Journal of Chromatography A, 1431 (2016) 122–130

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 30 December 2015

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.

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/

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.

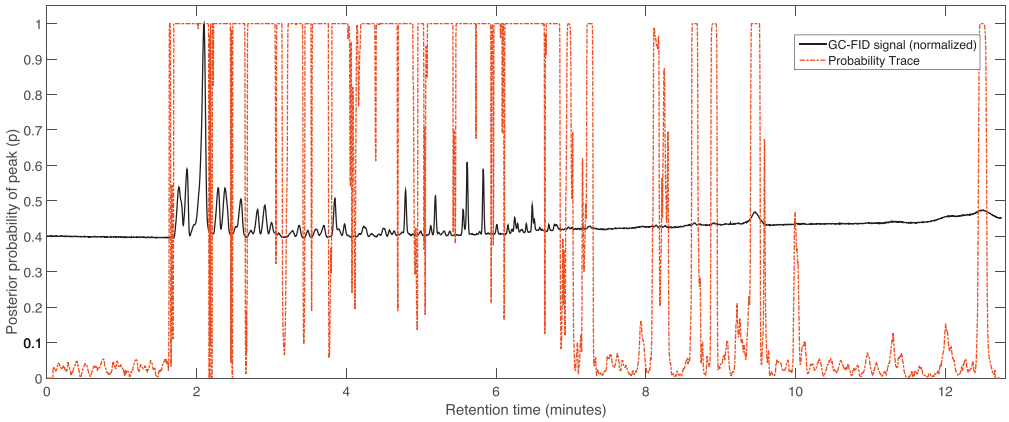

Fig. 1. Example probability trace over chromatogram.

2.3. Weights calculated by PW method

Previously published work [17] has addressed the calculation of a posterior probability that any point in a chromatogram is affected by a peak. This method for probabilistic peak detection requires an exhaustive set of models to be defined over a set window of data points. The window size is set to $2n+1$ points where n is the specified peak width σ_{peak} expressed in terms of number of data points. The set of models each take the form shown in Eq. (8)(below). These models are evaluated as to their suitability in describing the data (k, y_k), with $k \in [i-n, i+n]$ and y_k being the intensity of the point k.

$$y_k = a + bk + c_{i-n}g_{i-n}(k) \ldots + c_{(i+n)}g_{(i+n)}(k) + \epsilon_k. \tag{8}$$

This general model consists of a linear term to account for baseline drift and one or more Gaussian terms to account for the presence of chromatographic peaks within the scope of the window centered at i where,

$$g_j(k) = \begin{cases} \dfrac{1}{\sigma_{peak}\sqrt{2\pi}} e^{-(k-j)^2/2\sigma^2_{peak}}, & \text{if there is a peak centered at point } j \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

The parameters c in Eq. (8) affect the peak height and are also fit by least squares, but the peak centers are fixed in terms of their index j into the window k. An exhaustive set of peak configurations is defined for a given window. Exhaustive in the sense that all points inside the window are candidate peak center locations, and more than one peak center is allowed within the chromatographic window, in which case every possible combination of peak center locations is represented by one distinct model. This set of models are all fit by least squares to the chromatographic region. The necessary parameters for this method may be estimated visually from a chromatographic signal. The width of an ideal peak can be counted in terms of points to arrive at σ_{peak} and σ_ϵ may be calculated from a blank chromatographic run. The maximum number of peaks allowed per window must be bound for computational reasons, this is further discussed in Section 3.3. Robustness to parameter estimation errors is discussed in the original publication [17].

The residuals observed after fitting the model set are expected to be normally distributed with a mean of 0 and standard deviation equal to that of the intensities of baseline points (σ_ϵ). A likelihood can then be assigned to each model based on the product of its residuals evaluated against $\mathcal{N}(0, \sigma_\epsilon)$. Combining the resulting likelihoods with prior probabilities derived from the statistical theory of component overlap [15] is achieved using Bayes' rule. The weight vector required for penalized least squares regression as per Eq. (2) is directly taken as $1-p$, where p is the posterior probability of a point being affected by a chromatographic peak, as calculated by the method specified in [17]. In this way more weight is placed on points deemed to have a low probability of being affected by a chromatographic peak. In this case the values for p (substituted into Eq. (3) as the diagonal elements of W) are not iteratively updated but rather provided as appropriate weights. While very similar to the Mixture Model method described in Section 2.2, this method may be expected to yield more precise posterior probabilities due to the exhaustive set of peak configurations considered. Note that the number of possible models will depend on the width of chromatographic peaks and the number of co-eluting peaks allowed in the model construction; both of these are user-provided parameters. The baseline estimation problem may be directly solved by Eq. (3), once again requiring λ to moderate between the smoothness of the resulting baseline and its conformity to the observed chromatogram. This method is substantially more computationally intensive than the MM approach. Furthermore, in order to yield an advantage in accurately modelling complex patterns of co-eluting peaks the exhaustive model set must include models representing numerous peak configurations.

Fig. 1 shows an example of the probability trace calculated over a chromatographic signal that contains substantial peak coelution. This figure is shown here illustratively, later results directly compare the weights computed by each method.

