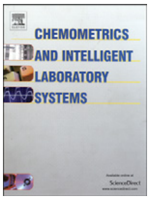
Chemometrics and Intelligent Laboratory Systems 139 (2014) 97–108

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



Baseline correction of high resolution spectral profile data based on exponential smoothing

Xinbo Liu [a], Zhimin Zhang [a,*], Yizeng Liang [a,**], Pedro F.M. Sousa [b], Yonghuan Yun [a], Ling Yu [c]

[a] Institute of Chemometrics and Intelligent Analytical Instruments, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China
[b] Faculty of Sciences and Technology, University of Algarve, Portugal
[c] Shanghai Tobacco Group Co., Ltd., Shanghai, PR China

ARTICLE INFO

Article history:
Received 28 April 2014
Received in revised form 9 September 2014
Accepted 30 September 2014
Available online 20 October 2014

Keywords:
Baseline correction
Exponential smoothing
High resolution spectra
Fast preprocessing

ABSTRACT

Extraction of qualitative and quantitative information from large amounts of analytical signals is difficult with drifted baselines, especially in multivariate analysis. Baseline drift obscures, "fuzzy" signals, and even deteriorates analytical results. In order to obtain accurate and clear results, some effective methods should be proposed and implemented to perform baseline correction before further data analysis. However, most of the classic methods require user's intervention or are prone to variability, especially with low signal-to-noise signals in large data. In this study, a novel baseline correction algorithm based on two-side exponential smoothing algorithm and iterative fitting strategy is proposed. In addition, the iteratively smoothing strategies were creatively implemented in progressively smoothing the residuals between fitted baseline and original signals. This method, named Automatic Two-side Exponential Baseline correction algorithm (ATEB), does hardly require user intervention and prior information, such as peak detection. It's worth noting that the innovative ATEB algorithm has some obvious advantages, especially, when it comes to the processing speed and corrected accuracy of high resolution spectral data with large scale dataset. After a series of benchmarks with high resolution spectral datasets and comparisons with several other popular methods, using various kinds of analytical signals (including hepatocellular carcinoma, MALDI-TOF mass spectrometry, coronary heart disease serum, NMR spectrum and GC–TOF-MS data), the proposed method is found to be accurate, fast, flexible and easy to use on real datasets.

1. Introduction

In general, the baseline drift is usually one of the main issues in chromatograms, mass spectra, Nuclear Magnetic Resonance (NMR) spectra and other spectral data analyses, especially for chemometric multivariate analysis, since the signals from these analytical instruments commonly consist of chemical information, baseline and random noises. Moreover, the baseline drift affects significantly some fundamental chemometric algorithms. Therefore it is necessary to fit the baseline and subtract the background from the analytical signal to alleviate its negative influence. It is worth noting that the influence of the background becomes more difficult to fit and subtract from extremely high resolution datasets, such as NMR spectra and Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectra (MALDI-TOF-MS). According to literature, the classic baseline correction method consists of manually selecting the start and end of a signal peak, and using a piecewise linear approximation to fit a curve as the baseline [1]. However, piecewise approximation is obviously time-consuming and requires much work especially for large scale dataset, and the accuracy depends on the users' experience. As a consequence, several flexible algorithms have been proposed for baseline fitting. Thus, literature from many fields has been published, mainly involving chromatography, vibrational spectroscopy, MALDI-TOF MS, NMR, digital signal processing and statistics.

First of all, let's start from some classic corrected measures for common spectrum. It was Pearson and Walter who proposed the first often cited baseline correction estimation method in 1970 [2]. This classic algorithm works iteratively and inspects which points lie in a specific interval related to their standard deviation, distinguishing the peak points from baseline points simultaneously. Although the algorithm is computationally efficient, it relies on the choice of two parameters (denoted by μ and ν), convergence criterion, and finally the use of a type of smooth curve fitted to the estimated baseline points. Slight mistake in the parameters would lead to unacceptable results. Following the research step of Pearson, many excellent researchers focused their views on improving the baseline correction methods. Liang et al. [3] introduced the roughness penalty method to decrease the influence of the measurement noise, and consequently improved the signal detection and resolution of chemical components with very low concentrations. Later, Shao et al. proposed another novel approach, focusing on the determination of the component number of overlapping

