ELSEVIER

Chemometrics and Intelligent Laboratory Systems 29 (1995) 213–221

Chemometrics and intelligent laboratory systems

Detection and quantification of embedded minor analytes in three-way multicomponent profiles by evolving projections and internal rank annihilation

Bjørn Grung, Olav M. Kvalheim *

Department of Chemistry, University of Bergen, N-5007 Bergen, Norway

Received 9 January 1995; accepted 18 April 1995

Abstract

Modern analytical instrumentation often leads to data arrays with more than two dimensions or directions. Such N-way data ($N > 2$) needs special resolution methods for optimising the amount of analytical information. In this work, a new interactive method designed to work with three-way data is presented. The method, evolving projections by optimised search (EPOS), presents a combined graphical and numerical way of resolving a three-way data array into the analytical profiles of the pure analytes. The method involves an internal rank annihilation step, which can be performed in several ways. The graphic interactive procedure used in this work compares favourably with Lorber's noniterative rank annihilation method. Thus, our method is significantly better for resolution of analytes with low relative concentration, especially in the presence of heteroscedastic noise. The EPOS method is tested on several simulated data sets to assess its performance. A peak purity example is carried out to show a case where two-way methods are unable to provide a unique solution, whereas EPOS gives correct results.

Keywords: Embedded minor analytes; Three-way multicomponent profiles; Evolving projections; Internal rank annihilation

1. Introduction

Multicomponent problems can be classified as white, grey, and black depending on the amount of a priori information available about the coexisting analytes [1]. The most difficult mixture problem from the analyst's point of view is the one posed by the black system, i.e. a system where neither the number of analytes present, nor their identity are known. The problem to be solved for a black system is thus (i) to decide the number of analytes, (ii) to quantify the analytes, and (iii) to identify the analytes.

For mixtures, multidetection chromatography, such as liquid chromatography with diode array detection and gas chromatography with infrared or mass spectral detection, represents a common approach to attack the resolution and quantification problem for black systems. This works well in many cases. Problems arise, however, if analytes have similar chromatographic retention time and spectra. In such cases, chemometrics methods for curve resolution may be helpful [1]. Thus, methods such as SIMPLISMA [2–

* Corresponding author.

SSDI 0169-7439(95)00038-0

4], iterative target transformation factor analysis (IT-TFA) [5–8], and evolving factor analysis (EFA) [9,10] have all proven valuable for the analysis of black multicomponent systems.

Recently, the heuristic evolving latent projections (HELP) method was developed [11]. This technique has been used in numerous applications [12–15]. The HELP method utilises the selective regions (regions with signal from only one analyte), the zero-concentration regions (regions where one analyte is absent) [10], and zero-component regions (regions where none of the analytes contributes to the signal) [11] to resolve the profiles of the pure analytes uniquely. These definitions will be used in this paper. In case of no selective regions, such regions may be created by differentiation [16–18]. Thus, if the instrumental profile at a specific retention time or wavelength contains signal from two analytes, the contribution from one of the analytes is removed by differentiation if the signal from that analyte has a local maximum at the specified retention time or wavelength. However, profiles for black systems where one chromatographic peak is embedded in another may be impossible to resolve uniquely by two-way methods. This situation arises if the peak maximum of the major analyte does not overlap with the peak from the embedded analyte and the spectral direction contains no selective information. Such a situation is depicted in Fig. 1. A unique solution for such cases demands three-way data.

Some instrumental techniques, such as liquid chromatography connected to excitation–emission fluorescence spectroscopy, produce three-way data for a single sample. Furthermore, hyphenated techniques can lead to three-way data when several samples of similar composition are analysed and compared.

A common problem in the pharmaceutical industry is that of detecting and quantifying possible impurities in drugs [12]. These impurities often appear as hidden minor peaks in a chromatogram. As discussed above, two-way curve resolution techniques may not be optimal for extracting the information from such data structures. Thus, there is a need for methods devoted solely to handling these structures if the information they contain is to be satisfactorily extracted.

