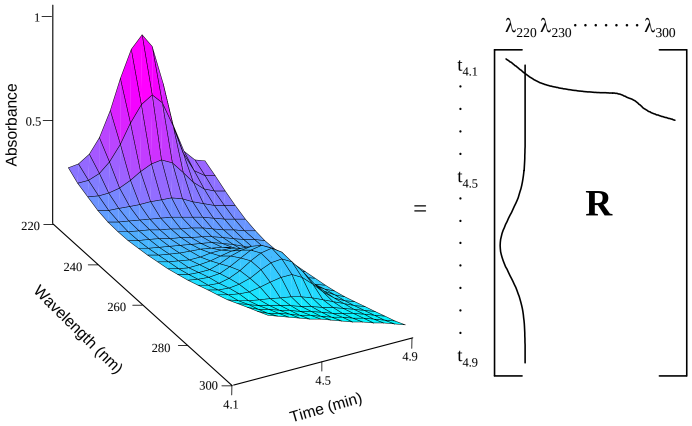
Application of the generalized rank annihilation method (GRAM) to second-order liquid chromatographic data

Enric Comas Lou

Doctoral Thesis



Tarragona 2004

Application of the generalized rank annihilation method (GRAM) to second-order liquid chromatographic data

Doctoral Thesis

ROVIRA I VIRGILI UNIVERSITY

Rovira i Virgili University
Department of Analytical Chemistry and Organic Chemistry

Application of the generalized rank annihilation method (GRAM) to second-order liquid chromatographic data

Dissertation presented by

Enric Comas Lou

to receive the degree

Doctor of the Rovira i Virgili University

European PhD

Tarragona, 2004

Supervisor

Dr. Joan Ferré Baldrich

While writing this thesis I was supervised by Dr. Joan Ferré, and the work was undertaken in the Chemometrics and Qualimetrics research group, headed by Prof. F. Xavier Rius, from the Analytical Chemistry and Organic Chemistry Department at the Rovira i Virgili University in Tarragona, Spain.

It included a research stay in the Department of Chemistry at East Carolina University, USA, in the laboratory of Prof. Paul J. Gemperline, in spring 2002. It also included a scientific collaboration with Dr. Rosa Ana Gimeno, Dr. Rosa Maria Marcé and Prof. Francesc Borrull, of the Chromatography and Environmental Applications research group, from our own department.

The thesis was sponsored by the 'Agència de Gestió d'Ajusts Universitaris i Recerca' (2001FI 00638) from the Catalan Government.

ROVIRA I VIRGILI UNIVERSITY
Department of Analytical Chemistry
and Organic Chemistry

Dr JOAN FERRÉ BALDRICH, Associate professor of the Department of Analytical Chemistry and Organic Chemistry at Rovira i Virgili University

CERTIFIES:

The Doctoral Thesis entitled: 'APPLICATION OF THE GENERALIZED RANK ANNIHILATION METHOD (GRAM) TO SECOND-ORDER LIQUID CHROMATOGRAPHIC DATA', presented by ENRIC COMAS LOU to receive the degree of Doctor of the Rovira i Virgili University, European PhD, has been carried out under my supervision, in the Department of Analytical Chemistry and Organic Chemistry at Rovira i Virgili University, and all the results presented in this thesis were obtained in experiments conducted by the above mentioned student.

Tarragona, November 2004

Dr Joan Ferré Baldrich

Thanks to all the helpful people who have been available
for a few words of wisdom throughout the last few years.

This thesis is warmly dedicated to Luis García and Patricia Rodríguez,
for their friendship and support.

Love makes the world go round,
research makes it go forward

(E. Emmet Reid, Invitation to chemical research)

TABLE OF CONTENTS

Chapter 1

Introduction and Objectives

1.1 INTRODUCTION

The first part of this chapter briefly introduces the importance of second-order calibration in quantitative analysis using chromatographic data. This short bibliographic revision is used to justify the objectives of the thesis (section 1.3). Sections 1.4 and 1.5 contain, respectively, the structure of the thesis and the references cited in this chapter.

1.2 SECOND - ORDER CALIBRATION IN QUANTITATIVE CHROMATOGRAPHIC ANALYSIS

Analytical chemistry is constantly evolving to meet the changing demands of our society [1]. Trends are moving towards increasing the automation of the analyses and determining more analytes, in more complex matrices, faster, with lower detection limits, and using smaller samples and less reagents.

Even though the analytical problems are varied, most situations require some of the constituents of the sample to be identified (qualitative analysis) or their concentration to be determined (quantitative analysis). Often, chromatography is the analytical technique of choice. It is the most widely used separation technique in chemical laboratories and in the chemical process industry [2-3]. Among the different kinds of chromatography, High Performance Liquid Chromatography (HPLC) is one of the most versatile. It is the key separation technique for the analysis of polar and high molecular mass compounds which are not amenable by gas chromatography.

Quantitative HPLC analysis is traditionally carried out by measuring univariate signals (e.g., absorbance at one wavelength) along time. Then, the height or area of the chromatographic peak is related to the concentration of the analyte by means of univariate calibration [4]. Univariate calibration is simple and well-known statistically, but requires the measurements to be selective for the analyte of interest and not influenced by interferences (apart from a constant background contribution) [5]. When complex samples are analyzed, such as environmental or clinical samples, the sample matrix may contain new, unexpected interferences that were not present when the chromatographic separation was optimized. If such interferences coelute with the analyte of interest, the non-selective measurements [6] will cause biased predictions. Such bias can be avoided by selecting an adequate instrumental set up and optimizing the experimental conditions. For example by:

- Cleaning-up the sample. Specific pre-concentration columns have been developed, using immunosorbents and molecular imprinted polymers (MIPs) [7,8], which selectively retain a compound or a family of compounds, and avoid the injection of other components into the chromatographic column.

- Coupling two different chromatographic techniques (multidimensional chromatography) [9-11].

- Using adequate detectors [12], like the diode array detector (DAD), the excitation-emission fluorescence detector (EEMs) or the mass spectrometry detector (MS) and selecting the detection channel in which the interferences have no contribution.

- Changing the chromatographic separation conditions: chromatographic column, mobile phase composition, temperature, pressure, etc.

- Derivatizing, i.e., adding a component that reacts either with the analyte of interest or the interference, before or after the chromatographic separation [13].

Most options require extra time and resources to achieve the adequate selectivity. Moreover, some solutions (i.e., modifying the composition of the mobile phase) may improve selectivity for the actual sample but do not prevent a new interference in the next sample from coeluting with the analyte of interest. For example, the optimal experimental conditions, in the determination of water pollutants in samples from different sources, were found to be different for each sample [14].

Alternatively, the analyst may take advantage of the multiple measurements that the DAD and MS detectors provide. These detectors are becoming commonplace in the analytical laboratories and can record the spectrum of the effluent along time. Hence, the chromatogram is represented by a data matrix, time × detection channel. Such type of data is called second-order data, in contrast to the first-order data (single absorbance measurement along time) or zero-order data (a single absorbance measurement). Second-order data is often underused for quantitative HPLC analysis: quantitation is performed by using only one channel of the measured spectrum, and the spectral dimension is only used for qualitative analysis (i.e., identification of the analyte).

In this thesis we focus our interest on using the full structure of the second-order data matrix. With second-order calibration methods, the concentration of the analytes of interest can be determined in an overlapped peak [15] and selective data are not needed. This reduces the time and cost of the analyses since there is no need to obtain selective data.

Concretely, we focus our attention on HPLC-DAD data and the Generalized Rank Annihilation Method (GRAM) [16,17]. GRAM is a second-order calibration method relatively simple and fast, which only requires a standard (calibration sample) to quantify the analyte in a test sample. GRAM decomposes the two data matrices to obtain: the chromatographic profiles, the spectra of the analytes and the relative concentration for each component. Hence, with GRAM both quantitative and qualitative information are obtained simultaneously.

For applying GRAM, the data from the measured standard and the measured test sample must follow a certain mathematical structure, called *trilinearity*. Variation in retention time and shape of the chromatographic profiles between different runs and samples produce lack of fit in the data structure that leads to unacceptable predictions. This is a reason why, despite its advantages, GRAM is difficult to use in routine applications [18,19]. Hence, research must be directed towards obtaining more reproducible data and correcting retention time mismatch when it occurs. Another reason for not using GRAM is that the input from a trained analyst is needed. The analyst must indicate the number of systematic variations present both in the peak of the standard and in the peak of the test sample. This is required because GRAM is a calibration method based on latent variables (factors). Each latent variable will account for a systematic source of variation in the measured data. Normally, for each analyte present in an overlapped peak, a latent variable is needed. Baseline changes along the peak due, for example, to the change in composition of the mobile phase will also need of a latent variable to describe it. Mathematical methods must be developed that help the analyst to identify the correct number of factors. Finally, the routine application of GRAM requires an adequate outlier detection system that should warn against possible biased predictions due to the lack of fit of the data to the trilinear structure.

1.3 OBJECTIVES OF THE THESIS

The objective of this thesis is to develop new methods to help in detecting and correcting the time shift, deciding the optimal number of factors, and detecting outlying samples in GRAM. These improvements are directed to increase the confidence and automation of the application of GRAM in HPLC-DAD analyses, and hence, increase the speed and reduce the cost of these analyses. In other words, these developments should help in obtaining more information from HPLC-DAD data that are usually recorded but underused.

More specifically, we developed:

1) A retention time shift correction method to align the chromatographic profile of the analyte of interest between different runs and samples. This method is based on Iterative Target Transformation Factor Analysis (ITTFA), which decomposes the peak in the profiles of its constituent analytes.

2) A graphical criterion to determine the number of factors for the GRAM model based on an internal weighting parameter of the GRAM algorithm (α).

3) Two criteria to check whether the measured data have the required trilinear structure, and, hence, estimations are correct. One criterion is based on α, and one criterion is based on the net analyte signal (NAS).

In addition,

4) We have applied GRAM to quantify water pollutants in environmental water samples from the area of Tarragona (Spain). Specifically, GRAM was used to determine aromatic sulfonates, pesticides and phenols in different water samples from the area of Tarragona (Spain).

5) We have compared the GRAM estimations with the ones provided by other second-order methods like Parallel Factor Analysis (PARAFAC) and Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS).

1.4 STRUCTURE OF THE THESIS

The thesis is based on the papers published in international journals. These papers have been edited to give uniform format and uniform mathematical notation along the thesis.

The contents have been structured in six chapters.

- Chapter 1. *Introduction and objectives* contains the introduction, objectives and structure of the thesis.

- Chapter 2. *Second-order chromatographic data. Theoretical background.* Section 2.2 describes the concept of zero-, first- and second-order data and the instruments that generate them. Section 2.3 describes the mathematical structure that will be assumed for second-order HPLC-DAD data. Section 2.4 introduces the theoretical background of curve resolution methods [20]. These methods are mainly used for qualitative analysis and for determining the purity of chromatographic peaks. The Iterative Target Transformation Factor Analysis method is described since it is the base of the method used for correcting the retention time shift developed in Chapter 3. Section 2.5 introduces the second-order calibration methods and Section 2.6 reviews the evolution of the Generalized Rank Annihilation Method during the last twenty-five years, written as an extract of the paper, *Generalized Rank Annihilation Method, a tutorial, J. Ferré, N.M. Faber, E. Comas, F.X. Rius, J. Chromatogr. A (to be submitted).*

- Chapter 3. *Practical aspects in the application of GRAM* deals with three aspects that must be considered to obtain accurate predictions using GRAM. Section 3.2 studies the retention time shift between samples and runs. This is a common effect in liquid chromatography that makes the GRAM estimations incorrect. The paper *Time shift correction in second-order liquid chromatographic data with iterative target transformation factor analysis. E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius, Anal. Chim. Acta 470 (2002) 163-173*, develops a method to correct the retention time shift.

Section 3.3 studies the determination of the number of factors needed to build a GRAM model. A new graphical method is presented in the paper: *Graphical criterion for assessing trilinearity and selecting the optimal number of factors in the Generalized Rank Annihilation Method using liquid chromatography-diode array detection data. E. Comas, J. Ferré, F.X. Rius, Anal. Chim. Acta 515 (2004) 23-30.*

Finally, section 3.4 considers the detection of outlying samples. A new method to detect outliers in GRAM, based on the Net Analyte Signal (NAS), is presented in the paper: *Outlier detection in the Generalized Rank Annihilation Method applied to chromatographic data. E. Comas, J. Ferré, F.X. Rius, Anal. Chem. Submitted.*

A paper in preparation is also included, which compares two strategies to determine the amount of noise in a second-order peak.

- Chapter 4. *Application of GRAM to the determination of water pollutants* contains two papers that show applications of GRAM in complex situations. In the first one, *Using second-order calibration to identify and quantify aromatic sulfonates in water by high-performance liquid chromatography in the presence of coeluting interferences, E. Comas, R. A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius, J. Chromatogr A 988 (2003) 277-284*, GRAM is used to determine aromatic sulfonates with ion-pair liquid chromatography. Due to the polarity of these compounds, the time required for a complete chromatographic separation was large (more than 45 minutes). With GRAM, the time of chromatographic separation was lower than 8 minutes, since a complete separation was not necessary for quantification.

In the second paper, *Quantification from highly drifted and overlapped peaks using second-order calibration methods, E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius, J. Chromatogr A 1035 (2004) 195-202*, GRAM was applied to quantify peaks over a highly drifted baseline. The analytes of interest, pesticides and phenols, eluted overlapped at a high band due to the humic and fulvic acids. GRAM was also compared with two other second-order calibration methods: Parallel Factor Analysis (PARAFAC) and Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS).

- Chapter 5. *Conclusions* contains the conclusions of the thesis. The advantages and limitations of the proposed methodologies are discussed and suggestions for further research are outlined.

- The *Appendix* contains the chemical structure of the studied compounds, the list of the abbreviations used in the thesis and the list of papers and meeting presentations given by the author during the period of development of this thesis.

1.5 REFERENCES

[1] R. Keller, J.M. Mermet, M. Otto, H.M. Winder, Analytical Chemistry, Wiley-VCH, New York, 1998.

[2] C.F. Poole, The Essence of Chromatography, Elsevier, Amsterdam, 2003

[3] Encyclopedia of Separation Science, Academic Press, 2000.

[4] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Elsevier, Amsterdam, 1998.

[5] R. Boqué, J. Ferré, LC-GC Europe 17 (2004) 402-407.

[6] J.M. Davis, J.C. Giddings, Anal. Chem. 55 (1983) 418-424.

[7] B. Sellergren, F. Lanza, Techniques and Instrumentation in Analytical Chemistry, Elsevier 2001.

[8] E. Caro, R.M. Marcé, P.A.G. Cormack, D.C. Sherrington, F. Borrull, J. Chromatogr. A 1047 (2004) 175-180.

[9] I.D. Wilson, U.A.Th. Brinkman, J. Chromatogr. A 1000 (2003) 325-356.

[10] L. Mondello, A.C. Lewis, K.D. Bartle (Eds.) Multidimensional Chromatography, Wiley, New York, 2002.

[11] A.E. Sinha, B.J. Prazen, R.E. Synovec, Anal. Bioanal. Chem. 378 (2004) 1948-1951.

[12] M.S. Lee, LC/MS Applications in Drug Developments, Wiley-interscience series on mass spectrometry, 2002.

[13] G.Lunn, L.C. Hellwig, Handbook of Derivatization Reactions for HPLC, John Wiley, New York, 1998

[14] R. Bossi, K.V. Vejrup, B.B. Mogensen, W.A. Asman, J. Chromatogr. A 957 (2002) 27-36.

[15] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) A782-A791.

[16] E. Sanchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496-499.

[17] N.M. Faber, Anal. Bioanal. Chem 372 (2002) 683-687, and references therein.

[18] R.B. Poe, S.C. Rutan, Anal. Chim. Acta. 283 (1993) 845-853.

[19] S. Li, P.J. Gemperline, K. Briley, S. Kazmierczak, J. Chromatogr. B 665 (1994) 213-233.

[20] A. de Juan, R. Tauler, Anal. Chim. Acta. 500 (2003) 195-210.

Chapter 2

Second-order chromatographic data. Theoretical background

[redacted]

2.1 INTRODUCTION

The aim of this chapter is to introduce the theoretical background of the methods used in the thesis. Section 2.2 introduces the nomenclature that is used to classify the data that can be obtained from an analytical instrument: zero-order, first-order and second-order data. Section 2.3 introduces the bilinear decomposition. Sections 2.4 to 2.6 focus on the methods that use second-order data. Section 2.4 deals with curve resolution methods that only use one sample and their objective is qualitative analysis. Iterative Target Transformation Factor Analysis (ITTFA) is described since it is used in Chapter 3 as a part of a method for correcting the time shift. Section 2.5 deals with second-order calibration methods where the objective is quantitative analysis. Of those, the Generalized Rank Annihilation Method (GRAM) is explained and its evolution in the last twenty-five years is reviewed in Section 2.6.

2.2 ZERO-, FIRST- AND SECOND-ORDER DATA

Sánchez and Kowalski [1,2] established a terminology to name and classify the experimental measurements and the analytical instruments that generate them. When a sample is analyzed, we can measure a single value (e.g., an absorbance at one wavelength), a value over time (e.g., an absorbance over time, which gives a chromatogram) or a series of values over time (e.g., a spectrum over time). Mathematically, these data are arranged as a scalar, a vector or a matrix of values respectively, which we will refer to as ‘zero-’, ‘first-‘ and ‘second-’ order data. This classification of the data is also applied to the analytical instruments that generate them, and calibration methods that use these data. Table 1 shows examples of some instruments that generate these types of data.

Table 1. Nomenclature of analytical data and instruments.

Data order	Data is arranged as a	Data type	Instrument
Zero	Scalar	Absorbance at one wavelength Voltage	Colorimeter pH meter
First	Vector	Chromatogram UV/Vis / NIR spectrum	GC-FID UV/Vis / NIR spectrophotometer
Second	Matrix	Spectrochromatogram Spectra from a kinetic study Two-dimensional chromatogram	HPLC-DAD, GC-MS Spectrophometer measuring over time LC × LC, GC × GC, LC × GC

Figure 1 represents the different orders of data. Lower-order data can also be obtained from higher-order instruments. For example, in the chromatographic peak of Figure 1 (second-order data), a slice at a given wavelength gives the chromatogram (first-order data). In turn, we normally only use the height or area of that chromatographic peak (zero-order data) for quantification.

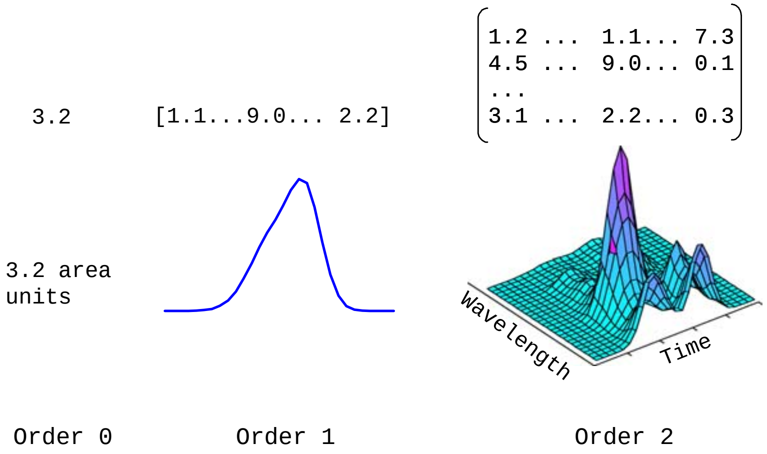


Figure 1. Representation of the different orders of the data.

When the order of the data increases, the cost of the technique and the complexity of the mathematical / statistical data processing also increase. However, the following benefits are obtained: (i) we can detect if other components (interferents) also contribute to the measured signal, and (ii) we can quantify the analyte of interest in the presence of those interferences, by mathematically deconvolving the signal [3,4]. Table 2 summarizes these abilities.

Table 2. Capabilities of data of different orders.

Data order	Detection of interferences	Quantification in the presence of non-calibrated interferences	Common type of calibration
Zero	No	No	Univaritate linear regression
First	Yes	No	Multivariate calibration
Second	Yes	Yes	Second-order calibration

To benefit from first and second-order data, the mathematical algorithms must be able to work with that data structure. Measuring a second-order peak like the one shown in Figure 1, but using the time and wavelength dimensions separately, will not enable us to make predictions in the presence of uncalibrated interferences.

A way of dealing with interferences is using **first-order data** and model the interferences. For this, we need: (i) a series of standards with a known concentration of the analyte of interest, in which the interferences are also present, and (ii) measure, for both the standards and the test sample, at least as many instrumental responses as interferents contribute to the signal. The use of first-order data to build calibration models is known as multivariate calibration. This type of calibration is widely used with spectroscopic data [5]. In chromatographic analysis, the chromatogram also constitutes first-order data. Unfortunately, obtaining reliable chromatograms to perform multivariate calibration is quite difficult because of the lack of reproducibility in retention times from sample to sample.

When a test sample contains an interferent that was not considered in the standards, a multivariate model will lately give biased predictions of the concentration of the analyte of interest (like in univariate calibration). However, unlike with zero-order data, we can detect the presence of the non-calibrated interference, either visually (e.g., if the peak of the analyte of interest has one shoulder, maybe an interferent is eluting with the analyte) or using more complex diagnostics [5]. But we cannot know what effect the interference had on the prediction and, therefore, we cannot correct the inaccurate prediction. With second-order calibration this limitation is overcome.

With **second-order data** we can predict the concentration of an analyte in a sample even in the presence of unknown interferents which were not present in the calibration standards. This useful property is called 'second-order advantage' [3]. In addition to the improved quantitative information, a second-order chromatogram can also be used to obtain qualitative information, such as whether the peak is pure. If the peak is not pure, we can calculate the number of compounds present in the mixture and, with the help of standards or a reference library, identify them.

Notation

Throughout this thesis, bold uppercase letters indicate matrices (second-order data), e.g. **A**; bold lowercase letters indicate vectors (first-order data), e.g. **a**; italic uppercase letters indicate scalars (zero-order data), e.g. *A* [6,7]. Transposition of a matrix or vector is symbolized by a superscripted 'T', e.g. $\mathbf{A}^T$. For a given matrix **A**, the matrices $\mathbf{A}^{-1}$ and $\mathbf{A}^{+}$ stand for its inverse and pseudoinverse, respectively. In full rank matrices $\mathbf{A}^{+} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$.

2.3 BILINEARITY

2.3.1 Bilinear data

Second-order data is produced by instruments that give a matrix of responses for each analyzed sample. Such matrices are obtained by measuring a multivariate response over the variation of a certain magnitude. The response matrix of a pure compound (an individual component, analyte *k*) $\mathbf{R}_k$ is bilinear if it can be expressed as an outer product of two vectors, representing the compound responses in each of the two orders:

$$\mathbf{R}_k = \mathbf{h}_k\mathbf{y}_k^T = d_k\mathbf{x}_k\mathbf{y}_k^T \qquad (1)$$

The scale of $\mathbf{h}_k$ and $\mathbf{y}_k$ is arbitrary. If they are scaled so that $\mathbf{x}_k\boldsymbol{y}_k^T$ is the component response at unitary concentration, then d_k is the concentration of the analyte. Alternatively, both $\mathbf{x}_k$ and $\mathbf{y}_k$ can be normalized to length one. Then, d_k is a scale factor that is proportional to the concentration. To the extent that the experimental noise can be neglected, $\mathbf{R}_k$ has rank one (which is called the pseudo-rank or chemical rank). For some instrumentation the pseudo-rank of a measurement can be assessed a priori, e.g., LC–UV usually gives a pseudo-rank 1 response per analyte due to the specific properties of the LC and UV instruments. In the equations bellow, the noise term has been omitted for simplicity. Table 3 shows examples of such instruments and the measurements of both orders.

Table 3. Techniques that produce bilinear data.

Technique	**order 1** ($\mathbf{x}_k$)	**order 2** ($\mathbf{y}_k$)
Fluorescence emission/excitation	Spectral profile	Spectral profile
GC-IR, LC-UV, HPLC-DAD, FIA-DAD etc.	Elution profile	Spectral profile (absorbance spectrum)
HPLC-MS, GC-MS	Elution profile	Spectral profile (mass spectrum)
GC × GC	Elution profile	Elution profile

Focusing the attention on HPLC-DAD, the measured chromatogram is a large second-order data matrix. From it, the zone where the analyte of interest elutes (sometimes overlapped with interferents) is selected (matrix **R** of dimensions J_1 retention times × J_2 wavelengths). If this matrix corresponds to a calibration standard or a test sample, it will be designated $\mathbf{R}_c$ and $\mathbf{R}_t$ respectively. Sometimes the location of the peak is undetermined and wider or narrower zones can be selected by different users. In chapter 3, we will see that choosing the exact zone is desirable but not critical when algorithms based on second-order data are used.

2.3.2 Bilinear decomposition

In most cases, the peak of the test sample $\mathbf{R}_t$ is a mixture of K analytes, each one contributing as in Eq 1. The objective of the bilinear decomposition is to decompose $\mathbf{R}_t$ as a linear combination of the contribution of each of the K analytes:

$$\mathbf{R}_t = \sum_{k=1}^{K} \mathbf{h}_k \mathbf{y}_k^T = \mathbf{H}\mathbf{Y}^T \qquad (2)$$

where **H** (J_1 × K) and **Y** (J_2 × K) contain the column and row profiles of $\mathbf{R}_t$ with a chemical meaning (e.g., elution profile and spectra respectively for HPLC-DAD data). **H** and **Y** may also contain the profile of a varying baseline, which can be treated as an analyte.

Eq 2 has two unknowns (**H** and **Y**). Hence, the decomposition of $\mathbf{R}_t$ is subject to ambiguities [8], i.e., $\mathbf{R}_t$ can be reproduced by using response profiles differing in shape (rotational ambiguity) or in magnitude (intensity ambiguity) from the (true) ones sought, leading to a range of feasible bands [9,10]. Chemical knowledge about the system being studied can improve the mathematical solution and reduce the number of feasible solutions. This knowledge is introduced as constraints on the possible solutions [11]. The typical constraints in chromatography are the non-negativity of the chromatographic profiles and spectra, and the unimodality (only one maximum) of the chromatographic profiles. Other constraints have been developed recently regarding the concept of local rank [11], closure [12], etc.

Another way of reducing the number of possible solutions in Eq 2 is to add new equations based on the responses of calibration samples and solve the system of equations simultaneously. Hence, the decomposition of Eq 2 can be achieved by using either $\mathbf{R}_t$ alone or $\mathbf{R}_t$ together with some calibration samples. Generally, the methods that only use $\mathbf{R}_t$ are called curve resolution methods. These methods provide qualitative information, such as the chromatographic profiles of the analytes in the peak and their spectra. When more than one sample is used, quantitative information can also be obtained, provided that the reference values in these other samples are known.

H and **Y** can be determined, basically, either by (i) decomposing $\mathbf{R}_t$ in factors, and transforming them to **H** and **Y**, or (ii) by the iteratively improving initial estimation of **H** and **Y**.

i) Decomposition of* $\mathbf{R}_t$*, via the singular value decomposition (SVD)

A liner combination of **H** and **Y** can be estimated by applying SVD [13]:

$$\mathbf{R}_t = \mathbf{USV}^T \qquad (3)$$

where the normalized columns of **U** ($J_1 \times K$) span the same space as the columns of $\mathbf{R}_t$, the normalized columns of **V** ($J_2 \times K$) span the same space as the rows of $\mathbf{R}_t$, and **S** is a $K \times K$ diagonal matrix of scaling factors (in non-increasing order) called singular values. **U**, **V**, and **S** have been truncated to include only the K significant factors. For an HPLC-DAD peak, the columns of **U** are linear combinations of the real elution profiles and the columns of **V** are linear combinations of the spectra.

H and **Y** can be determined from **U**, **V** and **S** by:

i.1) Projecting target vectors onto the space spanned by the columns of **U**. The Iterative Target Transformation Factor Analysis (ITTFA) [14-17] uses the SVD to span the vectorial space of the profiles. Then, a target vector is successively projected and modified until it is explained by the vectorial space of the profiles. Such a modified vector will likely correspond to a chromatographic profile of one of the analytes in $\mathbf{R}_t$. Once the chromatographic profiles of all the analytes in the peak (**H**) have been estimated, **Y** is estimated by solving Eq 2 via least squares.

i.2) Rotating **U**, **V** *and* **S** through an appropriate transformation matrix **T**:

$$\mathbf{H} = \mathbf{UST}$$
$$\mathbf{Y}^{T} = \mathbf{T}^{-1}\mathbf{V}^{T} \qquad (4)$$

In this case, the problem of finding the right profiles is reduced to obtaining **T**. This is the approach used in GRAM [18], for which several formulas exist. Eq 3 is used when the analytes included in the calibration standard are also included in $\mathbf{R}_t$. A more general formulation decomposes the matrix $\mathbf{Q} = \mathbf{R}_t + \alpha\ \mathbf{R}_c$ by SVD, in order to span the space of all the analytes in both the calibration and the test sample (see Section 2.6).

i.3) Finding pure component regions. Evolving Factor Analysis (EFA) [19-22] and its variants [23-27] use this approach. Briefly, those methods apply SVD to different parts of the peak, changing the size of the time window considered and studying the evolution of the eigenvalues (the squares of the singular values in Eq 3). From those methods, the evolution of the system can be found out. Eq 2 can be solved if there selective regions exist where only one analyte is present.

***ii) Use initial estimations of the profiles* H *or* Y**. In these approaches Eq 2 is solved by least squares and constraints are applied to every iteration to improve **H** and **Y** successively. Examples are the calibration methods Parallel Factor Analysis (PARAFAC) [28] and Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) [29,30].

The next section gives a short overview of the curve resolution methods that only use $\mathbf{R}_{t,}$ since one of them, ITTFA, is used in this thesis. The last section gives an overview of the second-order calibration methods, which use $\mathbf{R}_t$ and one or more calibration sample. Since GRAM is the calibration method used in this thesis, it is fully explained in section 2.6.

2.4 CURVE RESOLUTION METHODS

Some curve resolution methods work with first-order data, i.e., the chromatogram measured at only one wavelength. These approaches include the application of neural networks [31], genetic algorithms [32], differential signal detection [33] and the use of sets of equations that model the chromatographic peak [34,35]. The main limitation of these methods is that they must assume the number of analytes in each peak and the shape of the chromatographic profiles. Meyer [36,37] discussed how to measure the area of a peak that elute overlapped with an interference in different experimental situations. However, these situations were limited to overlapped peaks that only contain the analyte of interest and one interference.

The problem of curve resolution can be treated in a more effective way using second-order data and 'Self-modeling curve resolution' (SMCR) methods. [11,38,39]. These methods are powerful approaches whose ultimate goal is to determine the number of components in an overlapped chromatographic peak as well as the spectrum and chromatographic profile of each compound, without assumptions regarding peak shape, location, or identity [40].

Some reviews explain the state-of-the-art of curve resolution methods [11,38]. These methods have been extensively used in many industries, and especially in the pharmaceutical industry [41-47], for example, to determine the presence of impurities in drugs.

2.4.1 Iterative Target Transformation Factor Analysis (ITTFA)

ITTFA is the curve resolution method used in the retention time shift correction method developed in section 3.2. The ITTFA algorithm has four main steps:

Step 1) Singular value decomposition (SVD) of the overlapped peak under study $\mathbf{R}_t$.

$$\mathbf{R}_t = \mathbf{USV}^T \tag{5}$$

Step 2) Estimation of the number of components in the peak and their position.

2.1 $\mathbf{U}_1$ is designed as containing only the first column of $\mathbf{U}$.

2.2 A normalized target vector $\mathbf{x}_{target}$ with the shape of a chromatographic profile is proposed. Possible shapes are Gaussian peaks of different size, needle peaks [48] and triangular peaks.

2.3 $\mathbf{x}_{target}$ is projected into the space described by the column of $\mathbf{U}_1$:

$$\mathbf{x}_{projected} = \mathbf{U}_1\mathbf{U}_1^T\ \mathbf{x}_{target} \tag{6}$$

2.4 The norm of the difference of both vectors is calculated as:

$$d = \|\ \mathbf{x}_{projected} - \mathbf{x}_{target}\ \| \tag{7}$$

When $\mathbf{x}_{target}$ is described by the column of $\mathbf{U}_1$, $\mathbf{x}_{projected}$ will be similar to $\mathbf{x}_{target}$ and the difference d will be small. Then $\mathbf{x}_{target}$ will be representative of the peak of one of the analytes in $\mathbf{R}_t$. If $\mathbf{x}_{target}$ is far from the real one, d will be larger.

2.5. Different $\mathbf{x}_{target}$ are tested, in different positions along the time axis. For each $\mathbf{x}_{target}$ Steps 2.3 and 2.4 are repeated.

2.6. 1- d is represented as a function of the position of the maximum of $\mathbf{x}_{target}$ (see Figure 2).

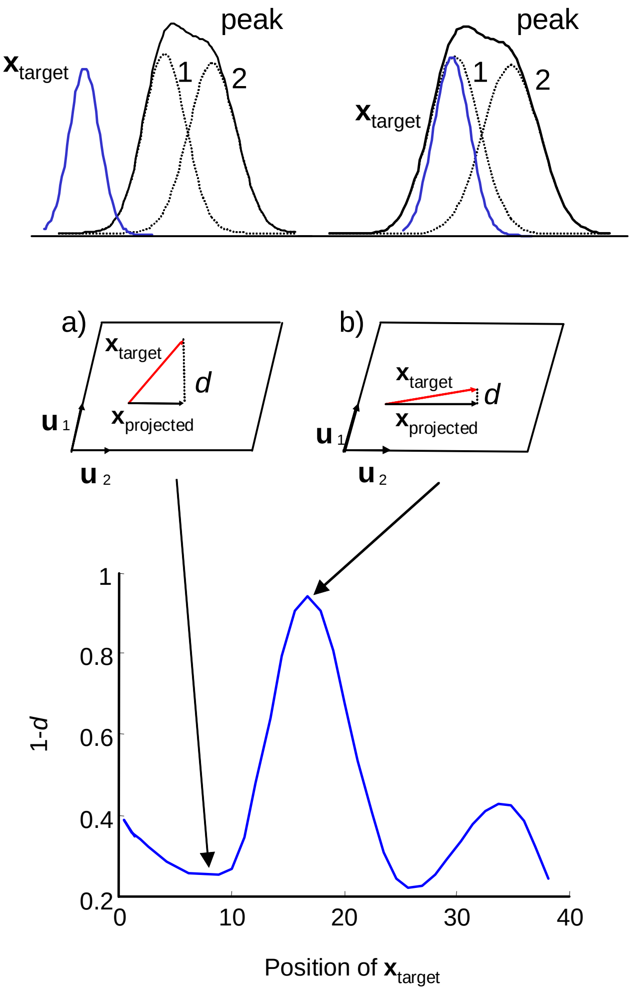


Figure 2. Steps 2.3 to 2.6 when $\mathbf{R}_t$ contains two analytes and two factors are considered. In case (a) $\mathbf{x}_{target}$ is not well explained by **U**, whereas in case (b) $\mathbf{x}_{target}$ is almost fully explained by **U**.

2.7. Steps 2.3 to 2.6 are repeated considering 2, 3, etc, columns of **U**: $\mathbf{U}_2$, $\mathbf{U}_3$, etc. For each number of factors considered, 1- *d* is represented against the position of $\mathbf{x}_{target}$. Each maximum suggests a location of the peak of one analyte. The optimal number of factors *A* is the one when the number of maxima does not increase when the number of factors is increased by one. Then, the number of factors corresponds to the number of components in the peak, and the positions of the maxima indicate the approximate situation of the maxima of the chromatographic profiles.

Step 3) Determination of the chromatographic profiles **H**.

3.1 A matrix $\mathbf{U}_A$ is created where the number of columns is the number of factors *A* determined in Step 2, and the profile $\mathbf{x}_{target}$ that corresponds to the estimated position of the peak of the analyte is selected.

3.2 $\mathbf{x}_{target}$ is projected onto the space spanned by the columns of $\mathbf{U}_A$.

$$\mathbf{x}_{projected} = \mathbf{U}_A\mathbf{U}_A^T\,\mathbf{x}_{target} \qquad (8)$$

$\mathbf{x}_{projected}$ is a tentative chromatographic profile of one of the analytes. It will probably have negative values and more than one maximum, which is not the expected shape for a chromatographic profile.

3.3 Non-negativity and unimodality constraints are applied to obtain $\mathbf{x}_{projected,\ constrained}$.

3.4 $\mathbf{x}_{projected,\ constrained}$ is considered as a new $\mathbf{x}_{target}$ in Eq 8. Steps 3.2 and 3.3 are repeated again until $d = \| \mathbf{x}_{projected,\ constrained} - \mathbf{x}_{target} \|$ is small enough, i.e., the algorithm converges. After convergence, $\mathbf{x}_{projected}$ corresponds to the chromatographic profile of one analyte.

3.5 Steps 3.2 – 3.4 are repeated for each analyte whose position was determined in Step 2. For each analyte a chromatographic profile is found. All the profiles are arranged in **H**.

Step 4) Determination of the spectra **Y** by solving Eq 2:

$$\mathbf{Y} = (\mathbf{H}^{+}\,\mathbf{R})^{T} \qquad (9)$$

2.5 SECOND-ORDER CALIBRATION METHODS

The objective of second-order calibration methods is quantitative. They relate the response variables to the variation of the concentration of the analytes of interest. They use one or more calibration samples and the test sample, in which the concentration is unknown [49-55]. GRAM only needs one calibration sample, while other methods like PARAFAC and MCR-ALS can use several calibration samples. Figure 3 shows how the data are arranged in GRAM, PARAFAC and MCR-ALS. The data in GRAM and PARAFAC are arranged in cubes whereas in MCR-ALS, the data are structured as an augmented matrix. Although the first goal of MCR-ALS is qualitative analysis (curve resolution), the concentration of unknown samples can be determined from the height or the area of the resolved chromatographic profiles. GRAM, PARAFAC and MCR-ALS also provide qualitative information, which is necessary in order to identify the profile of the analyte of interest. Other methods like n-PLS [54], do not provide qualitative information.

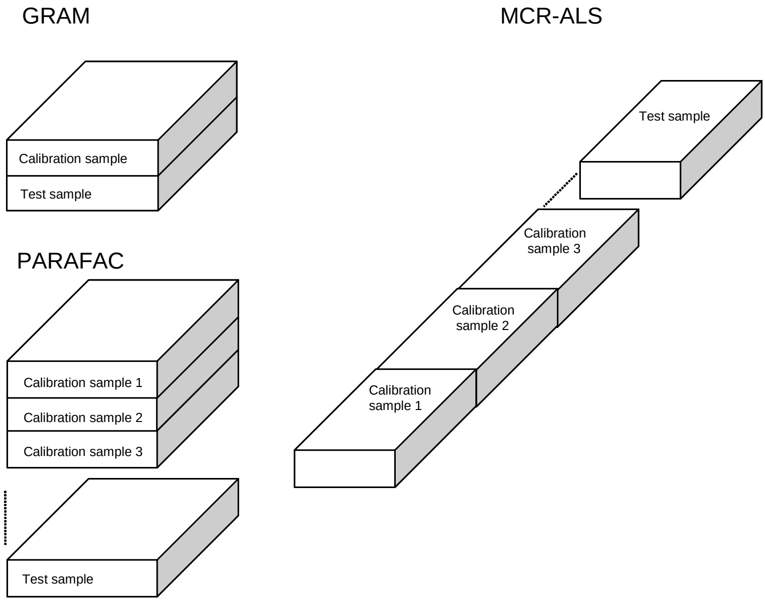


Figure 3. Data arrangement in GRAM, PARAFAC and MCR-ALS. Each slide represents a second-order data matrix (e.g. a second-order chromatographic peak).

GRAM is the method used in this thesis. The next section contains a review of GRAM and its applications, as well as some experimental aspects that must be taken into account to obtain accurate predictions with GRAM.

2.6 GENERALIZED RANK ANNIHILATION METHOD*

The Generalized Rank Annihilation Method (GRAM) stands behind a calibration and curve resolution method that has periodically received attention in analytical chemistry in the last twenty-five years. GRAM is one of the few calibration methods that have been developed within the Chemometrics field. Bruce Kowalski, one the authors, considered GRAM as one of the most important achievements in his career [56].

2.6.1 Theory

Rank annihilation factor analysis (RAFA)

GRAM is based on rank annihilation. Rank annihilation (RA), also called rank annihilation factor analysis (RAFA) [57], was developed by Ho et al. [58]. The principle behind RAFA is that if Eq 2 is followed, the contribution of the analyte of interest to the rank is one. Hence, if we iteratively subtract different amounts of the response of the analyte $\mathbf{R}_c$ from $\mathbf{R}_t$, when the chemical rank of

$$\mathbf{E}(q) = \mathbf{R}_t - q\mathbf{R}_c \qquad (10)$$

* *(extracted from the paper: 'Generalized Rank Annihilation Method, a tutorial'*

J. Ferré, N.M. Faber, E. Comas, F.X. Rius

to be submitted to Journal of Chromatography A) is reduced by one, (i.e., rank($\mathbf{R}_t$ − $q\mathbf{R}_c$) = rank($\mathbf{R}_t$) − 1 = K− 1), then q is the concentration of the analyte k in $\mathbf{R}_t$ relative to its concentration in $\mathbf{R}_c$, i.e., $q = c_{k,t}/c_{k,c}$. In practice, the eigenvalues of $\mathbf{EE}^T$ are monitored for different values of q. The decrease in the rank is indicated by one of the eigenvalues *approaching* zero. The eigenvalue does not become exactly zero because of errors in the data. Hence, RAFA can estimate the concentration of an analyte in a sample of unknown matrix composition using only the measured response of a pure standard of known concentration ($\mathbf{R}_c$) or its best rank one approximation [59]. However, RAFA does not yield the profiles **H** and **Y** of the K analytes in $\mathbf{R}_t$.

The original RAFA involved an iterative refinement to obtain a precise estimation of the concentration. Norgaard and Ridder [60] used the modified Simplex method for finding it. Lorber [61] showed that the reduction in rank could be expressed as a generalized eigenvalue problem and thus the solution could be found directly by SVD. In this method, $\mathbf{R}_c$ must contain only one component. Later, Lorber [62] extended the applicability of RAFA to cases in which $\mathbf{R}_c$ of a single component is characterized by a rank greater than one.

