Journal of Chromatography A, xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Leveraging probabilistic peak detection to estimate baseline drift in complex chromatographic samples

Martin Lopatka[a,c,*], Andrei Barcaru[b], Marjan J. Sjerps[a,c], Gabriel Vivó-Truyols[b]

[a] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Postbus 94248, 1090 GE Amsterdam, The Netherlands
[b] Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Postbus 94248, 1090 GE Amsterdam, The Netherlands
[c] Netherlands Forensic Institute, Postbus 24044, 2490 AA Den Haag, The Netherlands

ARTICLE INFO

Article history:
Received 30 September 2015
Received in revised form
18 December 2015
Accepted 23 December 2015
Available online xxx

Keywords:
Baseline correction
Asymmetric least squares
Chromatography
Chemometrics
Peak detection
Data preprocessing

ABSTRACT

Accurate analysis of chromatographic data often requires the removal of baseline drift. A frequently employed strategy strives to determine asymmetric weights in order to fit a baseline model by regression. Unfortunately, chromatograms characterized by a very high peak saturation pose a significant challenge to such algorithms. In addition, a low signal-to-noise ratio (i.e. $s/n < 40$) also adversely affects accurate baseline correction by asymmetrically weighted regression.

We present a baseline estimation method that leverages a probabilistic peak detection algorithm. A posterior probability of being affected by a peak is computed for each point in the chromatogram, leading to a set of weights that allow non-iterative calculation of a baseline estimate. For extremely saturated chromatograms, the peak weighted (PW) method demonstrates notable improvement compared to the other methods examined. However, in chromatograms characterized by low-noise and well-resolved peaks, the asymmetric least squares (ALS) and the more sophisticated Mixture Model (MM) approaches achieve superior results in significantly less time. We evaluate the performance of these three baseline correction methods over a range of chromatographic conditions to demonstrate the cases in which each method is most appropriate.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The use of chemometric methods for preprocessing chromatographic data has become a ubiquitous component in analytical chemistry. A common convention from the chemometric school of thought is to consider a chromatogram as being composed of three (additive) components, namely: signal, baseline, and noise [1,2]. An abundance of literature has been published regarding the preprocessing of chromatographic data [1–4] such that informative features can be extracted from the raw chromatographic data. This involves removing the perturbing components from the chromatographic signal, these artifacts are baseline drift and noise. A typical objective of data preprocessing methods is to arrive at a set of peak areas corresponding to particular compounds of interest, which forms the basis of subsequent multivariate data analysis.

Chromatographic baseline drift impedes the accurate quantification and interpretation of analytical data. The most widespread approaches use an asymmetrically weighted least squares regression procedure to determine the best fit baseline model [1–3,5–8]. For asymmetric weighting, a prerequisite condition is the determination of a series of weights intended to emphasize the effect of points belonging to baseline regions while suppressing the influence of points affected by peaks [9–11]. The probability that a point belongs to the baseline is used to minimize a penalty function of asymmetrically weighted deviations from a baseline of variable smoothness. Many strategies exist to estimate what points belong to the baseline, comprehensive summaries are given in [10,6]. Including several non-parametric methods for baseline correction [12,6].

We note that despite their widespread use, asymmetrically weighted least squares regression approaches often fail when chromatograms get very dense [11] or very noisy [5,13]. Even more recent work [14], using a Mixture Model formulation to assign points to a baseline component have difficulty when peak density becomes very high. Unfortunately, this is often the case with real chromatography; realistic estimates of component separations in chromatographic systems have suggested that for real applications, peak co-elution (i.e. regions of high density) are unavoidable [15,16]. We present a new method for the estimation of baseline

* Corresponding author.
E-mail addresses: m.lopatka@uva.nl (M. Lopatka), a.barcaru@uva.nl (A. Barcaru), m.j.sjerps@uva.nl (M.J. Sjerps), g.vivotruyols@uva.nl (G. Vivó-Truyols).

http://dx.doi.org/10.1016/j.chroma.2015.12.063
0021-9673/© 2016 Elsevier B.V. All rights reserved.

Please cite this article in press as: M. Lopatka, et al., Leveraging probabilistic peak detection to estimate baseline drift in complex chromatographic samples, J. Chromatogr. A (2015), http://dx.doi.org/10.1016/j.chroma.2015.12.063

which leverages a probabilistic approach to peak detection [17]. This method shows promising results, especially in cases with high peak density and a low signal to noise ratio. Results are shown for a range of chromatographic saturations [15] and chromatograms containing a variety of peak heights, from very near the noise level to 100 times greater.

2. Theory

The asymmetric least squares solution (ALS) proposed by Eilers and Boelens [9,10] strives to estimate a smoothed signal z of length n and sampled uniformly. The estimation of z strives to balance two conditions: consonant with y (the raw measurement from the chromatograph) having the same dimensionality as z and greater smoothness characteristics. These two characteristics are balanced by an additional parameter λ, which must be tuned by hand, however likely varies in the range $10^2 \leq \lambda \leq 10^9$. Ultimately, the ALS method also requires the introduction of a weight vector w, having the same dimension as the signal y. This strategy has inspired many variations [5,14,18] on determining the weight vector w required to estimate a smooth baseline that allows points unaffected by peaks to exert a greater influence on the resulting curve. In this paper we will compare the asymmetric least squares solution (ALS) [10], a subsequently published Mixture Model based estimation (MM) of these weights [14], and finally our own method, using a probabilistic peak detection [17] strategy, henceforth abbreviated as PW for *peak weighted*.

