Chemometrics and Intelligent Laboratory Systems 45 (1999) 171–176

Chemometrics and
intelligent
laboratory systems

Subwindow factor analysis

Rolf Manne [a,*], Hailin Shen [b], Yizeng Liang [b]

[a] *Department of Chemistry, University of Bergen, N-4007 Bergen, Norway*
[b] *College of Chemistry and Chemical Engineering, Human University 410082 Changsha, China*

Abstract

The method of subwindow factor analysis (SFA) is introduced as a solution to the problem of directly extracting component spectra from overlapping structures obtained from hyphenated chromatography without first resolving concentration profiles. This is of advantage when a complete resolution cannot be obtained or is of less interest in the analytical situation. The method is based upon comparisons of chromatographic regions (subwindows) which have only one eluting component in common. The paper presents the theory and an application to a structure with 4 overlapping components from a data set from a mixture of polyaromatic hydrocarbons recorded by high-performance liquid chromatography diode array detection (HPLC-DAD). © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Subwindow factor analysis; High-performance liquid chromatography diode array detection; Polyaromatic hydrocarbons

Contents

[*] Corresponding author.

PII: S0169-7439(98)00101-4

1. Introduction

The resolution of overlapping structures obtained from hyphenated chromatography, e.g., high-performance liquid chromatography diode array detection (HPLC-DAD), is an active area of research. From the rank of submatrices one may determine the number of pure components at any one time, and using only this knowledge one is able to resolve overlapping concentration profiles as well as spectra of the pure compounds in a mixture.

In recent papers Cuesta Sánchez et al. [1] and Malinowski [2] point to the problem that certain resolution methods are sensitive to the accurate determination of window limits. With these methods spectra of the pure compounds were obtained from the experimental data matrix **X** and calculated concentration profiles for the pure compounds **C** through a linear regression step

$$\mathbf{S} = \mathbf{X}'\mathbf{C}(\mathbf{C}'\mathbf{C})^{-1} \tag{1}$$

which may be expressed with the help of the generalized inverse as

$$\mathbf{S} = \mathbf{X}'(\mathbf{C}')^{+} \tag{2}$$

When $\mathbf{C}'\mathbf{C}$ is close to singular small errors in **C** may become large in $\mathbf{C}^{+}$ and **S**. This situation is unfortunate since the spectral information is generally more important than the concentration profiles for the identification of unknown constituents.

It is by no means necessary to calculate the spectra of pure compounds from resolved concentration profiles. Thus, with the HELP method [3,4] one identifies selective regions where only one compound elutes and obtains some spectra of pure compounds directly. When both the spectrum and the concentration profile have been identified for a given compound the signal from that compound is removed from the data. Then new selective regions appear, and the process continues.

Another method is the ‘key spectra resolution method for hyphenated two-data’ (KSRMHT) of Xu et al. [5] where the spectra of pure compounds are obtained from the solution of a system of linear equations involving a set of key spectra.

The method of subwindow factor analysis presented in this paper is a method for obtaining directly the spectra of pure compounds using window information for an overlapping structure in hyphenated chromatography. Its basis lies in a theorem previously formulated by one of the authors [6]: ‘‘If for every interferent the concentration window of the analyte has a subwindow where the interferent is absent, then it is possible to calculate the spectrum of the analyte’’. The idea is to select two (or more) subwindows where the only common spectral component is that of the analyte and then to identify this component. This procedure can be undertaken without previous knowledge of the concentration profiles and may be advantageous when a complete resolution is impossible or of less interest in the analytical context.

In this paper we first present a terminology of subwindows, i.e., those parts of the elution window where an analyte coexists with a given set of interfering substances. The following sections contain the theoretical derivation of the method as well as an application to a structure with 4 components from a HPLC-DAD data set obtained from a mixture of polyaromatic hydrocarbons. Further theoretical details are given in Appendix A.

2. Classification of subwindows

We will here present a terminology for the classification of subwindows. We are interested in the subdivisions of the window of the analyte with respect to interferents, i.e., co-eluting compounds.

In the normal situation in chromatography compounds are eluted successively and with little variation of the chromatographic peak widths. An interfering compound which starts to elute before the analyte will appear in a chromatogram to the left of the analyte and will be called a left interferent. In the same way, an interfering compound which continues to elute after the analyte has stopped eluting will be called a right interferent.

