Adaptive Asymmetric Least Squares baseline estimation for analytical instruments

Sergio Oller-Moreno*†‡, Antonio Pardo‡, Juan Manuel Jiménez-Soto†‡, Josep Samitier‡§ and Santiago Marco†‡
*Contact email: soller@ibecbarcelona.eu
†Signal and Information Processing for Sensing Systems, Institute for Bioengineering of Catalonia. Barcelona, Spain
‡Dept. Electronics, University of Barcelona, Barcelona, Spain
§Nanobiotechnology, Institute for Bioengineering of Catalonia. Barcelona, Spain

***Abstract*—Automated signal processing in analytical instrumentation is today required for the analysis of highly complex biomedical samples. Baseline estimation techniques are often used to correct long term instrument contamination or degradation. They are essential for accurate peak area integration. Some methods approach the baseline estimation iteratively, trying to ignore peaks which do not belong to the baseline. The proposed method in this work consists of a modification of the Asymmetric Least Squares (ALS) baseline removal technique developed by Eilers and Boelens. The ALS technique suffers from bias in the presence of intense peaks (in relation to the noise level). This is typical of diverse instrumental techniques such as Gas Chromatography–Mass Spectrometry (GC-MS) or Gas Chromatography–Ion Mobility Spectrometry (GC–IMS). In this work, we propose a modification (named *psalsa*) to the asymmetry weights of the original ALS method in order to better reject large peaks above the baseline. Our method will be compared to several versions of the ALS algorithm using synthetic and real GC signals. Results show that our proposal improves previous versions being more robust to parameter variations and providing more accurate peak areas.**

I. Introduction

Several instrumental techniques such as GC–MS or GC–IMS produce signals consisting of chemical information, baseline and noise. In order to extract reliable chemical information, such as a list of peak positions and peak areas, it is crucial to denoise the signal and remove its baseline. Analysts usually manually select the peak boundaries and fit a curve to them to estimate the baseline of each peak, however manual baseline estimation is very expensive when the number of peaks in the signal increases (such as in complex biological samples) or when there is a large number of signals to analyse. Moreover, the analyst adds a subjective component to peak identification that depends on her/his expertise. For this cases, an automatic baseline estimation method is needed.

There are many automatic baseline estimation methods published, such as methods based on polynomial fitting [1], methods based on weighted least squares [2]–[4] or methods based on wavelets [5].

In this work, a modification to the Asymmetric Least Squares method is proposed. This modification is presented in order to improve the performance of the baseline estimation when large peaks are present in the signal.

This paper is organized as follows: First a description of the synthetic and real datasets is given on section II. Then, on section III the three methods under comparison are described: in section III-A the original ALS method [6] is summarized; in section III-B the improved method *airPLS* [3] is described; and finally our proposed method "*Peaked Signal's Asymmetric Least Squares Algorithm*" (*psalsa*) is detailed in section III-C. Results and discussion are shown on sections IV and V respectively. Finally some conclusions are given.

II. Data description

Two Gas Chromatography datasets are used to compare the different methods: On the one hand, a synthetic dataset offers the possibility to assess objectively the performance of the different methods, as we know the real baseline added to the synthetic signal and therefore we can compute the error of the different baseline estimations. On the other hand, a real dataset lets us check how the different methods perform on *real world* samples, which inevitably are more complex than synthetic chromatograms.

A. Synthetic dataset

A dataset with $N_{\text{synth}} = 100$ samples was generated. Each synthetic chromatogram lasted 30 minutes long with a sampling frequency of 2 Hz. Each sample was the combination of three components: a baseline, noise and a signal consisting of the addition of several peaks.

A synthetic chromatogram is shown at figure 1

1) Peak model: In order to generate the signal, several peaks are generated and placed randomly on the signal. As peak density of 0.25 peaks/s is chosen, a total of 450 peaks/sample are generated.

Peaks are modelled following a Generalized Exponential (GEX) function. The generalized exponential function [7] is an empirical peak model that has been used successfully [8] to describe chromatographic peaks, taking into account factors such as peak shape and peak asymmetry.