In the following sections, the main idea of each method is explained, adhering as much as possible to the original notation used in published materials. The original cited sources should be consulted for in-depth explanation. All three methods approach the problem of assigning a probability that a particular point belongs to the baseline. The PW method may be seen as the most sophisticated due to complexity of the approach leading to the determination of this probability (denoted as p). The MM model relies on the Expectation Maximization (EM) algorithm to arrive at this posterior probability, denoted r in later sections. The ALS method may be seen as coarsely approximating this probability in terms of a high or low value, its quantity denoted by the parameter ρ introduced later. The computation of these probabilities sits at the crux of the differences observed in the performance between the methods and is further explored throughout this manuscript.

2.1. Weights calculated by ALS method

The original asymmetric least squares approach (ALS) relies on a minimization of the objective function $S_\lambda(y, z, w)$ such that z does not deviate from the original signal y to a great extent, while exhibiting a greater degree of smoothness than y. This concept of smoothness is expressed as:

$$\Delta^2_{z_i} = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) \tag{1}$$

$\Delta^2_{z_i}$ [1] calculated for $i = 3, 4, \ldots n$ to avoid inaccuracies at the beginning of the signal. The objective function $S_\lambda(y, z, w)$ in Eq. (2) is defined as

$$S_\lambda(y, z, w) = \sum_i w_i (y_i - z_i)^2 + \lambda \sum_i (\Delta^2_{z_i})^2 \tag{2}$$

An iterative solution is proposed in [10] that must give considerably more weight to values that lie below the trend line (since peaks will lie above the trend line). So iterative reweighing is performed such that $w_i = \rho$ if $y_i > z_i$ and $w_i = 1 - \rho$ otherwise. Here ρ is a scalar, user defined parameter where $0 < \rho < 1$. This introduces a degree of coarseness to the weights, as each element in w may only assume one of two values based on the choice of ρ. Finally, the values for the baseline trend are expressed in Eq. (3).

$$(W + \lambda D^T D)z = Wy \tag{3}$$

where $W = diag(w)$ and $D = I \circ (\Delta^2_z * (\Delta^2_z)^T)$, these can get quite large as w has the same dimensionality as the chromatographic signal. The baseline values z can then be solved explicitly via a Cholesky factorization.

2.2. Weights calculated by MM method

The Mixture Model method (MM) [14], relies on a calculation of the posterior probability that a point in the chromatogram belongs to the baseline. Points (y_i) associated with the baseline are assumed to be drawn from the normal density function $g(y_i|\mu, \sigma)$ where μ denotes the mean and σ, the standard deviation. Those points contained in chromatographic peaks are assumed to follow an unknown probability density $h(y_i - \mu)$. Therefore, for every point y_i in the chromatogram a probability for belonging to the baseline can be calculated as shown in Eq. (4).

$$r_i = \frac{\pi g(y_i|\mu, \sigma))}{(\pi g(y_i|\mu, \sigma) + (1 - \pi)h(y_i - \mu))}, \tag{4}$$

where μ is the mean unknown background level, σ is the unknown standard deviation and π is the unknown prior mixing proportion $\pi \in [0, 1]$. To estimate the components of the mixture, the authors used the Expectation Maximization (EM) algorithm. In the *E step*, the current values of the parameters are used to calculate the posterior probabilities (r_i in Eq. (4)) for baseline and peaks of each point in the chromatogram. Further, in the *M step*, the parameters r_i, μ, σ, and the estimate for $h(\cdot)$ are updated, given the calculated probabilities. The baseline estimate μ is modeled using P-splines (i.e., penalized B-splines). In this implementation the objective function S_λ is minimized:

$$S_\lambda(y, \alpha) = ||y - B\alpha||^2 + \lambda ||D_d \alpha||^2, \tag{5}$$

where y is the original signal, B is the basis splines matrix with dimensionality $(n \times m)$, containing m splines for the n data points present in chromatogram. The coefficients α (an $m \times 1$ vector) are fit, λ is a penalty parameter, and D_d is a matrix containing the coefficient of the dth order differencing operator. For a summarization of the signal with precisely 7 splines $m = 7$ and $d = 3$, the matrix D_3 can be written as follows:

$$D_3 = \begin{pmatrix} -1 & 3 & -3 & 1 & 0 & 0 & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -3 & 1 & 0 \\ 0 & 0 & 0 & -1 & 3 & -3 & 1 \end{pmatrix}$$

Posterior probabilities of belonging to the baseline are now introduced in Eq. (6) as weights, where r_i is an $n \times 1$ vector. The introduction of posterior probabilities means that the objective function gets modified into the following form:

$$S^*_\lambda = (y - B\alpha)^T R(y - B\alpha) + \lambda ||D_d \alpha||^2, \tag{6}$$

with $R = diag(r_i)$. Hence, the solution for the coefficients is:

$$\hat{\alpha} = (B^T R B + \lambda D_d^T D_d)^{-1} B^T R y \tag{7}$$

Here we used $d = 3$ as recommended in [14]. Low values of r_i indicate little influence on the baseline. This method is similar to the ALS method with the exception of the probabilistic weights in the matrix R. The MM algorithm uses the ALS approach to calculate initialization values for the EM algorithm.

[1] $\Delta^2_{z_i}$ term notation is adopted from original publication [10], is not a squared term.

Please cite this article in press as: M. Lopatka, et al., Leveraging probabilistic peak detection to estimate baseline drift in complex chromatographic samples, J. Chromatogr. A (2015), http://dx.doi.org/10.1016/j.chroma.2015.12.063

Download English Version:

https://daneshyari.com/en/article/7610368

Download Persian Version:

https://daneshyari.com/article/7610368

Daneshyari.com