The Generalized rank annihilation method (GRAM)

The RAFA method can only quantitate a single analyte at a time. Ho et al. [63] presented the simultaneous multicomponent rank annihilation (SMRA), which generalized the RA procedure to allow the simultaneous computation of the concentrations of all known components. Sanchez and Kowalski [18] generalized Lorber's method into the generalized rank annihilation factor analysis (GRAFA) (later called generalized rank annihilation method, GRAM) in which several components could be present/absent in both the calibration and the test samples.

In GRAM, only two samples are needed: $\mathbf{R}_c$ and $\mathbf{R}_t$. The GRAM algorithm has five main steps:

1) Addition of $\mathbf{R}_c$ and $\mathbf{R}_t$:

$$\mathbf{Q} = \alpha\, \mathbf{R}_c + \mathbf{R}_t \qquad (11)$$

The columns and rows of **Q** span the vectorial space of the spectra and chromatographic profiles of the analytes.

The weighting factor α was introduced by Faber et al. [64] to estimate and correct the possible bias of the predictions. The original derivation of GRAM considers α = 1, and this has been used along this thesis expect in the paper in section 3.3, in which α is used to determine the number of factors in GRAM. For simplicity, we consider α = 1 in the equations below.

2) SVD of **Q**:

$$\mathbf{Q} = \mathbf{USV}^{\mathrm{T}} + \mathbf{E} \tag{12}$$

U, **V**, and **S** are truncated for the number of factors selected to calculate the model.

3) Resolution of the eigenvalue equation:

$$(\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}_{t}\mathbf{V})\ \mathbf{T} = \mathbf{T\Phi} \tag{13}$$

where **T** is the matrix of the eigenvectors and $\mathbf{\Phi}$ is the diagonal matrix of the eigenvalues.

4) Reconstruction of the chromatographic profiles and spectra

$$\begin{aligned} \mathbf{H} &= \mathbf{UST} \\ \mathbf{Y} &= \mathbf{VT}^{-1} \end{aligned} \tag{14}$$

This step is necessary to identify the analyte of interest. The comparison of the predicted spectra **Y** with the spectra measured in standards is used to identify the analyte of interest. **H** and **Y** contain all the profiles of all the components present both in $\mathbf{R}_c$ and $\mathbf{R}_t$. In the appendix it can be seen that $\mathbf{R}_c$ and $\mathbf{R}_t$ can be written as a combination of **H** and **Y**.

5) Determination of the predicted concentration of the analyte of interest *k*.

$$c_{t,k} = \frac{\mathbf{\Phi}_k c_{c,k}}{1-\mathbf{\Phi}_k} \tag{15}$$

where $\mathbf{\Phi}_k$ is the diagonal element of $\mathbf{\Phi}$ regarding the analyte of interest, and $c_{c,k}$ is the concentration of the analyte of interest in the calibration standard.

Appendix 1 shows the deduction of Eq 13.

2.6.2 Practical considerations for applying GRAM

To confidently apply GRAM, the data in $\mathbf{R}_c$ and $\mathbf{R}_t$ must be trilinear. Some experimental aspects can introduce non-liniearities to the data, which makes the GRAM predictions wrong. This section reviews the model assumptions, possible violations and methods for solving them.

Model assumptions

The correct quantification with RAFA and GRAM requires the data to follow these prerequisites:

- The response matrix of a pure compound must be bilinear, i.e. Eq 1.

- The signals from the different analytes are additive.

- The bilinear data matrices of the standard and the unknown mixture as a group must also be trilinear, i.e., the pure analyte response at unit concentration has the same form in both samples.

- The columns of **H** (and also **Y**) must be linearly independent. This means that, for HPLC-DAD data, if the spectra or elution profiles of overlapped components are identical, GRAM cannot find a solution.

- The ratio of concentrations calibration/unknown must be different for the target analytes (i.e., $c_{o,k}/c_{t,k} \neq c_{o,j}/c_{t,j}$ for every $k \neq j$). Otherwise, the components are not correctly resolved.

- The total number of unique component signals in the sample and standard data matrices cannot exceed the smallest dimension of the data matrices.

Violation of model assumptions

Same ratio of concentrations calibration/unknown

The calibration standard in GRAM can be either a pure standard of known concentration or and aliquot of the unknown sample with an added known concentration of analyte (standard addition). The standard addition approach (spiked samples) is usually preferred [65,66] since this ensures that two components do not have the same ratio of concentrations between samples. However, this situation of same ratio of concentrations is hardly found in real samples.

Retention time and peak shape irreproducibilities

To give correct predictions, GRAM requires the profiles in **H** and **Y** of the analytes of interest be the same both in $\mathbf{R}_c$ and $\mathbf{R}_t$. This means that, for example, for HPLC-DAD data, the retention times and peak shapes for these analytes must be identical in the two samples. However, analyses of hyphenated chromatographic/spectral data sets often contain retention time and peak shape irreproducibilities [67], especially when gradient elution programs are used [65]. The match in analyte's peak shape and width can be maximized by eliminating chemical matrix effects through standard addition [68,69].

Some authors have addressed the problem of chromatographic retention time precision for second order chromatographic data [70-72]. These methods are commented in the section 3.2 of the thesis.

Pseudo-rank higher than one and not rank linear additivity.

When the pseudo-rank of a pure analyte response is not one, or the rank linear additivity does not hold (such in a described FIA system [73]), RAFA and GRAM do not work well. Rank linear additivity means that if analyte 1 gives a rank *r1* response and analyte two a rank *r2* response, then the mixture of the two analytes gives a rank *r1* + *r2* response. The direct generalization of GRAM for situations with pure analyte responses of pseudo-rank higher than 1 is Nonbilinear Rank Annihilation (NBRA) [74,75]. Quantification is still possible, but resolving individual profiles is not possible. Both GRAM and NBRA break down if the rank linear additivity property does not hold. A mathematical treatment of the properties of GRAM, NBRA and the relation to rank linear additivity is given by Kiers and Smilde [76]. Several methods have been described for complicated second-order calibration, that is, cases where the rank 1 property and perhaps even rank linear additivity do not hold. One of these methods is multivariate curve resolution (MCR) with restrictions. Another method uses restricted Tucker3 models to calibrate the complicated second-order system [77].

Determination of rank

GRAM requires an input estimate of the number of components present both in $\mathbf{R}_c$ and $\mathbf{R}_t$. The different ways to determine the number of factors are developed in section 3.3 of the thesis.

While GRAM fails if the number of components considered is less than the actual number, in the literature there is no agreement about if a number larger than the optimal number has a negative influence in the predictions. Wilson et al. [74] found that the concentration estimates obtained for the analytes of interest via a variation of the GRAM algorithm are not very sensitive to the inclusion of a few additional factors. As long as sufficient factors are included to describe the responses of all the analytes, the computation of the concentration estimates is stable. Later, Li et al. [65] found these conclusions to be true for simulated data and real data. In those cases, Li et al. [78] found that standard addition experiments can be especially useful for selecting the proper number of principal components. The difference in recoveries can be used as a criterion for estimating the number of principal components to be used. For the ternary mixture of EEM, Frenich et al. [69] did not found a significant influence of the number of factors (between 3 to 5) on the estimated spectral profiles.

Complex-valued solution to eigenvalue problem

GRAM can yield complex eigenvalues and eigenvectors. Faber et al. [79] commented the possibilities of obtaining complex eigensolutions and degenerate eigensolutions. They also showed that complex solutions should not arise for components present in both samples if the data follow the assumed linear additive model. In case they arise, Li et al. [65] showed an improvement to the GRAM algorithm that uses two similarity transformations for eliminating the imaginary part in the eigenvalues and eigenvectors when they are obtained.

2.6.3 Applications of GRAM

The main benefit for the analytical chemist is that GRAM allows quantifying an analyte in a sample without knowing the identity or amount of the other components (interferents) that also contribute to the instrumental response. This advantage has different readings. In chromatography, this property involves that GRAM can mathematically resolve and quantify partially resolved peaks. Since the compounds do not need to be completely separated from the interferences, sample preparation procedures can be simpler and run times can be shorter [65,68] than the ones based on univariate calibration. Moreover, with GRAM the quantification can be carried out with only one calibration sample (a pure standard or a spiked sample). In this context, GRAM has been applied to the chromatographic analysis of a variety of clinical and environmental samples. In HPLC-DAD data GRAM could accurately predict spectra and concentrations of components that are totally overlapped such as drugs of abuse in clinical samples [68,78] or polycyclic

aromatic hydrocarbons (PAHs) in water samples [66] among others [80,81]. It has also found applications on gas chromatography-selected-ion monitoring GC/GC-SIM [82], comprehensive GC×GC data [83] and bimodal HPLC-DAD data of polycyclic aromatic hydrocarbons in which data were acquired from two different chromatographic systems simultaneously and combined to form one data matrix [84]. Gross et al. [85] used GRAM for prediction in parallel column liquid chromatography with a single multi-wavelength absorbance detector. Fraga et al. [68] used GRAM for the high-speed quantitative analysis of aromatic isomers in a jet fuel sample using comprehensive two-dimensional gas chromatography (GC × GC) using the standard addition method and an objective retention time alignment algorithm. Fraga et al. [86] evaluated the theoretical enhancement provided by application of the GRAM for the analysis of unresolved peaks in comprehensive 2-D separations. They concluded that the use of GRAM should increase the number of analyzable peaks for all forms of comprehensive 2-D separations.
GRAM has also been applied to excitation-emission fluorescence spectroscopy. Frenich et al. [69] used GRAM for the resolution and quantitation of ternary mixture of pesticides with overlapped spectra. They illustrated its application in the analysis of real water samples containing the target pesticides.

RAFA, a predecessor of GRAM, has been applied to solve a variety of problems, both in excitation-emission fluorescence [58,59], LC/UV data [57], thin Layer Chromatography-reflectance imaging spectrophotometry [87] and flow injection analysis (FIA) system with a pH gradient [60]. RAFA was used for spectrophotometric study of complex formation equilibriums [88] with different complexation stoichiometries and spectral overlapping of involved components and also for determination of rate constants from two-way kinetic-spectral data [89].

Windig and Antalek [90,91] developed a modification of GRAM, which they called direct exponential curve resolution algorithm (DECRA), in which the data set from one single experiment is used to build the two data sets needed in GRAM. Only one experiment is needed when the contribution of the components in the mixture spectra is of a decaying exponential character. DECRA was used with pulsed gradient spin echo (PGSE) nuclear magnetic resonance (NMR) data and ultraviolet/visible data [91-93]. DECRA has also been used for rapid estimation of rate constants using on-line short-wavelength near-infrared (SW-NIR) measurements [94] and UV-vis spectra [95] when the contribution of the different species in the mixture spectra is of exponentially decaying character [96].

Appendix 1. Deduction of the GRAM equations.

If $\mathbf{R}_c$ and $\mathbf{R}_t$ are bilinear, they can be written as the outer product of the chromatographic profiles at unit concentration ($\mathbf{X}$), times the concentration, times the normalized spectra ($\mathbf{Y}$). Then the sum matrix in Eq 11 can be written as (considering α=1):

$$\mathbf{Q} = \mathbf{R}_c + \mathbf{R}_t = \mathbf{X}\mathbf{C}_c\mathbf{Y}^T + \mathbf{X}\mathbf{C}_t\mathbf{Y}^T = \mathbf{X}(\mathbf{C}_c+\mathbf{C}_t)\mathbf{Y}^T = \mathbf{H}\mathbf{Y}^T \qquad (16)$$

where $\mathbf{H}$ are the 'real' elution profiles and $\mathbf{Y}$ the 'real' spectra. $\mathbf{H}$ and $\mathbf{Y}$ contain the profiles of all the analytes present in both matrices ($\mathbf{R}_c$ and $\mathbf{R}_t$). $\mathbf{C}_c$ and $\mathbf{C}_t$ are diagonal matrices that contain the relative concentration for the different analytes in $\mathbf{R}_c$ and $\mathbf{R}_t$. If one analyte is not present in $\mathbf{R}_c$ or $\mathbf{R}_t$, its corresponding element in $\mathbf{C}_c$ or $\mathbf{C}_t$ is zero.

The calibration and test samples can also be expressed in terms of $\mathbf{H}$ and $\mathbf{Y}$:

$$\begin{aligned} \mathbf{R}_c &= \mathbf{H}\,\boldsymbol{\Pi}\,\mathbf{Y}^T \\ \mathbf{R}_t &= \mathbf{H}\,\boldsymbol{\Phi}\,\mathbf{Y}^T \end{aligned} \qquad (17)$$

where $\Pi = (\mathbf{C}_c + \mathbf{C}_t)^{-1}\,\mathbf{C}_c$ and $\boldsymbol{\Phi} = (\mathbf{C}_c + \mathbf{C}_t)^{-1}\,\mathbf{C}_t$. It can be seen that $\Pi + \boldsymbol{\Phi} = \mathbf{I}$ (identity matrix):

$$\mathbf{Q} = \mathbf{R}_c + \mathbf{R}_t = \mathbf{H}\boldsymbol{\Pi}\mathbf{Y}^T + \mathbf{H}\boldsymbol{\Phi}\mathbf{Y}^T = \mathbf{H}(\boldsymbol{\Pi}+\boldsymbol{\Phi})\mathbf{Y}^T = \mathbf{H}\mathbf{Y}^T \qquad (18)$$

The goal of GRAM is to find $\mathbf{H}$ and $\mathbf{Y}$. The space spanned by the rows and columns of $\mathbf{Q}$ is found via SVD of $\mathbf{Q}$.

$$\mathbf{Q} = \mathbf{USV}^{\mathrm{T}} + \mathbf{E} \quad \text{(see Eq 12)}$$

A transformation matrix (**T**) must be found that converts the abstract profiles into the real ones:

$$\mathbf{USTT}^{-1}\mathbf{V}^{\mathrm{T}} = \mathbf{HY}^{\mathrm{T}} \tag{19}$$

Hence

$$\mathbf{H} = \mathbf{UST} \tag{20}$$

$$\mathbf{Y}^{\mathrm{T}} = \mathbf{T}^{-1}\mathbf{V}^{\mathrm{T}} = \mathbf{V}(\mathbf{T}^{-1})^{\mathrm{T}}$$

If we isolate **Φ** from Eq 17, we find

$$\mathbf{H}^{+}\mathbf{R}_{t}(\mathbf{Y}^{\mathrm{T}})^{+} = \mathbf{\Phi} \tag{21}$$

where

$$\mathbf{H}^{+} = (\mathbf{UST})^{+} = \mathbf{T}^{-1}\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}$$

$$(\mathbf{Y}^{\mathrm{T}})^{+} = (\mathbf{T}^{-1}\mathbf{V}^{\mathrm{T}})^{+} = \mathbf{VT} \tag{22}$$

updating Eq 21

$$\mathbf{T}^{-1}\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}_{t}\mathbf{VT} = \mathbf{\Phi} \tag{23}$$

$$(\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}_{t}\mathbf{V})\ \mathbf{T} = \mathbf{T\Phi} \tag{24}$$

which is an eigenvalue equation , where ($\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{R}_{t}\mathbf{V}$) is a square matrix *K* × *K*, **T** is the matrix of eigenvectors and **Φ** the matrix of eigenvalues.

The analyte of interest is identified by spectral comparison of **Y** and the spectra measured in standards. The predicted concentration in the test sample ($c_{t,k}$) can be found from the k_{th} diagonal element of **Φ** :

$$\mathbf{\Phi}_{k} = \frac{c_{t,k}}{c_{c,k} + c_{t,k}} \tag{25}$$

The predicted concentration of the analyte $c_{t,k}$ is found as

$$c_{t,k} = \frac{\mathbf{\Phi}_k c_{c,k}}{1-\mathbf{\Phi}_k} \qquad (26)$$

Alternatively, the eigenvalue problem equation can be solved using $\mathbf{R}_c$ instead of using $\mathbf{R}_t$:

$$\mathbf{T}^{-1}\mathbf{S}^{-1}\mathbf{U}^{T}\mathbf{R}_c\mathbf{V}\mathbf{T} = \mathbf{\Pi} \qquad (27)$$

$$(\mathbf{S}^{-1}\mathbf{U}^{T}\mathbf{R}_c\mathbf{V})\ \mathbf{T} = \mathbf{T}\ \mathbf{\Pi} \qquad (28)$$

where the predicted concentration corresponds to:

$$\mathbf{\Pi}_k = \frac{c_{c,k}}{c_{c,k} + c_{t,k}} \qquad (29)$$

and

$$c_{t,k} = \frac{c_{c,k}(1-\mathbf{\Pi}_k)}{\mathbf{\Pi}_k} \qquad (30)$$

2.7 REFERENCES

[1] E. Sanchez, B. R. Kowalski, J. Chemom. 2 (1988) 247-263.

[2] E. Sanchez, B. R. Kowalski, J. Chemom. 2 (1988) 254-280.

[3] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) A782-791.

[4] R. Boqué, J. Ferré, LCGC Europe 17 (2004) 402 -407.

[5] H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons, 1989.

[6] H. A. L. Kiers, J. Chemom. 14 (2000) 105-122.

[7] R. A. Harshman 15 (2001) 689-714.

[8] W. H. Lawton, E. A. Sylvestre, Technometrics 13 (1971) 617-633.

[9] P.J. Gemperline, Anal. Chem. 71 (1999) 5398-5404.

[10] R. Tauler, J. Chemom. 15 (2001) 627-646.

[11] A. de Juan, R. Tauler, Anal. Chim. Acta 500 (2003) 195-210.

[12] A. de Juan, S. Navea, J. Diewok, R. Tauler, Chemom. Intell. Lab. Syst. 70 (2004) 11-21.

[13] E.R. Malinowski, Factor Analysis in Chemistry, 3rd Ed, John Wiley & Sons Inc, New York, 2002.

[14] P.J. Gemperline, J. Chem. Inf. Comput. Sci. 24 (1984) 206-212.

[15] R.A. Roscoe, P.K. Hopke, Comp. & Chem. 5 (1981) 1-7.

[16] B.G.M. Vandeginste, W. Derks, G. Kateman, Anal. Chim. Acta. 173 (1985) 253-264.

[17] B.G.M. Vandeginste, F. Leyten, M. Gerritsen, J. W. Noor, G. Kateman, I. Frank, J. Chemom. 1 (1987) 57-71.

[18] E. Sanchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496-499.

[19] M. Maeder, Anal. Chem. 59 (1987) 527-530.

[20] H.R. Keller, D.L. Massart, Chemom. Intell. Lab. Syst. 12 (1992) 209-224.

[21] A.C. Whitson, M. Maeder, J. Chemom. 15 (2001) 475-484.

[22] M. Maeder, A. Zilian, Chemom. Intell. Lab. Syst. 3 (1988) 205-213.

[23] R. Manne, H.L. Shen, Y.Z. Liang, Chemom. Intell. Lab. Syst. 45 (1999) 171-176.

[24] F.C. Sanchez, S.C. Rutan, M.D.G. Garcia, D.L. Massart, Chemom. Intell. Lab. Syst. 36 (1997) 153-164.

[25] S. Gourvenec, D.L. Massart, D.N. Rutledge, Chemom. Intell. Lab. Syst. 61 (2002) 51-61.

[26] C. Ritter, J.A. Gilliard, J. Cumps, B. Tilquin, Anal. Chim. Acta. 318 (1996) 125-136.

[27] A.K. Elbergali, R.G. Brereton, A. Rahmani, Analyst 121 (1996) 585-590.

[28] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149-171.

[29] R. Tauler, D. Barceló, Trends Anal. Chem. 12 (1993) 319-327.

[30] R. Tauler, Chemom. Intell. Lab. Syst. 30 (1995) 133-146.

[31] H. Miao, M. Yu, S. Hu, J. Chromatogr. A 749 (1996) 5-11.

[32] X. Shao, Z. Chen, X. Lin, Chemom. Intell. Lab. Syst. 50 (2000) 91-99.

[33] S. Nakamura, J. Chromatogr. A 859 (1999) 221-225.

[34] P. Nikitas, A. Pappa-Louisi, A. Papageorgiou. J. Chromatogr. A 912 (2001) 13-29.

[35] S. Jurt, M. Shär, V.R. Meyer, J. Chromatogr A 929 (2001) 165-168.

[36] V.R. Meyer, Chromatographia 40 (1995) 15-22.

[37] V.R. Meyer, J. Chromatogr Sci. 33 (1995) 26-33.

[38] J. Jiang, Y. Liang, Y. Ozaki, Chemom. Intell. Lab. Syst 71 (2004) 1-12.

[39] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Elsevier, Amsterdam, 1998.

[40] M.F. Delaney, Anal. Chem. 56 (1984) 261R-277R.

[41] F.C. Sanchez, M.S. Khots, D.L. Massart, J.O. de Beer, Anal. Chim. Acta. 285 (1994) 181-192.

[42] K. de Braekeleer, J.R. Torres-Lapasio, D.L. Massart, Chemom. Intell. Lab. Syst. 52 (2000) 45-59.

[43] A.G. Frenich, J.R. Torres-Lapasio, K. de Braekeleer, D.L. Massart, J.L.M. Vidal,M.M. Galera, J. Chromatogr. A 855 (1999) 487-499.

[44] F.C. Sanchez, M.S. Khots, D.L. Massart, Anal. Chim. Acta. 290 (1994) 249-258.

[45] H.R. Keller, D.L. Massart, Anal. Chim. Acta. 246 (1991) 379-390.

[46] K. Wiberg, M. Andersson, A. Hagman, S.P. Jacobsson J. Chromatogr. A 1029 (2004) 13–20.

[47] P.V. Zomoren, H. Darwinkel, P.M.J. Coenegracht, G.J. Jong, Anal. Chim. Acta. 487 (2003) 155-170.

[48] A. de Juan, B. van den Bogaert, F.C. Sanchez, D.L. Massart, Chemom. Intell. Lab. Syst. 33 (1996) 133-145.

[49] R. Bro, Multi-way analysis in the food industry, Models, algorithms, and applications. Doctoral dissertation, University of Amsterdam, 1998.

[50] N.M. Faber, R. Bro, P.K. Hopke, Chemom. Intell. Lab. Syst. 65 (2003) 119-137.

[51] M. Linder, Bilinear regression and second order calibration, Doctoral dissertation, Stockholm University, 1998.

[52] A. Smilde, R. Bro, P. Geladi, Multi-way analysis. Applications in the chemical sciences, Wiley 2004.

[53] A. de Juan, R. Tauler, J. Chemom. 15 (2001) 749-772.

[54] R. Bro, J. Chemom. 10 (1996) 47-61.

[55] V. Pravdova, F. Estienne, B. Walczak, D. L. Massart, Chemom. Intell. Lab. Syst. 59 (2001) 75-88.

[56] P. Geladi, K. Esbensen, J. Chemom. 4 (1990) 337-354.

[57] M. McCue, E. R. Malinowski, J. Chromatogr. Sci. 21 (1983) 229-234.

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

[59] C. N. Ho, G. D. Christian, E. R. Davidson, Anal. Chem. 52 (1980) 1071-1079.

[60] L. Norgaard, C. Ridder, Chemom. Intell. Lab. Syst. 23 (1994) 107-114 .

[61] A. Lorber, Anal. Chim. Acta. 164 (1984) 293-297.

[62] A. Lorber, Anal. Chem. 57 (1985) 2395-2397.

[63] C. N. Ho, G. D. Christian, E. R. Davidson, Anal. Chem. 53 (1981) 92-98.

[64] N.M. Faber, J. Ferré, R. Boqué, Chemom. Intell. Lab. Syst. 55 (2001) 67-90.

[65] S. Li, P. J. Gemperline, K. Briley, S. Kazmierczak, J. Chromatography B 665 (1994) 213-233.

[66] R. A. Gimeno, E. Comas, R. M. Marcé, J. Ferré, F.X. Rius, F. Borrull, Anal. Chim. Acta 498 (2003) 47-53.

[67] D. H. Burns, J. B. Callis, G. D. Christian, Anal. Chem. 58 (1986) 2805-2811.

[68] C. G. Fraga, B. J. Prazen, R. E. Synovec, Anal. Chem. 72 (2000) 4154 - 4162.

[69] A. G. Frenich, D. P. Zamora, M. M. Galera, J. L. M. Vidal, Anal. Bioanal. Chem. 375 (2003) 974-980.

[70] B. Grung, O.M. Kvalheim, Anal. Chim. Acta. 304 (1995) 57-66.

[71] B. J. Prazen, R. E. Synovec, B. R. Kowalski, Anal. Chem. 70 (1998) 218-225.

[72] E. Comas, R. A. Gimeno, J. Ferré, R. M. Marcé, F. Borrull, F. X. Rius, Anal. Chim. Acta 470 (2002) 163-173.

[73] A. K. Smilde, R. Tauler, J. Saurina, R. Bro, Anal. Chim. Acta. 398 (1999) 237-251.

[74] B. E. Wilson, E. Sanchez, B. R. Kowalski J. Chemom. 3 (1989) 493-498.

[75] B. E. Wilson, W. Lindberg, B. R. Kowalski, J Am. Chem. Soc. 111 (1989) 3797-3808.

[76] H. A. L. Kiers, A. K. Smilde, J. Chemom. 9 (1995) 179-195.

[77] A.K. Smilde, Y. Wang, B.R. Kowalski, J. Chemom. 8 (1994) 21-36.

[78] S. Li, J. C. Hamilton, P. J. Gemperline, Anal. Chem. 64 (1992) 599-607.

[79] N. M. Faber, L. M. C. Buydens, G. Kateman, J. Chemom. 8 (1994) 147-154.

[80] E. Sanchez, L. S. Ramos, B. R. Kowalski, J. Chromatogr 385 (1987) 151-164.

[81] E. Comas, R. A. Gimeno, J. Ferré , R. M. Marce, F. Borrull, F. X. Rius, J. Chromatogr. A 988 (2003) 277-284.

[82] C. G. Fraga, J. Chromatogr. A 1019 (2003) 31-42.

[83] B. J. Prazen, C. A. Bruckner, R. E. Synovec, B. R. Kowalski, J. Microcolumn Sep. 11 (1999) 97-107.

[84] L. S. Ramos, E. Sanchez, B. R. Kowalski, J. Chromatogr 385 (1987) 165-180.

[85] G. M. Gross, B. J. Prazen, R. E. Synovec, Anal. Chim. Acta 490 (2003) 197-210.

[86] C. G. Fraga, C. A. Bruckner, R. E. Synovec, Anal. Chem. 73 (2001) 675-683.

[87] M.L. Gianelli, D. H. Burns, J. B. Callis, G. D. Christian, N. H. Andersen, Anal. Chem. 55 (1983) 1858-1862.

[88] H. Abdollahi, F. Nazari, Anal. Chim. Acta 486 (2003) 109-123.

[89] Z. Zhu, J. Xia, J. Zhang, T. Li, Anal Chim. Acta 454 (2002) 21-30.

[90] B. Antalek, W. Windig, J. Am. Chem. Soc. 118 (1996) 10331-10332.

[91] W. Windig, B. Antalek, Chemom. Intell. Lab. Syst. 37 (1997) 241-254.

[92] W. Windig, B. Antalek, Chemom. Intell. Lab. Syst. 46 (1999) 207-219.

[93] W. Windig, B. Antalek, L. J. Sorriero, S. Bijlsma, D. J. Louwerse, A. K. Smilde, J. Chemom. 13 (1999) 95-100.

[94] S. Bijlsma, D. J. Louwere, W. Windig, A. K. Smilde, Anal. Chim. Acta 376 (1998) 339-355.

[95] S. Bijlsma, D. J. Louwere, A. K. Smilde, J. Chemom. 13 (1999) 311-329.

[96] S. Bijlsma, A. K. Smilde, J. Chemom. 14 (2000) 541-560.

Chapter 3

Practical aspects in the application of GRAM

[redacted]

3.1 INTRODUCTION

This chapter considers three aspects that must be taken into account in order to obtain accurate predictions with GRAM: (i) the alignment of chromatographic peaks, (ii) the selection of the number of factors to build the model and, (iii) the detection of outliers. The research developed to solve these aspects is presented as published papers. A paper in preparation is also included, which compares two strategies to determine the amount of noise in a chromatographic peak. This is needed in the outlier detection method. For simplicity, each section contains its own references.

3.2 ALIGNMENT OF CHROMATOGRAPHIC PEAKS

3.2.1 Introduction

To obtain acceptable predictions with GRAM, the analyte of interest must elute at the same retention time in the calibration and in the test sample [1-2]. In the cases studied in this thesis, the retention times varied a few seconds between different samples and runs. This difference was large enough to make the GRAM predictions incorrect. In the case described on page 70, a time shift of 2 seconds lead to prediction errors of 30%. This prediction error is also affected by the degree of overlap of the analyte of interest with interferences (section 3.4).

Several methods tackle the problem of peak alignment [3-8] in chromatography. This problem can also be encountered in other techniques and methodologies, like the optimization of batch processes [9-12] for which own methods for solving time shift between repetitions have been developed.

3.2.2 Experimental aspects that cause retention time shift

The main experimental aspects that can bring about retention time shift in HPLC are [13]:

1) Changes in the mobile phase composition caused by temperature and pressure fluctuations, variations in flow-rate and gradient dispersion.

2) Imprecise injection.

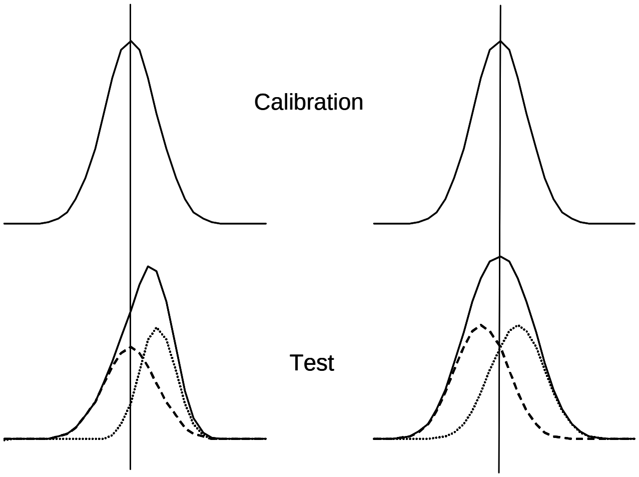
3) Degradation of the stationary phase.

4) Column overloading due to the over-injected amount or some components with a high concentration.

5) Possible interaction between analytes.

3.2.3 Detection and correction of the retention time shift

If the peaks of the calibration and test samples are pure, time shift can be detected by visually comparing the position of the maximum of each peak. In overlapped peaks, the chromatographic profile of the analyte of interest is unknown and the observed maximum corresponds to the sum of the analyte and the interference (Figure 4). Hence, visual inspection is not accurate enough to correct the time shift.



Observed:	shifted	non-shifted
Reality:	non-shifted	shifted

Figure 4. Visual inspection cannot detect time shift in overlapped peaks.

The methods available solve the time shift problem by selecting a time window for the peak of the calibration standard ($\mathbf{R}_c$) and different time windows for the peak of the test sample ($\mathbf{R}_t$). Then, an appropriate criterion is used to indicate in what particular $\mathbf{R}_t$ the underlying elution profiles are aligned with those in $\mathbf{R}_c$.

One criterion is to examine the elution profiles estimated by GRAM. Provided that the number of factors is correct, negative parts in the estimated profiles might indicate retention time shift. In which case, different time windows are tested for $\mathbf{R}_t$ until the shape of the profiles is as expected. However, sometimes these negative parts may not be significant. Even, small retention time shifts may not produce negative parts. Hence, this criterion is not accurate enough to detect time shift.

A second criterion is based on Bessel's inequality [14]. $\mathbf{R}_t$ is selected for different windows in the chromatogram. At each position, the SVD of $\mathbf{R}_t$ is calculated, to obtain the column space (spanned by the columns of **U**). A pure chromatographic profile from the calibration sample is projected on the column space and Bessel's inequality is calculated. Several positions are tested until the maximum in Besse'l inequality is found. In the non-shifted position, the **U** matrix fully explains the chromatographic profile of the calibration sample, i.e., the profile of the calibration sample is included in the test sample. When retention time shift exists, the pure chromatographic profile is not explained by the **U** space, and there is not such agreement. The major limitation of this method is that the pure chromatographic profile from a pure standard is needed. Moreover, this method may be influenced by the noise in the profile of the standard.

A third criterion, developed by Prazen et al. [15,16], consists of building an augmented matrix by adding column-wise the calibration sample and the test sample. This is repeated for different time windows of $\mathbf{R}_t$. The SVD of each augmented matrix is calculated and the singular values are studied by calculating the percentage of residual variance [15]. When the peaks are aligned, one singular value is associated to one analyte. When the peaks are not aligned, more singular values are needed. The percentage of residual variance for each singular value is represented against the time shift. When the peaks are aligned a minimum appears in the curve (Figure 5).

This method has been extensively used [17,18] and it is implemented in commercial software [19]. It works smoothly when the calibration peak is pure. However, when the calibration peak contains several analytes, which are also present in the test sample, several minima can appear (one for each analyte) that may lead to confusion. For example, in the paper presented in the section 3.2.6, the calibration sample contained two analytes, and when one analyte was aligned, the other was misaligned and vice versa.

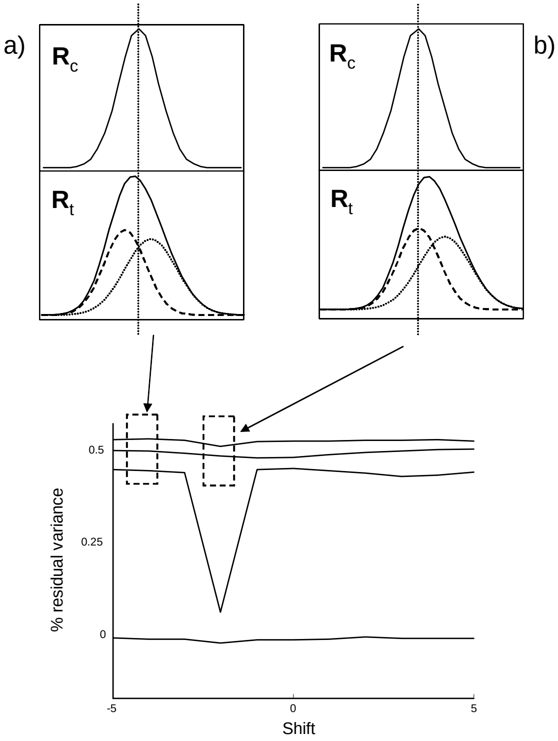


Figure 5. Representation of Prazen's time shift correction algorithm. $\mathbf{R}_c$ is a pure peak and $\mathbf{R}_t$ contains the analyte of interest plus an interference. (a) The analyte of interest (---) is shifted, and two singular values are needed to describe the analyte, plus one for the interference. (b) The analyte is aligned, hence one singular value describes the analyte, plus one for the interference.

3.2.4 A new method for correcting the time shift

We developed a method based on ITTFA (section 2.4.1). Both $\mathbf{R}_c$ and $\mathbf{R}_t$ are decomposed via ITTFA. The peak of the analyte of interest is identified in $\mathbf{R}_c$ and $\mathbf{R}_t$ and a time window is selected in $\mathbf{R}_t$ so that the peak in $\mathbf{R}_c$ and $\mathbf{R}_t$ is aligned. The advantage of this method is that the calibration peak can contain more than one analyte and that the correction is selectively done for the analyte of interest. This work is published in the paper shown in section 3.2.6.

3.2.5 References

[1] P.B. Poe, S.C. Rutan, Anal. Chim. Acta 283 (1983) 845-853.

[2] S. Li, P.J. Gemperline, K. Briley, S. Kazmierczak, J. Chromatogr. B 665 (1994) 213-233.

[3] P.J. Gemperline, J.H. Cho, B. Archer, J. Chemom. 13 (1999) 153-164.

[4] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, J. Chromatogr. A 961 (2002) 237-244.

[5] R.J.O. Torgrip, M. Alberg, B. Karlberg, S.P. Jacobson, J. Chemom. 17 (2003) 573-582.

[6] A. Bogomolov, M. McBrien, Anal. Chim. Acta 490 (2003) 41-58.

[7] G. H. Webster, T.L. Cecil, S.C. Rutan, J. Chemom. 3 (1988) 21-32.

[8] R. Andersson, M.D. Hämäläien, Chemom. Intell. Lab. Syst. 22 (1994) 49-61.

[9] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta 456 (2002) 77-92.

[10] E. Furusjo, L.G. Danielsson, Chemom. Intell. Lab. Syst. 50 (2000) 63-73.

[11] H. Ramaker, E. van Sprang, J.A. Westerhuis, A.K. Smilde, Anal. Chim. Acta 498 (2003) 133-153.

[12] M. Maeder, Y. Neuhold, A. Olsen, G. Puxty, R. Dyson, A. Zilian, Anal. Chim. Acta 464 (2002) 249-259.

[13] F. Gong, Y. Liang, Y. Fung, F. Chau, J. Chromatogr. A 1029 (2004) 173-183.

[14] B. Grung, O.M. Kvalheim, Anal. Chem. 304 (1995) 57-66.

[15] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218-225.

[16] B.J. Prazen, C.A. Bruckner, R.E. Synovec, B.R. Kowalski, J. Microcolumn Sep. 11 (1999) 97-107.

[17] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 73 (2001) 5833-5840.

[18] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 72 (2000) 4154-4162.

[19] PLS_Toolbox 3.0 for Matlab. Eigenvector Research Inc.

3.2.6 Paper

E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius

Time shift correction in second-order liquid chromatographic data with iterative target transformation factor analysis

Analytica Chimica Acta 470 (2002) 163

Time shift correction in second-order liquid chromatographic data with iterative target transformation factor analysis

Enric Comas, R. Ana Gimeno, Joan Ferré, Rosa M. Marcé
Francesc Borrull, F. Xavier Rius

Department of Analytical and Organic Chemistry, Rovira i Virgili University
Pl. Imperial Tarraco 1, 43005, Tarragona, Spain

ABSTRACT

When the generalized rank annihilation method (GRAM) is applied to liquid chromatographic data with diode-array detection, an important problem is the time shift of the peak of the analyte in the test sample. This problem leads to erroneous predictions. This time shift can be corrected if a time window is selected so that the chromatographic profile of the analyte in the test sample is trilinear with the peak of the analyte in the calibration sample. In this paper we present a new method to determine when this condition is met. This method is based on the curve resolution with iterative target transformation factor analysis (ITTFA). The calibration and test matrices are independently decomposed into profiles and spectra, and aligned before GRAM is applied. Here we study two situations: first, when the calibration matrix has one analyte and second, when it has two analytes. When the calibration matrix has two analytes, we selectively determine the time window for the analyte to be quantified. There were considerably fewer prediction errors after correction.

Keywords: HPLC, Time shift, ITTFA, Time window, Selectivity, GRAM

1. INTRODUCTION

With high performance liquid chromatography (HPLC), we can separate and quantify many analytes in a sample in one single analysis. Despite the great effort involved in optimizing the chromatographic conditions of the separation with standard solutions, new test samples, especially natural samples, such as river water, may contain unknown compounds that may overlap with the analytes of interest. Modifying the parameters of the chromatographic method to avoid such interference can be very costly and time consuming, and it is not always the best choice if it has to be done for each new test sample. The necessary selectivity, without the complete separation of the interferences, can be mathematically achieved by calibrating with second-order data. This type of data can be obtained, for instance, by HPLC with diode array detection (DAD). A matrix of responses is obtained for each chromatographic peak by recording the spectrum of the eluting compounds at each retention time.

Of the second-order calibration algorithms that allow quantification in the presence of non-calibrated components (known as the 'second-order' advantage) [1], the generalized rank annihilation method (GRAM) [2] requires only two matrices: one from the calibration sample (either a pure standard or a real sample with known concentration of the analyte) and one from the test sample. This makes GRAM a very useful quantification method for chromatographic data when the number of analyses is important. However, GRAM has a serious limitation in routine chromatographic analysis: the data matrices containing the peak of the analyte in the calibration sample and in the test sample must be trilinear [3, 4], i.e. the chromatographic profiles of the analytes of interest in the test sample and in the calibration sample must be proportional. This means that in both samples the analyte must elute at the same time, which is not so common in practice because imprecision in injection timing, fluctuations in temperature and changes in flow rate introduce time shifts in the peaks. Although other calibration methods that are robust to time shift have been developed [5, 6], GRAM has been widely studied [4, 7-11]. Expressions are available for calculating figures of merit [12], such as sensitivity, selectivity and limit of detection, as well as for removing the bias in the predictions and calculating the variance of the predicted concentrations [13].

Some algorithms have been developed to correct the time shift in second-order data by selecting the right time window for the test matrix [14-16]. The crucial point is to define the criterion that indicates when the profiles of the analyte of interest in the two matrices coincide. Prazen et al. [14] used the calibration sample matrix augmented with the peak of the test sample. The eigenvalues of the augmented matrix were calculated and plotted for different time windows of the test sample. A minimum in the plot indicates the optimal window. This method gives unique solutions when the calibration sample is a pure peak, i.e. the standard of the analyte of interest. However, when the peak of interest in the calibration matrix overlaps with others also present in the test matrix, such as when the calibration sample is the spiked test sample, it may show more than one minimum and indeterminations may appear.

This paper presents a new method for correcting the time shift of second-order HPLC-DAD data, based on the curve resolution of the peaks. We use it to determine the concentration of three polycyclic aromatic hydrocarbons, whose peaks elute overlapped in the chromatographic analysis, from a mixture.

2. THEORY

2.1. Notation

Boldface uppercase letters represent matrices, e.g. **A**; italic letters represent scalars, e.g. *a*; superscript `T' represents transposition.