The GEX model is given by:

$$f(t) = h \left(\frac{t - t_0}{t_m - t_0} \right)^{b-1} \exp \left\{ \frac{b-1}{a} \left[1 - \left(\frac{t - t_0}{t_m - t_0} \right)^{a} \right] \right\} \tag{1}$$

with $a > 0$ and $b > 1$ are constants, h is the peak height, t_m is the location of peak maximum and t_0 is the time where the peak starts emerging from the baseline.

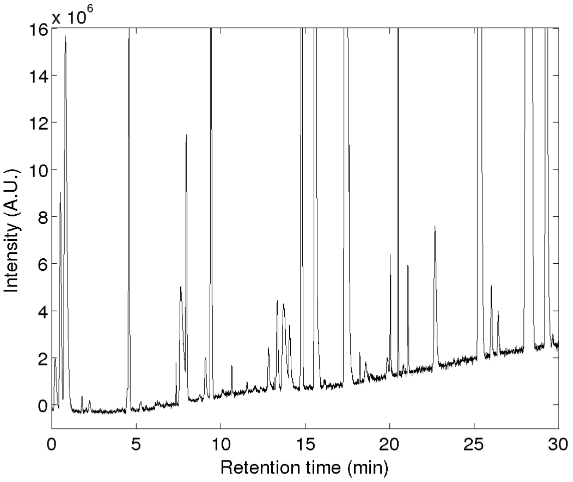

Fig. 1. Synthetic chromatogram

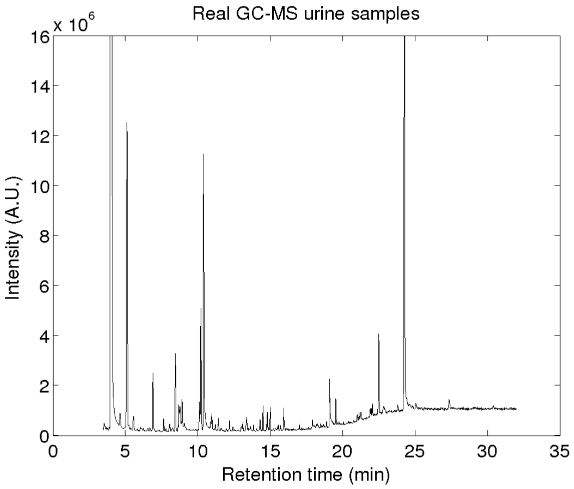

Fig. 2. Real urine samples

Peak model parameters are sampled from different probability distributions as follows:

- a: Uniform distribution with $\min = 0.5$ and $\max = 2$.
- b: Uniform distribution with $\min = 5$ and $\max = 8$.
- h: LogNormal distribution with $\mu = \log(400)$ and $\sigma = \log(200)$.
- t_0: Uniform distribution in the retention time range.
- t_m: $t_0 + 2 +$ Poisson distribution of $\lambda = 4$

By using those values we obtain a set of synthetic peaks similar to those of the real dataset.

2) Baseline model: In order to generate a realistic baseline, it is generated as the addition of three contributions:

$$b(t) = \text{ArcTan}(t) + \text{Linear}(t) + \text{Sinusoidal}(t) \tag{2}$$

where

$$\begin{aligned}\text{ArcTan}(t) &= A_{\text{low}} + \frac{2(A_{\text{high}} - A_{\text{low}})}{\pi} \cdot \arctan\left(\frac{\pi \cdot (t - t_0)}{t_r}\right)\\ \text{Linear}(t) &= mt + n\\ \text{Sinusoidal}(t) &= A \sin(2\pi f \cdot t + \varphi)\end{aligned}$$

The parameters for each baseline contribution are chosen from random uniform distributions in the following ranges:

- ArcTan: $A_{\text{low}} \in [2, 3] \cdot 10^5$, $A_{\text{high}} \in [1, 1.5] \cdot 10^6$, $t_0 \in [1100, 1300]$, $t_r \in [300, 700]$
- Linear: $m \in [3.5, 6] \cdot 10^5$, $n \in [4, 7] \cdot 10^5$
- Sinusoidal: $A \in [5, 30] \cdot 10^4$, $f \in [0.9, 1.4] \cdot 10^{-3}$, $\varphi \in [-\pi, \pi]$

3) Noise model: Gaussian noise with $A = 100+200t$, $\mu = 0$ and $\sigma = 400$ has been added to the signal. The amplitude increases with the retention time to simulate the fact that the end of the chromatogram is more noisy than the beginning.

