Talanta 83 (2011) 1147–1157

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Outlier detection for the Generalized Rank Annihilation Method in HPLC-DAD analysis

Joan Ferré [a,*], Enric Comas [b]

[a] Department of Analytical Chemistry and Organic Chemistry, Universitat Rovira i Virgili, Marcel·lí Domingo s/n, Campus Sescelades, 43007 Tarragona, Spain
[b] The Dow Chemical Company, Tarragona, Spain

ARTICLE INFO

Article history:
Available online 22 August 2010

Keywords:
GRAM
HPLC-DAD
Outlier
Second-order
Trilinearity

ABSTRACT

The Generalized Rank Annihilation Method (GRAM) is a second-order calibration method that is used in chromatography to quantify analytes that coelute with interferences. For a correct quantification, the peak of the analyte in the standard and in the test sample must be aligned and have the same shape (i.e., have a trilinear structure). Variations in retention time and shape between the two peaks may cause the test sample to behave as an outlier and produce an incorrect prediction. This situation cannot be detected by checking the coincidence of the recovered spectrum with the known spectrum of the analyte because the spectral domain is not affected. It cannot be detected either by checking if the recovered profile is correct (i.e., unimodal and positive). Several plots are presented to detect such outliers. The first plot compares the particular elution profiles in the standard and in the test sample that are recovered by least-squares regression from the spectra estimated by GRAM. The calculated elution profiles from both peaks should coincide. A second plot uses the elution profiles and spectra calculated by GRAM to define the vector space spanned by the interferences. The measured peaks in the standard and in the test sample are projected onto the space that is orthogonal to the space spanned by the interferences. These projections are proportional (up to the noise) if data are trilinear. The proportionality is checked graphically from the first singular vector of the projected peaks, or from the plot of the orthogonal signal versus the net sensitivity. The use of these graphs is shown for simulated data and for the determination of 4-nitrophenol in river water samples with liquid chromatography/UV–Vis detection.

1. Introduction

The Generalized Rank Annihilation Method (GRAM) is an algorithm for the qualitative and quantitative analysis of second-order bilinear data. An exhaustive bibliographic revision of GRAM and applications to different analytical techniques can be found in Ref. [1]. In chromatography GRAM is used to quantify analytes that coelute with unexpected interferences in complex samples and also in some high speed GC × GC analyses [2] and parallel column liquid chromatography analyses [3] in which full resolution is sacrificed to obtain other benefits, such as less time of analysis. GRAM only requires one calibration standard and does not make any assumptions with respect to the shape of the underlying profiles in the overlapped chromatographic peaks. For each component that is in the peak of the test sample and in the peak of the calibration standard, GRAM gives its elution profile, its spectrum and the relative concentration test/calibration. This ability to quantitate an analyte in the presence of interferences that are in the test sample but not in the calibration standard is known as the second-order advantage, and is obtained with bilinear second-order data. In chromatography, these data are obtained by recording a multivariate signal over time such as an UV–Vis spectrum (as it is done in HPLC-DAD) or a mass spectrum (as it is done in GC–MS). The peak of interest is then a two-dimensional data array, where the first dimension is time and the second is detection channel. These type of data can also be obtained by recording another elution profile, as it is done in GC × GC, but for simplicity we will focus on data from an HPLC-DAD system.

The results of GRAM for chromatographic data (i.e., predicted spectra, profiles and concentrations) may be degraded by the instrumental noise (specially heteroscedastic noise), the similarity of elution profiles (low chromatographic resolution between the analyte and the interferences), the similarity of spectra (low spectral selectivity) and the similarity of the concentration ratios in the standard and the test sample. In addition, predictions are affected by the wrong selection of the number of systematic contributions in the data and by data inconsistencies such as run-to-run time shift and peak broadening. Small non-linearities in the responses will also increase error if the concentration of the analyte in the test sample differs greatly from the concentration in the standard [4].

* Corresponding author. Tel.: +34 977 559564; fax: +34 977 558446.
E-mail address: joan.ferre@urv.cat (J. Ferré).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.08.007

And finally, although GRAM can theoretically handle any number of overlapped constituents, in reality the predictions are degraded as more interferences must be modeled due to the reduction of the net signal-to-noise ratio and the increase in analyte interactions and other potential sources of non-linearity. Several studies have shown how these factors influence de GRAM results. [2,5–7].

The mentioned problems are not unique of GRAM. Most first- and second-order calibration methods also suffer from them. Fortunately, some problems can be prevented. For example, matrix effects can be reduced with standard addition [3,7–12], and time shift can be corrected with rank alignment algorithms [13]. Others are minimized by the fact that GRAM is based on latent variables, which increases the robustness of the results to slight deviations from the ideal behavior. However, the list of causes that may degrade the prediction is still large and the analyst needs to be confident that GRAM can give correct predictions before it is used in daily analyses.

Like in univariate and multivariate calibration, confidence in the GRAM predictions is gained through validation and outlier diagnostics. Validation checks that the defined analytical procedure (here based on GRAM) can correctly analyze future samples of similar matrix. Outlier diagnostics warn the analyst whether the new data deviates from what was validated and that the predictions are not reliable.

The use of GRAM is often validated by comparing its predictions with the result of analyzing representative samples with an alternative, more time-consuming, method of analysis [2,10,14,15] or by recovery essays with spiked samples [16]. However, like in other types of calibration, validation alone may not provide sufficient confidence in the predictions of future samples. Despite the method has been validated, a test sample may behave different than expected. The sample may belong to a different population than the validation samples, have extreme values or suffer from some gross error of the analytical procedure that leads to a misleading instrumental response. On top of that, GRAM calculates a model for each test sample so previously successful models do not guarantee that the new model calculated for the next test sample will be successful. Each new sample requires the number of factors in the GRAM model to be decided, and preprocessing algorithms (e.g., for retention time correction) be applied with the settings that are adequate for that sample. This is different from univariate calibration and multivariate calibration, where the model is validated and used for predicting a large number of samples until model updating or model transfer is needed. Hence, outlier diagnostics are needed to flag if the validated method with GRAM can fail with the new sample.

Outliers in GRAM are easily detected when the input measurements are unrealistic, when the predictions are clearly erroneous out of the expected range (e.g., negative concentrations) and when the calculated profiles lack physical meaning (i.e., two maxima in one elution profile). The special concern is for those outliers whose predictions are reasonable (i.e., in the expected range of values) but biased. It also important to note one important difference between outliers in GRAM and outliers in multivariate calibration. In multivariate calibration, a sample is an outlier if interferences that were not included in the standards contribute to the instrumental response. In GRAM such a sample is not an outlier because the model is calculated with the test sample and the standard together. The signal of the interferences is modeled and quantification is possible as long as the selectivity is sufficient in both orders (i.e., the elution profile and the spectrum of the analyte are different enough from the profiles and spectra of the interferences).

The main reason for a sample be an outlier in GRAM is to deviate from the trilinearity requirement described in Eqs. (1) and (2) in Section 2. Trilinearity involves that: (i) the measured peak can be bilinearly decomposed as a sum of contributions of the different analytes and (ii) the elution profile and the spectrum of the analyte of interest are the same in the standard and in test sample, except for a scaling factor related to the concentration. Causes of deviations are irreproducibility of the profiles on both dimensions, deviations from linearity due to matrix effects, non-rank additivity [17] and non-bilinearity [17]. The latter two are not usually encountered for second-order instruments that are known to yield bilinear data, such as HPLC-DAD. Matrix effects can be handled by standard addition. Hence, the most common signal inconsistencies are reduced to variations in the shape of the profiles on both dimensions. When the second dimension corresponds to spectroscopic measurements, which are relatively highly repeatable, the main signal inconsistencies reduce to time shift and peak shape variation.

The simplest and most used outlier detection diagnostic in GRAM is to verify that the predicted concentrations and estimated elution profiles and spectra are physically realistic [2,3,7,15,18]. The predicted concentrations should be in the validated range of concentrations. The estimated elution profiles should be unimodal and non-negative over the considered time window. The estimated spectra should be non-negative and like the spectra obtained from measuring the pure analytes. The degree of coincidence can be checked, for example, with the correlation coefficient [2]. When these requirements are met, the confidence that the GRAM predictions are valid is high [2,18]. However, these comparisons are not always sufficient. Apparently correct elution profiles and spectra can be obtained even when the data are not trilinear and the prediction errors are large. Other outlier diagnostics check whether the peaks follow the trilinear model in Eqs. (1) and (2). A first measure of lack of trilinearity is given by the difference between the measured peak and the predicted peak. Large systematic differences indicate bad model fit either because the number of factors in GRAM is underestimated or because the data lacked trilinearity. Random differences between the two matrices are considered to indicate accuracy of the concentration estimate [19]. However, these differences evaluate the fit of the whole chromatographic peak and not the specific analyte we are quantifying. Hence, residuals may be non-random but predictions for the analyte of interest be accurate. A related option is to project one peak onto the space spanned by the rows and columns of the other peak [20]. The projection should recover the projected peak within the noise. This method 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 [21]. Their rank is the same if the data are trilinear. Evaluating a significant increase in rank is sometimes difficult because small non-linearity is distributed over the relevant eigenvectors/eigenvalues. 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 [22]. This criterion is only partially related to the quality of the predictions and is one more to add to the diagnostics pool.

This paper presents another graphical criterion for detecting outliers in GRAM for second-order HPLC-DAD data. It is specially suited when the measurements along the second axis are satisfactorily accurate and most signal inconsistencies (responsible for outliers) are due to shifts along the time axis.