3. Materials and methods

3.1. Simulated chromatograms

In order to perform a quantitative assessment of the performance of the three methods for baseline estimation we require a known ground truth for the baseline component. Since this is impossible for real analytical data, we simulate representative chromatographic signals conforming the assertions discussed in Section 1 regarding the composition of chromatographic signals. Chromatograms with a fixed length of 1000 points are simulated by the addition of the three components via Eq. (10). The desired chromatographic saturation can be achieved by setting the number of peaks desired (t) according to the following equality: $\alpha = \frac{t}{1000/peak\,width}$, where α is the saturation of a chromatogram as defined in [15]. The locations of the t peaks are randomly drawn from a uniform distribution over the range 1:1000. Peak heights v are sampled from a normal distribution $v \sim \mathcal{N}(50, 250)$. This creates a matrix A with t rows and 1000 columns, each row containing one Gaussian peak centered at a random column index. The noise addition component $\eta \sim \mathcal{N}(0, 5)$ is then added. The baseline term β is

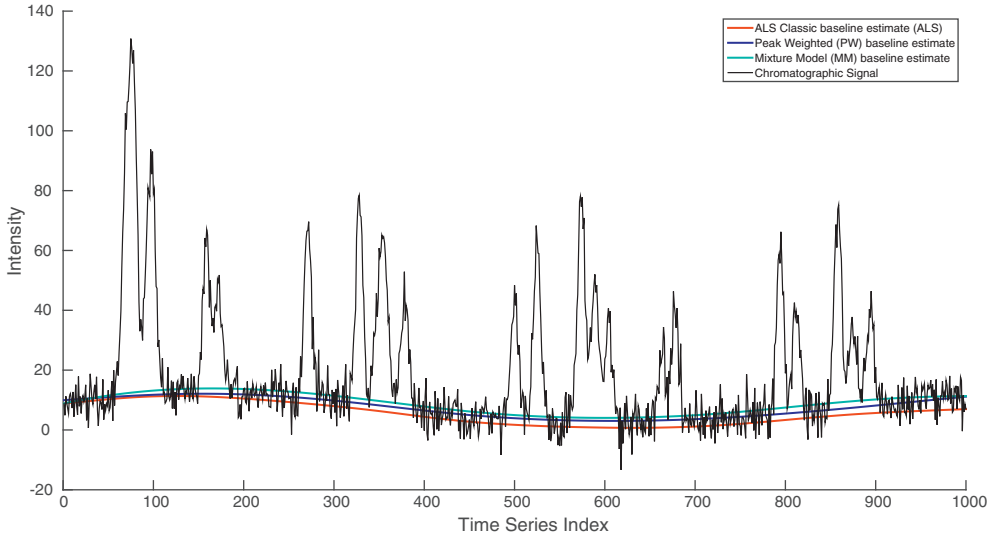

Fig. 2. Example baseline estimates. Simulation parameters: $\sigma_{peak\,width} = 4$, $\sigma_{noise} = 5$, mean $s/n = 10$, saturation, $(\alpha) = 0.5$.

generated by the following method: 5 seed points are draw from a uniform distribution $s_1, \ldots, s_5 \sim \mathcal{U}(-0.5\bar{v}, 0.5\bar{v})$, where $\bar{v}$ is the mean of peak height values (before the addition of noise). These points are randomly assigned to locations, again according to $\mathcal{U}(1, 1000)$. To arrive at a baseline component with 1000 points, a 4th degree polynomial is fit to these 5 seed points and added to the simulated peaks and noise. Thus we obtain the chromatographic signal vector y:

$$y = \sum_{j=1}^{t} A_j + \eta + \beta \tag{10}$$

Fig. 2 in Section 4, shows an example of simulated chromatogram, generated by Eq. (10), and the results of three baseline estimation techniques performed on the simulated signal.

3.2. Real chromatograms

3.2.1. GC-FID screening of fire debris samples

A one dimensional GC-FID chromatogram typical of a screening method used for fire debris investigation is also examined. The sample itself is a mixture of common substrates typically found in residential structure fires (clothing, furniture textile, wood, paper, laminated particle board, painted gypsum board, carpet, and under carpet insulation) the sample was treated with 15 mL of gasoline and allowed to burn for 8 min. This leads to the formation of many volatile pyrolosis products. Analysis was performed from a 0.5 mL headspace injection of the sample using an Agilent 6890N Gas Chromatograph with a Flame Ionization detector (FID). Injection temperature was set to 250°C. Constant flow rate was 2.0 mL/min with split ratio 20. The oven ramping program was 80 °C held for 2 min then 80–225°C at 40 °C/min and held at 225 °C for 5 min. Capillary column used was an Agilent DB-624 with the following dimensions: 30 m × 0.32 mm × 1.80 μm.

3.2.2. GCxGC-FID analysis of ignitable liquids

A comprehensive chromatographic separation (GCxGC-FID) typical of fire debris investigation for ignitable liquid identification was also examined. The complexity of the matrix in this cases necessitated a second dimension separation to fully resolve all peaks. Nonetheless, trace compounds occurring in low abundance are important to the investigative objective for such samples. The system used was an Agilent Technologies 6890N with 1D column DB5 30 m × 0.25 μm × 0.25 μm × and 2nd dimension column: DB17 1 m × 0.25 μm × 0.25 μm × . Temperature programming was as follows: 40°C held for 0.2 min, ramping to 130°C at 2°C per minute, increase to 250°C at 8 degrees per minute, hold at 250°C for 5 min. Injection temperature and FID detector temperature were both set to 250°C. Modulation period was 4 s.

