ELSEVIER

Chemometrics and Intelligent Laboratory Systems 38 (1997) 183–196

Chemometrics and intelligent laboratory systems

Second-order multivariate curve resolution applied to rank-deficient data obtained from acid–base spectrophotometric titrations of mixtures of nucleic bases

A. Izquierdo-Ridorsa, J. Saurina, S. Hernández-Cassou, R. Tauler *

University of Barcelona, Department of Analytical Chemistry, Diagonal 647, 08028 Barcelona, Spain

Received 4 April 1996; revised 22 July 1996; accepted 8 March 1997

Abstract

Rank-deficient data matrices, obtained from simulated spectrophotometric acid–base titrations of mixtures of up to four nucleic bases (adenine, cytosine, hypoxanthine and uracil), were analyzed by second-order multivariate curve resolution. The analysis of these individual mixture data matrices gives a rank value of $n + 1$, where n is the number of nucleic bases present in the system. This number is, however, lower than $2n$, the number of spectrometrically active species theoretically present in the systems under study, since each nucleic base is expected to give two species, a protonated and a deprotonated species. This rank deficiency is solved when more than one titration is simultaneously analyzed by second-order multivariate curve resolution. Full rank recovery is achieved when the titration of the mixture of n nucleic bases and other $n - 1$ titrations, each one corresponding to a different base, are simultaneously analyzed. Results obtained by second-order multivariate curve resolution indicate that for a total resolution of the system full rank is necessary. However, resolution and quantitative determinations of individual nucleic bases in mixtures in the presence of interferences can be achieved (with a prediction error lower than 2% in most cases) even in the case of rank deficiency.

Keywords: Acid–base titrations; Multivariate curve resolution (second-order); Data matrices

1. Introduction

The quantitative determination of mixtures of nucleic bases is of great interest. Nowadays, the most widely used techniques are HPLC and capillary electrophoresis, which permit the analysis of both majority and minority bases [1–6]. However, these analyses are quite time consuming and other simpler methods should be applied when the interest is only in majority bases, as in the analysis of pharmaceuticals which contain mixtures of nucleic bases or derived compounds (nucleosides and nucleotides). Among these simpler methods the use of first order multivariate calibration for the study of a mixture of nucleic bases, with spectrophotometric detection, has been proposed [7]. The experimental data used were the UV spectra obtained for the different mixtures. Analysis of mixtures of nucleic bases could also focus on their acid–base properties. Potentiometric acid–base titrations have traditionally been used for the quantitative analysis of mixtures of compounds that present acid–base properties, but this procedure

* Corresponding author. Tel.: +34-3-4021545; fax: +34-3-4021233; e-mail: roma@quimio.qui.ub.es.

PII S0169-7439(97)00034-8

can only be applied successfully to very simple mixtures. For the analysis of more complex mixtures, first-order multivariate calibration has also been proposed [8], using the potentiometric values obtained through acid–base titrations as the experimental data. However, methods based on first-order multivariate calibration are mainly applicable to routine analysis since the preparation of a calibration matrix is time-consuming. Furthermore, the mixtures used as standards have to be similar (in interferences present, concentration ranges, etc.) to the unknown mixtures, in order to model correctly the relationship between analyte concentration and instrumental response. In this paper a multivariate curve resolution method is proposed for the analysis of mixtures of nucleic bases. In contrast to first-order multivariate calibration, this method has the advantage of simultaneously taking into account the spectroscopic and the acid–base behaviour of the substances and is thus more highly selective. It also avoids the need for a great number of calibration samples, because the one containing the analyte of interest is enough. The method and working conditions that are proposed here can be applied whenever mixtures of compounds presenting acid–base behaviour and absorption spectrum are to be analyzed.

In this paper, the experimental data are obtained from simulated spectrophotometric acid–base titrations of mixtures of up to four nucleic bases (adenine A, citosine C, hypoxanthine H and uracil U), arranged in individual single titration data matrices and in augmented multiple titration data matrices. The systems studied here are closed-reaction systems i.e. those where the total concentration is kept constant during the titration for each acid–base pair considered.

Multivariate curve resolution methods are commonly used to resolve the components present in unknown mixtures. When they are applied to ordered data matrices, a profile on each order (mode) of the data matrix for each component in the mixture is recovered. For instance, in the case of the multivariate curve resolution of a data matrix obtained in an acid–base spectrometric titration, two profiles are recovered for each species: a pH distribution profile and a species spectrum. Both profiles characterize the species in the system and identify it. We have recently shown [9,10] that multivariate curve resolution can easily be extended to the study of several correlated data matrices, each obtained under different conditions and therefore providing independent information. The proposed method of simultaneous treatment of several correlated data matrices by multivariate curve resolution is related to other second-order multivariate methods such as generalized rank annihilation (GRAM) [11], trilinear decomposition (TLS) [12] and Tucker and Tucker restricted three-way data analysis methods [13]. The proposed method shares some of the advantages of all these second-order data analysis methods such as the simultaneous resolution and quantification of the components in the analyzed mixtures and the possibility of performing the quantitative determinations in the presence of unknown and uncalibrated interferences (second order advantage). In previous papers, it has been shown that the proposed second-order multivariate curve resolution method is a powerful tool for species resolution and quantification of many types of unresolved chemical mixtures and is especially suitable for chemical data structures which do not fulfil the too high requirements of other second-order methods, such as trilinearity or strict analyte bilinearity.

This paper is intended to show the application of second-order multivariate curve resolution to the study of rank-deficient chemical data. The effects that rank-deficient data produce both in the resolution and in the quantification of unresolved mixtures are extensively studied. The presence of a rank deficiency has been observed for other kinds of experimental data [14,15]. The most important case is, perhaps, that of reacting mixtures where the number of independent reactions is lower than the number of response-active (absorbing) species. Mixtures of substances with acid–base behaviour (like nucleic bases) also give rank deficient data when spectrophotometric pH titrations are performed. A similar situation of a rank deficiency in this type of acid–base data mixtures has also been described [15] in the analysis of a mixture of three hydroxybenzaldehydes using a flow injection analysis system with a pH gradient.