An abnormal situation arises when an interferent is both left and right. The chromatogram of interfer-

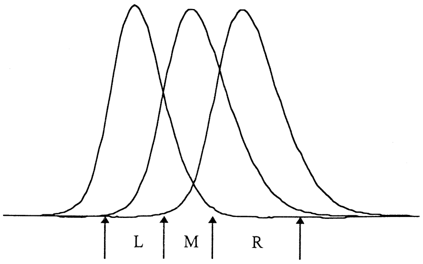

Fig. 1. An illustration of the subwindows of the middle peak.

ent is thus embedding the chromatogram of the analyte. It is known [6] that the presence of embedding interferents makes it impossible to resolve the spectrum of the analyte using window information alone. We will not consider this situation further. In the reverse situation the chromatogram of the interferent is embedded in that of the analyte. This situation creates no significant problem for obtaining the spectrum of the analyte.

We will now consider the various kinds of subwindows that may appear in the chromatogram. The most favourable case, the selective region, is a subwindow where the analyte appears without interferents. In this case the spectrum of the analyte is obtained without further analysis. A subwindow where, in addition to the analyte, there are left interferents but no right interferents will be called a left subwindow. It occurs in the left part of the chromatographic window of the analyte. In the same way, when there are right interferents but no left interferents we will speak of a right subwindow. A subwindow with both left and right interferents will be situated in between the left and right ones. We will call that a middle subwindow. Fig. 1 shows a typical situation.

When there are embedded interferents care has to be taken that subwindows are assigned in such a way that there is only the analyte which elutes both in the left and in the right subwindows.

3. Direct resolution of spectra

The first step in the resolution process is the identification of subwindows. This task is in principle identical to the identification of windows in other window-based resolution methods (e.g., EFA [7,8], orthogonal projection [9], WFA [10], HELP [3,4]) and will not be dealt with here in detail. Our method of choice is rank analysis with a fixed-size moving window [11], also known as eigenstructure tracking analysis (ETA) [12]. The difference from other methods lies in how the elution limits are combined into windows or subwindows. This will be discussed further below.

The rank analysis gives the number of chemical components of the left and right subwindows, say m and n, respectively. The number of components in the combination of left and right subwindows is $m+n-1$ since the analyte is common to both. One may then find an orthogonal basis $\{\boldsymbol{e}_1,\ \boldsymbol{e}_2,\dots \boldsymbol{e}_m\}$ spanning the spectral direction of the left subwindow and a similar basis $\{\boldsymbol{f}_1,\ \boldsymbol{f}_2,\dots \boldsymbol{f}_n\}$ spanning that of the right subwindow. The procedure to use here is singular-value decomposition. The vectors corresponding to the largest singular values are most likely to represent the chemical information. At the risk of increasing the noise-level one may well try bases $\{\boldsymbol{e}_i\}$ and $\{\boldsymbol{f}_j\}$ which are slightly larger than given by the chemical rank.

We now wish to find a vector $\boldsymbol{v}$ which is common to both subspaces. We write for an ideal case

$$\boldsymbol{v}=\Sigma \boldsymbol{e}_i a_i=\Sigma \boldsymbol{f}_j b_j \tag{4}$$

or

$$\boldsymbol{v}=\mathbf{E}\,\boldsymbol{a}=\mathbf{F}\,\boldsymbol{b} \tag{5}$$

Here $\mathbf{E}$ and $\mathbf{F}$ are matrices with m and n columns, respectively. In reality, $\mathbf{E}\boldsymbol{a}$ and $\mathbf{F}\boldsymbol{b}$ are not identical, and we instead search for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ which minimize the squared norm

$$N=\|\mathbf{E}\,\boldsymbol{a}-\mathbf{F}\,\boldsymbol{b}\|^2=\boldsymbol{a}'\mathbf{E}'\mathbf{E}\,\boldsymbol{a}+\boldsymbol{b}'\mathbf{F}'\mathbf{F}\,\boldsymbol{b}-2\,\boldsymbol{a}'\mathbf{E}'\mathbf{F}\,\boldsymbol{b} \tag{6}$$

under the conditions $\boldsymbol{a}'\boldsymbol{a}=\boldsymbol{b}'\boldsymbol{b}=1$.