3.3. Computational considerations

All computations carried out in the course of this research were performed on an analytical computer with an Intel® i7 3.6 GHz CPU and 32Gb of installed system memory, running Microsoft© Windows 7 Professional and Matlab version 8.4.0.150421 (R2014b) 64-bit. Chromatographic signals were acquired using Leco ChromaTOF software version 4.32 and exported in comma-separated values (csv) format for subsequent import into Matlab. The computation of a baseline for a 1000 point simulated chromatogram were as follows: ALS – 0.007 s, MM – 0.08 s, PW 0.001s. These benchmarks were averaged from 100 repeated measurements of different 1000 point chromatograms. The actual probabilistic peak detection required to apply the PW method took an additional 12.4 s (on average). For a larger chromatogram i.e. 50,000 data points, corresponding to a 20 min run time with a detector operating at 40 Hz, the time required to perform the probabilistic peak detection increases to approximately 5.5 min, whereas the baseline estimation for the ALS and MM methods remain under 1 s. For the PW method, these processing times are related to the size of the exhaustive model set (described in Section 2.3) and grow exponentially with the number of coeluting peaks allowed in the model set and with expected peak width. As will be demonstrated in the following results, the efficacy of each baseline correction method will vary depending on certain characteristics of the chromatography. An important practical consideration is whether the nature of the chromatographic signal warrants the use of the PW method over the much faster alternatives (ALS and MM), which may yield comparable performance in certain situations.

4. Results and discussion

4.1. Initial observations

Initial exploration into baseline correction on experimental data showed seemingly comparable performance for MM method

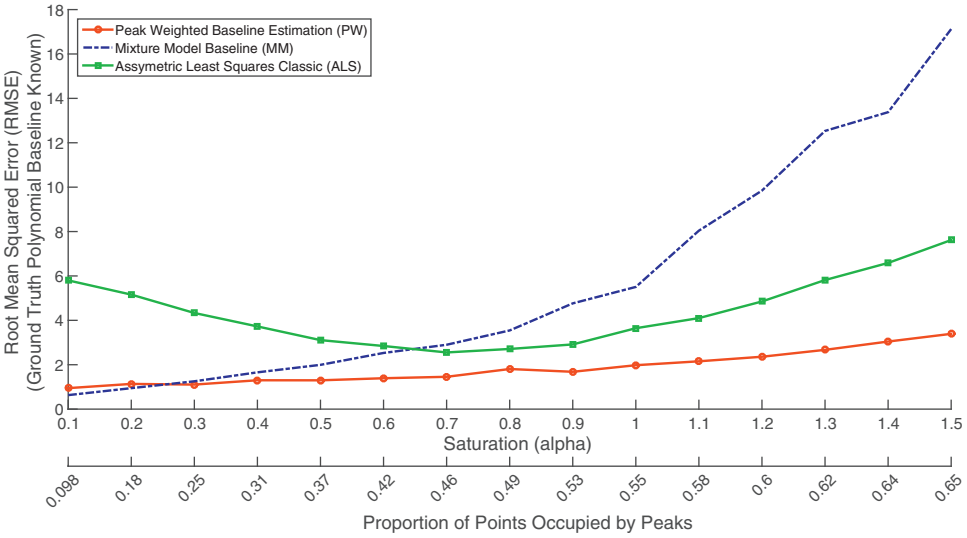

Fig. 3. Mean performance of three methods on 100 simulated chromatograms per saturation conditions. RMSE calculated versus known polynomial baseline.

and PW methods. Both provided baseline estimates deemed more desirable (by an expert analyst) than those achieved using the ALS method. However, small differences were observed in regions where a high density of (partially) coeluting peaks were seen. This effect was exacerbated when the signal to noise ratio was fairly low (i.e. $s/n < 40$).

In order to thoroughly investigate the particular chromatographic conditions in which the methods diverge in their performance, the saturation of simulated chromatograms was varied over a range deemed realistic (α values from 0.1 to 1.5) for routinely observed chromatography. Furthermore, the parameter for additive noise was set in order to achieve a range of peak heights where the signal to noise ratio remained around 50.

Results are also presented for an analytical domain in which chromatography is frequently encountered with a high peak density and fairly low signal to noise ratio, forensic fire debris analysis. For such analyses accurate baseline correction is crucial for the detection of ignitable liquid residue. Small peaks near the baseline noise, corresponding to low abundance compounds, are often the most interesting. Unfortunately, these peaks are the most sensitive to variations in baseline estimation.

4.2. Simulated chromatograms

Using the chromatogram simulation procedure described in Section 3.1, 100 simulated chromatograms are generated for each saturation interval over the range [0.1, 0.2, . . ., 1.5]. These 100 chromatograms are each subjected to baseline estimation by all three of the candidate methods, ALS, MM, and PW. The ground truth baseline term β can then be compared to the estimated baselines. The root mean squared error (RMSE) is calculated over the chromatographic range.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(z_i - \beta_i)^2}{n}}, \tag{11}$$

where n is the number of individual data points in the chromatogram, z is the estimated baseline by whichever method is being evaluated and must have exactly n points, β is the polynomial baseline that was added when the simulated chromatogram was generated by Eq. (10), as such is considered the ground truth. Fig. 2 shows an example of the estimated baselines for each method over a simulated chromatogram with moderate saturation and Fig. 3 shows the performance of each baseline estimation method as a function of the varying chromatographic saturation. Fig. 4 shows the variation observed between specific chromatograms simulated with a common saturation level.