2. Data sets under study

The main purpose of this paper is the study of the effects that a rank deficiency has upon resolution and

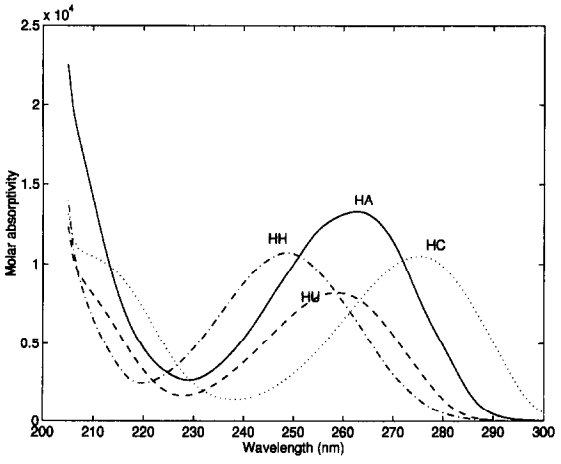

Fig. 1. Absorption spectra for the protonated forms of adenine (HA), cytosine (HC), hypoxanthine (HH) and uracil (HU).

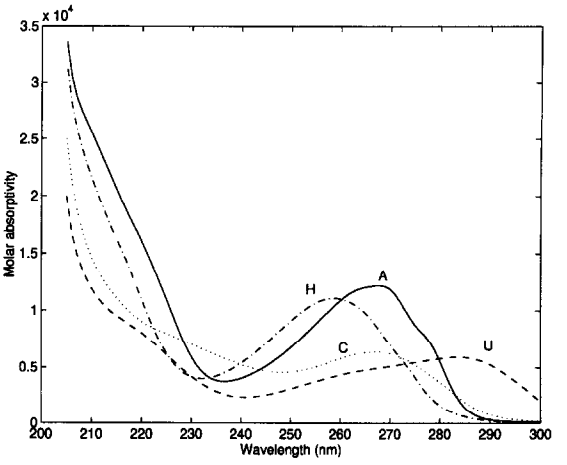

Fig. 2. Absorption spectra for the deprotonated forms of adenine (A), cytosine (C), hypoxanthine (H) and uracil (U).

quantification when multivariate curve resolution methods are applied and the evaluation of the best strategy in these cases. Simulated data were used in order to better compare and understand the results obtained. The main advantage of this kind of data is that the results obtained from the mathematical treatment can be easily compared with the theoretical ones, since these are known. Unambiguous conclusions about the behaviour of these systems can then be extracted and can be applied after when real systems are studied.

Four nucleic bases were used: adenine, cytosine, hypoxanthine and uracil. For each of them only one acid–base equilibrium was considered, with pK_a values of 4.02, 4.56, 8.51 and 8.93, respectively. Figs. 1 and 2 respectively show the absorption spectra for the protonated and deprotonated forms of the four nucleic bases considered. All the spectra are taken between 205 and 300 nm, at 1 nm intervals. Fig. 3 includes the concentration profiles for the four nucleic bases. It can be observed that the system of the four bases under study presents a high degree of overlapping of both spectra and distribution profiles.

Different spectroscopic titrations were simulated, each containing a different mixture of nucleic bases. Table 1 contains the concentrations of the nucleic bases present in these titrations.

The data obtained from each titration consist of the absorbance values measured between 205 and 300 nm at successive pH values (between pH 2 and pH 10.4, at 0.4 pH intervals), which can be considered as successive titration points. For each titration the data are arranged in a matrix **D**, which has a number of columns (n) equal to the number of wavelengths used and a number of rows (m) equal to the number of titration points (i.e. pH values). For the titration of a single nucleic base, matrix **D** was calculated using the

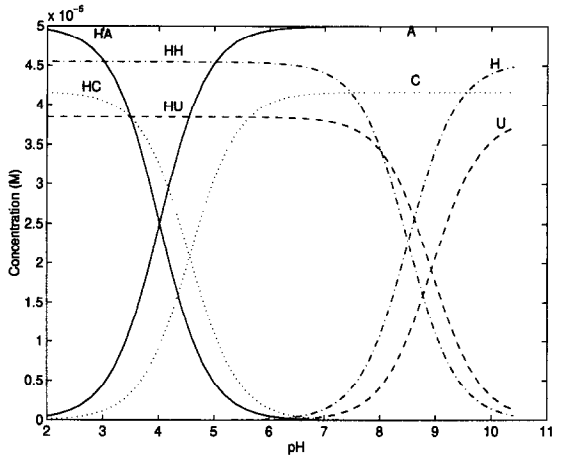

Fig. 3. Concentration profiles of adenine (5.00×10^{-5} M), cytosine (4.17×10^{-5} M), hypoxanthine (4.55×10^{-5} M) and uracil (3.85×10^{-5} M). HX and X (where X is A, C, H or U, see notation in Figs. 1 and 2) refer, respectively, to the protonated and deprotonated forms of each nucleic base.

Table 1
Composition of the individual titrations

Titration		Concentration ($\times 10^5$ M)			
number	symbol	adenine	cytosine	hypoxanthine	uracil
1	A	5	—	—	—
2	C	—	5	—	—
3	H	—	—	5	—
4	U	—	—	—	5
5	AU	5	—	—	5
6	AC	5	5	—	—
7	ACU	5	5	—	5
8	AHU	5	—	5	5
9	$ACHU_1$	5	5	5	5
10	$ACHU_2$	5	5	25	5
11	$ACHU_3$	5	25	25	5
12	$ACHU_4$	25	25	5	25

Lambert–Beer law in matrix form, as described below:

$$\boldsymbol{D} = \boldsymbol{C}\boldsymbol{S}^{\mathrm{T}}$$

The dimensions of these three matrices are $\mathbf{D}(m \times n)$, $\mathbf{C}(m \times 2)$ and $\mathbf{S}(n \times 2)$. The superscript T indicates the transposed matrix. Matrix **S** contains the individual spectra of the protonated and deprotonated forms of the nucleic base (see Figs. 1 and 2). Matrix **C** contains the concentrations of the protonated and deprotonated forms of the nucleic base at the different pH values (see Fig. 3). Throughout the titration the total concentration (c_a) of nucleic base was kept constant. At each pH value a closure condition controlled by the mass action law had to be fulfilled: the total concentration must be the sum of the concentrations of protonated and deprotonated nucleic base. From the known pK_a value and for a certain total concentration of nucleic base the concentrations of protonated (BH) and deprotonated (B) base are calculated at each pH value by applying the mass action law:

$$[\mathrm{BH}] = c_a[\mathrm{H}^+]/(\mathrm{K}_a + [\mathrm{H}^+])$$

$$[\mathrm{B}] = \mathrm{K}_a c_a/(\mathrm{K}_a + [\mathrm{H}^+])$$

For the titrations that contain more than one nucleic base, the absorbance data are calculated as follows:

$$\boldsymbol{D} = \sum \boldsymbol{C}_i \boldsymbol{S}_i^{\mathrm{T}}$$

$i = 1, \ldots, 4$, since up to four different nucleic bases may be present in the simulated systems under study. In these systems the concentrations of protonated and deprotonated forms for each nucleic base are related by their respective mass action law as mentioned above. It is important to note that this closure condition is fulfilled independently for each nucleic base.