B. Real samples

Chromatograms from a GC–MS dataset of human urine samples were used to test the proposed algorithm. Figure 2 shows samples from this dataset, notice the logarithmic scale on the y axis showing peaks orders of magnitude larger than the rest of the signal.

Samples were analysed at the PCB (Barcelona Scientific Park) premises, using a gas chromatograph – mass spectrometer (Focus GC–DSQ II) from Thermo Scientific equipped with a quadrupole analyser and an electron multiplier detector. The capillary column used was DB-624 (60 m × 0.32 mm i.d.) coated with 6 % cyanopropylphenyl 94 % dimethylpolysiloxane (film thickness 1.8 µm). The temperature program of the chromatographic oven began at 60 °C (2 min), ramped to 220 °C at 8 °C min^{-1} and held for 5 min. The injection port was maintained at 220 °C throughout the experiments.

III. Method description

In 1987, Newey and Powell introduced [9] Asymmetric Least Squares (ALS) in order to construct statistical tests for homoskedasticity applying them to Econometrics. More recently, Eilers et al. applied ALS for baseline estimation in connection to Parametric Time Warping alignment [10], and presented it in detail [6]. In 2010, Zhang et al. presented *airPLS* [3] which improved the weights of the original ALS method. Additionally, J Peng et al. [4] presented a different improvement to the original ALS method focusing on baseline estimation with multiple samples.

A. Original Asymmetric Least Squares

Given a signal y of length m, ALS aims to estimate a signal z smoother than y but still similar to it. ALS proposes a model-free cost function given by:

$$S = \sum_i d_i^2 + \lambda \sum_i \left(\Delta^2 z_i\right)^2 \tag{3}$$

where $d_i = y_i - z_i$ are the residuals of the estimation and $\Delta^2 z_i = z_i - 2z_{i-1} + z_{i-2}$.

The first term in S accounts for the *fidelity* from z to y, while the second term imposes *smoothness* to z. Smoothness is controlled by parameter λ, usually chosen between $10^2 \leq$

$\lambda \leq 10^9$. The cost function can be generalized by introducing weights w:

$$S = \sum_i w_i d_i^2 + \lambda \sum_i \left(\Delta^2 z_i\right)^2 \tag{4}$$

These weights w are introduced so as, if properly defined, will be able to reject penalizations to the cost function produced by regions where the signal is above the estimated baseline (i.e. peaked regions).

The proposed definition of w is based on a parameter p which is usually chosen as $0.001 \leq p \leq 0.1$:

$$w_i = \begin{cases} p & \text{if } d_i > 0 \\ 1 - p & \text{otherwise} \end{cases} \tag{5}$$

As one can see from the definition of w_i and the values of p, regions where the signal is placed above the baseline will have a much smaller contribution to the penalty.

Minimization of 4 leads to:

$$\left(W + \lambda D'D\right) z = Wy \tag{6}$$

where $W = diag(w)$ and D being the difference matrix: $Dz = \Delta^2 z$. As there is no model imposed on z, there will be m equations forming a sparse system, where only the diagonal end two sub-diagonals above and below it are non-zero.

A solution to eq. 4 can be found by iterating. Given an initial set of weights $w_i = 1$, an initial estimation for z_i can be computed. From z_i, weights are computed and used to get a new estimation for z. Less than 20 iterations are needed for a proper estimation of z.

According to [6], a proper value for p may be validated by considering the histogram of the residuals d, so as the noise components are normally distributed near zero and peaks are represented in the histogram as a positive asymmetric component. The right value for p will produce a baseline that cuts the noise instead of fitting below or above it.

B. airPLS *correction*

Zhang et al proposed in [3] an improvement to the definition of w with two objectives: To remove the parameter p, simplifying the usage of the algorithm; and to improve the quality of the estimation by adapting the weights depending on the distance from the signal to the baseline.

The definition of the weight vector w for *airPLS* is as follows:

$$w_i = \begin{cases} 0 & \text{if } d_i > 0 \\ \exp\left(\frac{-t \cdot |d_i|}{\sum_{d_i<0} |d_i|}\right) & \text{otherwise} \end{cases} \tag{7}$$

where t is the current iteration. With this definition of weights, regions of the signal where the signal is above the estimated baseline are ignored at the next iteration. For the rest of the weights, the further the signal is from the baseline the least it contributes to the penalty.