Since $\mathbf{E}'\mathbf{E}=\mathbf{I}_\mathrm{m}$ and $\mathbf{F}'\mathbf{F}=\mathbf{I}_\mathrm{n}$ (unit matrices of dimension $m\times m$, and $n\times n$, respectively) we obtain

$$N=2-2\,\boldsymbol{a}'\mathbf{E}'\mathbf{F}\,\boldsymbol{b} \tag{7}$$

It may be shown (see Appendix A) that N is minimized if $\boldsymbol{a}$ and $\boldsymbol{b}$ are the left and right singular vectors, respectively, associated with the first (largest)

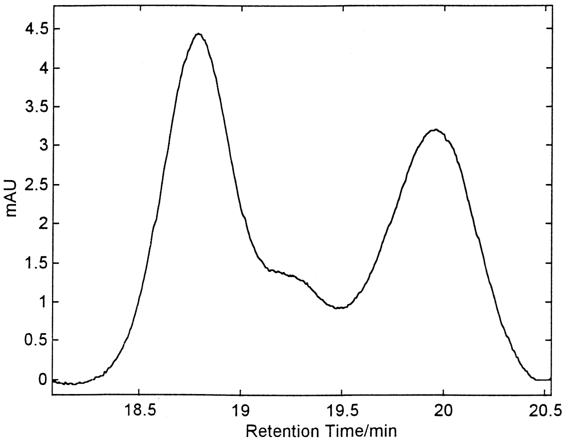

Fig. 2. Average concentration profile of the analyzed data set.

singular value d_1 of the matrix $\mathbf{E}'\mathbf{F}$. Inserting this result in (6) we obtain

$$N = 2(1 - d_1) \tag{8}$$

The singular values d_i are in the range $0 \leq d_i \leq 1$, and the larger the value of d_1 the closer is the agreement between $\mathbf{E}\boldsymbol{a}$ and $\mathbf{F}\boldsymbol{b}$.

An advantage of this method for determining $\boldsymbol{v}$ is that it makes possible a control that this vector, and this vector only, is common for the left and right two subwindows. One thus obtains two solutions $\mathbf{E}\boldsymbol{a}$ and $\mathbf{F}\boldsymbol{b}$ which may be plotted and which have to agree if the solution is to be accepted. If there is no common vector the largest singular value d_1 will be significantly less than 1. On the other hand, if there are two or more common vectors, the second singular value d_2 will also be close to 1. In both cases, one lacks information for the unique identification of the spectral vector $\boldsymbol{v}$.

In some cases, the subwindow limits are uncertain. Also, a subwindow may be too narrow, i.e., its rank is lower than the number of chemical components. In such cases it may be expedient to set the outer limits of the left and right subwindows wider than suggested by the rank map and then vary the inner limits until a good fit is obtained.

It is also possible to add spectral vectors for already resolved interferents to a subwindow thus insuring that the vectors $\{\boldsymbol{e}_i\}$ and $\{\boldsymbol{f}_j\}$ describe the spectra of all interferents as well as the analyte.

4. Resolution of concentration profiles

Since the purpose of the present procedure is the direct resolution of component spectra we wish to leave this topic somewhat open. One way is to use the known spectra and linear regression expression in analogy with (1) and (2) giving the result

$$\mathbf{C} = \mathbf{XS}(\mathbf{S}'\mathbf{S})^{-1} = \mathbf{X}(\mathbf{S}')^{+} \tag{9}$$

The quality of the concentration profiles (non-negativity, unimodality) may then be used as a criterion for the accuracy of the obtained spectra. Alternatively, one may use the window information available for a direct solution, by, e.g., EFA, orthogonal projection, or WFA methods [7–10]. One thus does not have to solve all spectra before one can obtain the first concentration profiles.

5. Numerical

The procedures outlined above have been implemented in Matlab 4.2 and applied to a data set selected from an HPLC-DAD recording of polyaromatic hydrocarbons in the atmosphere at Hong Kong Baptist University. The experimental details are available separately [13,14]. The complete data set contains 47 components. For the present example we have chosen a time interval where, after stripping of earlier and later eluting components, there are 4 unresolved components. The average elution profile (Fig. 2) shows only two maxima with a third component weakly discernible between the two peaks. The subwindows used are listed in Table 1. Concentration profiles were obtained by linear regression, Eq. (9).

Table 1
Subwindows used in the resolution of the experiment

Component	Left subwindow (min.)	Right subwindow (min.)
1	18.166–18.897	18.166–19.359
2	18.904–19.319	19.490–19.801
3	19.325–19.622	19.808–20.270
4	19.491–20.484	20.277–20.484

6. Results and discussion

The resolved spectra and concentration profiles are shown in Fig. 3. The figure shows the mean of the spectra from the left and right subwindows, respectively. In all cases, the overlap between the spectra from the two subwindows (= largest singular value d_1) is greater than 0.99. For the first and last peaks, where the spectra are fully determined by selective regions, this result is to be expected. For the two middle peaks the next singular value d_2 takes the values 0.39 and 0.20, respectively. As with the results of other window-based resolution methods the spectra are obtained in an arbitrary scale. They should consequently be normalized. In our figures we have chosen to normalize to constant Euclidean norm.

