Analytica Chimica Acta 1201 (2022) 339605

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Critical comparison of background correction algorithms used in chromatography

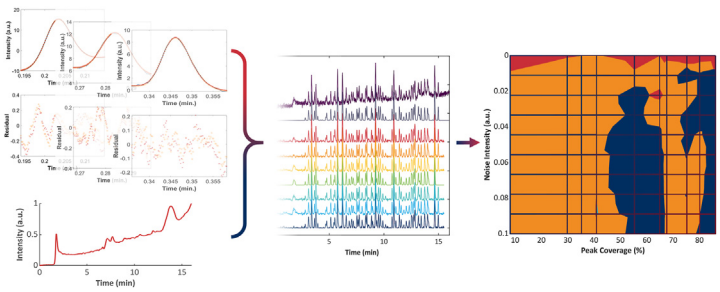
Leon E. Niezen [a, b, *], Peter J. Schoenmakers [a, b], Bob W.J. Pirok [a, b]

[a] Analytical Chemistry Group, van 't Hoff Institute for Molecular Sciences, Faculty of Science, University of Amsterdam, Science Park 904, 1098, XH Amsterdam, the Netherlands
[b] Centre for Analytical Sciences Amsterdam (CASA), the Netherlands

HIGHLIGHTS

- A software application was developed that allows for the comparison of smoothing and drift correction algorithms.
- It can generate hybrid (part experimental, part simulated) data for other comparison studies and allows anyone to do so.
- The need for a large and varied common data set against which all correction algorithms will be tested is highlighted.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 22 September 2021
Received in revised form
10 February 2022
Accepted 12 February 2022
Available online 18 February 2022

Keywords:
Background correction
Noise filtering
Smoothing
Chemometrics
Data processing
Pre-processing

ABSTRACT

The objective of the present work was to make a quantitative and critical comparison of a number of drift and noise-removal algorithms, which were proven useful by other researchers, but which had never been compared on an equal basis. To make a rigorous and fair comparison, a data generation tool is developed in this work, which utilizes a library of experimental backgrounds, as well as peak shapes obtained from curve fitting on experimental data. Several different distribution functions are used, such as the log-normal, bi-Gaussian, exponentially convoluted Gaussian, exponentially modified Gaussian and modified Pearson VII distributions. The tool was used to create a set of hybrid (part experimental, part simulated) data, in which the background and all peak profiles and areas are known. This large data set (500 chromatograms) was analysed using seven different drift-correction and five different noise-removal algorithms (35 combinations). Root-mean square errors and absolute errors in peak area were determined and it was shown that in most cases the combination of sparsity-assisted signal smoothing and asymmetrically reweighted penalized least-squares resulted in the smallest errors for relatively low-noise signals. However, for noisier signals the combination of sparsity-assisted signal smoothing and a local minimum value approach to background correction resulted in lower absolute errors in peak area. The performance of correction algorithms was studied as a function of the density and coverage of peaks in the chromatogram, shape of the background signal, and noise levels. The developed data-generation tool is published along with this article, so as to allow similar studies with other simulated data sets and

Abbreviations: 1D, One-dimensional; 2D, Two-dimensional; 1D-LC, One-dimensional liquid chromatography; 2D-LC, Two-dimensional liquid chromatography; airPLS, Adaptive iteratively reweighted penalized least squares; ANN, Artificial neural network; asLS, Asymmetrical least squares; arPLS, Asymmetrically reweighted penalized least squares; BEADS, Baseline estimation and denoising using sparsity; DAD, Diode-array detector; EMG, Exponentially modified gaussian; FIR, Finite impulse response; LMV, Local minimum value; LC, Liquid chromatography; MairPLS, Modified adaptive iteratively reweighted penalized least squares; MM, Mixture model; MPLS, Morphologically weighted penalized least squares; PLS, Penalized least-squares; RID, Refractive-index detector; RMSE, Root-mean-square error; RPLC, Reversed-phase liquid chromatography; SASS, Sparsity-assisted signal smoothing; SSE, Sum-of-squares error.

* Corresponding author. Postbus 94157, 1090, GD Amsterdam, the Netherlands.
E-mail address: L.E.Niezen@uva.nl (L.E. Niezen).

https://doi.org/10.1016/j.aca.2022.339605

possibly other algorithms. The rigorous assessment of correction algorithms in this work may facilitate further automation of data-analysis workflows.

1. Introduction

Spectroscopic or chromatographic data can generally be assumed to consist of three components, (*i*) low-frequency baseline drift, (*ii*) high-frequency noise and (*iii*) relevant chemical information, typically with a frequency between that of drift and noise. The latter two contributions together are also commonly described as "background". Often, there is more background than chemical information present in a signal, as each data point contains a background contribution. In such a case, or if the background is of a frequency very similar to that of the relevant signals, problems may occur with the interpretation of the data. For example, peak detection may be hindered, and errors in classification, discrimination, and, especially, quantification, may occur [1–7]. It is, therefore, desirable to perform baseline-drift correction and noise removal to ensure a correct interpretation of the data, unless peak detection can be performed in such a way that it is not hindered by the presence of noise and drift.

A large number of background-correction algorithms have been developed [8–25]. Examples of baseline-drift-correction algorithms include many of the penalized least-squares (PLS) methods, including asymmetrical least squares (asLS) [16], asymmetrically reweighted penalized least squares (arPLS) [24], adaptive iteratively reweighted penalized least squares (airPLS) [19], modified airPLS (MairPLS), and morphologically weighted penalized least squares (MPLS) [10] as well as other techniques, such as iterative polynomial fitting [26,27], Corner-Cutting [9], Backcor [11,12] and baseline estimation and denoising using sparsity (BEADS) [14]. Additionally, methods based on Fourier filtering and on wavelets have also been developed [22,28,29] as well as less conventional methods based on the use of neural networks [30,31]. Although many background-correction methods have been proposed, comparisons between the performance of these are scarce and often inadequate.

Firstly, in many cases the background-correction methodologies developed for use on spectroscopic or chromatographic data are compared only qualitatively to two or three other methods using experimental data, while quantitative comparisons tend to be limited to small sets of simple simulated data [19,24]. Consequently, it is not clear which background-correction methods perform best. Instead, a trial-and-error approach is routinely taken, in which three or four methods are arbitrarily selected and applied to a small test set of data. The (qualitatively) best performing one is used for the correction of all further measurements. If the test set is representative for all data and good methods are selected, such an approach can work reasonably well. However, this is by no means guaranteed and when correction is required for large numbers of measurements automation of background correction in data-analysis workflows is susceptible to errors. This is especially relevant when data-analysis methods, such as classification, discrimination or clustering are employed. In such cases, incorrect background correction can lead to erroneous results and incorrect conclusions.

Secondly, most approaches have been developed for specific datasets, such as Backcor, which was originally intended for the background correction of optical spectra [11]. While this is understandable, it induces the risk of a data-dependent bias in performance when evaluating the different methods. However, since quantitative comparisons are virtually non-existent, the magnitude of this risk cannot be assessed.

Thirdly, there are no data sets available for an objective comparison of background-correction approaches. Authors have generally employed specific datasets or simulated data. The latter is a pragmatic solution, which has the advantage that the ground-truth values for peak characteristics (*e.g.* peak area, shape) and background are known. This allows quantification of the extent of information loss as a result of the correction. A common criticism against the use of simulated data is that it is thought to be less representative than real data. In many cases this can be deemed true, as simple polynomial, sinusoidal or linear baselines are used, along with Gaussian peak shapes. Ideally, a large set of generated data that is sufficiently varied should be used against which all methods can be benchmarked.

In this work we aim to rigorously compare a number of recently developed background correction methodologies (*i.e.* baseline-drift correction as well as noise removal) in a comprehensive and critical manner. For this purpose, we used experimental data on backgrounds and peaks to create large sets of hybrid (part experimental, part simulated) data. Methods that featured very long computation times (*e.g.* several minutes or more for a one-dimensional signal of approximately 20,000 data points) were discarded in the present work after an initial evaluation, as our eventual objective is to apply the most-appropriate algorithms to two-dimensional liquid chromatography (2D-LC). For the same reason, only methods with no more than three different input parameters were assessed, to avoid significant manual tuning of the parameters to obtain satisfactory results.