2. Theory

2.1. Notation

Boldface uppercase letters represent matrices, boldface lowercase letters indicate column vectors and italic letters indicate scalars. Superscripts 'T', '−1' and '+' indicate transposition, inverse

and Moore–Penrose pseudoinverse, respectively. Column vectorisation of a matrix is indicated by 'vec'. A 'hat', e.g., $\hat{\mathbf{H}}$, indicates reconstructed/predicted data when it is needed to differentiate them from the measured or underlying data. The analyte of interest is designated as 'analyte k'. $\mathbf{I}$ is the identity matrix of appropriate size.

2.2. Model assumptions

In HPLC-DAD, the chromatogram of an analyzed sample is a time × wavelength matrix of responses obtained by measuring spectra over time. From that chromatogram, the data matrix $\mathbf{R}_t$ ($J_1 \times J_2$) of the peak of interest is extracted. This peak contains the analyte of interest to be quantified plus unknown interferences that coeluted with the analyte. GRAM uses as a calibration standard a single data matrix $\mathbf{R}_c$ (of the same size as $\mathbf{R}_t$) with a known concentration of the analyte of interest (c_c). $\mathbf{R}_c$ can be obtained either by analyzing a pure standard, or by analyzing an aliquot of the test sample after adding a known amount of the analyte (standard addition).

In GRAM it is assumed that both peaks, $\mathbf{R}_c$ and $\mathbf{R}_t$, are the sum of responses of each constituent, and that the response of each constituent can be factored out as the product of a column ($\mathbf{x}$) and row ($\mathbf{y}$) profile. It is also assumed that the constituents in the sample do not interact (i.e., the profiles of one constituent are independent on the presence of the other constituents). These assumptions can be written as:

$$\mathbf{R}_c = \sum_{k=1}^{K} c_{c,k}\mathbf{x}_k\mathbf{y}_k^T = \mathbf{X}\mathbf{C}_c\mathbf{Y}^T \quad (1)$$

$$\mathbf{R}_t = \sum_{k=1}^{K} \mathbf{c}_{t,k}\mathbf{x}_k\mathbf{y}_k^T = \mathbf{X}\mathbf{C}_t\mathbf{Y}^T \quad (2)$$

For convenience, the sum is over the total number of constituents K present both in $\mathbf{R}_c$ and $\mathbf{R}_t$. If a particular constituent is not present in $\mathbf{R}_c$ (or in $\mathbf{R}_t$), its corresponding concentration $c_{c,k}$ (or $c_{t,k}$) is zero. In matrix notation, $\mathbf{X}$ ($J_1 \times K$) and $\mathbf{Y}$ ($J_2 \times K$) contain, in columns, the profiles $\mathbf{x}$ and $\mathbf{y}$ of the K constituents and the diagonal matrices $\mathbf{C}_c$ ($K \times K$) and $\mathbf{C}_t$ ($K \times K$) contain the concentrations. With this convention, for example, if $\mathbf{R}_t$ contains the same constituents as $\mathbf{R}_c$ plus one interferent, all the diagonal elements of $\mathbf{C}_t$ are nonzero and the element of $\mathbf{C}_c$ corresponding to that interferent is zero. Other systematic sources of variation, such as a baseline, can also be introduced in $\mathbf{X}$ and $\mathbf{Y}$ as analytes.

2.3. The GRAM algorithm

Different GRAM algorithms have been proposed. The one used here is based on Sánchez and Kowalski [20] and Faber et al. [23]. The steps in GRAM are:

(1) Calculation of the sum matrix $\mathbf{R}$

$$\mathbf{R} = \alpha\mathbf{R}_c + \mathbf{R}_t \quad (3)$$

where α is a weight parameter that can be optimized to reduce prediction bias. By default, $\alpha = 1$ will be used here. This sum matrix gathers in one place all the profiles so that step 2 below can model all those contributions. Note that $\mathbf{R} = \alpha\mathbf{X}\mathbf{C}_c\mathbf{Y}^T + \mathbf{X}\mathbf{C}_t\mathbf{Y}^T = \mathbf{X}(\alpha\mathbf{C}_c + \mathbf{C}_t)\mathbf{Y}^T = \mathbf{H}\mathbf{Y}^T$ where the columns of $\mathbf{H}$ are the elution profiles in $\mathbf{X}$ multiplied by the concentration.

(2) Singular value decomposition (SVD) of $\mathbf{R}$:

$$\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{V}^T + \mathbf{E} \quad (4)$$

where the matrices of singular vectors ($\mathbf{U}$,$\mathbf{V}$) and singular values ($\mathbf{S}$) have been truncated for F relevant factors and $\mathbf{E}$ is the matrix of residuals. Ideally, F should be equal to the number of systematic variations in $\mathbf{R}$, i.e., equal to the total number of analytes K.

(3) Solution of the eigenvalue problem:

$$(\mathbf{S}^{-1}\mathbf{U}^T\mathbf{R}_t\mathbf{V})\mathbf{T} = \mathbf{T}\boldsymbol{\Phi} \quad (5)$$

where $\boldsymbol{\Phi}$ ($F \times F$) is a diagonal matrix of eigenvalues and $\mathbf{T}$ is the matrix of eigenvectors.

(4) Recovery of the elution profiles $\hat{\mathbf{H}}$ ($J_1 \times F$) and the spectral profiles $\hat{\mathbf{Y}}$ ($J_2 \times F$):

$$\hat{\mathbf{H}} = \mathbf{U}\mathbf{S}\mathbf{T} \quad (6)$$

$$\hat{\mathbf{Y}} = \mathbf{V}(\mathbf{T}^{-1})^T \quad (7)$$

Note that the columns in $\hat{\mathbf{H}}$ and $\hat{\mathbf{Y}}$ are not necessarily in the same order as in Eqs. (1) and (2). Also, the scale is undetermined. Usually, the spectral profiles are normalized and the scaling constant is introduced in $\hat{\mathbf{H}}$.

(5) The eigenvalue in $\hat{\boldsymbol{\Phi}}$ that corresponds to the analyte k, $\hat{\Phi}_k$, is the ratio between the concentration of analyte k in $\mathbf{R}_t$ and the concentration of that analyte in $\mathbf{R}$, $\hat{\Phi}_k = c_{t,k}/c_{t,k} + \alpha c_{c,k}$. By inserting the known concentration of analyte k in the standard, the predicted concentration of analyte k is:

$$\hat{c}_{t,k} = \frac{\alpha c_{c,k}\hat{\Phi}_k}{1 - \hat{\Phi}_k} \quad (8)$$

Note that the eigenvalue $\hat{\Phi}_k$ that corresponds to the analyte k is not necessarily the kth diagonal element of $\boldsymbol{\Phi}$. In order to know what eigenvalue corresponds to the analyte, the columns of $\hat{\mathbf{Y}}$ must be compared with the spectrum of analyte k measured from a pure standard.

2.4. Outlier detection

Outlier detection is based on testing the trilinearity of the measured peaks. Two methods are presented here.

The first method is based on recovering the individual elution profiles in $\mathbf{R}_c$ and $\mathbf{R}_t$. The profiles in $\hat{\mathbf{H}}$ (Eq. (6)) are a compromise fit of the true profiles in $\mathbf{R}_c$ and $\mathbf{R}_t$ and do not inform about the particular profiles in each matrix. Since the spectral domain is not affected by deviations in the time domain, $\hat{\mathbf{Y}}$ (Eq. (7)) is a good estimation of the spectra in both measured peaks and can be used to obtain the particular elution profiles by solving Eqs. (1) and (2):

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

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

If $\mathbf{R}_c$ and $\mathbf{R}_t$ are trilinear, the columns in $\hat{\mathbf{H}}$, $\hat{\mathbf{H}}_c$ and $\hat{\mathbf{H}}_t$ are all multiple (the concentration) of the same underlying profiles $\mathbf{X}$. This can be seen by plotting the normalized columns of $\hat{\mathbf{H}}$, $\hat{\mathbf{H}}_c$ and $\hat{\mathbf{H}}_t$. If they coincide this will increase the reliability of the predicted concentrations. If they do not coincide, the test sample behaves as an outlier because of, for example, elution time shift or peak broadening. Model underfitting can also be detected with this plot.

The second method is a variation of the previous one, but focuses on how the prediction is calculated. It is based on the idea of the net analyte signal (NAS), [24] that is the part of the measured response that is used for prediction. The NAS of analyte k can be 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 $\hat{\mathbf{R}}_c$ and $\hat{\mathbf{R}}_t$ are the peaks recovered from the calculated profiles:

$$\hat{\mathbf{R}}_c = \hat{\mathbf{H}}\hat{\mathbf{\Pi}}\hat{\mathbf{Y}}^T \quad (13)$$

$$\hat{\mathbf{R}}_t = \hat{\mathbf{H}}\hat{\mathbf{\Phi}}\hat{\mathbf{Y}}^T \quad (14)$$

with $\hat{\mathbf{\Phi}} + \alpha\hat{\mathbf{\Pi}} = \mathbf{I}$ and $\mathbf{P}_H$ and $\mathbf{P}_Y$ are the projection matrices defined as

$$\mathbf{P}_H = \mathbf{I} - \hat{\mathbf{H}}_{-k}\hat{\mathbf{H}}^+_{-k} \quad (15)$$

$$\mathbf{P}_Y = \mathbf{I} - \hat{\mathbf{Y}}_{-k}\hat{\mathbf{Y}}^+_{-k} \quad (16)$$

where $\hat{\mathbf{H}}_{-k}$ and $\hat{\mathbf{Y}}_{-k}$ are $\hat{\mathbf{H}}$ and $\hat{\mathbf{Y}}$ without the column of analyte k. $\mathbf{P}_H$ projects the columns of the peak onto the subspace that is orthogonal to the subspace spanned by $\hat{\mathbf{H}}_{-k}$ (the profiles of the interferences). Hence, at every wavelength, only the part of the elution profile of analyte k that cannot be described as a linear combination of the profiles in $\hat{\mathbf{H}}_{-k}$ remains. This signal is unique for the analyte and can be used for quantification. Similarly, the columns of $\hat{\mathbf{Y}}_{-k}$ span the subspace of the spectra of the interferences. $\mathbf{P}_Y$ projects the rows of the peak onto the subspace that is orthogonal to the subspace spanned by $\hat{\mathbf{Y}}_{-k}$. At every retention time, only the part of the spectrum of analyte k that is not a linear combination of the spectra in $\hat{\mathbf{Y}}_{-k}$ remains. Note that $\mathbf{P}_H$ and $\mathbf{P}_Y$ are also valid if $\hat{\mathbf{H}}_{-k}$ and $\hat{\mathbf{Y}}_{-k}$ are not the pure profiles of the interferences but a linear combination of them because they span the same vectorial subspace.