Having the current iteration t in the exponent forces the weights to be smaller on each iteration, making more significant the *smoothness* term as iterations go on.

The criteria set by *airPLS* to stop iterating is given by either a maximum number of 20 iterations or by:

$$\sum_{d_i<0} |d_i| < 0.001 \sum_{\forall i} |y_i| \tag{8}$$

The featured *airPLS* version 2.0 for MATLAB was used as the reference implementation. In this version, a p value is used to set the weights of points found at the beginning and at the end of the spectra as the adaptation of the weights does not give good estimates close to the signal limits.

C. Proposed method: psalsa

We propose a different definition for the weights much more similar to the original ALS algorithm. ALS is not able to fit very intense peaks because, even though a small p value is chosen, the peak residual is big enough to contribute significantly to the baseline fit. If a smaller value of p is used instead, then the baseline fits completely below the noise instead of cutting through it. Therefore, an adaptive value for the weights depending on the residuals is required and the following definition is proposed:

$$w_i = \begin{cases} p \cdot e^{-\frac{d_i}{k}} & \text{if } d_i > 0 \\ 1 - p & \text{otherwise} \end{cases} \tag{9}$$

The difference with respect to the original ALS method is on the positive residuals, where p is pondered by $\exp(-\frac{d_i}{k})$. Peak regions will show large residuals getting smaller weights, whereas noise regions will present small residuals and weights close to p. k is an additional parameter that controls the exponential decay of the weights. An easy way to estimate k is by setting it to the peak height we want to start rejecting. Note that by taking the limit $k \to \infty$ we recover the traditional ALS method.

As the original ALS method does, the criteria used by *psalsa* to stop iterating is given by either a maximum number of iterations (usually 20) or when the residuals do not change of sign with respect to the previous iteration.

IV. Results

A. Synthetic chromatograms

The three described methods were applied to the synthetic chromatograms. An example of how the different baseline estimations look like is shown on figure 3.

In order to estimate the best parameters for each method, the parameter space was swept. For each sweep, the root mean square error (RMSE) was computed applying equation 10 to each sample. The RMSE values were averaged obtaining a global RMSE. The optimal parameter values for the synthetic database were chosen as the parameters with the smallest global RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} \left(z_i - b_i\right)^2}{m}} \tag{10}$$

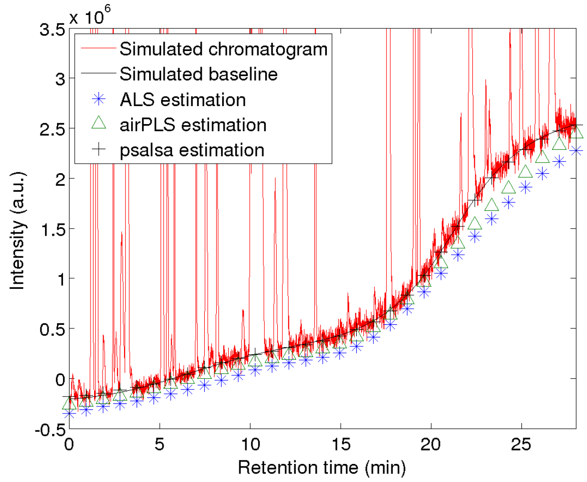

Fig. 3. Region of a synthetic sample showing different baseline estimations.

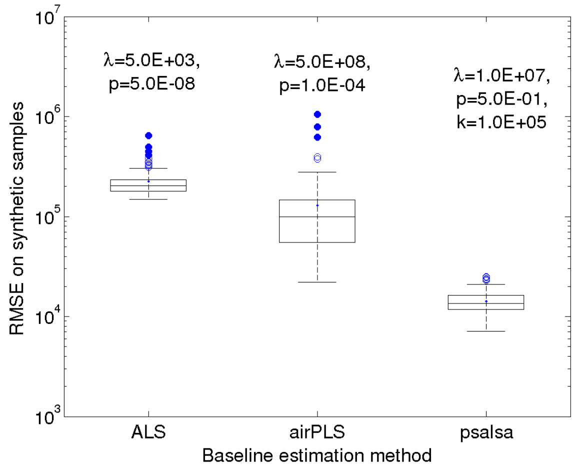

Fig. 4. Comparison of the three methods for synthetic chromatograms

In equation 10, z_i refers to the estimated baseline and b_i to the simulated baseline. m is the signal length.