A random error with a standard deviation of 0.001 absorbance units was added to the absorbance data. No systematic error, such as baseline drift was included in the data. In the study of the rank of the different matrices pure absorbance data free from random error were also analyzed.

Since the individual matrices have the two orders in common i.e. they share their row and column space (absorbance data were obtained at the same pH and wavelengths values for all the individual titrations), the second-order multivariate curve resolution method can be applied to the simultaneous analysis of more than one data matrix. In these cases the individual matrices are arranged in an augmented one. Fig. 4 indicates the two possible structures for an

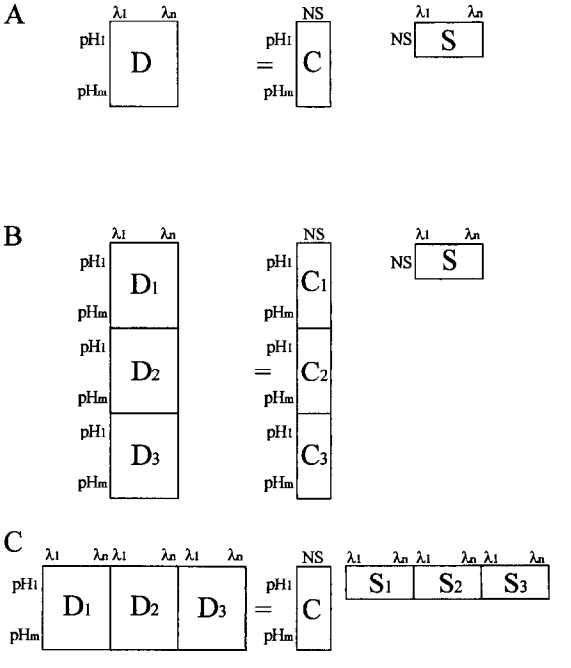

Fig. 4. Structures for: (A) individual, (B) augmented column-wise and (C) augmented row-wise data matrices. Augmented matrices are indicated using Matlab notation: $[\mathbf{D}_1; \mathbf{D}_2; \mathbf{D}_3]$ and $[\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3]$ refer to column-wise and row-wise augmented matrices respectively.

augmented matrix $\mathbf{D}_{aug}$. It can be augmented column-wise (wavelength-wise, keeping the wavelength values in common) or row-wise (pH-wise, keeping the pH values in common). In the first case, spectra must be measured at the same wavelengths in all the submatrices; in the second one, the pH values must coincide. The notation used to indicate column-wise or row-wise augmentation is explained in Fig. 4.

3. Data treatment

Although more detailed descriptions are given in other papers [10,16–20], the main steps of the second-order multivariate curve resolution procedure can be summarized as follows:

(1) Arrangement of the spectral individual and/or augmented data matrices. This step has been described in detail in Section 2.

(2) Rank analysis of the data matrices for the determination of the number of species present in each data matrix (individual or augmented). Since the analyzed data contain only randomly distributed error, the indicator function proposed by Malinowski [21] and the visual inspection of magnitudes of matrix singular values were used indistinctly. It should be pointed out that Malinowski's function fails when other contributions such as baseline drift or heterocedastic noise are present in the data. On the other hand, if the chemical components make a much larger contribution to the data variance than noise, background or baseline changes, the number of chemical components can still be estimated easily by visual inspection of plots of singular values.

(3) Initial estimation of the concentration profiles or of the pure spectra of these species. They can be indistinctly used as initial input to the alternating least squares (ALS) optimization. Depending on the data structure and data selectivity, it is easier to obtain more reliable estimations of either concentration profiles or pure spectra profiles. These are derived from techniques based on either the detection of 'purest' variables [22,23] or evolving factor analysis [24,25]. Local rank analysis, by techniques such as fixed size moving window evolving factor analysis [26], can be used to detect both selectivity and windows or regions of species existence or non- existence.

(4) Alternating least-squares (ALS) optimization of the concentration profiles and of the pure spectra based on compliance with the Beer's law. Depending on the nature and structure of the data, different constraints can be applied during the optimization. Some of the constraints have been specifically developed for second-order data, in order to take advantage of this kind of data. These constraints will also depend on whether the optimization is based on concentration or spectra initial estimations.

The constraints which may be applied to the ALS optimization when concentration initial estimations are used (as is mostly the case in the present paper) are:

(i) Concentration profiles and UV-visible spectra of species present in the system must be non-negative.

(ii) Zero concentration windows: this constraint is applied to the initial estimation of the concentration profiles of each species, by keeping to zero the concentration in those regions where the species is known not to be present. In this way possible selective regions and possible regions where a certain species is known to exist and/or not exist can be fixed during the optimization.

(iii) Correspondence between common species in the different data matrices.

(iv) Pure spectra of common species present in different titrations are equal. This requires the titrations to be carried out under the same experimental conditions of temperature and solvent composition.

(v) Concentration profiles of common species present in different titrations can have equal shapes: this condition is fulfilled in the case of acid–base equilibria, where the shape of the species distribution profiles does not depend on the concentration. This condition, however, is not fulfilled in general for multiequilibria reaction based systems [18].

Constraints (iii), (iv) and (v) are specifically applied to second–order data. Both constraints (iv) and (v) are fulfilled for second order data with a trilinear structure (two orders in common) [12]. For second order data with one order in common between matrices either constraint (iv) or constraint (v) is fulfilled, depending on which is the order in common. The use of these constraints implies an improvement of the resolution conditions. If the rank conditions of resolution are achieved for one species in one matrix, then the resolution is also obtained for the same species in

the other matrices (even if resolution conditions did not hold for this species in those matrices). Furthermore, with the second-order multivariate curve resolution approach, the number of possible solutions is also highly limited.