We propose an interactive latent projection technique called evolving projections by optimised search (EPOS) to solve problems of this kind for three-way

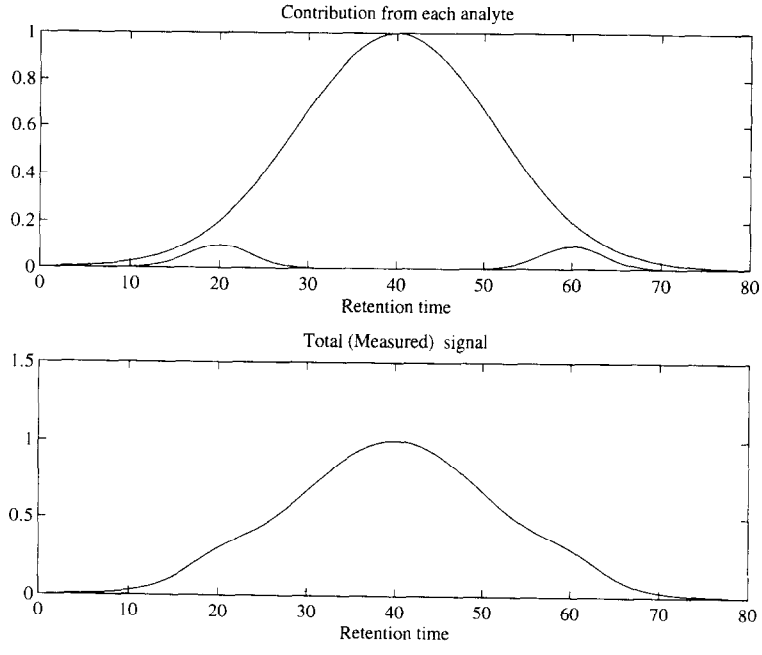

Fig. 1. The problem of peak purity. The large analyte completely overlaps the two minor ones.

data. Examples showing the potential of the method for resolving such cases, will be carried out in detail. The aim of the method is thus to detect the number of chemical components in the sample(s), to resolve their analytical profiles for the purpose of identification, and, to find their relative concentrations under the assumptions of equal total response.

2. Theory

This section starts by providing the formulas for decomposition of 2- and 3-way data in terms of the analytical profiles of the contributing analytes. We then proceed with a brief description of procedures for determination of rank trends in the evolving multicomponent profile. The rank information is then utilised for quantification and resolution through a rank annihilation step.

Two basic assumptions must be fulfilled in order for our procedure to work: (i) the data must be bilinear and (ii) selective regions must exist for at least one analyte.

2.1. 2- and 3-way decomposition

The decomposition formula

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

where $\mathbf{C}$ frequently contains chromatographic profiles and $\mathbf{S}$ spectral profiles of the pure analytes in the multicomponent profile represented by the data matrix $\mathbf{X}$, is central in all two-way resolution methods. The superscript T denotes transposition and $\mathbf{E}$ is a noise matrix. Equation 1 highlights the bilinear structure of the data. If either $\mathbf{C}$ or $\mathbf{S}$ is found, a straightforward least-squares calculation resolves the other set of profiles.

A three-way data array needs three different sets of resolved profiles for the pure analytes, e.g. spectra, chromatograms and concentration profiles, for its decomposition. Quite analogously to Eq. (1) for two-way data, the decomposition into profiles of pure analytes for a three-way array can be written as

$$\underline{\mathbf{X}} = \sum_{a=1}^{A} \mathbf{c}_a \otimes \mathbf{m}_a \otimes \mathbf{s}_a + \underline{\mathbf{E}} \tag{2}$$

where $\mathbf{m}_a$ denotes the third type of profile. Underlining and bold imply a three-way matrix. The symbol $\otimes$ indicates outer products.

The goal of resolving the data into profiles of the pure analytes using EPOS can be divided into three sub tasks: (i) establishing an overview of the local rank (number of analytes) throughout the data set; (ii) utilisation of this information to obtain all but one set of profiles for all the pure analytes; and, (iii) matrix multiplications (and possibly normalisation) to acquire the last profile set. A simple rearrangement of Eq. (2) subsequently produces the final profile. Eq. (3) shows this reordering for a situation where $\mathbf{M}$ is the unknown.

$$\mathbf{M}\underline{\mathbf{I}} = (\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\underline{\mathbf{X}}\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1} \tag{3}$$

Here $\underline{\mathbf{I}}$ is a three-way identity matrix of dimensions $A \times A \times A$.