2. Theory

A variety of different background-correction methods have been compared in this work. Here a brief overview of the theory behind each method is given. The noise-removal methods that were used prior to drift correction include well-known smoothing methods, such as Savitsky-Golay smoothing (SG) [8], Whittaker smoothing [32], finite-impulse-response (FIR) low-pass filtering [33], and wavelet filtering [34], as well as the more novel sparsity-assisted signal smoothing (SASS) [21]. After smoothing, drift-correction methods, *viz.* asymmetric least-squares (asLS) [16] and two conceptually similar methods, including adaptive iteratively reweighted penalized least-squares (airPLS) [19], asymmetric reweighted penalized least-squares (arPLS) [24] were applied. Additionally, the mixture model (MM) [18], a method based on iterative polynomial fitting with an asymmetric cost function (Backcor [11]), a method based on local minimum values (LMV) [15] and a recent method based on the use of an artificial neural network (ANN), henceforth referred to as the Autoencoder [31], were also included.

2.1. Drift-correction methods

2.1.1. Backcor

Many drift-correction methods are based on a polynomial-fitting approach, with the signal drift being described by a

polynomial of a certain order. Such approaches cannot easily be automated, as this would require automatic detection of peaks and selection of background regions in the raw data. The correction method should itself be capable of determining which points belong to the drifted baseline and which do not. In Backcor this is achieved by iteratively fitting a polynomial through the entire signal and utilizing asymmetric forms of typically symmetric cost functions, such as the Huber or truncated quadratic cost functions, to penalize data points falling above the fit less harshly than those that fall below the fit [11,12]. As a result, minimizing such a cost function results in positive peaks being automatically filtered out during the fitting procedure, since they have a lower cost. Because the noise around the drift is assumed to be normally distributed, this method still relies on a user-defined threshold to distinguish between noise and peaks.

The condition for the asymmetric truncated quadratic cost function is as follows.

$$\varphi = \begin{cases} d^2, & d < s \\ s^2, & d \geq s \end{cases} \tag{1}$$

With $d = x_i - z_i$, x the original data, z the fitted data, d the difference between the fit and the data for the i-th datapoint, φ the cost, s the user-defined threshold. The user additionally has to choose the cost function and the degree of the polynomial. Note that when the asymmetrical truncated quadratic cost function is used together with a threshold s of zero the approach is equivalent to the iterative polynomial fitting approaches [26,27]. It can be seen that d^2 corresponds to conventional least squares, unless the difference exceeds the set threshold, above which d^2 becomes constant (and equal to s^2). Minimizing the sum of the φ for all datapoints allows determining the polynomial coefficients and hence the drift.

2.1.2. Local-minimum value

The local-minimum-value approach relies on the presence of local-minimum values (LMV's), which are data points lower in intensity than adjacent data points [15]. The approach first establishes an "initial background" consisting of local minimum values, and then removes any points above this initial estimate by using a moving window and median-based outlier detection. The latter relies on a threshold based on the amount of noise in the signal. The drift is then obtained by linear interpolation between the areas that were considered as outliers (peak regions). In this approach the peak regions are therefore detected based primarily on the noise in the signal and the chosen window width, which should be chosen based on the peak width.

2.1.3. Asymmetric least-squares

Many well-known background-correction algorithms are based on the use of penalized least squares (PLS). The PLS algorithm relies on balancing the fit of a model to the data, F, given by the sum of squares error (SSE), and its roughness (R) by adjusting a smoothing parameter, λ:

$$\varphi = F + \lambda R = \sum_{i=1}^{m} (x_i - z_i)^2 + \lambda \sum_{i=2}^{m} (\Delta z_i)^2 = \mathrm{x} - \mathrm{z}^2 + \lambda \mathrm{D}\mathrm{z}^2 \tag{2}$$

Where x_i is the ith data point in the signal, x, and z_i is the ith point of the fitted data, z. The difference between adjacent fitted data points is therefore given by Δz_i. This method as such cannot be used for background-drift correction, as it requires prior information on the locations of peaks in the signal. If these locations are known a binary mask or "weighted matrix" can be created, which ensures that only the background drift is modelled [35,36].

$$(W + \lambda D'D)z = Wx \tag{3}$$

$$z = (W + \lambda D'D)^{-1} Wx \tag{4}$$

Where W is a diagonal matrix with weight vector w_i on its diagonal, λ is the smoothing parameter and D is a difference matrix such that Dz = Δz. In case a binary mask is used w_i consists of solely ones and zeroes to differentiate between peaks and baseline, respectively. However, in principle the weights may be any value between zero and one depending on how the weights are established. Furthermore, in case of the asLS, arPLS, and airPLS algorithms the determination of these weights is based on an iterative process where weights are selected based on the difference from the fitted baseline. For the initial fit no penalty is given (weights are all equal to one). Points far away from this initially determined baseline are then given smaller weight and hence will have less influence on the fit. These weights are then used to solve **Equation (4)** once again, and new weights are established. This process is continued until the weights become invariable. In asLS, developed by Eilers et al. [16], the weights are established using an asymmetry parameter (p), which allows for the weights associated with positive and negative deviations from the baseline to be different (smaller and larger, respectively). This approach is summarized in **Equation (5)**.

$$w_i = \begin{cases} p, & x_i > z_i \\ 1 - p, & x_i \leq z_i \end{cases} \tag{5}$$

The asymmetry parameter p can vary between 0 and 0.5, with 0.5 resulting in a conventional fit, while anything smaller than 0.5 will result in the peaks being taken into account less.

2.1.4. Adaptive iteratively reweighted penalized least-squares

In the case of airPLS weights are selected based on an exponential function (**Equation (6a)**) [19]. The algorithm is terminated once the difference between the signal and the fitted vector $|d_t|$ falls below a user-selected threshold, i.e. when condition (**Equation (6b)**) is met.

$$w_i = \begin{cases} 0, & x_i \geq z_i \\ e^{t(x_i - z_i)/|d|}, & x_i < z_i \end{cases} \tag{6a}$$

$$|d_t| < 0.001|\mathrm{x}| \tag{6b}$$

With t being the iteration index and $d = x_i - z_i$, as earlier defined.

2.1.5. Asymmetric reweighted penalized least-squares

In the asymmetrically reweighted penalized least-squares (arPLS) algorithm, developed by Baek et al. [24] the weights are established based on a logistic function, as shown in **Equation (7)**. This method functions in essentially the same way as asLS and airPLS, but is claimed to be better at establishing the drift in the presence of noise, due to how the weights are selected.

$$w_i = \begin{cases} logistic(d,\ \mu_i, \sigma_i), & x_i \geq z_i \\ 1, & x_i \leq z_i \end{cases} \tag{7}$$

$$logistic(d,\ \mu_i, \sigma_i) = \frac{1}{1 + e^{2(d - (-\mu + 2\sigma))/\sigma}} \tag{8}$$

With $d = x_i - z_i$, μ and σ are the mean and standard deviation of, which is the part of d where the condition $x_i < z_i$ is met. This allows for weights above and below the signal to be the same, while any signal higher than the noise mean will receive a progressively lower

weight. The baseline is established once the weights become invariable, once again depending on a set threshold.

2.1.6. Mixture model

The mixture model estimates the baseline by calculating the posterior probability that a point belongs to the baseline [18]. The entire signal is assumed to be constructed from a mixture of two probability densities, one of which is normal (and corresponds to the baseline) and one of which is unknown, corresponding to the peaks. To estimate both components of the signal/mixture, a so-called Expectation-Maximization algorithm is used. In the first step of this algorithm the posterior probabilities are calculated, after which the baseline is modelled using P-splines (penalized B-splines). The coefficients (or penalties α) of these P-splines are determined by minimizing the following objective function:

$$\varphi = (x - B\alpha)^T P(x - B\alpha) + \lambda D\alpha^2 \tag{9}$$

$$\widehat{\alpha} = (B'PB + \lambda D'\mathrm{D})^{-1} B'\mathrm{Rx} \tag{10}$$

In which x corresponds to the data, B corresponds to an n × m cubic spline basis of m number of splines, λ is the smoothing parameter, $P = diag(p_i)$ and p_i is the posterior probability for the ith data point to belong to the baseline. These posterior probabilities are calculated from

$$p_i = \frac{\pi g(x|\mu,\sigma)}{\pi g(x|\mu,\sigma) + (1-\pi)h(x-\mu)} \tag{11}$$

where $g(x|\mu,\sigma)$ is the normal density function (baseline + normally distributed noise with background level, μ, and standard deviation, σ), $h(x-\mu)$ the unknown density function (peaks), and π an unknown mixing ratio. This approach is conceptually similar to the previous three methods, with the posterior probabilities used as the weights. Once again, the method differs in how these weights are determined.

2.1.7. Autoencoder

The Autoencoder method is based on the use of deep learning algorithms and aims to concomitantly denoise and drift correct the input data [31]. The method achieves this by using a large number (in the order of several thousands) of differentiable or adaptable filters, which can be fine-tuned as long as a representative and large data set is available on which to train the method. Naturally this method is therefore limited by the data on which it was trained. However, recently Kensert et al. [31] have shown that by using a model trained on a large set of simulated data the successful drift correction and smoothing of experimental data may be achieved. In the present study we have included their pre-trained model to further assess how well this method can perform without performing extensive initial training. This is interesting because if the method is sufficiently flexible it could allow for unsupervised background correction, which is typically very difficult to achieve.

