Available online at www.sciencedirect.com
ScienceDirect
Journal of Chromatography A, 1158 (2007) 258–272
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Towards unsupervised analysis of second-order chromatographic data: Automated selection of number of components in multivariate curve-resolution methods

G. Vivó-Truyols [a,*], J.R. Torres-Lapasió [b], M.C. García-Alvarez-Coque [b], P.J. Schoenmakers [a]

[a] *Polymer-Analysis Group, van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands*

[b] *Departament de Química Analítica, Facultat de Química, Universitat de València, Dr. Moliner, 50, 46100 Burjassot, Spain*

Available online 12 March 2007

Abstract

A method to apply multivariate curve-resolution unattendedly is presented. The algorithm is suitable to perform deconvolution of two-way data (e.g. retrieving the individual elution profiles and spectra of co-eluting compounds from signals obtained from a chromatograph equipped with multiple-channel detection: LC–DAD or GC–MS). The method is especially adequate to achieve the advantages of deconvolution approaches when huge amounts of data are present and manual application of multivariate techniques is too time-consuming. The philosophy of the algorithm is to mimic the reactions of an expert user when applying the orthogonal projection approach—multivariate curve-resolution techniques. Basically, the method establishes a way to check the number of significant components in the data matrix. The performance of the method was superior to the Malinowski *F*-test. The algorithm was tested with HPLC–DAD signals.

Keywords: Multivariate curve-resolution; Unsupervised; Orthogonal projection approach; Autocorrelation; Durbin–Watson

1. Introduction

The emergence of hyphenated chromatography systems (e.g. HPLC-diode-array-UV, GC–MS) and of comprehensive two-dimensional separation methods (e.g. LC × LC, GC × GC) has multiplied the sheer amounts of data produced in the laboratory. The task of interpreting these enormous heaps of data, so as to generate meaningful information, requires powerful data-analysis tools. During several decades chemometricians have been occupied by developing algorithms for this purpose—and they still are. In most cases the signal associated with each specific compound must be identified and extracted from the entire data set. The most popular family of techniques in this context involves multivariate curve-resolution. With these techniques, the separation of the contributions of the interferences from those of the analytes is theoretically possible when the analytical technique does not provide complete selectivity. In other words, chemometrics can complement the imperfect signal separation achieved by chemical methods. The capability of retrieving the analyte signal in the presence of interferences that are accounted for during the calibration has been called the first-order advantage. The second-order advantage refers to the possibility of correctly determining an analyte in the presence of interferences that were not accounted for [1].

The application of multivariate methods to large, complex datasets is, however, difficult, because it requires too much user interaction and supervision. Multivariate methods can only be applied locally to small parts of the dataset. In hyphenated chromatographic techniques, this implies applying multivariate curve-resolution within a small time window, and to then move this along the entire chromatogram. Such a procedure, which implies a (large) number of consecutive applications of multivariate techniques, should be fully automated to be practical. Asking for user intervention each time a multivariate technique is (locally) applied dramatically decreases their usefulness in the work-flow. The greater the amounts of data, the greater the need for automation. Biosystems data analysis (i.e. "omics") constitutes one example in which automation is absolutely necessary,

* Corresponding author. Tel.: +31 20 525 6576.
E-mail address: vivo@science.uva.nl (G. Vivó-Truyols).

0021-9673/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2007.03.005

because the ambition is to detect (and, if possible, to quantify) all components in the sample. This particular area constitutes one of the targets of the present study. When the number of sample components is overwhelming, unsupervised multivariate curve-resolution can be the answer.

In the present context, automation is meant to imply avoiding user interaction each time the multivariate technique is applied locally (to a limited data region). This is not an easy task, as it involves actions at different levels. For instance, at the data pre-processing level corrections for base-line distortions are hard to automate. Yet, the result critically affects the subsequent multivariate analysis. If the base-line correction is imperfect, the multivariate analysis will yield biased results. Another task that is hard to automate is accounting for small shifts in retention times, when data arising from more than one chromatographic run (e.g. samples and calibrants) are processed together. This necessitates a so-called chromatogram-alignment step. One way to circumvent this problem is to apply multivariate curve-resolution methods independently to each chromatogram. Because data pre-processing becomes simpler, this approach is preferred in the present study. Another step that demands user intervention is the validation step. Normally, the user inspects the results of the multivariate technique in a particular local region. A decision is taken to decide whether they are satisfactory, normally based on a visual inspection of the residuals.