2.2. Rank mapping by slicing

Information about number of analytes, i.e. the chemical rank [13], especially location of selective regions in the multicomponent profile $\mathbf{X}$, is crucial for the resolution of profiles acquired from evolving systems. For two-way multicomponent profiles this information is usually expressed in the form of a rank map obtained by moving a window systematically through the data [19–21]. A three-way generalisation of the moving window techniques would lead to a prohibitive number of possible sub matrices to be examined. Furthermore, selectivity, which is crucial for unique resolution by EPOS, is usually not equally distributed in all directions. Instead, we have proposed procedures whereby the data array is either collapsed in one direction (summing matrices element-wise in the chosen direction) or sliced into complete two-way matrices one direction at-the time. The resulting matrices are subsequently examined by a variety of graphic procedures [22].

Rank mapping based upon collapsing or slicing takes into account that it is not necessary to derive a rank map of the complete multicomponent profile. Rather, the task is to obtain a clear picture of the *rank trend*. Thus, whether the exact start of an interesting region is, e.g. at retention time 50 or 53 is usually of

no importance. The crucial point is to make sure that we are able to pick slices representing unambiguously all the different rank situations in the multicomponent profiles.

2.3. Internal rank annihilation

The rank map provides an overview of the analytes' elution patterns. The next step is to use this information to obtain $N-1$ profiles for every analyte. In our case N is equal to 3, so two sets of profiles, e.g. spectra and concentration profiles, should be extracted.

Unless the experiments that produce the data are of particularly poor quality, there will always be at least one selective matrix (one retention time) for at least one analyte, either in the beginning or end of the retention interval for an unresolved peak cluster. The first score and loading vector from PCA of this matrix provides two of the profiles of this analyte. Multiplying these two profiles together column by row yield a matrix $\mathbf{X}_1$ containing only information about this analyte. This procedure removes random noise.

Next, we use the rank map to select a retention time where the signal is composed of this analyte together with another analyte. The matrix at this retention time is called $\mathbf{X}_{12}$. In order to extract the pure profiles for the second analyte from $\mathbf{X}_{12}$, the contribution from the first analyte needs to be subtracted. The task at hand is stated in Eq. (4):

$$\mathbf{X}_2 = \mathbf{X}_{12} - k\mathbf{X}_1 \tag{4}$$

In order to obtain the matrix $\mathbf{X}_2$ with contribution only from the second analyte, the estimate for k can be determined by rank annihilation. Rank annihilation was first proposed by Ho et al. [23–25]. Lorber proposed a noniterative approach [26], which was recently found inferior to a graphic iterative and interactive approach developed by the present authors [27]. Traditionally, one of the primary conditions for performing rank annihilation is that standards of the analytes to be quantified are available. For three-way data arrays, resolved standard profiles can be extracted directly from selective regions in the data. Thus, for peak purity problems our proposed solution can be described as an internal rank annihilation performed by means of an interactive graphic search (IRA).

Our way of carrying out the IRA step is to start with two guesses for k, a lower concentration limit $k=0$ and an upper concentration limit, e.g. $k=10$. The two matrices produced by the subtractions $\mathbf{X}_{12} - k\mathbf{X}_1$ are then decomposed by PCA, and the resulting score and loading vectors from the two major principal components are plotted. If one of our initial guesses is correct, the corresponding score and loading plots reveal structure only in the first principal component, and k equals the guess we made. This situation is of course not likely to arise, and so the search proceeds. A fast procedure for this search is to halve the interval between the lower and upper limit on k after each try.

The IRA procedure is repeated until an acceptable estimate is found. One then continues to resolve the profiles for other analytes. In case of overlap among more than two analytes at the time, so-called stripping techniques may be necessary, either by removing the contribution of an analyte using Eq. (4) or by means of orthogonal projections [18]. After this step, all but one of the N profile sets are completely resolved.