$\hat{\mathbf{R}}^*_c$ and $\hat{\mathbf{R}}^*_t$ have rank 1 and are proportional

$$\hat{\mathbf{R}}^*_t = \frac{\hat{c}_t}{c_c}\hat{\mathbf{R}}^*_c \quad (17)$$

so that the predicted concentration $\hat{c}_t$ can be found by inserting the known concentration of the standard c_c in the previous equation. The NAS cannot be used to detect outliers because $\hat{\mathbf{R}}^*_c$ and $\hat{\mathbf{R}}^*_t$ are calculated from $\hat{\mathbf{H}}$ and $\hat{\mathbf{Y}}$ only and any deviations from trilinearity are already embedded in $\hat{\mathbf{H}}$ and $\hat{\mathbf{Y}}$. However, replacing in Eqs. (11) and (12) the fitted $\hat{\mathbf{R}}_c$ and $\hat{\mathbf{R}}_t$ by the measured peaks:

$$\mathbf{R}^*_c = \mathbf{P}_H\mathbf{R}_c\mathbf{P}_Y \quad (18)$$

$$\mathbf{R}^*_t = \mathbf{P}_H\mathbf{R}_t\mathbf{P}_Y \quad (19)$$

gives the part of the measured peaks that is orthogonal to the space spanned by the interferences. This part includes the NAS plus the projection of the error:

$$\mathbf{R}^*_c = \mathbf{P}_H(\hat{\mathbf{R}}_c + \mathbf{E}_c)\mathbf{P}_Y = \hat{\mathbf{R}}^*_c + \mathbf{E}^*_c \quad (20)$$

$$\mathbf{R}^*_t = \mathbf{P}_H(\hat{\mathbf{R}}_t + \mathbf{E}_t)\mathbf{P}_Y = \hat{\mathbf{R}}^*_t + \hat{\mathbf{E}}^*_t \quad (21)$$

If $\mathbf{R}_c$ and $\mathbf{R}_t$ are trilinear and the sufficient number of factors has been selected, $\mathbf{E}^*_c$ and $\mathbf{E}^*_t$ will be random and small, and $\mathbf{R}^*_t$ and $\mathbf{R}^*_c$ will be almost proportional because the NAS signal will dominate in Eqs. (20) and (21). If $\mathbf{R}_c$ and $\mathbf{R}_t$ are not trilinear, $\mathbf{E}^*_c$ and $\mathbf{E}^*_t$ will be larger and different, and the difference between $\mathbf{R}^*_t$ and $\mathbf{R}^*_c$ will be systematic. This difference can be easily observed from the SVD of $\mathbf{R}^*_t$ and $\mathbf{R}^*_c$.

$$\mathbf{R}^*_c = \mathbf{U}_c\mathbf{S}_c{\mathbf{V}_c}^T + \mathbf{E}_c \quad (22)$$

$$\mathbf{R}^*_t = \mathbf{U}_t\mathbf{S}_t{\mathbf{V}_t}^T + \mathbf{E}_t \quad (23)$$

By defining $\mathbf{u}_t$, $\mathbf{u}_c$, $\mathbf{v}_t$, $\mathbf{v}_c$ as the first column of $\mathbf{U}_t$, $\mathbf{U}_c$, $\mathbf{V}_t$ and $\mathbf{V}_c$, respectively, if data are trilinear, $\mathbf{u}_t$ and $\mathbf{u}_c$ should coincide, and $\mathbf{v}_t$ and $\mathbf{v}_c$ should coincide. This gives confidence to the predicted concentrations. Otherwise, the test sample peak $\mathbf{R}_t$ can be flagged as an outlier.

The use of these diagnostics is commented below for two cases: (a) when data are trilinear, (b) when the lack of trilinearity is due to time shift.

3. Experimental

3.1. Simulations

Eqs. (1) and (2) were used to simulate a pure chromatographic peak for the standard ($\mathbf{R}_c$) and an overlapped peak for the test sample ($\mathbf{R}_t$). The elution profiles ($\mathbf{X}$) and normalized spectra ($\mathbf{Y}$) are shown in Fig. 1. $\mathbf{R}_c$ was simulated with diag($\mathbf{C}_c$) = [1.5 0 0] where diag means the elements in the diagonal. This is a pure peak with the analyte at concentration 1.5. $\mathbf{R}_t$ was simulated with diag($\mathbf{C}_t$) = [1 1 0.5]. This peak contains the analyte at concentration 1, an overlapped interference and a small baseline drift. Note that $\mathbf{R}_c$ and $\mathbf{R}_t$ fulfill the trilinearity requirement because $\mathbf{X}$ and $\mathbf{Y}$ are the same. To simulate retention time shift, a new $\mathbf{R}_t$ was calculated with the same $\mathbf{Y}$ and $\mathbf{C}_t$ as the previous $\mathbf{R}_t$ but with profiles in $\mathbf{X}$ centered at three time steps earlier. Considering that in the chromatograms described in the measured data section, a spectrum was recorded every 0.4 s, three time steps correspond to 1.2 s. Such time shifts, and even larger, are possible in routine measurements especially for compounds with large retention times. White noise (0.1% in relation to the maximum of the peak) was finally added to every simulated matrix. Fig. 2 shows the trilinear peaks of the calibration standard $\mathbf{R}_c$ and of the test sample $\mathbf{R}_t$. The shifted peak (not shown) looks like $\mathbf{R}_t$ but three time steps earlier.

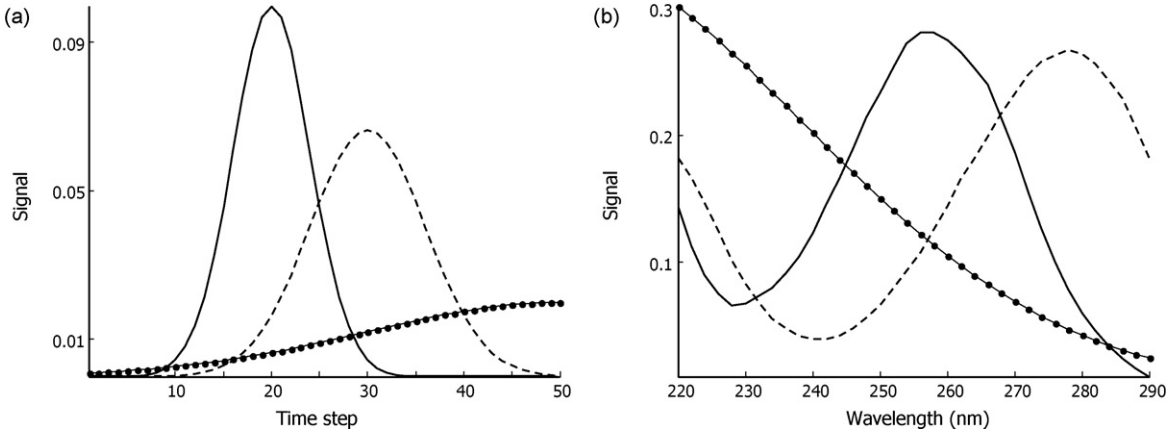

Fig. 1. (a) Simulated elution profiles. (b) Simulated normalized spectra. (–) Analyte, (– –) interference, (●–●) baseline.

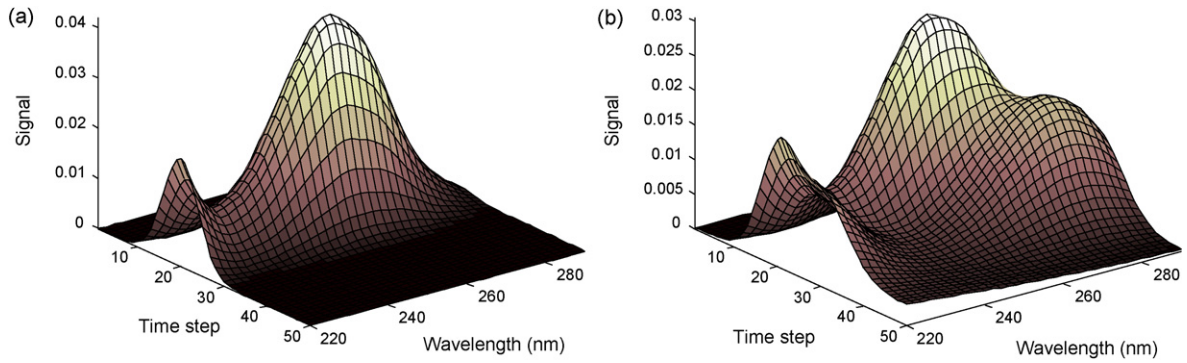

Fig. 2. (a) Simulated calibration standard peak. (b) Simulated test sample peak.