In order to compare the three algorithms, figure 4 shows a boxplot of the RMSE distribution for the different methods in their optimal settings.

As the *psalsa* algorithm uses an additional parameter to control the exponential decay of the weights, we wanted to check the influence of that parameter on the overall performance. Several values of k chosen on a wide range of orders of magnitude were tested, obtaining figure 5.

B. Real samples

The three methods were applied to real samples. Figure 6 and figure 7 show the estimated baselines on different regions of a real urine sample.

V. Discussion

The original ALS algorithm was not designed specifically to fit signals with peaks several orders of magnitude above the

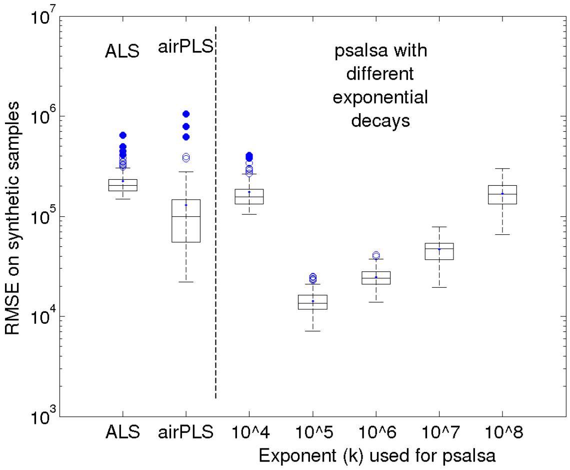

Fig. 5. Performance comparison of different exponents for *psalsa*. ALS and airPLS optimal results are shown for comparison

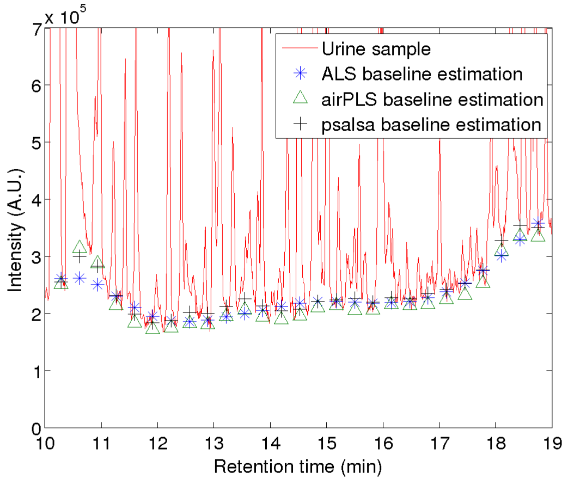

Fig. 6. Comparison of the three methods for real samples

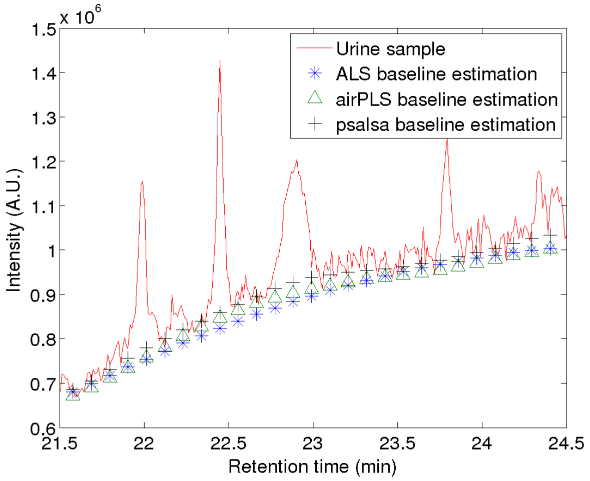

Fig. 7. Comparison of the three methods for real samples

baseline. Considering eq. 4, one may notice that even though a small value for w_i is given for $d_i > 0$, given a large enough d_i, its contribution to S may still be dominant, producing an estimation of the baseline which contains part of the peak area. In order to avoid the over-fitting, the value for p has to be chosen so as the baseline is not over-fitted to the peaks, instead of being chosen so as it cuts through the noise as described by [6]. This means that the value for p will have to be smaller, leading to baseline estimations below the real baseline. Given that the estimation is below the baseline, a flexible baseline will be easier to adapt to the real baseline whenever possible, that is the reason why λ values are smaller in the ALS method with respect to the other methods.

