Journal of Chromatography A, 1507 (2017) 1–10

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Assisted baseline subtraction in complex chromatograms using the BEADS algorithm

J.A. Navarro-Huerta[a], J.R. Torres-Lapasió[a,*], S. López-Ureña[b], M.C. García-Alvarez-Coque[a]

[a] Department of Analytical Chemistry, Faculty of Chemistry, Universitat de València, c/ Dr. Moliner 50, 46100 Burjassot, Spain
[b] Department of Mathematics, Faculty of Mathematics, Universitat de València, c/ Dr. Moliner 50, 46100 Burjassot, Spain

ARTICLE INFO

Article history:
Received 21 March 2017
Received in revised form 24 May 2017
Accepted 26 May 2017
Available online 29 May 2017

Keywords:
HPLC
Complex samples
Baseline subtraction
BEADS
Working parameters
Autocorrelation

ABSTRACT

The data processing step of complex signals in high-performance liquid chromatography may constitute a bottleneck to obtain significant information from chromatograms. Data pre-processing should be preferably done with little (or no) user supervision, for a maximal benefit and highest speed. In this work, a tool for the configuration of a state-of-the-art baseline subtraction algorithm, called BEADS (Baseline Estimation And Denoising using Sparsity) is developed and verified. A quality criterion based on the measurement of the autocorrelation level was designed to select the most suitable working parameters to obtain the best baseline. The use of a log transformation of the signal attenuated artifacts associated to a large disparity in signal size between sample constituents. Conventional BEADS makes use of trial-and-error strategies to set up the working parameters, which makes the process slow and inconsistent. This constitutes a major drawback in its successful application. In contrast, the assisted BEADS simplifies the setup, shortens the processing time and makes the baseline subtraction more reliable. The assisted algorithm was tested on several complex chromatograms corresponding to extracts of medicinal herbs analysed with acetonitrile-water gradients, and a mixture of sulphonamides eluted with acetonitrile gradients in the presence of the non-ionic surfactant Brij-35 under micellar conditions.

1. Introduction

Modern high-performance liquid chromatography (HPLC) instruments are able to provide highly complex signals in routine analysis, from which the relevant information should be extracted [1]. In these analyses, the data processing step constitutes a bottleneck, constraining sample throughput [2,3]. Problems such as noisy signals, coeluting peaks (sometimes highly overlapped), peak shifts and the presence of irregular baselines should be addressed. The operations to handle these problems should be done preferably with little (or no) user supervision for a maximal benefit and highest speed.

The aim of this work is to improve the baseline subtraction in chromatograms of high complexity, with complete drift suppression and little analyst supervision. Very recently, a new algorithm called "Baseline Estimation And Denoising using Sparsity" (BEADS) was proposed [4,5], which presents as novelty the capability of performing a full decomposition of chromatograms in net signal (i.e., the pure signals of the analytes and their accompanying compounds), baseline and noise. The baseline is modelled as a low frequency signal and the noise as a high frequency contribution, while the peaks of analytes are described as sparse signals, whose first and second derivatives are also sparse (a vector signal is classified as "sparse" when most of its elements are zero). For this purpose, BEADS requires that the user specify several parameters to ensure that the recovered signals have chemical meaning (e.g., positive signals for all analytes). It should be noted that most baseline subtraction algorithms also require some user inputs. This is the case of the mixture models based on splines proposed by Rooi and Eilers [6], the adaptive iteratively reweighted penalised least squares (airPLS) [7], and the backcor algorithm [8].

The authors of BEADS validated it in comparison with the airPLS and backcor algorithms [4]. The three methods yielded reasonable estimates of the baselines, but BEADS offered the best performance. Indeed, in our trials with a variety of chromatograms, BEADS was verified to provide excellent results in complex situations. However, we found some issues that make its routine application to real samples difficult, which should be addressed.

[*] Corresponding author.
E-mail address: jrtorres@uv.es (J.R. Torres-Lapasió).

http://dx.doi.org/10.1016/j.chroma.2017.05.057
0021-9673/© 2017 Elsevier B.V. All rights reserved.

The triple decomposition of chromatograms in BEADS is done essentially by using highly efficient frequency filters, which makes the outline easier and the calculation faster. Moreover, the algorithmic framework is based on majorization-minimization [4], which converges quickly regardless of the set of values used in its initialization. The result of the combination of these techniques is a highly efficient algorithm that saves memory. Another advantage is that, in contrast to other baseline algorithms [8], the set of baselines obtained by BEADS is not described as a parametric family of functions. This feature confers BEADS an extreme flexibility to accommodate any baseline, whatever its complexity.

The limitations of BEADS can be classified in two categories. First, it requires a careful adjustment of the working parameters to properly process real signals of different origin. This operation may be difficult for highly complex signals, owing to the instability of the adjustment process (i.e., small changes in the parameters may lead to very different baselines). Secondly, chromatograms must fulfil some conditions (described in detail in Section 3.1), mandatory for the application of BEADS, but hardly fulfilled in practice with real chromatograms.

In this work, we analyse comprehensively the limitations of BEADS, and propose some solutions, which improve the results and reliability of this algorithm and contribute to make it more robust, faster and easier to apply to chromatograms of real highly complex samples of different origin, with little supervision.

2. Experimental

2.1. Reagents

In order to explore the correct subtraction of the baseline, several fingerprints of medicinal herbs were processed, corresponding to extracts in hot water of horsetail and decaffeinated teas obtained in the laboratory. For the chromatographic analysis, hydro-organic gradients were prepared with acetonitrile (Scharlab, HPLC grade, Barcelona, Spain) and water. This was buffered at pH 3 with 0.01 M sodium dihydrogen phosphate (Sigma, Roedermark, Germany) and a suitable amount of 0.01 M HCl (Scharlab). The chromatographic signals of extracts of red peony root, taken from Ref. [7], were also processed.

The influence of negative peaks associated with refractometric void volume signals was studied using chromatograms for a mixture of 15 sulphonamides: sulphaguanidine, sulphanilamide, sulphacetamide, sulphadiazine, sulphathiazole, sulphapyridine, sulphamera-zine, sulphamethazine, sulphamethizole, sulphamonomethoxine, sulphachloropyridazine, sulphamethoxazole, sulphisoxazole, sulphadimethoxine and sulphaquinoxaline, eluted with an acetonitrile gradient in the presence of Brij-35 (Sigma, St. Louis, MO, USA), buffered at pH 3 with 0.01 M sodium dihydrogen phosphate. All solutions were filtered through 0.45 μm Nylon membranes from Micron Separations (Westboro, MA, USA), before their injection into the chromatographic system.

2.2. Preparation of extracts of medicinal herbs

The extracts of horsetail and decaffeinated teas were obtained following the recommendations of Dumarey et al. [9]. For this purpose, 20 ml of nanopure water was added to 0.2 g of ground sample, and boiled in the absence of light. The extracts were filtered through 0.2 μm membrane filters from Pall Gelman Laboratory (Karlstein/Main, Germany), to finally fill 2 ml vials for chromatographic analysis.

2.3. Apparatus, columns and software

An Agilent modular instrument (HP 1100, Waldbronn, Germany) was used, consisting of quaternary pump, automatic injector, temperature controller, and variable wavelength UV-visible detector. The chromatograms of the medicinal herbs and mixtures of sulphonamides were detected at 210 and 254 nm, respectively. The column temperature was fixed at 25 °C. The injection volume was 10 μl, and the flow rate was kept constant at 1 ml/min, in all instances.