IRA has been found to compare favourably with Lorber's direct procedure when the concentration of the minor analyte decreases and the level of heteroscedasticity of noise increases [27]. If the noise level is high, a more accurate solution of Eq. (4) can be achieved by using the reconstructed $\mathbf{X}_1$ and $\mathbf{X}_{12}$ from PCA on the collapsed matrices, i.e. the two matrices obtained by summing in one direction all the selective matrices having contributions from analyte 1, and, similarly, all the matrices with significant contributions from both analytes 1 and 2.

2.4. Resolution of the final set of profiles

The last set of pure analyte profiles is easily calculated by means of eq. 3. The procedure used in this step varies according to the data characteristics. For samples containing the same analytes in varying proportions and analysed by multidetection chromatography, the chromatographic and spectral profiles are normalised. Assuming the same chromatographic and spectroscopic response for all chemical components, the third set of profiles then displays the correct relative concentrations of the analytes. In order to obtain absolute concentrations, one needs standards of the analytes.

The resolved profiles can also be used for identification purposes, whereafter quantification using standard samples can be performed.

3. Experimental

The EPOS method has been tested on a number of data sets, which were all simulated by means of MATLAB. Spectra and chromatograms were generated as sums of Gaussian peaks. An example is shown in detail (see results and discussion), along with tabulated results for some other data sets. Every data set is simulating multidetection chromatographic profiles for several samples of the same two analytes. The dimensions are $10 \times 80 \times 80$. The concentration of the minor analyte was varied between 1% and 15%. In all data sets, the minor analyte is completely embedded by the major one in the chromatographic direction, making it a difficult problem to solve with conventional curve resolution methods. Both homoscedastic and heteroscedastic noise patterns were added. The size of the noise relative to the maximum total signal is around 0.03% for these data. All quantitative information is inherent in the sample direction, meaning that the response level of the chromatograph and spectrometer is assumed equal for the analytes. Table 1 presents information on the data sets. As the next section will show, EPOS produces excellent results for all these systems.

Table 1
The design of the data sets

Data set	Range of relative concentration of minor analyte (in percent)	Type of noise
1	4–13	Homoscedastic
2	4–13	Heteroscedastic
3	1–3	Homoscedastic
4	1–3	Heteroscedastic

4. Results and discussion

Chromatograms, spectra and concentration profiles for the two analytes in the first data set are

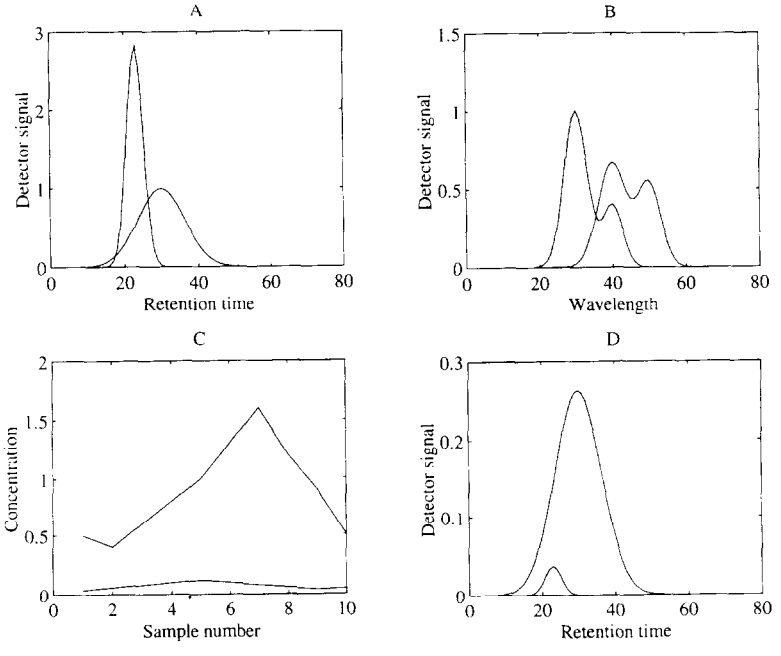

Fig. 2. An overview of the data set analysed in this work: (a) shows noise free chromatographic profiles normalised to constant sum; (b) the noise free spectral profiles normalised to constant sum, while (c) depicts the concentration profiles; (d) illustrates chromatographic profiles averaged with respect to samples. The minor analyte is completely overlapped by the larger one.