When spectral estimations are used, another important constraint can be applied in addition to constraints (i), (iii), (iv) and (v) described above:

(vi) Keeping the species spectrum of a certain species constant during the ALS optimization; this constraint is applied to improve the resolution and the accuracy of the quantification in those cases where the shape of the spectrum of a certain species is previously well known. The shape of the spectrum of this species is fixed during optimization and only the spectra of the remaining species may be modified. In the present study, this constraint is applied in the quantitative study, for the determination of the analyte in the presence of interferences, using a pure standard for the analyte (this corresponds to the analysis of augmented data matrices such as **[AU; A]**, **[ACU; A]** and **[ACHU; A]**). In such cases, the shapes of the spectra of the protonated and deprotonated analyte can be unambiguously recovered from the standard data matrix and this estimation is thus the best result.

(5) Quantification of the analyte is possible when the proposed ALS optimization method is simultaneously applied to standard and unknown mixture data matrices, all arranged together in an augmented data matrix. When the matrix relative to the unknown sample is analyzed simultaneously with that of a standard, the ratio between the areas of the concentration profiles for a particular species (the protonated or the deprotonated form of a certain nucleic base) in the different samples will give the ratio between the concentrations for this species in the different titrations. Since one concentration (that in the standard) is known, the absolute concentration in the unknown sample can then be evaluated easily. The area under the concentration profile of either the protonated or deprotonated form of the nucleic base of interest is also proportional to the total nucleic base (analyte) concentration (fulfillment of the mass action law). Quantification can then be performed from the concentration profile of either the protonated or deprotonated form. In this paper both concentration profiles are used in the quantification process.

The ALS procedure applied in this paper is implemented in a series of MATLAB subroutines.

4. Results and discussion

4.1. Rank analysis

Rank analysis of the individual and augmented data matrices was carried out by singular value analysis and by Malinowski's indicator function. This study was performed both on pure data and on data with an added random error (standard deviation 0.001 absorbance units).

Table 2 contains the singular values obtained for the different individual data matrices free from random error. Analogous results are obtained when random error is added to the data.

A preliminary aspect to consider in the analyzed data is that the rank of the data matrix obtained in the titration of a single nucleic base (analyte) is two (see Table 2), since during the acid–base titration two species, the protonated and deprotonated forms of the nucleic base, are formed. Therefore, the usual assumption for bilinear data, that the analyte provides a rank one data matrix, fails in this case. In fact this situation is the common one in many reaction-based systems such as equilibrium and kinetic systems [16]. In the literature, systems where a single analyte provides a data matrix with a rank higher than one have been called non-bilinear systems [13] and have traditionally been considered more difficult to resolve and

Table 2
Singular values for the individual data matrices [a], free from random error

A	AU	AC	ACU	AHU	ACHU
35.72	51.91	58.54	74.65	72.52	94.97
4.80	6.85	7.74	8.74	10.47	11.16
— [b]	1.83	0.65	2.73	1.79	3.83
—	—	—	0.44	0.34	0.38
—	—	—	—	—	0.28
—	—	—	—	—	—
Theoretical number of components					
2	4	4	6	6	8

[a] See text for the description of the matrix notation.
[b] Hyphens indicate a zero singular value.

to use for quantitative determinations using current second-order and three-way data analysis methods. However, when the concept of analyte is replaced by the concept of chemical species, a system considered non-bilinear in terms of the analyte becomes bilinear in terms of the species. Bilinear based methods can then also be used for resolution and quantification of the species instead of the analyte. The relation between the species and the analyte concentrations can be established a posteriori for a particular type of reaction. In the case of acid–base reactions, for example, it can easily be shown that the concentration of any of the basic or acidic species formed during the acid–base reaction is directly proportional to the total concentration of the acid or base (analyte).

The results included in Table 2 show that, with the exception of the data matrices with a single nucleic base, the rank of the matrices is lower in all the other cases than the total number of absorbing species present in each of them (rank deficiency). This is explained by the nature of closed reaction systems [27] (i.e. systems in which the total concentration is kept constant). Data matrices of closed systems have been shown to be rank deficient when operations like mean centring or differentiation are applied to them [27]. Furthermore, a rank deficiency in the original data may occur in the specific case of closed systems with reacting mixtures, in which the different concentration profiles are linearly dependent due to the governing equilibrium reactions, when the number of absorbing species is higher than the number of independent reactions [14].

For an example, the four-nucleic base system has been chosen (matrix **ACHU**). The rank of this data matrix is five, meaning that only five independent sources of data variation (factors) are detected. At the beginning of the titration (pH 2) a mixture of the four protonated species is present (see Fig. 3) and described by a single factor. As the titration advances, (i.e. pH is increased) there is the successive formation of the deprotonated forms of the four nucleic bases (see Fig. 3). Each deprotonation process, which is ruled by an equilibrium constant, is related to an independent source of data variation. Therefore, the number of independent factors must be five in agreement with the calculated rank for this matrix. This is also true of all the other studied matrices and it can be stated that for systems containing n different compounds with acid–base behaviour (and thus, n reactions and $2n$ absorbing species), the rank of the corresponding individual data matrix will be $n + 1$.

Two methods have been proposed to achieve the rank augmentation of a rank-deficient data matrix [14]: (1) matrix augmentation by simultaneous analy-

Table 3
Indicator function [21] values ($\times 10^7$) for the column-wise (wavelength) augmented data matrices [a], containing random error (standard deviation 0.001 a.u.)

a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*
138	105	165	123	127	212	154	154	143
38.2	35.7	65.6	67.2	57.7	87.5	83.7	89.1	83.1
15.9	15.4	45.8	41.8	31.8	44.8	47.1	45.9	46.6
1.46	1.18	13.2	13.5	15.9	22.0	25.3	25.9	25.6
1.47	1.19	1.48	8.39	8.32	3.06	15.8	16.1	15.4
1.48	1.20	1.50	1.22	1.23	1.52	1.57	4.97	7.27
1.49	1.22	1.51	1.24	1.24	1.53	1.24	3.24	5.03
1.51	1.24	1.53	1.25	1.26	1.54	1.26	1.28	1.28
1.53	1.26	1.54	1.26	1.27	1.56	1.28	1.29	1.30
Number of significant components								
4	4	5	6	6	6	7	8	8
Theoretical number of components								
4	4	6	6	6	8	8	8	8

a = [AU; A], *b* = [AU; A; U], *c* = [ACU; A], *d* = [ACU; A; C], *e* = [ACU; A; C; U], *f* = [ACHU; A], *g* = [ACHU; A; C], *h* = [ACHU; A; C; H], *i* = [ACHU; A; C; H; U].