2.2. Correction procedure

When GRAM is applied to HPLC-DAD data, it is assumed that the $J_1 \times J_2$ matrices of measured responses of the calibration ($\mathbf{R}_c$) and test ($\mathbf{R}_t$) samples can be expressed as:

$$\mathbf{R}_c = \mathbf{X}\mathbf{C}_c\mathbf{Y}^T + \mathbf{E}_c \quad (1)$$

$$\mathbf{R}_t = \mathbf{X}\mathbf{C}_t\mathbf{Y}^T + \mathbf{E}_t \quad (2)$$

where the columns of $\mathbf{X}$ ($J_1 \times K$) and $\mathbf{Y}$ ($J_2 \times K$) are the normalized profiles and the normalized spectra, respectively, K the total number of constituents in both matrices, $\mathbf{C}_c$ and $\mathbf{C}_t$ are $K \times K$ diagonal matrices of concentration-related scale factors, and $\mathbf{E}_c$ and $\mathbf{E}_t$ are error matrices. J_1 is the number of spectra in the time window where the analyte of interest is included, and it is determined for the calibration sample.

In the method we propose, iterative target transformation factor analysis (ITTFA) [17-19] is used to decompose each individual matrix into the profiles and spectra of the analytes. In this case, the test matrix $\mathbf{R}_{t,TW2}$ ($J_{1,TW2} \times J_2$) (the subscript 'TW2' indicates that this is not the same matrix as $\mathbf{R}_t$ in Eq 2 is initially selected from the chromatogram so that its time window is arbitrarily wider than for the calibration matrix $\mathbf{R}_c$, i.e. $J_{1,TW2} > J_1$. This ensures that $\mathbf{R}_{t,TW2}$ includes the peak of the analyte to be quantified. Then $\mathbf{R}_c$ and $\mathbf{R}_{t,TW2}$ are individually decomposed with ITTFA:

$$\mathbf{R}_c = \mathbf{H}_c \mathbf{Y}_c^T + \mathbf{E}_c \tag{3}$$

$$\mathbf{R}_{t,TW2} = \mathbf{H}_t \mathbf{Y}_t^T + \mathbf{E}_{t,TW2} \tag{4}$$

where the columns of $\mathbf{H}_c$ ($J_1 \times K$) and $\mathbf{H}_t$ ($J_{1,t} \times K_t$) are not normalized profiles and a different number of analytes may be found for each matrix. Then, $\mathbf{H}_c$ and $\mathbf{H}_t$ are plotted and the profile of the analyte of interest is identified by comparing the calculated spectra $\mathbf{Y}_c$ and $\mathbf{Y}_t$ with the spectrum of the pure analyte. The maximum of each profile is separated by Δt time steps. The final $\mathbf{R}_t$ ($J_1 \times J_2$) is selected starting at an elution time that is Δt from the starting elution time of $\mathbf{R}_c$. This correction improves the trilinearity of the data and GRAM can be applied with more guarantees.

3. EXPERIMENTAL SECTION

3.1. Chemicals and samples

We studied three analytes: (A) Benzo[*b*]fluoranthene, from Aldrich Chemie (Beere, Belgium); (B) benzo[*k*]fluoranthene, from Fluka (Buchs, Switzerland) and (C) benzo[*a*]pyrene, from Sigma (Alcobendas, Spain), all with a purity of over 98%. Standard solutions of each compound at a concentration of 500 mg l^{-1} were prepared in HPLC-gradient grade acetonitrile (SDS, Peypen, France) and stored at 4 °C. All the working solutions used in this study were prepared by dilution. We analyzed standards of each analyte, as well as mixtures of two components (A+B) and three components (A+B+C). These working solutions contained the compounds at a concentration of 1 mg l^{-1}.

3.2. Instrumental

We used an HP1100 series HPLC system (Agilent technologies, Waldbronn, Germany) for the analysis. This consisted of a degasser, a binary pump, an oven, a diode-array detector (DAD) and a manual injector with a 20 μl-loop. The chromatographic column was a 15cm × 0.46 cm Eclipse XDB-C8 with a 5 μm particle size (Hewlett-Packard, Barcelona, Spain). Acetonitrile was the mobile phase. This was delivered at a flow rate of 1.5 ml min^{-1} and the column temperature was 40 °C.

For detection, the spectra were recorded between 220 and 300 nm, every 0.4 nm. A spectrum was collected every 0.4 s, i.e. five spectra were measured every 2 s. Data were recorded from 0 to 2 min. In these conditions, the analytes eluted approximately from 1.29 to 1.55 min (from 77 to 93 s).

In routine analysis of natural samples, it is not unusual for the analyte of interest to coelute with interferences. Since the samples used here were synthetic, we created the necessary data by selecting the chromatographic conditions so that the analytes eluted overlapped. Although overlapping was forced, the time shift between the successive analyses was not created artificially and was the result of repeated injections.

3.3. Software

The GRAM routine belongs to the N-way toolbox of R. Bro and C. Andersson and was downloaded from their website [20]. The ITTFA algorithm and the shift correction algorithm were made in house subroutines for MATLAB version 6 [21].

3.4. Data analysis

Our objectives were to quantify B in the test mixture A+B, using the standard of analyte B as a calibration sample, and to quantify B in the test mixture A+B+C using either the standard of analyte B or the mixture A+B as a calibration sample. In this second case, we also determined the concentration of A. We analyzed each sample three times to estimate the variability of the experiment.

We considered two time windows. Time window TW1 was from 1.29 to 1.55 min (from 77 to 93 s) and was always used for $\mathbf{R}_c$ (i.e. J_1=40). Time window TW2 was wider from 1.25 to 1.60 min (from 74 to 96 s) and was used for $\mathbf{R}_{t,TW2}$ (i.e. $J_{1,TW2}$=50).

In the three situations we studied, we first applied GRAM to the measured data using time window TW1 for both $\mathbf{R}_c$ and for $\mathbf{R}_t$, which is the desired calibration situation. We then selected $\mathbf{R}_{t,TW2}$ with time window TW2 and corrected the time shift. We then applied GRAM for the test matrix that had been corrected. We validate the method by analyzing the improvement in the prediction errors and by comparing the spectra and chromatographic profiles calculated with GRAM with real ones from the pure standards. We also show the results of quantification using the areas of the peaks obtained with ITTFA, which is insensitive to time shift (since each matrix is tackled separately).

4. RESULTS AND DISCUSSION

Fig. 1 shows the measured chromatographic profiles of the three pure analytes (A, B and C), of a mixture of two components (A+B) and of a mixture of three components (A+B+C) at the interval described by time window TW1. The selected wavelength (254 nm) is the usual one for determining these analytes [22]. The analytes in the eluted mixtures overlapped from the chromatographic column. This lack of resolution cannot be avoided by selecting a different wavelength, since the three analytes absorb at all the recorded wavelengths (see Fig. 2).

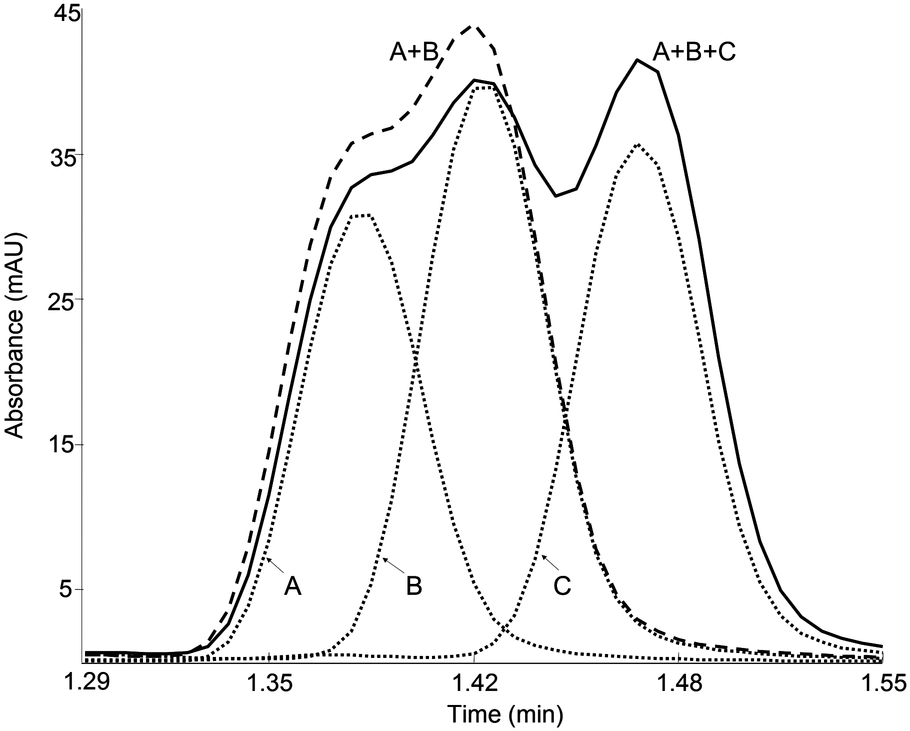


Fig. 1. Chromatographic profiles measured at 254 nm. (. . .) Individual standards (A, B, C); (- - -) mixture A+B; (–) mixture A+B+C. The concentration of each analyte is 1 mg l^{-1} in all samples.

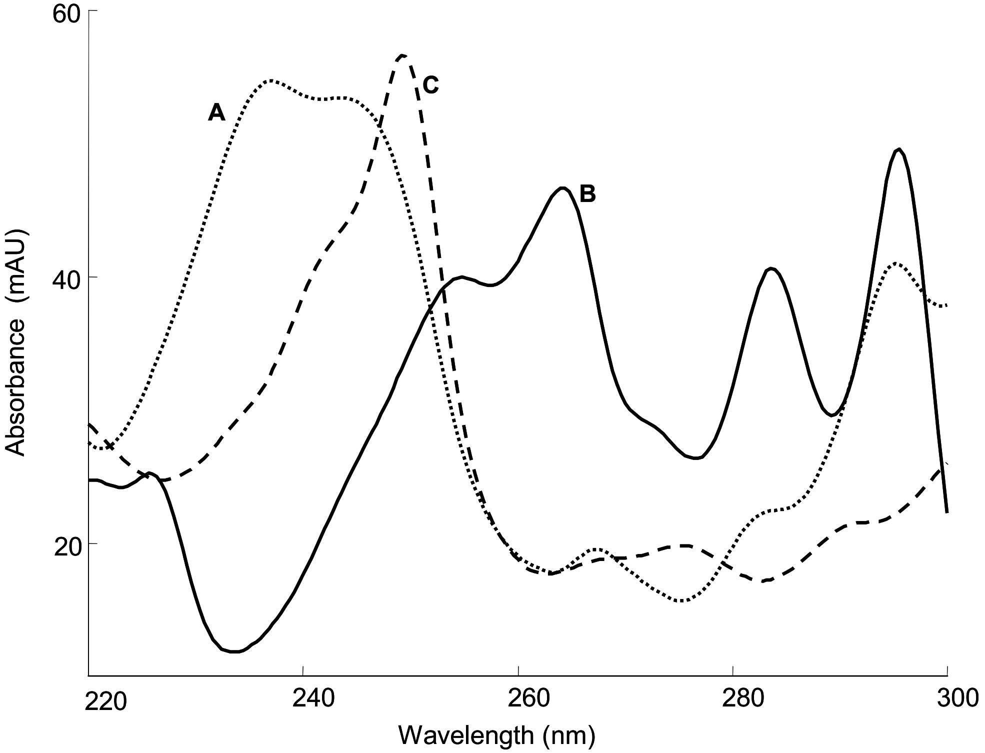


Fig. 2. Spectra of analytes (. . .) A, (−) B and (- - -) C at 1 mg l^{-1}.

4.1. Quantification of B in the mixture A + B using standard B as a calibration matrix

Fig. 3 shows the calibration matrix (standard B, $\mathbf{R}_c$) and the test matrix (mixture A+B, $\mathbf{R}_t$) in TW1 conditions. The overlap in $\mathbf{R}_t$ makes it difficult to identify the position of the analyte of interest (B) and difficult to detect and correct any time shift.

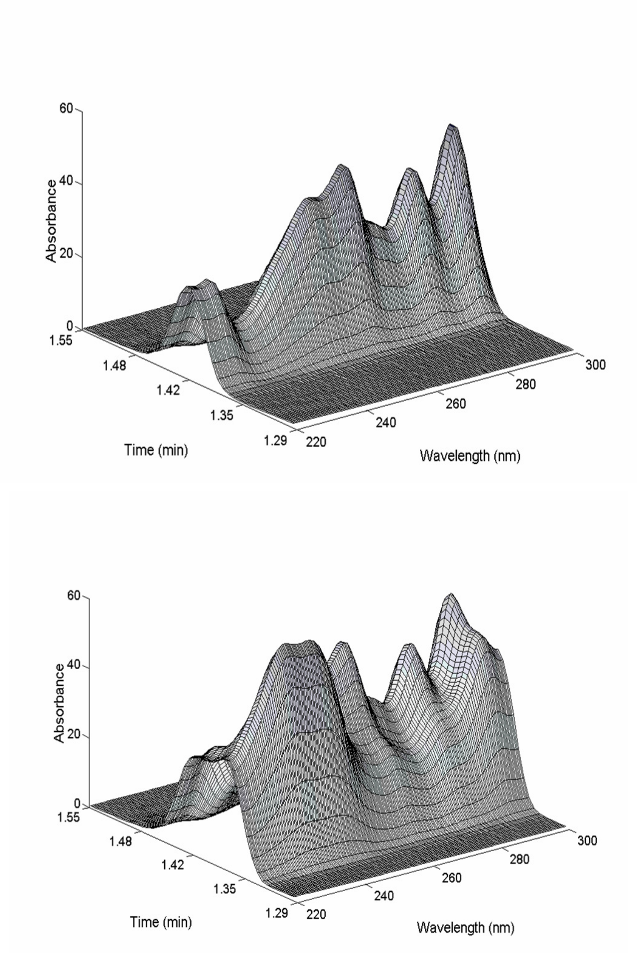


Fig. 3. Measured matrices, (a) calibration sample $\mathbf{R}_o$ (B); (b) test sample $\mathbf{R}_t$ (A+B) in TW1 conditions.

Fig. 4 shows the chromatographic profiles of analyte B calculated by ITTFA for $\mathbf{R}_c$ and the profiles of A and B calculated for $\mathbf{R}_{t,TW2}$. The profiles were assigned to the analytes by comparing the calculated spectra with the spectrum of the standard (the correlation coefficient in all cases was higher that 0.99). We can see that the profile of the analyte B in $\mathbf{R}_{t,TW2}$ is shifted in Δt=1 time steps to a higher time than B in $\mathbf{R}_c$. Therefore, we finally selected matrix $\mathbf{R}_t$ from the chromatogram starting at one time step shifted in relation to TW1. In this way, the maximum of the profile for analyte B in the test matrix (thick dashed line) calculated with ITTFA coincides with that in the calibration matrix.

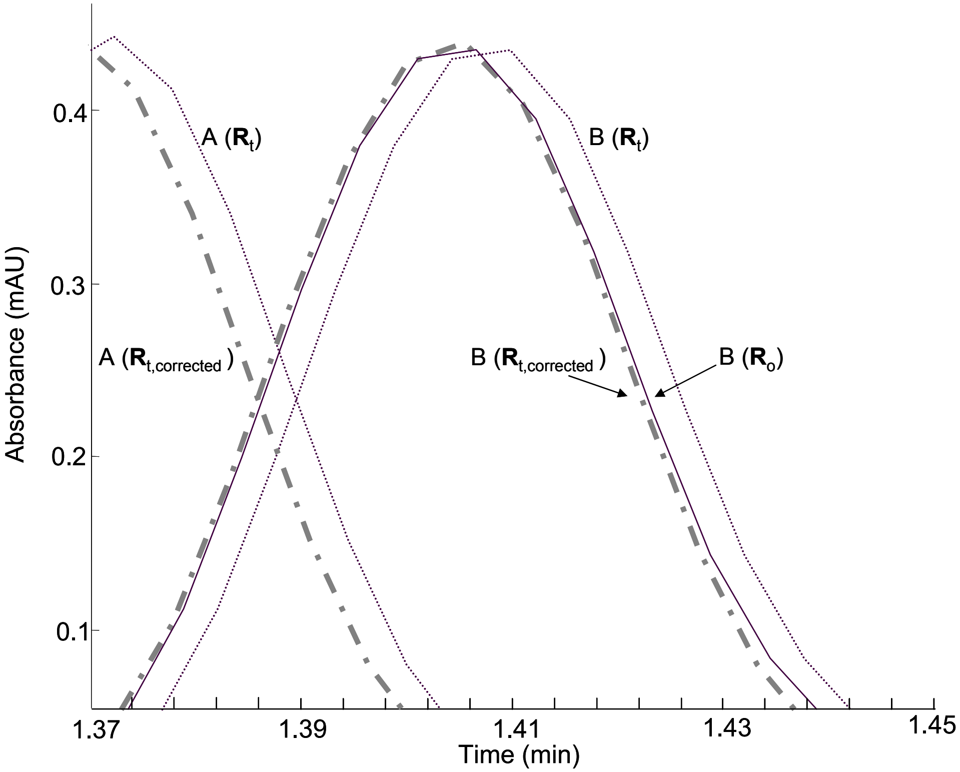


Fig. 4. (−) ITTFA calculated profile of B in the calibration matrix ($\mathbf{R}_c$) before time shift correction; (. . .) ITTFA calculated profiles of A and B in the prediction matrix ($\mathbf{R}_{t,TW2}$); (- · - · -) ITTFA calculated profiles of A and B in the test matrix after the time shift correction ($\mathbf{R}_t$ corrected).

Table 1 shows the relative percentage error in the concentration predicted by GRAM applied to the data before shift correction (BSC) (conditions TW1 for $\mathbf{R}_c$ and $\mathbf{R}_t$), and applied to the shift-corrected data (SC) for three repeated measurements of the mixture A+B. As expected, the errors always decreased after shift correction. The improvement was greater when the measured peaks were shifted two time steps (repetition number 3). After the correction procedure, the calculated error in the three cases was of the same magnitude as the variability in the different repetitions.

Table 1. Relative percentage error in the GRAM prediction of analyte B in the mixtures A+B and A+B+C using standard B as a calibration sample

Mixture (test sample)	Repetition	% error BSC	Δt	% error SC	% error ITTFA
A+B	1	1.7	1	0.2	2.7
	2	1.8	1	0.2	1.9
	3	3.2	2	0.4	1.8
A+B+C	1	5.9	1	1.2	3.4
	2	6.5	2	0.2	3.4
	3	6.5	1	0.9	0.3

Δt: time shift (number of units) determined by ITTFA; BSC: data before shift correction; SC: shift corrected data; % Error ITTFA: quantification using the areas calculated by ITTFA.

4.2. Quantification of B in the mixture A + B + C using standard B as a calibration matrix

Table 1 also shows the results of predicting B in the mixture A+B+C. Initially, the prediction errors were as high as 6.5%. This was mainly because the peaks of the analytes overlapped a great deal, and small differences in the time shift produced large errors in GRAM. The ITTFA results show that the shift was only 0.4 s (one time step), which shows that the sensitivity of GRAM to the shift in the data is significant. The prediction errors dropped to as low as 0.2% after the right time window for $\mathbf{R}_t$ was determined.

With curve resolution methods, we can also quantify the analyte of interest from the area of the resolved peaks [23]. To do this, we compared the areas under the profiles calculated by ITTFA with the areas for the pure standards. Here, time shift was not a disadvantage because the two samples were dealt separately. In this case, we considered TW1 conditions for both $\mathbf{R}_c$ and $\mathbf{R}_t$. The results are also shown in Table 1. For the mixture A+B, the prediction errors were of the same magnitude as those obtained when we applied GRAM before correcting the time shift. For the mixture A+B+C they were not so high, but they were still higher than the errors for GRAM after the time shift had been corrected. These results may be due to the fact that ITTFA is not a calibration method but a curve resolution method with rotational and scale indeterminations. Moreover, the exact profiles cannot always be satisfactorily calculated by ITTFA, because the shape of the peak, and therefore its area, varies according to the number of factors used for ITTFA, and because the different resolution of the peaks affects the value of the area. The low error for repetition number 3 (0.3) was attributed to chance. All of these problems restrict correct quantification by the ITTFA calculated profiles.

4.3. Quantification of A and B in the sample A + B + C using sample A + B as a calibration matrix

As well as calibrating with the pure standard, we can also use GRAM to quantify several coeluting analytes of the test sample in the same analysis with a calibration matrix that also contains all the analytes to be quantified.

Time shift correction when the calibration matrix has several coeluting analytes that are also present in the test sample is more difficult than in the previous situation because the relative retention time between the analytes may vary slightly from one sample to another. In this case, the shift can be corrected for each individual analyte. The next example shows that with ITTFA, only the analyte of interest can be corrected. The calibration matrix contains analytes A+B. The test matrix contains analytes A+B+C, where A and B must be quantified and C is interference.

Table 2 shows the error in the predicted concentrations for three repetitions of the calibration sample and two repetitions of the test sample. In all cases, the results were significantly repetitive. The difference was due to random effects, such as noise or variance in the overlap of the test analytes. When we applied GRAM to the matrices in TW1 conditions for $\mathbf{R}_c$ and $\mathbf{R}_t$, the prediction error was up to 6.5% for analyte A and up to 20.4% for analyte B. The large error for analyte B shows that the data do not follow (1) and (2). When the time shift correction was applied for analyte A, we detected no shift between the different matrices. We were unable, therefore, to improve the predictions with GRAM, so they are not shown here. For analyte B, the shift was detected by ITTFA and, once the difference in the elution time was corrected, the percentage of prediction error dropped to around 4%.

However, the error in the concentration of A increased to 30%. This shows that a global correction is not possible and the correction has to be done for each analyte. It also shows that when one profile is aligned, the other is misaligned because the matrix is `moved'.

Table 2. Relative percentage error in the GRAM prediction of A and B in a mixture A+B+C using mixture A+B as a calibration sample

Calibration sample (number of repetition)	Prediction sample (number of repetition)	BSC			Corrected B	
		A (% error)	B (% error)	Δt	A (% error)	B (% error)
1	1	3.9	13.5	1	11.3	4.2
	2	6.3	20.4	2	23.1	0.1
2	1	4.5	18.0	2	26.0	2.3
	2	6.5	9.0	2	30.2	6.0
3	1	0.4	16.3	3	33.1	4.4
	2	4.6	8.5	3	25.9	3.6

Δt: time shift (number of units) determined by ITTFA; BSC: data before shift correction.

We validated these results by comparing the chromatographic profiles and the spectra calculated by GRAM before and after the time shift correction. The chromatographic profile of analyte B calculated with GRAM (Fig. 5a) had a notable negative part, which is impossible for this type of analysis. This suggests that the data do not follow the trilinear model assumed by GRAM. One important reason for this is the time shift of the peaks of $\mathbf{R}_t$ with respect to $\mathbf{R}_c$. The calculated spectrum of B (Fig. 5b) was also quite different from the spectrum of the standard (Fig. 2). Fig. 5c and d show the chromatographic profiles and the spectra after the shift has been corrected. Now the profile for analyte B has only one maximum and no negative part, which is the expected shape for this kind of data. The calculated spectrum was also like the measured spectrum from pure B. The negative part observed for the chromatographic profile of analyte A was due to the fact that this peak was misaligned when analyte B was corrected for the time shift. This agrees with the large errors for A. These results show that the shift for each analyte to be quantified needs to be corrected before GRAM is applied.

The success of the method will depend on the ability of ITTFA for decomposing the data matrix according to Eq. (3). Problems can arise when the peak of the analyte of interest is either highly overlapped with or embedded in the peaks of the interferences and its response is relatively low compared to the response of the interferences. In addition, high collinearities between the spectra of the analytes will make the identification of the analyte of interest more ambiguous.

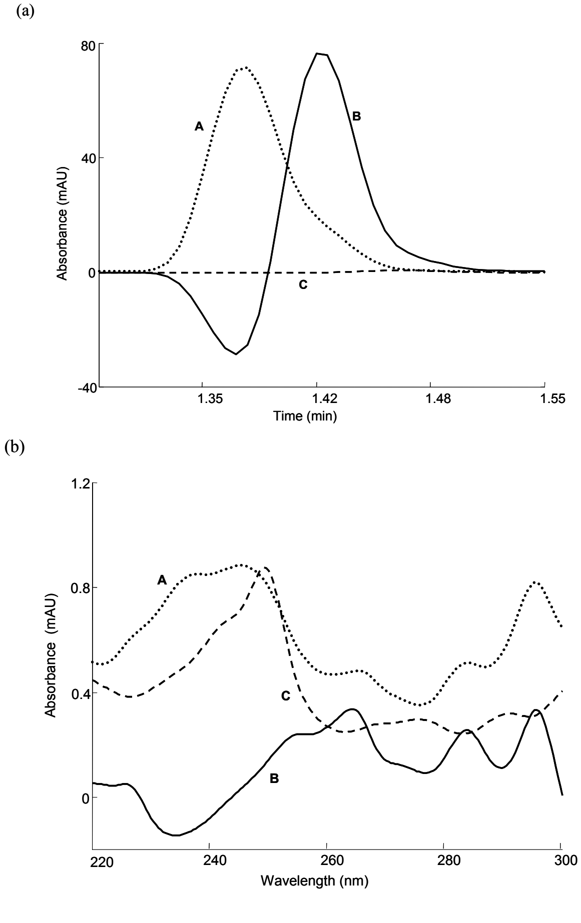


Fig. 5. Chromatographic profiles and spectra calculated with GRAM. Before shift correction (a and b), and after shift correction (c and d).

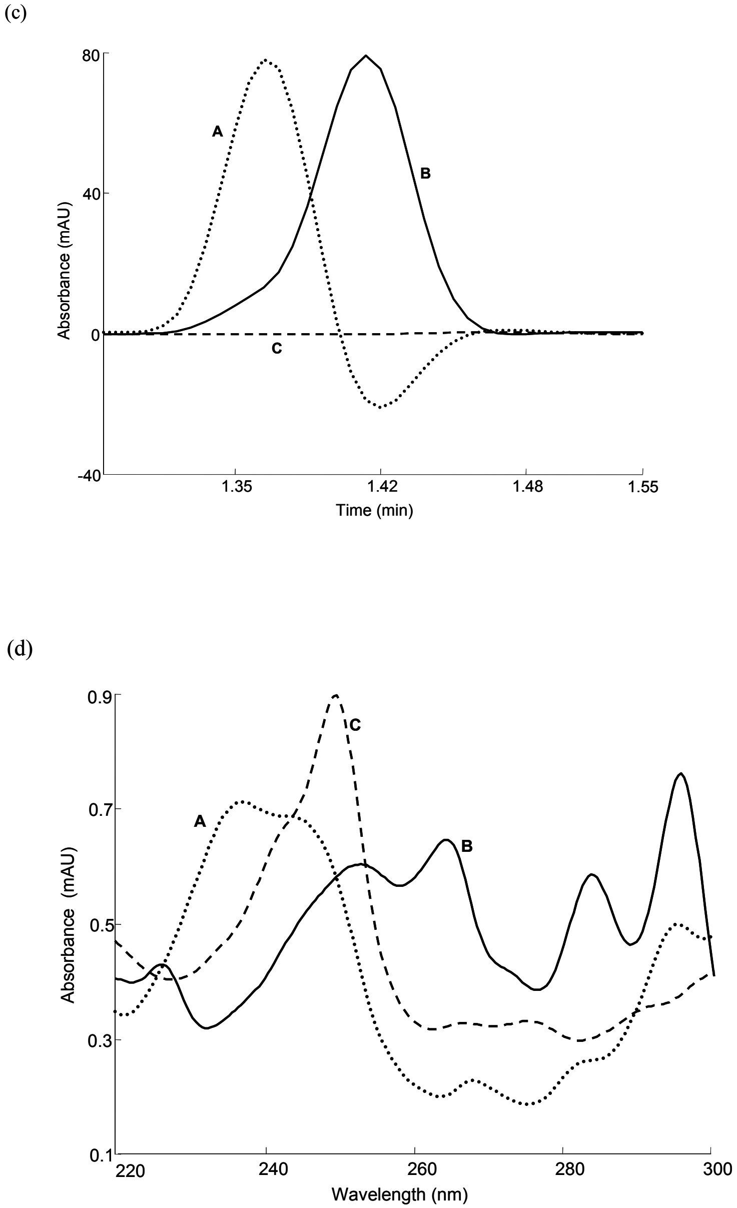


Fig. 5. *(Continued)*

5. CONCLUSIONS

We present a time shift correction method for second-order liquid chromatographic data based on determining the right time window. The correction is made after the calibration and test matrix are individually decomposed by ITTFA. This method can selectively correct the analyte of interest, thus making the corrected results more precise. Although it may not completely correct the lack of trilinearity, it can improve it so that prediction errors are lower. Variations of this procedure could be based on other curve resolution methods, which also decompose the individual matrices into profiles and spectra.

ACKNOWLEDGEMENTS

The authors would like to thank the Spanish Ministry of Education, Culture and Sports (Project BQU 2000-1256) for economic support and the CIRIT, of the Catalan Government, for providing Enric Comas with a doctoral fellowship.

REFERENCES

[1] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) A782.

[2] E. Sánchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496.

[3] R.B. Poe, S.C. Rutan, Anal. Chim. Acta 283 (1993) 845.

[4] L. Scott Ramos, E. Sánchez, B.R. Kowalski, J. Chromatogr. 385 (1987) 165.

[5] R. Bro, C.A. Andersson H.A.L. Kiers, J. Chemom. 13 (1999) 295.

[6] J. Saurina, S. Hernandez-Cassou, R. Tauler A. Izquierdo-Ridorsa, Anal. Chem. 71 (1999) 2215.

[7] N.M. Faber, L.M.C. Buydens, G. Kateman, J. Chemom. 8 (1994) 147.

[8] N.M. Faber, L.M.C. Buydens, G. Kateman, J. Chemom. 8 (1994) 181.

[9] N.M. Faber, L.M.C. Buydens, G. Kateman, J. Chemom. 8 (1994) 273.

[10] M.J.P. Gerritsen, H. Tanis, B.G.M. Vandeginste, G. Kateman, Anal. Chem. 64 (1992) 2042.

[11] E. Sánchez, L. Scott Ramos, B.R. Kowalski, J. Chromatogr. 385 (1987) 151.

[12] N.M. Faber, R. Boqué, J. Ferré, Chemom. Intell. Lab. Syst. 55 (2001) 91.

[13] N.M. Faber, J. Ferré, R. Boqué, Chemom. Intell. Lab. Syst. 55 (2001) 67.

[14] B. J Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218.

[15] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Anal. Chem. 70 (1998) 2796.

[16] B. Grung, O.M. Kvalheim, Anal. Chim. Acta 304 (1995) 57.

[17] P.J. Gemperline, Anal. Chem. 58 (1986) 2656.

[18] B.G.M. Vandeginste, F. Leyten, M. Gerritsen M, J.W. Noor, G. Kateman, J. Frank, J. Chemom. 1 (1987) 57.

[19] P.K. Hopke, Chemom. Intell. Lab. Syst. 6 (1989) 7.

[20] http://www.models.kvl.dk/source/nwaytoolbox, June 2002.

[21] MATLAB, The Mathworks, South Natick, MA, USA.

[22] F. Sun, D. Littlejohn, M.D. Gibson, Anal. Chim. Acta 364 (1988) 1.

[23] J. Saurina, S. Hernández-Cassou, R. Tauler, A. Izquierdo-Ridorsa, Anal. Chem. 71 (1999) 126

3.3 SELECTION OF THE NUMBER OF FACTORS

3.3.1. Bibliographic revision

GRAM requires the number of factors to be specified. This number will be the number of columns in the calculated matrices of the elution profiles (**H**) and spectra (**Y**).

A factor (or latent variable), as used in first-order and second-order calibration, is a variable made by linearly combining variables [1-3]. Such a linear combination describes a systematic variation in the chromatographic peak caused by an eluting analyte or a baseline change.

Several methods have been developed to determine the number of factors to be used in second-order calibration methods. Some authors studied the lack of fit in the PARAFAC model when it was unfolded in the different directions, and different number of factors were considered [4-6]. Dable and Booksh [7] tested different kinds of noise distribution to determine the number of factors. Malinowski [3] developed statistical tests, like the F-test, to determine the number of significant factors from the SVD decomposition. The number of factors can also be found by checking the lack of fit of the reconstructed data from the GRAM estimations and the measured data. However, no direct information about the analyte of interest is obtained. The general fit may be unacceptable, but the fit exclusive for the analyte of interest be sufficient. This is the case, for example when one interferent is in both the calibration and the test sample, and the interference is not trilinear, whereas the analyte of interest is trilinear. In this case, GRAM predicts correctly, despite of the significant difference between the reconstructed data from the GRAM predictions and the measured data.

Gerritsen et al. [8] used the correlation of the estimated profiles $\mathbf{h}_i$ and $\mathbf{y}_i$ with library profiles $\mathbf{h}_k$ and $\mathbf{y}_k$ as a criterion:

$$S = \sum_{k=1}^{K} [\ \mathrm{corr}\,(\mathbf{h}_{i,}\, \mathbf{h}_{k})^2 + \mathrm{corr}\,(\mathbf{y}_{i,}\, \mathbf{y}_{k})^2]$$

where corr(·) is the correlation coefficient between the two vectors and $\mathbf{h}_k$ and $\mathbf{y}_k$ are the concentration profile and UV spectrum of compound *k* that can be obtained, for example, from the HPLC-UV data of a single component. The number of principal components is the one that gives the highest value of *S*.

3.3.2 Graphical criterion to determine the number of factors in GRAM

Faber et al. [9] used a weight parameter (α) in the GRAM algorithm in order to calculate and correct the bias in the predictions. We can use this parameter for selecting the number of factors.

When the peaks are correctly aligned (i.e., trilinear data) and the right number of factors is selected, all the GRAM models with a different value of α will predict the same concentration for the test sample. However, when trilinearity is not fulfilled or the number of factors is not correct, GRAM yields different predictions when α varies. This trend can be followed graphically, representing the evolution of the predictions for different number of factors and different values of α. This is shown in the paper in section 3.3.4.

3.3.3 References

[1] A. Lorber, Anal. Chem. 57 (1985) 2395-2397.

[2] N.M. Faber, A. Lorber, B. R. Kowalski, J. Chemom. 11 (1997) 419-461.

[3] E.R. Malinowski, Factor Analysis in Chemistry, 3rd ed, John Wiley & Sons Inc, New York, 2002.

[4] H. Xie, J. Jiang, N. Long, G. Shen, H. Wu, R. Yu, Chemom. Intell. Lab. Syst. 66 (2003) 101-115.

[5] Z. Chen, Y. Liang, Y. Li, H. Qian, R. Yu, J. Chemom. 13 (1995) 15-30.

[6] Z. Chen, Z. Liu, Y. Cao, R. Yu, Anal. Chim. Acta, 444 (2001) 295-307.

[7] D.K. Dable, K.S. Booksh, J. Chemom. 15 (2001) 591-613.

[8] M. J. P. Gerritsen, H. Tanis, B. G. M. Vandeginste, G. Kateman, Anal. Chem. 64 (1992) 2042-2056.

[9] N.M. Faber, J. Ferré, R. Boqué, F.X. Rius, Chemom. Intell. Lab. Syst. 55 (2001) 67-90.

3.3.4 Paper

Graphical criterion for assessing trilinearity and selecting the optimal number of factors in the generalized rank annihilation method using liquid chromatography-diode array detection data

E. Comas, J. Ferré, F.X. Rius

Analytica Chimica Acta 515 (2004) 23-30

Graphical criterion for assessing trilinearity and selecting the optimal number of factors in the generalized rank annihilation method using liquid chromatography–diode array detection data

Enric Comas, Joan Ferré, F. Xavier Rius

Department of Analytical and Organic Chemistry. Rovira i Virgili University
Pl. Imperial Tarraco 1, 43005, Tarragona, Spain

ABSTRACT

A weight parameter (α) was introduced into the Generalized Rank Annihilation Method (GRAM) to calculate and reduce the bias in the predicted concentration. Here we show that α can be used as an indicator to determine whether the trilinearity assumptions are met and to select the right number of factors to calculate the model.

The procedure is to calculate several GRAM models by varying α and the number of factors. When the experimental data are trilinear and the right number of factors is used, α does not affect the predicted concentration. If the condition of trilinearity is not met or the correct number of factors is not used, the predicted concentration changes when α changes. A graph shows this behavior.

Both simulated and real data were checked for trilinearity. Deviations from the ideal mathematical model, such as the time shift or different shape of the chromatographic profiles, were simulated. These parameters have an effect but the greatest effect was produced by the selection of the number of factors.

Keywords: GRAM, Trilinearity, Number of factors, Weight parameter, Graphical criterion.

1. INTRODUCTION

In the analysis of environmental and biological samples by high performance liquid chromatography (HPLC), the compound to be quantified sometimes elutes overlapped with interferences. There are two main ways to properly quantify this compound. One is to change the chromatographic conditions. This implies spending time and resources because it must be done for that specific analyte in that specific sample, and not always successful results are obtained. Moreover, the new optimized conditions may not be suitable for another sample that may contain a different interferent. Another way is to use second-order calibration methods to quantify overlapped peaks.

Second-order calibration uses second-order data, which can be easily obtained by using an HPLC–diode array detector (DAD) instrument. A spectrum is measured at each retention time and a matrix of responses is therefore obtained for each analyzed peak.

The special features of the generalized rank annihilation method (GRAM) [1-3] make it a particularly suitable method for extracting information from this kind of data. Only two samples, a calibration sample and a test sample, are required. The calibration sample peak can be one of the standards used in the optimization of the chromatographic process. The test sample peak is the already measured, overlapped one, so more experimental work is not needed. Also, in comparison with other second-order calibration methods, figures of merit such as the sensitivity and the limit of detection can be easily calculated [4,5].

However, GRAM has some mathematical requirements before proper quantification can be carried out. Firstly, and most importantly, the data must follow a trilinear model (see Section 2). Experimental factors such as the time shift and the variation in the shape of the profiles introduce non-trilinearities into the data. Secondly, a suitable number of factors must be selected in order to calculate the predictions. This number is related to the number of analytes in both peaks and it is required to separate the signal into the systematic part of the data, described by the model, and the random part included in the residuals.

Several methods have been proposed to determine the number of factors in second-order calibration [6-10]. These algorithms are mainly applied to the parallel factor analysis (PARAFAC) model, where more than two samples are used. In brief, they study the change of fit in the model used (for example, PARAFAC) when different models are tested by considering different subsets (split-half analysis) or when the structure of the data is unfolded or not. In Ref. [10], the number of factors is set after different kinds of noise structure are added to the data.

Here we present a graphical criterion for checking the trilinearity of the data and choosing the right number of factors for the GRAM model. Faber et al. [5] introduced a weight parameter (α) into the algorithm to determine the bias and figures of merit. α has no effect in the predicted concentration only if the data are trilinear and a suitable number of factors is used. By studying the variation in the predicted concentration when calculating several GRAM models for several values of α and numbers of factors, we obtain an indication of the trilinearity of the data. If the predicted concentration depends on the value of α, GRAM produces misleading results. Some solutions will be reported to act in this case.

We will use simulated HPLC-DAD data to show how different experimental aspects affect the results of GRAM and how they are detected by varying α. We will study the influence of the high level of noise, the time shift and the different shape of the profiles. Finally, we will test a real case.

2. THEORY

In the following discussion, we will use the following conventions: bold uppercase letters to indicate matrices, e.g. **A**; italic uppercase letters to indicate scalars, e.g. *A*; and superscript T to indicate transposition.

The GRAM equations can be found elsewhere [1,11]. As a summary, GRAM only requires two samples, the calibration sample peak ($\mathbf{R}_c$) and the test sample peak ($\mathbf{R}_t$). Each peak is decomposed as a product of three matrices, corresponding to the chromatographic profiles (**X**), the relative concentration (**C**) and the spectra (**Y**). There is also a term for the error (**E**):

$$\mathbf{R}_c = \mathbf{X}\,\mathbf{C}_c\,\mathbf{Y}^T + \mathbf{E}_c\,,\qquad \mathbf{R}_t = \mathbf{X}\,\mathbf{C}_t\,\mathbf{Y}^T + \mathbf{E}_t \qquad (1)$$

This decomposition is called trilinear and assumes that the response at each time-wavelength in $\mathbf{R}_c$ and $\mathbf{R}_t$ is the addition of the individual response of each analyte at this time–wavelength. **X** and **Y** are the same for the decomposition of $\mathbf{R}_c$ and $\mathbf{R}_t$. This implies that the profile and the spectra are the same (same position and same size). The only difference is due to the concentration.

The decomposition is done by solving an eigenvalue problem in which the eigenvectors are related to the profiles (chromatographic and spectral) and the eigenvalues are related to the concentrations.

Fig. 1 shows this decomposition. GRAM is both a calibration method and a curve resolution method, i.e. both quantitative and qualitative information is obtained.

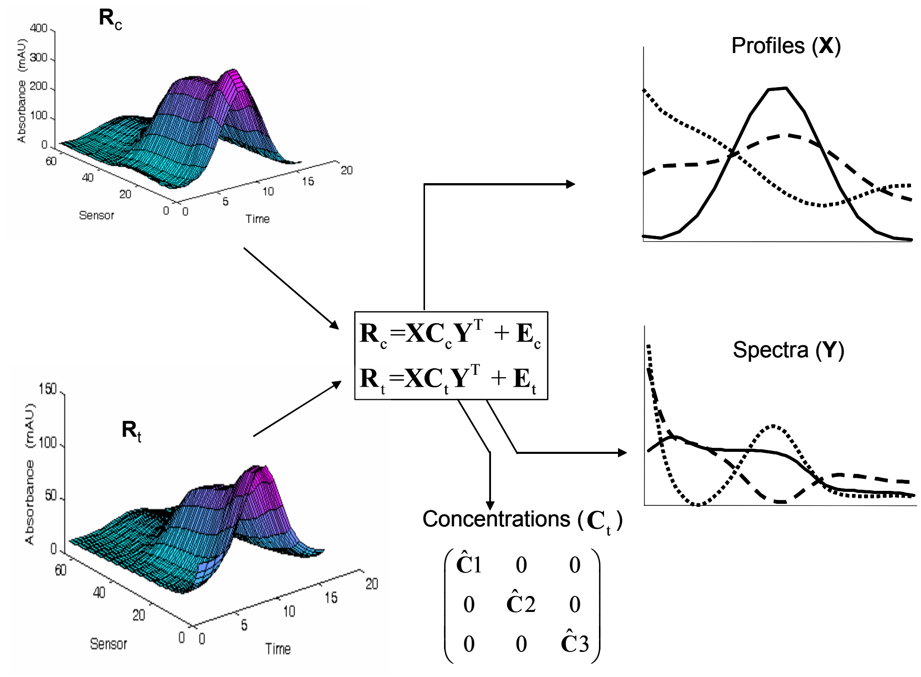


Fig. 1. GRAM decomposition of the calibration sample peak ($\mathbf{R}_c$) and the test sample peak ($\mathbf{R}_t$) into chromatographic profiles (**X**), spectra (**Y**) and relative concentration (**C**). The number of factors, in this case, is 3.