3. Experimental

Experimental backgrounds were obtained from various sources. Background 5 was measured on an Acquity system purchased from Waters (Milford, MA, USA) using refractive-index detection, while all other backgrounds (1–4) were measured on an Agilent 1260 system using diode-array UV detection, purchased from Agilent (Waldbronn, Germany). Backgrounds 1, 3, and 5 were obtained from empty modulations in comprehensive two-dimensional liquid chromatography (LC × LC) runs, while backgrounds 2 and 4 were blank measurements in one-dimensional LC.

Signal simulation has been performed using MATLAB 2018a purchased from MathWorks (Natick, MA, USA), on a Dell XPS13 Laptop purchased from Dell (Round Rock, TX, USA). Background correction and automatic parameter determination were performed using MATLAB 2020a on a Dell Alienware Area 51–9829 R2 PC.

The developed tool is available as a downloadable application on: https://cast-amsterdam.org/software/

4. Results & discussion

4.1. Establishing experimental data for use in simulation

The major disadvantage of simulated data is that the complexity of experimental data may be oversimplified. Conversely, using experimental data may complicate the comparison of background-correction methods, as the ground-truth values (*e.g.* true peak areas) are not known. Therefore, we developed a library of simulated data which was based on experimental data. The workflow comprised three steps (see Fig. 1). (*i*) A background was selected from a pool of blank experiments; (*ii*) Varying degrees of white noise were drawn from a Gaussian distribution; (*iii*) A number of peaks were added, with a shape extracted from experimental data by curve-fitting.

4.1.1. Establishing experimental background and adding noise

The first step in the creation of the simulated data was the establishment of the low frequency drift component. As drift can be highly unpredictable and, therefore, difficult to model, an empirical approach was taken, where the background signals were obtained by compiling a library of different blanks from a variety of chromatographic experiments. This library can be further expanded with future research. Naturally, such experimental backgrounds contain an initial amount of noise in addition to the drift component. To establish an estimate of the initial noise in these backgrounds, the median absolute deviation (MAD) was used [37]. This is a robust measure of the deviation around the local average (i.e. the noise) present in the signal and is calculated using **Equation (12a)**. However, in the presence of a baseline and peaks it has been suggested that a more representative value can be obtained by calculating the MAD from the first derivative of the background signal as given by **Equation (12b)** [15].

$$\sigma = k^{*}median|x_i - median(x)| \ (i = 1, \ldots, N) \tag{12a}$$

$$\sigma = k^{*}median|dx_i - median(dx)| \ (i = 1, \ldots, N) \tag{12b}$$

In which dx is the derivative of the signal and dx_i is the ith point in this derivative, k is a (constant) scaling factor which for normally distributed data equals 1.4826. For an overview of the five experimental backgrounds that were used see Supplementary Material **Figure S-1**, section S-1.

In some cases, the experimental backgrounds contained one or more system peaks. These were manually removed from the signals by curve fitting and subtraction, followed by smoothing across this range. This was deemed necessary, because our approach ideally requires an experimental background that contains only low frequency drift and a small amount of initial noise. Their removal had to be performed manually and was hence tedious and time-consuming. However, when algorithms are not compared as presented in this work, the removal of such peaks is not required, as long as these are positive. Only in the presence of negative peaks will this be critical for most drift correction algorithms, as such peaks are generally treated as background drift.

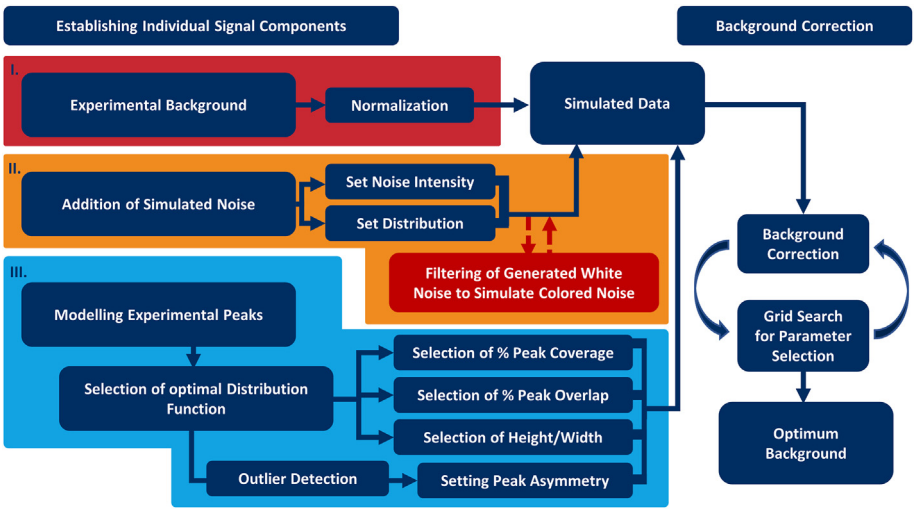


Fig. 1. Scheme illustrating the developed approach for data simulation and subsequent background correction.

The experimental backgrounds were perturbed with additional white noise, which was simulated as numbers randomly drawn from a normal distribution.

4.1.2. Establishing experimental peak shape

For the last step in the creation of the library (step III in Fig. 1) experimental peak shapes were extracted from real chromatographic data. It is of key importance that the peaks are accurately modelled. This may be achieved by fitting empirical peak-shape models or distribution functions, such as the Gaussian, exponentially modified Gaussian (EMG), or Pearson distributions to the data. For chromatographic data deviations from the expected ideal Gaussian profile are expected, due to heterogenous mass-transfer kinetics and non-linear adsorption isotherms [38]. Such deviations usually come in the form of tailing (or sometimes fronting) peak profiles. For each peak the extent of tailing and/or fronting may be different. To describe all possible peaks mathematically, a function must be used that is flexible enough to describe any amount of tailing and fronting. Several comparisons of distribution functions have previously been performed [39–42]. From these studies a general consensus emerged that the EMG distribution described chromatographic peak shapes most accurately [42]. In the present study these common distributions were also evaluated, along with several alternatives, such as Gaussian, Bi-gaussian [43,44], Pearson VII [45] and Modified Pearson VII [46] distributions. To successfully perform curve fitting two requirements must be met, *i.e.* (*i*) the approximate peak location must be known, and (*ii*) no background must be present [47]. In this study, the first prerequisite was met through manual selection of peak locations. While this can be performed automatically using peak-detection approaches, this would induce a risk of overlooking overlapping or small peaks. This may result in incorrect fitting for overlapping peaks. To meet the second requirement, either some form of background correction must be applied or data containing little or predictable background drift must be used. We opted to perform a linear background correction from the first to the last point in the selected peak regions. This involves the assumption that within the region of the peak the baseline does not show significant curvature. There are cases in which the approach cannot be used to describe peaks, for example when a large number of overlapping peaks is present, or when the background drift is significantly non-linear directly under the peak.

After the locations and peak regions were established, each peak was subjected to a least-squares curve-fitting procedure with 15 different distributions. In case of overlapping peaks, all peaks in the selected region were included and curve fitting was performed with two or more distributions of the same type. The possibility that overlapping peaks required different types of distribution functions was not considered in the current study, however even with a single distribution function it is still possible to describe a variety of peak shapes. The goodness-of-fit of the distributions to the experimental peaks was assessed using the Akaike information criterion (AIC) calculated using **Equation (13)**:

$$AIC = n^{*}\ln\frac{SSE}{n} + 2K \tag{13}$$

Where n is the number of data points, SSE is the sum of squared errors, in our case normalized for peak height, and K corresponds to the number of variables in the distribution function.

As an example, the results of this fitting approach applied to a selection of peaks from a single 2D-LC modulation (second-dimension chromatogram) containing significant background drift are shown in Table 1 along with the types of distribution tested. The AIC values for the five best-performing distribution functions for each peak are shown in Fig. 2, along with the individual fits for peaks 1–5.

The results shown in Table 1 and Fig. 2 are representative of those obtained on large numbers of treated datasets. The modified Pearson VII and EMG distributions were found capable of describing a range of different peak shapes. Other distribution functions performed well in some cases, but were less generally applicable. These conclusions are in agreement with previous studies on this subject [39,41]. Because the modified Pearson VII distribution [46] provided a slightly better fit compared to the EMG function this distribution was chosen for the creation of peaks in the simulated data. This distribution is described by **Equation 14**

$$f(x) = \mathrm{A}\left(1 + \frac{(x-\mu)^2}{m(\sigma + A_s(x-\mu))^2}\right)^{-m} \tag{14}$$

in which μ corresponds to the mean of the distribution (the retention time), σ indicates the width of the peak, m is a parameter related to the kurtosis of the peak, covering a range between a fully Gaussian and a Lorentzian peak shape, A_s describes the asymmetry, or the extend of fronting or tailing, of the peak and A corresponds to the height of the peak.