shown in Fig. 2(a)–(c). Fig. 2(d) depicts the average amount of the analytes recorded at the chromatographic detector with time. This example illustrates the situation of embedded minor peaks.

Slicing the three-way data in the retention time direction and using the procedures developed in [22] to examine the rank changes in the retention time direction gave the following results: The first analyte starts to elute around retention time 9 and the second around 18. The second analyte disappears around time 34, while the first is exhausted at around 57. Due to the influence of the noise level on the detection limit, the detected values for appearance of analytes are around three retention time points higher than the correct values (the values chosen for the simulations). Similarly, the detected values for disappearance are approximately three retention time points too low.

After having established an overview of the rank changes in the retention time direction, the next step is to extract the profiles of the pure analytes. While the second analyte is completely overlapped throughout the retention time direction, the rank map shows selective retention times for the first analyte. Our method needs only one selective retention time to be successful, provided that the S/N-ratio is acceptable, so the situation looks promising.

As the rank map has shown the presence of two analytes where one has selective regions and the other not, the internal rank annihilation technique has to be applied. This means that we need to pick one retention time matrix containing signal from the first analyte only and another one with signal from both analytes. It is best to pick matrices with good S/N-ratios, so a good choice is the matrices corresponding to retention times 36 (single-analyte region) and 24 (two-analyte region). The two matrices are hereafter called $\mathbf{X}_{36}$ and $\mathbf{X}_{24}$.

A PCA analysis of $\mathbf{X}_{36}$ produced the scores and loadings illustrated in Fig. 3. These vectors correspond to the concentration profile and the spectral profile for the first analyte. The two vectors in Fig. 3 are multiplied together to yield a retention time matrix containing signal from the first analyte only. The amount of noise in the new matrix (called $\mathbf{X}_1$) is less than in $\mathbf{X}_{36}$.

The next step is to solve Eq. (4) using internal rank annihilation, i.e. to remove from the matrix $\mathbf{X}_{24}$ the contribution from the first analyte. Our initial guesses

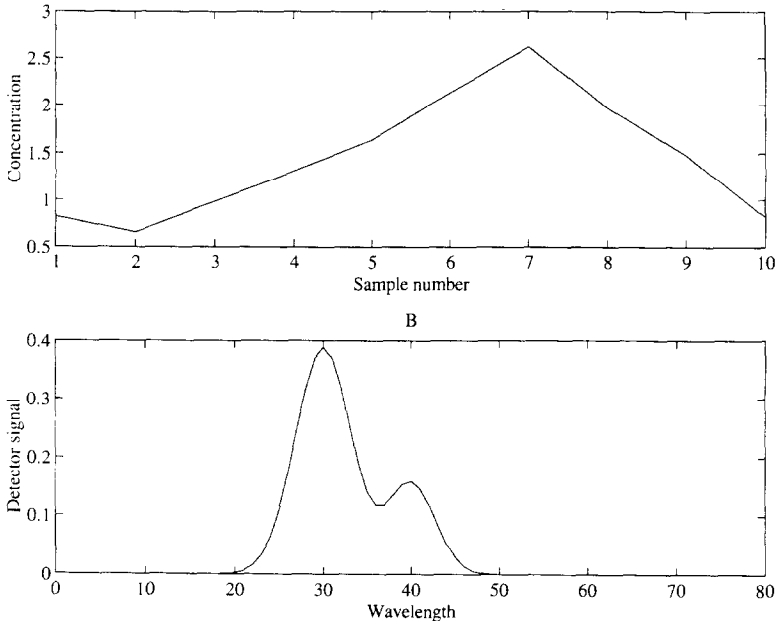

Fig. 3. The first score (a) and loading (b) vector for the matrix $\mathbf{X}_{36}$. The profile shapes correspond to the concentration and spectral profile of the major analyte.