An OpenLAB CDS LC ChemStation (Agilent, B.04.03 revision) was used for the acquisition of chromatographic signals. Raw chromatograms were processed without any correction by the ChemStation software, unless those associated to the default working parameters, such as autobalance in the pre-run, 5% zero offset, or attenuation to 1000 mAU. Matlab 2016b (The MathWorks Inc., Natick, MA, USA) was applied for data treatment. The Matlab function [5] (which is included in the Supplementary material of Ref. [4]) was used for the conventional application of BEADS.

3. Results and discussion

3.1. Limitations of BEADS

As indicated, BEADS makes the simultaneous decomposition of a signal $\mathbf{y}$ in three contributions:

$$\mathbf{y} = [y_1, y_2, .., y_n] = \mathbf{c} + \mathbf{b} + \mathbf{e} \quad (1)$$

where $\mathbf{c}$, $\mathbf{b}$ and $\mathbf{e}$ make reference to the sparse chromatogram, baseline and noise vectors computed by BEADs, which depend on a set of working parameters $\mathbf{p}$. The working parameters are the cutoff frequency (f_c, which constitutes the boundary between the baseline and the rest of contributions), asymmetry (r, which penalizes the negative values) and regularization parameters (λ_0, λ_1 and λ_2, which control the sparsity of vector $\mathbf{c}$). An additional parameter is the amplitude (A), which multiplies the regularization parameters; thus, the regularization parameters are actually $A \times \lambda_i$, which makes the ratios among the λ_i parameters independent of their magnitude.

The adaptability of BEADS to real baselines is noteworthy, but its application has the following limitations, especially severe for complex chromatograms:

(i) Requirement of the same signal intensity for the first and last points in the chromatogram (i.e., periodicity of the signal).
(ii) Abnormal risings of the baseline under major signals in chromatograms where the analytes exhibit extreme variations in signal size. The overall appearance of the computed baseline is wavy (see figures discussed in Section 3.2), instead of having a smooth trend at large scale.
(iii) Problematic processing of chromatograms containing sporadic negative peaks, such as those corresponding to refractometric signals, or those observed in chromatograms obtained using indirect UV–visible detection. This forces a careful adjustment of the working parameters for each sample.
(iv) Dependence among the working parameters. The baseline is particularly susceptible to the selected cutoff frequency at low frequencies, which results in an unstable adjustment process. This situation is worsened by the wide range of values to be explored, which in some cases comprises several orders of magnitude (a typical chromatogram composed of 10,000 points can involve exploring cutoff frequencies over 4 orders of magnitude).
(v) Need for each chromatogram of a particular adaptation of the working parameters (i.e., each set of parameters is translated

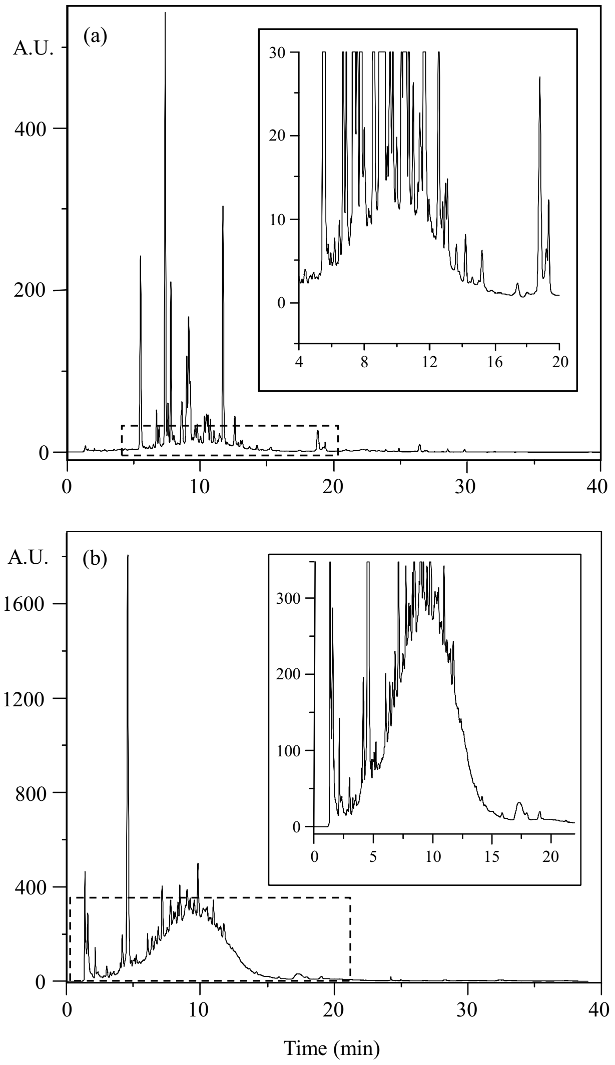

Fig. 1. Chromatographic fingerprints for extracts of: (a) horsetail tea, and (b) decaffeinated tea, before being processed. The chromatograms were obtained with a 20–60% (v/v) acetonitrile gradient reaching the upper concentration in 10 min. The upper inserts magnify the central regions of the chromatograms to highlight the complexity of the baseline associated with the matrix and gradient program.

in a different baseline). Fortunately, related samples may share similar parameter values.

For the development of the assisted BEADS, we used a set of 65 multi-analyte chromatograms, all of them with severe problems in their respective baselines. Three of these chromatograms are shown in Figs. 1 and 2. Those in Fig. 1 were obtained in our laboratory, and correspond to extracts of horsetail and decaffeinated teas (Fig. 1a and b, respectively). The separation was carried out with a Zorbax Eclipse XDB C18 column (150 mm × 4.6 mm I.D., 5 µm particles, Merck, Darmstadt, Germany), using gradient elution where the acetonitrile content was increased from 20 to 60% (v/v), in a gradient time of 10 min, while the pH was kept at a nominal value of 3. A chromatogram taken from Ref. [7] (Fig. 2), corresponding to an extract of red peony root, was also analysed. This chromatogram belongs to a set of 10 chromatograms originally processed by the authors to subtract baselines, using airPLS [7], the so-called "faster algorithm for betweenness centrality" (FABC) proposed by Cobas et al. [10], and the alternating least squares (ALS) algorithm [11].

The chromatogram of the horsetail tea sample (Fig. 1a) is used to illustrate the performance of the solutions proposed in this work to analyse complex signals, particularly the selection of the best working parameter values to be used by BEADS.

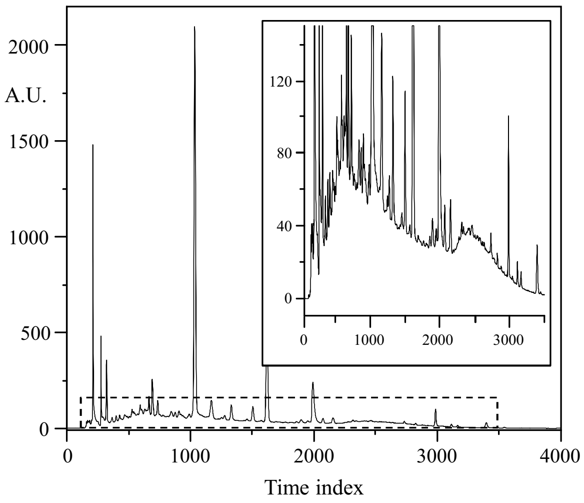

Fig. 2. Chromatographic fingerprint corresponding to an extract of red peony root, used in Refs. [4,7], before being processed. The upper insert magnifies the central region of the chromatograms to highlight the complexity of the baseline associated with the matrix and gradient program.

