expected of a chromatographic system.

It appears that the prediction of $K_3Fe(CN)_6$ concentrations in the presence of heparin in general is better than the predictions for dextran in the presence of heparin. This can be explained by the almost identical resolution profiles of the two high molecular weight substances used. Comparing the predicted and actual concentrations obtained by applying the two methods on the 12 test samples, the root mean square error of prediction (RMSEP) can be calculated. The RMSEP for the prediction of dextran is 0.03 mg ml^{-1} and that for $K_3Fe(CN)_6$ is below 0.01 mg ml^{-1}. The predictions obtained by tri-PLS are slightly better than those obtained by PARAFAC, but both the PARAFAC and the tri-PLS algorithms give satisfactory predictions in the presence of unknown chemical interferents. However, tri-PLS does not possess the uniqueness properties of PARAFAC and it is therefore not possible without external information actually to find the pure spectra of the analytes. If the investigated samples contain specific chemical components of interest, the PARAFAC algorithm might therefore be preferable.

Acknowledgements

The author thanks Professor Elo H. Hansen for valuable comments and suggestions. Rasmus Bro is thanked for helpful comments and for placing the Matlab m-files for PARAFAC and tri-PLS on the Internet. The Danish Ministry of Agriculture and Fisheries is acknowledged for financial support. This work was performed in collaboration with the Danish Institute for Fisheries Research, Department of Seafood Research.

References

[1] K.S. Booksh and B.R. Kowalski, Anal. Chem., 66 (1994) 782A.
[2] P. MacLaurin, P.J. Worsfold, P. Norman and M. Crane, Analyst, 118 (1993) 617.
[3] O. Hernandez, A.I. Jimenez, F. Jimenez and J.J. Arias, Anal. Chim. Acta, 310 (1995) 53.
[4] O. Hernandez, F. Jiminez, A.I. Jimenez and J.J. Arias, Analyst, 121 (1996) 169.
[5] C. Ridder and L. Nørgaard, Chemom. Intell. Lab. Syst., 14 (1992) 297.
[6] L. Nørgaard and C. Ridder, Talanta, 41 (1994) 59.
[7] L. Nørgaard and C. Ridder, Chemom. Intell. Lab. Syst., 23 (1994) 107.
[8] I.E. Bechmann, Anal. Chim. Acta, 320 (1996) 155.
[9] J. Ruzicka and G. D. Christian, Analyst, 115 (1990) 475.
[10] E.R. Malinowski, Factor Analysis in Chemistry, 2nd edn, Wiley, New York, 1991.
[11] E. Sanchez and B.R. Kowalski, Anal. Chem. 58 (1986) 496.
[12] A.K. Smilde, P.H. van der Graaf and D.A. Doornbos, Anal. Chim. Acta, 235 (1990) 41.
[13] A.K. Smilde, Chemom. Intell. Lab. Syst., 15 (1992) 143.
[14] A.K. Smilde, Y.-D. Wang and B.R. Kowalski, J. Chemom., 8 (1994) 21.
[15] R. Bro, J. Chemom., 10 (1996) 47.