3.2. Measured data

Data from a previous study was used. The objective is the determination of 4-nitrophenol in a river water sample from the Ebre river (Spain). The experimental details are described in Ref. [16]. Briefly, a river water sample free of 4-nitrophenol (the original study included also other analytes) was collected and spiked at 1.01 ppb of 4-nitrophenol to produce a test sample. The calibration standard was another aliquot of that water sample spiked a 5.07 ppb of the analyte. The samples were submitted to HPLC-DAD analysis after solid-phase extraction (SPE). Sample chromatograms are plotted in Fig. 3 in Ref. [16].

All calculations were performed with subroutines made in house for Matlab (MathWorks, MA, USA).

4. Results and discussion

4.1. Simulations

The system $\mathbf{R}_c$ and $\mathbf{R}_t$ represents a probable quantification situation: $\mathbf{R}_c$ is a pure peak of the analyte (easily available from a pure standard, probably one of the standards used for univariate calibration) and $\mathbf{R}_t$ is the peak of the analyte in the test sample that unexpectedly coeluted with an interference. A small baseline drift has also been added. The optimal model dimensionality for this system is three because there are three systematic contributions (analyte, interference and baseline). Fig. 3 shows the estimated elution profiles ($\hat{\mathbf{H}}$) and the normalized spectra ($\hat{\mathbf{Y}}$), Eqs. (6) and (7), for the three-factors model. The spectrum of the analyte coincides with the underlying spectrum in Fig. 1(b). The elution profile of the analyte coincides with the source profile (Fig. 1(a)) multiplied by the concentration in the sum matrix $\mathbf{R}$, i.e., 2.5. The recovered profiles and spectra of the interference and of the baseline are different than those used for simulation. This is to be expected because in GRAM, when two or more constituents have the same ratio of concentrations between the standard and the test sample, their columns in $\hat{\mathbf{H}}$ and $\hat{\mathbf{Y}}$ may be linear combinations of the true underlying profiles. This happens here with the interference and the baseline. They are both in $\mathbf{R}_t$ but not in $\mathbf{R}_c$.

The first outlier test compares the elution profiles for the analyte k in $\hat{\mathbf{H}}_c$ and $\hat{\mathbf{H}}_t$. In this case, the normalized profiles are not shown because they coincide among them, and are also equal to the normalized profile in Fig. 3. This indicates that the peak of analyte satisfies the trilinearity condition.

Fig. 4(a) and (b) shows matrices $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$, that contain the part of the signal in $\mathbf{R}_c$ and $\mathbf{R}_t$ that is orthogonal to the profiles of the interference and the baseline. This orthogonal signal includes the NAS of the analyte of interest in both modes plus the projected noise. Since noise is low and $\mathbf{R}_c$ and $\mathbf{R}_t$ are trilinear, $\mathbf{E}_c^*$ and $\mathbf{E}_t^*$ are also small so $\mathbf{R}_c^* \approx \hat{\mathbf{R}}_c^*$ and $\mathbf{R}_t^* \approx \hat{\mathbf{R}}_t^*$. Hence, noise apart, $\mathbf{R}_t^*$ and $\mathbf{R}_c^*$ are proportional if $\mathbf{R}_c$ and $\mathbf{R}_t$ are trilinear. The proportionality can be checked by approximating $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ by their first right and left singular vectors, $\mathbf{R}_c^* \approx \mathbf{u}_c\, \mathbf{s}_{c,1}\, \mathbf{v}_c{}^T$ and $\mathbf{R}_t^* \approx \mathbf{u}_t\, \mathbf{s}_{t,1}\, \mathbf{v}_t{}^T$ (Eqs. (22) and (23)). $\mathbf{u}_c$ is the normalized version of the net elution profile at each wavelength in $\mathbf{R}_c^*$ and $\mathbf{v}_c$ is the normalized version of the net spectral profile at each retention time in $\mathbf{R}_c^*$. Similarly, $\mathbf{u}_t$ and $\mathbf{v}_t$ are the common net elution profile and net spectral profile in $\mathbf{R}_t^*$. Since $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ are proportional, $\mathbf{u}_c$ and $\mathbf{u}_t$ coincide (Fig. 4(c)), and $\mathbf{v}_c$ and $\mathbf{v}_t$ coincide (Fig. 4(d)). This indicates that the trilinearity requirement is fulfilled and gives confidence to the predicted concentration, that in this case was 1.00.

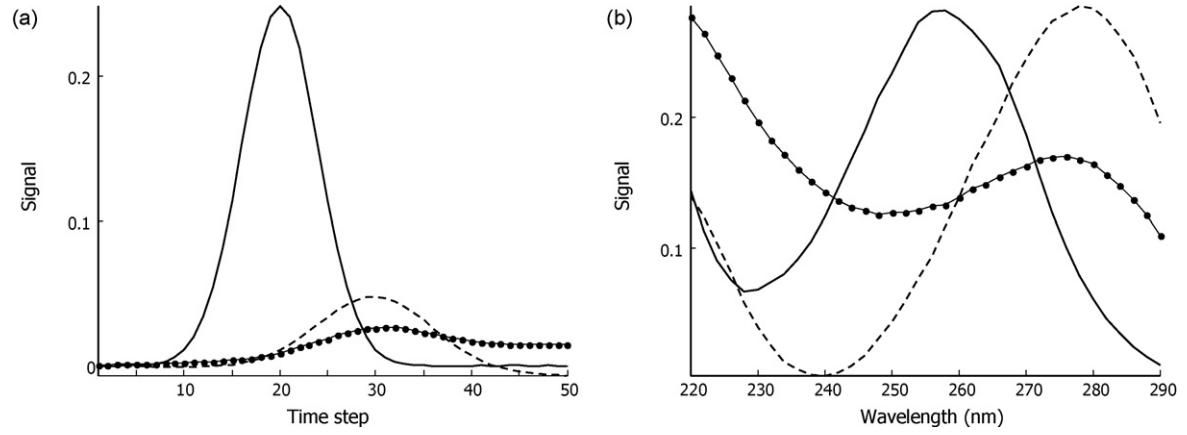

Fig. 3. Trilinear simulated data. GRAM model with three factors. (a) Estimated elution profiles $\hat{\mathbf{H}}$ (b) Estimated normalized spectra $\hat{\mathbf{Y}}$. (–) Analyte, (– –) interference, (●–●) baseline.

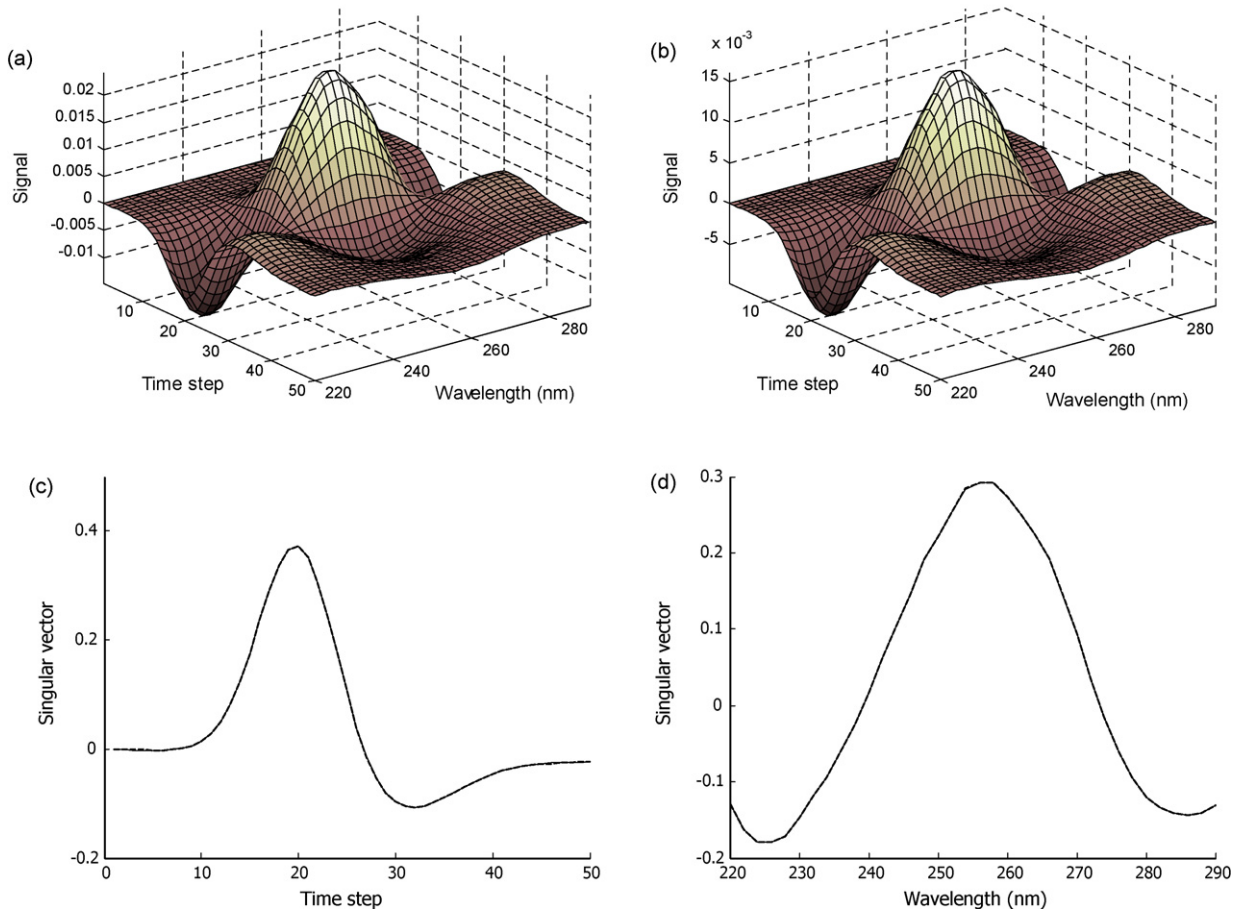

Fig. 4. Trilinear simulated data. (a) Orthogonal signal for the analyte of interest in the calibration standard peak $\mathbf{R}_c^*$ (b) Orthogonal signal for the analyte of interest in the test sample peak $\mathbf{R}_t^*$ (c) Left singular vectors $\mathbf{u}_c$ and $\mathbf{u}_t$ from $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ (d) Right singular vectors $\mathbf{v}_c$ and $\mathbf{v}_t$ from $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$.