Multivariate curve-resolution methods have proven successful for the deconvolution of (partially overlapping) components of a mixture analysed by second-order instruments [2]. One of the most popular multivariate curve-resolution methods is the so-called orthogonal projection approach in combination with alternating least squares (OPA–ALS) [3]. Alternatives to OPA – with common features – can be found in the literature (e.g. SIMPLISMA [4] or Evolving Factor Analysis [5]). It is not the purpose of this work to compare all these methods, but to automate the most flexible one. We have selected OPA–ALS as the core of our approach to the automatic processing of two-way signals. It has been demonstrated [6] that the OPA method, when compared to other methods, performs quite satisfactory in situations of strong overlap. OPA is, however, normally a highly interactive process. It requires experienced users and the complete analysis of large complex data sets is barely feasible. Even more problematic is the fact that human intervention introduces some variations from person to person, making the final results poorly reproducible.

One main step that requires user intervention in OPA–ALS is the selection of initial spectra estimates via OPA analysis prior to submit to ALS. In the OPA method, the user determines sequentially the significant spectra under the signal, so the number of co-eluting species is collaterally obtained. This implies the manual selection of the number of compounds that contribute to the signal matrix. Theoretically, the number of co-eluting compounds in a data matrix (i.e. the pseudo-rank) can be computed in several ways [7,8]. The most usual approach is the Malinowski *F*-test [9]. However, one of the premises for applying this test is that the instrument noise should be uncorrelated [10,11], which is difficult to meet in practice. In a recent report [12], noise autocorrelation for different instruments was studied in order to optimise Savitzky–Golay filters. All instruments tested yielded autocorrelated noise. This decreases the potential applicability of the Malinowski test.

A study aimed at decreasing the extent of human interaction in selecting the number of components and the initial estimates (i.e. spectra) in the OPA method was reported by Gourvénec et al. [13]. The matrix pseudo-rank was estimated by examining the pattern found in the dissimilarity vector with increasing number of components, using the Durbin–Watson (DW) test. We have found, however, that this test yields correct results only if the instrument noise is uncorrelated. The approach thus suffers from the same limitations as the Malinowski test.

A more general solution is provided in the present work for instruments producing autocorrelated noise. The noise autocorrelation of the instrument is inspected and this information is used in a second step to analyse the noise pattern of the dissimilarity. The method can be used unattended as many times as needed, provided that the noise autocorrelation of the instrument remains constant (a reasonable assumption for common instruments). The proposed method gives similar results as the Malinowski test in case of uncorrelated (white) noise and performs better with autocorrelated (coloured) noise. Thus, for most analytical instruments the gain in accuracy is manifest.

2. Theory

2.1. General description of orthogonal projection approach–alternating least squares

As mentioned in Section 1, the OPA–ALS method was selected for automation in this work. The purpose of this section is not to explain the algorithms in detail. We will only give a general overview, which is necessary for the next sections. For more details see ref. [14]. In order to introduce the approaches, an example was taken from high-performance liquid chromatography with diode-array UV detection (HPLC–DAD), in which two compounds co-elute. The approach is not restricted to this type of experiment. OPA–ALS can in principle be applied to separate the individual contributions to any kind of second-order signal, provided they are bilinear [2].

Let us define $\mathbf{Y}$ as the matrix corresponding to the experimental signal of an HPLC–DAD injection containing n_c co-eluting compounds that exhibit different spectra. If the signal was measured at n_t retention times, and the spectra contain n_w wavelengths, $\mathbf{Y}$ will have the dimensions $n_t \times n_w$. If bilinearity holds, $\mathbf{Y}$ can be decomposed as follows:

$$\mathbf{Y} = \mathbf{PD} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{P}$ ($n_t \times n_c$) is the elution profile matrix, with the signal of each compound arranged in each column, $\mathbf{D}$ ($n_c \times n_w$) the spectra matrix, which contains a spectrum of one compound in each row and $\boldsymbol{\varepsilon}$ represents the noise. In our example, since only two compounds are present, $\mathbf{P}$ has the dimensions $n_t \times 2$ and contains the HPLC peak profiles of the two compounds, and $\mathbf{D}$ has the dimensions $2 \times n_w$ and contains the respective spectra.