4.2.1. Parameter selection

The three methods compared each rely on the tuning of particular parameters. To perform this tuning, an arbitrary moderate saturation chromatogram was used ($\alpha = 0.8$). For the ALS method the tuning strategy proposed in the original publication [10] was used to arrive at the parameters $\lambda = 10^{6.95}$ and $\rho = 0.04$. For the MM method, parameter tuning was manually performed by an expert analyst following the guidelines suggested in the original publication [14]. The ground truth was unknown to the expert in order to reduce bias in determining an acceptable baseline. The PW method operates in two steps, first the probabilistic peak detection is deployed to generate a probability trace over the chromatogram [17], this process requires 4 parameters which were set to $\sigma_{peak} = 4$, $\sigma_{noise} = 5$, $n = 3$, and $\alpha = 0.37$. The choice of a fixed α parameter of 0.37, this particular value was suggested in earlier work on statistical theory of component overlap [15]. In the second step a baseline is calculated based on the posterior probabilities, for this step the same value for λ was used for the PW baseline detection as was with the ALS method. A trend in the performance of each method is clearly visible as the chromatograms become more saturated and fewer points belong only to the baseline. The ALS method struggles to maintain flexibility as points may only be assigned a weight of either ρ or $1 - \rho$. This method performs a fixed number of iterative point-wise weight assignments coupled to the evaluation of the objective function given in Eq. (2). The two parts of this function (smoothness and adherence to the original signal) are modulated by the lambda parameter, no explicit assumption regarding the difference between peak and baseline regions is made. The MM and PW methods explicitly model the distributions for the *noise* and *peak*.

Differences in performance of ALS are demonstrated by the decrease in efficiency observed at both high and low saturation levels in Fig. 3. The optimal parameter combination in terms of smoothness and adherence to the original signal is shown to be sensitive to the density of peaks. As the number of peaks increases, points assigned a regression weight of $1 - \rho$ may be erroneously assigned a weight of ρ. Likewise, as the peak density decreases we observe the opposite kind of erroneous weighting. Points likely to be erroneously weighted are often very small peaks or high points

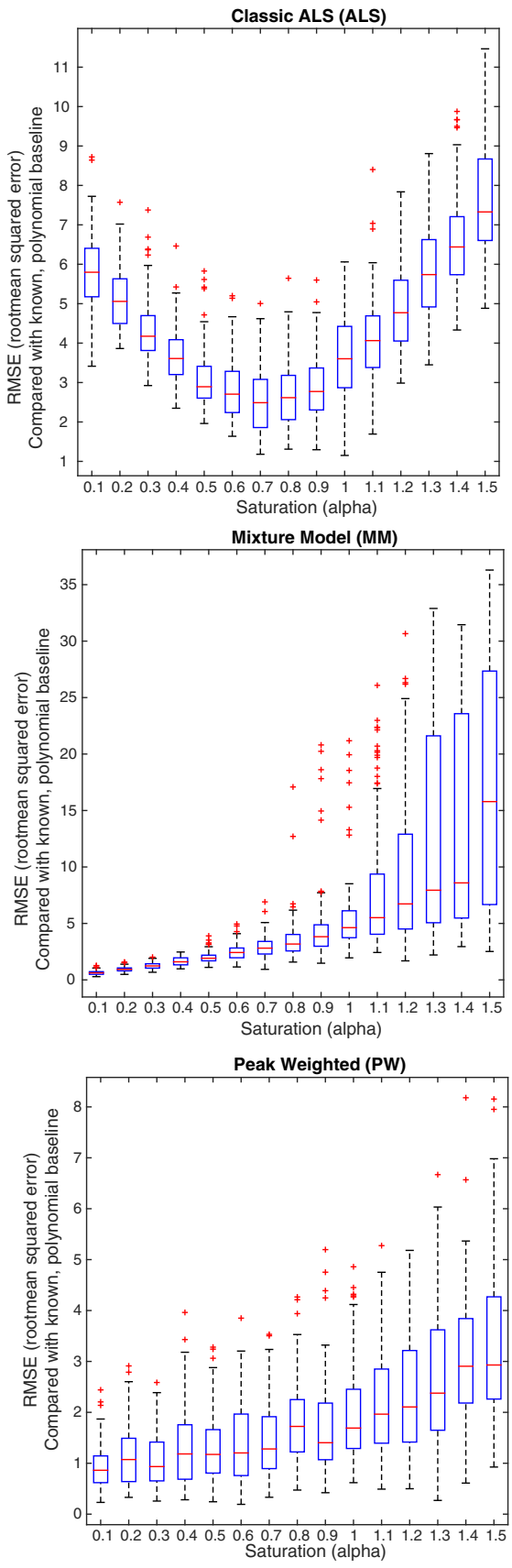

Fig. 4. Boxplots showing the variation within simulation conditions, horizontal red lines indicate median values over 100 simulations, blue box edges indicate the 25th and 75th percentile, and whiskers reach most extreme points not considered outliers.