In order to decompose $\mathbf{R}_c$ and $\mathbf{R}_t$ according to Eq. (1), Faber et al. [5] weighted the calibration sample peak as

$$\mathbf{Q} = \mathbf{R}_t + \alpha\mathbf{R}_c \qquad (2)$$

where the eigenvalue problem is solved on **Q**. We can use α to calculate and correct the bias in the prediction and to calculate figure of merit.

The concentration in the test sample peak (c_{t}) is calculated as

$$c_{\mathrm{t,k}} = \frac{\alpha c_{\mathrm{c,k}} \mathbf{\Pi}_{\mathrm{k}}}{1 - \mathbf{\Pi}_{\mathrm{k}}} \qquad (3)$$

where $c_{\mathrm{t,k}}$ is the concentration of the analyte *k* in the calibration sample peak and $\mathbf{\Pi}_{\mathrm{k}}$ is the corresponding eigenvalue.

When the data follow the trilinear model and the right number of factors is selected, the value of α has no influence on $c_{\mathrm{t,k}}$ because it is compensated by the value of $\mathbf{\Pi}_{\mathrm{k}}$. In this case, GRAM predicts correctly [11-12].

Therefore, trilinearity can be checked by studying whether $c_{\mathrm{t,k}}$ changes when α varies.

Several experimental aspects can affect the GRAM decomposition (Eq. (1)). One of these is the time shift between the profiles, i.e. the profile of the analyte of interest does not elute at exactly the same retention time in both peaks. Another is the different shape of the chromatographic profiles in both samples. Other aspects that can affect the decomposition are the noise or the complexity of the overlapped peak.

3. EXPERIMENTAL

3.1. Simulated data

The calibration sample peak and the test sample peak were simulated following Eq. (1). The chromatographic peaks (**X**) were assumed to be Gaussian with 30 data points. The spectra (**Y**) were taken from the study of Zscheile et al. [13] and correspond to the analysis of ribonucleic acids.

For simplicity, in all the experiments it was considered that $\mathbf{R}_c$ only contained a standard of the analyte of interest and $\mathbf{R}_t$ contained the same analyte plus an interferent. The concentration of the analyte of interest in $\mathbf{R}_c$ and $\mathbf{R}_t$ was 1 ppm. Several aspects that can be found in practical chromatographic analysis were studied. Table 1 summarizes the different experiments. Indicated for $\mathbf{R}_c$ and $\mathbf{R}_t$ are the maximum of the chromatographic peak (center) (different maxima indicate time shift); the width (σ) of the profile measured as the standard deviation of the Gaussian peak (a different value involves a different shape); the noise, expressed as percent of the maximum response of the matrix and the number of factors tested for each experiment.

Table 1. Parameters for the simulated data

Experiment	Matrix	Analyte of interest		Interference		% Noise	Factors	Studied Effect
		Center	σ	Center	σ			
1	$\mathbf{R}_c$	15	4	-	-	1	**2**	Trilinearity
	$\mathbf{R}_t$	15	4	10	3			
2	$\mathbf{R}_c$	15	4	-	-	1	**1**	
	$\mathbf{R}_t$	15	4	10	3			
3	$\mathbf{R}_c$	15	4	-	-	**5**	1,2,3	Noise
	$\mathbf{R}_t$	15	4	10	3			
4	$\mathbf{R}_c$	15	4	-	-	**20**	1,2,3	
	$\mathbf{R}_t$	15	4	10	3			
5	$\mathbf{R}_c$	**15**	4	-	-	1	1,2,3	Time shift
	$\mathbf{R}_t$	**14**	4	10	3			
6	$\mathbf{R}_c$	**15**	4	-	-	1	1,2,3	
	$\mathbf{R}_t$	**16**	4	10	3			
7	$\mathbf{R}_c$	15	**4**	-	-	1	1,2,3	Shape
	$\mathbf{R}_t$	15	**5**	10	3			
8	$\mathbf{R}_c$	15	**4**	-	-	1	1,2,3	
	$\mathbf{R}_t$	15	**3**	10	3			

For each experiment the maximum (center) and the standard deviation of the chromatographic profiles in $\mathbf{R}_c$ and $\mathbf{R}_t$, the percent of white noise added to the data, the number of factors tested and the studied effect are indicated. Differences are written in bold.

Fig. 2 shows the cases we studied. We can see the effects of the time shift, the noise and the different shape. For simplicity, only the analyte of interest at one wavelength is depicted. In practice we have a matrix and interferents.

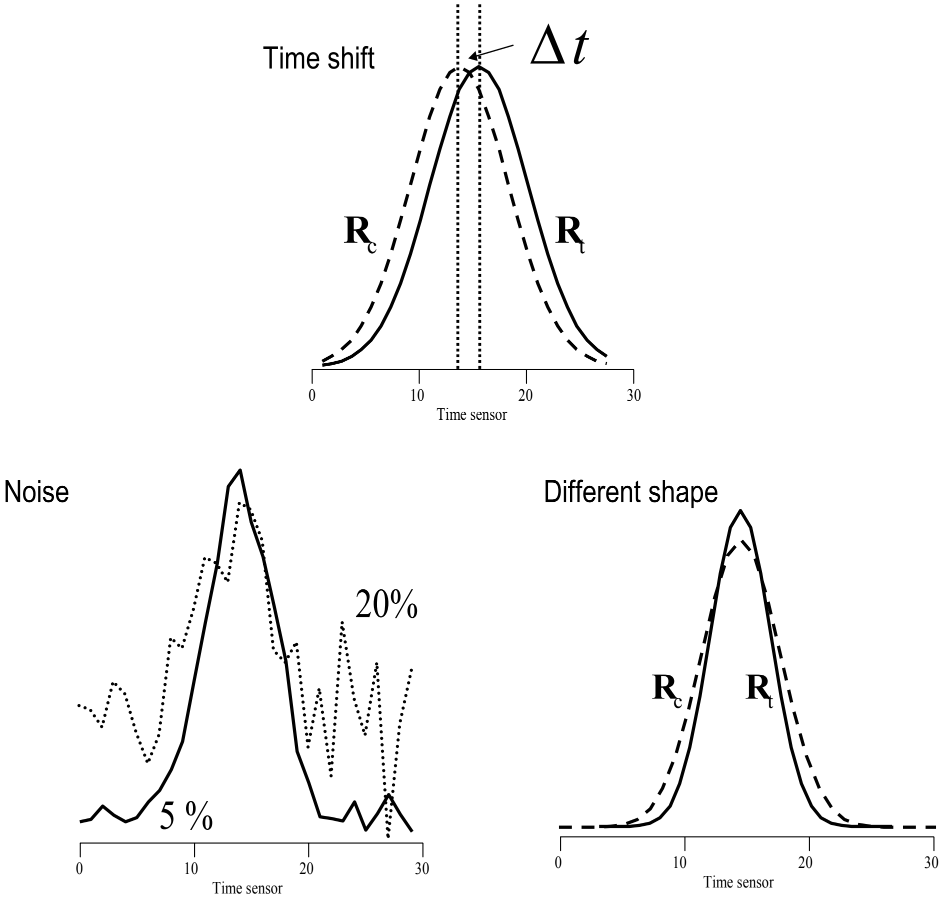


Fig. 2. Different aspects that can affect trilinearity. For simplicity the profiles are only depicted at one wavelength.

3.2. Aromatic sulfonates

Aromatic sulfonates are widely used in the dye and tannery industries. Their toxicology is not yet defined but their high solubility in water makes it very difficult for them to be removed from the wastewater and they are thought to have a contaminant effect.

The experimental chromatographic conditions for determining aromatic sulfonates are described in Ref. [11]. We studied one aromatic sulfonate: the 6-amino-4-hydroxy-2-naftalensulfonate. Fig. 3 shows the chromatographic profile measured at 230 nm and the peak containing the analyte of interest.

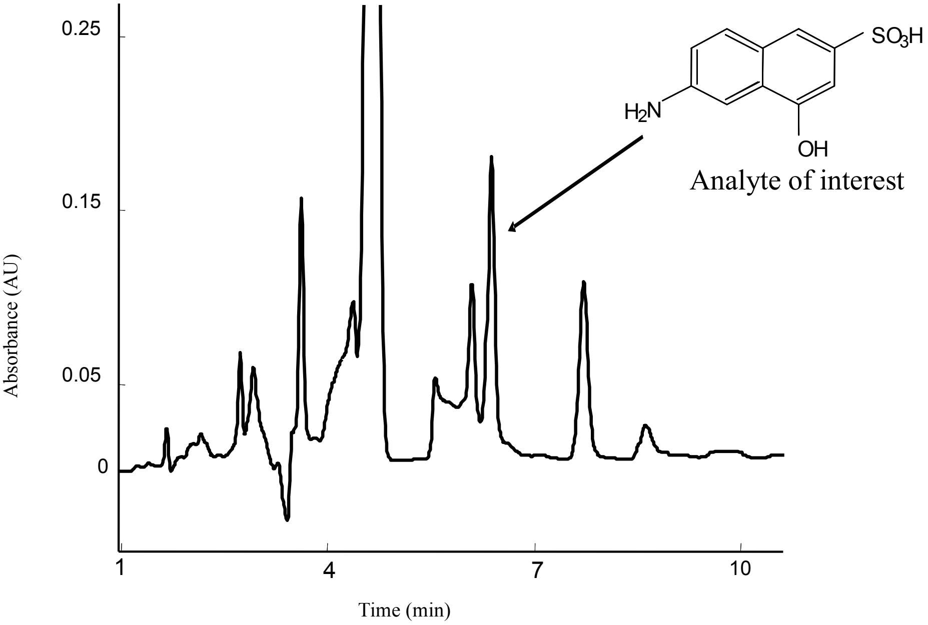


Fig. 3. Chromatographic profile of the water sample containing aromatic sulfonates measured at 230 nm. The peak of the analyte of interest is indicated.

In both the simulated and the real data, the predicted concentration was calculated by changing the value of α for each number of factors. GRAM was tested for different numbers of factors, from 1 to 3 in simulated data, and from 1 to 4 in the real aromatic sulfonates data. As a rule of thumb, the value of α should be around the value of c_t / c_c in order to test both the cases where $\mathbf{R}_t$ has more weight than $\mathbf{R}_c$ and vice versa. Here, in all cases, the value of α was from 0.1 to 3, in 0.1 steps, i.e. we calculated 30 different models for each factor.

3.3. Software

All calculations were done using in-house subroutines for MATLAB [14] version 6.

4. RESULTS AND DISCUSSION

4.1. Simulated data

Table 2 summarizes the mean predicted concentration and its range of variation, expressed as the relative variation (%) with respect to the mean value, for the 30 models calculated by varying α. The variation is the difference between the maximum value and the minimum value.

Table 2. Results for the simulated data

Experiment	Optimal number of factors	c_t	Range of variation (%)	Effect
1	2	0.99	0.02	Trilinear , optimal number of factors
2	1	2.80	45.6	Trilinear, underfitting
3	2	0.99	0.5	5 % noise
4	2	1.04	6.5	20 % noise
5	2	1.09	4.7	Shift -
6	2	0.87	2.8	Shift +
7	2	0.94	2.2	σ +1
8	2	1.08	2.8	σ -1

For each experiment the optimal number of factors, the mean predicted concentration (c_t) and its range of variation along the 30 GRAM models (%) are indicated. The last column indicates which effect was studied at each experiment.

4.1.1. Trilinear data (Experiments 1 and 2)

Here the data were trilinear because the chromatographic profiles in both matrices were proportional, i.e. they eluted at the same retention time and had the same shape. Fig. 4 shows the predicted concentration against the value of α for 1 and 2 factors. When the data were trilinear and the right number of factors was selected, the predicted concentration hardly varied in all the tested models (Experiment 1), i.e. the value of α did not influence the predictions. The horizontal line in the plot for two factors suggests that the data are trilinear and the correct number of factors is two. This gives confidence to the concentration predicted by GRAM.

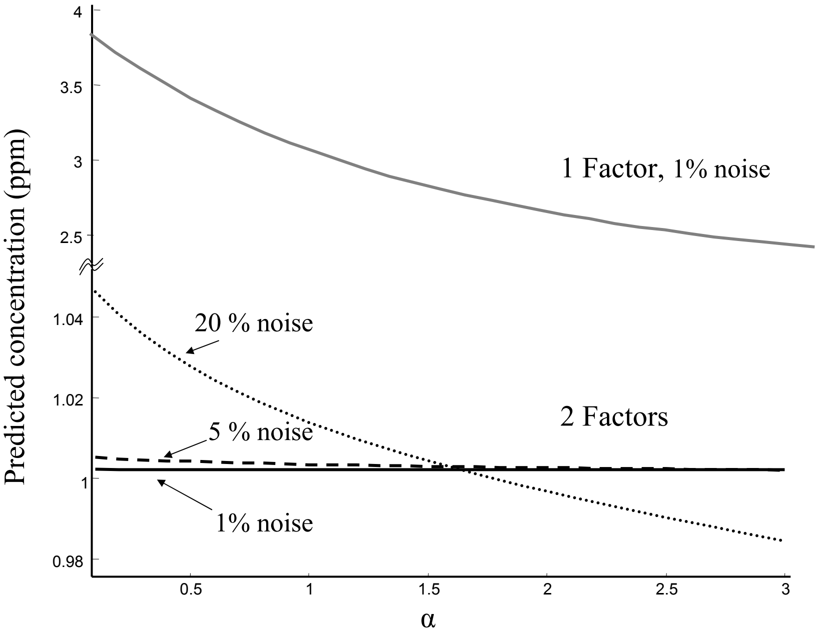


Fig. 4. Predicted concentration by GRAM (Experiments 1–4). Effect of the noise and the wrong selection of the number of factors.

When the number of factors was 1 (underfitting), the predicted concentration depended on the value of α, even though the data followed the trilinear model. Wrongly selecting the number of factors strongly affected the predicted concentrations and led to wrong predictions.

Once we have assessed that we can confidently apply GRAM to this data using two factors, we can select the optimum value of α (α_{opt}). The calculation of α_{opt} has a meaning only when the trilinearity has already been assessed as before. As we can see in Fig. 4, the value of α did not change the predicted concentration. However, it did affect the variance and the bias of the predicted concentration.

The value of α_{opt} is calculated by an iterative process [5]. First a GRAM model is calculated with α_1=1, and a c_t is predicted. A new α_2 is calculated as $\alpha_2 = c_c / c_t$. Another GRAM model is then performed with the new value of α and a new c_t is predicted. This process is done iteratively until convergence.

In all the simulations in this study without the presence of noise, the theoretical value of α was 1. In experiment 1 the calculated α_{opt} was 1.002, which was very close to the real value.

The following sections take into account the different deviations from trilinearity that may be found in chromatographic analyses.

4.1.2. Presence of high noise (Experiments 3 and 4)

The greater the increase in white noise in the data, the greater the change in the predicted concentration values. We carried out tests with several levels of noise up to 20%. Fig. 4 shows the results when we considered two factors (which is the optimal value). Strong dependence on the predicted concentration was observed when the noise level reached 20%. In our experience, 20% of noise is seldom found in this kind of analysis (see Fig. 2) unless the concentration of the analyte is at the limit of detection. The influence of the noise was low because the singular value decomposition step in GRAM situated the noise in the factors that are not used for prediction.

However, it is important to note that this variation (about 6%; see Table 2) was much smaller than the one produced by wrongly selecting the number of factors (about 45%). Notice the scale of Fig. 4.

4.1.3. Time shift (Experiments 5 and 6)

The cases where the peak of the analyte of interest in $\mathbf{R}_t$ is shifted to the left (shift −, in Table 2) and to the right (shift +, in Table 2) are shown. In practice this represents a shift of only 0.8 s (see the experimental section of [15]). Time shift is common in HPLC. It is due to imprecisions in injection timing, fluctuations in temperature and changes in the flow rate. Fig. 5 plots the predicted concentration against α when 1–3 factors were considered. In all cases, dependency was observed. The variations were greatest when 1 and 3 factors (wrong values) were considered. The plot shows that the data were not perfectly trilinear, and that the predictions will have a wider variability if the GRAM model is used.

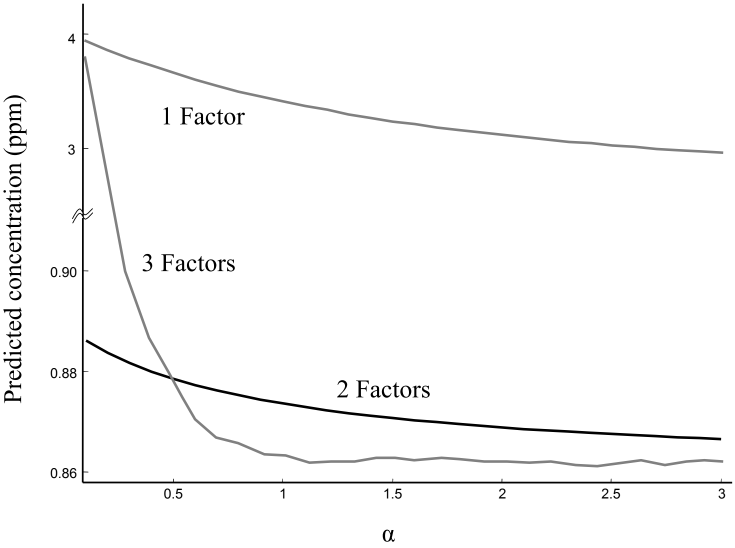


Fig. 5. Predicted concentration by GRAM. Effect of the time shift when 1–3 factors are considered. In all cases, a dependence on α is observed.

Fig. 5 is especially useful because in overlapped peaks the time shift cannot be detected and corrected just by plotting the peak. This is because what is observed is not the individual response, but the sum of the profiles of the different analytes in the peak. It is therefore recommended that existing algorithms be applied to correct the time shift [15-16] before GRAM is applied.

4.1.4. Different shape of the chromatographic profiles (Experiments 7 and 8)

The chromatographic profiles of the analyte of interest in $\mathbf{R}_c$ and $\mathbf{R}_t$ must be proportional. In practice, the shape may be different if the composition of the matrix is different. Table 2 shows that shape has an effect, since the predicted concentration depends on α when the shape varies. The standard addition method is recommended for reducing this effect.

From the results in Table 2 we can conclude that although these effects influence the predictions, the results are often still useful for practical analysis because the variability introduced in the results is often acceptable. This shows that when the data are not completely trilinear, GRAM can still provide useful results. More shifted profiles (for instance by 4 s) or a bigger difference in shape produce misleading results.

Other factors can make it impossible to decompose Eq. (1): for example, when the spectra of the different analytes are very collinear or when the experimental data are rank deficient. This method shows that the data are not trilinear and by correcting the possible shift or doing standard addictions the variations are still observed. In this case, GRAM is not useful for quantifying and other methods that can handle non-linear data, such as MCR-ALS or Tucker3 [17-18], should be used.

4.2. Aromatic sulfonates data

GRAM was first applied to the peaks when the same time window for the calibration sample and the test sample was used. GRAM was tested from 1 to 4 factors. Fig. 6 shows the change in the predicted concentration for different values of α and different numbers of factors. The large change in the predicted concentration with α for all the tested factors suggests that the measured data are not suitable for GRAM. For two factors, which is the one with the least variation, the mean concentration was 0.23 ppm, but the range of variation was 11%.

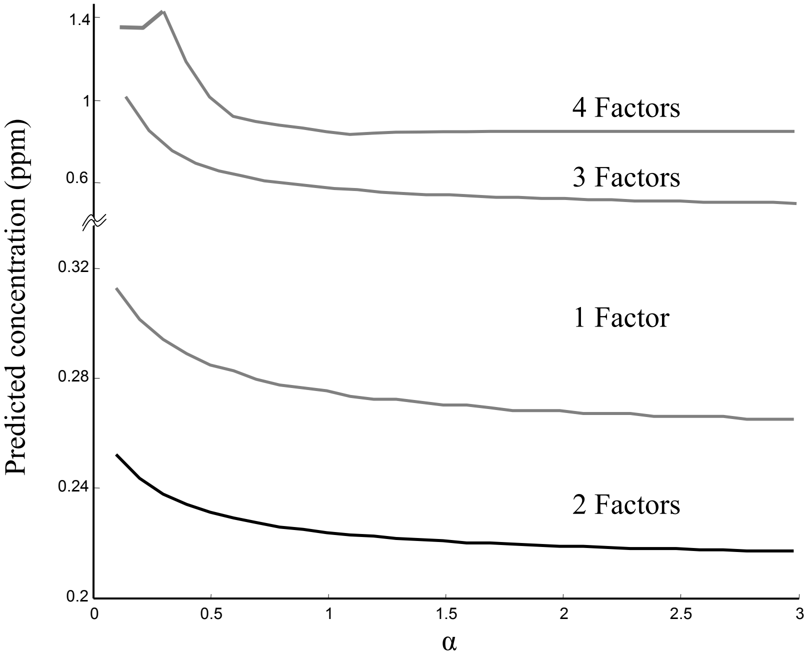


Fig. 6. Predicted concentration by GRAM for the raw aromatic sulfonates data.

Before excluding the use of GRAM, we checked for time shift in the profiles. We used a previously developed algorithm [15] and found this effect. The algorithm selected a time window in $\mathbf{R}_t$ so that the maximum of the chromatographic profile of the analyte of interest coincided with the maximum of the profile in $\mathbf{R}_c$.

After correcting the time shift, the plot of the predicted concentration versus α showed significantly better results than when the data not corrected for the time shift were used (see Fig. 7). When two factors were used, the concentration hardly changed with different values of α. This suggests that the data follow the trilinear model and the results obtained from GRAM are reliable. The mean concentration was 0.23 ppm and the range of variation was 1%.

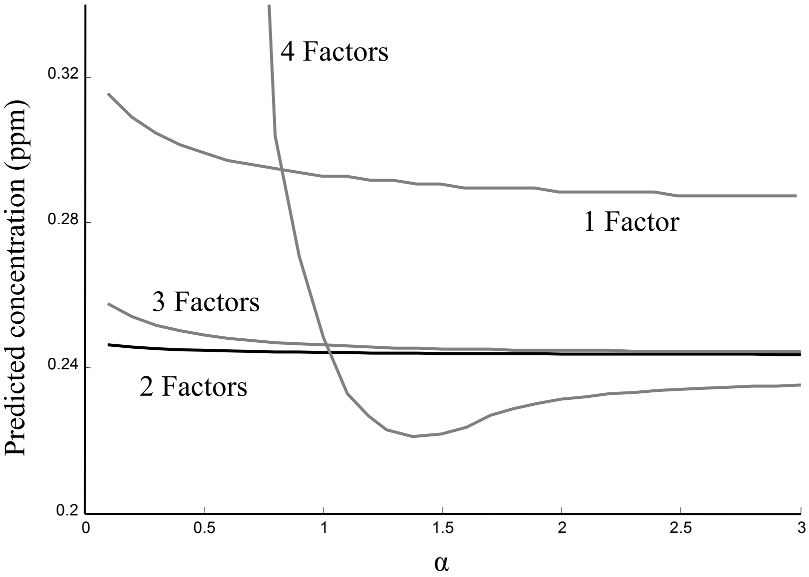


Fig. 7. Predicted concentration by GRAM for the aromatic sulfonates data after correcting the time shift of the chromatographic profile of the analyte of interest in the test sample peak.

If the number of factors was wrongly selected (1 or 3), the variation was around 50%.

As this data set had been used in a previous study, we knew that the real concentration of the analyte was 0.24 ± 0.1 ppm. This was determined by univariate calibration after the complete resolution of the peak (tedious experimental work). As we can see in Fig. 7, the value predicted by GRAM fully agreed with the one obtained by univariate calibration. Here the value of α_{opt} was 1.64.

5. CONCLUSIONS

GRAM is a second-order calibration method that can extract information from overlapped peaks in HPLC–DAD data. However, the data must satisfy certain mathematical requirements (trilinearity) and the number of factors to build the model must be selected properly.

In chromatography, the time shift and the different shape of the peaks are common factors that cause lack of trilinearity. A graphical representation of the predicted concentration versus the value of α can detect whether the data are trilinear and what the right number of factors is. These calculations indicate the quality of the final results. We have shown that when the data slightly deviate from trilinearity, useful results are still obtained and that wrongly selecting the number of factors has the greatest effect.

Constructing the graph requires calculating several GRAM models by changing the value of α and the number of factors, but this can be done very quickly.

ACKNOWLEDGEMENTS

The authors would like to thank R.A. Gimeno, R.M. Marcé and F. Borrull, from our own department, for recording and providing the aromatic sulfonates data. We also would like to thank the Spanish Ministry of Education, Culture and Sports (Project BQU 2000-1256) for economic support and the AGAUR, of the Catalan Government, for providing Enric Comas with a doctoral fellowship.

REFERENCES

[1] E. Sánchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496.

[2] N.M. Faber, L.M.C. Buydens, G. Kateman, J. Chemom. 8 (1994) 147.

[3] L.S. Ramos, E. Sánchez, B.R. Kowalski, J. Chromatogr A. 385 (1987) 165.

[4] N.M. Faber, R. Boqué, J. Ferré, Chemom. Intell. Lab. Syst. 55 (2001) 91.

[5] N.M. Faber, J. Ferré, R. Boqué, Chemom. Intell. Lab. Syst. 55 (2001) 67.

[6] D.J. Louwerse, A.K. Smilde, H.A.L. Kiers, J. Chemom. 13 (1999) 491.

[7] Z. Chen, Y. Liang, J. Jiang, Y. Li, H. Qian, R. Yu, J. Chemom. 13 (1999) 15.

[8] Z. Chen, Z. Liu, Y. Cao, R. Yu, Anal. Chim. Acta 444 (2001) 295.

[9] H. Xie, J. Jiang, N. Long, G. Shen, H. Wu, R. Yu, Chemom. Intell. Lab. Syst. 66 (2003) 101.

[10] B.K Dable, K.S. Booksh, J. Chemom. 15 (2001) 591.

[11] E. Comas, R.A. Gimeno, J.Ferré, R.M Marcé, F. Borrull, F.X. Rius, J.Chromatogr A. 988 (2003) 277.

[12] S. Li, P.J. Gemperline, K. Briley, S, Kazmierczak, J. Chromatogr B 665 (1994) 213.

[13] F.P. Zscheile, H.C. Murray, G.A. Baker, R.G. Peddicord, Anal. Chem. 34 (1962) 1776.

[14] Matlab, The Mathworks, South Natick, MA, USA.

[15] E. Comas, R.A. Gimeno, J.Ferré, R.M Marcé, F. Borrull, F.X. Rius, Anal. Chim. Acta. 470 (2002) 163.

[16] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218.

[17] A. de Juan, R. Tauler, J. Chemom. 15 (2001) 749.

[18] N. M. Faber, R. Bro, P. K. Hopke, Chemom. Intell. Lab. Syst. 65 (2003) 119.

3.4 OUTLIER DETECTION IN GRAM

As in other analytical methodologies, the use of GRAM in the analysis of future samples also requires validation tests. Such validation may involve analyzing validation samples of a known analyte concentration, with a matrix similar to the test samples that will be found in the future [1]. However, as opposed to zero- and first- order calibration, in which the validated model is used for all the future samples (except periodical updates), a GRAM model is calculated for each new test sample. This makes the validation of the methodology more complex: validating one GRAM model for one test sample does not imply that the model for the next test sample will also be valid. The main reason is the lack of trilinearity (due to, for example, retention time shift or variation of the peak shape) of the test sample peak with respect to the peak in the calibration standard. Biased predictions may also be caused by the wrong selection of the number of factors needed to build the GRAM model (which may vary from one test sample to another because the number of interferents in the overlapped peak may vary). Hence, as happens in zero- and first- order calibration, outlier detection tools are also needed in GRAM to prevent us from reporting largely biased predictions.

Since the GRAM solution is guaranteed to be acceptably correct when the trilinearity requirement is fulfilled [2], the basic outlier detection tool must be directed to checking whether such trilinearity exists. Here we present an outlier detection criterion based on the Net Analyte Signal (NAS) calculated from the GRAM estimations. The NAS is calculated both for the calibration sample and for the test sample. Both NAS's are proportional if both samples were modeled correctly. Otherwise, there are reasons to suspect lack of trilinearity and the fact that the predictions may be incorrect. This outlier detection tool is developed in the paper *Outlier detection in the Generalized Rank Annihilation Method applied to chromatographic data.* E. Comas, J. Ferré, F.X. Rius. Analytical Chemistry, submitted. The detection of outliers through this criterion is based on the visual inspection of the correlation of the NAS of $\mathbf{R}_t$ and the net sensitivity (NAS of $\mathbf{R}_c$ at unit concentration) in the net analyte signal regression plot (NASRP) [3]. A measurement of the noise is needed to determine when the lack of fit observed in the NASRP is acceptable or not. Such measurement of noise may be obtained from either flat region of the chromatogram or by using the bilinear structure of the peaks of the calibration and test samples. This last method compares the estimated noise along the time axis with the estimated noise along the wavelength axis. Since the lack of trilinearity is mainly caused by time shift, the estimated noise along the wavelength axis can be considered as a pure estimation of the noise, whereas the estimation of noise along the time axis may contain systematic variations when retention time shift exists. The comparison of both estimated "noises" may be used to detect that the test sample is an outlier. This criterion is explained in the paper *Estimation of the net noise in a second-order chromatographic peak* (in preparation), included at the end of this section.

References

[1] R. Boqué, A. Maroto, J. Riu, F.X. Rius, Grasas y Aceites 53 (2002) 128-143.

[2] N.M. Faber, Anal. Bioanal. Chem 372 (2002) 683-687.

[3] J. Ferré, F.X. Rius, Anal. Chem. 70 (1998) 1999-2007.

3.4.1 Paper

Outlier detection in the Generalized Rank Annihilation Method applied to chromatographic data.

E. Comas, J. Ferré, F.X. Rius.

Analytical Chemistry, submitted.

Outlier detection in the generalized rank annihilation method applied to chromatographic data

Enric Comas, Joan Ferré, F. Xavier Rius

Department of Analytical and Organic Chemistry. Rovira i Virgili University
Pl. Imperial Tarraco 1, 43005, Tarragona, Spain

ABSTRACT

The Generalized Rank Annihilation Method (GRAM) can be used in HPLC-DAD to determine the concentration of the analyte of interest when it elutes overlapped with unknown interferences. Retention time shift and peak broadening between measurements may cause the peak of the test sample to behave as an outlier, thus producing an incorrect GRAM prediction. Here we present a method based on the second-order net analyte signal (NAS) to assess the quality of the GRAM predictions and detect such outliers. The chromatographic and spectral profiles predicted by GRAM are used to define the space spanned by the interferences and their orthogonal counterparts. The projections of the calibration and test sample peaks onto this space are proportional if the trilinear model, assumed by GRAM, is followed. Proportionality is checked by the regression of both unfolded projections. The slope of the fitted straight line is equal to the GRAM prediction. The size and distribution of the residuals indicate the degree of fit of the data to the assumed trilinear model. Systematic trends in the residuals indicate a lack of trilinearity and predictions with a large error. Simulated data were used to test this method with respect to retention time shift and peak broadening. Analytical data from the determination of two water pollutants were studied with the outlier detection method. In one case, a retention time shift problem was detected. After correction, the prediction error was reduced from 25% to 2%. In the other, the data were acceptably trilinear.

Key-words: GRAM, HPLC-DAD, outlier, second-order NAS, trilinearity.

INTRODUCTION

In High Performance Liquid Chromatography with Diode Array Detection (HPLC-DAD), the separation is often optimized for mixtures of pure standards. Univariate calibration is then used to quantify the analyte of interest in test samples based on the area or height of the peak. When samples with complex matrices are analyzed (such as environmental and clinical samples), the analyte of interest often elutes overlapped with unexpected compounds. The experimental protocol must then be modified to achieve the selectivity required in univariante calibration. This may involve e.g. cleaning-up, adding specific compounds that react either with the analyte of interest or with the interferences, modifying the chromatographic parameters or changing the detection channel or detector [1].

The Generalized Rank Annihilation Method (GRAM) [2,3] is a good alternative to such additional experimental work. This calibration method can predict the concentration of the analyte of interest in an overlapped peak even when the background signal varies from sample to sample. Qualitative information, i.e. the elution profiles and the spectra, is also obtained. To obtain these advantages, the spectrum of the effluent must be measured during the separation. In this way, a matrix of absorbances (time × wavelength) for each peak is obtained. This is known as second-order data [4]. Such data are recorded almost by default today, since separations are often monitored by measuring the entire spectrum. This enables the analyte to be identified both from its retention time and its spectrum. When overlap is detected, GRAM can be applied without additional work because the necessary data have already been recorded. Moreover, GRAM only requires one calibration sample.

GRAM has been applied to techniques such as NMR [5], fluorescence [6] and UV-vis [7] spectroscopies and chromatography [8,9]. It was recently applied in HPLC-DAD analyses to determine polycyclic aromatic hydrocarbons (PAHs) [10,11] and aromatic sulfonates [12] in water as well as pesticides and phenolic compounds over a highly drifted baseline [13].

Though GRAM may significantly reduce analysis time and costs, it is hardly used in routine HPLC analyses. Partly this is for practical reasons such as the lack of adequate commercial software and trained analysts. These may be circumvented as the interest of chromatographers in this method grows. Other reasons are technical. GRAM predictions are easily affected by the irreproducibility of the separations. The retention time shift of the elution profiles and the change in their shape (e.g. peak broadening) increase prediction error. For example, a prediction error of as much as 30% was obtained in the analysis of PAHs when the peak in the test sample shifted three seconds in relation to the peak in the standard [10]. Analysts will therefore rarely use GRAM in daily analyses unless they are confident of the analyte concentration it predicts. This confidence is gained by checking the procedure with validation samples and having adequate outlier detection diagnostics to warn against biased predictions.

External validation of the method must be done by analyzing reference samples that are representative of the test samples [14]. Recovery essays in which test samples are spiked are also possible. This type of validation is not sufficient, however, to provide complete confidence in the predicted concentration. One reason for this is that the GRAM model is calculated for each new test sample, so successful previous models do not guarantee correct quantification for the peaks of the test samples at hand. Outlier detection therefore plays a primary role in the application of GRAM to HPLC-DAD analyses.

A test sample is an outlier if it has extreme values or if it does not accommodate to the calibration model. A test sample may not follow a calculated model for several reasons. In univariate and multivariate calibration, one reason for this is the presence of unknown interferences that contribute to the instrumental response but were not considered when the model was calculated. Such a sample is not an outlier in GRAM because, as the model is built especially for that test sample, the signals of the interferences are 'modeled'. Quantification is possible as long as the selectivity is sufficient in both orders (i.e. the elution profile and the spectrum of the analyte of interest are different enough from the profiles and spectra of the interferences). The main reason why a test sample is an outlier in GRAM is that it deviates from trilinearity. Trilinearity involves: (i) that the measured peak can be bilinearly decomposed as a sum of contributions of the different analytes and (ii) that, except for a scaling factor related to the concentration, the elution profile and spectrum of the analyte of interest are the same in both the standard and test samples. When trilinearity is fulfilled, the GRAM predictions are accurate [15]. Retention time shift and peak shape variation can break down trilinearity and cause incorrect predictions [16]. The sample may also be an outlier if the ratios of the concentrations of two analytes in the calibration and test samples are the same [17]. This last requirement is hardly found in real samples or can be avoided by using standard additions.

The simplest outlier detection method in GRAM is to verify that the estimated elution profiles/spectra are as expected. The elution profiles should be unimodal and non-negative. The spectra should be those obtained by measuring pure standards of the analytes. The degree of coincidence can be checked with the correlation coefficient [18] or, equivalently, with the dissimilarity value [19]. However, these comparisons are not sufficient (see below). Apparently, correct elution profiles and spectra can be obtained even when the data are not trilinear and the prediction errors are large [10]. More advanced outlier detection tools check whether the peaks follow the trilinear model. A first measure of lack of trilinearity is given by the difference between the measured peak and the predicted peak. Large differences mean bad model fit either because the number of factors in GRAM is wrong or because the data lack trilinearity. However, these differences evaluate the peak globally and do not relate directly to the specific analytes we wish to quantify. We may obtain non-random residuals but accurate predictions. A related tool is to project one peak onto the space spanned by the rows and columns of the other peak [2]. The projection should recover the projected peak within the noise. However, this tool is limited when the two peaks contain different interferences. Although the sum peak may be used to span the calibration space, small deviations from trilinearity are still difficult to detect. A third tool is to compare the chemical rank of the augmented matrices by joining the calibration and test sample matrices both column-wise and row-wise [20]. Their rank is the same if the data are trilinear. However, evaluating a significant increase in rank is difficult because small non-linearities are distributed through the relevant eigenvectors/eigenvalues and a few others. Recently, a visual criterion was proposed to assess the trilinearity of HPLC-DAD data and find the correct number of factors to calculate a GRAM model [16]. However, this criterion is only partially related to the quality of the predictions and more advanced tests are still needed.

Here we report a new graphical criterion for detecting outliers in GRAM for HPLC-DAD data. It can be used to internally assess the quality of the predictions. The criterion is inspired from an outlier detection method developed for the classical least-squares (CLS) model [21]. CLS is the extension of the univariante Lambert-Beer's Law to multivariate calibration. If the test sample follows the calculated CLS model, the part of its spectrum that is orthogonal to the spectra of the modeled interferences is proportional to the vector of regression coefficients (or net sensitivity vector). This characterizes the regression model. Biased predictions due to unmodeled interferences can be detected because this proportionality does not exist. The same principle is applied here to GRAM for HPLC-DAD data, thus extending this test to second-order calibration. The spectra and elution profiles estimated by GRAM are used to define the space spanned by the interferences. Projecting the peak of the standard and the test sample onto this space produces two matrices that are proportional if both samples follow the trilinear model. This is checked by examining the linear fit of one signal with respect to the other after the matrices are unfolded. For trilinear data, the residuals of the fit are within the noise range. Otherwise, (e.g. in retention time shifted data), systematic trends are observed in the residuals and the reliability of the predictions can be questioned.

Simulated data were used to study the ability of the criterion to detect lack of trilinearity due to time shift and peak broadening. Real data from the analysis of an aromatic sulfonate and a phenolic compound in water samples were used to demonstrate the utility of this tool.

THEORY

1. Notation

Boldface uppercase letters represent matrices, boldface lowercase letters indicate column vectors and italic letters indicate scalars. Transposition of a matrix or vector is symbolized by a superscripted 'T'. Vectorisation of a matrix (i.e., stacking its columns from left to right) is indicated by 'vec'. For a given matrix **A**, the matrices $\mathbf{A}^{-1}$ and $\mathbf{A}^{+}$ stand for its inverse and Moore-Penrose pseudoinverse, respectively. A 'hat', e.g., $\hat{\mathbf{A}}$, was added to the reconstructed calibration and test data to differentiate them from the measured ones. The analyte of interest is designated as 'analyte *k*'. **I** is the identity matrix of the appropriate size.

2. The GRAM algorithm

GRAM uses the peak of the analyte of interest in a calibration sample ($\mathbf{R}_c$), whose known concentration is c_c, and the peak of interest in the test sample ($\mathbf{R}_t$). This peak also contains contributions from one or more interferences. Both $\mathbf{R}_c$ and $\mathbf{R}_t$ have size ($J_1 \times J_2$) where the J_1 columns correspond to the retention times and the J_2 rows correspond to the wavelengths. $\mathbf{R}_c$ can be obtained by measuring either a pure standard [2,10] or a spiked sample [11].

The algorithm can be found in reference 2 but is briefly given here for the sake of completeness:

1) Calculate the sum matrix **R** ($J_1 \times J_2$)

$$\mathbf{R} = \mathbf{R}_c + \mathbf{R}_t \qquad (1)$$

2) Calculate the singular value decomposition (SVD) of **R** [22]:

$$\mathbf{R} = \mathbf{USV}^{T} + \mathbf{E} \qquad (2)$$

where the matrices of singular vectors (**U**,**V**) and singular values (**S**) have been truncated for *F* factors [9]. Several tools have been developed to determine the number of factors [22,23]. Here, the F-test [22] was used. The residual matrix **E** contains the non-modeled contributions.

3) Solve the eigenvalue problem:

$$(\mathbf{S}^{-1}\mathbf{U}^{T}\mathbf{R}_{t}\mathbf{V})^{T}\mathbf{T}=\mathbf{T\Phi} \qquad (3)$$

The diagonal matrix of eigenvalues (**Φ**) has the relative concentration for each analyte (*f*) in the calibration (c_c) and test ($\hat{c}_t$) samples:

$$\mathbf{\Phi}_f = \frac{\hat{c}_{t,f}}{\hat{c}_{t,f} + c_{c,f}} \qquad (4)$$

4) Calculate the elution profiles **H** ($J_1 \times F$) and the spectral profiles **Y** ($J_2 \times F$):

$$\mathbf{H}=\mathbf{UST} \qquad (5)$$

$$\mathbf{Y}=\mathbf{V}(\mathbf{T}^{-1})^{T} \qquad (6)$$

Usually, the spectral profiles are normalized and the scaling constant is introduced in **H**. **H** and **Y** include the profiles of all the components in $\mathbf{R}_c$ and $\mathbf{R}_t$, so **R** can be predicted as $\hat{\mathbf{R}} = \mathbf{HY}^{T}$.