3.2. Monitoring the autocorrelation to explore the BEADS working parameters

The quality of the results offered by BEADS depends critically on the correct selection of the working parameters, especially the cutoff frequency, which has a major influence in the returned baseline. This relies on the fact that the main principle of BEADS is a decomposition based on the frequency. The other working parameters exhibit milder variations. BEADS parameters are conventionally adjusted by trial-and-error, and when one parameter is modified, others are collaterally misadjusted. This makes the process slow and unpredictable when a chromatogram with unknown characteristics (without any information about the correct frequency) is processed.

To facilitate the selection of the working parameters when the original BEADS algorithm is used, auxiliary plots (see Fig. 3 and Figs. S1–S5 in the Supplementary material) were designed. The auxiliary plots assist in the fast and reliable selection of the working parameters, which makes the application of the original BEADS troublesome. The plots are based on the measurement of the autocorrelation, which can be defined as the correlation of a signal with a delayed copy of itself [12]. Therefore, this property measures the similarity between consecutive data points in a series, such as the measurements taken at regular time intervals, as is the case of a chromatogram.

The auxiliary plot described in this section will be particularised to the case of the cutoff frequency. We will assume that the other BEADS parameters are more or less correctly set, although this is not so necessary in practice. Our hypothesis is that the removal of a certain feature from the chromatogram, such as the trend in the baseline, even being imperfect, would produce an alteration in the autocorrelation level. If the selected cutoff frequency were correct, by subtracting from the total chromatographic signal the contributions of the sparse chromatogram and baseline estimated by BEADS, only noise would remain. Ideally, this noise should not show any autocorrelation. At any other relatively close cutoff frequency, the decomposition in sparse chromatogram and noise will not be perfect, and some autocorrelation will persist. Therefore, the observation of the changes in the autocorrelation level will help to

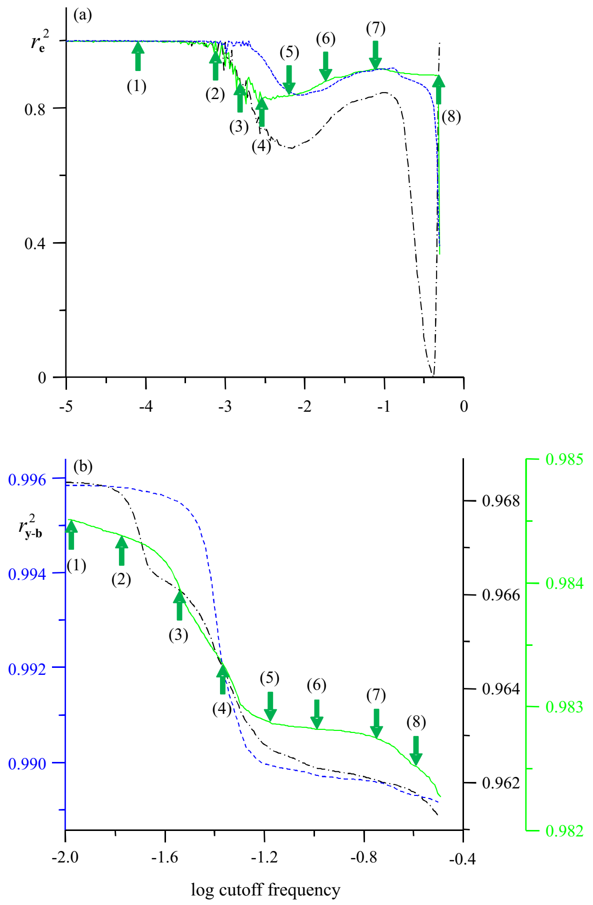

Fig. 3. Autocorrelation plots expressed as r^2 and the logarithm of the cutoff frequency used for the subtraction of the baseline: (a) original scale where was calculated from the noise, and (b) logarithmic scale where was calculated from the baseline corrected signal. Extracts: horsetail tea (continuous line and far right *y*-axis in (b)), decaffeinated tea (short dashed line and left *y*-axis), and red peony root (dotted dashed line and near right *y*-axis). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

select the features to be removed. It also will reveal that a certain feature has been removed in a particular parameter domain.

In this work, we have measured the autocorrelation based on the Durbin-Watson (*DW*) statistic [13]:

$$DW = \frac{\sum_{i=2}^{n} (d_i - d_{i-1})^2}{\sum_{i=1}^{n} {d_i}^2} \tag{2}$$

Conventionally, this statistic is applied to regression analysis and smoothing, d_i being the difference between the raw signal for point *i* minus the fitted (or smoothed) signal for that point. Therefore, d_i estimates the lack of fit in the case of fitting, and the noise in the case of smoothing. In the case of applying BEADS, an estimation of the noise can be obtained from Eq. (1):

$$\mathbf{e} = \mathbf{y} - \mathbf{c} - \mathbf{b} \tag{3}$$

This operation is equivalent to subtracting a real signal from the adjusted or smoothed signal, and allows the direct application of Eq. (2), making the difference vector $\mathbf{d} = \mathbf{e}$.

The *DW* statistic can be developed as follows:

$$DW = \frac{\sum_{i=2}^{n} d_i^2 - 2\sum_{i=2}^{n} d_i\, d_{i-1} + \sum_{i=2}^{n} d_{i-1}^2}{\sum_{i=1}^{n} d_i^2} \approx \frac{2\sum_{i=2}^{n} d_i^2 - 2\sum_{i=2}^{n} d_i\, d_{i-1}}{\sum_{i=1}^{n} d_i^2} = \frac{2\sum_{i=2}^{n} d_i^2}{\sum_{i=1}^{n} d_i^2} - \frac{2\sum_{i=2}^{n} d_i\, d_{i-1}}{\sum_{i=1}^{n} d_i^2} = 2 - 2r \tag{4}$$

where *r* measures the autocorrelation level for vector **d**. *DW* tends to 0 for a perfect positive correlation ($r = 1$), and to 4 for a perfect negative correlation ($r = -1$). As indicated, when the contributions of the analytes (i.e., sparse chromatogram) and baseline are perfectly subtracted from the raw chromatogram, only noise will remain. If this is white noise, it should not exhibit any autocorrelation and *DW* will tend to 2, since $r = 0$. In practice, the autocorrelation experiences a drop around the optimal cutoff frequency, without necessarily reaching a null value.

Monitoring the *DW* statistic, which ranges between 0 and 4, there is low probability of reaching the ideal $DW = 2$ (denoting null autocorrelation). In order to measure the autocorrelation, the following expression:

$$r^2 \approx \frac{(2 - DW)^2}{4} \tag{5}$$

was found more convenient in practice. If the signal pre-treatment assures that the first and last points in the chromatogram match ($d_1 = d_\text{n}$), Eq. (5) will be exact (i.e., not an approximation).

Fig. 3a plots the autocorrelation measured as r^2 (estimated from the noise **e**) versus the cutoff frequency in a logarithmic scale, for the analysis of the extracts of the three samples described above (Figs. 1 and 2). This figure should be analysed together with Fig. 4, which shows the baselines obtained by BEADS corresponding to the cutoff frequencies marked in Fig. 3a, where several regions can be observed. When the selected cutoff frequency is too low (point (1)), nearly flat baselines are subtracted that scarcely affect the autocorrelation (only the vertical shift of the whole chromatogram is corrected, which is insufficient in most situations). Intermediate cutoff frequencies tend to eliminate several contributions of the baseline at large scale (points (4) and (5)). In this region, the cutoff frequency will be the ideal. Beyond these frequencies, BEADS tends to eliminate gradually the contribution of the analytes and baseline signals, attenuating the peaks (point (7)), until only noise remains (point (8)). At even higher cutoff frequencies, all contributions including the noise would be eliminated, leaving a null vector (in this case, r^2 cannot be calculated, because Eq. (2) becomes undefined owing to the division by zero). In other words, the baseline to be subtracted would be equal to the raw signal. The observation of the results in Figs. 3a and 4 leads to the conclusion that the observed minimum in the autocorrelation plot points out the optimal cutoff frequency (that one giving rise to the best baseline subtraction), which is specific for each sample.