Table 1
AIC values obtained for a selection of distribution functions for the fitted peaks shown in Fig. 2.

Distribution	Peak 1	Peak 2	Peak 3	Peak 4	Peak 5	Peak 6	∑AIC
Modified Pearson VII	-1.03×10^3	-1.77×10^3	-1.13×10^3	-1.28×10^3	-1.10×10^3	-9.86×10^2	-7.20×10^3
Exponentially Modified Gaussian	-1.04×10^3	-1.39×10^3	-1.11×10^3	-1.24×10^3	-1.08×10^3	-1.01×10^3	-6.91×10^3
Exponentially Broadened Gaussian	-9.75×10^2	-1.37×10^3	-1.13×10^3	-1.23×10^3	-1.08×10^3	-1.00×10^3	-6.29×10^3
BiGaussian	-8.89×10^2	-1.37×10^3	-1.13×10^3	-9.69×10^2	-1.06×10^3	-8.55×10^2	-5.59×10^3
Mixed Gaussian/Lorentzian	-7.55×10^2	-9.59×10^2	-1.29×10^3	-9.26×10^2	-8.28×10^2	-7.98×10^2	-5.38×10^3
Pearson	-7.56×10^2	-9.45×10^2	-1.13×10^3	-9.12×10^2	-8.30×10^2	-7.70×10^2	-5.30×10^3
Log-normal	-7.78×10^2	-9.58×10^2	-1.12×10^3	-8.72×10^2	-8.39×10^2	-7.81×10^2	-5.30×10^3
Gaussian	-7.47×10^2	-9.44×10^2	-1.13×10^3	-8.63×10^2	-8.22×10^2	-7.70×10^2	-5.14×10^3
Logistic	-6.64×10^2	-9.35×10^2	-1.23×10^3	-8.24×10^2	-7.56×10^2	-7.07×10^2	-4.08×10^3
Exponentially Broadened Lorentzian	-5.35×10^2	-1.06×10^3	-6.86×10^2	-9.08×10^2	-6.35×10^2	-6.09×10^2	-3.82×10^3
Lorentzian	-5.04×10^2	-8.98×10^2	-6.65×10^2	-8.30×10^2	-5.86×10^2	-5.73×10^2	-3.04×10^3

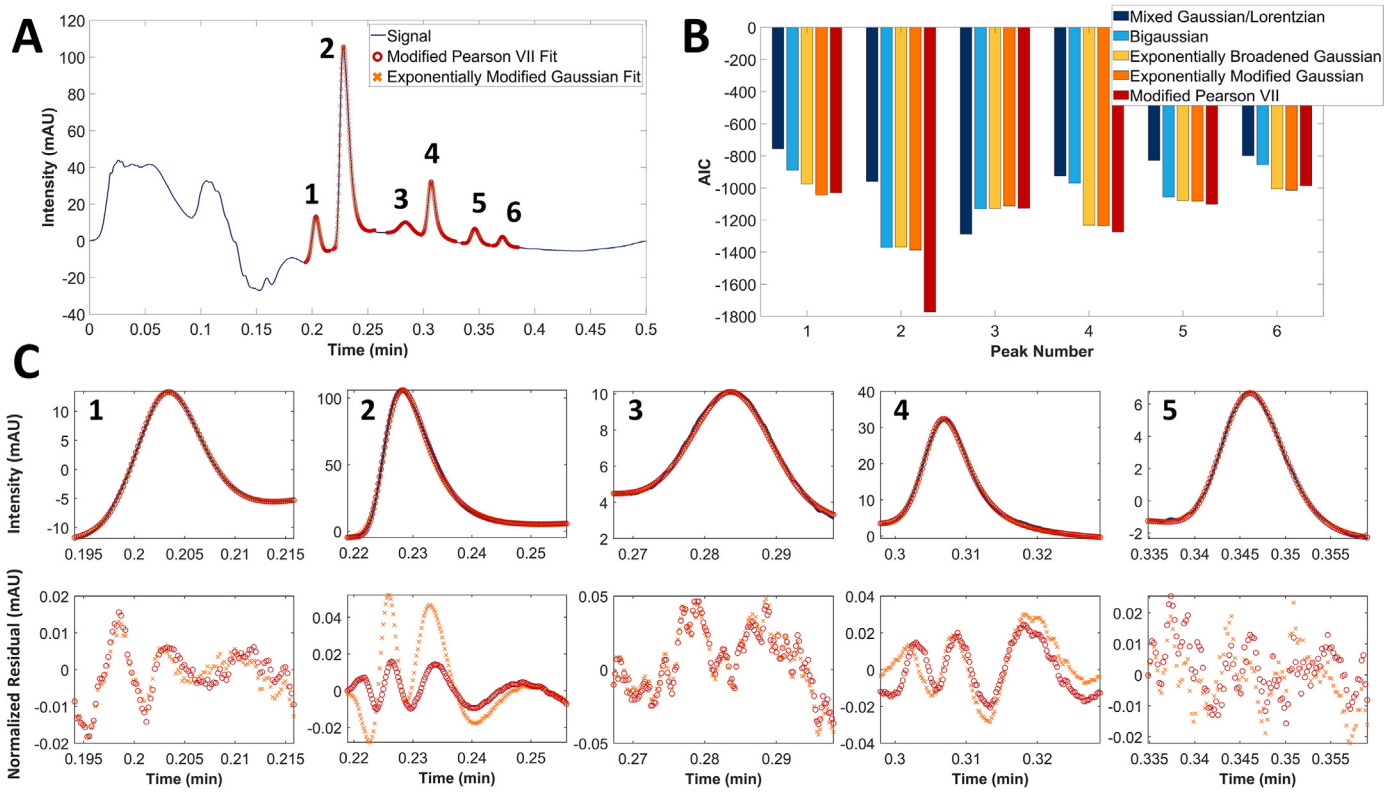


Fig. 2. A) Fit for Modified Pearson VII and EMG distributions on experimental data, B) AIC values for the five best distribution models for each of the fitted peaks and C) zoomed-in fits and residuals for five individual peaks.

4.1.3. Simulation of model chromatograms and spectra

Chromatograms or spectra are generated based on several input parameters as summarized in Table 2. The first set of parameters is used to describe peak shape and height while the second set of parameters dictates the intensity of the noise and the number of peaks in the chromatogram as well as their spacing.

The first set of input parameters determine the peak shape. These are drawn randomly and uniformly from the established maximum and minimum of a respective shape-defining parameter, *i.e. m* or *As*. As certain peaks in the input data may be significantly more asymmetrical than other ones, an outlier test was first performed on the *m* and *As* parameters obtained from the fitting procedure. Parameters were marked as outliers if they were more than three times the median absolute deviation away from the median. This test was performed to ensure that the simulated signals would represent realistic signals *i.e.* it was assumed that most peaks in a chromatogram would be of a typical shape. Outliers were subsequently removed. To ensure that peak heights are consistent between all signals, the experimental background drift was first "min-max normalized" and peak heights were uniformly selected from a range between 0.1 and 1. However, in realistic data often some peaks are much higher or lower in intensity than the "average peak". To simulate this, 10% of the generated peaks were randomly selected to have an intensity of either 5 or 200% of the maximum intensity of the drift (*i.e.* 0.05 or 2). Peak widths were chosen based on the fitted peaks. This seems justified, since for signals of different length (*e.g.* slow one-dimensional chromatograms or very fast second-dimension chromatograms) the parameters from the first set (*m* and *As*) did not seem to change significantly, rather the peak width itself changed. However, it should be considered that in the present study primarily gradient-RPLC data of small, uncharged molecules were used to model the peaks, hence peak widths are expected to be relatively constant. The peak-shape parameters may change significantly in other modes of chromatography.

Table 2
All parameters to be selected for signal generation.

Parameter Set	Parameter	Symbol	Range	Description
1st Set:	Kurtosis	m	3.5–51	Describes peak kurtosis
	Asymmetry	A_s	0.01–0.28	Describes peak asymmetry
	Amplitude	A	0.1–1	Multiplier for peak height
	Width	σ	0.007–0.008	Peak Width
	Position	μ	Random	Peak Retention Time
2nd Set:	Min. Peak Spacing		10%	Minimum space between two adjacent peaks (*e.g.* $\mu_1 and \mu_2$)
	Peak Coverage		10–100%	% of data points corresponding to peaks within selected region
	Noise Intensity		0–0.1	Amplitude of noise added to the chromatogram
	Background		1,2,3,4 or 5	Experimental blank
	Region Selection		Based on blank t_0	Region in which peak μ are generated.