5) Find the columns of **H** and **Y** that correspond to analyte *k*. This can be done by visually comparing **Y** with the spectrum of analyte *k* measured from a pure standard. A dissimilarity value [19] can be used. A value under 0.0141 means that the correlation between the two spectra is above 0.9999 [18]. The corresponding diagonal element in the eigenvalues matrix $\mathbf{\Phi}_k$ can be used to calculate the predicted concentration as

$$\hat{c}_t = \frac{c_c \mathbf{\Phi}_k}{1-\mathbf{\Phi}_k} \qquad (7)$$

Eq 3 can also be solved by considering $\mathbf{R}_c$ instead of $\mathbf{R}_t$, i.e. $(\mathbf{S}^{-1}\mathbf{U}^{T}\mathbf{R}_{c}\mathbf{V})^{T}\mathbf{T} = \mathbf{T\Pi}$. In this case the matrix of eigenvalues, for each analyte *f* is:

$$\mathbf{\Pi}_f = \frac{c_{c,f}}{\hat{c}_{t,f} + c_{c,f}} \quad (8)$$

The predicted calibration and the test matrices are

$$\hat{\mathbf{R}}_c = \mathbf{H\Pi Y}^T \quad (9)$$

$$\hat{\mathbf{R}}_t = \mathbf{H\Phi Y}^T \quad (10)$$

where $\mathbf{\Pi} + \mathbf{\Phi} = \mathbf{I}$

3. Net Analyte Signal (NAS)

The net analyte signal (NAS) [24] is the part of the measured response that is used for prediction. The NAS of analyte *k* is calculated for both the calibration sample and the test sample as:

$$\hat{\mathbf{R}}_c^* = \mathbf{P}_H \hat{\mathbf{R}}_c \mathbf{P}_Y \quad (11)$$

$$\hat{\mathbf{R}}_t^* = \mathbf{P}_H \hat{\mathbf{R}}_t \mathbf{P}_Y \quad (12)$$

where

$$\mathbf{P}_H = \mathbf{I} - \mathbf{H}_{-k}\mathbf{H}_{-k}^+ \quad (13)$$

$$\mathbf{P}_Y = \mathbf{I} - \mathbf{Y}_{-k}\mathbf{Y}_{-k}^+ \quad (14)$$

are projection matrices, $\mathbf{H}_{-k}$ and $\mathbf{Y}_{-k}$ where $\mathbf{H}$ and $\mathbf{Y}$ are without the column of analyte *k*. $\mathbf{P}_H$ and $\mathbf{P}_Y$ project a peak onto the space that is orthogonal to the spectra and elution profiles of the interferences. The double projection involves the NAS taking into account the net contribution in each order. The rank of the NAS matrices, which only have the contribution of the analyte of interest, is 1. Combining equations 9 and 10 with 11 and 12 we find that

$$\hat{\mathbf{R}}_t^* = \frac{\hat{c}_t}{c_c} \hat{\mathbf{R}}_c^* \quad (15)$$

The proportionality constant is the ratio between the predicted concentration $\hat{c}_t$ and the concentration in the calibration standard. Hence, the net sensitivity matrix is found as the NAS of the calibration sample divided by its concentration:

$$\mathbf{S}^* = \frac{\hat{\mathbf{R}}_c^{\ *}}{c_c} \tag{16}$$

4. Trilinearity assessment and outlier detection

When both $\mathbf{R}_c$ and $\mathbf{R}_t$ are trilinear, their rows and columns are spanned by the calculated spectra and elution profiles up to the noise level.

In this case, the projection of the measured peak $\mathbf{R}_t$ onto the space that is orthogonal to that spanned by the interferences

$$\mathbf{R}_t^* = \mathbf{P}_H \mathbf{R}_t \mathbf{P}_Y = \mathbf{P}_H\ (\hat{\mathbf{R}}_t + \mathbf{E}_t)\ \mathbf{P}_Y = \hat{\mathbf{R}}_t^{\ *} + \mathbf{E}_t^* \tag{17}$$

will be approximately equal to the NAS (eq. 12), since the values of $\mathbf{E}_t^*$ will be random and small ($\mathbf{E}_t^*$ contains the part of $\mathbf{R}_t$ not explained by $\hat{\mathbf{R}}_t$). When $\mathbf{R}_c$ and $\mathbf{R}_t$ deviate from trilinearity, these deviations are embedded in **H** and **Y,** which no longer span the rows and columns of $\mathbf{R}_c$ and $\mathbf{R}_t$. Hence, the projected test sample peak $\mathbf{R}_t^*$ contains the part that is used for prediction ($\hat{\mathbf{R}}_t^*$) plus contributions from the lack of trilinearity, and $\mathbf{E}_t^*$ becomes large and systematic.

A visual measure of the relative appearance of $\mathbf{E}_t^*$ is obtained by regressing vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ with a non-intercept straight line model (net analyte signal regression plot). The slope of the fitted line is the concentration predicted by GRAM, $\hat{c}_t$ (see the supporting information) since vec$\mathbf{E}_t^*$ is orthogonal to vec$\mathbf{S}^*$ (Figure 1). The quality of the fit indicates the trilinearity of the data and can be used to detect outliers. For non-trilinear data (e.g. time shifted peaks), systematic patterns are observed in the residuals of the regression. For trilinear data, the residuals will be approximately random.

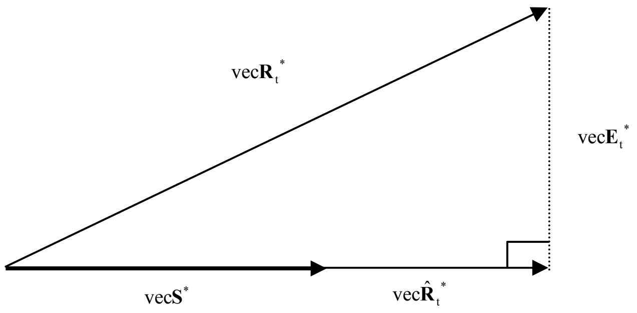

$\mathrm{vec}\mathbf{R}_t^*$
$\mathrm{vec}\mathbf{E}_t^*$
$\mathrm{vec}\mathbf{S}^*$
$\mathrm{vec}\hat{\mathbf{R}}_t^*$

Fig 1. Relationship between vec$\mathbf{S}^*$ and vec$\mathbf{R}_t^*$. $\mathbf{R}_t^*$ is decomposed as $\hat{\mathbf{R}}_t^* + \mathbf{E}_t^*$.

The significance of the residuals of regressing vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ can be measured by comparing them with an estimation of the net noise. The data will be trilinear when these residuals are comparable to the net noise. The noise (**N**) is estimated from a time window without analytes. This can either be a region as close as possible to the peak of the analyte of interest or the same time window of the analyte in a blank sample. The net noise ($\mathbf{N}^*$) is then calculated as

$$\mathbf{N}^* = \mathbf{P}_H\mathbf{N}\mathbf{P}_Y \tag{18}$$

Since the analyte of interest is not present, the values in $\mathbf{N}^*$ are non-systematic and the variations are related to the amount of noise.

5. Reduction of prediction error by correcting the retention time shift

A probable reason why $\mathbf{R}_t$ is an outlier is the retention time shift with respect to $\mathbf{R}_c$. Retention time shift correction methods [10,25] have been developed to align the elution profile of the analyte of interest in both the standard and test samples. The method of Comas et al. [10] will be used here. Briefly, both $\mathbf{R}_c$ and $\mathbf{R}_t$ are decomposed independently by Iterative Target Transformation Factor Analysis (ITTFA). The analyte of interest in both matrices is identified by spectral comparison. The time window of $\mathbf{R}_t$ is shifted so that the decomposed profile of the analyte of interest is aligned with the elution profile of this analyte in $\mathbf{R}_c$.

EXPERIMENTAL SECTION

Simulated data

Trilinear data. Second-order chromatographic peaks (**R**) were simulated for the elution profiles (**H**) and spectra (**Y**), $\mathbf{R} = \mathbf{H}\mathbf{Y}^T$. The peak of the analyte in a pure standard ($\mathbf{R}_c$) was simulated as a Gaussian peak with 75 time points, standard deviation $\sigma_1 = 4$ and maximum centered at the time step 25. As the chromatograph in our laboratory records a spectrum every 0.4 seconds, 75 time points correspond to 30 seconds. Its spectrum was that of Adenine from reference 26. The peak was weighted to simulate a concentration of 1 μg l^{-1}. The peak in the test sample ($\mathbf{R}_t$) was simulated to contain the same analyte as $\mathbf{R}_c$ but at 0.6 μg l^{-1} (same shape and position as in $\mathbf{R}_c$) plus a highly overlapped interference (resolution less than 0.2): a Gaussian peak with 75 time points and $\sigma_2 = 4$ but centered at the time step 35. The spectrum was that of Guanine from reference 26. The interference was simulated at 5 μg l^{-1}, so the analyte of interest is a minor component in $\mathbf{R}_t$. White noise (1% in relation to the maximum of the profile) was added to $\mathbf{R}_c$ and $\mathbf{R}_t$.

Non-trilinear data. Non-linearities are not usually found in the spectral mode in HPLC-DAD data, so we focused on the retention time variations. Retention time shift was simulated by modifying the position of the elution profiles in $\mathbf{R}_t$. The time window for $\mathbf{R}_t$ was shifted from 1 to 10 time steps to both shorter and longer elution times. Time shift of 10 units corresponds to 4 seconds. Shifts of 1 or 2 seconds in daily routine measurements may be common. The variation in the shape of the profiles (σ) was also considered. For the analyte of interest in $\mathbf{R}_t$, σ_2 was varied from 3.2 to 4.8, at 0.2 intervals. σ_1 was not changed. The value of σ_2 / σ_1 changed form 0.8 to 1.2.

Measured data

Two water pollutants were analyzed: the aromatic sulfonate 1-amino-6-naftalensulfonate and the phenolic compound resorcinol.

Aromatic sulfonates data. Aromatic sulfonates are used in the dye and tannery industries. Their toxicology is not yet defined, but their high solubility in water makes it very difficult to remove them from wastewater and they are thought to have a contaminating effect. The sample was collected from the output of the sewage treatment plant in Tarragona (Spain). The chromatographic conditions were optimized with pure standards to simultaneously determine six aromatic sulfonates [12]. Of these, 1-amino-6-naftalensulfonate (retention time = 7.8 minutes) eluted overlapped with unknown interferences when the test sample was analyzed, and was quantified by GRAM. To obtain a reference value for external validation, the chromatographic conditions were varied in order to fully isolate the analyte of interest and use univariate calibration. The concentration found was 0.089 μg l^{-1}. The analysis lasted 45 minutes. This second optimization process was tedious and did not guarantee that the analyte of interest would be isolated from other interferences in future test samples.

Resorcinol data. Resorcinol is potentially hazardous both for the environment and for human health. It is regulated by the European Union (EU) to ensure good quality bathing and drinking water. Resorcinol was analyzed in water samples from the Ebre River (Spain) as described in reference [13]. Due to its low concentration, a preconcentration step by solid phase extraction (SPE) was carried out. The SPE process also retained humic and fulvic acids, which caused a large peak at the beginning of the chromatogram and baseline drift where the analyte of interest elutes. To obtain the reference value, sodium sulfite (Na_2SO_3) was added to react with the humic and fulvic acids and make them elute separately from the analyte of interest. The concentration found with univariate calibration was 5.6 ± 1.4 μg l^{-1}. However, as sodium sulfite does not always remove the baseline drift, we tested GRAM in this situation. Unlike with aromatic sulfonates, $\mathbf{R}_c$ was not obtained from a pure standard but from a standard addition of the analyte of interest to the test sample.

In both data sets, chromatographic separation was carried out using an HP1100 system (Agilent Technologies, Waldbronn, Germany). This system consisted of a degasser, two isocratic pumps, a manual injector provided with a 20 μl loop, a column oven and a DAD. Each pump was used to deliver one fraction of the mobile phase. Separation was carried out using a 25 × 0.46 cm Kromasil 100 C_{18} chromatographic column with a 5 µm particle size (Teknokroma, Barcelona, Spain). The spectrum of the effluent was recorded between 220 and 300 nm every 0.4 nm.

Software

We made the GRAM and the second-order NAS methods subroutines in house for MATLAB version 6 [27].

RESULTS AND DISCUSSION

Simulated data

Figure 2 shows the variation in the prediction error of the GRAM estimations at different retention time shifts. Each curve corresponds to a different shape of the elution profile of $\mathbf{R}_t$ (peak broadening). The estimated concentration was highly influenced by the retention time shift. For example, a time shift of -4 seconds, led to a prediction error of 60%. The effect of shape variation in the prediction error was not so large: the prediction error was less than 10% in all cases, irrespective of the retention time shift. In our experience, such shape variations are not found in practical analysis.

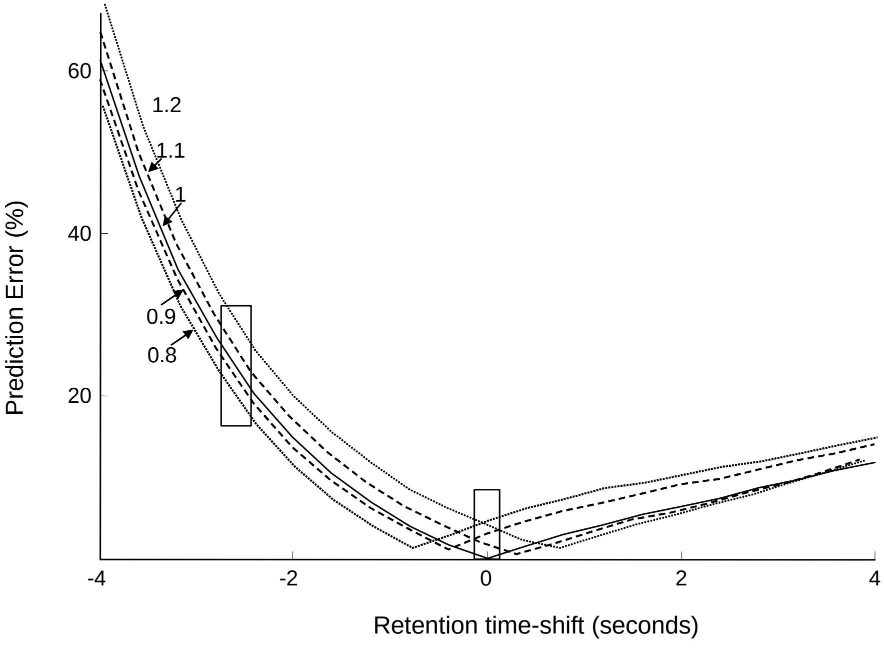


Fig 2. Prediction error (%) from the GRAM estimations against the retention time shift. Each curve corresponds to a different σ_2/σ_1 ratio, from 0.8 to 1.2. The two cases are indicated at retention time shifts of 0 and -2.4 seconds.

The curves in Figure 2 show that the effects of positive retention time shifts on prediction error are different from those of negative retention time shifts. This is because the prediction error caused by the lack of trilinearity is affected by the overlap between the analyte and the interferences. In these simulations, the interference is eluted after the analyte of interest. A shift to earlier elution times (the negative shift in Figure 2) reduces the overlap between the analyte of interest in $\mathbf{R}_c$ and $\mathbf{R}_t$ and also increases the overlap of the interference in $\mathbf{R}_t$ with the analyte of interest in $\mathbf{R}_c$ (see Figure 3). This increases the prediction error. On the other hand, when $\mathbf{R}_t$ elutes later (the positive shifts in Figure 2), the interference overlaps less with the analyte in $\mathbf{R}_c$, and prediction errors are lower.

As an illustration, this method of assessing trilinearity and detecting outliers was applied to two cases: the trilinear case and, as an example of non-trilinear data retention, data time shifted by -2.4 seconds. Both situations are shown in Figure 2. All the other situations in Figure 2 are analogous to these cases.

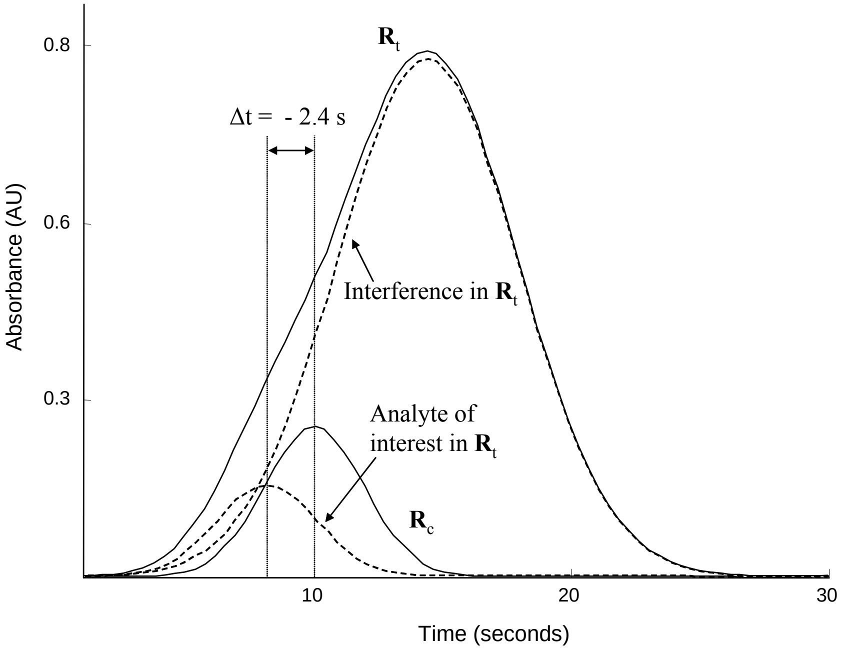


Fig 3. Elution profiles of $\mathbf{R}_c$ and $\mathbf{R}_t$ at 220 nm in -2.4 seconds retention time shifted data. Dashed lines indicate the underlying profiles in $\mathbf{R}_t$.

Trilinear data.

In this case, the elution profiles regarding the analyte of interest are non-shifted and have the same shape in $\mathbf{R}_c$ and $\mathbf{R}_t$. Figure 4 shows the regression of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$. The fit is satisfactory and the slope is the concentration estimated by GRAM. Residuals are distributed randomly, which indicates that the values in $\mathbf{E}_t^*$ are small and non-significant compared to those in $\hat{\mathbf{R}}_t^*$. The correlation coefficient was 0.9998.

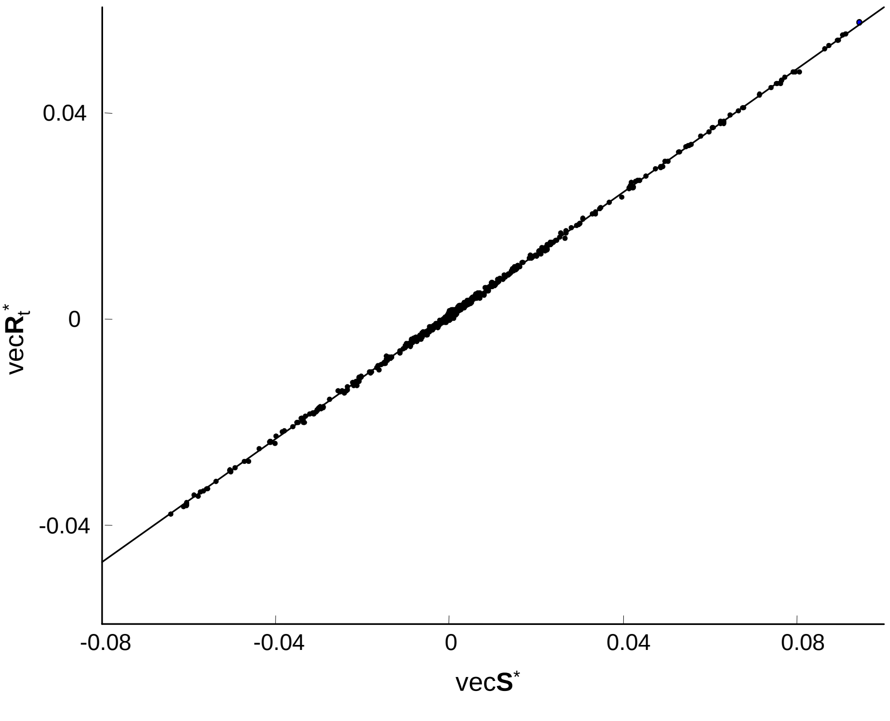


Fig 4. Net analyte signal regression plot for trilinear data.

.

Non-trilinear data.

Figure 3 shows the elution profiles of $\mathbf{R}_c$ and $\mathbf{R}_t$ at 220 nm for a retention time shift of -2.4 seconds. The dashed lines are the underlying elution profiles in $\mathbf{R}_t$. Since the main contribution in $\mathbf{R}_t$ is the interference, the true retention time shift of the analyte of interest cannot be detected by comparing $\mathbf{R}_c$ and $\mathbf{R}_t$. GRAM analysis yields a similar estimated spectrum of Adenine to the one used to simulate the peak, with a dissimilarity value of less than 0.0141 (Figure 5). Therefore, the lack of trilinearity along the time dimension did not affect the result over the wavelength dimension and, despite the retention time shift, the qualitative analysis is correct. The good agreement between the two spectra may lead us to incorrectly accept the quantitative result, which, due to the retention time shift, has a large prediction error of 22%. This shows that spectral comparison is a poor method of outlier detection for GRAM. Also, no evidence of a lack of trilinearity was observed in the elution profiles estimated by GRAM since they were unimodal and nonnegative.

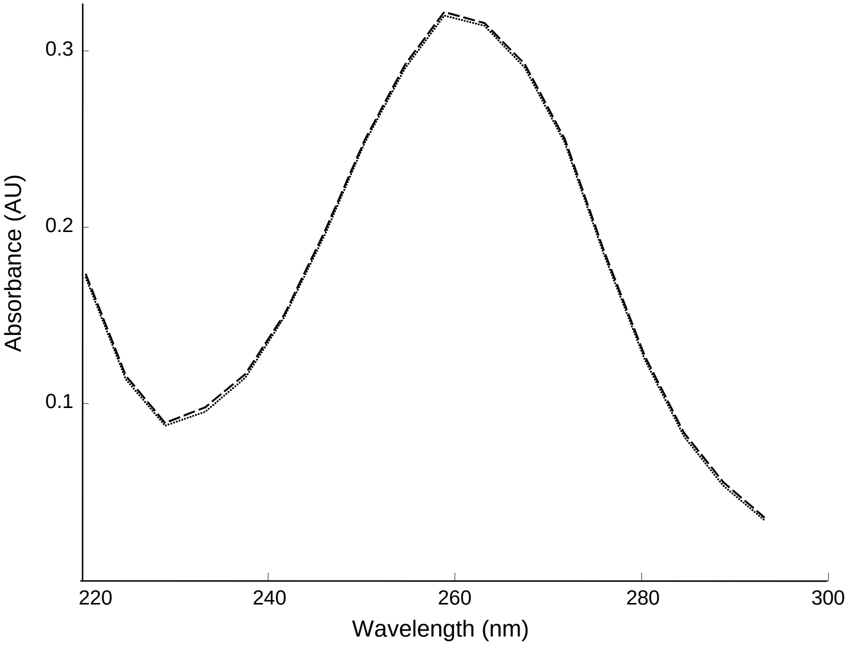


Fig 5. Estimated Adenine spectrum by GRAM (**--**) in -2.4 seconds retention time shifted data and the one used for simulation (··).

Non-trilinear data can be easily detected by considering the NAS. Figure 6 shows $\mathbf{S}^*$ and $\mathbf{R}_t^*$. The peak from 1 to 15 seconds corresponds to the analyte of interest. The flat region from 20 to 30 seconds corresponds to the zone where only the interference eluted (see Figure 3). The signal from the interference was removed in the projection step (eqs. 11 and 17) and only the one signal related to the analyte of interest used for quantification remains. The negative parts on the surface are due to the orthogonality of the projection. Figure 7 shows vec$\mathbf{S}^*$ and vec$\mathbf{R}_t^*$. If $\mathbf{R}_c$ and $\mathbf{R}_t$ were trilinear, $\mathbf{S}^*$ and $\mathbf{R}_t^*$ would be proportional and the two plotted lines would be identical except for a scaling constant (the predicted concentration). In our case, due to the significant contribution of $\mathbf{E}_t^*$ in $\mathbf{R}_t^*$, the lack of trilinearity translated into a shift between the two curves. The proportionality of $\mathbf{S}^*$ and $\mathbf{R}_t^*$ can be better checked by plotting one curve against the other (see Figure 8). The trend in the residuals indicate lack of trilinearity i.e. the values of $\mathbf{E}_t^*$ are significant. Hence, $\mathbf{R}_t$ is a second-order outlier and the predictions are erroneous. The circular pattern observed in the residuals is due to the elements of $\mathbf{S}^*$ where the signal increases when confronted to the elements of $\mathbf{R}_t^*$ where, due to the retention time shift, the signal decreases. The many points around (0,0) correspond to the part of the peak in which the analyte of interest is not present and the net signal is therefore zero. Fewer points have extreme values which correspond to the largest net signal.

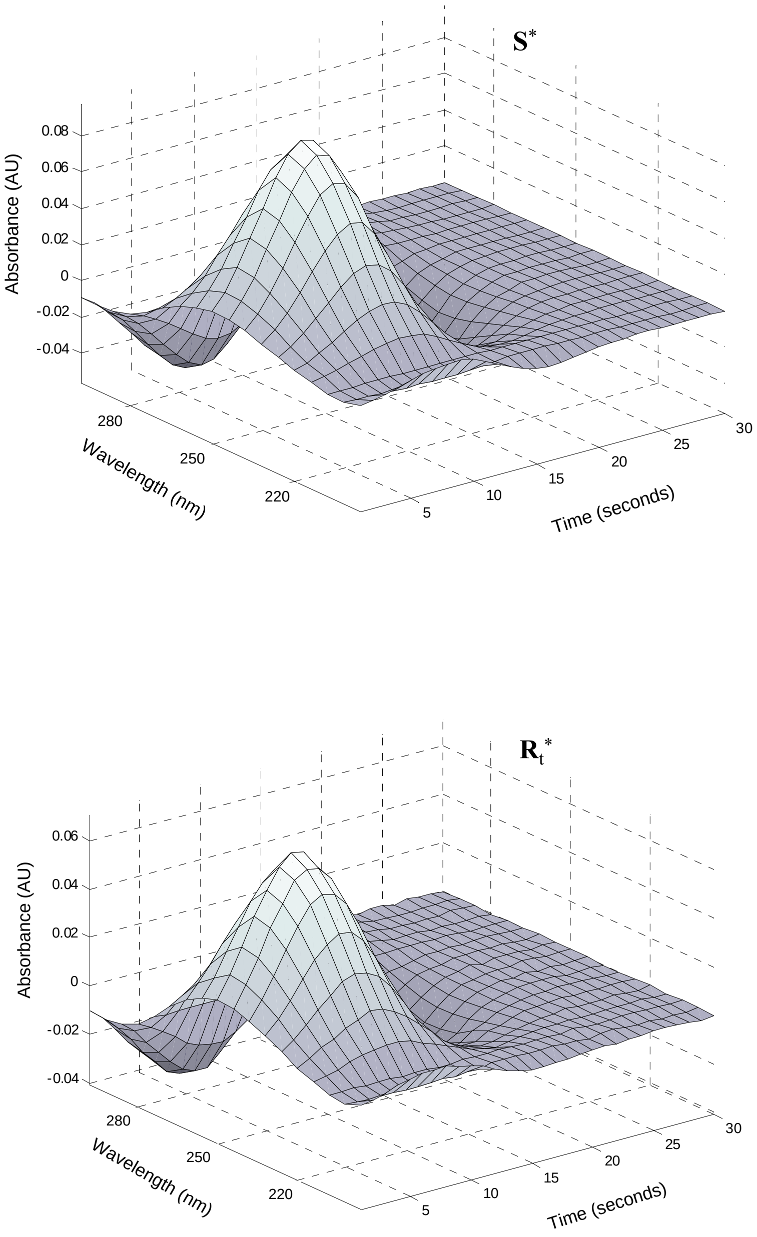


Fig 6. **S*** (6a) and **R**$_t^*$ (6b) in -2.4 seconds retention time shifted data.

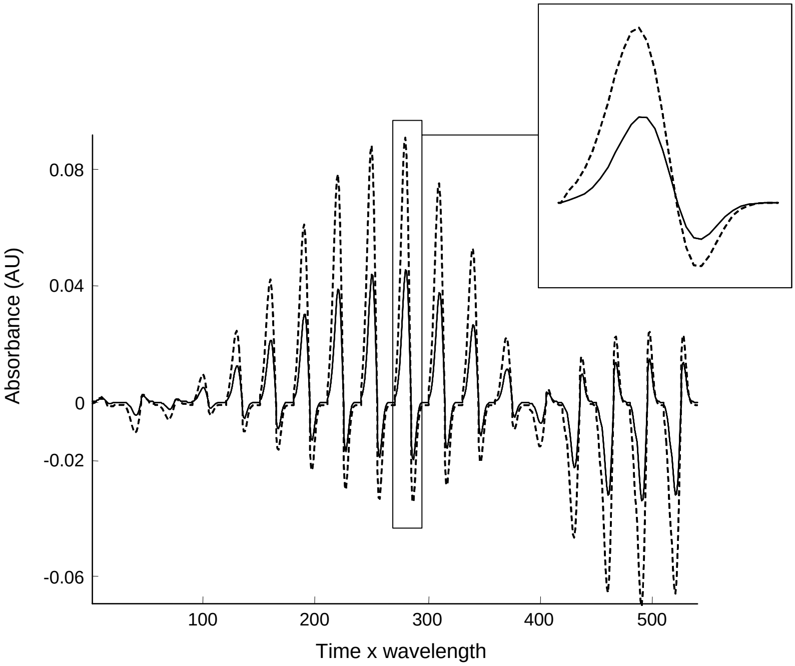


Fig 7. vec**S*** (--) and vec**R**$_t^*$ (—) .

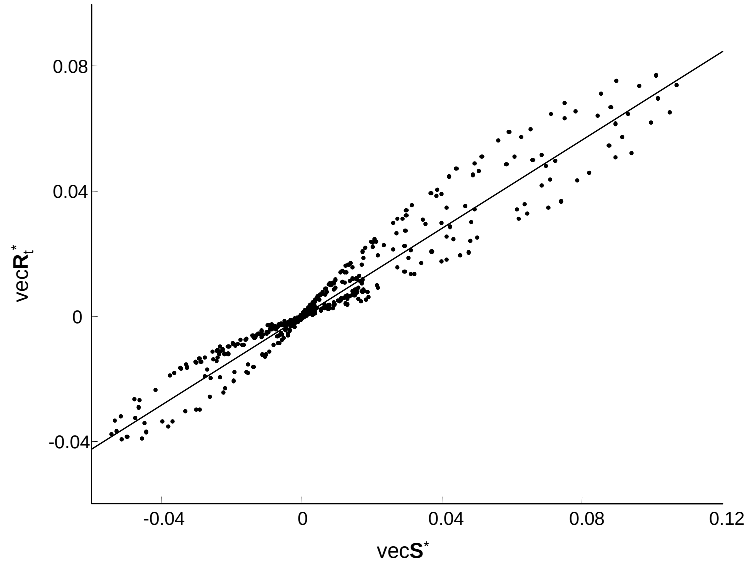


Fig 8. Net analyte signal regression plot in -2.4 seconds retention time shifted data.

In our simulations, the prediction error was above 5 % when the correlation between vec$\mathbf{R}_t^*$ and vec$\mathbf{S}^*$ was less than 0.999. Although this cannot be extrapolated as a strict rule because the error of the predicted concentration is affected by the degree of overlap of the interferences and the amount of noise this correlation provides a rough indication of the sensitivity of the regression tool for detecting large prediction errors. In all simulations, the dissimilarity value for the spectra was calculated. The estimated spectra were always similar to the simulated one, which led to the false conclusion that the predictions were correct.

Measured data

Aromatic sulfonate data. Figure 9 shows the chromatogram measured at 230 nm. The elution window of the analyte of interest was visually selected from 7.1 to 7.6 min for both $\mathbf{R}_c$ and $\mathbf{R}_t$. The net noise was estimated from the flat area between 6.4 and 6.9 min from $\mathbf{R}_t$. Here, **N** does not contain any signal that is not modeled in the peak of interest. The trilinearity test was first applied to the raw data (Figure 10a) for a GRAM model with three factors. The correlation coefficient was 0.823. The large and systematic residuals suggest that the concentration predicted by GRAM (0.065 μg l^{-1}) was inaccurate. The value found with univariate calibration in new optimized conditions was 0.089 μg l^{-1}. The prediction error was therefore 25%. Figure 10b shows the regression plot after the retention time shift was corrected. Two factors were sufficient to build the GRAM model. The residuals no longer show a systematic pattern. To check the significance of these residuals, their values were compared to an estimation of the net noise. Figure 11 plots the residuals of vec$\mathbf{R}_t^*$ against those of vec$\mathbf{S}^*$ and those corresponding to vec$\mathbf{N}^*$ before (11a) and after (11b) the time shift is corrected. To simplify the figure, only one of every two data points are plotted. We can see that in the first case the residuals are much larger than the net noise but in the second case this difference is not so great. Therefore, $\mathbf{R}_t$ is an outlier in the first case but not in the second case, which suggests that the predicted value (0.086 μg l^{-1}) after the retention time shift has been corrected, is reliable. This value is close to the value of 0.089 μg l^{-1} found with univariate calibration.

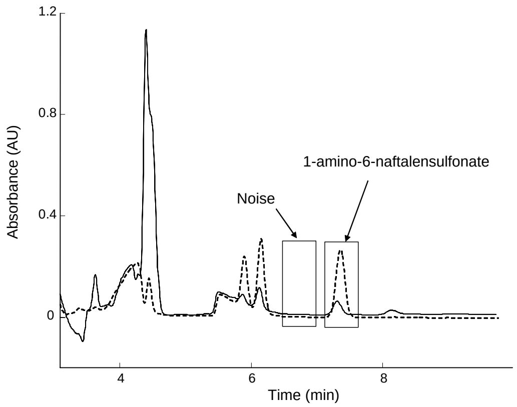


Fig 9. Chromatograms measured at 230 nm of the sewage water sample (–) and of the mixture of standards (--).

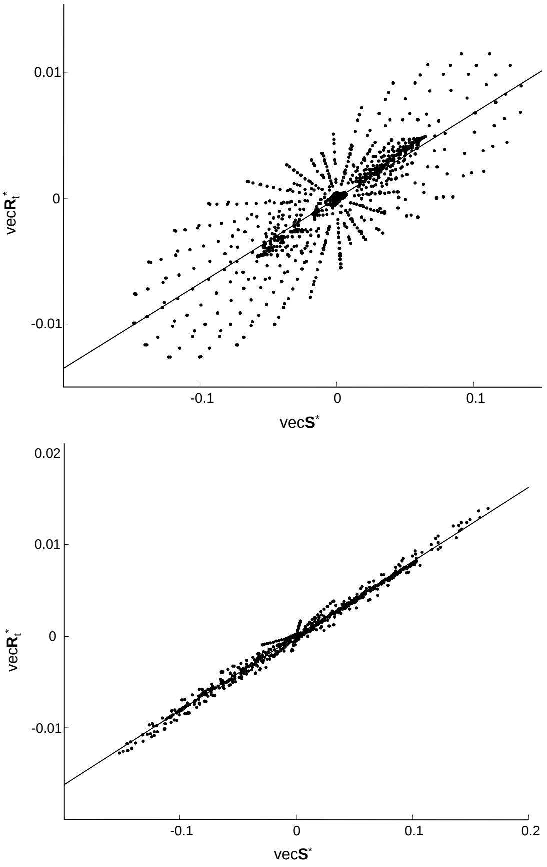


Fig 10. Aromatic sulfonates data. Net analyte signal regression plot in raw data (10a) and after the retention time shift is corrected (10b).

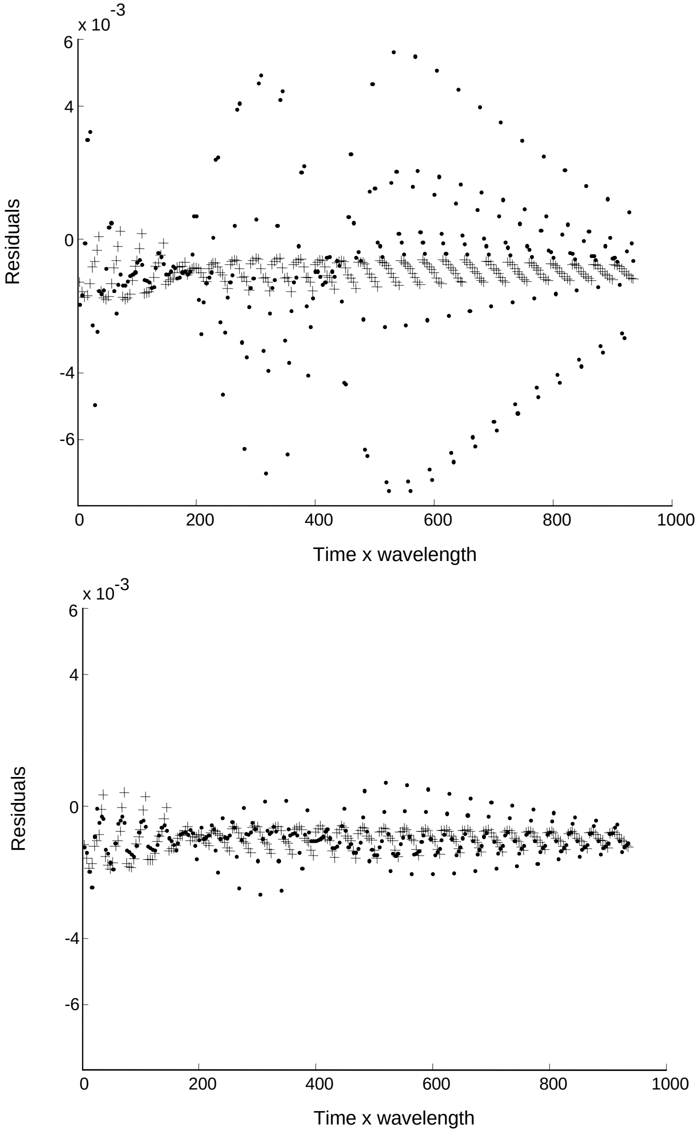


Fig 11. Aromatic sulfonates data. Residuals of the regression of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ (·), and the noise (+) of the raw data (11a) and after the retention time shift is corrected (11b).

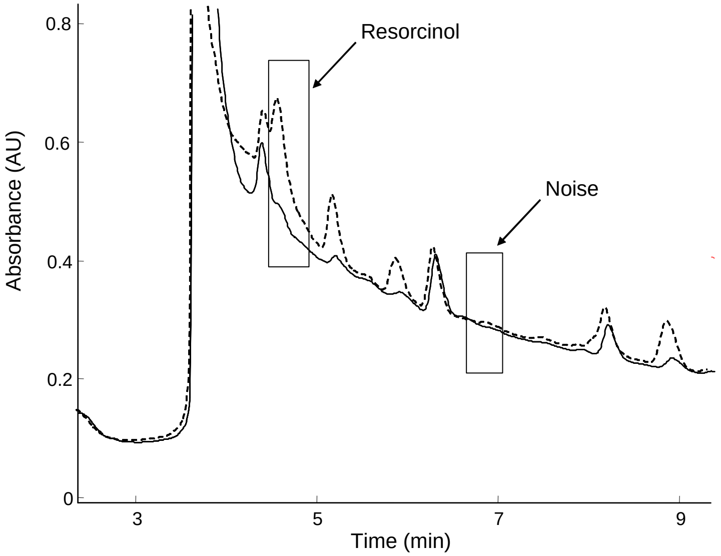
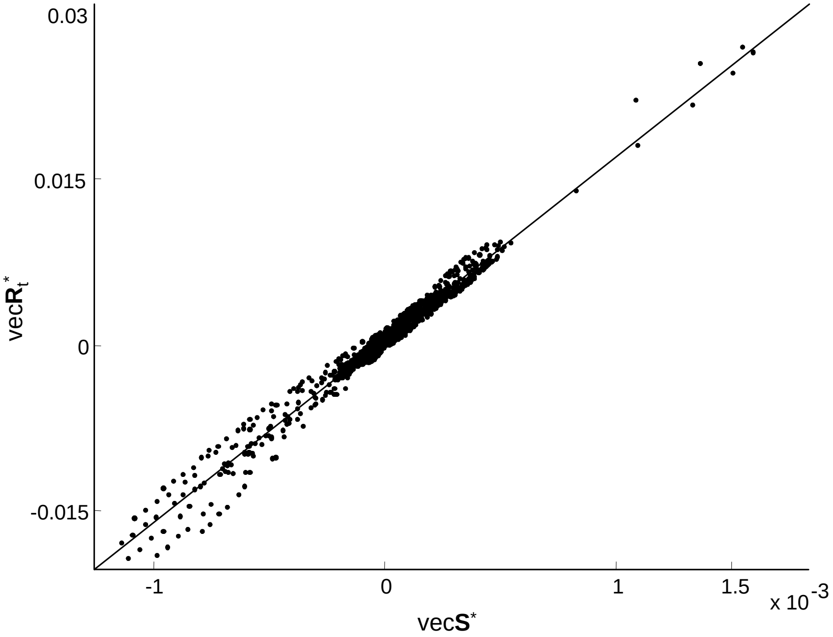


Fig 12. Chromatograms at 250 nm of the river water sample (–) and the standard addition sample (--).

Resorcinol data. Figure 12 shows the chromatogram at 250 nm of the river water sample and the standard addition sample. Three factors were needed to build the GRAM models. The outlier tool applied to the raw measured data and retention time shift corrected data are shown in Figure 13. The peak only shifted one time step (0.4 seconds). After time shift correction, the residuals were slightly smaller, which suggests that the trilinear behavior of the calibration and test sample improved. The net contribution of the noise was estimated using the signal from 6.8 to 7.1 min from $\mathbf{R}_t$. A difficulty may arise when the baseline varies from the position of the peak of interest to the position of the blank zone. A different baseline would involve a contribution that was not considered in GRAM, so the net analyte signal in the elution profile ($\mathbf{P}_H\mathbf{N}$) direction may have a systematic contribution. However, in the area in which the peak elutes and where the blank was considered, the background spectrum is probably the same and its net signal $\mathbf{NP}_Y \approx 0$. Therefore, $\mathbf{N}^* = \mathbf{P}_H\mathbf{NP}_Y$ will remove the signal even if the baseline in the bank zone is different.