[a] See text for the description of the matrix notation.

sis of multiple process runs, (2) matrix perturbation by addition of a single species or of mixtures of them during the process or reaction. The first approach will be applied here: rank augmentation can be achieved by simultaneous analysis of the titration matrix of the mixture of nucleic bases together with the titration matrices of single nucleic bases. Both augmented column-wise (wavelength-wise) or row-wise (pH-wise) data matrices are studied.

Tables 3 and 4 give the results of the indicator function calculated respectively for the column-wise and row-wise augmented matrices in the presence of random error. Analogous results are obtained in the absence of error. The indicator function gives a minimum value for the number of independent detected chemical species or chemical rank.

For all the systems studied there is an increase in rank when an augmented column-wise (wavelength-wise) matrix is considered. On the other hand, when an augmented row-wise (pH- wise) matrix is considered its rank does not increase, independently of the number of single nucleic base titration matrices included in the augmented data matrix.

The different effects of row-wise and column-wise data augmentation on the rank of the corresponding augmented matrices must be related to the particular structure of the data analyzed in the present work.

Each individual data matrix **D** can be considered as the product of two matrices, $\mathbf{D} = \mathbf{C}\mathbf{S}^T$ (see above). Since for each nucleic base the concentrations of protonated and deprotonated forms are constrained by closure and mass action law, the columns of **C** cannot change independently. For each nucleic base, if one of the two concentration profiles is known the other is automatically defined. However, each species is characterized by an independent spectrum and consequently there is no linking relationship between the rows of $\mathbf{S}^T$.

It can be demonstrated that when a matrix **Z** can be decomposed into the product of two matrices, $\mathbf{Z} = \mathbf{PQ}$, then $\text{rank}(\mathbf{Z}) \leq \min(\text{rank}(\mathbf{P}), \text{rank}(\mathbf{Q}))$. In the case under study the rank of the experimental matrix must be equal or lower to that of the matrix of concentration profiles since the matrix of spectra is always of full rank. Similar conclusions were obtained by Amrhein et al [14]. In fact, rank analysis of the matrix of concentration profiles that corresponds to a mixture of the four nucleic bases and contains eight concentration profiles, gives a value of five. This value coincides with that obtained for the experimental data matrix **ACHU** (see Table 2).

Row-wise and column-wise augmented data matrices can also be decomposed as the product of a matrix of concentration profiles and a matrix of spectra (Fig. 4) and consequently rank of the augmented data matrix is related to that of the matrix of concentration profiles. Row-wise matrices are decomposed in an individual matrix **C** which is rank-deficient due to the dependence between species (for a mixture of the four nucleic bases its rank is five). On the con-

Table 4
Indicator function [21] values ($\times 10^7$) for the row-wise (pH) augmented data matrices [a], containing random error (standard deviation 0.001 a.u.).

a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*
547	492	660	597	547	820	720	685	633
177	215	204	173	213	340	297	330	310.1
6.21	6.29	55.0	59.4	53.4	63.5	64.2	60.6	58.6
6.44	6.55	6.52	6.62	6.61	25.9	22.7	25.0	32.0
6.70	6.83	6.78	6.90	6.90	6.85	6.96	6.95	6.94
6.97	7.13	7.07	7.21	7.22	7.13	7.26	7.27	7.27
Number of significant components								
3	3	4	4	4	5	5	5	5
Theoretical number of components								
4	4	6	6	6	8	8	8	8

a = [AU, A], *b* = [AU, A, U], *c* = [ACU, A], *d* = [ACU, A, C], *e* = [ACU, A, C, U], *f* = [ACHU, A], *g* = [ACHU, A, C], *h* = [ACHU, A, C, H], *i* = [ACHU, A, C, H, U].

[a] See text for the description of the matrix notation.

trary, column-wise augmented matrices are decomposed into an augmented **C** matrix that consists on independent matrices of concentration profiles. Therefore, the rank of the augmented **C** matrix is increased when new information is included in the different individual concentration matrices included in the augmented column-wise **C** matrix. As an example, the rank of the augmented **C** matrix that contains the rank-five matrix of the eight concentration profiles of the four nucleic bases and the rank-two matrix of the two concentration profiles (protonated and deprotonated) of one of the nucleic bases, is six. The rank is increased by one because one additional concentration profile (a new column in matrix **C**) is needed to explain the change in concentration of the protonated form of the nucleic base which is fully resolved in the single nucleic base rank-two data matrix and unresolved in the four nucleic base rank-five data matrix. This is in agreement with the results obtained in the analysis of augmented column-wise data matrices such as [**ACHU**; **A**].

There is rank augmentation when the individual matrices included in the column-wise augmented one correspond to mixtures containing different proportions of the nucleic bases. In this paper mixtures containing a single nucleic base have been used in order to get rank augmentation, but analogous results are obtained if different mixtures containing the same nucleic bases are analyzed, provided that the proportion between the bases is different in each mixture. The inclusion of each of these individual matrices increases the rank of the previous matrix by one. It can be deduced that the rank of the augmented column-wise matrices is equal to $r + k$, for $k < n$, where r is the rank of the mixture of nucleic bases titration matrix, k is the number of different single nucleic base titration matrices added to the system and n is the total number of nucleic bases contained in the mixture. When the information concerning $n - 1$ titrations of single different nucleic bases is included in the augmented data matrix, this matrix becomes a full rank.

4.2. Species resolution

4.2.1. Individual data matrices

(a) Individual data matrices containing a single nucleic base: [**A**], [**H**], [**C**] or [**U**].

The resolution (i.e. recovery of spectral and concentration profiles) of the individual data matrices containing one of the four nucleic base acid–base titrations (full rank two matrices) can be achieved in all cases without ambiguities because full selectivity is present in such systems [10]: at the most acidic pH values only the acidic form is present and at the most basic pH values only the basic form is present. The deduction of the species distribution and spectra was achieved in the four cases by application of the ALS optimization.

(b) Individual data matrices containing a mixture of nucleic bases: [**AC**], [**ACU**], [**ACHU**]...

For individual data matrices containing a mixture of two or more nucleic bases, the total absence of selectivity either in the pH or in the wavelength direction (Fig. 1, Fig. 2 and Fig. 3) and the existence of a rank deficiency in the data matrix, hinder resolution of all the species when these matrices are individually analyzed.