The second set of input parameters determine the number of peaks, their locations and the amount of overlap allowed between individual peaks. Firstly, regions where no peaks are to be generated, can be selected. Such regions were selected based on the experimental measurements. Secondly, a minimum peak spacing is selected, based on peak widths, as well as a total peak coverage. The peak overlap was calculated based on the width at 5% of the peak height and it was chosen in accordance with the selected peak coverage. For example, a minimum spacing of 0% may result in completely overlapping peaks, while a minimum spacing of 100% will result in a signal where all peaks are completely separated. The chosen total peak coverage is defined as the number of data points containing information on peaks, again measured from the width at 5% of the peak height, to the total region available for peak generation. The coverage, width and shape roughly determine the number of peaks that must be generated, but the coverage alone does not account for varying amounts of peak overlap. The signal coverage is calculated based on the minimum spacing only. This means that the actual coverage will always be lower than the chosen value. To account for this the actual peak coverage is determined once more after the full signal has been generated.

By adding the generated peaks to the experimental background perturbed with additional white noise, many realistic chromatograms can be rapidly created. To allow other researchers to retrieve data from earlier publications (*e.g.* the present study) the tool features the setting of a seed, which is a number that serves to initialize the random number generator. This ensures that all "random" signal generation is controlled and reproducible. The same "random" signals can be generated again at any time, when required for future work.

4.2. Background correction

Using the developed data-simulation tool, background-correction methods could now be accurately compared. To this end different signals were created with peak coverages of 10, 21, 32, 43, 54, 66, 77, 88, 99 and 110%. Additionally, noise was added to each signal at ten different levels with "intensity" or standard deviation (σ) of 0–0.1 (up to $\approx$ 20% of the average peak height). Minimum peak spacing in all signals was set to 10%. This ensures the occurrence of severely overlapping peaks, while there were no peaks generated at exactly the same location.

Realistic data were generated in this manner for each of five different experimental backgrounds, resulting in 500 different signals for background comparison. A small representative fraction of the signals simulated in this way (different noise and coverage levels, three different experimental baselines) are shown in Fig. 3.

To compare the background-correction and smoothing methods it is vital that the input parameters for each method are set such that the estimated background is as close as possible to the actual background. For the smoothing methods ideal input parameters were obtained by minimizing the root-mean-square error (RMSE, given by **Equation (15)**) between the simulated noise and the noise obtained from subtracting the smoothed signal from the original signal, using a grid-search approach within manually defined constraints. For the drift-correction methods a similar approach was used where the RMSE was minimized between the known drift component and the background as determined by a drift-correction algorithm.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(b_i - z_i)^2}{n}} \tag{15}$$

Where b_i corresponds to i-th data point in the known background (either noise or drift component), z_i corresponds to the i-th data point in the estimated background and n corresponds to the total number of data points in either z or b. For both smoothing and drift-correction methods this RMSE was minimized separately, starting with optimization of the smoothing, followed by optimization of the drift correction. Examples of some of the corrections obtained in this way are illustrated in Fig. 4. In this case Savitsky-Golay was used for smoothing, while the result of six different drift-correction methods are shown.

Based on visual inspection of this specific signal the LMV and arPLS methods seem to perform slightly better than the other methods. The next step was to quantitatively compare all methods using the entire collection of generated data (500 chromatograms).

4.3. Quantitative comparison of correction performance

4.3.1. Influence of the smoothing method

Using the simulated data and automatic parameter selection, a quantitative comparison was made between all methods or combinations of methods by evaluating the RMSE obtained for each signal as well as the % error in obtained peak areas. First, the influence of the smoothing method, applied before the drift correction, was investigated based on the calculated RMSE values. This resulted in a response surface where RMSE as a function both the added noise intensity, and the peak coverage could be visualized. A comparison was made by overlaying all surfaces and maintaining only the lowest RMSE values. For many of the smoothing methods the response surfaces were fairly similar (see Supplementary Material **Figure S-2**, section S-3), indicating that there were only minor differences in the performance of the smoothing methods. Background 5 (see Supplementary Material **Figure S-1**) yielded larger deviations. However, this may be explained by the fact that this is by far the shortest signal in terms of the number of data points, as a result of a very low detector frequency. Therefore, the way in which we added noise (by randomizing each data point) may not be realistic in this case and the frequency of peaks in this signal is much closer to the frequency of the noise. Overall, the

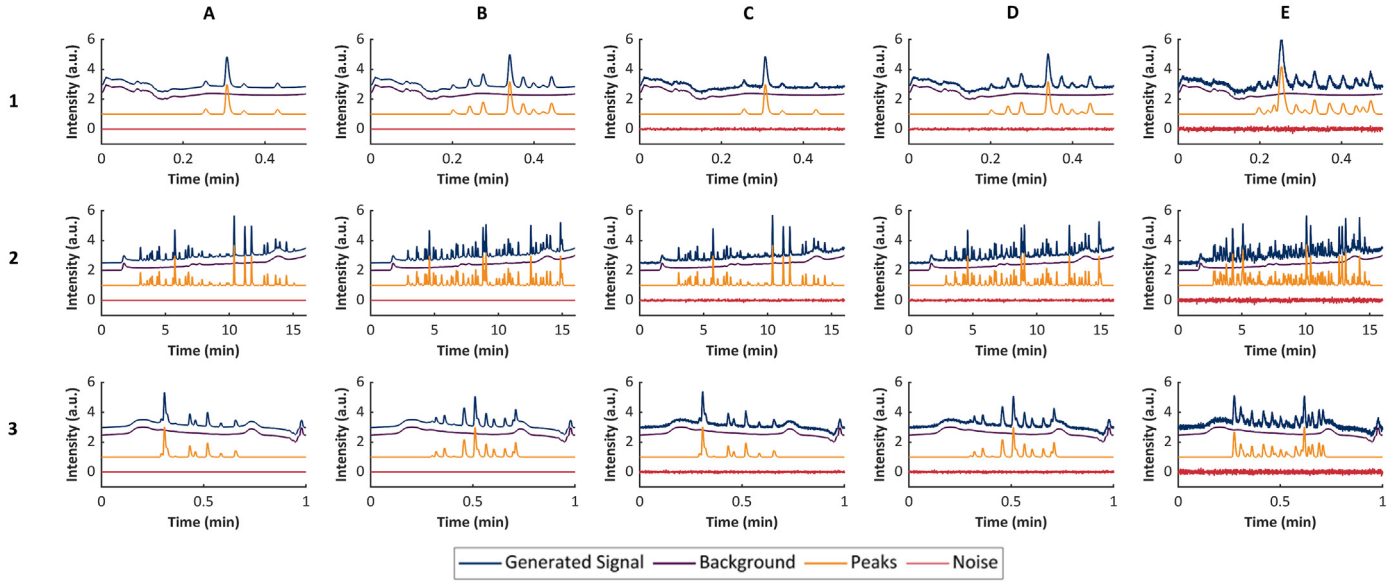


Fig. 3. Representative examples of the generated data. A) no added noise + low peak coverage (44%), B) no added noise + medium peak coverage (76%), C) medium added noise ($\sigma = 0.044$) + low peak coverage (44%), D) medium added noise ($\sigma = 0.044$) + medium peak coverage (76%), E) high added noise ($\sigma = 0.078$) + high peak coverage (92%). Numbers indicate the different backgrounds.