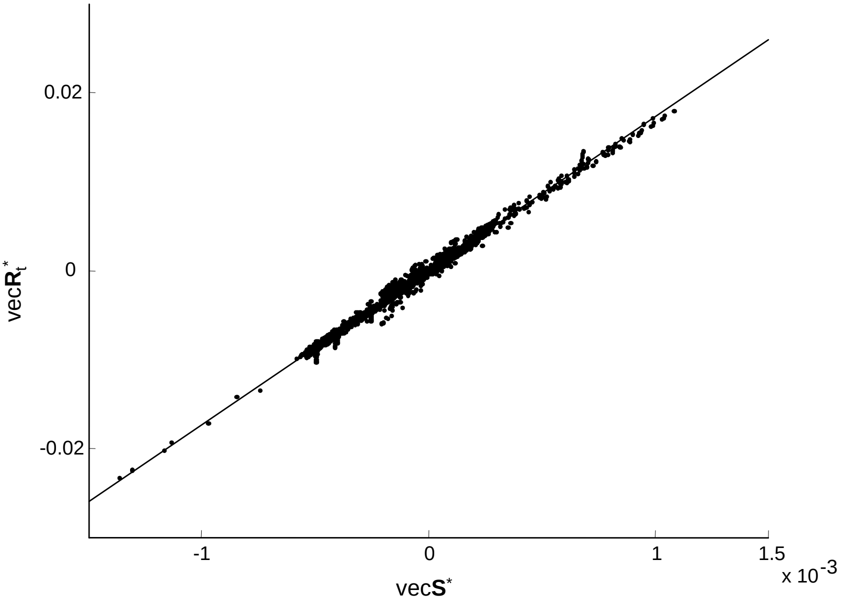


Fig 13. Resorcinol data. Net analyte signal regression plot in raw data (13a) and after the retention time shift is corrected (13b).

Figure 14 compares the net noise with the residuals of the proposed criterion. We can see that after the retention time shift was corrected, the differences between the residuals of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ and vec$\mathbf{N}^*$ were slightly smaller than before the retention time shift was corrected, which confirms the improvement in trilinearity. The error in the GRAM predictions was checked against the reference value (5.6 µg l^{-1}), which was obtained using univariante calibration in new modified separation conditions. In the raw data, GRAM predicted 4.8 µg l^{-1} and in the time shift corrected data, it predicted 4.9 µg l^{-1}. The difference is not very large because the retention time shift was only one time step. 4.9 µg l^{-1} is acceptable and similar to 5.6 µg l^{-1} for this type of analysis in which online preconcentration is a major source of variability in the data.

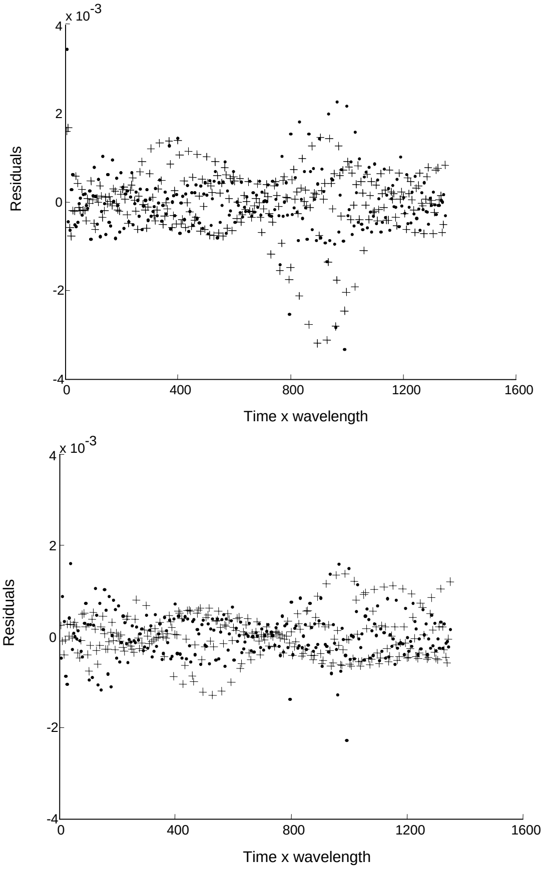


Fig 14. Resorcinol data. Residuals of the regression of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ (·), and the noise (+) of the raw data (14a) and after the retention time shift is corrected (14b).

CONCLUSIONS

In the analysis of HPLC-DAD data, retention time shift and peak broadening can make a test sample behave as an outlier in GRAM. Although the qualitative analysis may still be possible, the quantitative analysis may be seriously affected. We have demonstrated a visual tool for detecting such outliers that can be used to internally assess the quality of the predictions. The criterion is based on checking the proportionality between the net sensitivity matrix and the projected test sample peak in the net analyte signal regression plot. When the data are not trilinear, the residuals are large and show a systematic pattern. When the data are trilinear, the residuals are small and random. Since this method is based on the net contribution of the analyte of interest, it can be useful either when the calibration peak is pure or when it is a mixture of analytes obtained from standard additions.

ACKNOWLEDGMENTS

We would like to thank R.A. Gimeno, R.M. Marcé and F. Borrull, from our own department, for providing the measured data. We also would like to thank the Spanish Ministry of Science and Technology (Project BQU 2003-01142) for economic support and the AGAUR, of the Catalan Government, for providing E. Comas with a doctoral fellowship.

REFERENCES

[1] N. Masqué, M. Galià, R.M. Marcé, F. Borrull, J.Chromatogr A. 803 (1998) 147-155.

[2] E. Sánchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496-499.

[3] J. Ferré, N.M Faber, E. Comas, F.X. Rius, *Generalized Rank Annihilation Method, A tutorial.* In preparation.

[4] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) A782-A791.

[5] A. Nordon, P.J. Gemperline, C.A. McGill, D. Littlejohn, Anal. Chem. 73 (2001), 4286-4294.

[6] A.G. Frenich, D.P. Zamora, M.M. Galera, J.M.L. Vidal, Anal. Bioanal. Chem. 375 (2003) 974-980.

[7] Z. Lin, K.S. Booksh, L.W. Burgess, B.R. Kowalski, Anal. Chem. 66 (1994) 2552-2560.

[8] S. Li, P.J. Gemperline, K. Briley, S. Kazmierczak, J. Chromatogr B. 655 (1994) 213-233.

[9] A.E. Sinha, B.J. Prazen, R.E. Synovec, Anal. Bioanal. Chem. 378 (2004) 1948-1951.

[10] E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius, Anal. Chim. Acta. 470 (2002) 163-173.

[11] R.A. Gimeno, E. Comas, R.M. Marcé, J. Ferré, F.X. Rius, F. Borrull, Anal. Chim. Acta. 498 (2003) 47-53.

[12] E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius, J.Chromatogr A. 988 (2003) 277-284.

[13] E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius, J.Chromatogr A. 1035 (2004) 195-202.

[14] EURACHEM 1998. *The fitness for purpose of Analytical methods. A laboratory guide to method validation and related topics.* EURACHEM secretariat, Teddington, Middlesex.

[15] N.M. Faber, Anal. Bioanal. Chem. 372 (2002) 683-687.

[16] E. Comas, J. Ferré, F.X. Rius, Anal. Chim. Acta. 515 (2004) 23-30.

[17] E. Sánchez, L.S. Ramos, B.R. Kowalski, J. Chromatogr. 385 (1997) 151-164.

[18] J. Ferré, F.X. Rius, Quim. Anal. 15 (1995) 259-252.

[19] R. Gargallo, R. Tauler, F. Cuesta-Sanchez, D.L. Massart, Trends Anal. Chem. 15 (1996) 279-289.

[20] R. Tauler, A.K. Smilde, B.R. Kowalski, J. Chemom. 9 (1995) 31-58.

[21] J. Ferré, F.X. Rius, Anal. Chem. 70 (1998) 1999-2007.

[22] E.R. Malinowski, Factor Analysis in Chemistry, 3rd ed, John Wiley & Sons Inc: New York, 2002.

[23] A. Elbargali, J. Nygren, M. Kubista, Anal. Chim. Acta. 379 (1999) 143-158.

[24] N.M. Faber, A. Lorber, B.R. Kowalski, J. Chemom. 11 (1997) 419-461.

[25] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218-225.

[26] F.P. Zscheile, H.C. Murray, G.A. Baker, R.G. Peddicord, Anal. Chem. 34 (1962) 1776-1780.

[27] Matlab, The Mathworks, South Natick, MA, USA. Version 6.5

Supporting information

Proof that the slope (b) of the regression of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ is the concentration predicted by GRAM ($\hat{c}_t$).

The slope of a regression line is

$$b = \frac{\text{vec}\mathbf{S}^{*\text{T}}\text{vec}\mathbf{R}_{\text{t}}^{*}}{\text{vec}\mathbf{S}^{*\text{T}}\text{vec}\mathbf{S}^{*}} \qquad \text{(a1)}$$

since $\mathbf{R}_t^* = \hat{\mathbf{R}}_{\text{t}}^{*} + \mathbf{E}_t^*$, and $\text{vec}\mathbf{S}^{*\text{T}}\text{vec}\mathbf{E}_{\text{t}}^{*} = 0$ (see Figure 1):

$$b = \frac{\text{vec}\mathbf{S}^{*\text{T}}\text{vec}\hat{\mathbf{R}}_{\text{t}}^{*} + \text{vec}\mathbf{S}^{*\text{T}}\text{vec}\mathbf{E}_{\text{t}}^{*}}{\text{vec}\mathbf{S}^{*\text{T}}\text{vec}\mathbf{S}^{*}} = \frac{\text{vec}\mathbf{S}^{*\text{T}}\text{vec}\hat{\mathbf{R}}_{\text{t}}^{*}}{\text{vec}\mathbf{S}^{*\text{T}}\text{vec}\mathbf{S}^{*}} \qquad \text{(a2)}$$

From Eqs 11 and 12, it is found that:

$$\hat{\mathbf{R}}_c^* = \mathbf{P}_H\ \mathbf{H\Pi Y}^T\ \mathbf{P}_Y \qquad \text{(a3)}$$

$$\hat{\mathbf{R}}_t^* = \mathbf{P}_H\ \mathbf{H\Phi Y}^T\ \mathbf{P}_Y \qquad \text{(a4)}$$

Operating:

$$\hat{\mathbf{R}}_t^* = \frac{\mathbf{\Phi}}{\mathbf{\Pi}}\ \hat{\mathbf{R}}_c^* \qquad \text{(a5)}$$

where $\mathbf{\Pi}$ and $\mathbf{\Phi}$ are diagonal matrices (FxF). The diagonal elements of $\mathbf{\Pi}$ and $\mathbf{\Phi}$ with regard to the analyte of interest are $\frac{c_c}{c_c + \hat{c}_t}$ and $\frac{\hat{c}_t}{c_c + \hat{c}_t}$ respectively.

Hence, $\hat{\mathbf{R}}_t^* = \frac{\hat{c}_t}{c_c}\ \hat{\mathbf{R}}_c^* = \hat{c}_t\,\mathbf{S}^*$ (a6)

Updating Eq a3 with vec$\hat{\mathbf{R}}_t^* = \hat{c}_t$ vec$\mathbf{S}^*$

$$b = \hat{c}_t\,\frac{\text{vec}\mathbf{S}^{*\text{T}}\text{vec}\mathbf{S}^{*}}{\text{vec}\mathbf{S}^{*\text{T}}\text{vec}\mathbf{S}^{*}} = \hat{c}_t \qquad \text{(a7)}$$

3.4.2 Paper in preparation

Net noise estimation in a second-order chromatographic peak.

Net noise estimation in a second-order chromatographic peak

(in preparation)

ABSTRACT

A recently proposed method for outlier detection in GRAM compares the Net Analyte Signal (NAS) of the test sample and the NAS of the calibration sample. The NAS matrices are calculated from the GRAM estimated profiles. To make the NAS's comparison useful, an estimation of the amount of noise in the measured data is needed. In this paper two strategies to estimate the noise are compared: (i) using a flat region in the chromatogram, and (ii) using the bilinear structure of the peaks for estimating the noise in the time domain and in the spectral domain.

Keywords: noise estimation, NAS, GRAM, outlier, trilinearity.

INTRODUCTION

In a previous paper (*Outlier detection in GRAM applied to chromatographic data, Anal. Chem. submitted*) we developed a criterion to determine whether or not the calibration sample and the sample are trilinear. If they are not, the test sample is detected as an outlier for GRAM. The criterion is based on regressing the unfolded Net Analyte Signal (NAS) of the test sample, vec$\mathbf{R}_t^*$, against the NAS of the calibration sample at unit concentration, vec$\mathbf{S}^*$. The slope of the fitted line is the concentration estimated by GRAM. The degree of fit is an indicator of the trilinearity of the data. When data are trilinear, the residuals of the straight line are random and comparable with the amount of noise. When the residuals are large and systematic, the data are not trilinear and the concentration estimated by GRAM may be incorrect. Visual inspection of the residuals of the regression of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ may show systematic trends that suggest that the GRAM estimations are wrong. For example, Figure 8 in section 3.4.5 (page 132) shows those large and systematic residuals. However, it is always desirable to use a criterion that is more rigorous than the visual inspection of the residuals to determine their significance. The residuals are compared with the net noise of the measured data. Hence, to determine whether the residuals of the fit are significant, an estimation of the amount of noise of the data is needed.

This paper compares two methods for estimating the noise. The advantages and disadvantages of each one are commented on in the analysis of three water pollutants.

Another strategy to estimate the amount of noise is by using replicates of the analysis and calculating the standard deviation of each response (at *I* time sensor and *J* wavelength) in the different replicates. However, all replicates must be completely aligned, and the differences only due to the experimental noise. Hence, time shift can make the determination of the standard deviation of each response incorrect.

THEORY

Two methods are proposed for estimating the noise:

Method 1. Estimation of the noise using a part of the chromatogram free of the presence of the analyte. This zone (**B**) can be either the same time window where the analyte elutes but in the chromatogram of a blank sample or, from the chromatogram of the test sample, a zone close to the analyte of interest.

The net noise ($\mathbf{B}^*$) is estimated as:

$$\mathbf{B}^* = \mathbf{P}_H\mathbf{B}\mathbf{P}_Y \qquad (1)$$

where $\mathbf{P}_H$ and $\mathbf{P}_Y$ are the projection matrices described in section 3.4.1 (page 120).

When environmental and biological samples are analyzed, it is not easy to find a blank sample, i.e., a similar sample without the analyte of interest. For that reason, in the following examples, we estimated **B** from the test sample in a time window where no analyte was present.

Notice that $\mathbf{P}_H\mathbf{B}$ will cancel the contributions in the time axis that have been modeled and are not due to the analyte of interest. Therefore, if **B** contains a systematic variation not included in the model (typically a baseline variation different that one used for modeling), it will significantly contribute to $\mathbf{P}_H\mathbf{B}$. However, if the background spectrum is the same as used for calibration, the net signal in the spectral profile direction will be cancelled $\mathbf{B}\mathbf{P}_Y \approx 0$, hence, $\mathbf{B}^* = \mathbf{P}_H\mathbf{B}\mathbf{P}_Y$ will remove the signal even if the baseline in the blank zone is different than the baseline in the calibrated peaks.

Method 2. Net noise estimation from the NAS matrices of the analyte of interest ($\mathbf{R}_t^*$ and $\mathbf{S}^*$).

The residuals of the regression of vec$\mathbf{R}_t^*$ and vec$\mathbf{S}^*$ are estimated by regressing the rows and columns of $\mathbf{R}_t^*$ against the rows and columns of $\mathbf{S}^*$ respectively, i.e., along the time domain and along the spectral domain (Figure 1). To obtain an estimation of the net noise in the chromatographic profile direction, each column of $\mathbf{R}_t^*$ (each net chromatographic profile) is regressed against the first elution profile of $\mathbf{S}^*$

(although any column of $\mathbf{S}^*$ can be used, since $\mathbf{S}^*$ has rank 1). The residuals ($\mathbf{e}_p$i) from each regression are stacked in a vector ($\mathbf{e}_p$). To obtain an estimation of the net noise in the spectral direction, each row of $\mathbf{R}_t^*$ is regressed against the first row of $\mathbf{S}^*$. The residuals ($\mathbf{e}_s$i) from each regression are stacked in a vector ($\mathbf{e}_s$). Figure 1 shows the procedure.

The lack of trilinearity is mainly caused along the time domain, but not along the spectral domain (that is the reason why the spectral comparison is not enough to validate the GRAM estimations, see section 3.4.1). The residuals in the spectral direction are only due to the noise, whereas in the chromatographic profile direction the residuals are due to the noise and the possible presence of effects that introduce systematic contributions. $\mathbf{e}_p$ corresponds to the residuals of the regression of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$. Hence, when $\mathbf{e}_p$ and $\mathbf{e}_s$ are comparable, the data are trilinear.

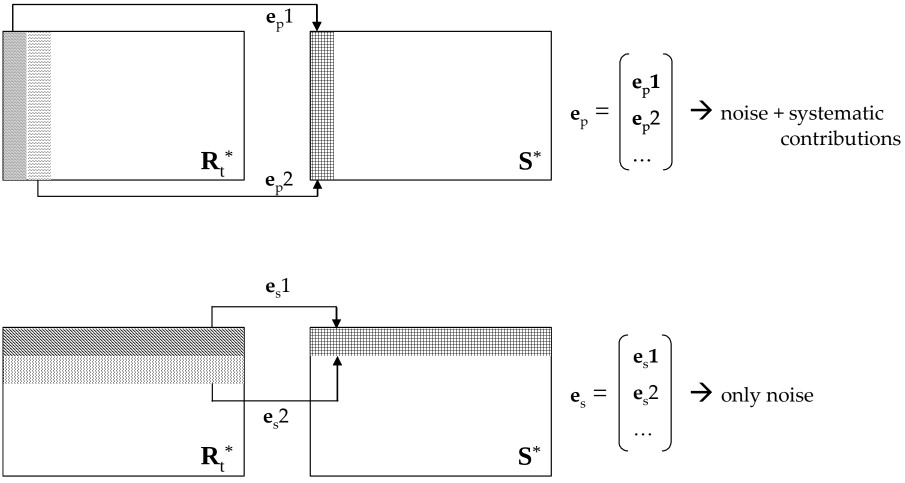


Figure 1. Net noise estimation from the NAS matrices.

EXPERIMENTAL SECTION

Samples

Three analytes were studied: 1-amino-6-naftalensulfonate, from a waste water sample, and resorcinol and phenol, from a river water sample. Figure 2 shows the two chromatograms.

1-amino-6-naftalensulfonate is used in the dye and tannery industry. This analyte is potentially hazardous. The usual way to analyze it is by ion-pair liquid chromatography. This technique is quite tedious and GRAM can reduce the separation time, since completely resolved peaks are not necessary.

Resorcinol and phenol are two polar water pollutants that can be found in river waters. As their concentration level was very low, a preconcentration step was needed prior to the chromatographic separation. This preconcentration step was not selective enough, and the humic and fulvic acids were also retained in the preconcentration cartridge and eluted at the beginning of the chromatogram, producing a high band (see Figure 2b). Resorcinol and phenol eluted in that varying baseline, making the application of univariate calibration difficult. By the application of GRAM this problem was overcome (see chapter 4 of the thesis for details).

RESULTS AND DISCUSSION

1-amino-6-naftalensulfonate

Figure 3 shows the regression of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ obtained for the GRAM model built with the peaks indicated in Figure 2. The slope of the fitted line is the predicted concentration, which is 0.082 μg l^{-1}. This result can be acceptable if the residuals are only due to the noise. If residuals are due to a lack of trilinearity, the test sample is detected as an outlier. Hence, we need to compare the magnitude of the residuals with an estimation of the noise. These residuals of the fit of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ are shown in Figure 4a together with the net noise (vec$\mathbf{B}^*$, Eq 1) estimated from the "blank" time window shown in Figure 2a. The residuals of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ are larger than the net noise, indicating that there is a lack of trilinearity and that the test sample is an outlier, and we cannot be confident of the GRAM prediction.

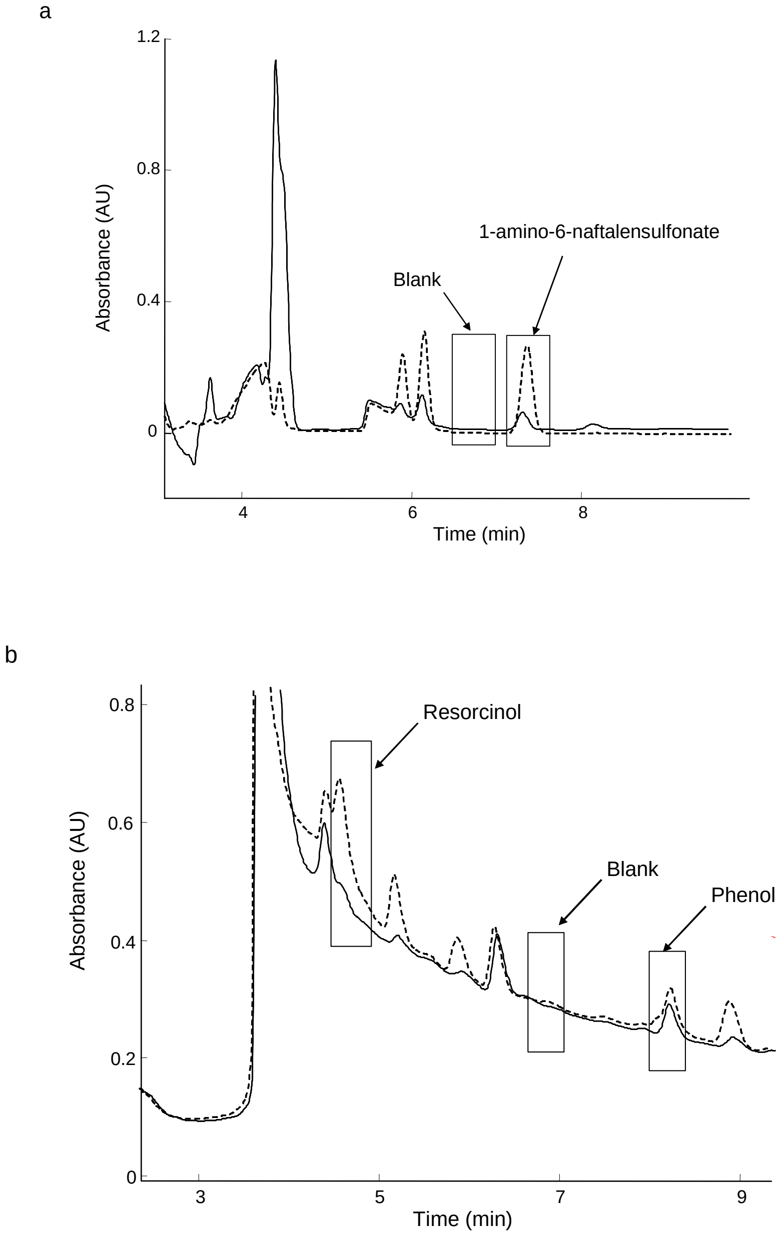


Figure 2. Chromatograms from the calibration (--) and test samples (–). (a) waste water sample and (b) river water sample measured at 240 nm. The analytes being studied and the “blank” time-window used to calculate the net noise are indicated.

The same conclusions can be drawn by considering the net noise estimated from the NAS matrices along the chromatographic profile direction (▪) and the spectral direction (+) (Figure 4b). The residuals along the elution profile direction (▪) are the same as the residuals of the fit (•).

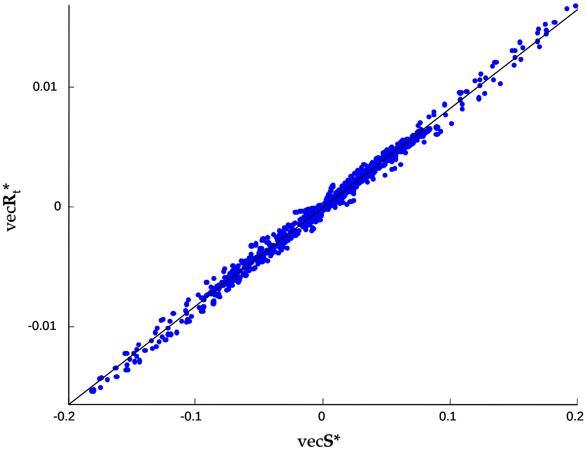
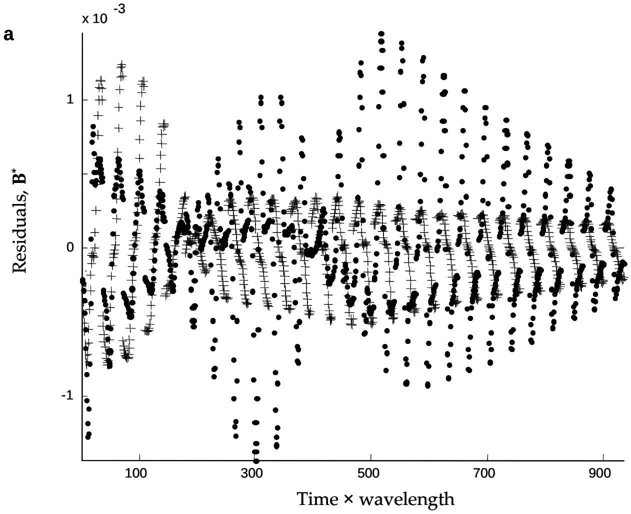


Figure 3. vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ for the 1-amino-6-naftalensufonate.



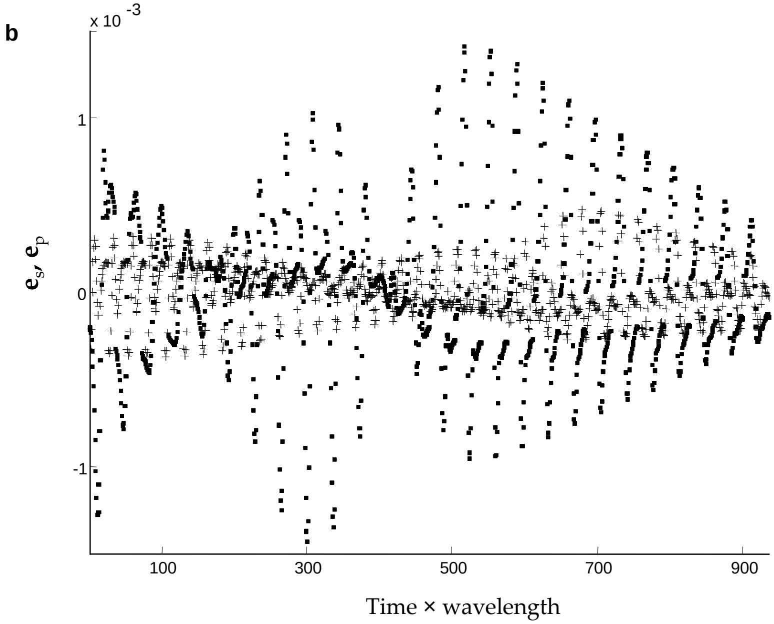


Figure 4. (a) vec$\mathbf{B}^*$ (+) and the residuals of the fit (●). (b) $\mathbf{e}_s$ from the spectral direction (+) and $\mathbf{e}_p$ from the time direction (■).

Resorcinol and phenol

As can be seen in Figure 2b, the time window of the blank was determined from an area where no analyte was present, but the baseline was different than in the windows of resorcinol and phenol. This shows the difficulty of selecting the blank zones in this kind of samples.

Focusing on resorcinol, Figure 5 shows the regression of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$. The slope corresponds to a predicted concentration, 4.69 μg l^{-1}. The fit is better than in the case of 1-amino-6-naftalensulfonate presented in Figure 3. The net noise estimated by projecting the blank zone is not significantly different than the residuals of the regression of vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ (Figure 6a). In addition, no differences are observed between the noise calculated in the time direction and in the spectral direction (Figure 6b). Hence we conclude that the data are trilinear and we can be confident of the concentration predicted with GRAM.

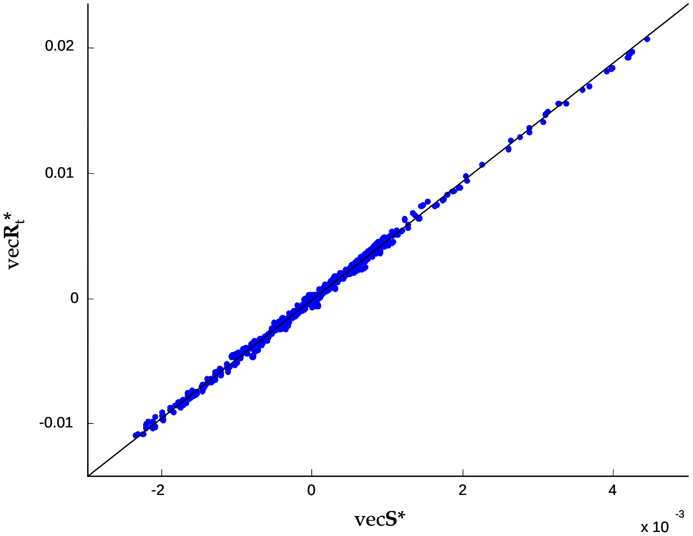
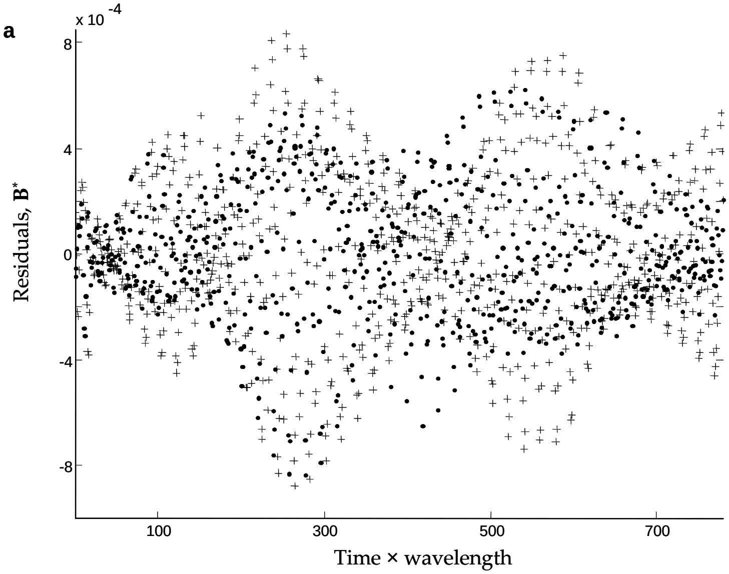


Figure 5. vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ of resorcinol.



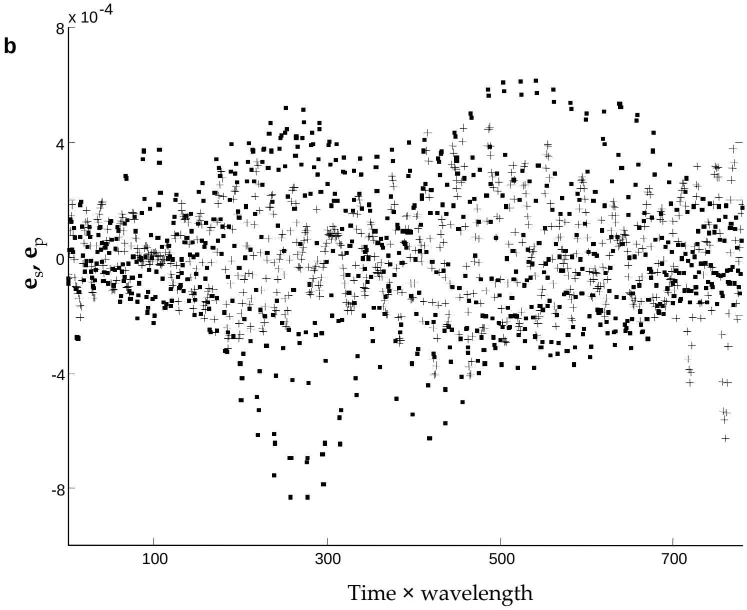


Figure 6. (a) vec$\mathbf{B}^*$ (+) and the residuals of the fit (●). (b) $\mathbf{e}_s$ from the spectral direction (+) and $\mathbf{e}_p$ from the time direction (■).

Focusing on phenol, Figure 7 shows the regression of vec$\mathbf{R}_t^*$ against $\mathbf{S}^*$. The predicted concentration is 5.32 μg l^{-1}. The apparently large residuals may indicate a lack of trilinearity. When we compare the residuals of the fit with the net noise estimated from the "blank" time window (Figure 8a), we see that the net noise is much larger than the residuals of the fit. This indicates that the "blank" zone **B** contained contributions that were not modeled by GRAM and produced systematic variations in $\mathbf{B}^*$. Notice that the **B** used here is the same as the one used for resorcinol. However, for resorcinol the background spectrum of **B** was also in the resorcinol peak. Hence, for resorcinol, $\mathbf{B}^*$ was random and not significant. This background spectrum is not present in the phenol peak, and $\mathbf{B}^*$ contains the projection of this systematic variation. This renders method 1 useless.

By considering method 2, the residuals found in the chromatographic profile direction and in the spectral direction (Figure 8b) are similar, and we conclude that the large residuals are probably due to random noise and the data are trilinear.

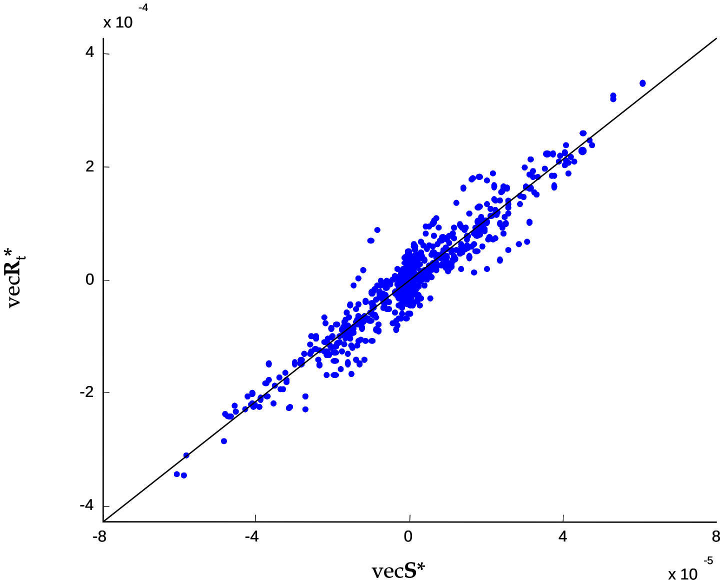
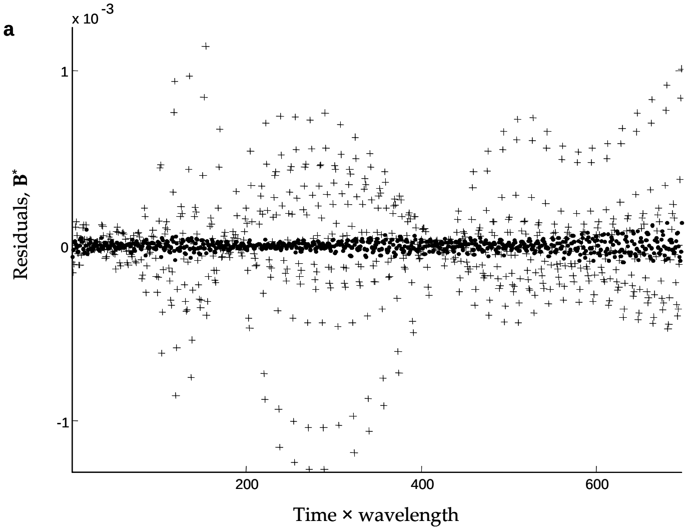


Figure 7. vec$\mathbf{R}_t^*$ against vec$\mathbf{S}^*$ for phenol.



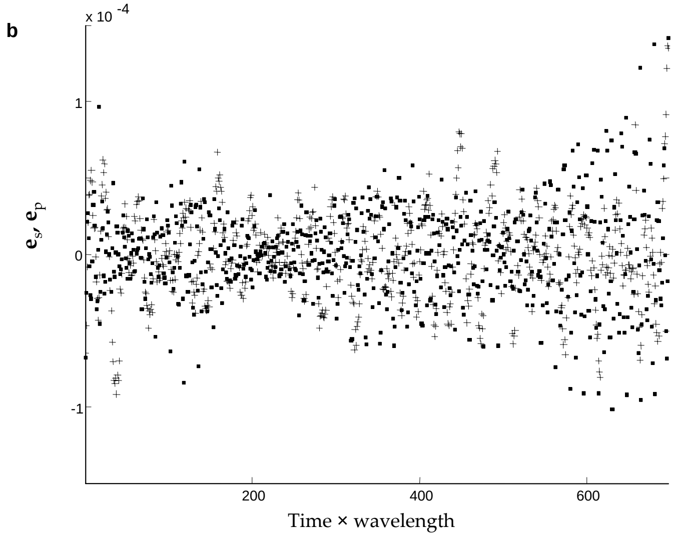


Figure 8. (a) vec**B*** (+) and the residuals of the fit (●). (b) $\mathbf{e}_s$ from the spectral direction (+) and $\mathbf{e}_p$ from the time direction (■).

CONCLUSIONS

Two methods for estimating the net noise needed for detecting outliers in GRAM have been compared. Method 1 estimates the noise from a blank sample or from a "blank" window in the test sample where the analyte is not present. This zone **B**, at least in one direction (elution profiles or spectra), must be also included in $\mathbf{R}_t$ or $\mathbf{R}_c$ in order to obtain $\mathbf{B}^*$ random. When this zone cannot be found in the chromatogram, method 1 cannot be used.

Method 2 estimates the net noise directly from the NAS of the calibration and the NAS of the test sample and it is not necessary to measure a blank region. This makes method 2 preferable to method 1.

Chapter 4

Application of GRAM to the determination of water pollutants

4. APPLICATION OF GRAM TO THE DETERMINATION OF WATER POLLUTANTS

4.1 Introduction

The aim of this chapter is to present two practical cases, where GRAM, PARAFAC and MCR-ALS were applied to the analysis of water pollutants. The analytes of interest were, in all cases, at very low concentration levels in very complex matrices. The water samples were taken from a river and a waste water plant.

The presence of these analytes in the samples can be dangerous for the environment and for human health, making their determination important.

Two studies are included. In the first one, aromatic sulfonates are determined. The chromatographic conditions were optimized to determine several components using pure standards. The total analysis lasted less than 8 minutes. When a real sample was analyzed, some analytes eluted overlapped with interferences, making their analysis by univariate calibration not possible. GRAM was applied in the above conditions. To externally validate the GRAM predictions, the chromatographic conditions were changed in order to isolate the analytes of interest in those specific samples. It took over 45 minutes to achieve this, increasing the time of separation six fold.

The second case represents another challenge for chromatographists; the determination of analytes that elute overlapped at a high band. This band typically appears when polar compounds are to be analyzed and a pre-concentration step is done. The pre-concentration step is not selective enough and many polar compounds, like the humic and fulvic acids are also retained in the preconcentration step.

After the chromatographic separation, it is not easy to determine the area of each peak. Moreover, it cannot always be assumed that the signal recorded at a range of retention times is selective for the analyte of interest, which is required for applying univariate calibration.

The change of the experimental conditions, to validate the predictions, was not successful. Another strategy was used, the addition of sodium sulfite (Na_2SO_3) to the sample. This compound reacts with the humic and fulvic acids producing non-polar compounds that do not elute at the beginning of the chromatogram. However, it should be taken into account that this compound can also react with the analytes of interest and so is not always effective. The effect of the sulfite depends on the composition of the water samples.

4.2 Paper

Using second-order calibration to identify and quantify aromatic sulfonates in water by high-performance liquid chromatography in the presence of coeluting interferences.

E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius.

Journal of Chromatography A 988 (2003) 277 – 284.

Using second-order calibration to identify and quantify aromatic sulfonates in water by high-performance liquid chromatography in the presence of coeluting interferences

Enric Comas, R. Ana Gimeno, Joan Ferré, Rosa M. Marcé,

Francesc Borrull, F. Xavier Rius

Department of Analytical Chemistry and Organic Chemistry, Rovira i Virgili University
Imperial Tàrraco, 1, 43005, Tarragona, Spain

ABSTRACT

We used the Generalized Rank Annihilation Method (GRAM), a second-order calibration method, to quantify aromatic sulfonates in water with high-performance liquid chromatography (HPLC) when interferences coeluted with the analytes of interest. With GRAM, we can quantify in only two chromatographic analyses, one for a calibration sample and one for the unknown sample. The calculated concentrations were not statistically different to those obtained when the chromatographic separation of the unknown sample was modified in order to completely separate the analyte from the interferences before univariate calibration. With GRAM, the concentrations are determined much more quickly because a complete resolution is not required.

Keywords: Generalized rank annihilation method, Chemometrics, Co-elution, Second-order calibration, Sulfonates.

1. INTRODUCTION

Many factories discharge their wastewater into rivers or directly into the sea after a treatment process to eliminate the most common contaminants. In the tannery and dye industries, aromatic sulfonates are widely used and are highly soluble in water. They are difficult to remove completely by the treatment process and have been found in effluent waters [1]. Little is known about their toxicity but they have a low biodegradability, so they are potentially hazardous for the aquatic environment. It is therefore important to monitor them in these kinds of samples.

As the polarity of these compounds is high, the most common analytical technique is ion-pair liquid chromatography with UV–Vis or fluorescence detection [2, 3]. This technique is not sensitive enough to quantify these compounds in real samples, so an enrichment step is needed before the chromatographic analysis. The most common preconcentration technique is ion-pair solid-phase extraction using highly crosslinked polymeric sorbents such as isolute ENV+, which has a high retention for the most polar analytes [1].

In natural waters, other polar compounds can also be retained in the solid-phase extraction process and coelute with the analytes of interest during the chromatographic analysis. This coelution may produce strongly biased quantifications when the concentration is determined with univariate calibration, which requires highly selective measurements. When coelution is detected, the conditions of the HPLC method must be optimized again from the unknown sample until the analyte of interest elutes separately from the interferences. This may be difficult if the properties of the analyte and interferences are similar and is an important outlay of time and resources. Also, since the interferences depend on the source of the sample, it may be cumbersome to optimize the conditions for each particular analyte and every unknown sample.