It is interesting to compare the present method with the HELP method [3,4]. In the latter the spectra are found successively from selective regions which appear after resolved components have been deleted by the stripping technique. These selective regions are identical to the left or right subwindows as defined here. At first sight it might appear as if HELP is able to resolve the data from only one subwindow while the present method requires two. This is not completely correct. Assume that the left subwindow of an analyte has become a selective region after stripping of left interferents. For this process one has to know where the last left interferent stops to elute, i.e. the left limit of the right subwindow. The right limit of the right subwindow has to be known when one determines the concentration profile of the analyte. The two methods thus use the same information.

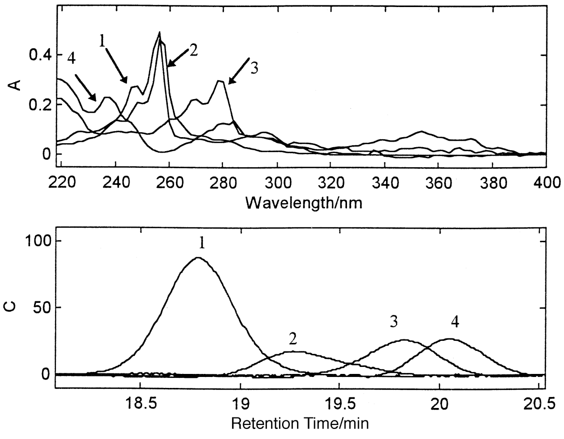

Fig. 3. Resolved spectra and concentration profiles.

If one subwindow of an analyte is absent or not clearly discernible the concentration profile of the analyte may be seen as embedded in that of one of the interferents. In that case complete resolution of the spectrum of the analyte by any window-based technique is impossible without the use of additional modelling information.

The present method does not appear to be very sensitive to the choice of the limits of subwindows. It is immediately clear that one does little harm by setting the outer limits of the subwindows too wide. That means only that one obtains a better description of the spectra of the interferents. Our test case was recorded with very high time resolution (370 time steps for the whole structure considered). Under these conditions it appears as if one can obtain good results also with some contamination of unwanted interferents, i.e. right interferents in the left subwindow and vice versa.

There are several aspects of the present method which warrant further investigation. Together with established window-based methods [7–10] it makes possible the resolution of both the spectrum and the concentration profile of an analyte without assuming knowledge of those of any other compound in the mixture. This is similar to the purpose of rank annihilation factor analysis (RAFA) [15]. By this method one may obtain from 2-way data the concentration of an analyte in a mixture relative to a standard without assuming any knowledge of other compounds present. The well-known problem of RAFA in connection with chromatography is that concentration profiles are subject to drift from one run to another. One possibility for overcoming this problem is now to resolve both the spectrum and the concentration profile of the analyte in the sample without making a full resolution of all the interferents as some other approaches require. A relative concentration scale is obtained by normalizing both the spectrum and the profile and determining the contribution of the resolved component to the full data matrix by RAFA. Comparison with a known standard, which would be resolved in the same way but with a somewhat different concentration profile of the analyte, would fix the concentration scale. As long as the elution window and the subwindows of the analyte can be properly identified such a procedure would be unaffected by chromatographic shifts.

Further applications are shown in a companion paper [16].

7. Conclusion

The present paper thus shows that, provided information of window limits is available, that spectral resolution of two-way data from hyphenated chromatography is possible without previous resolution of the concentration profiles. We believe this result will be of value for the continued exploration of data obtained with this technique.

Acknowledgements

This work is supported by the Chinese National Foundation Committee of Natural Science. R.M. acknowledges the generous support and the hospitality shown to him during his visit at Hunan University September-October 1997.