4.2.2. Augmented data matrices

(a) Full-rank augmented column-wise data matrices: [**AC**; **A**], [**ACU**; **A**; **C**], [**ACHU**; **A**; **C**; **H**]...

When an augmented full-rank matrix is analyzed, species resolution is accomplished for all the species present in the system, including for those species whose nucleic base individual titration is not included in the augmented matrix. As an example, a complete recovery of the spectra and concentration profiles of both adenine and cytosine is accomplished from the analysis of the augmented matrix [**AC**; **A**], with correlations higher than 0.9999 between the theoretical and ALS calculated spectra and concentration profiles. Analogous results are obtained from the analysis of any other full-rank augmented matrix.

(b) Rank deficient augmented column-wise data matrices: [**ACU**; **A**], [**ACHU**; **A**], [**ACHU**; **A**; **C**]...

A correct resolution is now achieved only for those species which are simultaneously present in the matrix of the mixture and in the matrix of the single nucleic base titration. As an example, when the augmented matrices [**ACU**; **A**] and [**ACHU**; **A**] are analyzed, a good resolution (with correlations higher than 0.9999) is obtained only for the protonated and deprotonated forms of adenine. Fig. 5 shows the concentration profiles obtained from the analysis of the augmented matrix [**ACU**; **A**].

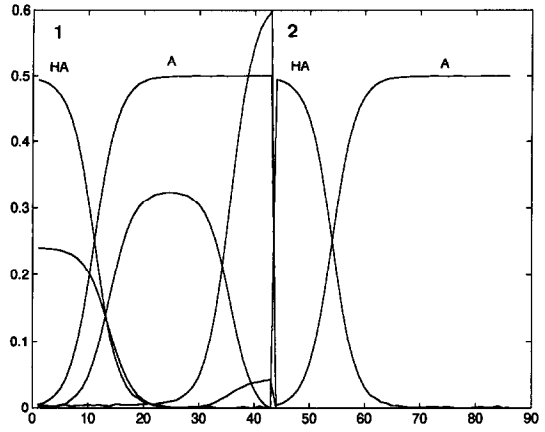

Fig. 5. Concentration profiles obtained from the analysis of the augmented matrix [**ACU**; **A**]. Titration 1: Concentration profiles for matrix **ACU** and titration 2: concentration profiles for matrix **A**. HA and A refer to the protonated and deprotonated forms of adenine, respectively.

4.3. Quantification

The study of quantification with the ALS multivariate curve resolution method in systems with rank deficiency was exhaustively analyzed for different mixtures of nucleic bases. The quantification of the analyte in the unknown mixture is performed by the ALS simultaneous analysis of standards and unknowns.

4.3.1. Quantification of single analyte mixtures

Single nucleic base titrations give single analyte matrices of rank two, since the acid–base titration gives two species, the acidic and basic forms of the nucleic base. When this matrix is analyzed in comparison with that of a standard (containing a known concentration of the nucleic base to be determined), quantification is perfect (analysis of augmented matrices such as $[\mathbf{A}_{c1}; \mathbf{A}_{c2}]$). This is a very simple system, in which, since there is no rank deficiency, a complete recovery of both the concentration and the spectral profiles is accomplished and quantification is achieved by comparison of areas of concentration profiles, from either the acidic or basic form of the nucleic base.

4.3.2. Quantification of binary nucleic base mixtures

When a single mixture data matrix of two nucleic bases is analyzed, no quantification is possible. In-

Table 5
Percentage of prediction error for several mixtures of nucleic bases, deduced using either concentration or spectra profiles as initial estimates. The concentration of all the nucleic bases in all the matrices is $5x10^{-5}$ M

Analyzed matrix	Initial concentration estimates			Initial spectra estimates		
	% prediction error [a]		% fitting error [b]	% prediction error [a]		% fitting error [b]
	HX [c]	X		HX	X	
[AU; A]	0.13	< 0.1	0.159	0.14	< 0.1	0.165
[AU; U]	0.14	< 0.1	0.185	< 0.1	< 0.1	0.185
[AC; A]	< 0.1	< 0.1	0.136	< 0.1	0.47	0.146
[AC; C]	< 0.1	< 0.1	0.147	0.16	0.76	0.153
[ACU; A]	< 0.1	< 0.1	0.141	0.25	1.15	0.146
[ACU; C]	< 0.1	< 0.1	0.154	0.12	0.97	0.138
[ACU; U]	< 0.1	< 0.1	0.145	0.23	< 0.1	0.149
[ACHU; A]	0.14	0.16	0.124	0.14	1.53	0.198
[ACHU; C]	< 0.1	0.18	0.197	< 0.1	2.77	0.176
[ACHU; H]	< 0.1	0.10	0.130	6.10	0.43	0.129
[ACHU; U]	< 0.1	0.53	0.134	15.3	1.30	0.174

[a] The percentage of prediction error is calculated as follows: % P.E. = $(c_i^* - c_i)/c_i * 100$, where c_i^* and c_i refer respectively to the calculated and real concentration of the nucleic base that is quantified.

[b] The percentage of fitting error is calculated as follows: % F.E. = $\sqrt{\Sigma_i \Sigma_f (d_{if}^* - d_{if})^2 / \Sigma_i \Sigma_f d_{if}^2} * 100$, where d_{if}^* and d_{if} refer respectively to the calculated and real absorbance data.

[c] HX and X refer respectively to the protonated and deprotonated forms of the nucleic base, which are present simultaneously in both individual matrices.

stead, an augmented data matrix is built, where the information of one of the two nucleic bases is given at least at two independent conditions. In this case the rank-deficiency is completely eliminated in the augmented matrix (see rank analysis and species resolution above) and the ambiguities in the recoveries of the species profiles for all the species in the different data matrices are almost solved. For this reason quantification is also achieved simply by comparing areas of concentration profiles of the same species in the different data matrices. A mixture of adenine and uracil (matrix **AU**) was chosen for the study of quantification of one of these compounds in the presence of the other, which acted as an interference. Firstly, uracil was the compound to be determined and adenine the unknown interferent. A data matrix of the unknown mixture was analyzed simultaneously with a data matrix of pure uracil of known concentration (augmented matrix [**AU**; **U**]). While the mixture sample matrix **AU** has a rank of three, the addition of the information of the standard sample matrix **U** in the column-wise augmented data matrix [**AU**; **U**] increases the rank by one. In this case rank deficiency is completely broken with the added information of the standard.