Therefore, on the analysed signals, the parameters which minimize the RMSE on the ALS method are chosen to be able to properly fit the large peaks, instead of according to their theoretical purpose.

On the other hand, the *airPLS* algorithm is able to cope with large peaks, as it gives $w_i = 0$ for $d_i > 0$. Unfortunately, that approach again leads to baselines fitted below the noise level instead of cutting through it. The *airPLS* algorithm was designed with the aim of removing the p parameter, and indeed p contribution is less relevant to the final estimation than the contribution of p at the original ALS algorithm, as it is only used at the boundaries of the signal.

Finally, *psalsa* algorithm does not suffer the issues of the original *ALS* method, as the exponential modulation reduces the contribution to S of the large peaks. This makes it possible to use p to enforce that the baseline crosses the noise level, instead of fitting below it. p value is not comparable directly to the *ALS* method, as its contribution is modulated by the exponential.

Even though *psalsa* requires an additional parameter (k) to control the exponential decay of the weights, figure 5 shows that the RMSE value is smaller on *psalsa* on a range of three orders of magnitude, making it easy to provide a value for k that improves ALS results.

When applying the three methods on real samples, we can confirm how *psalsa* is able to estimate a baseline cutting through the noise, instead of being under-fitted as happens with the other two methods.

VI. Conclusion

The *psalsa* algorithm is able to fit baselines of chromatographic signals with large peaks, improving the results of the original *ALS* method and the *airPLS* algorithm without adding computational complexity. Even though it requires an additional parameter k, it is easy to provide a reasonable estimate by relating it to the height of the peaks.

The proposed algorithm has been applied to both synthetic and real chromatograms obtaining successful estimations in both cases.

The algorithm is being adapted to work with Ion Mobility Spectrometry datasets and results will be published in a near future.

Acknowledgments

This work is financially supported by the Spanish project SMART-IMS (TEC2011-26143).

S.Oller, S. Marco and A. Pardo are part of a consolidated research group by the Generalitat the Catalunya (2009SGR-0753). S. Oller should acknowledge the Spanish “Ministerio de Economía y Competitividad” for the FPU PhD scholarship.

References

[1] M. L. Salit and G. C. Turk, “A drift correction procedure,” *Analytical chemistry*, vol. 70, no. 15, pp. 3184–3190, 1998.

[2] P. H. Eilers, “A perfect smoother,” *Analytical Chemistry*, vol. 75, no. 14, pp. 3631–3636, 2003.

[3] Z.-M. Zhang, S. Chen, and Y.-Z. Liang, “Baseline correction using adaptive iteratively reweighted penalized least squares,” *Analyst*, vol. 135, no. 5, pp. 1138–1146, 2010.

[4] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, and J. Tan, “Asymmetric least squares for multiple spectra baseline correction,” *Analytica chimica acta*, vol. 683, no. 1, pp. 63–68, 2010.

[5] X.-G. Shao, A. K.-M. Leung, and F.-T. Chau, “Wavelet: a new trend in chemistry,” *Accounts of chemical research*, vol. 36, no. 4, pp. 276–283, 2003.

[6] B. H. F. M. Eilers Paul H. C. (2005) Baseline correction with asymmetric least squares smoothing. http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf.

[7] A. Felinger, *Data Analysis and Signal Processing in Chromatography*. Elsevier, 1998.

[8] R. A. Vaidya and R. D. Hester, “Deconvolution of overlapping chromatographic peaks using constrained non-linear optimization,” *Journal of Chromatography A*, vol. 287, pp. 231–244, 1984.

[9] W. K. Newey and J. L. Powell, “Asymmetric least squares estimation and testing,” *Econometrica: Journal of the Econometric Society*, pp. 819–847, 1987.

[10] P. H. Eilers, “Parametric time warping,” *Analytical Chemistry*, vol. 76, no. 2, pp. 404–411, 2004.