Appendix A

The minimization problem of Eq. (7) can be expressed in terms of a matrix of dimension $(m+n)\times(m+n)$ which may be written in block matrix form as

$$\begin{aligned} N &= [\,\boldsymbol{a}' \; -\boldsymbol{b}'\,]\begin{bmatrix} \mathbf{E}'\mathbf{E} & \mathbf{E}'\mathbf{F} \\ \mathbf{F}'\mathbf{E} & \mathbf{F}'\mathbf{F} \end{bmatrix}\begin{bmatrix} \boldsymbol{a} \\ -\boldsymbol{b} \end{bmatrix} \\ &= [\,\boldsymbol{a}' \; -\boldsymbol{b}'\,]\begin{bmatrix} \mathbf{I}_{\mathrm{m}} & \mathbf{E}'\mathbf{F} \\ \mathbf{F}'\mathbf{E} & \mathbf{I}_{\mathrm{n}} \end{bmatrix}\begin{bmatrix} \boldsymbol{a} \\ -\boldsymbol{b} \end{bmatrix} \\ &= \boldsymbol{p}'\,\mathbf{A}\,\boldsymbol{p} \end{aligned} \qquad \text{(A1)}$$

Working with unnormalized vectors $\boldsymbol{p}$ we may write the minimization of N as the minimization of a Rayleigh quotient $\boldsymbol{p}'\mathbf{A}\boldsymbol{p}/\boldsymbol{p}'\boldsymbol{p}$. This problem has the solution equal to the smallest eigenvalue of $\mathbf{A}$. The structure of this problem is well-known both from statistics (canonical correlation [17]) and from quantum chemistry (corresponding orbitals [18] and Coulson-Rushbrooke theory of alternant hydrocarbons [19]). Let $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ be the left and right singular vectors of $\mathbf{E}'\mathbf{F}$ associated with the singular value $d_i > 0$, i.e.

$$\boldsymbol{u}_i'\mathbf{E}'\mathbf{F}\,\boldsymbol{v}_i = d_i \qquad \text{(A2)}$$

It is then straight-forward to show that $\begin{bmatrix} \boldsymbol{u}_i \\ -\boldsymbol{v}_i \end{bmatrix}$ is an eigenvector of $\mathbf{A}$ with eigenvalue $1-d_i$ and that $\begin{bmatrix} \boldsymbol{u}_i \\ +\boldsymbol{v}_i \end{bmatrix}$ is an eigenvector with eigenvalue $1+d_i$. The remaining eigenvalues of $\mathbf{A}$ are equal to 1 and are associated with singular vectors with $d_i = 0$. They are of no interest in the present context. With d_1 being the largest singular value, the smallest eigenvalue of $\mathbf{A}$ is thus $1-d_i$.

References

[1] F. Cuesta Sánchez, S.C. Rutan, M.D. Gil Garcia, D.L. Massart, Chemometr. Intell. Lab. Syst. 36 (1997) 153–164.

[2] E.R. Malinowski, J. Chemometrics 10 (1996) 273–279.

[3] O.M. Kvalheim, Y.-z. Liang, Anal. Chem. 64 (1992) 936–945.

[4] Y.-z. Liang, O.M. Kvalheim, H.R. Keller, D.L. Massart, P. Kiechle, F. Erni, Anal. Chem. 64 (1992) 946–953.

[5] J. Xu, Z. Guo, Y.-Z. Liang, R. Yu, J. Chemometrics 10 (1996) 63–76.

[6] R. Manne, Chemometr. Intell. Lab. Syst. 27 (1995) 89–94.

[7] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbuhler, Talanta 32 (1985) 1133.

[8] M. Maeder, A.D. Zuberbuhler, Anal. Chim. Acta 181 (1986) 287.

[9] A. Lorber, Anal. Chem. 58 (1986) 1167.

[10] E.R. Malinowski, J. Chemometrics 3 (1989) 29–40.

[11] H.R. Keller, D.L. Massart, Anal. Chim. Acta 246 (1991) 279.

[12] Y.-z. Liang, O.M. Kvalheim, A. Rahmani, R.G. Brereton, J. Chemometrics 7 (1993) 265–279.

[13] Y.-z. Liang, White, Gray and Black Multicomponent Systems and their Chemometric Algorithms, Hunan Publishing House of Science and Technology, Changsha, China, 1994 (in Chinese), p. 205.

[14] H. Shen, Y. Liang, R. Yu, X. Li, X. Sun, Science in China (Ser. B) 41 (1998) 21–29.

[15] C.-N. Ho, G.D. Christian, E.R. Davidson, Anal. Chem. 50 (1978) 1108–1113.

[16] H. Shen, R. Manne, D. Chen, Y. Liang, Chemometr. Intell. Lab. Syst. 45 (1998) .

[17] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Chap. 10, Academic Press, London, 1979.

[18] A.T. Amos, G.G. Hall, Proc. R. Soc. A 263 (1961) 483–493.

[19] C.A. Coulson, G.S. Rushbrooke, Proc. Cambridge Philos. Soc. 36 (1940) 193.