Fig. 5(a) and (b) shows the predicted profiles when $\mathbf{R}_t$ suffers from time shift. Since the spectral mode is not affected by time shift, the spectrum of the analyte is recovered correctly and coincides with the one used for simulation. On the other hand, the time mode is affected by time shift, and the underlying elution profiles in $\mathbf{R}_c$ and $\mathbf{R}_t$ do not coincide. Despite this, the recovered elution profile of the analyte seems correct (positive and unimodal) except for some small negative values from time step 35 to 50 that could be attributed to a baseline variation in the recorded peak and be ignored. In other words, the visual inspection of the recovered elution profile and spectrum of the analyte would not warn that $\mathbf{R}_t$ is an outlier. The elution profiles and spectra of the interference and baseline can neither be used to detect this problem. The negative part in the spectrum of the interference could be due to the fact that these profiles are linear combinations of the true underlying profiles so these deviations from the ideal shape should not be taken into account. For this system, the predicted concentration was 1.14, a 14% error that would pass unnoticed.

Fig. 6 shows the elution profiles for the analyte of interest in $\hat{\mathbf{H}}_c$, $\hat{\mathbf{H}}_t$ and $\hat{\mathbf{H}}$. The difference between them indicates that the test

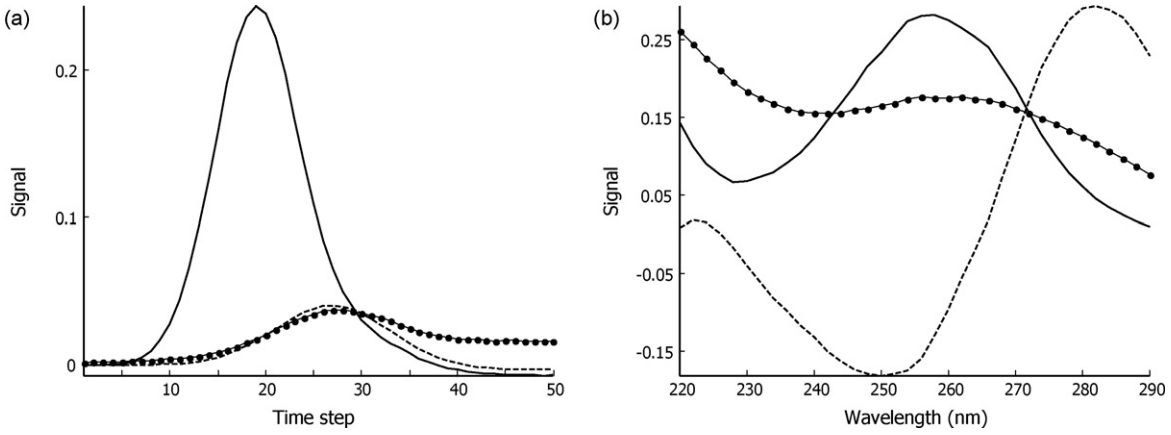

Fig. 5. Time shifted simulated data. GRAM model with three factors. (a) Estimated elution profiles $\hat{\mathbf{H}}$. (b) Estimated normalized spectra $\hat{\mathbf{Y}}$. (–) Analyte, (– –) interference, (●–●) baseline.

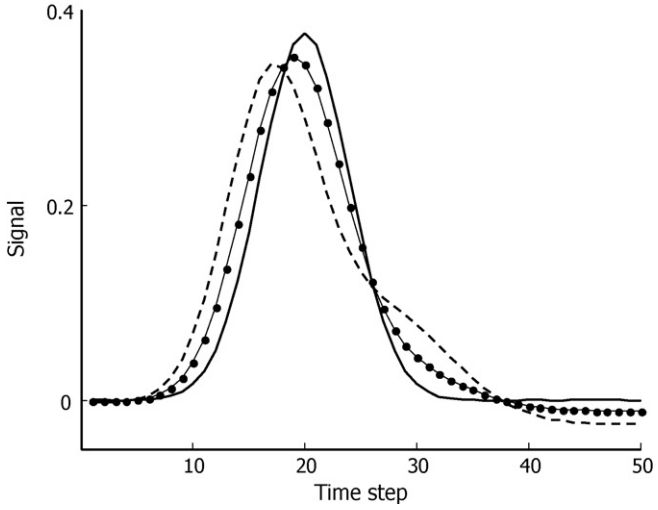

Fig. 6. Normalized elution profile of the analyte in $\hat{\mathbf{H}}_c$ (−), $\hat{\mathbf{H}}_t$ (− −) and $\hat{\mathbf{H}}$ (●–●).

sample is an outlier. Note how the profile $\hat{\mathbf{H}}$ is a compromise of the profiles in $\hat{\mathbf{H}}_c$ and $\hat{\mathbf{H}}_t$. The outlying situation can also be detected from the plots of the orthogonal signal. Since the underlying elution profile of the analyte in $\mathbf{R}_c$ is not aligned with its profile in $\mathbf{R}_t$, each profile has a different part that is orthogonal to the space spanned by the profiles of the interference and the baseline. Hence, $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ (Fig. 7(a) and (b)) are no longer proportional (noise apart) and one surface is shifted in time with respect to the other. This can be better observed because the singular vectors $\mathbf{u}_c$ and $\mathbf{u}_t$ are different (Fig. 7(c)), so $\mathbf{R}_t$ is flagged as an outlier. Note that the orthogonal contribution in the spectral mode is not affected because the time shift did not affect the spectra (Fig. 7(d)). After $\mathbf{R}_t$ has been flagged as an outlier, a time shift correction algorithm could be used, that would reveal a shift of three time steps. Correction of this shift would lead to the results discussed in the previous section.

4.2. Measured data

Fig. 8 shows the peaks of 4-nitrophenol in the standard and in the test river water sample. The SVD of each peak separately indicated that the peaks were not pure so quantitation with GRAM was justified. Fig. 9 shows the chromatographic profiles and spectra estimated by the GRAM model with three factors. The estimated spectrum of the analyte perfectly matches the spectrum recorded for 4-nitrophenol in a pure standard. This is used to identify what of the three elution profiles corresponds to the analyte. The other two spectra seem reasonable although they could be linear combinations of the true spectra. The elution profile of the analyte is unimodal and mostly positive and seems correct. The agreement of the spectrum and the satisfactory elution profiles could suggest, at a first glance, that the prediction, that is 0.91 ppb, is reliable. The outlier diagnostics, however, will flag the sample as an outlier.

Fig. 10(a) shows the elution profile of the analyte of interest in $\hat{\mathbf{H}}_c$, $\hat{\mathbf{H}}_t$ and $\hat{\mathbf{H}}$. The separation between the profiles in $\hat{\mathbf{H}}_c$ and $\hat{\mathbf{H}}_t$ indicate that $\mathbf{R}_t$ is an outlier. Note also that the profile in $\hat{\mathbf{H}}$ is more similar to the profile in $\hat{\mathbf{H}}_c$ than the profile in $\hat{\mathbf{H}}_t$. This is because $\alpha = 1$