[*] Corresponding author. Tel.:+86 731 88830824, +86 13973110646 (Mobile).
[**] Corresponding author. Tel.:+86 731 8830824; fax: +86 731 8830831.
E-mail addresses: zhangzhimin.csu@gmail.com (Z. Zhang), yizeng_liang@263.net (Y. Liang).

http://dx.doi.org/10.1016/j.chemolab.2014.09.018
0169-7439/© 2014 Published by Elsevier B.V.

chromatograms and baseline corrections, relying on wavelet transform for de-noising [4–8]. In order to correct the background of the measured spectra during elution in chromatograms, asymmetric least squares (ALS) was also introduced by Boelens et al. [9]. Subsequently, Cheung et al. advocated a similar method for preprocessing pyrolysis–gas chromatography–differential mobility spectrometry (Py–GC–DMS) data, via asymmetric least squares (ALS) to eliminate any unavoidable baseline drift [10]. A new idea of morphological weighted penalized least squares (MPLS) algorithm was recommended by Li and Zhan [11], which was successfully applied in the baseline correction of GC–TOF-MS datasets.

Pay attention to the area of vibrational spectroscopy, there also exist a great number of researchers who have proposed a series of algorithms for baseline fitting in it. Firstly, Lieber et al. proposed an approach using least-squares polynomial fitting technique to avoid defects of simple curve fitting [12]. Then, Mazet et al. modified Lieber's method, designing it to minimize a non-quadratic cost function, which was proved to be faster and simpler [13]. Regarding near infrared spectroscopy analysis, Schechter introduced a useful method for the fluctuating non-linear background [14]. Morháč developed a non-linear iterative peak clipping algorithm to correct the baseline of various kinds of spectra, such as IR, NIR and Raman [15]. Zhang et al. succeeded in suppressing fluorescent background in Raman spectroscopy using wavelet and penalized least squares algorithm [16,17]. Liland K.H. proposed a customized baseline correction method which successfully applied in Raman spectra on melted fat from pork adipose tissue [18]. Moreover, lifting wavelet has been applied in baseline corrections for Raman and NMR datasets by Liu and Shao [19].

To the best of our knowledge, many methods previously proposed by other analysts could be effectively applied to small datasets, such as low resolution spectra. However, when it comes to the large scale dataset with high resolution spectra, the research progress has kept rather a slow pace. As early as 1990s, Dietrich et al. applied the second derivative to the signal for peak detection and successfully fitted a NMR baseline with a fifth degree polynomial [20]. Soon afterwards, Moore and Jorgenson recommended a method using a median filter with a very broad window [21]. Even though Moore's method was simple and practical, only peaks with wide baseline segments can be successfully fitted in NMR signals. In 2005, Mirre E. et al. [22] innovatively applied modified asymmetric least-squares algorithm to analyze the reliability of human serum protein profiles generated with C8 magnetic beads assisted MALDI-TOF mass spectra. A practical algorithm designated as adaptive iteratively reweighted Penalized Least Squares (airPLS) was also promoted by Zhang et al. [23,24], by iteratively changing the weights of sum-squared errors between fitted baseline and original signals. Recently, Marcelo R. et al. [25] developed a simple orthogonal background correction (OBGC) method to correct the complex diode array detector (DAD) background signals in fast online comprehensive two-dimensional liquid chromatography (LC × LC). Subsequently, Kuoching Wang et al. [26] presented a novel Distribution-Based Classification method, Baseline Corrector, for automatically estimating the baselines of metabolomics 1D proton NMR spectra. Liu et al. innovatively proposed a novel baseline correction method combing statistic quantile regression algorithm with iterative strategy named selective iteratively reweighted quantile regression (SirQR) [27] which was successfully applied to large datasets, such as GC–TOF-MS and NMR signals. In general, these baseline estimators have been proven fast and flexible in some extent, and some methods can be effectively implemented to different kinds of analytical signals as well.