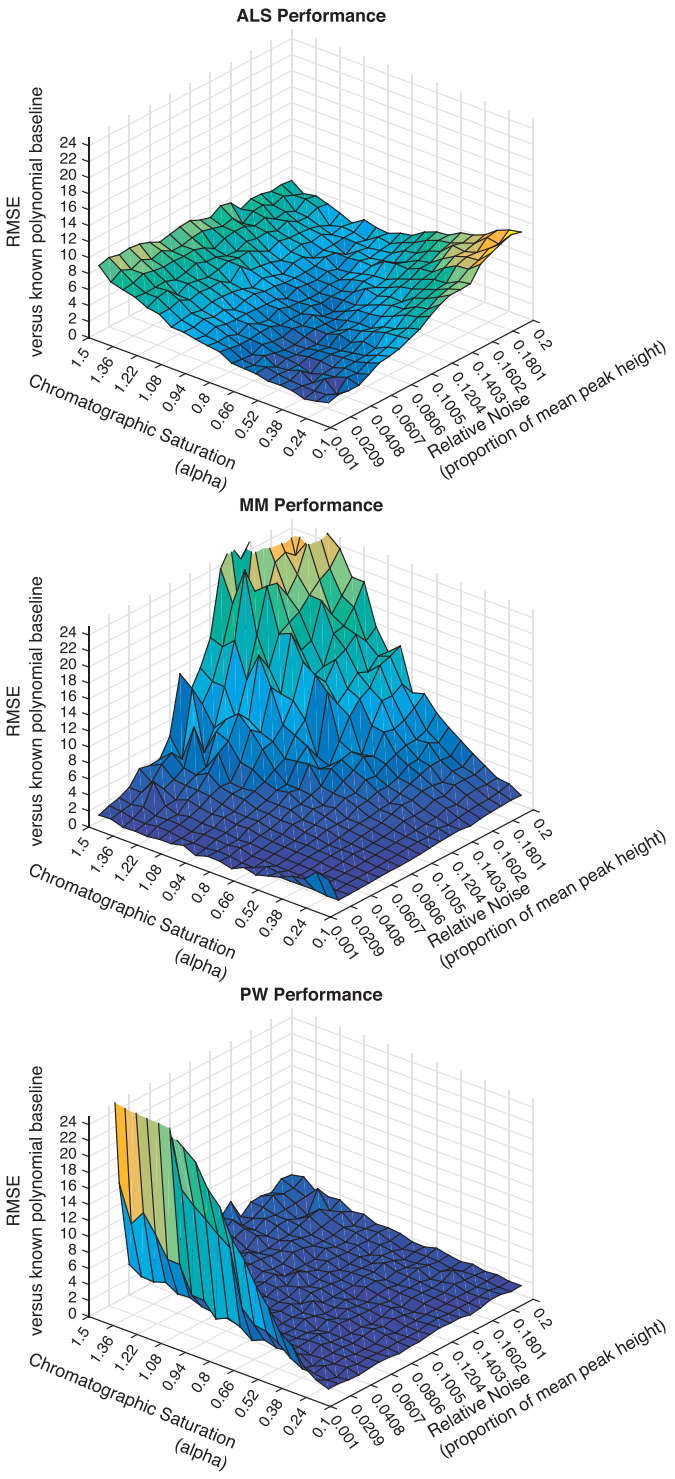

Fig. 5. Comparison of performance versus known polynomial baseline as simulated chromatograms vary in signal-to-noise ratio and chromatographic saturation (α).

in baseline noise. Nonetheless, the ALS method shows consistent performance across the entire range of saturations. The best performance is seen at $\alpha = 0.8$ where approximately 50% of the points in the chromatogram are unaffected by peaks. The coarse reweighing strategy is well suited for this case as it will naturally assign half the points a weight of ρ and the other half $1 - \rho$.

We observe a divergence between the performance of the MM and PW methods as the saturation increases with the PW method outperforming MM except in very sparse chromatograms. These