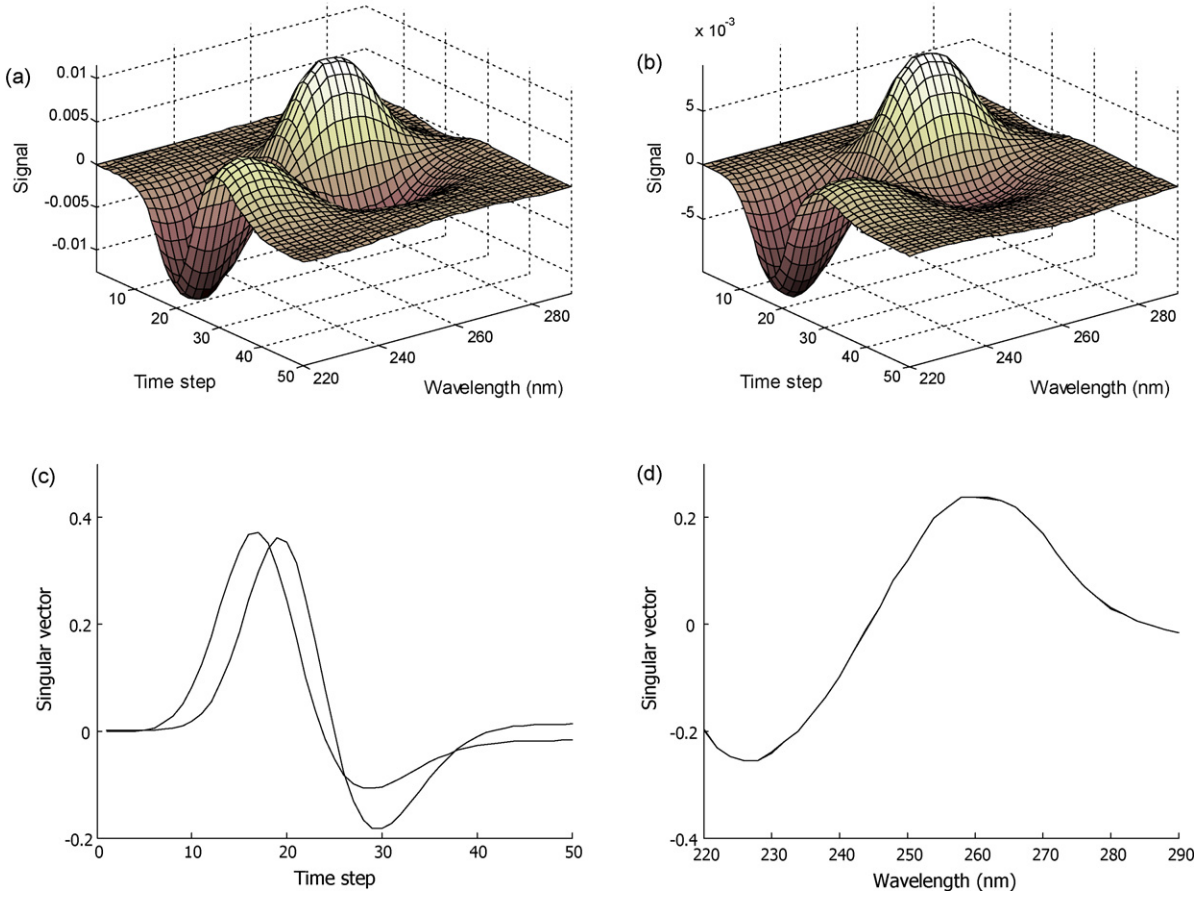

Fig. 7. Time shifted simulated data. (a) $\mathbf{R}_c^*$, (b) $\mathbf{R}_t^*$. (c) The difference between the left singular vectors $\mathbf{u}_c$ and $\mathbf{u}_t$ from $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ highlight the time shift problem. (d) Right singular vectors $\mathbf{v}_c$ and $\mathbf{v}_t$ from $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$.

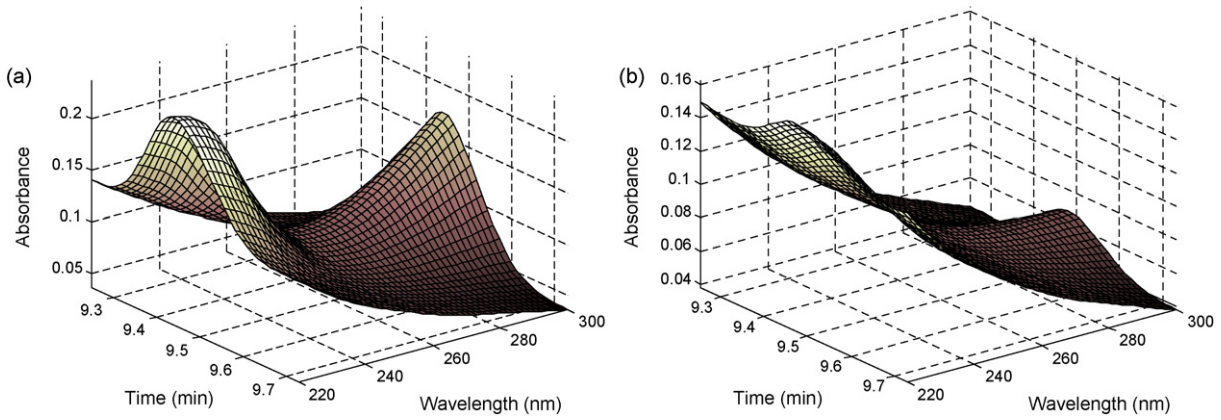

Fig. 8. River water sample. (a) Measured 4-nitrophenol in the standard. (b) Measured 4-nitrophenol in the test sample.

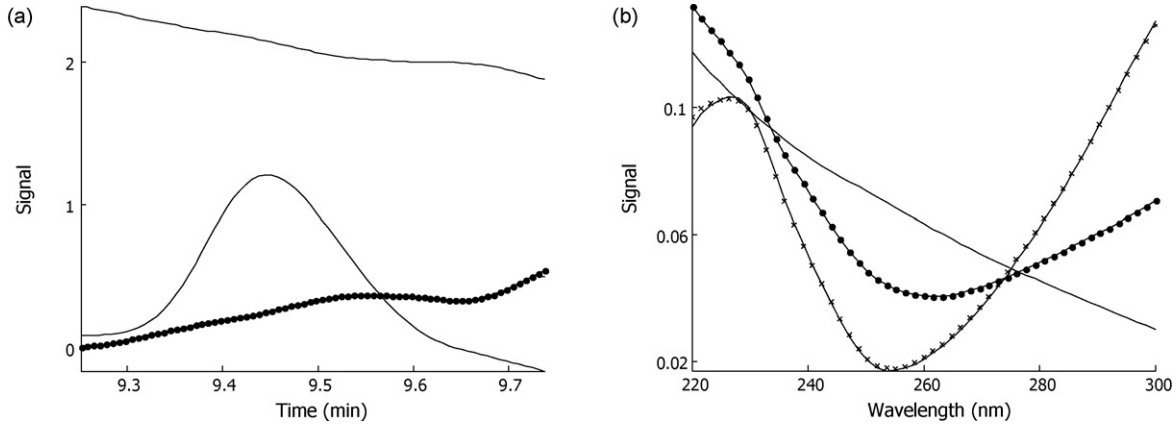

Fig. 9. River water sample. GRAM model with three factors. (a) Estimated elution profiles $\hat{\mathbf{H}}$. (b) Estimated normalized spectra $\hat{\boldsymbol{Y}}$. The (×) indicate the reference spectrum of the analyte obtained by measuring a pure standard of 4-nitrophenol.

was used in Eq. (3). Hence, $\mathbf{R}_c$ (that is much larger than $\mathbf{R}_t$) dominated in the sum matrix $\mathbf{R}$ and in the SVD in Eq. (4). If $\alpha = 0.1$ had been used, the profile in $\hat{\mathbf{H}}$ would be more similar to the profile in $\hat{\mathbf{H}}_t$ than to the profile in $\hat{\mathbf{H}}_c$. $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ in Fig. 11(a) and (b) also lead to the conclusion that $\mathbf{R}_t$ is an outlier. Note that one surface is shifted with respect to the other. The fact that $\mathbf{u}_c$ and $\mathbf{u}_t$ do not coincide in Fig. 11(c) confirms this. Again, the spectral mode was not affected by time shift and the singular vectors of $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ for the spectral mode, $\mathbf{v}_c$ and $\mathbf{v}_t$, agree. The time shift was corrected by moving the time window of $\mathbf{R}_t$ in the chromatogram of the test sample one unit a time until the calculated profiles agreed. Other algorithms, such as the one by Prazen et al. [13] could also be used. Fig. 12 shows the elution profiles and spectra estimated by GRAM after the time window for $\mathbf{R}_t$ had been shifted six time steps with respect to the time window in $\mathbf{R}_c$. Again, the profile of 4-nitrophenol seems correct and its spectrum perfectly matches the reference spectrum of this analyte. Fig. 10(b) now shows that the elution profiles for the analyte of interest in $\hat{\mathbf{H}}_c$, $\hat{\mathbf{H}}_t$ and $\hat{\mathbf{H}}$ agree. The surfaces of $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ in Fig. 13(a) and (b) are now similar, which is confirmed by the match between $\mathbf{u}_c$ and $\mathbf{u}_t$ (Fig. 13(c)). These plots suggest that the

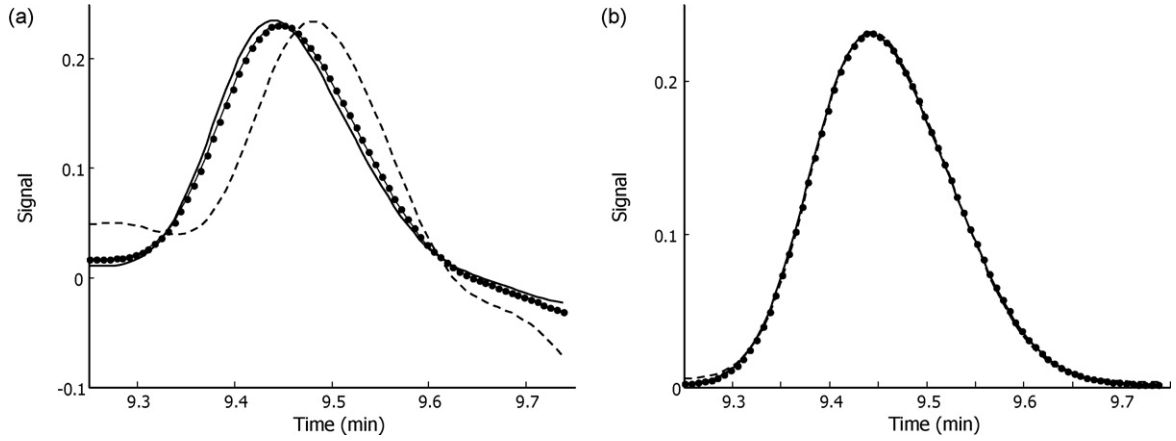

Fig. 10. Normalized elution profile of 4-nitrophenol in $\hat{\mathbf{H}}_c$ (−), $\hat{\mathbf{H}}_t$ (− −) and $\hat{\mathbf{H}}$ (•−•) for the raw measured data (a) and after correcting the retention time shift (b).