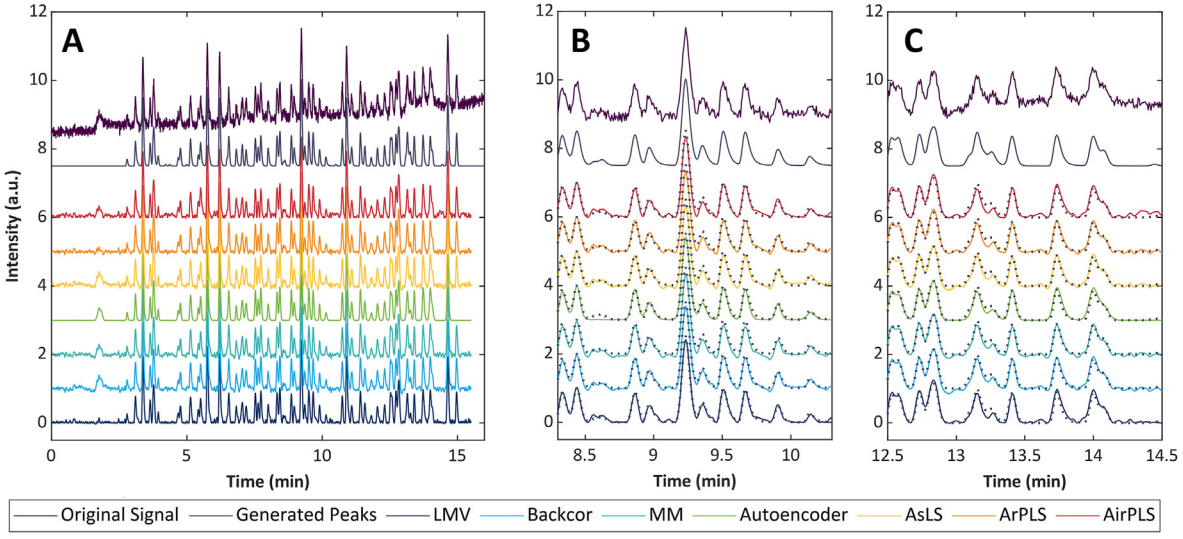


Fig. 4. A) Correction of generated data (added noise of 0.078, coverage of 67.5%) by the combination of Savitsky-Golay (window width: 23, polynomial order: 6) smoothing followed by drift correction using the LMV (window width: 5), Backcor (s: 0.0579, polynomial order: 18), MM (number of b-splines: 73, λ: 10^5), Autoencoder, AsLS (λ: 10^5, p: 0.0306), arPLS (λ: 3×10^5) and airPLS (lambda: 3×10^4) methods. B) Expansion of the region 8.3–10.3 min. C) Expansion of the region 12.5–14.5 min. For information on the methods see section 2. Dotted lines in (B) and (C) correspond to the generated (true) peak signals.

degree of noise and/or the choice of the smoothing algorithm appear to only marginally affect the performance of the various background-correction algorithms.

4.3.2. *Influence of the drift correction method*

To provide a clearer overview of the influence of noise and peak coverage for the different-drift correction methods the RMSE was calculated between the known drift and the background determined by the algorithms. This was performed both with and without prior smoothing to evaluate how well the different algorithms performed in the presence of additive noise. Calculating the RMSE between the sum of known noise and drift components and the estimated background illustrates clearly that most methods cannot perform smoothing and drift removal simultaneously. In this case the Autoencoder method performed best. However, when the RMSE is calculated between the estimated background and the known drift component it is shown that most methods do not perform worse at determining the underlying signal drift in the presence of additive noise. This data is included in the Supplementary Material (**Figure S-4** and **Figure S-5**). Because the Autoencoder was capable of describing both noise and drift while many of the other algorithms could not it was decided to first perform smoothing using SASS for the comparison of the methods. In Fig. 5 the RMSE values obtained using background 2, initially smoothed using SASS are illustrated.

From this comparison arPLS is seen to be the best performing method in most cases. However, this conclusion was found to depend on the background used. For other backgrounds the LMV or

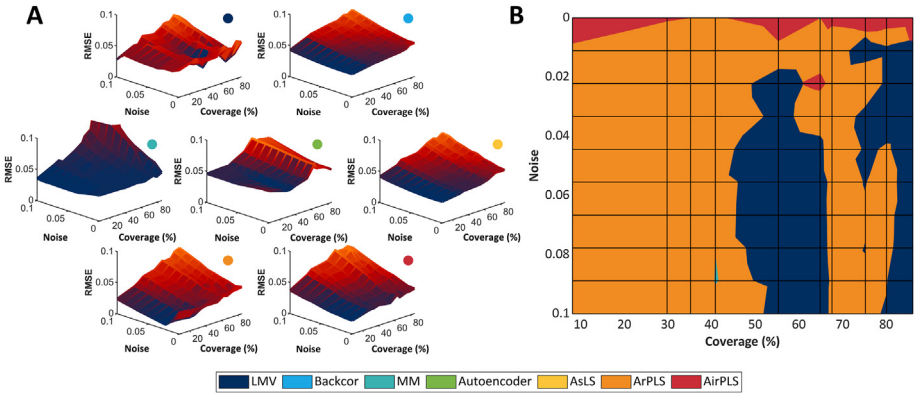


Fig. 5. A) RMSE surfaces obtained for the various drift-correction methods in combination with the SASS smoothing algorithm and for background 2 in Fig. 3. Methods are indicated by the coloured dots. B) Bottom view (lowest values) resulting from the overlaid RMSE surfaces. For an explanation of the methods see section 2.

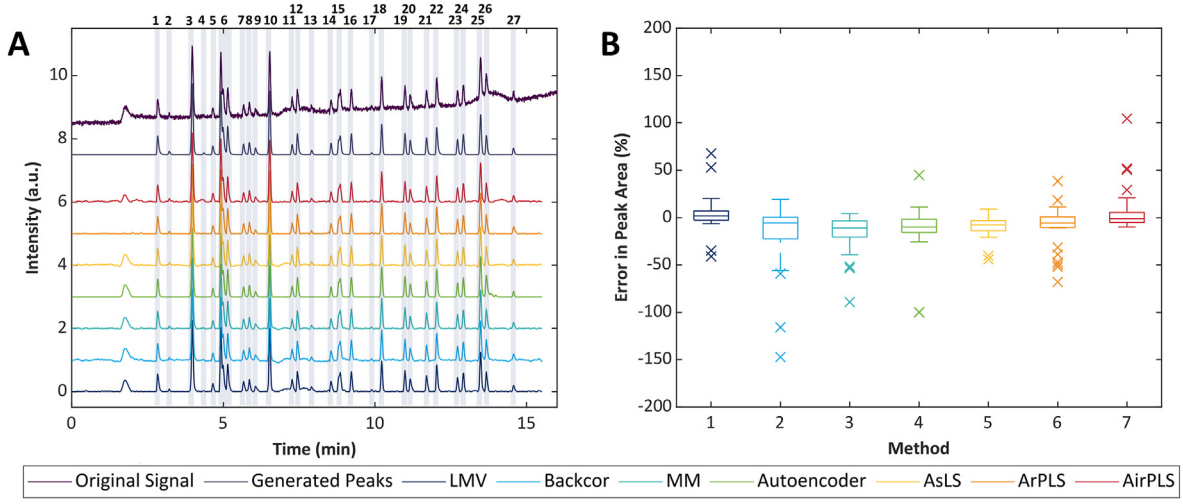


Fig. 6. A) Signal with (in blue) markings indicating peak regions used for the % error calculation, the different traces represent (from top to bottom) the uncorrected signal and the generated peaks, followed by signals corrected using the LMV (window width: 113), Backcor (s: 0.1271, polynomial order: 20), MM (number of b-splines: 100, λ: 10^5), Autoencoder, AsLS (λ: 10^5, p: 0.0510), arPLS (λ: 5.4×10^3) and airPLS (lambda: 4×10^4) methods. B) Error in peak area for each method. Regions 3, 6, 15 and 25 contain overlapping peaks; all other regions are individual peaks. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Backcor algorithms performed better than arPLS. However, if arPLS did not perform best, it was usually the second-best method and arPLS showed consistently good performance for all backgrounds investigated in this study (for an overview of all minimum RMSE surfaces see Supplementary Material **Figure S-3**).

4.3.3. Determination of error in peak areas

It is interesting to know which method results in the smallest error in peak area. Therefore, the error in peak area obtained after correction has been evaluated. These errors were determined and compared to an approximate peak area obtained by trapezoidal integration from the simulated peaks. In case of overlapping peaks these were treated as one, therefore this is a "peak region" comparison rather than a comparison of individual peaks. This avoids reliance on curve fitting and the possibility of baseline or noise resulting in incorrect results. The regions were selected using the simulated peaks. The error in peak areas obtained from this approach for each method and for the signal with a peak coverage of 35.2% and a noise intensity of 0.033 from background 2 is portrayed in Fig. 6. This is a representative case, in which a few peak regions contain overlapping peaks, but many peaks are isolated.

From Fig. 6-B it is clear that Backcor performed worse than the other methods (indicated by the larger spread) followed by the MM and Autoencoder methods. The best-performing methods were LMV, asLS and airPLS, with mean errors of 3.8, 9.9 and 7.9% respectively. In many cases errors were still substantial, especially for low-intensity peaks, as could be expected. Peaks in regions 2, 4, 5, 13 and 17 showed the largest relative errors of up to nearly −150%, which implies that the drift is significantly over-corrected in this region relative to the peak's height, resulting in a much smaller determined peak area. However, the peaks in these regions typically also had very low signal-to-noise ratios (SNR) on the order of ~1.5−3. Relative errors of 10 or even 20% in peak area were quite common, also for peaks of average height (for more in-depth figures of % error vs peak height for the different methods see Supplementary Material **Figure S-6)**. This surprising finding highlights the importance of appropriate drift correction and noise removal when accurate quantification is desired. The errors were similar when a different smoothing method was used prior to the drift correction. The results for the same signal and selected peak regions as in Fig. 6, but with four different noise intensities (0.02, 0.06, 0.08, and 0.1), are illustrated in Fig. 7.