Mathematical separation is an alternative to chromatographic separation [4]. Diode array detectors (DAD) can record the UV–Vis spectra at every retention time, and a matrix (elution time×wavelength) is obtained for each peak to be quantified. Applying second-order calibration algorithms to this data matrix can: (a) indicate whether the peak of the analyte of interest contains coeluting interferences, (b) determine the number of coeluting species, (c) determine which species are present on the basis of their spectral features—qualitative analysis—and (d) determine the concentration of the analyte of interest in the overlapping peaks (known as the 'second order' advantage) [5].

Of the second-order calibration algorithms that allow quantification in the presence of non-calibrated components, the Generalized Rank Annihilation Method (GRAM) [6] is very useful for chromatographic data, where the number of analyses is important. It only requires two data matrices. One of these is from a calibration sample, i.e. the spectra measured at the different retention times of the peak of the analyte obtained by analyzing either a pure standard or a sample with a known added concentration of the analyte. The other is the spectra measured at the different retention times of the peak from the unknown sample. Moreover, GRAM has been widely studied [7-14] and mathematical expressions are available for calculating figures of merit [15] and the variance of the predicted concentrations [16].

In its application to HPLC–DAD data, it was pointed out that the different elution times of the analytes of interest between analysis is an important problem that leads to misleading results [13]. For this reason, the application of GRAM to experimental chromatographic data in routine analysis is not as straightforward. Here we report a systematic methodology for routine quantification using GRAM. This includes a previous time shift correction step with a recently developed algorithm [17] that allows a selective correction of the time shift depending on the analyte of interest.

This methodology was applied to an implemented in-house routine method for the determination of six aromatic sulfonates. When analyzing a sample of water from a sewage treatment plant in Tarragona (Spain), the peak of two of the analytes of interest overlapped with interferences. While the other four could be determined by univariate calibration, the quantification of the two other analytes required modifying the separation conditions until the peaks were completely resolved. This paper shows that it is possible to quantify the unresolved peaks with GRAM without more experimental work. Statistical tests are used to assess that the concentrations found by both GRAM and full resolution of the peaks are comparable.

2. THEORY

This section briefly describes the chemometrical tools we have used in this paper. There is a more detailed explanation of the algorithms in the cited references.

We will use these conventions: bold uppercase letters to indicate matrices, e.g. **A**; italic lowercase letters to indicate scalars, e.g. *a*; and superscript T to indicate transposition.

For every analyzed sample, the peak of the analyte of interest (either pure or overlapped with interference) is represented by a matrix **R** (time × wavelength), where the element r_{ij} represents the absorption measured at the *i*th retention time and the *j*th wavelength.

2.1. Generalized Rank Annihilation Method (GRAM)

For GRAM, the calibration matrix ($\mathbf{R}_c$) is the spectra at each retention time of the peak of the analyte obtained by analyzing the pure standard. The concentration of the analyte of interest ($c_{c,k}$) is known. The prediction matrix is the spectra at each retention time of the peak of the analyte in the unknown sample ($\mathbf{R}_t$). Both matrices are the same size ($J_1 \times J_2$) and it is assumed that they can be expressed as:

$$\mathbf{R}_c = \mathbf{X}\mathbf{C}_c\mathbf{Y}^T + \mathbf{E}_c$$

$$\mathbf{R}_t = \mathbf{X}\mathbf{C}_t\mathbf{Y}^T + \mathbf{E}_t$$

where $\mathbf{X}$ ($J_1 \times K$) and $\mathbf{Y}$ ($J_2 \times K$) contain the normalized chromatographic profiles and spectra, respectively, K is the total number of analytes in both matrices, $\mathbf{C}_c$ and $\mathbf{C}_t$ are $K \times K$ diagonal matrices of concentration related scale factors, and $\mathbf{E}_c$ and $\mathbf{E}_t$ are $J_1 \times J_2$ error matrices. Calibration and prediction with GRAM is a four-step process [16]:

1. Singular value decomposition of the matrix $\mathbf{Q}=\mathbf{R}_t+\mathbf{R}_c$ as $\mathbf{Q}=\mathbf{USV}^T+\mathbf{E}$. This equation is calculated only for a number of factors equal to the total number of analytes contained in both matrices.

2. Resolution of the eigenvalue problem $(\mathbf{S}^{-1}\mathbf{U}^T\mathbf{R}_t\mathbf{V})^T\mathbf{T}=\mathbf{T\Pi}$, where the diagonal elements of $\mathbf{\Pi}$ are the eigenvalues π_k and $\mathbf{T}$ is the matrix of eigenvectors.

3. Calculation of the chromatographic profiles (peak shapes) $\mathbf{X}=\mathbf{UST}$ and the pure spectra $\mathbf{Y}=\mathbf{V}(\mathbf{T}^{-1})^T$.

4. Calculation of the concentration of the analyte k in the unknown sample:

$$c_{t,k} = \frac{c_{c,k}\mathbf{\Pi}_k}{1-\mathbf{\Pi}_k}$$

In Step 4, we need to assign which of the calculated eigenvalues corresponds to the analyte of interest. We do this by calculating the correlation coefficient between the spectrum of the pure analyte (available from the peak of the pure standard) and the spectrum calculated with GRAM in $\mathbf{Y}$. The eigenvalue associated with the spectrum with the highest correlation is used for prediction in step 4.

2.2. Time shift correction for the unknown sample peak

One requirement that prevents GRAM from being used in routine chromatographic analysis is that the data matrices containing the peak of the analyte in the calibration sample and in the unknown sample must be trilinear [10, 18]. This means that the chromatographic profile of the analyte in the unknown sample must have the same shape and elute at the same time as in the calibration sample matrix. Of these two requirements, complete coincidence of retention time is not common in practice, because imprecision in injection timing, fluctuation in temperature, and changes in flow-rate introduce time shifts in the peaks. The characteristics of ion-pair chromatography also largely influence the time shift. Several approaches exist for solving the problem of the time shift in different chromatographic runs [14, 19] and improve trilinearity. We applied a recently developed time shift correction algorithm to $\mathbf{R}_t$ before we applied GRAM [17].

The algorithm used is based on selecting the correct time window for $\mathbf{R}_t$. Both $\mathbf{R}_c$ and $\mathbf{R}_t$ are individually decomposed into pure spectra and concentration profiles using Iterative Target Transformation Factor Analysis (ITTFA) [20-22]. The peak of the analyte of interest in both matrices is located and a time window for $\mathbf{R}_t$ is selected so that both matrices are aligned with respect to the analyte of interest. This alignment is made so that the maximum of the profile of the analyte of interest in both matrices occurs at the same time. To apply GRAM, $\mathbf{R}_c$ and $\mathbf{R}_t$ must have the same number of rows (time units) and columns (wavelengths). However, to correct the time shift, $\mathbf{R}_t$ is first selected at a wider time window than the calibration matrix to ensure that the profile of the analyte is contained in the selected window. Using this methodology the matrices are selectively aligned with regard to the analyte of interest.

3. EXPERIMENTAL

3.1. Reagents, standards and samples

3-Amino-1-benzenesulfonate, 6-amino-4-hydroxy-2-naphthalenesulfonate, 6-amino-1-hydroxy-3-naphthalenesulfonate, 1-amino-6-naphthalenesulfonate, 1-naphthalenesulfonate and 2-naphthalenesulfonate were obtained as free acids or sodium salts from Fluka (Buchs, Switzerland) or Aldrich Chemie (Beerse, Belgium). Standard solutions of 1000 mg l^{-1} of each compound were prepared in Milli-Q quality water. To increase solubility, we added several drops of sodium hydroxide 0.1 N. All samples used in this study were prepared from these solutions.

We used disodium hydrogen phosphate (Panreac, Barcelona, Spain), sodium dihydrogen phosphate (Probus, Badalona, Spain), phosphoric acid 85% (Probus, Badalona, Spain), tetrabutylammonium bromide (Fluka, Buchs, Switzerland), methanol (HPLC grade, SDS, Peypen, France) and acetonitrile (HPLC gradient grade, SDS, Peypen, France) to prepare mobile phase and samples.

Samples were collected from the output of the sewage treatment plant in Tarragona (Spain) in precleaned amber glass bottles, filtered through a 0.45-μm membrane filter and kept at 4 °C until analysis. Although the 6-amino-1-hydroxy-3-naphthalenesulfonate (A) and the 1-amino-6-naphthalenesulfonate (B) had been previously found [1] in this kind of wastewater, they were not present in the analyzed sample. Therefore, the samples were spiked at 0.08 and 0.15 mg l^{-1} to ensure their presence and test the usefulness of GRAM.

3.2. Instrumental

Chromatographic analyses were carried out using an HP1100 series system (Agilent Technologies, Waldbronn, Germany) equipped with a Rheodyne manual injector with a 20-μl injection loop, a degasser, a binary pump, a column oven and a diode-array detector. The chromatographic column was a 25.0 cm×0.46 cm Kromasil 100 C_{18} with a 5-μm particle size (Teknokroma, Barcelona, Spain).

The enrichment was carried out using a solid-phase extraction manifold (Teknokroma, Barcelona, Spain) connected to a vacuum pump (Gast Manufacturing Company, Buckinghamshire, UK).

3.3. Experimental conditions

3.3.1. Chromatographic conditions

3.3.1.1. Conditions 1

These conditions correspond to the in-house implemented method optimized for the determination of the six aromatic sulfonates indicated in the Reagents, standards and samples section. The optimal separation of a standard sample containing the six aromatic sulfonates was carried out under isocratic conditions at 30 °C with a flow-rate of 1 ml min^{-1}. The aqueous component of the mobile phase was a Milli-Q water solution containing 8 m*M* of disodium hydrogen phosphate, 8 m*M* of sodium dihydrogen phosphate and 7 m*M* of tetrabutylammonium bromide. Its pH was adjusted to 6.5 with phosphoric acid and the resulting solution was filtered through a 0.45-μm membrane filter [2]. The organic component was acetonitrile (30%). The spectra from the effluent of the chromatographic system were recorded between 220 and 300 nm, every 0.4 nm. The spectra were recorded every 0.4 s. The analysis lasted 17 min.

When we analyzed the wastewater sample, the 6-amino-1-hydroxy-3-naphthalenesulfonate (A) and the 1-amino-6-naphthalenesulfonate (B) eluted overlapped with other interferences, so we concentrated specifically on quantifying these two analytes.

3.3.1.2. Conditions 2

These conditions were determined for the wastewater sample in order to fully separate the analytes A and B that, in Conditions 1, overlapped with interferences. In this case, the optimal composition of the mobile phase was 22% acetonitrile and the chromatographic separation lasted 65 min. Absorbance was measured at 250 nm because this wavelength was selective for the analytes of interest.

3.3.2. Solid-phase extraction

Before solid-phase extraction, tetrabutylammonium bromide was added to the sample in a concentration of 3 m*M* as an ion-pairing reagent and the pH was adjusted to 7 with a disodium hydrogen phosphate/sodium dihydrogen phosphate buffer to ensure the ion-pair formation. The preconcentration cartridge, an Isolute ENV+ cartridge (International Sorbent Technology, Mid. Glamorgan, UK), was conditioned with 5 ml of acetonitrile and 5 ml of Milli-Q water. Then 50 ml of sample was preconcentrated at a flow-rate of 5 ml min^{-1}. Finally, the retained analytes were eluted with 5 ml of methanol. Solvent was eliminated with a nitrogen carrier stream and the analytes were redissolved with 1 ml of the chromatographic mobile phase. In these conditions, recoveries (of the six aromatic sulfonates) were between 50 and 90%, with %RSD between 4 and 8%.

3.4. Software

All calculations were done using in-house subroutines for MATLAB [23] version 6.

4. RESULTS AND DISCUSSION

4.1. Detection of overlap

Fig. 1 shows the superposed chromatographic profiles recorded from 220 to 300 nm of the wastewater spiked at 0.08 ppm of A and B. The vertical lines indicate the expected elution time of A and B that had been found with standards. Overlap of the peaks of these analytes was detected by visual inspection of the spectra over time and calculation of the chemical rank for each peak. A closer look at the peaks reveals that they are time shifted with respect of the peaks from the standards. Fig. 2a shows the profile of A obtained from the pure standard of 0.4 ppm of A (**R**$_c$). Fig. 2b shows the peak of analyte A overlapping with other interferences in the wastewater sample analysis, which was later used for prediction with GRAM. No selective wavelengths were found, so quantification using univariate calibration may be largely biased. With conditions 1, we used the GRAM to determine the concentration of A and B.

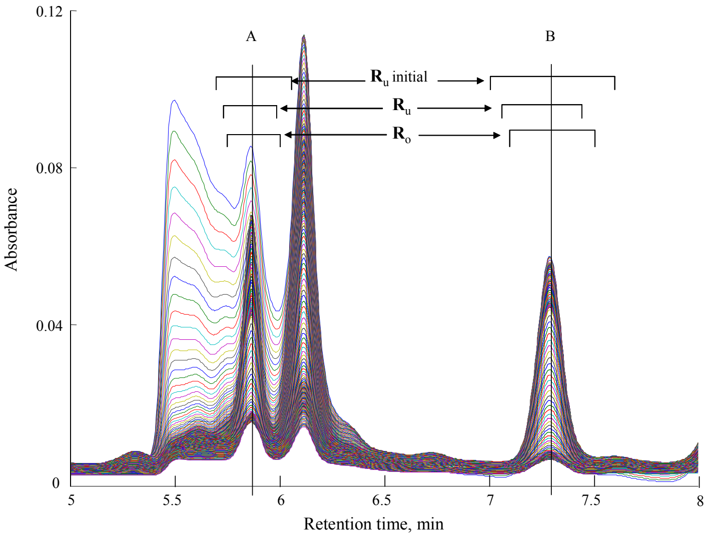


Fig. 1. Superposed chromatographic profiles of the wastewater recorded from 220 to 300 nm spiked with 0.08 ppm of A and B from 5 to 8 min. The vertical lines indicate the expected elution time for both analytes determined with standards. The time windows selected for **R**$_c$ and **R**$_t$ (before and after time shift correction) are indicated.

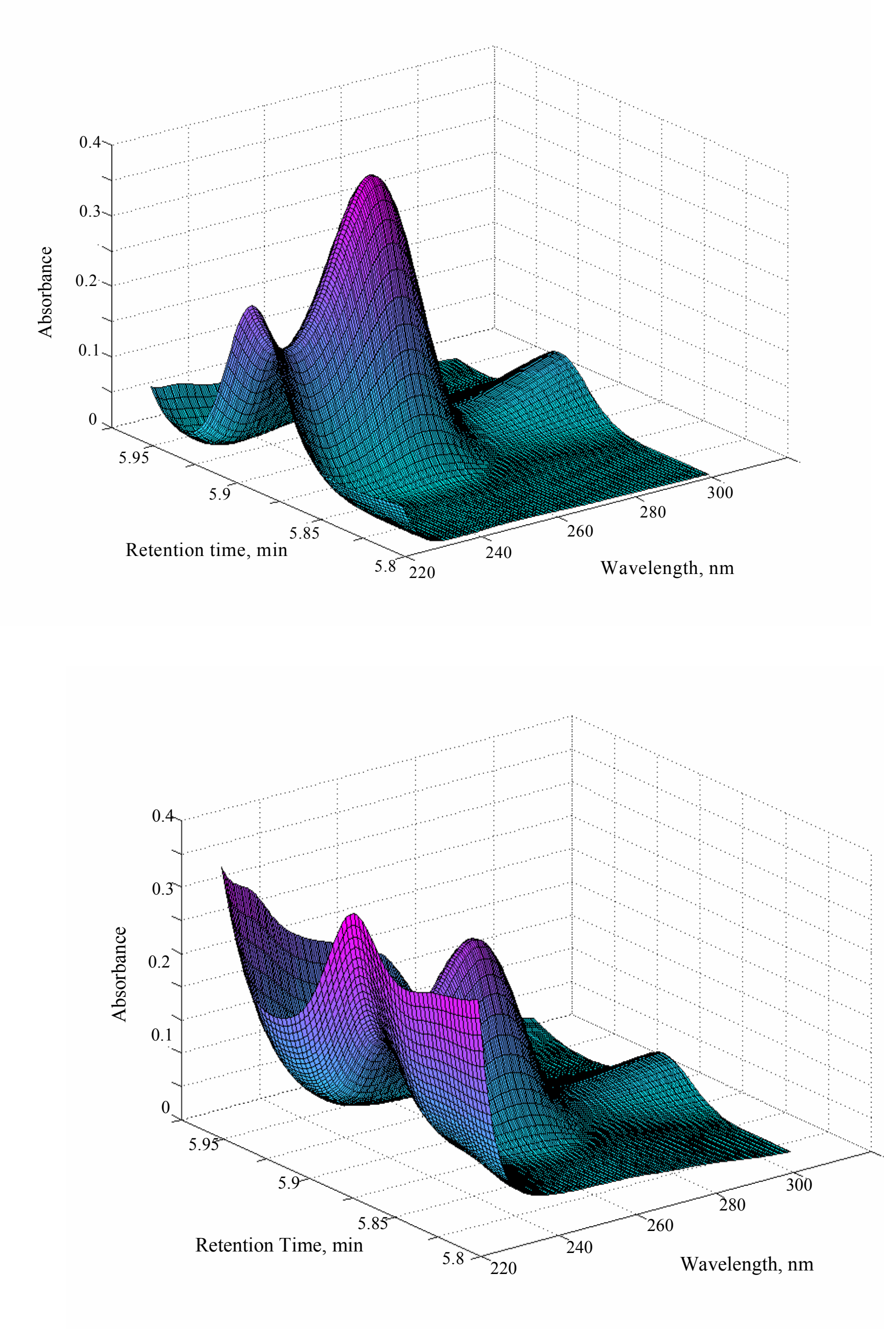


Fig. 2. Peak of analyte A in the wavelength range studied. (a) A pure standard of A. (b) Wastewater, where A elutes overlapping with interferences.

4.2. Time shift correction and GRAM

For the calibration matrices $\mathbf{R}_c$, we considered the time window where each analyte elutes. In this case, it was from 5.75 to 6.01 min for A and from 7.10 to 7.52 min for B. To correct the time shift in the wastewater sample, we selected a window that was 10 time steps wider on both sides, i.e. from 5.68 to 6.07 min for A and from 7.03 to 7.59 min for B. Fig. 1 schematically shows the time windows of $\mathbf{R}_c$ and $\mathbf{R}_t$ before and after we applied the time shift correction. Notice that for applying GRAM, the time window for $\mathbf{R}_t$ was the same size as $\mathbf{R}_c$.

In all cases, we calculated GRAM with two factors. Fig. 3 compares the spectra calculated by GRAM when determining A using the same time window for both matrices, i.e. not taking into account the time shift, and the spectra calculated by GRAM once the time shift was corrected. The calculated spectra of analyte A are very similar in both cases. The correlation coefficients of the spectrum of A in the pure standard and both calculated spectra were higher than 0.996. However, the shape of the spectrum of the interference was like that obtained in the non-spiked water only when the time shift was corrected. Results were similar for analyte B, whose correlation coefficient between the GRAM calculated spectrum of B and the spectrum of B measured in a standard sample, was higher than 0.999. If GRAM was applied without a correction of the time shift, considering the same time window for $\mathbf{R}_c$ and for $\mathbf{R}_t$, large prediction errors, around 30% were obtained.

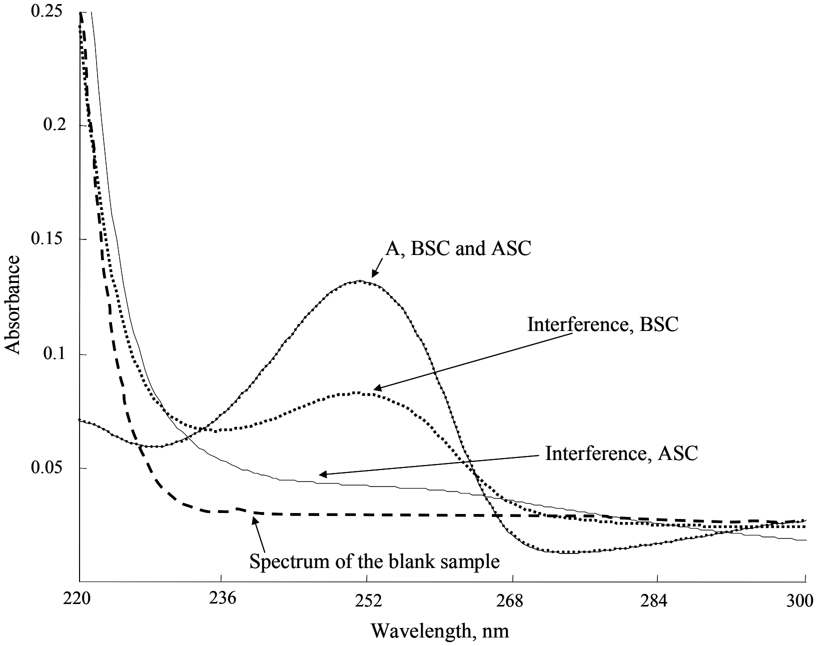


Fig. 3. Calculated spectra with GRAM in the determination of analyte A. (· · ·) Before shift correction (BSC) of **R**u; (—) after shift correction (ASC) of **R**u; (- - -) spectrum of a blank sample, where analyte A was not present.

In the initially optimized conditions 1, we recorded three replicate data matrices for the calibration sample (the standard contained 0.4 ppm of A and 0.4 ppm of B) and three for the unknown sample. Therefore, we were able to calculate nine different GRAM models (after the time shift had been corrected) by combining each calibration and each unknown sample matrix at each spiked level. To calculate the mean concentration and the precision (expressed as standard deviation) of the method, the nine models were divided into three groups of three models each, as shown in Table 1. All the models in each group are independent, since no matrix is repeated. From each group, the mean and the standard deviation of the predicted concentration are calculated. A pooled variance [24] was calculated as:

$$s^2 = \frac{\sum_i (n_i - 1)s_i^2}{\sum (n_i - 1)}$$

where n_i=3 is the number of elements in each group. The denominator corresponds to the degrees of freedom that were used in the statistical test (see next section). In this case there were six degrees of freedom. As an example, Table 1 contains the results for analyte A in the sample spiked at 0.08 ppm.

Table 1. Mean value and standard deviation of the GRAM models for the analyte A spiked at 0.08 ppm

	Group 1	Group 2	Group 3
$\mathbf{R}_c$ / $\mathbf{R}_t$	1 – I 2 – II 3 – III	1 – II 2 – III 3 – I	1 – III 2 – I 3 – II
Mean concentration	0.0650	0.0648	0.0650
Standard deviation (s_i)	0.0059	0.0006	0.0052
Grand mean (calculated concentration)	0.065		
Standard deviation	0.003		

Three groups of three independent models were analyzed, combining each calibration $\mathbf{R}_c$ (1,2,3) and prediction $\mathbf{R}_t$ (I, II, III) matrices.

4.3. Validation

We compared the predicted concentration values obtained by GRAM with the values obtained by univariate calibration. The experimental conditions were again optimized for the water sample so that analytes A and B eluted separately from any interference. Under these conditions 2, which we have specified in the Chromatographic conditions section, the test sample was measured three times. We carried out univariate calibration at 250 nm using standard solutions of A and B with concentrations ranging from 0 to 0.2 ppm. Linearity was very acceptable for this range, with determination coefficients (R^2) of 0.9984 and 0.9990 for A and B, respectively.

Table 2 shows the predicted concentration values obtained by GRAM and the values obtained by univariate calibration. An *F*-test was used to evaluate the precision of both methodologies. With a confidence interval of 95%, no significant differences were observed, i.e. at this level of significance, both strategies provide the same precision.

Table 2. Mean concentration and standard deviation obtained by GRAM and univariate calibration for analytes A and B spiked at two concentration levels. *t*-Test indicates the calculated *t*- value and the minimal alpha so that $t_{\text{calculated}} < t_{\text{tabulated}}$.

Analyte	Spiked concentration (ppm)	GRAM (conditions 1)		Univariate Calibration (conditions 2)		*t* - test	
		Calculated concentration	Standard deviation	Calculated concentration	Standard deviation	calculated	Minimal alpha (%)
A	0.08	0.065	0.003	0.065	0.005	0.01	62
A	0.15	0.167	0.007	0.173	0.005	1.09	77
B	0.08	0.084	0.003	0.089	0.002	2.05	95
B	0.15	0.171	0.005	0.166	0.003	1.35	84

We used a two-sided *t*-test to compare the results obtained with GRAM with those obtained with univariate calibration. This comparison was not carried out using the value of the initial spiked concentration in order to avoid errors due to the irreproducibility of the extraction and chromatographic processes.

In all cases, the results were similar for a confidence interval of 95%. This proves that, for the studied cases, GRAM can be used for quantification and that the results obtained with this method are similar to those obtained with univariate calibration. It is important to note that the peaks of the analytes A and B eluted in less than 8 min. In this case, their shapes were sufficiently similar among the different samples to enable good quantification. Nevertheless, in future applications of GRAM, it must be considered that if the analyte of interest elutes at much higher retention times, the shape of its chromatographic profile may vary from one sample to another, causing a significant deviation from the trilinearity and unreliable predictions. Usually, this can be detected by comparing the spectrum calculated by GRAM and the spectrum of the pure standard.

5. CONCLUSIONS

We have shown that GRAM can be used to quantify aromatic sulfonates in environmental samples with HPLC–DAD when the peak of the analyte of interest is not completely resolved from the other interferences. As it requires only two analyses, GRAM is an efficient alternative to the tedious and time-consuming chromatographic separation of the analytes followed by univariate calibration. The problem of time shift between the calibration and the unknown sample in GRAM can be solved, and results are similar to univariate calibration. GRAM can be applied to samples from different sources without any extra experimental work. With univariate calibration, optimization must be done for each individual analyte in every sample, which in practice is almost impossible.

ACKNOWLEDGEMENTS

The authors would like to thank the Spanish Ministry of Education, Culture and Sports (Project BQU 2000-1256) for economic support and the CIRIT, of the Catalan Government, for providing Enric Comas with a doctoral fellowship

REFERENCES

[1] M.C. Alonso, D. Barceló, Anal. Chim. Acta 400 (1999) 211.

[2] R.A. Gimeno, R.M. Marcé, F. Borrull, Chromatographia 53 (2001) 22.

[3] S. Fichtner, F.Th. Lange, W. Schmidt, H.J. Brauch, Fresenius J. Anal. Chem. 353 (1995) 57.

[4] R. Bro, J.J. Workman, P.R. Mobley, B.R. Kowalski, Appl. Spectrosc. 32 (1997) 237.

[5] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) A782.

[6] E. Sanchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496.

[7] N.M. Faber, L.M.C. Buydens, G. Kateman, J. Chemom. 8 (1994) 181.

[8] N.M. Faber, L.M.C. Buydens, G. Kateman, J. Chemom. 8 (1994) 273.

[9] N.M. Faber, L.M.C. Buydens, G. Kateman, J. Chemom. 8 (1994)147.

[10] L.S. Ramos, E. Sanchez, B.R. Kowalski, J. Chromatogr. 385 (1987) 165.

[11] M.J.P. Gerritsen, H. Tanis, B.G.M. Vandeginste, G. Kateman, Anal. Chem. 64 (1992) 2042.

[12] E. Sanchez, L.S. Ramos, B.R. Kowalski, J. Chromatogr. 385 (1987) 151.

[13] S. Li, P.J. Gemperline, K. Briley, S. Kazmierczak, J. Chromatogr. B 665 (1994) 213.

[14] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218.

[15] N.M. Faber, R. Boqué, J. Ferré, Chemom. Intell. Lab. Syst. 55 (2001) 91.

[16] N.M. Faber, J. Ferré, R. Boqué, Chemom. Intell. Lab. Syst. 55 (2001) 67.

[17] E. Comas, R.A. Gimeno, J. Ferré, R. M Marcé, F. Borrull, F.X. Rius, Anal. Chim. Acta 470 (2002) 163.

[18] R.B. Poe, S.C. Rutan, Anal. Chim. Acta 283 (1993) 845.

[19] B. Grung, O.M. Kvalheim, Anal. Chim. Acta 304 (1995) 57.

[20] P.J. Gemperline, Anal. Chem. 58 (1986) 2656.

[21] B.G.M. Vandeginste, F. Leyten, M. Gerritsen, J.W. Noor, G. Kateman, J. Frank J. Chemom. 1 (1987) 57.

[22] P.K. Hopke, Chemom. Intell. Lab. Syst. 6 (1989) 7.

[23] Matlab, The Mathworks, South Natick, MA, USA.

[24] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam (1997).

4.3 Paper

Quantification from highly drifted and overlapped peaks using second-order calibration methods.

E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius.

Journal of Chromatography A 1035 (2004) 195 – 202.

Quantification from highly drifted and overlapped chromatographic peaks using second-order calibration methods

**Enric Comas, R. Ana Gimeno, Joan Ferré, Rosa M. Marcé,
Francesc Borrull, F. Xavier Rius**

*Department of Analytical Chemistry and Organic Chemistry, Rovira i Virgili University
Pl. Imperial Tarraco, 1, 43005, Tarragona, Spain*

ABSTRACT

For determining low levels of pesticides and phenolic compounds in river and wastewater samples by high performance liquid chromatography (HPLC), solid phase extraction (SPE) is commonly used before the chromatographic separation. This preconcentration step is not necessarily selective for the analytes of interest and it may retain other compounds of similar characteristics as well. In this case, we present, humic and fulvic acids caused a large baseline drift and overlapped the analytes to be quantified. The inaccurate determinations of the area of the peaks of these analytes made it difficult to quantify them with univariate calibration. Here we compare three second-order calibration algorithms (generalized rank annihilation method (GRAM), parallel factor analysis (PARAFAC) and multivariate curve resolution–alternating least squares (MCR–ALS)) which efficiently solve this problem. These methods use second-order data, i.e., a matrix of responses for each peak, which is easily obtained with a high performance liquid chromatography–diode array detector (HPLC–DAD). With these methods, the area does not need to be directly measured and predictions are more accurate. They also save time and resources because they can quantify analytes even if the peaks are not resolved. GRAM and PARAFAC require trilinear data. Biased and imprecise concentrations (relative standard deviation, %R.S.D.=34) were obtained without correcting the time shift. Hence, a time shift correction algorithm to align the peaks was needed to obtain accurate predictions. MCR–ALS was the most robust to the time shift. All three algorithms provided similar mean predictions, which were comparable to those obtained when sulfite was added to the samples. However, the predictions for the different replicates were more similar for the second-order algorithms (%R.S.D. = 3) than the ones obtained by univariate calibration after the sulfite addition (%R.S.D. = 13).

Keywords: Peak overlap, Second-order calibration, GRAM, PARAFAC, MCR–ALS, Water analysis, Uncertainty reduction, Pesticides, Phenolic compounds.

1. INTRODUCTION

High performance liquid chromatography with diode array detection (HPLC–DAD) is routinely used for the qualitative and quantitative analysis of natural samples. In optimized separation conditions, each chromatographic peak ideally corresponds to a single compound. Actually, peaks may overlap, particularly when the samples are environmental and biological and have a complex matrix. In this case, quantification with univariate calibration requires special attention in order to neither incorporate bias nor reduce precision.

One such case is shown in Fig. 1. The chromatogram is of a water sample from a sewage treatment plant, which is studied in this paper. The analytes of interest are two phenolic compounds (resorcinol and phenol) and two pesticides (oxamyl and methomyl). These compounds are potentially hazardous for the environment and human health, so they are regulated by the European Union (EU) to ensure good quality bathing [1] and drinking water [2]. Because of their low concentrations, a preconcentration step by solid phase extraction (SPE) is carried out before the chromatographic separation [3, 4]. The SPE process also retained humic and fulvic acids because their polarity was similar to that of the analytes of interest. This caused a large peak at the beginning of the chromatogram (around 3–4 min) and baseline drift. This baseline drift considerably increases the uncertainty of the predicted concentration of resorcinol if univariate calibration is used, since it is not possible to know where the peak starts and finishes. Since the baseline cannot be defined precisely, both the area and the height of the peak will be uncertain. Moreover, univariate calibration requires selective measurements, i.e., the area or height of the peak must be due only to the analyte of interest. Here, it is difficult to check whether other compounds of similar polarity coeluted with the analyte of interest, since the spectrum at each retention time also contains the contribution of the humic and fulvic acids. Hence, the peak purity parameter that is commonly found in the software of the HPLC instrument will fail.

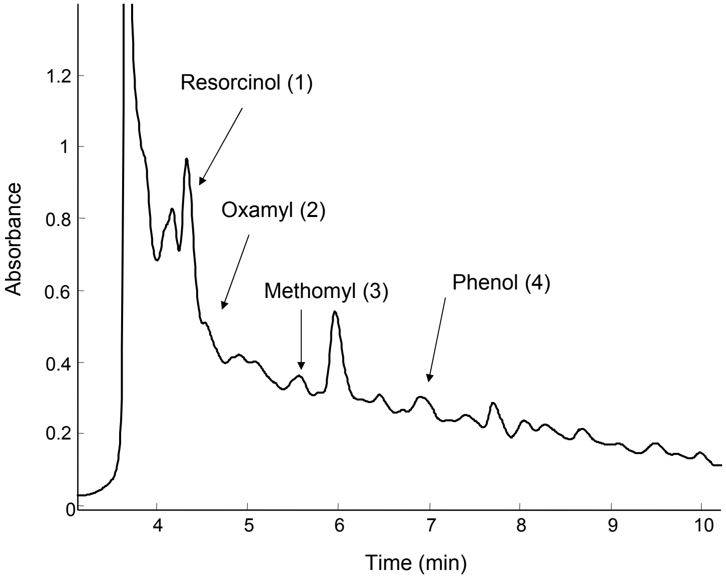


Fig. 1. Chromatographic profile of the sewage treatment plant sample measured at 240 nm. The analytes of interest are indicated.

The analytes of interest can be determined more precisely by changing the experimental conditions to achieve full resolution. This involves spending time and resources and there is no guarantee that the separation will be complete. In particular, resorcinol is difficult to isolate from humic and fluvic acids because their chemical properties are similar.

A second option is to add sodium sulfite (Na_2SO_3) to the sample before it is preconcentrated [5]. This compound reacts with the humic and fulvic acids and makes them elute separately from the analytes of interest. However, the effect of sodium sulfite depends on the sample matrix and in some cases, such as the analyses of water from a sewage treatment plant (see below), it is not useful.

In this paper, we study and apply a third solution: the chemometric processing of the peak, in order to obtain the net contribution of the analyte of interest. This can be done with a variety of mathematical approaches. Basically, when the detection is based on absorbance responses in the UV-Vis region, they can be classified into two groups: those based on mono-channel detection, i.e, one absorbance value measured at each retention time; and those based on multi-channel detection, i.e., a UV-Vis spectrum measured at each retention time.

The approaches that use mono-channel detection include neural networks [6], genetic algorithms [7], differential signal detection [8] and the development of a set of equations that model the chromatographic peak [9, 10]. One of the drawbacks of these methods is that they must assume that the chromatographic profile has a particular shape and that each peak has a number of analytes. Meyer [11, 12] fully discussed how the area of the peak should be measured for different experimental situations. However, these conditions were limited to overlapping peaks containing only the analyte of interest and a single interference.

Here, we show that multi-channel detection with HPLC–DAD instruments can be used to treat this problem in a more efficient way. Since we can measure the spectrum at each retention time, a matrix of absorbances can be obtained for each peak analyzed: a second-order data matrix. Each row of the matrix is a spectrum measured at each retention time. Each column is a chromatographic profile at one wavelength. Fig. 2 shows the second-order data matrix of the resorcinol peak.

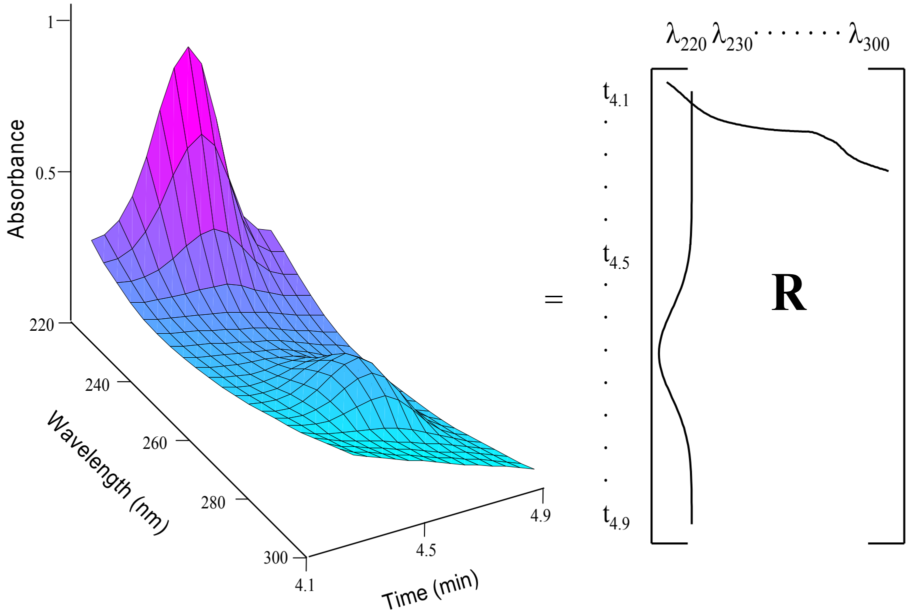


Fig. 2. Second-order data and its equivalence in matrix notation for the resorcinol peak.

Several algorithms can be used to predict the analyte concentration in a non-resolved peak using second-order data [13-17]. Here, we compare the performance of the three that are most commonly used: generalized rank annihilation method (GRAM) [13], parallel factor analysis (PARAFAC) [14] and multivariate curve resolution–alternating least squares (MCR–ALS) [15]. They make quantification possible even if the test sample contains interferences that are not considered in the calibration samples. This is known as 'the second-order advantage' [18]. This advantage is particularly looked for in our case where the interferences in the sample are the humic and fulvic acids.

Mitchell and Burdick [19] argued that PARAFAC and MCR–ALS have better properties and that their results are more reliable than those of GRAM. Recently, Faber [20] compared them in a simulation study and concluded that GRAM can also be a useful option in many cases. Here, we extend Faber's study to a real case: the analysis of water samples from the Ebre river (Spain), and from a sewage treatment plant in Tarragona (Spain). Hence, the objective of this paper is two-fold: (a) to demonstrate that GRAM, PARAFAC and MCR–ALS can be used to quantify from highly drifted and overlapping peaks and (b) to point out in which situations one method is better than the others. Their results were externally validated by a reference methodology based on chromatographic optimization and univariate calibration.

2. EXPERIMENTAL SECTION

2.1. Reagents and standards

The compounds studied were: (1) resorcinol (Sigma, Madrid, Spain), (2) oxamyl (Riedel-de-Haën, Seelze, Germany), (3) methomyl (Riedel-de-Haën), (4) phenol (Aldrich Chemie, Beere, Belgium), (5) 4-nitrophenol (Aldrich Chemie), (6) 2,4-dinitrophenol (Aldrich Chemie). They are all more than 97% pure. Standard solutions at a concentration of 2000 mg l^{-1} were prepared in acetonitrile (SDS, Peypen, France) for compound 1 and methanol (SDS) for the other compounds. These solutions were stored at 4 °C. All the working solutions were prepared by diluting these standard solutions. Analytes 1 to 4 were to be determined. Analytes 5 and 6 were included to test the reproducibility of the system.

HPLC gradient grade acetonitrile (SDS) was used for the mobile phase in the chromatographic separation and the extraction process. Ultra pure water was prepared by ultra filtration with a Milli-Q water purification system (Millipore, Bedford, MA, USA). Hydrochloric acid (Probus, Barcelona, Spain) was used to adjust the pH of the mobile phase and the samples. In the validation of the results obtained by the second-order algorithms, sodium sulfite (Probus) was added to reduce the peak at the beginning of the chromatogram caused by humic and fulvic acids in the water samples.

2.2. Samples

Samples were collected from the Ebre River (Spain) and from the output of the sewage treatment plant in Tarragona (Spain) in precleaned amber glass bottles. The pH of these samples was adjusted to 2.5 with hydrochloric acid in order to prevent the compounds of interest from being in ionic form. They were filtered through a 0.45 µm membrane filter and kept at 4ºC until analysis.

The analytes of interest have only occasionally been found in this kind of samples [4]. To ensure that they were actually present, the samples were spiked at different levels of concentrations. One aliquot of the river-water sample was spiked at 5 µgl^{-1} for resorcinol and at 1 µg l^{-1} for the other analytes. This sample was taken as the test sample. In the same way, one aliquot was spiked at 20 µg l^{-1} for resorcinol and at 5 µg l^{-1} for the other analytes. This sample was taken as the calibration sample.

The sample from the sewage treatment plant was treated in the same way. Here the levels were 20 µg l^{-1} for resorcinol and 5 µg l^{-1} for the other analytes in the test sample, and 80 µg l^{-1} for resorcinol and 20 µg l^{-1} for the other analytes in the calibration sample.

2.3. Instrumental

The chromatographic separation was carried out using an HP1100 system (Agilent Technologies, Waldbronn, Germany). This system consisted of a degasser, two isocratic pumps, a manual injector provided with a 20 µl loop, a column oven and a DAD. Each pump was used to deliver one fraction of the mobile phase. Separation was carried out using a 25 cm×0.46 cm Kromasil 100 C_{18} chromatographic column with a 5 µm particle size (Teknokroma, Barcelona, Spain).

For on-line SPE, an Applied Biosystems pump (Ramsey, USA) was used to preconcentrate samples through a stainless steel precolumn (10 mm×3 mm, i.d.) (Free University, Amsterdam, The Netherlands), which was laboratory-packed with isolute ENV + sorbent (International Sorbent Technology, Mid. Glamorgan, UK).

Chromatographic and extraction systems were on-line coupled by means of a Rheodyne 7010 valve. The set-up of the system allowed the compounds retained in the extraction cartridge to be eluted with only the organic part of the mobile phase [21]. This set-up was used to prevent the peaks from broadening out because of the low elutropic force of the mobile phase.