As mentioned above, many different kinds of chemometric algorithms have been proposed and implemented for treating different kinds of analytical signals, including both classic methods and novel algorithms. Thus, it might be a good idea to change our view to the other analytical fields, for instance, learning something from statistics and digital signal processing. It is noteworthy that Roger Koenker proposed a general approach by employing l_1 regularization methods to estimate quantile regression models for longitudinal data [28]. Eilers et al. developed a fast and effective smoothing algorithm based on penalized quantile regression for the Comparative Genomic Hybridization (CGH) signals [29]. Yu et al. suggested a novel quantile-based Bayesian maximum entropy (QBME) method to account for the non-stationary and non-homogeneous characteristics of ambient air pollution dynamics [30]. In addition, Mencía and Sentana et al. promoted a new algorithm using a location-scale mixture of normal representation of the asymmetric Laplace distribution, transferring different flexible modeling concepts from Gaussian mean regression to Bayesian semiparametric quantile regression [31,32]. Simon Luo and Dave Hale proposed a new digital signal analytical method where a vector shift field was used to represent non-vertical deformations in a seismic image flattening [33]. Robert G. Brown proposed an exponential smoothing method for predicting demand inventory control by an electric computing system [34]. In practice, the exponential smoothing algorithm was first suggested by Robert Goodell Brown in 1956 [35], and then expanded by Charles C. Holt in 1957 [36]. Although the estimates of this exponential smoothing method proposed by Robert are not statistically efficient, they are economically efficient considering the cost of computation. Meanwhile, Prajakta S. Kalekar [37] introduced Holt-Winters Exponential Smoothing algorithm that concentrates on the analysis of seasonal time series data to analyze two models including the Multiplicative Seasonal Model and the Additive Seasonal Model. Furthermore, Joseph J. LaViola Jr. successfully presented a novel Filter-Based Predictive tracking algorithm “double exponential smoothing” for predictive tracking of user position and orientation [38]. When compared against Kalman and extended Kalman filter-based predictors with derivative free measurement models, this method runs approximately 135 times faster with equivalent prediction performance and simpler implementations [39].

According to the previous literature, polynomial fitting, penalized or weighted least square, wavelet, derivatives, and robust local regression have been widely adopted in analytic chemistry for baseline corrections. However, none of these algorithms are entirely perfect for all the practical applications. Each of them has some drawbacks in certain aspects. Firstly, simple manual polynomial fitting methods depend on the analysts' experience for accuracy. Although modified polynomial fitting method is suitable for the most cases, it cannot work well in low signal-to-noise and signal-to-background ratio signals. Secondly, the baseline correction algorithms based on wavelet only remove the baseline successfully when the transformed domain of the signal is well-separated. However, most of the real-world signals do not consent this hypothesis. Thirdly, robust local regression not only demands the specification of the bandwidth and tune parameters by the user, but also requires that the baseline should be smooth and vary slowly. Adaptive iteratively reweighted penalized least square (airPLS) seems to be the optimal automatic baseline correction method. However, airPLS depends on the penalized least squares, which is not robustness enough. Last but not least, the most important problem is that when the analysis signals become an extremely large scale with high resolution, many algorithms cannot offer to process them efficiently and effectively. On the other side, in the smoothing strategies, a classic smoothing algorithm designated penalized least squares was proposed by Whittaker in 1923 [40] , without setting zeroes to the weight vectors at positions corresponding to peak segments. A detailed baseline correction treatment with several related applications has been presented by Eilers [41]. However, the error value of minimized Q_2 is not robustness enough, and can be enlarged by square especially for real-world signals. Moreover, asymmetric least squares, which means asymmetric weights of least squares, has been widely applied to different kinds of baseline correction algorithms, such as ALS method by Eilers [42] and EBS(eliminate the background spectrum) method by Boelens et al. [9] Although the ALS algorithm is effective and useful to some extent, it has some drawbacks. On the one hand, two parameters, namely asymmetric and smoothing parameters, need to be optimized to obtain satisfactory results. On the