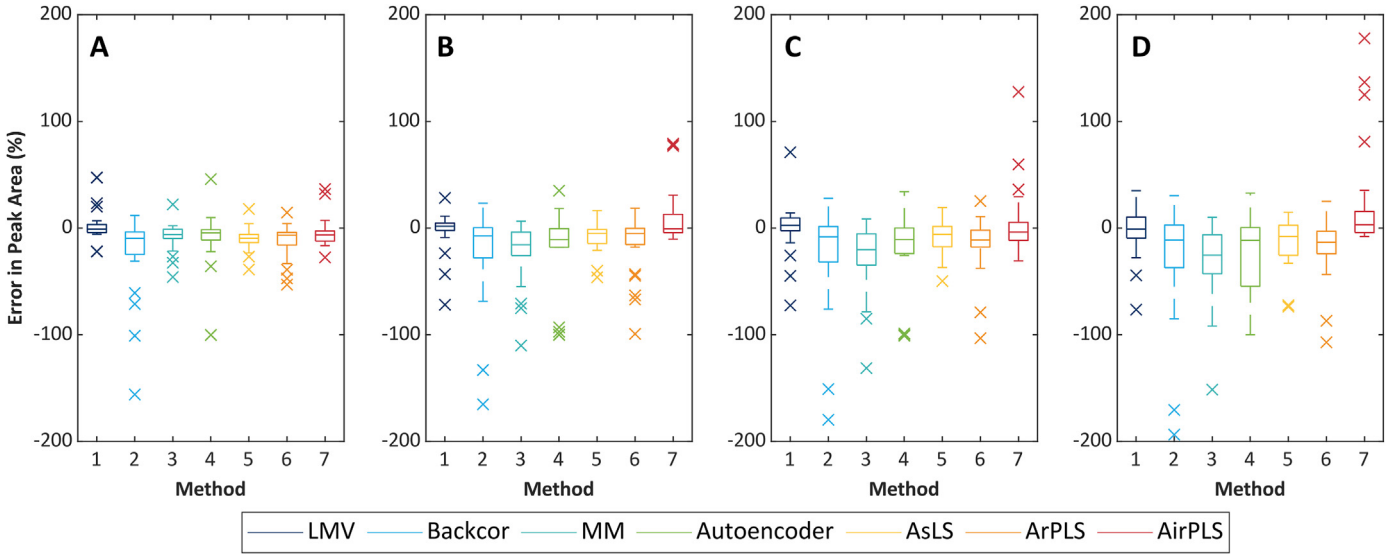


Fig. 7. Relative errors for six different-drift correction methods for the signal shown in Figure8-A for four levels of added noise corresponding to noise intensities of A) $\sigma = 0.02$, B) $\sigma = 0.06$, C) $\sigma = 0.08$ and D) $\sigma = 0.1$. Signals smoothed prior to this analysis using SASS.

It is seen in Fig. 7 that every method performs worse at higher noise levels, as is evident from the larger spread and the higher number of outliers. The deterioration of the performance is strongest for Backcor, MM and Autoencoder methods. Specifically in the case of the largest noise level it is likely that the deterioration of the Autoencoder's performance is because signals with this noise intensity (and correspondingly low SNR) were not included in the training set data. For all methods the reduced performance is not reflected in the corresponding RMSE surfaces (all of which appear relatively similar).

4.3.4. Outlook

Various improvements can still be made to the approach followed in this work and our evaluation method still has some limitations. The greatest challenges to the validity of the approach remain (*i*) the method used to extract the background, peaks and noise from the experimental input data, and (*ii*) whether the experimental data can be accurately recreated using simulations. We have avoided the first issue partially by using a set of experimental blanks for the simulations, which do not contain any peaks. The second point is especially critical when reconstructing peak shapes. In this work we have extracted peak shapes using an automatic curve-fitting approach. This is critically important, as manually evaluating the performance of background-correction and peak-detection algorithms will be incredibly time-consuming for large data sets. A further improvement can be parameter optimization. While a grid-search may currently be the best method for the determination of the parameters, there is still a certain risk that sub-optimal parameters are selected and the approach is very time-consuming. Furthermore, the influence of differently correlated noise on background-correction may be significant and this should be a subject of future study, as different detectors will produce different types of noise.

5. Conclusion

A data-simulation tool has been developed which makes it possible to compare different background-correction and peak-detection methods. We have used this tool to compare a variety of data (pre-)processing methods. From the methods compared in this study, a combination of SASS and arPLS most often resulted in the lowest RMSE. Based on visual inspection this combination also showed the best looking results. However, it did not result in the smallest errors in peak area. The combination of SASS and LMV methods performed best in this respect. In terms of speed the Backcor and LMV algorithms provided the fastest drift correction, generally with evaluation times of less than half a second, while the arPLS algorithm performed slowest. However, this algorithm still generally provided results in less than a second if the number of data points remained below 10 thousand. For the smoothing algorithms nearly all algorithms performed equally, typically with evaluation times of less than 0.1s. It should also be specially mentioned that while in this case a pre-trained Autoencoder model was used, the results of this method were still relatively good. This indicates that in the future it may be possible to perform automatic background correction using similar methods as long as the training set is sufficiently large. The best combination of methods seems to depend on the nature of the background, which implies that it cannot always be a-priori predicted. However, the present study has provided valuable tools and methods to improve quantification in case the true background is unknown.

CRediT authorship contribution statement

Leon E. Niezen: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Peter J. Schoenmakers:** Supervision, Funding acquisition, Project administration, Writing – review & editing. **Bob W.J. Pirok:** Supervision, Funding acquisition, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

LN acknowledges the UNMATCHED project, which is supported by BASF, DSM and Nouryon and receives funding from the Dutch Research Council (NWO) in the framework of the Innovation Fund for Chemistry (CHIPP Project 731.017.303) and from the Ministry of Economic Affairs in the framework of the "TKI-toeslagregeling". BP acknowledges the Agilent UR grant #4354. This work was performed in the context of the Chemometrics and Advanced Separations Team (CAST) within the Centre for Analytical Sciences Amsterdam (CASA). The valuable contributions of the CAST members are gratefully acknowledged. The authors would additionally like to thank Dr. Andrea F.G. Gargano, Mimi J. den Uijl and Denice van Herwerden for providing some of the experimental measurements used in this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2022.339605.