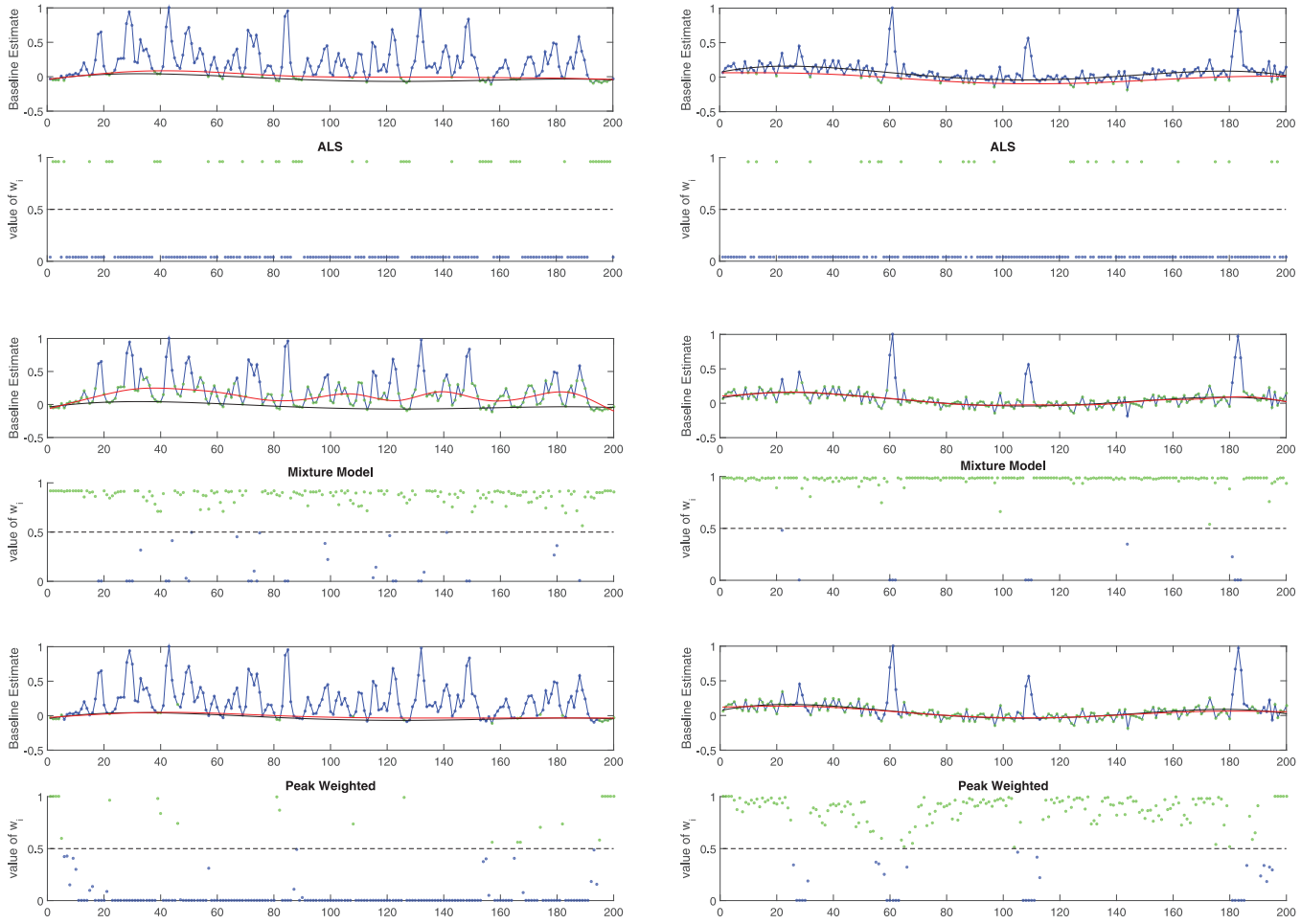

Fig. 6. Weight vectors for different approaches to baseline correction. Left column: High peak density example chromatogram ($\alpha = 1.5$). Right Column: Sparse chromatogram example ($\alpha = 0.1$), points colored based on thresholding at 0.5 green colored points have the greatest influence on baseline shape. (For interpretation of reference to color in this figure legend, the reader is referred to the web version of this article.)

results may be explained due to the way in which point-wise weights are assigned based on underlying *noise* and *peak* models. The PW method determines the weight of a point by how well (in terms of residuals) the point and surrounding points can be described by a peak shape model, independent of the peak height. The Expectation Maximization algorithm at the heart of the MM method pools the intensity of points, and iteratively fits a two-component Mixture Model to the data. As such, the range of low-value points that may be small peaks or large magnitude noise points may also be weighted incorrectly.

The results depicted in Figs. 3 and 4 suggest that the variation of the baseline points may have an effect on the performance of these three methods. In order to explore the combined effects of chromatographic saturation and signal to noise ratio on accurate baseline detection, we repeat the experimental conditions that were used to generate Figs. 3 and 4, this time also varying the additive noise contribution in the simulated chromatograms. Fig. 5 shows the surfaces relating to RMSE observed over a range of chromatographic saturation and signal-to-noise ratio conditions. Interestingly, the performance of the PW method suffers under *low* noise conditions. Reflecting on the observations in Fig. 1, it can be noted that non-peak areas are given some posterior probability of belonging to a peak area. This effect is exacerbated when a very large number of models must be considered to account for the exhaustive set of possible co-elution phenomena. When sparse chromatograms are examined with a large model set, the minimum possible posterior probability is increased by the influence of the prior probabilities. This may explain the performance drop in the low-noise and sparse conditions, as the α parameter is not tuned.

Further interesting findings seen in Fig. 5 show that each method performs optimally in a specific range of the parameter space. The PW method is capable of the best baseline estimation for high-noise and relatively crowded chromatograms, precisely the region where the MM model shows the worst performance. The ALS method shows the most consistent performance, but overall lower in the regions where other methods exhibit their optimal results. These findings should of course be considered in light of the computational requirements of the different methods. The optimal method for baseline correction should be selected based on the characterization of the chromatography in terms of noise and peak density and the computational resources available versus the acceptable error margin.

In chromatograms with more or less isolated peaks, for example at $\alpha = 0.5$ where peak co-elution is quite unlikely (only approximately14% of peaks overlap with other peaks), the MM method and PW method produce comparable results. However the divergence in performance in more densely populated chromatograms is apparent. Here the EM algorithm responsible for assignment of weights, the maximization (M-step) is subjected to high errors when the distributions pertaining to *noise* and *peak* exhibit high overlap. This occurs more frequently as the signal-to-noise ratio decreases, since small peaks will tend to resemble the noise. Fig. 6 depicts the weight vectors calculated by the three methods for one

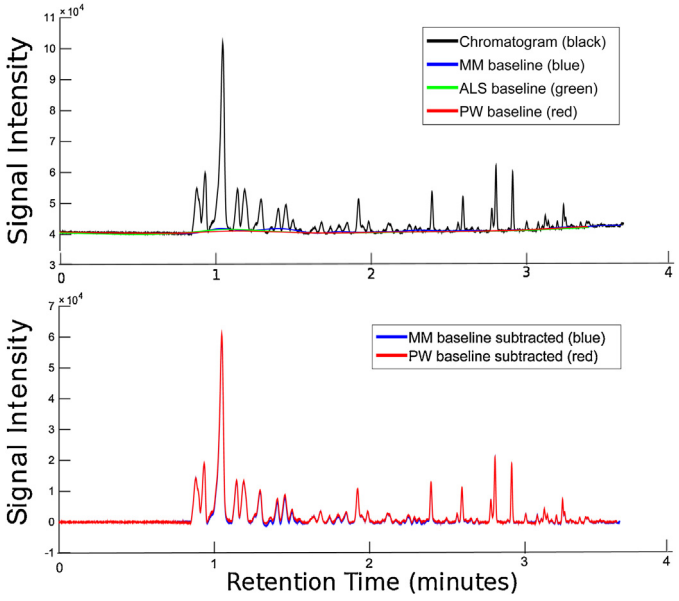

Fig. 7. Baseline correction performed on GC-FID separation of fire debris material.

example of a densely populated chromatogram and one example of a sparsely populated one.

Since the PW method does not couple the calculation of the weight vector w with the derivation of a baseline, the probabilistic method assigns a posterior probability invariant to peak height, as such it is possible that a small peak very near the noise may receive a very low weight in terms of its influence on the baseline. The densely populated example in Fig. 6(lower left) provides a good example as we see very few points are assigned a high weight for the baseline contribution. It is important to note that the chromatograms depicted in Fig. 6 are illustrative examples exhibiting characteristics that correspond to the extreme parameter values of the axes in Fig. 5. The inclusion of these examples may assist in the assessment of chromatographic conditions in which the PW and MM method may exhibit the most drastically different and most similar performance.