Four spectroscopically different species are considered in the system under study. The next step was the selection of the initial estimators for the concentration and spectra profiles of these species for the ALS optimization procedure. Both concentration and spectral initial estimations are used and compared. For uracil, the spectra or concentration profiles deduced from the individual analysis of matrix **U**, are considered the initial estimators. Considering that no information is known about the presence or responses in either order (concentration and spectra profiles) of the interferent, adenine, the initial concentration estimators were obtained by evolving factor analysis [24,25] of the sample matrix and its initial spectral estimators were evaluated by analysis of the purest variables [22,23] present in the sample matrix. The results are included in Table 5. They show correct quantification in both cases, independently of the initial estimators used. Analogous results were obtained when adenine was the analyte to be quantified and uracil was the unknown interference (analysis of the augmented matrix [**AU**; **A**]) and also for the mixtures of adenine and cytosine, which have very similar pK_a values (analysis of augmented matrices [**AC**; **A**] and [**AC**; **C**]). These results can be extended to the analysis of any other binary mixture of nucleic bases with identical results.

4.3.3. Quantification of tertiary nucleic acid–base mixtures

The next possible situation is the mixture of three nucleic bases. A mixture of three nucleic bases gives a rank of four and is therefore two units rank deficient. Matrix augmentation by the addition of two data matrices corresponding each one to the titration of one of the nucleic bases included in the mixture breaks rank deficiency and gives an augmented rank 6 matrix, allowing the resolution and recovery of all the species present, even of those not included in the matrices of titrations of individual nucleic bases. The results of this case, where rank deficiency is completely broken (adding two single nucleic base standard matrices to the three nucleic base mixture), are not given here since they are equivalent to those of the analysis of the binary mixture with one standard.

More interesting is the study of the tertiary mixture samples when only the information about one of the nucleic bases (the analyte to be determined) is included (matrix of single nucleic base) and the other nucleic two bases (four species) are the unknown interferences. In this case, rank deficiency is not eliminated (there are six species but the rank of the augmented matrix will be only five). Consequently, only the recovery of the concentration and spectral profiles of the analyte (present in both matrices) is assured. The number of species to be considered in the ALS treatment will be a maximum of five.

The results obtained from the analysis of the mixture of equimolar concentrations of adenine, cytosine and uracil (matrix [**ACU**]) when only one of these nucleic bases is included as standard (systems [**ACU**; **A**], [**ACU**; **C**] and [**ACU**; **U**]) are given in Table 5. A good quantification is achieved for all the nucleic bases present in the mixtures when either concentration profiles or spectra are used as initial estimations. These results show again that although the total resolution and quantification of rank-deficient systems is not possible, resolution and quantification of the common analyte in the simultaneously analyzed matrices is still possible.

Table 6
Percentage of prediction error for three mixtures that contain different concentration ratios of the four nucleic bases. Results obtained using concentration profiles as initial estimators. The concentration of the nucleic bases in the standard matrices is $5x10^{-5}$ M

Concentration ($\times 10^4$ M) [a]				[ACHU; A]		[ACHU; C]		[ACHU; H]		[ACHU; U]	
				% P.E [b]		% P.E		% P.E		% P.E	
A	C	H	U	HX [c]	X	HX	X	HX	X	HX	X
0.5	0.5	2.5	0.5	0.11	0.11	< 0.1	0.15	4.85	0.30	< 0.1	0.63
0.5	2.5	2.5	0.5	0.15	0.13	0.40	2.06	4.72	0.30	< 0.1	0.84
2.5	2.5	0.5	2.5	0.50	1.96	0.45	1.38	< 0.1	0.38	10.7	0.20

[a] Real concentration (M × 10000) for the four nucleic bases in the **ACHU** matrix.
[b] % P.E. is the percentage of prediction error (see Table 5).
[c] HX and X refer respectively to the protonated and deprotonated forms of the nucleic base, which are present in both individual matrices.

4.3.4. Quantification of quaternary nucleic acid–base mixtures

A mixture with equimolar concentrations of adenine, cytosine, hypoxanthine and uracil was studied, using both concentration and spectra profiles as initial estimations. As for tertiary systems, total resolution and quantification is achieved for full rank augmented data matrices (i.e. augmented matrices such as the full rank [**ACHU**; **A**; **C**; **H**] matrix, with 8 species, 3 analytes and 1 interferent). These results are obtained by applying the same principles and methods described above.

More interesting again is the analysis of the rank-deficient augmented matrices of this quaternary system (for instance when only one nucleic base titration is included in the augmented matrix, i.e. [**ACHU**; **A**] rank six augmented matrix). The results are shown in Table 5. In this more complex case, quantification is better when concentration profiles rather than spectra are used as initial estimations in the ALS optimization procedure, since a selectivity constraint (see description of constraints in the description of the ALS optimization procedure) can be applied only in the pH direction and not in the wavelength direction, where the different species absorb in all the wavelength range. In the pH direction there is always a certain selective pH range for each species, acidic or basic, which can be exploited during the ALS optimization. For the four nucleic base system, different mixtures in different concentration ratios were also analyzed, using the concentration profiles as initial estimations in the ALS optimization and applying selectivity constraints. The results are included in Table 6. It can be observed that when the concentration of the nucleic base in the mixture is different from that in the standard (5×10^{-5} M for all standards), worse results are obtained in certain cases. This is clearly related to the higher concentration extrapolation that must be performed using a single standard calibration method. Quantification of the four bases in the different mixtures was also performed using standards with a concentration 2.5×10^{-4} M giving a prediction error lower than 1% for all the nucleic bases present in the mixtures at higher concentration and worse results for the nucleic bases whose concentration was 5×10^{-5} M.

5. Conclusion

Rank analysis of the data matrices obtained from spectrophotometric acid–base titration of nucleic bases showed that whenever a single nucleic base is analyzed, a full rank data matrix is obtained with rank of 2 since two species are formed during the titration: the deprotonated and protonated forms of the nitrogen base. These two species are linked by a concentration closure and by the constraint of the mass action law. On the other hand, when mixtures of nucleic bases are analyzed, rank-deficient matrices are obtained. Rank augmentation is achieved when augmented matrices in the spectral order are analyzed whereas matrix augmentation in the pH order does not provide any rank increase. With augmented column-wise matrices there is a decrease of the rank-deficiency, and eventually full rank can be achieved. This depends on what individual matrices are included in the augmented one. Thus, the selection of

the experiments to carry out in order to break the rank deficiency is a matter of choice and it is related to the experimental design.