When the raw signal is processed with BEADS, small variations in the selected cutoff frequency may be translated into large variations in the baseline (see Fig. S6 in the Supplementary material). This behaviour is observed especially at intermediate frequencies, below the optimal one. In addition, cutoff frequencies above the optimal (frequencies (6) to (8)) make the baseline undesirably sen-

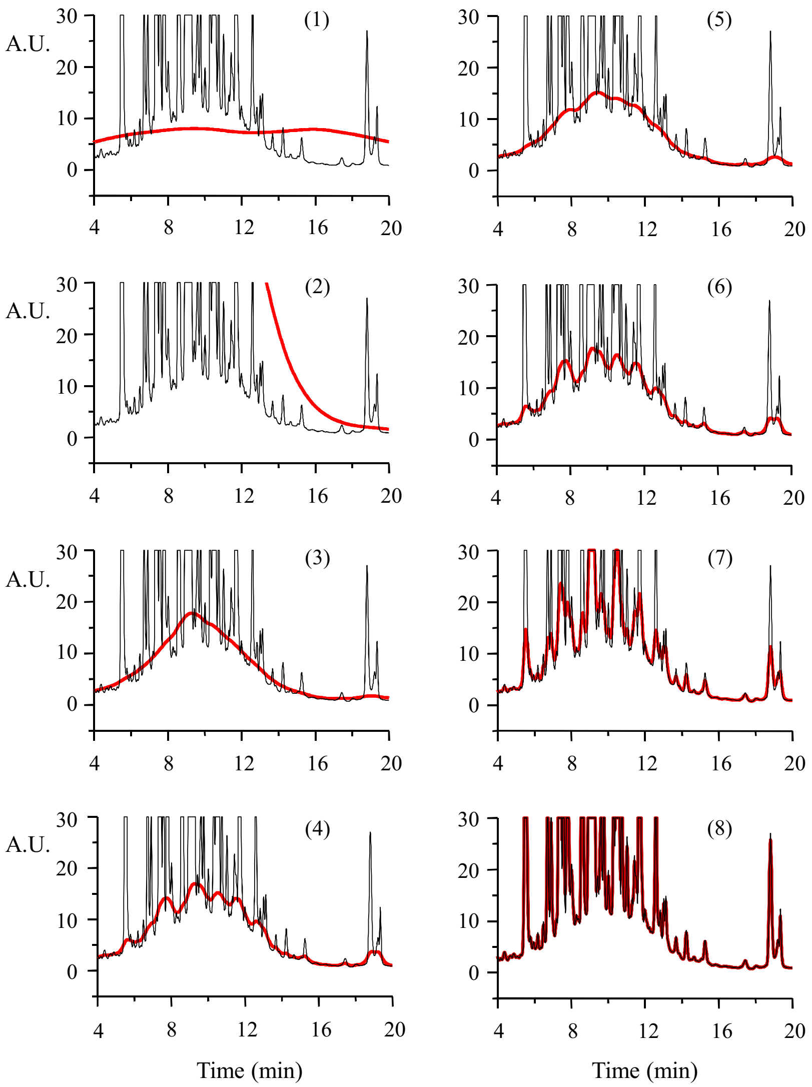

Fig. 4. Exploration of the cutoff frequency used to subtract the baseline from the chromatogram in Fig. 1a (horsetail tea), using the original scale. The frequency values correspond to the points marked in Fig. 3a. The estimated baseline has been overlapped on an enlarged section of the chromatogram.

sitive to the peak magnitude. The described disturbances can be cancelled, at least to some extent, by a careful adjustment of the asymmetry and regularization parameters.

To sum up, each feature eliminated from the chromatogram at a certain cutoff frequency results in a domain of characteristic r^2 values, starting by a value close to one when no contribution has been removed yet, up to values close to zero when even the noise has been removed. However, as observed for points (4) to (6) in Fig. 4, even at the best cutoff frequencies, some irregularities (ripples under the main peaks) remain in the baseline. This problem is addressed in Section 3.3.2.

3.3. Enhancements in the application of BEADS

As commented, particular working parameters for each type of sample and signal are needed for the routine application of BEADS. The quality of the results depends critically on the experience and skill of the analyst. In addition, the process may become excessively slow and prone to subjectivity. On the contrary, using the proposed auxiliary autocorrelation plots, described in Sections 3.2 and 3.4, BEADS can be quite easily adapted to any kind of sample, reducing the subjectivity in the selection of the working parameters, and providing always reliable results. In addition to the hard selection of the optimal parameters, other limitations of BEADS have been described, which make its practical use for baseline subtraction in complex chromatograms troublesome. Some proposals to overcome each limitation are indicated below.

3.3.1. Periodicity of the chromatogram

The correct application of BEADS requires periodic signals: if the signal values at the extremes of the chromatogram differ, artefacts will appear. In a first step, we considered solving the requirement of periodicity at the extremes of the chromatogram, through the subtraction of the straight-line that connects the first and last points. However, some problems may appear when the slopes at both extremes differ. Only a careful trial-and-error adjustment of the regularization parameters can mitigate this. We found that a more practical solution was the subtraction of a parabola, since it is able to fully cancel incidental differences in the slopes at the start and end of the chromatogram (see Fig. S7 in the Supplementary material).

BEADS is based on the use of high pass filters, which allow all features above a critical selected frequency survive (sparse chromatogram and noise), whereas the lower frequency features are cancelled (baseline, and any added feature to correct the periodicity problem, such as the parabola). The process of correction implies the treatment of a distorted signal (**y'**) with BEADS, where