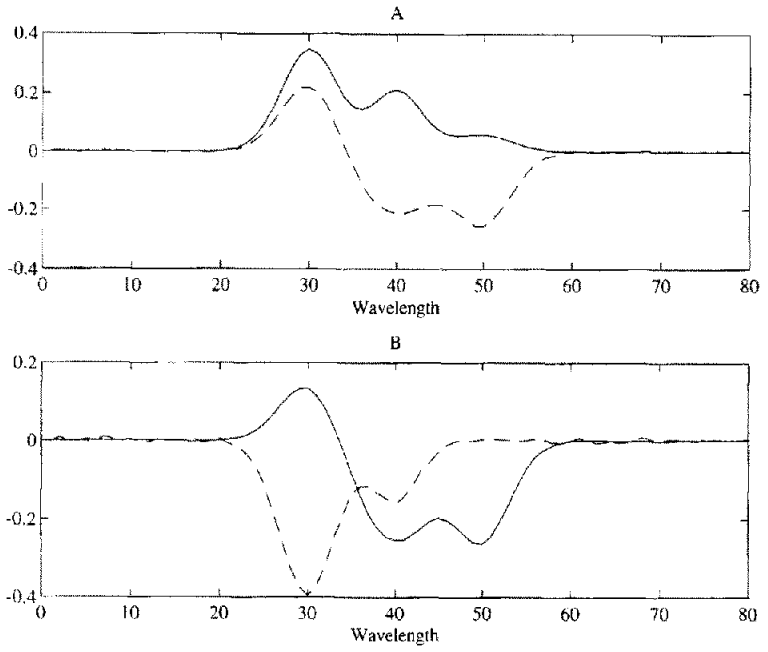

Fig. 4. Two attempts at trying to solve Eq. (4) – with $k = 0$ (a) and $k = 10$ (b). The plots show the first two loading vectors after a decomposition of the resulting matrix $\mathbf{X}_2$. Clearly, none of these attempts are correct – there is structure both in the first (solid line) and second (dotted line) loading vector.

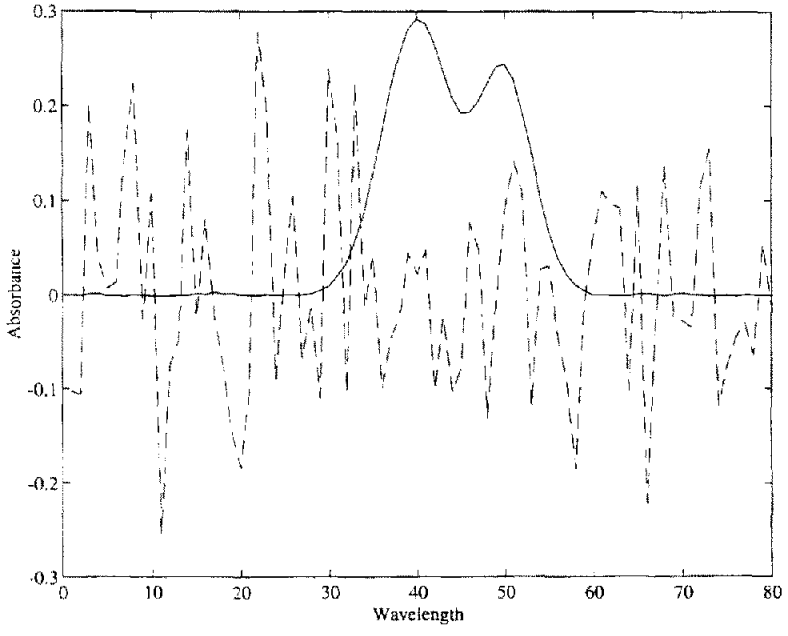

Fig. 5. Eq. (4) is solved. The solid line is the spectral profile belonging to the second, minor analyte. The noisy, dotted line is the second loading vector extracted from $\mathbf{X}_2$ when the correct k is used in Eq. (4).