4.3. Real chromatograms

Chromatograms with very high peak density pose a problem for accurate baseline estimation since very few points belong exclusively to the baseline region. Likewise, a low signal-to-noise ratio means that differentiating baseline points from peak regions is complicated by their similarity. We examine the performance of both MM and PW methods for baseline correction on a GC-FID chromatogram, seen in Fig. 7. While the performance of the two methods seems comparable, closer inspection reveals some undesirable effects in particular regions. For example, examining the 7th, 8th, 9th, and 10th visibly describable peaks, we see that the MM model estimates a higher than appropriate baseline in this region due to a lack of non-peak points. In the same region, the PW method weights all points such that they do not influence the baseline estimation, resulting in a lower trend line and higher peak areas for those peaks.

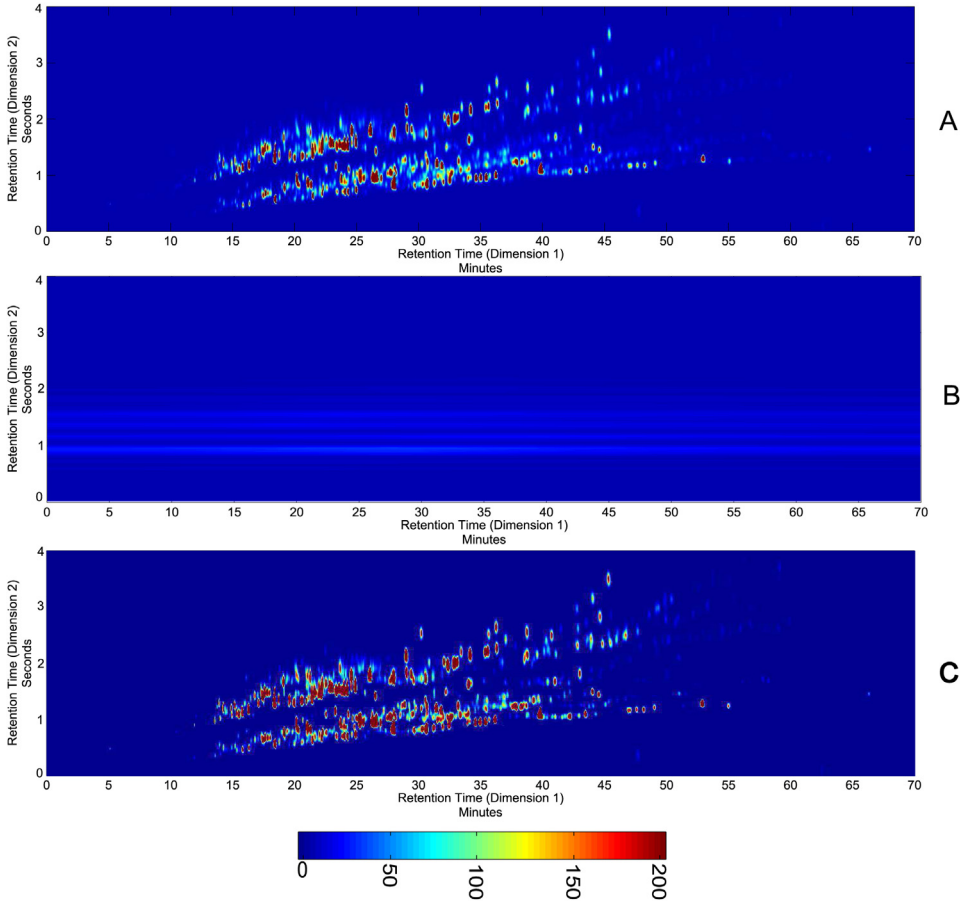

Fig. 8. Baseline correction performed row-wise on a comprehensive GCxGC-FID separation of fire debris material. (A) Original chromatogram, (B) PW estimated baseline, (C) corrected (baseline subtracted) chromatogram.

In comprehensive chromatographic separations baseline detection methods are largely adapted from one-dimensional methods [19,20]. The modulated nature of the signal may introduce several artifacts that complicate baseline estimation. The splitting and refocusing of each peak in a modulator leads to the introduction of smaller peak heights for compound portions occurring at relatively earlier and later modulations. We attempt to reduce the influence of baseline by operating on the comprehensive chromatogram orthogonally along the second dimension. This strategy has proven effective for similar analytical domains [21] where baseline *ridges* may appear as peaks in the modulated signal, but as gradual baseline in the second dimension. First the probabilistic peak detection is similarly performed orthogonally along the second dimension, then using the peak weighted (PW) method we subtract the baseline signal in the same manner. For real chromatographic data, the quality of the baseline correction can only be assessed qualitatively as no ground truth is known.

Fig. 8 shows a higher contrast between the small, late eluting peaks ($rt_1 > 45$ min) in the PW treated chromatogram compared to the original signal. The baseline estimate itself is plotted as well using the same color scale as the original signal, allowing inspection of the regions where significant baseline correction was performed. The region around $rt_1 = 50$ min and $rt_2 = 2.5$ s, shows clear increase in peak to noise contrast without the loss of small peaks for the PW method. Finally, peaks eluting in the highly saturated center of the chromatogram are visibly easier to deconvolve after PW baseline correction. While some research demonstrates natively two-dimensional base surface estimation, [22] the one dimensional methods used here were selected in order to maintain comparability with earlier results.