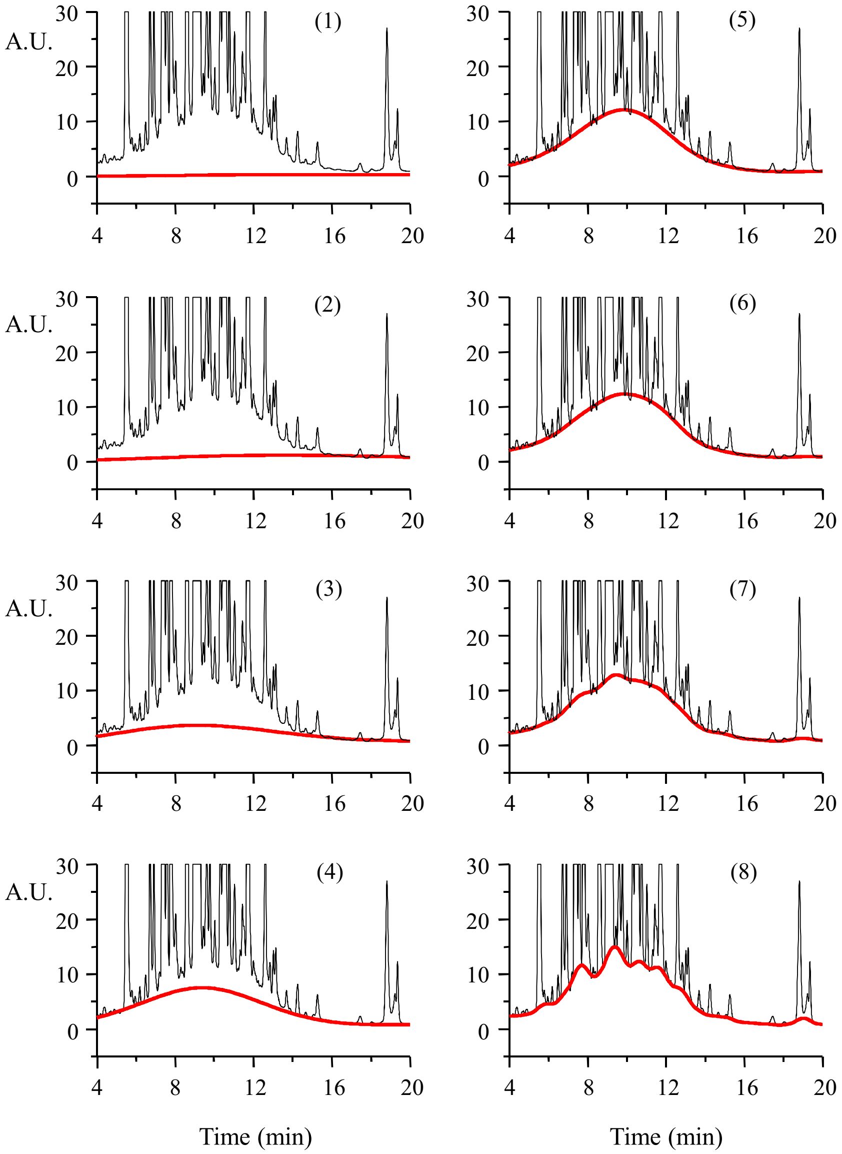

Fig. 5. Exploration of the cutoff frequency used to subtract the baseline from the chromatogram in Fig. 1a (horsetail tea), using the log transformation of the signal. The frequency values correspond to the points marked in Fig. 3b. The estimated baseline has been overlapped on an enlarged section of the chromatogram.

a parabola has been subtracted to the raw signal in order to make the slopes at the extremes identical. In these conditions and as a result of the high-pass filter, BEADS will give a correct estimation of the sparse chromatogram (**c**) and noise (**e**), but a biased baseline (**b'**):

$$\mathbf{y}' = \mathbf{c} + \mathbf{b}' + \mathbf{e} \tag{6}$$

The correct baseline ($\mathbf{b}_{\text{corr}}$) can be easily recovered by adding the parabola (**p**) previously subtracted:

$$\mathbf{b}_{\text{corr}} = \mathbf{b}' + \mathbf{p} = \mathbf{b}' + (\mathbf{y} - \mathbf{y}') \tag{7}$$

3.3.2. Chromatograms involving peaks with extremely different magnitude

As we showed in Fig. 4, chromatograms with extreme differences in peak size give rise to ripples when processed by BEADS. To eliminate the influence of the highest peaks on the baseline in such chromatograms, there are at least two solutions. The most straightforward treatment is clipping the highest peaks, so that the signal cannot exceed a selected height. We have explored this strategy with chromatograms of diverse complexity. Clipping works fine with relatively simple chromatograms, but for complex chromatograms with bulky baselines, the ripples remain (see Fig. S8 in the Supplementary material).

The second solution is using a log transformation of the signal, which is compatible with the operations of the original BEADS algorithm. The signal is transformed to the logarithmic scale after subtracting its minimal value, slightly increased with an arbitrary positive offset, ε:

$$\mathbf{z} = \log(\mathbf{y} - \min(\mathbf{y}) + \varepsilon) \tag{8}$$

The larger the offset, the less aggressive the pre-treatment. We decided to use an offset $\varepsilon = 1$. This value is appropriate regarding the magnitude of the signals being processed, which reach maxima around 500–10,000. Another reason for selecting $\varepsilon = 1$ is because if $y_i = \min(\mathbf{y})$, then $\log(y_i - \min(\mathbf{y}) + 1) = \log 1 = 0$.

The log transformation reduces the weight of the largest peaks along the BEADS operation, and as a result, the ripples of the baseline that appear under the main peaks totally disappear. Since the magnitude of the ripples in the baseline under the peaks is correlated to the size of the signals (the higher the peak, the higher the ripple), by operating in the logarithmic scale, the ripples will only be perceptible at very high frequencies. Naturally, after applying BEADS to the log transformation, the results should be back-transformed to the original scale.

The decomposition of the raw signal (**y**) and the log transformed signal (**z**), using BEADS, can be denoted as:

$$\mathbf{y} = \mathbf{c}_\mathbf{y} + \mathbf{b}_\mathbf{y} + \mathbf{e}_\mathbf{y} \tag{9}$$

$$\mathbf{z} = \mathbf{c}_\mathbf{z} + \mathbf{b}_\mathbf{z} + \mathbf{e}_\mathbf{z} \tag{10}$$

Considering Eq. (8), the back transformation of the BEADS results obtained in the logarithmic scale can be carried out as follows:

$$\mathbf{b}_{\mathrm{corr},\mathbf{y}} = 10^{\mathbf{b}_\mathbf{z}} + \min(\mathbf{y}) - \varepsilon \tag{11}$$

$$(\mathbf{c}+\mathbf{e})_{\mathrm{corr},\mathbf{y}} = \mathrm{y} - 10^{\mathbf{b}_\mathbf{z}} - \min(\mathbf{y}) + \varepsilon \tag{12}$$

Even if the correct BEADS working parameters are used, it will be not possible to differentiate the contributions of the sparse chromatogram and noise, once the log transformation has been applied, because it affects the sparsity of the signal and its derivative, and prevents the persistence of linearity (Eq. (9)), once returned to the original scale. This has another consequence: the best cutoff frequency cannot be selected from the noise. As will be shown in Section 3.4, it can be still obtained from $\mathbf{y} - \mathbf{b}_{\mathrm{corr},\mathbf{y}}$.

It should not be forgotten here that the objective is to correct the raw signal by removing the baseline. This is obtained by subtracting from the raw signal the back transformed baseline.

3.3.3. Sporadic negative signals

BEADS can be applied in different ways for the processing of asymmetrical signals (chromatograms with sporadic negative peaks). A first possibility is taking advantage of the symmetry parameter to set the level of tolerance to negative signals in the sparse chromatogram, using trial and error. This treatment is slow and has no guarantee of success. We propose an alternative, which consists in running BEADS repeatedly, in iterations, using a fixed value of the symmetry parameter suitable for positive signals. Along the iterations, and after each BEADS evaluation, those points below a certain threshold under the baseline are replaced by the corresponding values of the baseline found in the current iteration. This iterative replacement process is repeated until convergence, or up to a given maximal number of iterations is fulfilled. Proceeding in this way, not only the problems associated with the negative signals were eliminated, but also spurious contributions (which break the general trends) disappeared. This process does not affect the peak heights.

3.4. Autocorrelation plot using the baseline-corrected signal

As a consequence of the log transformation, the noise returned by BEADS cannot be used to estimate the autocorrelation. Instead, the signal corrected by subtracting the returned baseline can be used to monitor the changes in the baseline:

$$\mathbf{y}_{\mathrm{bcorr}} = \mathbf{c} + \mathbf{e} = \mathbf{y} - \mathbf{b} \tag{13}$$

Even though it is not possible to obtain an unbiased estimation of the noise in the original scale, Eq. (2) can still be applied to estimate the autocorrelation, by taking d_i as the difference between the $\mathbf{y}_{\mathrm{bcorr}}$ signal for points i and $i-1$. Therefore, the consistency of the variations around point i in a window of three points ($i-1$, i and $i+1$) is monitored. This means that there is no proper residual for making the comparison, and $\mathbf{y}_{\mathrm{bcorr}}$ includes a correlated contribution (the sparse chromatogram, **c**). However, as will be shown below, monitoring the autocorrelation of the baseline corrected signal can still be useful to set the best working parameters in BEADS.