Rank deficiency is a very important hindrance to resolution of this kind of systems. In order to achieve a particular species resolution in an unknown mixture the information relative to this species can be added in a column-wise augmented data matrix. As it could be expected, a total resolution is only achieved when full-rank matrices, either augmented or not, are analyzed.

On the other hand, quantification was carried out succesfully for nucleic bases in unknown mixtures (with unknown and uncalibrated interferences), with independence of whether the data matrices were full rank or rank-deficient. Rank-deficiency has proved to be of minor importance for quantification purposes by using the proposed second-order multivariate curve resolution method.

References

[1] C.W. Gehrke, K.C. Kuo, Ribonucleoside analysis by reversed-phase high-performance liquid chromatography, J. Chromatogr. 471 (1989) 3–36.

[2] R.C. Simpson, P.R. Brown, High-performance liquid chromatographic profiling of nucleic acid components in physiological samples, J. Chromatogr. 379 (1986) 269–311.

[3] A.S. Cohen, S. Terabe, J.A. Smith, B.L. Karger, High-performance capillary electrophoretic separation of bases, nucleosides and oligonucleotides: Retention manipulation via micellar solutions and metal additives, Anal. Chem. 59 (1987) 1021–1027.

[4] T.J. Kasper, M. Melera, P. Gozel, R.G. Brownlee, Separation and detection of DNA by capillary electrophoresis, J. Chromatogr. 458 (1988) 303–312.

[5] H.N. Cong, O. Bertaux, R. Valencia, T. Becue, T. Fournier, D. Biou, D. Porquet, Separation and characterization of the main methylated nucleobases from nuclear, cytoplasmic and poly(A)(+) RNA by high-performance liquid-chromatography and mass-spectrometry, J. Chromatogr. B-Biomed. Appl. 661 (1994) 193–204.

[6] Z.X. Zhao, J.H. Wahl, H.R. Udseth, S.A. Hofstadler, A.F. Fuciarelli, R.D. Smith, Online capillary electrophoresis electrospray-ionization mass-spectrometry of nucleotides, Electrophoresis 16 (1995) 389–395.

[7] R. Tauler, A. Izquierdo-Ridorsa, E. Casassas, Comparación de métodos de calibraci ón multivariante aplicados al estudio espectrofotométrico de mezclas de bases purínicas y pirimidínicas, An. Química 87 (1991) 571–579.

[8] W. Lindberg, B. Kowalski, Evaluation of potentiometric acid–base titrations by partial-least-squares calibration, Anal. Chim. Acta 206 (1988) 125–135.

[9] R. Tauler, A. Izquierdo-Ridorsa, E. Casassas, Simultaneous analysis of several spectroscopic titrations with self-modelling curve resolution, Chemom. Intell. Lab. Syst. 18 (1993) 293–300.

[10] R. Tauler, A. Smilde, B.R. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemom. 9 (1995) 31–58.

[11] E. Sanchez, L.S. Ramos, B.R. Kowalski, Generalized rank annihilation method. I. Application to liquid chromatography-diode array ultraviolet detection data, J. Chromatogr. 385 (1987) 151–164.

[12] E. Sanchez, B.R. Kowalski, Tensorial resolution: A direct trilinear decomposition, J. Chemom. 4 (1990) 29–45.

[13] A.K. Smilde, Y. Wang, B.R. Kowalski, Theory of medium rank second order calibration with restricted Tucker models, J. Chemom. 8 (1994) 21–36.

[14] M. Amrhein, B. Srinivasan, D. Bonvin, M.M. Schumacher, On the rank deficiency and rank augmentation of the spectral measurement matrix, Chemom. Intell. Lab. Syst. 33 (1996) 17–33.

[15] L. Norgaard, C. Ridder, Rank annihilation factor analysis applied to flow injection analysis with photodiode-array detection, Chemom. Intell. Lab. Syst. 23 (1994) 107–114.

[16] R. Tauler, A.K. Smilde, J.M. Henshaw, L.W. Burgess, B.R. Kowalski, Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. 2. Chemical speciation using multivariate curve resolution, Anal. Chem. 66 (1994) 3337–3344.

[17] R. Tauler, A. Izquierdo-Ridorsa, R. Gargallo, E. Casassas, Application of a new multivariate curve resolution procedure to the simultaneous analysis of several spectroscopic titrations of the copper(II)–polyinosinic acid system, Chemom. Intell. Lab. Syst. 27 (1995) 163–174.

[18] E. Casassas, R. Gargallo, A. Izquierdo-Ridorsa, R. Tauler, Application of a new multivariate curve resolution procedure to the study of the acid–base and copper(II) complexation equilibria of polycytidylic acid, React. Polym. 27 (1995) 1–14.

[19] E. Casassas, R. Tauler, I. Marqués, Interactions of H^+ and Cu(II) ions with poly(adenylic acid): Study by factor analysis, Macromolecules 27 (1994) 1729–1737.

[20] E. Casassas, R. Gargallo, I. Giménez, A. Izquierdo- Ridorsa, R. Tauler, Application of an evolving factor analysis-based procedure to speciation in the copper(II)–polyuridylic acid system, Anal. Chim. Acta 283 (1993) 538–547.

[21] E.R. Malinowski, Factor Analysis in Chemistry, Wiley, New York, 2nd ed., 1991.

[22] W. Windig, Self-modeling mixture analysis of spectral data with continuous concentration profiles, Chemom. Intell. Lab. Syst. 16 (1992) 1–23.

[23] W. Windig, D.A. Stephenson, Self-modeling mixture analysis of 2nd-derivative near-infrared spectral data using the Simplisma approach, Anal. Chem. 64 (1992) 2735–2742.

[24] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbuhler, Cal-

culation of equilibrium constants from multiwavelength spectroscopic data. III. Model free analysis of spectrophotometric and ESR titrations, Talanta 32 (1985) 1133–1139.

[25] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbuhler, Calculation of equilibrium constants from multiwavelength spectroscopic data. IV. Model free least squares refinement by use of evolving factor analysis, Talanta 33 (1986) 943–951.

[26] H.R. Keller, D.L. Massart, Peak purity control in liquid chromatography with photodiode-array detection by a fixed size moving window evolving factor analysis, Anal. Chim. Acta 246 (1991) 379–390.

[27] R.J. Pell, M.B. Seasholtz, B.M. Kowalski, The relationship of closure, mean centering and matrix rank interpretation, J. Chemom. 6 (1992) 57–62.