5. Conclusions

The original ALS algorithm [10] provided a basis for chromatographic baseline estimation and correction that drastically improved the quality of chromatographic data. Subsequent improvements [14] have emphasized the importance of accurately determining the weight that should be placed on each point in a chromatographic signal when a baseline is estimated. We demonstrate the application of a probabilistic peak detection method [17], decoupled from the baseline estimation problem, for this task. This method for peak detection evaluates an exhaustive set of peak co-elution models to determine the posterior probability that any point in a chromatogram is affected by a peak. These posterior probabilities are then used to compute a baseline estimate. Due to the increased computational demands of this approach it is important to note that comparable performance may be achieved using simpler methods when chromatographic peaks are more or less well resolved and a high signal to noise ratio is present. Improved performance is observed for chromatographic separations exhibiting high density of peaks and substantial peak co-elution. Furthermore, systems with a high dynamic range of peak heights or with a lower signal to noise ratio may require peak weighted baseline correction to arrive at an accurate baseline estimate. Peaks most susceptible to shape distortions and integration errors introduced by baseline estimation errors are those closest to the noise level. Any application domain where the accurate detection and integration of low abundance peaks is considered a high priority analytical may benefit from the PW approach described herein.

Acknowledgments

This research was funded by NWO grant number 727.011.006, under the research objectives of the COMFOR project. The authors acknowledge and appreciate the contribution of the Mixture Model baseline estimation code from Johan de Rooi and Paul Eilers, as well as their original work which inspired this approach. The authors also thank Eduard Derks (DSM resolve) for valuable discussion that was formative in the direction of this manuscript. The authors thank Andjoe Sampat and Brenda van Daelen for providing the GC-FID and GCxGC-FID chromatograms presented in Section 4.3.

References

[1] M. Daszykowski, B. Walczak, Use and abuse of chemometrics in chromatography, TrAC Trends Anal. Chem. 25 (11) (2006) 1081–1096.
[2] D.W. Cook, S.C. Rutan, Chemometrics for the analysis of chromatographic data in metabolomics investigations, J. Chemomet. 28 (9) (2014) 681–687.
[3] S. Castillo, P. Gopalacharyulu, L. Yetukuri, M. Orešič, Algorithms and tools for the preprocessing of LC–MC metabolomics data, Chemomet. Intell. Lab. Syst. 108 (1) (2011) 23–32.
[4] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A 1255 (2012) 3–11.
[5] P.G. Stevenson, X.A. Conlan, N.W. Barnett, Evaluation of asymmetric least squares baseline algorithm through the accuracy of statistical peak moments, J. Chromatogr. A 1284 (2013) 107–111.
[6] L.G. Johnsen, T. Skov, U. Houlberg, R. Bro, An automated method for baseline correction, peak finding and peak grouping in chromatographic data, Analyst 138 (2013) 3502–3511.
[7] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan, Asymmetric least squares for multiple spectra baseline correction, Anal. Chim. Acta 683 (2010) 63–68.
[8] S. He, W. Zhang, L. Liu, Y. Huang, J. He, W. Xie, P. Wu, C. Du, Baseline correction for raman spectra using an improved asymmetric least squares method, Anal. Methods 6 (2014) 4402–4407.
[9] P.H.C. Eilers, A perfect smoother, Anal. Chem. 75 (14) (2003) 3631–3636.
[10] P.H. Eilers, Parametric time warping, Anal. Chem. 76 (2) (2004) 404–411.
[11] D. Chang, C.D. Banack, S. Shah, Robust baseline correction algorithm for signal dense NMR spectra, J. Magn. Resonan. 187 (2007) 288–292.
[12] B.D. Prakash, Y.C. Wei, A fully automated iterative moving averaging (aima) technique for baseline correction, Analyst 136 (15) (2011) 3130–3135.
[13] X. Liu, Z. Zhang, Y. Liang, P.F. Sousa, Y. Yun, L. Yu, Baseline correction of high resolution spectral profile data based on exponential smoothing, Chemomet. Intell. Lab. Systems 139 (2014) 97–108.
[14] J.J. de Rooi, P.H. Eilers, Mixture models for baseline estimation, Chemomet. Intell. Lab. Sys. 117 (2012) 56–60.
[15] J.M. Davis, K.C. Giddings, Statistical theory of component overlap in multicomponent chromatograms, Anal. Chem. 55 (1983) 418–424.
[16] M. Martin, D.P. Herman, G. Guiochon, Probability distributions of the number of chromatographically resolved peaks and resolvable components in mixtures, Anal. Chem. 58 (1986) 2200–2207.
[17] M. Lopatka, G. Vivó-Truyols, M.J. Sjerps, Probabilistic peak detection for first-order chromatographic data, Anal. Chim. Acta 19 (817) (2014) 9–16.
[18] S.J. Baek, A. Park, Y.-J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, Analyst 140 (2015) 250–257.
[19] J.T. Matos, R.M. Duarte, A.C. Duarte, Trends in data processing of comprehensive two-dimensional chromatography: state of the art, J. Chromatogr. B 910 (2012) 31–45.
[20] K.M. Pierce, J.C. Hoggard, R.E. Mohler, R.E. Synovec, Recent advancements in comprehensive two-dimensional separations with chemometrics, J. Chromatogr. A 1184 (112) (2008) 341–352.
[21] M.R. Filgueira, C.B. Castells, P.W. Carr, A simple, robust orthogonal background correction method for two-dimensional liquid chromatography, Anal. Chem. 84 (15) (2012) 6747–6752.
[22] , in: Mixture models for two-dimensional baseline correction, applied to artifact elimination in time-resolved spectroscopy, Anal. Chim. Acta 771 (2013) 7–13.