Indeed, we have found that a plot of $r^2_{\mathbf{y}-\mathbf{b}}$ (Eq. (5) applied to the log transform), as a function of the cutoff frequency and considering the full chromatogram vector (Fig. 3b), is very useful to detect the most appropriate cutoff frequency. This plot should be compared with the plot in Fig. 3a, where the autocorrelation corresponds to the noise ($r^2_\mathbf{e}$), without applying any transformation to the data. Both plots show different patterns that depend on the use of the original scale (Fig. 3a), or the log transformation (Fig. 3b), and on the kind of data from which the autocorrelation is measured: the noise (Fig. 3a) or the baseline corrected chromatogram (Fig. 3b). When the noise is processed, the plot exhibits a minimum at intermediate cutoff frequencies (see Section 3.2), whereas the use of the baseline corrected chromatogram leads to a stepped plot. The value of $r^2_{\mathbf{y}-\mathbf{b}}$ decreases as the diverse baseline contributions to the chromatogram are removed. Each horizontal region in the plot corresponds to a consistent baseline returned by BEADS in a given frequency interval. When the contributions of the peaks of analytes, baseline and noise disappear completely, the autocorrelation of the residuals should be ideally $r^2_{\mathbf{y}-\mathbf{b}} = 0$. We have observed from a collection of chromatograms that the optimal cutoff frequency is close to the centre of the last step at higher frequencies, that is, around the last inflection point (point (6) in Fig. 3b). In practice, it is convenient to select slightly lower cutoff frequencies (i.e., a point between the beginning and the centre of the last horizontal region in the autocorrelation plot), to attenuate somehow the flexibility of the baseline.

Fig. 5 illustrates the impact of the cutoff frequency on the baseline subtraction, for the chromatogram of the horsetail tea extract (Fig. 4 corresponds to the results obtained with BEADS with the original scale, see also Fig. 3a). The baselines found by BEADS using different cutoff frequencies are overlapped on the chromatograms. For the same cutoff frequency, the baselines calculated from the direct signal (Fig. 4), and its log transformation (Fig. 5), do not match. The baselines found using the log transformation of the signal were much more satisfactory, for all assayed chromatograms. Also, an important aspect to remark is that the selection of the cutoff frequency becomes less critical when the signal is translated to the logarithmic scale. The cutoff frequencies (1) to (4) marked in Fig. 3b are too low, while frequencies (7) and (8) overfit the baseline (i.e., unreal ripples appear under the peaks). For frequencies (5) and (6), the baseline can be considered highly satisfactory.

The fine adjustments of the other working parameters (asymmetry, r, and regularization parameters, λ_0, λ_1 and λ_2) used in BEADS are given in the Supplementary material (Figs. S1–S5). As observed, in all instances, stepped plots are obtained and the optimal parameter value is close to an inflection point. However, by adjusting only the cutoff frequency after setting approximate values for the remaining parameters, highly satisfactory results were found in all assayed cases by operating with the logarithm of the signal.

3.5. Application of the assisted BEADS

Fig. 6 illustrates the baseline found after selecting the optimal cutoff frequency, using the log transformation of the signal, for full chromatograms of the three samples of medicinal herbs (Figs. 1 and 2). Fig. 7 shows the corresponding final baseline-corrected chromatograms. The result is highly satisfactory in all instances.

Fig. 8 shows the chromatogram of a mixture of sulphonamides (see Section 3.3.3), eluted with a gradient of acetonitrile from 0 to 20% (v/v), reaching the upper concentration in 30 min in the presence of 0.01 M Brij-35. The separation was carried out using a Chromabond C18 column (150 mm × 4.6 mm I.D., 5 μm particle diameter, Scharlab). In the chromatogram, a refractometric per-

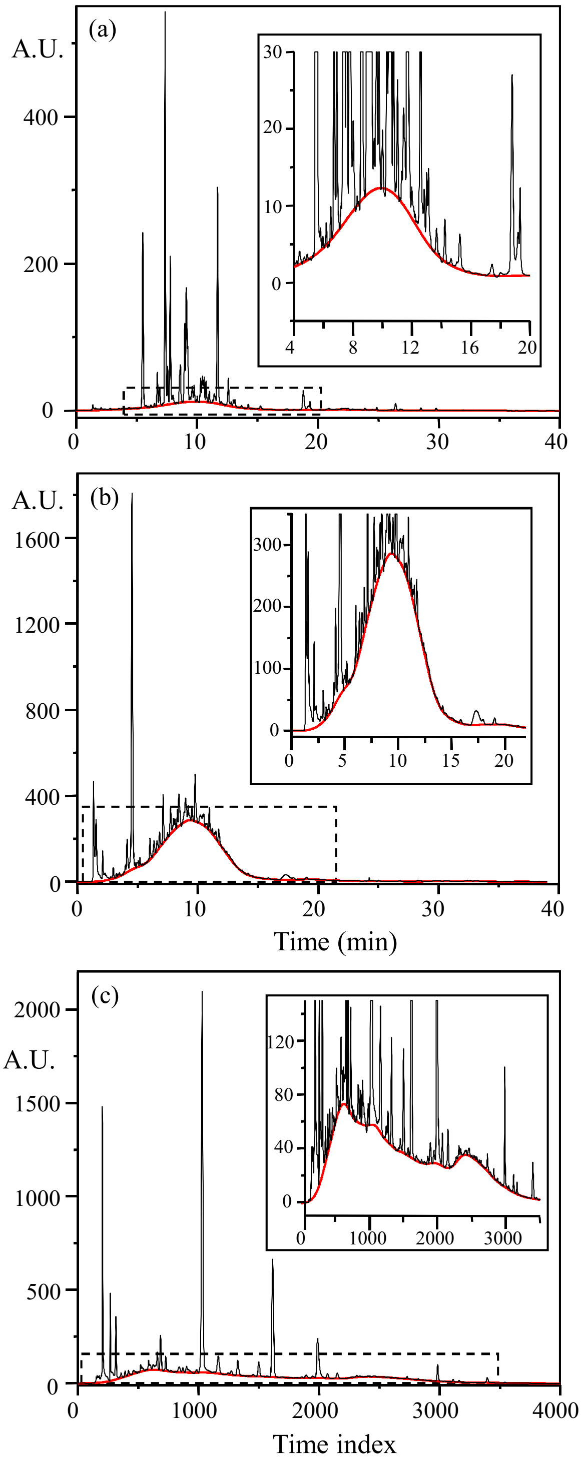

Fig. 6. Chromatographic fingerprints of medicinal herbs: (a) horsetail tea, (b) decaffeinated tea, and (c) extract of red peony root taken from Ref. [7], with the optimal baseline overlapped, using the assisted BEADS algorithm. Cutoff frequency: (a) 0.105 (see also Fig. 3b), (b) 0.132 and (c) 0.130. The upper inserts magnify the central regions of the chromatograms to allow a better inspection.

turbation associated with the mixing of the sample and mobile phase appears close to the void volume. Fig. 8a shows the baseline in successive iterations, where the points below the negative threshold are replaced by the respective predicted baseline points. In Fig. 8b, the baselines to be subtracted according to the original BEADS and applying the proposed approach are overlapped. The original BEADS required a modification of all working parameters