Table 2
Real and predicted relative concentrations of the minor analyte in the ten samples for data set 1

Real relative conc. of minor analyte	Predicted relative conc. of minor analyte	Error in percent
7.41	7.48	−1.04
13.04	13.13	−0.70
11.76	11.85	−0.76
11.11	11.20	−0.82
10.71	10.80	−0.84
7.14	7.21	−0.97
4.76	4.80	−0.90
4.76	4.81	−1.09
4.25	4.29	−0.88
9.09	9.15	−0.68

for k are 0 and 10. Two principal components are extracted from the resulting matrices. Their loading vectors are displayed in Fig. 4. Clearly, neither of these values are near the correct value. Both the first and second loading vector show structure – implying that we have either removed too much from $\mathbf{X}_{24}$ ($k = 10$) or too little ($k = 0$). The next attempts are $k = 5$, $k = 2.5$, and so on until the second loading vector is without structure. It may be necessary to perform the iterations several times to be sure of finding the best value for k. After several such runs, we ended up with a k value of 0.9999563. This situation is pictured in Fig. 5.

Qualitative concentration and spectral profiles are now established for both analytes. What remains is to resolve the chromatographic profiles and to obtain the quantitative results for the concentrations under the assumption of equal spectral and chromatographic response.

The chromatographic profiles are easily obtained by normalising the two matrices with spectral and concentration profiles and using Eq. (3). This produces the profiles in Fig. 2(d), and the qualitative information is now complete.

Finding the relative concentrations is easy using our knowledge and assumptions about the data. Thus, the quantitative information should lie in the sample direction. A rearrangement of Eq. (3) is used again. This time with normalised spectra and chromatograms as input along with $\mathbf{X}$. The results are shown in Table 2. The EPOS method shows excellent results for these data. The results for the other

Table 3
Real and predicted relative concentrations of the minor analyte in the ten samples for data set 2

Real relative conc. of minor analyte	Predicted relative conc. of minor analyte	Error in percent
7.41	7.47	−0.87
13.04	13.13	−0.66
11.76	11.84	−0.65
11.11	11.17	−0.56
10.71	10.78	−0.62
7.14	7.19	−0.70
4.76	4.79	−0.63
4.76	4.79	−0.51
4.25	4.27	−0.42
9.09	9.14	−0.55

Table 4
Real and predicted relative concentrations of the minor analyte in the ten samples for data set 3

Real relative conc. of minor analyte	Predicted relative conc. of minor analyte	Error in percent
1.00	1.00	−0.02
1.20	1.19	0.94
1.40	1.36	2.68
1.60	1.56	2.69
1.80	1.76	2.08
2.00	1.96	1.96
2.20	2.16	1.69
2.40	2.37	1.34
2.60	2.56	1.36
2.80	2.74	2.17

Table 5
Real and predicted relative concentrations of the minor analyte in the ten samples for data set 4. For this data set, the results from using Lorber's direct method are also showed

Real relative conc. of minor analyte	Predicted relative conc. of minor analyte	Error in percent	Error in percent (Lorber)
1.00	1.00	−0.49	4.58
1.20	1.20	−0.28	4.77
1.40	1.42	−1.20	3.89
1.60	1.62	−1.23	3.84
1.80	1.85	−2.67	2.47
2.00	2.04	−1.97	3.12
2.20	2.25	−2.13	2.96
2.40	2.45	−2.29	2.80
2.60	2.67	−2.52	2.57
2.80	2.87	−2.53	2.55

three data sets described in Table 1 are shown in Table 3, Tables 4 and 5. We see that for the data sets where the concentration range of the analyte is 4–13%, EPOS overestimates the concentrations. For the two data sets where the concentration range is 1–2.8%, EPOS underestimates the concentrations in one case, and overestimates in the other. The results also show that the quality of the resolved profiles and the quantification is independent of whether the noise is heteroscedastic or homoscedastic.