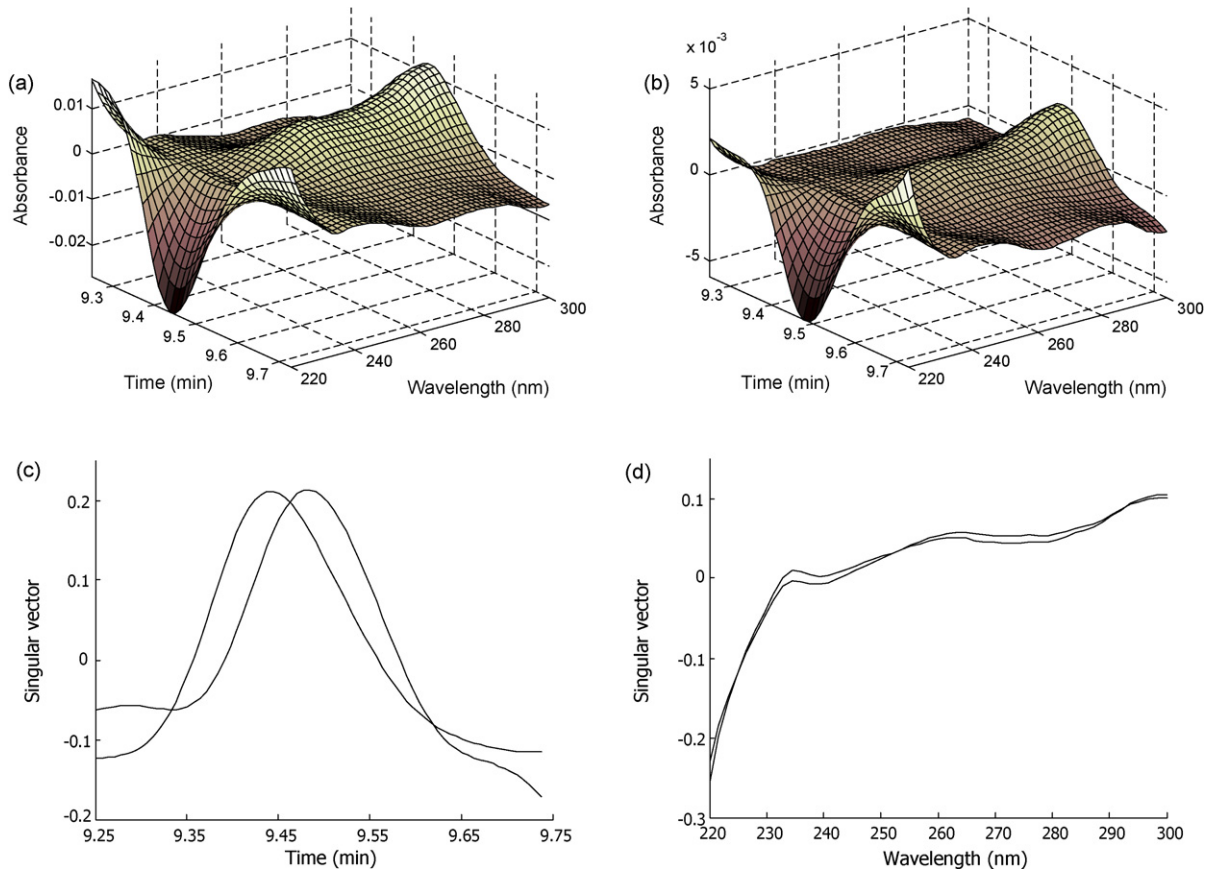

Fig. 11. River water sample (a) $\mathbf{R}_c^*$, (b) $\mathbf{R}_t^*$. (c) Left singular vectors $\mathbf{u}_c$ and $\mathbf{u}_t$ from $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$. (d) Right singular vectors $\mathbf{v}_c$ and $\mathbf{v}_t$ from $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$.

trilinearity has bee improved and give confidence to the predicted concentration, that was 1.04 ppb. This value was very close to the known spiked amount 1.01 ppb. This result was considered acceptable taking into account the dispersion of the results that the SPE step can produce.

An additional plot can be derived from the projected matrices. Fig. 14 shows the scatter plot of $vec(\mathbf{R}_t^*)$ versus $vec(\mathbf{S}^*)$, where $\mathbf{S}^*$ is the net sensitivity for analyte k defined as the NAS at unit concentration, $\mathbf{S}^* = \hat{\mathbf{R}}_c^*/c_c$. The slope of the least-squares fitted line in this plot is the predicted concentration. Fig. 14(a) corresponds to the original, non-trilinear data. Time shift makes $\mathbf{R}_t^*$ not be a multiple of $\mathbf{S}^*$, and the residuals are large and systematic. After correcting the time shift (Fig. 14(b)), the residual are much smaller and the dispersion is mainly due to the noise in the data, which increases the reliability of the predicted concentration.

The indicated plots can be also used to detect other sources of non-trilinearity such as peak broadening. Peak broadening (i.e., the peak of the analyte in test sample is aligned with the peak of the standard but it is not a multiple of it) also increases prediction error, but usually less than time shift because the shape of the elution profile of the analyte does not vary excessively from one run to the other. This case, although not shown here, can be detected with the same reasoning than retention time shift. If the profile of the analyte of interest in $\mathbf{R}_c$ is different than in $\mathbf{R}_t$, then $\hat{\mathbf{H}}_c$, $\hat{\mathbf{H}}_t$ and $\hat{\mathbf{H}}$ will not coincide, $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ will not be proportional and $\mathbf{R}_t$ will be flagged as an outlier.

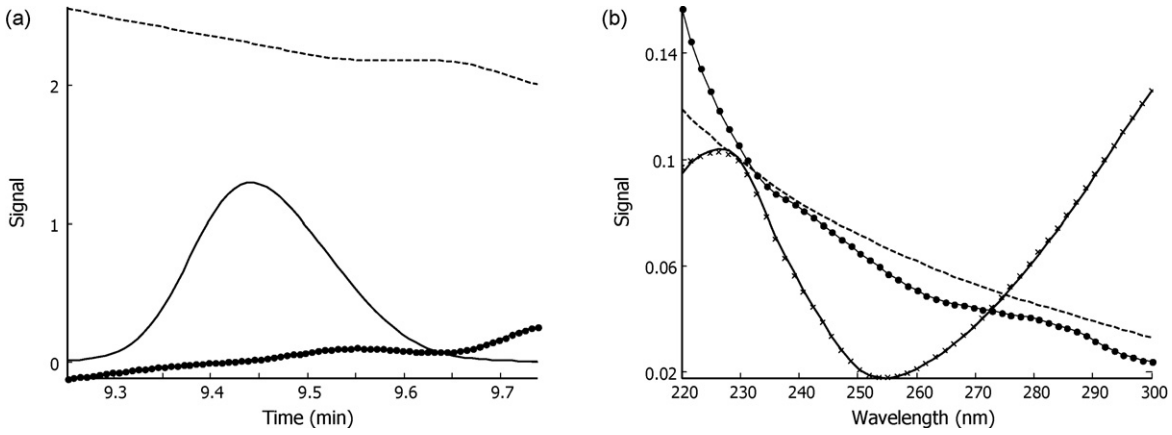

Fig. 12. River water sample after correcting the time shift. GRAM model with three factors. (a) Estimated elution profiles $\hat{\mathbf{H}}$. (b) Estimated normalized spectra $\hat{\mathbf{Y}}$. The (×) indicate the reference spectrum of the analyte obtained by measuring a pure standard of 4-nitrophenol.