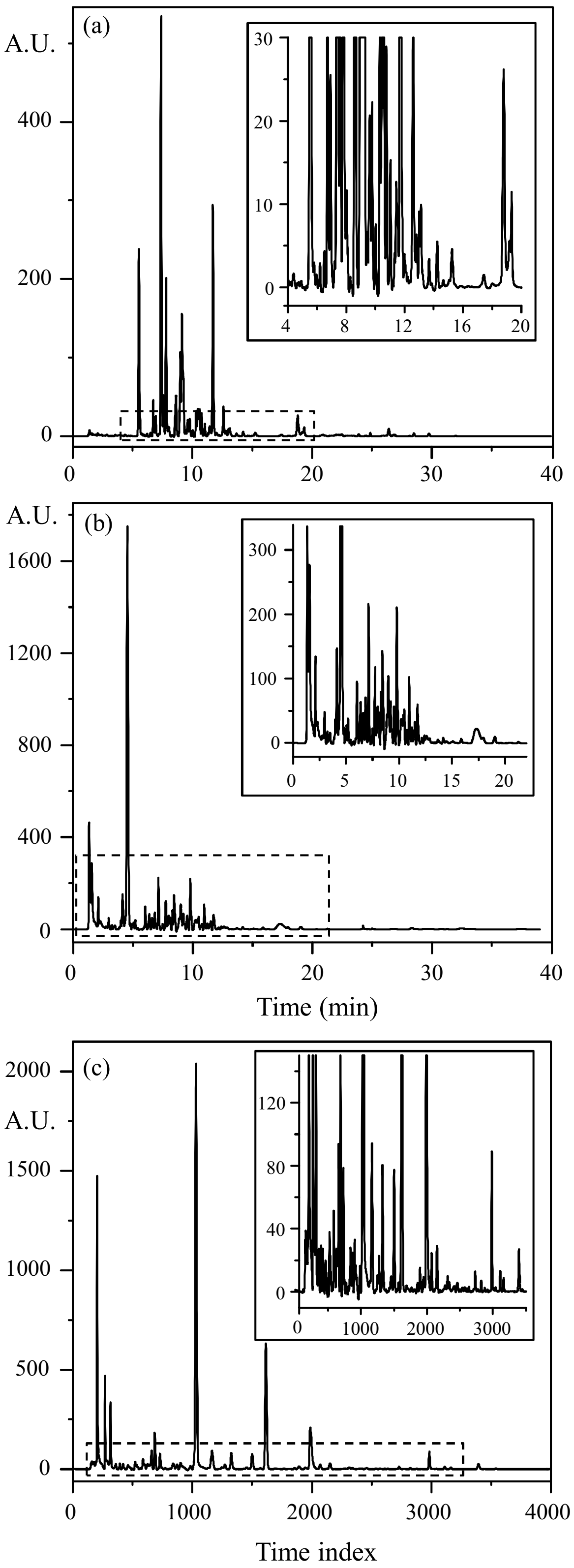

Fig. 7. Baseline corrected chromatograms for (a) horsetail tea, (b) decaffeinated tea, and (c) extract of red peony root (see Fig. 6 for the unprocessed signals and the found baselines). The upper inserts magnify the central regions of the chromatograms to allow a better inspection.

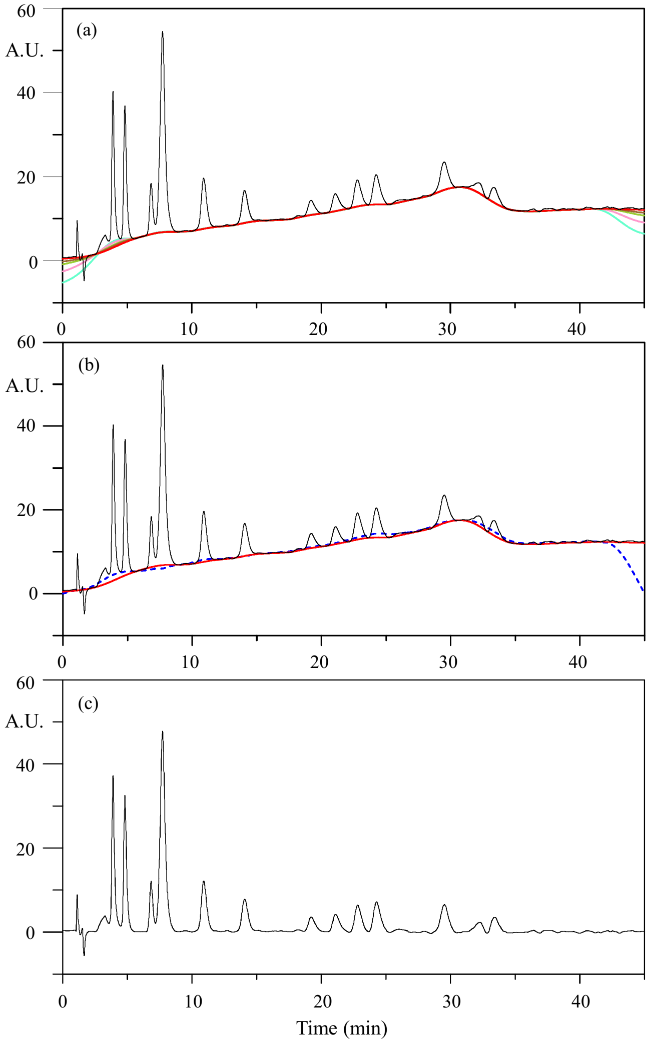

Fig. 8. Chromatogram showing refractometric negative peaks, corresponding to the elution of a mixture of 15 sulphonamides, using gradient elution with acetonitrile in the presence of Brij-35: (a) progress of the iterations showing the successive baselines up to reach convergence, (b) baseline obtained using the iterative substitution (continuous line) versus that one obtained with the original BEADS (dashed line). (c) Final chromatogram after baseline subtraction using the iterative substitution. In spite of the presence of a negative signal, the same value of symmetry parameter was used as in Figs. 6 and 7 (which showed only positive peaks).

by trial and error, and the compensation of the negative signal was less perfect. Fig. 8c shows the baseline-corrected chromatogram according to the proposed approach.

3.6. *Quantification*

Appraising properly the consequences of a global baseline correction on peak quantification is not easy, since they depend on a number of factors difficult to parametrize. For instance, the results depend on the mutual magnitude of the peaks and the size and frequencies of the baseline fluctuations. Thus, when a poor baseline is subtracted, the recovery error in a large signal can be much smaller than the corresponding error for a small signal with a good baseline correction. Other factors affecting the results are the peak location in the chromatogram (e.g., in an empty region or at the extremes of the chromatogram), the surroundings of the peak to be quantified (e.g., an isolated peak, a peak in a cluster or a peak in the neighborhood of a major constituent), the presence of noise or negative signals, among others.