2.4. Experimental conditions

2.4.1. Separation

Chromatographic separation was performed under gradient conditions. The mobile phase consisted of acetonitrile and Milli-Q water (pH 3 adjusted with hydrochloric acid to prevent the column degradation). The gradient started with 20% of acetonitrile and it was linearly increased to 55% in 20 min and then to 100% in 5 min. This percentage was maintained for 10 min to return to the initial conditions in 5 min. The column was equilibrated for 5 min. The temperature of the column was 65 °C and the mobile phase flow rate was 1 ml min^{-1}. The spectrum of the effluent was recorded between 220 and 300 nm every 0.4 nm. For univariate calibration, absorbance at 240 nm was used.

2.4.2. Solid phase extraction

The on-line solid phase extraction was as follows: the precolumn was first washed with 10 ml of acetonitrile and then with 10 ml of Milli-Q water (pH 2.5 adjusted with hydrochloric acid) at 4 ml min^{-1}; the position of the valve was changed and the tubes were then purged with the sample; finally, the appropriate volume of sample was preconcentrated at 4 ml min^{-1}. The retained analytes were eluted in back-flush mode by means of the acetonitrile of the mobile phase when the valve position was changed again. The sample volume preconcentrated was 100 ml for the river-water and 25 ml for the sewage treatment plant water.

For univariate calibration, 1 ml of sodium sulfite 10% (w/v) solution was added to the sample before it was preconcentrated in order to decrease the high peak that appears at the beginning of the chromatogram when the river-water was preconcentrated.

2.5. Algorithms

Three second-order calibration methods were considered: generalized rank annihilation method, parallel factor analysis and multivariate curve resolution–alternating least squares. The three methods decompose the chromatographic peak into pure chromatographic profiles and their corresponding spectra. By including samples with known concentration, they can be used as calibration methods. So, we tested how well they predicted the concentration of the analytes of interest when the baseline drift was large, this drift being caused by the presence of the humic and fulvic acids.

The equations can be found elsewhere [13-15]. Briefly, GRAM only needs the peak of the analyte from a calibration sample (which can be either a pure standard [22] or a spiked sample [23]), and the peak of interest in the test sample. This is very attractive for the routine use of chromatography, since there is no need to measure additional samples, which is an important saving of time and resources. The algorithm is non-iterative and based on the resolution of an eigenvalue problem. It is very fast (less than 1 s on a Pentium IV 1.4 GHz) and figures of merit can be calculated easily [24, 25].

PARAFAC and MCR–ALS are iterative methods and can work with more than two samples. They need initial estimations of the chromatographic profiles or the spectra to start the iterative process [26]. Here we used, as initial chromatographic profiles, the solutions of the evolving factor analysis [27] applied to the test sample. An attractive property of PARAFAC is that the decomposition of the peak is unique, with no rotational ambiguities. To improve the solutions from PARAFAC and MCR–ALS, constraints in the iterative process are imposed, based on the chemical knowledge of the system. For chromatographic peaks, we imposed that the chromatographic profiles and the spectra had to be non-negative and that the chromatographic profile of each analyte had to be unimodal (one maximum only).

GRAM and PARAFAC require perfect trilinear data whereas MCR–ALS does not. Trilinearity can be viewed as an extension of Beer's law to second-order data. This amounts to assuming that the measured peak is the sum of the individual peaks of each analyte and that the profile and the spectrum of one analyte are proportional in all the samples. However, trilinearity is not always accomplished in chromatography. For it to be so, the profile of the analyte of interest must elute at exactly the same retention time in all the samples. In practice, time shift is usual in this kind of analysis [22] because of imprecision in the injection or fluctuations of pressure and temperature in the on-line system. Moreover, as the chromatographic separation is done in gradient mode, time shift is even more significant than when isocratic conditions are used. Several methods have been proposed for correcting the time shift [28, 29]. Prazen et al. [28] plotted the eigenvalues of the augmented matrix containing the calibration sample peak and the test sample peak, for different time windows of the test sample. A minimum in the plot indicated the optimal window. Comas et al. [29] selected the time window of the test sample after the deconvolution of the calibration and the test samples independently, using a curve resolution method, the iterative target transformation factor analysis (ITTFA). Both methods were tested in a preliminary step and provided the same results. The one described by Comas et al. [29] was used

2.6. Validation of the results from second-order algorithms

Validation of the predictions from second-order calibration algorithms is currently an active area of research [22, 30]. The philosophy underlying these algorithms is different than for multivariate calibration methods such as partial least squares (PLS) or principal components regression (PCR). In multivariate calibration, calibration and prediction are independent steps. Hence, we can check the performance of the model before it is used for prediction. In second-order calibration, both calibration and prediction are performed in one step, and both calibration and prediction samples are used at the same time. That is to say, a new model is calculated for each sample analyzed. Hence, methods are needed to check that the model is calculated correctly and to guarantee as far as possible the accuracy of the predicted concentration in the test sample. This process, which is called internal validation, is possible thanks to the fact that the three methods studied provide the pure spectrum and the chromatographic profile of each analyte. If the calculated spectra are comparable with the true ones (known from standards), and the estimated chromatographic profiles are non-negative and unimodal, the confidence that the predictions are correct is greater.

2.7. Software

The PARAFAC routine belongs to the N-way toolbox of R. Bro and C. Andersson and was downloaded from their website [31]. The MCR–ALS routine belongs to the MCR toolbox of R. Tauler and A. de Juan and was downloaded from their website [32]. We made the GRAM and ITTFA algorithms subroutines in house for MATLAB version 6 [33].

2.8. Data acquisition and data processing

The following procedure was used:

(1) The reproducibility of the on-line preconcentration and separation system was estimated before the second-order calibration methods were applied and validated. Poor reproducibility would make the study meaningless.

(2) Each sample was analyzed by the on-line SPE–HPLC–DAD method, and the second-order chromatogram was recorded.

(3) For both the calibration (with known concentration of the analytes of interest) and test samples, we manually selected the time window in which each analyte of interest eluted. When the start and the end of the peak was uncertain (e.g., resorcinol in Fig. 1), we considered a wider range. The start and end of the peaks need not be precisely estimated for second-order

calibration algorithms. These algorithms also make quantification possible with only a fraction of the peak.

(4) The chromatographic profiles were aligned with a time shift correction algorithm [29]. This was necessary for GRAM and PARAFAC.

(5) GRAM, PARAFAC and MCR–ALS were applied to the corrected peaks.

(6) The predictions were internally validated by checking that the predicted spectra were similar to the spectra of the pure analytes, and that the chromatographic profiles were non-negative and unimodal. This gave confidence in the predictions.

(7) The predictions were externally validated. They were compared to the predictions obtained by adding sodium sulfite to the sample and using univariate calibration. External validation was only possible for the river-water sample. Sodium sulfite had no effect on the water from the sewage treatment plant.

3. RESULTS AND DISCUSSION

3.1. River-water sample

In order to check the reproducibility of the analytical procedure, we analyzed five replicates of the calibration sample (i.e., five aliquots of the same spiked sample). The reproducibility was checked both graphically and by measuring the area of the peaks. Measuring the area in highly drifted peaks was not easy so we added analytes 5 (4-nitrophenol) and 6 (2,4-dinitrophenol) to the sample. These compounds are less polar so they eluted at 9.5 and 12 min, respectively, far from the peak of the humic and fulvic acids. Fig. 3 shows the chromatographic profile of the five replicates measured at 240 nm.

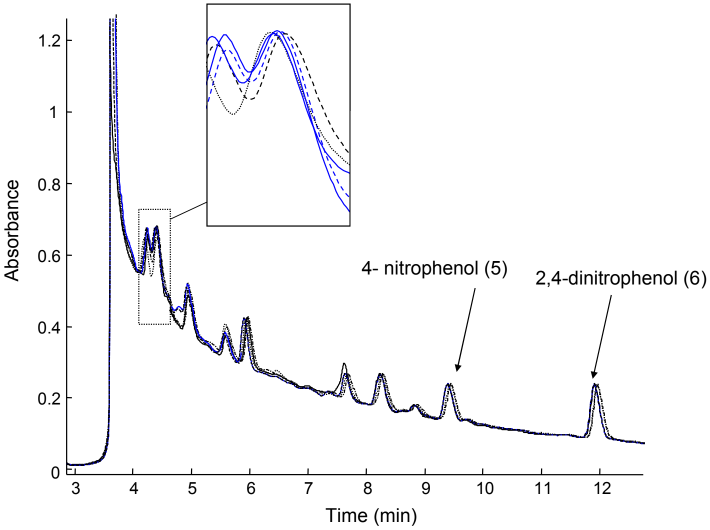


Fig. 3. Five replicates of the river-water sample with absorbance measured at 240 nm and used to check the reproducibility of the analytical system.

Table 1 shows the mean value of the area calculated by the integration algorithm of the HPLC instrument and its relative standard deviation (R.S.D.) expressed as a percentage. Taking into account the low concentration levels determined, the reproducibility of the on-line system is acceptable for this kind of analysis, and it is similar to what has already been reported [4].

Table 1. Area of the peaks in the different replicates of river-water

Analyte	Mean value	RSD (%)
Resorcinol	523.2	10.0
Oxamyl	741.2	2.5
Methomyl	663.5	3.4
Phenol	409.2	3.1
4-nitrophenol	665.5	2.1
2,4-dinitrophenol	1777	1.1

A closer look at the peak of resorcinol in Fig. 3 shows that the maximum of the peak in the different replicates was not at the same retention time, but that the maximum absorbance was the same. This time shift is usual in this kind of analysis and had to be corrected before GRAM and PARAFAC were applied.

Once the reproducibility had been assessed, the test samples were analyzed under the same conditions. The selected time ranges where each analyte eluted are shown in Table 2.

Table 2. Time range selected for each analyte

Analyte	Initial time (min)	Final time (min)
Resorcinol	4.16	4.83
Oxamyl	4.78	5.30
Methomyl	5.44	5.85
Phenol	8.05	8.66
4-nitrophenol	9.11	9.75
2,4-dinitrophenol	11.84	12.30

GRAM, PARAFAC and MCR–ALS were run with only two matrices, i.e., one calibration and one test sample. Since we used two replicates for the calibration sample and two for the test samples, we built four models for each algorithm and analyte. In MCR–ALS, the matrices were considered column-wise, i.e, the spectra were considered to be common in both matrices. In all cases the number of factors needed to run these algorithms corresponded to the sum of the number of analytes in both matrices. Several methods have been developed to determine the number of factors [30, 34-36]. The one used here was the *F*-test [36]. In all cases the number of factors was either 2 or 3, but never 1, which is what is required for univaritate calibration. Table 3 shows the mean predicted concentration (from the four models) and its relative standard deviation (%) when the same time window was considered for the calibration and test samples (before SC in Table 3) and after the time shift had been corrected (after SC). When the time shift was not corrected, the three methods gave substantially different predictions, especially for resorcinol and oxamyl. Also, the predicted concentrations are very dissimilar among replicates, resulting in an increase in the R.S.D. value. The reason for this is that GRAM and PARAFAC require trilinear data, whereas ALS does not. When the time shift was corrected, the predictions of the three methods were similar and the R.S.D. for each analyte was considerably reduced. GRAM and PARAFAC predicted very similar concentrations, and the four models provided close predictions. On the other hand, the predictions made by MCR–ALS with and without correction of the time shift are very similar. This was to be expected since MCR–ALS does not require the data to be trilinear in the time mode.

Table 3. Mean value (μg l^{-1}) and its R.S.D. (%) of the predicted concentration with second-order calibration methods, considering the same time range in the calibration and test sample (before SC) and after the time shift had been corrected (after SC)

Analyte	GRAM				MCR-ALS				PARAFAC			
	Before SC		After SC		Before SC		After SC		Before SC		After SC	
	Mean	RSD	Mean	RSD	Mean	RSD	Mean	RSD	Mean	RSD	Mean	RSD
Resorcinol	10.33	21.3	4.39	8.2	5.41	5.8	5.11	6.1	4.18	34.1	4.20	12.6
Oxamyl	0.98	3.6	0.95	3.4	1.18	13.0	1.12	1.8	0.89	21.6	0.96	3.8
Methomyl	0.98	5.1	1.00	3.1	1.04	1.1	1.04	1.0	0.95	5.7	1.00	3.9
Phenol	1.28	6.4	1.31	3.0	1.21	1.5	1.20	1.2	1.26	8.4	1.30	2.9

The reliability of the results was first checked by internal validation. Fig. 4 compares the spectra of resorcinol obtained with GRAM, PARAFAC and MCR–ALS. All three spectra are very similar, with correlation coefficients higher than 0.999 which shows that the results of the three methods are similar. An extensive study is being carried out in our laboratory to test which is the threshold value in the correlation coefficient to be confident of the predictions.

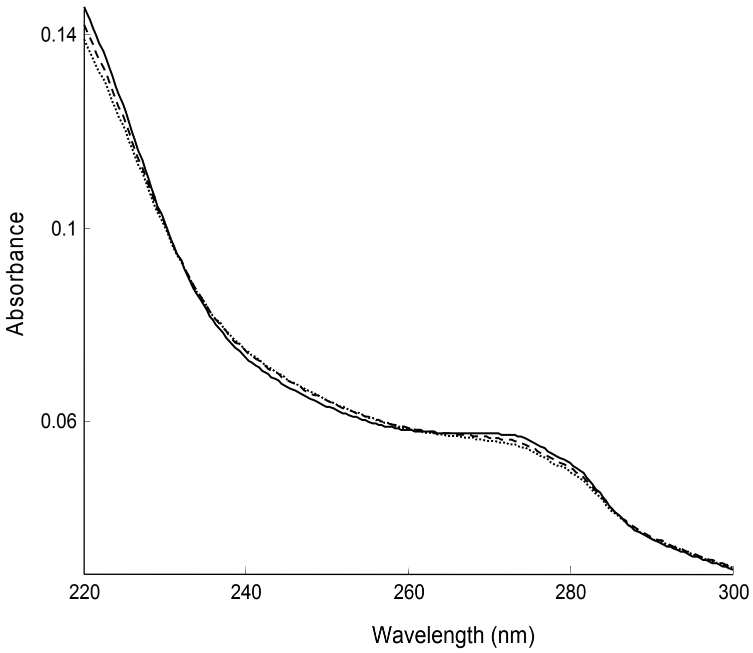


Fig. 4. Spectra provided by GRAM (–), PARAFAC (···) and MCR–ALS (- -).

Finally, the predictions were externally validated with univariate calibration. Sodium sulfite was added to the sample to decrease the large band corresponding to the humic and fulvic acids. Fig. 5 shows the chromatogram at 240 nm of the same sample before and after sodium sulfite had been added. The sulfite was successful at removing the peak of fulvic and humic acids and univariate calibration could be used since the area of each peak was determined more accurately.

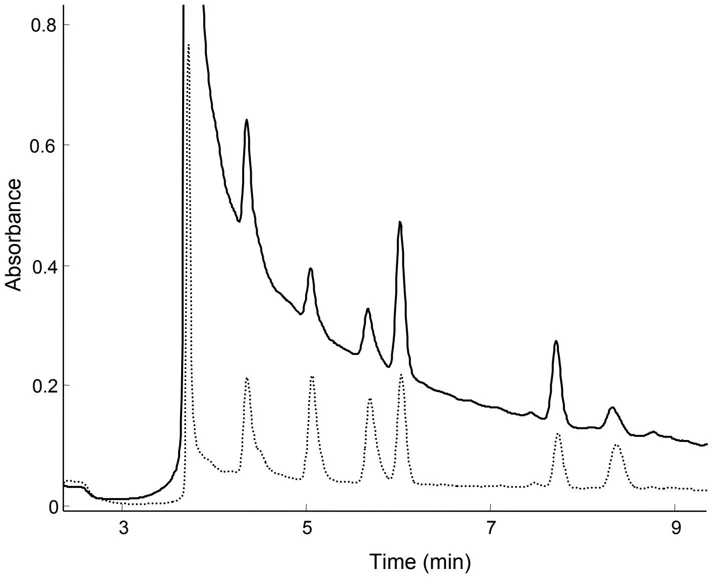


Fig. 5. Chromatographic profiles of the river-water measured at 240 nm before (–) and after (···) sodium sulfite was added.

The river-water sample was spiked at different concentration levels. Those samples were analyzed in the same conditions as the previous samples and the univariate models were constructed. For each spiked aliquot, three replicates were analyzed and the compounds studied were quantified. Table 4 shows the results.

Table 4. Mean predicted concentration (μg l^{-1}) and its R.S.D. found by univariate calibration in the water sample with added sodium sulfite

Analyte	Univariate calibration	
	Mean value	RSD (%)
Resorcinol	5.60	25.1
Oxamyl	1.09	13.2
Methomyl	1.10	8.1
Phenol	1.26	8.6

As we can see, the results are similar to those in Table 3. A two-sided *t*-test was used to compare the results obtained by the different methods with those obtained with univariate calibration. In all cases, the results were similar for a confidence interval of 95%. This validates the results obtained from the second-order calibration methods.

Hence, any of the three methods can be used, but the MCR–ALS has the advantage that the shift is not a problem as it is in GRAM and PARAFAC. As far as practical aspects of the algorithms are concerned, GRAM is faster and no initial estimations are needed, while MCR–ALS and PARAFAC are iterative and the time needed for completion depends on how similar the initial estimation and the final solution are.

3.2. Sewage plant water

Three calibration samples and three test samples were analyzed in accordance with the conditions in the Experimental section. Unlike the river-water samples, the addition of sodium sulfite had hardly any effect on the organic matter that produced the high band at the beginning of the chromatogram (Fig. 6). Hence, the peaks of the analytes could not be resolved chemically and the external validation could not be done as it was for the river-water sample.

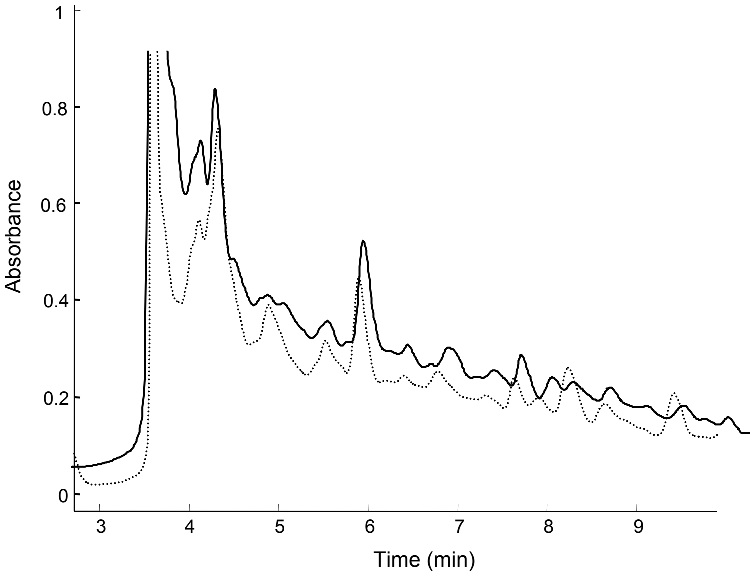


Fig. 6. Chromatographic profiles of the sewage treatment plant water measured at 240 nm before (–) and after (···) sodium sulfite was added.

The concentration of the analytes of interest could only be determined using second-order calibration. Table 5 shows the predictions in µg l^{-1}. For methomil, there was a large difference because when it eluted it was largely overlapped with other interferences. For the three methods, the predicted spectra were similar to the ones measured with standards, with correlation coefficients higher than 0.999. Unlike the river-water, where we recovered approximately the spiked amount, the predicted concentration for resorcinol was significantly larger than what we spiked. To check whether the analyte was present in that sample, we applied GRAM, PARAFAC and MCR–ALS using the non-spiked sample as a test sample. The predicted concentrations were 48.6, 47.3 and 44.4 µg l^{-1}, respectively. This agrees with the obtained value presented in Table 5, which corresponds to the concentration found in the non-spiked sample, plus the amount spiked (20 μg l^{-1}).

Table 5. Mean predicted concentration in sewage water (μg l^{-1}) and its R.S.D.

Analyte	GRAM		MCR-ALS		PARAFAC	
	Mean value	RSD (%)	Mean value	RSD (%)	Mean value	RSD (%)
Resorcinol	68.65	7.4	66.86	3.9	62.58	14.1
Oxamyl	5.21	1.7	5.28	7.8	4.96	2.0
Methomyl	6.87	49.2	6.61	35.1	6.39	13.7
Phenol	7.74	0.6	8.08	4.6	7.71	1.0

4. CONCLUSIONS

GRAM, PARAFAC and MCR–ALS were able to quantify overlapped and highly drifted chromatographic profiles. Such profiles can be found in the determination of compounds at very low concentrations (μg l^{-1}) in natural samples. With these methods it is not critical to assess where the peak starts or finishes. Of the three second-order calibration methods, GRAM is fast, and requires only two matrices and no initial estimations of the chromatographic profiles and the spectra of the analytes. Also the figures of merit can be easily calculated. On the other hand, PARAFAC and MCR–ALS are iterative. GRAM and PARAFAC require trilinear data, which is difficult to achieve in this kind of data because of the time shift in the chromatographic profiles.

ACKNOWLEDGEMENTS

The authors would like to thank the Spanish Ministry of Science and Technology (Project BQU 2003-01142) for economic support and the AGAUR, of the Catalan Government, for providing Enric Comas with a doctoral fellowship.

REFERENCES

[1] Council Directive 76/160/EEC concerning the quality of bathing water. Official Journal of the European Union, L 031, 05/02/1976, 1.

[2] Council Directive 80/778/EEC relating to the quality of water intended for human consumption, Official Journal of the European Union, L 229, 30/08/1980 11.

[3] D. Puig, D. Barceló, J. Chromatogr. A 733 (1996) 371.

[4] N. Masqué, E. Pocurull, R. M. Marcé, F. Borrull, Chromatographia 47 (1998) 176.

[5] N. Masqué, R.M. Marcé, F. Borrull, Chromatographia 48 (1998) 231.

[6] H. Miao, M. Yu, S. Hu, J. Chromatogr A 749 (1996) 5.

[7] X. Shao, Z. Chen, X. Lin, Chemom. Intell. Lab. Syst. 50 (2000) 91.

[8] S. Nakamura, J. Chromatogr A 859 (1999) 221.

[9] P. Nikitas, A. Pappa-Louisi, A. Papageorgiou, J Chromatogr A 912 (2001) 13.

[10] S. Jurt, M. Shär, V.R. Meyer, J. Chromatogr A. 929 (2001) 165.

[11] V.R. Meyer, Chromatographia 40 (1995) 15.

[12] V.R. Meyer, J. Chromatogr. Sci. 33 (1995) 26.

[13] E. Sánchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496.

[14] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149.

[15] R. Tauler, Chemom. Intell. Lab. Syst. 30 (1995) 133.

[16] N.M. Faber, R. Bro, P.K. Hopke, Chemom. Intell. Lab. Syst. 65 (2003) 119.

[17] R.Bro, Anal. Chim. Acta. 500 (2003) 185.

[18] K. S. Booksh, B. R. Kowalski, Anal. Chem. 66 (1994) A782.

[19] B.C. Mitchell, D.S. Burdick, Chemom. Intell. Lab. Syst. 20 (1993) 149.

[20] N. M. Faber, Anal. Bioanal. Chem. 372 (2002) 683.

[21] E. Pocurull, R. M. Marcé, F. Borrull, Chromatographia 41 (1995) 521.

[22] E. Comas, R.A. Gimeno, J.Ferré, R.M Marcé, F. Borrull, F.X. Rius, J.Chromatogr A. 988 (2003) 277.

[23] R.A. Gimeno, E. Comas, R.M. Marcé, J. Ferré, F.X. Rius, F. Borrull, Anal. Chim. Acta. 498 (2003) 47.

[24] N.M. Faber, R. Boqué, J. Ferré. Chemom. Intell. Lab. Syst. 55 (2001) 91.

[25] N.M. Faber, J. Ferré, R. Boqué, Chemom. Intell. Lab. Syst. 55 (2001) 67.

[26] R. Bro, Multi-way analysis in the food industry, Models, algorithms, and applications. Doctoral dissertation. University of Amsterdam (1998).

[27] A. de Juan, R. Tauler, Anal. Chim. Acta 500 (2003) 195.

[28] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218.

[29] E. Comas, R.A. Gimeno, J.Ferré, R.M Marcé, F. Borrull, F.X. Rius, Anal. Chim. Acta. 470 (2002) 163.

[30] R. Bro, H.A. L. Kiers, J. Chemom. 17 (2003) 274.

[31] http://www.models.kvl.dk/source/nwaytoolbox/download.asp. October 2003.

[32] http://www.ub.es/gesq/mcr/ndownload.htm. October 2003.

[33] Matlab, The Mathworks, South Natick, MA, USA.

[34] Z. Chen, Y. Liang, J. Jiang, Y. Li, H. Qian, R. Yu, J. Chemom. 13 (1999) 15.

[35] B.K Dable, K.S. Booksh, J. Chemom. 15 (2001) 591.

[36] Malinowski, E. R. *Factor Analysis in Chemistry*, 3rd Ed. John Wiley & Sons Inc. New York 2002.

Chapter 5

Conclusions

[redacted]

5.1 INTRODUCTION

This chapter contains the conclusions of the work presented in this thesis. Some suggestions are also indicated about future work on the application of GRAM and other second-order calibration methods to non-selective chromatographic data.

5.2 CONCLUSIONS

1. The Generalized Rank Annihilation Method can be used to identify and quantify analytes in non-selective HPLC-DAD data provided that the necessary "precautions" are taken into account (see below). When properly implemented, it may save time and resources as the analytes of interest do not need to be resolved. In addition, only one calibration sample is needed.

2. Applying GRAM requires that:

   a. The signal regarding the analyte of interest has to be trilinear, i.e., proportional between the different samples.
   b. The number of factors needed to build the model has to be properly selected.

Specific tools have been developed to ensure that each condition was fulfilled.

2.1. *Detection and correction of the retention time shift*

Retention time variability is common in HPLC-DAD data due to the lack of reproducibility of the separations. Small changes in the experimental conditions and in the manipulation of the sample may affect the reproducibility and introduce shift in the peak of the analyte of interest. The shift may be larger when an on-line preconcentration step is performed before the chromatographic analysis.

With the technique and instrument used in this thesis, and for the cases studied, the observed retention time shifts were not larger than 2 seconds. However, these variations were large enough to make the concentration predicted by GRAM incorrect, for example, leading to 30% of prediction error in the case of analyzing polycyclic aromatic sulfonates (PAHs). This prediction error is also influenced by the degree of overlap of the analyte of interest with the interferences.

We developed a method to correct the retention time shift, based on the application of the curve resolution method ITTFA. A time window in the test sample is selected so that the profile of the analyte of interest coincides with its profile in the calibration sample. The correction is done selectively for the analyte of interest. With this method, for example, the prediction errors were reduced from 30% to 2% in the analysis of PAHs.

2.2. *Determination of the number of factors used to calculate the GRAM model*

The wrong selection of the number of factors used to calculate the GRAM model may also lead to large prediction errors. These are higher when the selected number is lower than the optimal one (underfitting). When the number of factors is larger than the optimal one (overfitting), GRAM models noise and non-relevant information, but the predicted concentration is not as affected as in the underfitting case. This has been observed by other authors (see section 2.6).

We developed a graphical tool to determine the right number of factors: several GRAM models are tested by changing the number of factors and the weight parameter α. A large variation in the predicted concentration when α varies indicates underfitting or lack of trilinearity. A limitation of this criterion is that, for each model, the prediction that corresponds to the analyte of interest must be identified, which requires automatic algorithms that select which prediction is that of the analyte of interest.

2.3. *Check for trilinearity and outlier detection*

Two methods were studied to check if the data are trilinear *enough* to be used for calculating a GRAM model. In the first one we studied the relationship between α and the predicted concentration by GRAM, by testing different values of α. When the predicted concentration did not vary, the data were trilinear. We concluded that even when perfect trilinearity is not accomplished, useful results can still be obtained, and that the wrong selection of the number of factors produces the largest errors (see section 3.3).

A new outlier detection method for GRAM was also developed. Different from univariate and multivariate calibration, a sample that contains interferences not considered in the calibration step is not an outlier in GRAM. In GRAM, outlying samples are those where the analyte of interest in the test sample does not follow the same data structure as in the calibration sample.

With HPLC-DAD data the comparison of the measured and predicted spectra is not accurate enough for detecting such outliers. The reason is that the lack of trilinearity in chromatography is commonly present in the time domain but not in the spectral domain. Hence, although the elution profiles and the concentration may be wrong, the estimated spectra can still be correct. This means that such data can be used for qualitative analysis but not for quantitative anlaysis.

The developed criterion for detecting outliers is based on the projection of the measured matrices onto the Net Analyte Signal (NAS) space described by the GRAM estimations. The residuals of a fitted straight line indicate lack of trilinearity. To know whether the residuals are significative, an estimation of the noise is needed, which can be obtained either from a blank sample or from the projected peaks of the calibration and test sample onto the NAS space. The second option seems preferable (see section 3.4).

3. Applications of GRAM

GRAM was applied to the analysis of aromatic sulfonates through ion-pair liquid chromatography, and to the analysis of phenols and pesticides in water.

Both analytical challenges required a preconcentration step in order to detect the concentration levels found in the water samples. The separation conditions were optimized with standards. However, when the test samples were analyzed the analytes of interest eluted overlapped with interferences. The experimental chromatographic conditions were changed in order to isolate those components. The optimization must be done for each analyte in every sample, which in practice is almost impossible. By the application of GRAM all those problems were overcome.

Aromatic sulfonates. The application of GRAM was an efficient alternative to the tedious and time-consuming optimization of the ion-pair liquid chromatographic separation method followed by univariate calibration. GRAM was applied to the overlapped peaks, and after less than 8 minutes, the analytes of interest were determined. To validate the results, the experimental chromatographic conditions were changed in order to isolate those components, and the analysis lasted 45 minutes. In these conditions, we then applied univariate calibration. We found similar results when GRAM was applied to the former conditions. Using GRAM, in this case, we saved analysis time.

Phenolic compounds and pesticides. The analytes of interest eluted highly overlapped with a high band of humic and fulvic acids. We compared the GRAM predictions with two other second-order calibration methods, PARAFAC and MCR-ALS. No differences were found between the predictions of GRAM and the predictions of PARAFAC and MCR-ALS, in triliner data. However, GRAM seems preferable because GRAM is faster than PARAFAC and MCR-ALS. The speed of PARAFAC and MCR-ALS depends on the similarity of the initial estimations of the model results.

4. Real evaluation of GRAM

For the cases above, the separation time was reduced as no selective measurements were needed. Although this makes this step of the analysis faster, applying GRAM actually increased the total time of analysis. The reason is that important input is required from a trained analyst in most of the steps of the data analysis: (a) exporting the data from the chromatograph software to the software where GRAM is implemented (in this thesis, Matlab), (b) locating and selecting the peak of interest both in the calibration sample and in the test sample and creating the necessary data matrices for these peaks ($\mathbf{R}_c$ and $\mathbf{R}_t$), (c) determining the total number of contributions to both peaks (d) identifying which of the solutions of GRAM corresponds to the analyte of interest through spectral comparison (which requires having the spectrum of the analyte of interest available in Matlab environment), and (e) checking the possibility of the test sample being an outlier. These tasks are not as yet automatized. This makes the total time of analysis larger than the possible time required through optimizing the conditions and using univariate calibration. However, we have shown the possibility of using GRAM to solve those analytical problems and the implementation of these methods in the software of the chromatograph is just a technical aspect that can easily be overcome as interest in these methods increases.

Validation of the analytical method is another aspect that must be taken into account when GRAM is used for routine analysis. Figures of merit such as limit of detection, sensitivity, selectivity, linear range, etc., are calculated as a way of characterizing the performance of the analytical methods based on univariate and multivariate calibration. Although similar expressions are also available for calculating such figures of merit when GRAM is used for quantitation, their practical calculation and interpretation still needs further study.

5.3 SUGGESTIONS FOR FUTURE RESEARCH

This thesis focuses on the application of GRAM to nonselective HPLC-DAD data. The work presented here points towards a number of areas that still need to be studied in more detail. One area is related to the application of second-order methods to non-selective chromatographic data. Another is related to the application of those methods to other techniques and analytical situations.

1. Application of GRAM and other second-order calibration methods to non-selective chromatographic data.

   a) The calculation of the uncertainty of the GRAM predictions has been developed. The GRAM estimations are slightly biased. Faber et al. [Chem. Intell. Lab. Syst. 55 (2001) 67-90] showed the calculation and correction of that bias. As far as we know, those equations have not been tested in measured data yet. The difficulty is that an estimation of the standard deviation of the noise is needed. The net noise calculated from blank samples or from the NAS matrices as developed in section 3.4 could be possible candidates for these equations.

b) Develop user-friendly software that automates most of the aspects mentioned in point 4 (page 210) and increases the speed of the analysis based on GRAM.

2. Application of GRAM to other techniques

a) The application of GRAM and other second-order calibration methods to other techniques, such as chromatography with Mass Spectrometry (MS) detection or Capillary Electrophoresis (CE) with DAD detection has yet to be further investigated. Some aspects mentioned before, such as exporting/importing data for MS detector need to be studied.

b) The application of second-order calibration methods to CE with DAD detection seems to be the next step, because the recorded signal is of a similar kind to the one measured by HPLC-DAD. Although CE instruments provide much more irreproducible data than an HPLC instrument, CE has some advantages over HPLC-DAD, such as power of resolution and reduction of the time of analysis, which makes the research into the application of second-order methods to that technique worthwhile.

c) The application of second-order calibration algorithms to dynamic processes in which the signal (e.g. a spectrum) is measured within a process over time. Second-order data are obtained, and by the application of second-order calibration algorithms, the composition, the evolution, the kinetics and the mechanisms of the reactions can be determined on-line, without the use of off-line measurements to characterize the process and the chemical reactions.

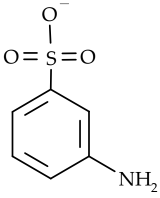
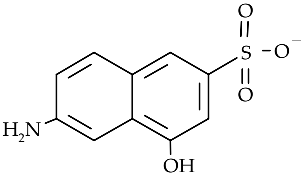
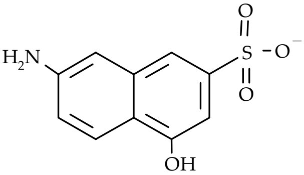
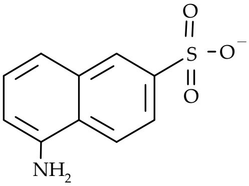
Appendix

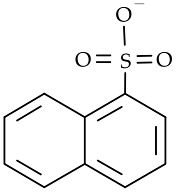
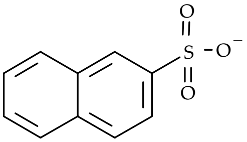
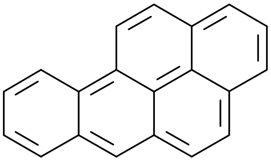
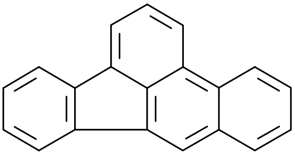
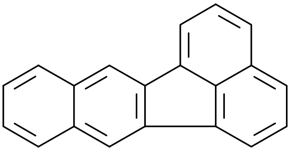
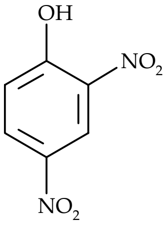
[redacted]

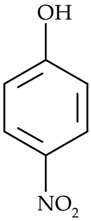
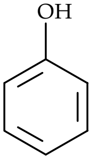
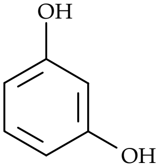
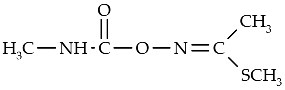
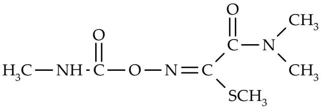
INTRODUCTION

This chapter contains three appendixes: the molecular structure of the compounds studied in this thesis, the abbreviations used, and the list of papers and meeting presentations given by the author during the thesis.

STUDIED ANALYTES

Name	Family	Structure
3-aminobenzenesulfonate	Benzenesulfonates	
6-amino-4-hydroxy-2-naphthalenesulfonate	Naphthalenesulfonates	
6-amino-1-hydroxy-3-naphthalenesulfonate	Naphthalenesulfonates	
1-amino-6-naphthalenesulfonate	Naphthalenesulfonates	

Name	Family	Structure
1-naphthalenesulfonate	Naphthalenesulfonates	
2-naphthalenesulfonate	Naphthalenesulfonates	
Benzo[a]pyrene	Polycyclic aromatic hydrocarbons	
Benzo[b]fluoranthene	Polycyclic aromatic hydrocarbons	
Benzo[k]fluoranthene	Polycyclic aromatic hydrocarbons	
2,4-dinitrophenol	Phenols	

Name	**Family**	**Structure**
4-nitrophenol	Phenols	
Phenol	Phenols	
Resorcinol	Phenols	
Methomyl	Pesticides	
Oxamyl	Pesticides	

ABBREVIATIONS

Abbreviations used in the thesis

ALS	Alternating Least Squares
BZS	Benzenesulfonate
CE	Capillary Electrophoresis
CR	Curve Resolution
DAD	Diode Array Detector
DECRA	Direct Exponential Curve Resolution Algorithm
EEM	Excitation Emission Matrix - Fluorescence
EFA	Evolving Factor Analysis
EU	European Union
FA	Factor Analysis
FSWEFA	Fixed Size Window Evolving Factor Analysis
GC	Gas Chromatography
GRAM	Generalized Rank Annihilation Method
HELP	Heuristic Evolving Latent Projection
HPLC	High Performance Liquid Chromatography
ITTFA	Iterative Target Transformation Factor Analysis
MCR	Multivariate Curve Resolution
MIP's	Molecular Imprinted Polymers
MS	Mass Spectrometry
NS	Naphthalenesulfonate
OPA	Orthogonal Projection Approach
PAHs	Polycyclic Aromatic Hydrocarbons
PARAFAC	Parallel Factor Analysis
PCA	Principal Components Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
RAFA	Rank Annihilation Factor Analysis
RSD	Relative Standard Deviation
SIMPLISMA	Simple-to-use interactive self-modeling mixture analysis

SMCR	Self-Modeling Curve Resolution
SPE	Solid Phase Extraction
SVD	Singular Value Decomposition
TFA	Target Factor Analysis
TLD	Trilinear Decomposition
TW	Time Window
UV-Vis	Ultraviolet visible

LIST OF PAPERS AND MEETING CONTRIBUTIONS

List of papers by the author presented in this thesis (in chronological order):

1. E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius.
   Time shift correction in second-order liquid chromatographic data with iterative target transformation factor analysis.
   Analytica Chimica Acta 470 (2002) 163 – 173.
   (Chapter 3)

2. E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius.
   Using second-order calibration to identify and quantify aromatic sulfonates in water by high-performance liquid chromatography in the presence of coeluting interferences.
   Journal of Chromatography A 988 (2003) 277 – 284.
   (Chapter 4)

3. E. Comas, J. Ferré, F.X. Rius.
   Graphical criterion for assessing trilinearity and selecting the optimal number of factors in the generalized rank annihilation method using liquid chromatography-diode array detection data.
   Analytica Chimica Acta 515 (2004) 23 – 30.
   (Chapter 3)

4. E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius.
   Quantification from highly drifted and overlapped chromatographic peaks using second-order calibration methods.
   Journal of Chromatography A 1035 (2004) 195 – 202.
   (Chapter 4)

5. E. Comas, J. Ferré, F.X. Rius.
   Outlier detection in the Generalized Rank Annihilation Method applied to chromatographic data.

Analytical Chemistry, submitted.
(Chapter 3)

6. J. Ferré, N.M. Faber, E. Comas, F.X. Rius.
   GRAM, a tutorial.
   Journal of Chromatography A, to be submitted, special issue 'Chemometrics in Chromatography'.
   (Chapter 2)

7. E. Comas, J. Ferré, F. X. Rius.
   Net noise estimation in a second order chromatographic peak.
   Paper in preparation.
   (Chapter 3)

The following papers have been omitted since their content is not related to the scope of the thesis. However, working on different projects has been very useful to broaden my knowledge of second-order data.

8. R.A. Gimeno, E. Comas, R.M. Marcé, J. Ferré, F.X. Rius, F. Borrull.
   Second-order calibration for determining polycyclic aromatic compounds in marine sediments by solvent extraction and liquid chromatography with diode array detection.
   Analytica Chimica Acta 498 (2003) 47 – 53.

9. B. Ma, P.J. Gemperline, E. Cash, M. Bosserman, E. Comas.
   Characterizing Batch Reactions with in-situ Spectroscopic Measurements, Calorimetry, and Dynamic Modeling.
   Journal of Chemometrics 17 (2003) 470 – 479.

Contributions to international meetings, directly related with the thesis:

1. E. Comas, J. Ferré, F.X. Rius.
   2D window selection in second order chromatographic data
   V Colloquium Chemometricum Mediterraneum – Ustica (Italy).
   Oral communication.

2. E. Comas, J. Ferré, F.X. Rius.
   A graphical criterion for assessing trilinearity and selecting the optimal number of factors in GRAM using HPLC-DAD data.
   V Colloquium Chemometricum Mediterraneum – Ustica (Italy).
   Poster communication.

Other meeting contributions not directly related with the thesis:

3. J. Ferré, E. Comas, F.X. Rius.
   Two-dimensional wavelet transform applied to second-order calibration.
   7th Chemometrics in Analytical Chemistry. CAC'00. Antwerp (Belgium).
   Poster communication.

4. P.J. Gemperline, B. Ma, E. Cash, S.D. Moore, E. Comas.
   Characterizing Batch Reactions with in-situ Spectroscopic Measurements, Calorimetry, and Dynamic Modeling.
   3rd International Chemometrics Research Meeting (ICRM 2002), Veldhoven (The Netherlands).
   Plenary lecture of the first author.

5. P.J. Gemperline, S.D. Moore, E. Comas, R.R. Reinhart, K. High, S. Alam.
   Fusing Data from Diverse Sources to Characterize Batch Reactions
   8th Chemometrics in Analytical Chemistry. CAC'02. Seattle, WA (United States).
   Poster communication.