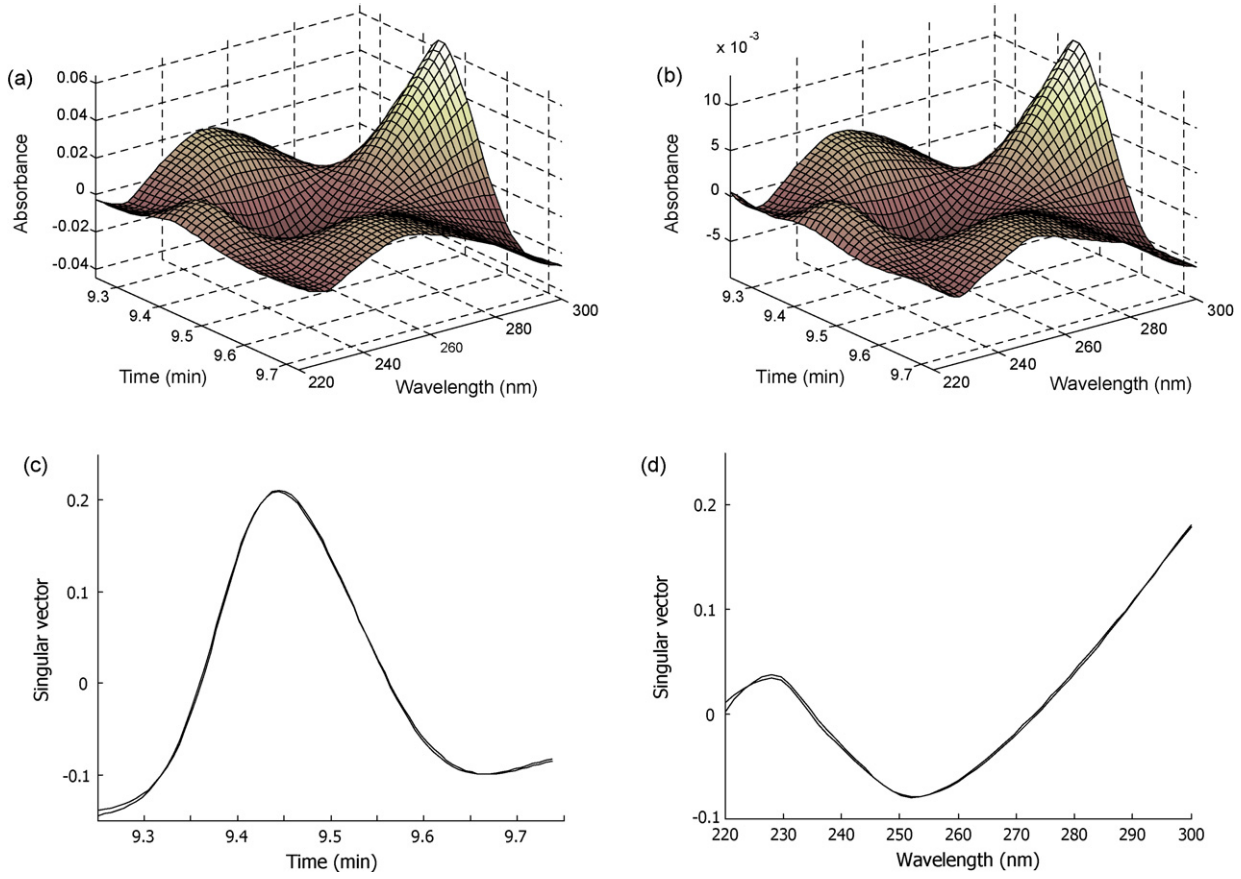

Fig. 13. River water sample after correcting the time shift. (a) $\mathbf{R}_c^*$, (b) $\mathbf{R}_t^*$. (c) Left singular vectors $\mathbf{u}_c$ and $\mathbf{u}_t$ from $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$. (d) Right singular vectors $\mathbf{v}_c$ and $\mathbf{v}_t$ from $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$.

A similar behavior can also be observed in these plots when $\mathbf{R}_t$ is not an outlier but the model is underfitted i.e., the GRAM model, Eq. (4), is calculated with less factors than needed. In an underfitted model, $\hat{\mathbf{H}}$ and $\hat{\mathbf{Y}}$ do not span the row and column spaces of $\mathbf{R}_c$ and $\mathbf{R}_t$ correctly. Hence, the calculated profiles often lack meaning or it is difficult to recognize the spectrum of the analyte in $\hat{\mathbf{Y}}$. This is a warning for checking the GRAM model and see whether this was caused by a shift in the elution profile (i.e., $\mathbf{R}_t$ is an outlier) or underfitting. Sometimes, however, the spectrum of the target analyte can be identified in $\hat{\mathbf{Y}}$ despite the model being underfitted. This may happen, for example, when the contribution of one the components in $\mathbf{R}_t$ is low. In that case, the analyte may be (erroneously) quantified. The proposed plots display that situation. If the profile of the analyte has been identified, matrices $\hat{\mathbf{H}}_{-k}$ and $\hat{\mathbf{Y}}_{-k}$ can be constructed with the remaining columns of $\hat{\mathbf{H}}$ and $\hat{\mathbf{Y}}$. In that case, $\mathbf{P}_H$ and $\mathbf{P}_Y$ will not completely remove the contribution of the interferences, $\mathbf{R}_c^*$ and $\mathbf{R}_t^*$ will not be proportional (up to the noise level) and the plot $vec(\mathbf{R}_t^*)$ versus $vec(\mathbf{S}^*)$, will reveal this situation. Overfitting, on the other hand, does not increase prediction error as much as underfitting does. Usually adding an extra factor leaves

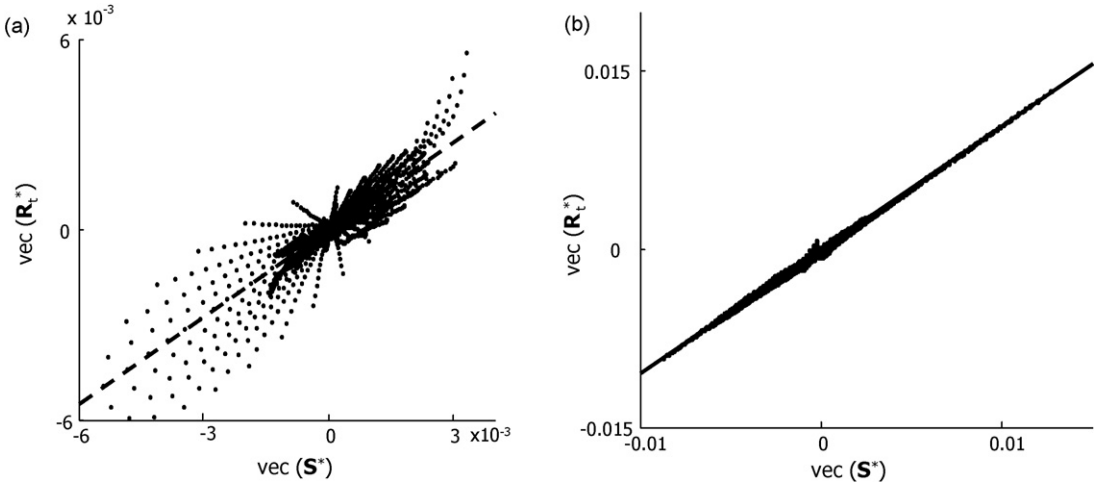

Fig. 14. River water sample. $vec(\mathbf{R}_t^*)$ versus $vec(\mathbf{S}^*)$. (a) Before correcting the time shift. (b) After correcting the time shift. In both cases, the slope of the fitted line is the predicted concentration.

the existing profiles mostly unchanged and adds a new one that accounts for noise, or one of the existing profiles is split in two similar profiles because of small data inconsistencies (for example, slight non-linearity or variation of the peak shape). Note, for example, that the GRAM model for 4-nitrophenol, after correcting the shift, has two elution profiles with a very similar spectrum. This may be indicating that the main sources of variation are two and that the third profile was forced to come out because the model was calculated with three factors. In fact, a GRAM model calculated with only two factors yields a very similar profiles and spectra than the model with three factors, and the prediction hardly changed (1.04 ppb). Finally, note that although slight overfitting may not affect the prediction, adding more factors than needed decreases the NAS and decreases the signal-to-noise ratio because the NAS must be orthogonal to more profiles. Hence, it is important to use the best number of factors.

5. Conclusions

Outlier detection diagnostics are needed for the routine application of GRAM. In HPLC-DAD analysis, retention time shift and peak broadening are two important sources of outliers. Although qualitative analysis may still be possible because the spectral mode is not affected by the lack of trilinearity, quantitative analysis may be seriously affected. Several tools for the detection of unreliable predictions have been presented. From them, time shift, peak broadening and underfitting can be detected. It is a good practice to monitor these plots for different number of factors in the GRAM model before the quantitative result is released.

Acknowledgments

We thank R.A. Gimeno, R.M. Marcé and F. Borrull, from the Universitat Rovira i Virgili, for providing the measured data. The support of the Spanish Ministerio de Educación y Ciencia, project CTQ2007-66918/BQU is acknowledged.

References

[1] J. Ferré, R. Boqué, N.M. Faber, in: S. Brown, R. Tauler, R. Walczak (Eds.), Comprehensive Chemometrics, vol. 2, Elsevier, Oxford, 2009, pp. 365–409.
[2] G.C. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 72 (2000) 4154.
[3] G.M. Gross, B.J. Prazen, R.E. Synovec, Anal. Chim. Acta 490 (2003) 197.
[4] Z. Lin, K.S. Booksh, L.W. Burgess, B.R. Kowalski, Anal. Chem. 66 (1994) 2552.
[5] R.B. Poe, S.C. Rutan, Anal. Chim. Acta 283 (1993) 845.
[6] L.S. Ramos, E. Sanchez, B.R. Kowalski, J. Chromatogr. 385 (1987) 165.
[7] C.G. Fraga, C.A. Bruckner, R.E. Synovec, Anal. Chem. 73 (2001) 675.
[8] S. Li, J.C. Hamilton, P.J. Gemperline, Anal. Chem. 64 (1992) 599.
[9] G.C. Fraga, J. Chromatogr. A 1019 (2003) 31.
[10] R.A. Gimeno, E. Comas, R.M. Marcé, J. Ferré, F.X. Rius, F. Borrull, Anal. Chim. Acta 498 (2003) 47.
[11] M. Jalali-Heravi, M. Vosough, Anal. Chim. Acta 537 (2005) 89.
[12] C.G. Fraga, C.A. Corley, J. Chromatogr. A 1096 (2005) 40.
[13] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218.
[14] E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius, J. Chromatogr. A 988 (2003) 277.
[15] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Anal. Chem. 70 (1998) 2796.
[16] E. Comas, R.A. Gimeno, J. Ferré, R.M. Marcé, F. Borrull, F.X. Rius, J. Chromatogr. A 1035 (2004) 195.
[17] H.A.L. Kiers, A.K. Smilde, J. Chemometr. 9 (1995) 179.
[18] C.G. Fraga, B.J. Prazen, R.E. Synovec, Anal. Chem. 73 (2001) 5833.
[19] B.E. Wilson, W. Lindberg, B.R. Kowalski, J. Am. Chem. Soc. 111 (1989) 3797.
[20] E. Sánchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496.
[21] R. Tauler, A.K. Smilde, B.R. Kowalski, J. Chemometr. 9 (1995) 31.
[22] E. Comas, J. Ferré, F.X. Rius, Anal. Chim. Acta 515 (2004) 23.
[23] N.M. Faber, J. Ferré, R. Boqué, Chemometr. Intell. Lab. Syst. 55 (2001) 67.
[24] K. Faber, A. Lorber, B.R. Kowalski, J. Chemometr. 11 (1997) 419.