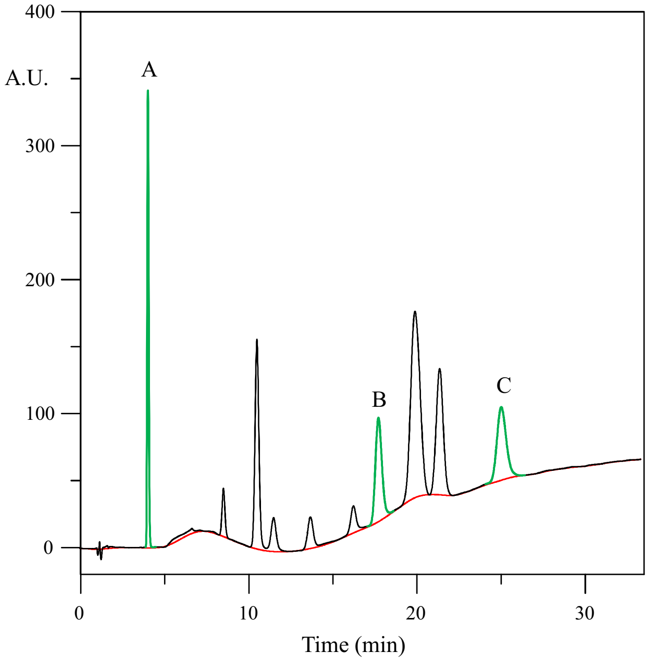

Fig. 9. Chromatogram of a sample containing 7 sulphonamides, eluted with Brij-35. The original signal is drawn in black, and the respective baseline corrected using the assisted BEADS is overlaid in red. Peaks A, B and C (marked in green) were added in independent artificial experiments (see text for more details). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 9 gives an idea of the errors that could be expected after BEADS baseline subtraction. The figure shows the chromatogram of 7 sulphonamides, eluted with Brij-35, overlaying the corresponding found baseline. Three artificial peaks (marked as A, B and C) were added in independent in silico experiments, for calculating the errors. The three peaks had the same area (35.00 units), asymmetry factor (1.23) and plate count (8700). For each peak, the chromatogram corresponding to the 7 sulphonamides plus the added peak were processed by BEADS to recover the baseline. The recovered area was then calculated after subtracting the overall signal and the baseline for the respective peak (global baseline correction). In addition, the area obtained by fitting the local baseline around each peak in the respective global chromatogram was also calculated (local baseline correction).

The relative errors for the three peaks (global and local corrections) were: Peak A (0.11%, 0.014%), peak B (2.5%, 3.7%), and peak C (4.9%, 4.0%). As expected, the magnitude of the errors is correlated with the retention time, since the peaks become wider and the weight of the area of the residual signals in the original chromatogram under the peak of interest is increased. Also, the relative errors obtained with a local baseline, which only considers the surroundings of the peak of interest, are usually smaller, since fitting a global baseline implies losing details in particular regions of the chromatogram. In spite of this, the magnitude of the errors is comparable for the global and local baselines.

4. Conclusions

The main problem of applying BEADS is the need to set properly the parameters for each specific signal. A correct setup of the BEADS working parameters in its original formulation is difficult to establish, particularly the cutoff frequency, which is by far the most critical working parameter. This work proposes an auxiliary autocorrelation plot to assist in the selection of the optimal cutoff frequency, which is also valid for adjusting the other working parameters. The irregularities in the baseline associated to large differences in scale between major and trace components (i.e., baseline ripples appearing under the main peaks) are solved by replacing the raw signal by its log transformation.

With the assisted BEADS, the selection of the optimal frequency is less critical. The subtraction of the baseline using straightforwardly BEADS requires some experience and a selection of the working parameters by trial and error, owing to the mutual dependence and sensitivity among them. In contrast, the use of the autocorrelation plot and the log transformation allows a fast, simple and reliable selection of the cutoff frequency and other working parameters. The third improvement is an iterative algorithm that discards sporadic negative signals breaking the general trend of the baseline, such as refractometric peaks or transitions associated to gradients.

Our long-term aim is the optimization of the separation conditions for complex samples, such as chromatographic fingerprints, whose baselines are notably irregular. The origin of these problematic baselines is the complexity of the matrix, together with the use of gradients to expedite the analyses. The evaluation of such chromatograms forced to search a method capable of adjusting very complex baselines. Ideally, the method should be reliable and require few or no user interaction. The assisted BEADS provided very satisfactory results in all assayed examples (about 65 chromatograms), and needs little supervision.

The decomposition of the net signal in sparse chromatogram, baseline and noise shows a certain level of mutual dependence, so that the net chromatogram has peaks significantly smaller, even after a correct baseline subtraction. Thus, for certain applications, such as the quantification of peaks, processing the net chromatogram is risky, and the noise can be overestimated in regions of the chromatogram where peaks are found. Therefore, it is preferable to subtract only the baseline and process the resulting signal by other methods able to eliminate the noise, such as the Savitsky-Golay smoother [14]. The whole process, from loading signals to obtaining the final table of results takes a few seconds.

BEADS, in its original formulation, was also suggested for signals of other nature, such as electrocardiograms. Therefore, the tools developed in this work may also improve signals coming from fields different from chromatography. It should be also mentioned that some of the proposed solutions are also valid for other baseline subtraction algorithms. Thus, for instance, the autocorrelation plots can be useful for configuring other parametric baseline estimation approaches.

As commented, BEADS may suffer from transient artifacts at signal end-points. These periodicity errors can be solved by reformulating the filter used in the original BEADS algorithm, as recently proposed by Selesnick [15].

Acknowledgements

This work was supported by Projects CTQ2016–75644-P and MTM2014–54388 (Ministry of Economy and Competitiveness, MINECO, Spain, and FEDER funds), and PROMETEO/2016/128 (Direcció General d'Universitat, Investigació i Ciència, Generalitat Valenciana, Spain). Sergio López-Ureña thanks the Ministry of Education and Culture and Sports, MECD of Spain for the FPU14/02216 grant. José Antonio Navarro thanks the Generalitat Valenciana for a pre-doctoral grant, associated to Project PROMETEO/2016/128.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.chroma.2017.05.057.

References

[1] S. Fanali, P. Haddad, C.F. Poole, M.L. Riekkola, Liquid Chromatography: Fundamentals and Instrumentation, 2nd ed., Elsevier, Amsterdam, 2017.
[2] A. Felinger, Data Analysis and Signal Processing in Chromatography, Elsevier, Amsterdam, 1998.
[3] Chemometrics in Chromatography, in: Ł. Komsta, Y. Vander Heyden, J. Sherma (Eds.), CRC Press, New York, 2017.
[4] X. Ning, I.W. Selesnick, L. Duval, Chromatogram baseline estimation and denoising using sparsity (BEADS), Chemometr. Intell. Lab. Syst. 139 (2014) 156–167.
[5] https://in.mathworks.com/matlabcentral/fileexchange/49974-beads–baseline-estimation-and-denoising-w–sparsity–chromatogram-signals-?requestedDomain=www.mathworks.com
[6] V. Mazet, C. Carteret, D. Brie, J. Idier, B. Humbert, Background removal from spectra by designing and minimizing a non-quadratic cost function, Chemometr. Intell. Lab. Syst. 76 (2005) 121–133.
[7] Z.M. Zhang, S. Chen, Y.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (2010) 1138–1146.
[8] J.J. de Rooi, P.H.C. Eilers, Mixture models for baseline estimation, Chemom. Intell. Lab. Syst. 117 (2012) 56–60.
[9] M. Dumarey, I. Smets, Y. Vander Heyden, Prediction and interpretation of the antioxidant capacity of green tea from dissimilar chromatographic fingerprints, J. Chromatogr. B 878 (2010) 2733–2740.
[10] J.C. Cobas, M.A. Bernstein, M. Martín Pastor, P.G. Tahoces, A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data, J. Magn. Reson. 183 (2006) 145–151.
[11] P.H.C. Eilers, Parametric time warping, Anal. Chem. 76 (2004) 404–411.
[12] I.E. Frank, R. Todeschini, The Data Analysis Handbook, Elsevier, Amsterdam, 1994.
[13] J. Durbin, G.S. Watson, Testing for serial correlation in least squares regression. III, Biometrika 58 (1971) 1–19.
[14] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[15] I. Selesnick, Sparsity-assisted signal smoothing (revisited), IEEE Int Conf. Acoust., Speech, Signal Proc. (2017) (March).