References

[1] R.C. Allen, M.G. John, S.C. Rutan, M.R. Filgueira, P.W. Carr, Effect of background correction on peak detection and quantification in online comprehensive two-dimensional liquid chromatography using diode array detection, J. Chromatogr. A 1254 (2012), https://doi.org/10.1016/j.chroma.2012.07.034.
[2] S. Samanipour, P. Dimitriou-Christidis, J. Gros, A. Grange, J.S. Arey, Analyte quantification with comprehensive two-dimensional gas chromatography: assessment of methods for baseline correction, peak delineation, and matrix effect elimination for real samples, J. Chromatogr. A 1375 (2015), https://doi.org/10.1016/j.chroma.2014.11.049.
[3] D.F. Thekkudan, S.C. Rutan, P.W. Carr, A study of the precision and accuracy of peak quantification in comprehensive two-dimensional liquid chromatography in time, J. Chromatogr. A 1217 (2010), https://doi.org/10.1016/j.chroma.2010.04.039.
[4] I. Latha, S.E. Reichenbach, Q. Tao, Comparative analysis of peak-detection techniques for comprehensive two-dimensional chromatography, J. Chromatogr. A 1218 (2011), https://doi.org/10.1016/j.chroma.2011.07.052.
[5] G. Vivó-Truyols, H.G. Janssen, Probability of failure of the watershed algorithm for peak detection in comprehensive two-dimensional chromatography, J. Chromatogr. A 1217 (2010), https://doi.org/10.1016/j.chroma.2009.12.063.
[6] S. Peters, G. Vivó-Truyols, P.J. Marriott, P.J. Schoenmakers, Development of an algorithm for peak detection in comprehensive two-dimensional chromatography, J. Chromatogr. A 1156 (2007), https://doi.org/10.1016/j.chroma.2006.10.066.
[7] G. Vivó-Truyols, J.R. Torres-Lapasió, A.M. van Nederkassel, Y. vander Heyden, D.L. Massart, Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part I: peak detection, J. Chromatogr. A 1096 (2005), https://doi.org/10.1016/j.chroma.2005.03.092.
[8] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964), https://doi.org/10.1021/ac60214a047.
[9] Y. Liu, X. Zhou, Y. Yu, A concise iterative method using the Bezier technique for baseline construction, Analyst 140 (2015), https://doi.org/10.1039/c5an01184a.
[10] Z. Li, D.J. Zhan, J.J. Wang, J. Huang, Q.S. Xu, Z.M. Zhang, Y.B. Zheng, Y.Z. Liang, H. Wang, Morphological weighted penalized least squares for background correction, Analyst 138 (2013), https://doi.org/10.1039/c3an00743j.
[11] V. Mazet, C. Carteret, D. Brie, J. Idier, B. Humbert, Background removal from spectra by designing and minimising a non-quadratic cost function, Chemometr. Intell. Lab. Syst. 76 (2005), https://doi.org/10.1016/j.chemolab.2004.10.003.
[12] V. Mazet, D. Brie, J. Idier, Baseline spectrum estimation using half-quadratic minimization, in: European Signal Processing Conference, 2015.
[13] J.A. Navarro-Huerta, J.R. Torres-Lapasió, S. López-Ureña, M.C. García-Alvarez-Coque, Assisted baseline subtraction in complex chromatograms using the BEADS algorithm, J. Chromatogr. A 1507 (2017), https://doi.org/10.1016/j.chroma.2017.05.057.
[14] X. Ning, I.W. Selesnick, L. Duval, Chromatogram baseline estimation and denoising using sparsity (BEADS), Chemometr. Intell. Lab. Syst. 139 (2014), https://doi.org/10.1016/j.chemolab.2014.09.014.
[15] H.Y. Fu, H.D. Li, Y.J. Yu, B. Wang, P. Lu, H.P. Cui, P.P. Liu, Y. bin She, Simple automatic strategy for background drift correction in chromatographic data analysis, J. Chromatogr. A 1449 (2016), https://doi.org/10.1016/j.chroma.2016.04.054.
[16] P.H.C. Eilers, H.F.M. Boelens, Baseline Correction with Asymmetric Least Squares Smoothing, Life Sciences, 2005.
[17] P.H.C. Eilers, A perfect smoother, Anal. Chem. 75 (2003), https://doi.org/10.1021/ac034173t.
[18] J.J. de Rooi, P.H.C. Eilers, Mixture models for baseline estimation, Chemometr. Intell. Lab. Syst. 117 (2012), https://doi.org/10.1016/j.chemolab.2011.11.001.
[19] Z.M. Zhang, S. Chen, Y.Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (2010), https://doi.org/10.1039/b922045c.
[20] S.E. Reichenbach, M. Ni, D. Zhang, E.B. Ledford, Image background removal in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 985 (2003) 47–56, https://doi.org/10.1016/S0021-9673(02)01498-X.
[21] I. Selesnick, Sparsity-assisted signal smoothing (revisited), in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2017, https://doi.org/10.1109/ICASSP.2017.7953017.
[22] F. Vogt, Data filtering in instrumental analyses with applications to optical spectroscopy and chemical imaging, J. Chem. Educ. 88 (2011), https://doi.org/10.1021/ed100984c.
[23] D.F. Thekkudan, S.C. Rutan, Denoising and Signal-To-Noise Ratio Enhancement: Classical Filtering, Comprehensive Chemometrics, 2009, https://doi.org/10.1016/b978-0-444-64165-6.02002-4.
[24] S.-J. Baek, A. Park, Y.-J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, The Analyst 140 (2015), https://doi.org/10.1039/c4an01061b.
[25] T.S. Bos, W.C. Knol, S.R.A. Molenaar, L.E. Niezen, P.J. Schoenmakers, G.W. Somsen, B.W.J. Pirok, Recent applications of chemometrics in one- and two-dimensional chromatography, J. Separ. Sci. 43 (2020), https://doi.org/10.1002/jssc.202000011.
[26] C.A. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, Appl. Spectrosc. 57 (2003), https://doi.org/10.1366/000370203322554518.
[27] F. Gan, G. Ruan, J. Mo, Baseline correction by improved iterative polynomial fitting with automatic threshold, Chemometr. Intell. Lab. Syst. (2006) 82, https://doi.org/10.1016/j.chemolab.2005.08.009.
[28] S. Cappadona, F. Levander, M. Jansson, P. James, S. Cerutti, L. Pattini, Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry, Anal. Chem. 80 (2008), https://doi.org/10.1021/ac800166w.
[29] M.F. Wahab, T.C. O'Haver, Wavelet transforms in separation science for denoising and peak overlap detection, J. Separ. Sci. 43 (2020), https://doi.org/10.1002/jssc.202000013.
[30] A. Mani-Varnosfaderani, A. Kanginejad, K. Gilany, A. Valadkhani, Estimating complicated baselines in analytical signals using the iterative training of Bayesian regularized artificial neural networks, Anal. Chim. Acta 940 (2016), https://doi.org/10.1016/j.aca.2016.08.046.
[31] A. Kensert, G. Collaerts, K. Efthymiadis, P. van Broeck, G. Desmet, D. Cabooter, Deep convolutional autoencoder for the simultaneous removal of baseline noise and baseline drift in chromatograms, J. Chromatogr. A 1646 (2021), https://doi.org/10.1016/j.chroma.2021.462093.
[32] E.T. Whittaker, On a new method of graduation, Proc. Edinb. Math. Soc. 41 (1922), https://doi.org/10.1017/s0013091500077853.
[33] C.M. Rader, B. Gold, Digital filter design techniques in the frequency domain, Proc. IEEE 55 (1967), https://doi.org/10.1109/PROC.1967.5434.
[34] A. Graps, An introduction to wavelets, IEEE Comput. Sci. Eng. 2 (1995), https://doi.org/10.1109/99.388960.
[35] J. Carlos 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), https://doi.org/10.1016/j.jmr.2006.07.013.
[36] Z.M. Zhang, S. Chen, Y.Z. Liang, Z.X. Liu, Q.M. Zhang, L.X. Ding, F. Ye, H. Zhou, An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy, J. Raman Spectrosc. 41 (2010), https://doi.org/10.1002/jrs.2500.
[37] M. Daszykowski, K. Kaczmarek, Y. vander Heyden, B. Walczak, Robust statistics in data analysis - a review, Basic concepts, Chemometrics and Intelligent Laboratory Systems 85 (2007), https://doi.org/10.1016/j.chemolab.2006.06.016.
[38] G. Guiochon, Attila Felinger, D. Golshan-Shirazi, A.M. Katti, Fundamentals of Preparative and Nonlinear Chromatography, Second, Elsevier, San Diego, 2006.
[39] M.L. Phillips, R.L. White, Dependence of chromatogram peak areas obtained by curve-fitting on the choice of peak shape function, J. Chromatogr. Sci. 35 (1997), https://doi.org/10.1093/chromsci/35.2.75.
[40] J. Olivé, J.O. Grimalt, Gram-Charlier and Edgeworth-Cramér series in the characterization of chromatographic peaks, Anal. Chim. Acta 249 (1991), https://doi.org/10.1016/S0003-2670(00)83005-6.
[41] J. Grimalt, H. Iturriaga, J. Olive, An experimental study of the efficiency of different statistical functions for the resolution of chromatograms with overlapping peaks, Anal. Chim. Acta 201 (1987), https://doi.org/10.1016/S0003-2670(00)85337-4.
[42] Y. Kalambet, Y. Kozmin, K. Mikhailova, I. Nagaev, P. Tikhonov, Reconstruction of chromatographic peaks using the exponentially modified Gaussian function, J. Chemometr. 25 (2011), https://doi.org/10.1002/cem.1343.
[43] Gottl Friedr, E.B.T. Lipps, G.T. Fechner, Kollektivmasslehre, Phil. Rev. 7 (1898),

https://doi.org/10.2307/2177148.
[44] K.F. Wallis, The two-piece normal, binormal, or double Gaussian distribution: its origin and rediscoveries, Stat. Sci. 29 (2014), https://doi.org/10.1214/13-STS417.
[45] IX, Mathematical contributions to the theory of evolution.—XIX. Second supplement to a memoir on skew variation, in: Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 1916, p. 216, https://doi.org/10.1098/rsta.1916.0009.
[46] G.R. McGowan, M.A. Langhorst, Development and application of an integrated, high-speed, computerized hydrodynamic chromatograph, J. Colloid Interface Sci. 89 (1982), https://doi.org/10.1016/0021-9797(82)90124-2.
[47] A. Felinger, Data Analysis and Signal Processing in Chromatography, First, Amsterdam, 1998.