Table 5 shows that for relative concentrations of the minor analytes above 2.5% and heteroscedastic noise, IRA [27] and Lorber's approach [26] perform equally well. For a relative concentration of the analyte below 2.5%, EPOS performs significantly better than the direct approach. We prefer the visual IRA approach as graphical methods are easier to understand and use.

5. Conclusions

The combined visual and numerical inspection of the data set makes EPOS a simple and fast method to use. The problem of embedded minor peaks is uniquely solved, and quantification of the embedded minor analyte is performed with excellent results under the assumption of equal chromatographic and spectral response for the analytes. The method is not restricted to three-way multicomponent profiles. All the principles of the methods are equally valid for data sets containing more than three directions. The only difference for N-way data with $N > 3$ is that the equations used are more complex.

Acknowledgements

B.G. is grateful for a Ph.D. grant on a strategic technology program in chemometrics awarded by The Norwegian Research Council (NFR).

References

[1] Y.-Z. Liang, O.M. Kvalheim and R. Manne, Chemom. Intell. Lab. Syst., 18 (1993) 235.
[2] W. Windig and J. Guilment, Anal. Chem., 63 (1991) 1425.
[3] W. Windig, C.E. Heckler, F.A. Agblevor and R.J. Evans, Chemom. Intell. Lab. Syst., 14 (1992) 195.
[4] W. Windig and D.A. Stephenson, Anal. Chem., 64 (1992) 2735.
[5] B.G.M. Vandeginste, W. Derks and G. Kateman, Anal. Chim. Acta, 173 (1985) 253.
[6] B.G.M. Vandeginste, R. Essers, T. Bosman, J. Reijnen and G. Kateman, Anal. Chem., 57 (1985) 971.
[7] P.J. Gemperline, J. Chem. Inf. Comput. Sci., 34 (1984) 206.
[8] P.J. Gemperline, Anal. Chem., 58 (1986) 2656.
[9] M. Maeder and A.D. Zuberbuhler, Anal. Chim. Acta, 181 (1986) 287.
[10] M. Maeder, Anal. Chem., 59 (1987) 527.
[11] O.M. Kvalheim and Y. Liang, Anal. Chem., 64 (1992) 936.
[12] Y.-Z. Liang, O.M. Kvalheim, H.R. Keller, D.L. Massart, P. Kiechle and F. Erni, Anal. Chem., 64 (1992) 946.
[13] Y.-Z. Liang, O.M. Kvalheim, A. Rahmani and R.G. Brereton, J. Chemom., 7 (1993) 15.
[14] M.D. Hämäläinen, Y.-Z. Liang, O.M. Kvalheim and R. Andersson, Anal. Chim. Acta, 271 (1993) 101.
[15] J. Toft, J., O.M. Kvalheim, F.O. Libnau and E. Nodland, Vibrat. Spectr., 7 (1994) 125.
[16] Y.-Z. Liang and O.M. Kvalheim, Anal. Chim. Acta, 276 (1993) 425.
[17] Y.-Z. Liang, R. Manne and O.M. Kvalheim, Chemom. Intell. Lab. Syst., 22 (1994) 229.
[18] Y.-Z. Liang and O.M. Kvalheim, Anal. Chim. Acta, 292 (1994) 5.
[19] P. Geladi and S. Wold, Chemom. Intell. Lab. Syst., 2 (1987) 273.
[20] H.R. Keller and D.L. Massart, Anal. Chim. Acta, 246 (1991) 379.
[21] J. Toft and O.M. Kvalheim, Anal. Chem., 65 (1993) 2270.
[22] B. Grung B. and O.M. Kvalheim, Chemom. Intell. Lab. Syst., 29 (1995) 223.
[23] C.-N. Ho, G.D. Christian and E.R. Davidson, Anal. Chem., 50 (1978) 1108.
[24] C.-N. Ho, G.D. Christian and E.R. Davidson, Anal. Chem., 52 (1980) 1071.
[25] C-N. Ho, G.D. Christian and E.R. Davidson, Anal. Chem., 53 (1981) 92.
[26] A. Lorber, Anal. Chim. Acta, 164 (1984) 293.
[27] B. Grung and O.M. Kvalheim, Anal. Chim. Acta, in press.