Analytica Chimica Acta 940 (2016) 56–64

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Estimating complicated baselines in analytical signals using the iterative training of Bayesian regularized artificial neural networks

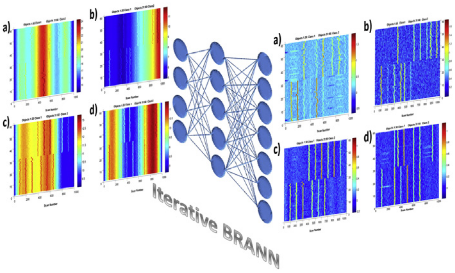
Ahmad Mani-Varnosfaderani [a, *], Atefeh Kanginejad [a], Kambiz Gilany [b], Abolfazl Valadkhani [c]

[a] Chemometrics and Chemoinformatics Laboratory, Department of Chemistry, Faculty of Sciences, Tarbiat Modares University, P.O. Box 14115-175, Tehran, Iran
[b] Reproductive Biotechnology Research Center, Avicenna Research Institute, ACECR, Tehran, Iran
[c] Department of Analytical Chemistry, Chemistry and Chemical Engineering Research Center of Iran, P. O. Box 14335-186, Tehran, Iran

HIGHLIGHTS

- A new baseline correction method based on Bayesian regularized artificial neural networks (iBRANN) has been developed.
- The proposed method can handle different baselines with cave, convex, curve-linear, triangular and sinusoidal patterns.
- Implementation of iBRANN on 1D and 2D data revealed the superiority of this method over iPF, airPLS, MPLS and CC techniques.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 July 2016
Received in revised form
21 August 2016
Accepted 30 August 2016
Available online 13 September 2016

Keywords:
Bayesian regularized neural networks
Baseline correction
Penalized least squares
projected difference resolution

ABSTRACT

The present work deals with the development of a new baseline correction method based on the comparative learning capabilities of artificial neural networks. The developed method uses the Bayes probability theorem for prevention of the occurrence of the over-fitting and finding a generalized baseline. The developed method has been applied on simulated and real metabolomic gas-chromatography (GC) and Raman data sets. The results revealed that the proposed method can be used to handle different types of baselines with cave, convex, curvelinear, triangular and sinusoidal patterns. For further evaluation of the performances of this method, it has been compared with benchmarking baseline correction methods such as corner-cutting (CC), morphological weighted penalized least squares (MPLS), adaptive iteratively-reweighted penalized least squares (airPLS) and iterative polynomial fitting (iPF). In order to compare the methods, the projected difference resolution (PDR) criterion has been calculated for the data before and after the baseline correction procedure. The calculated values of PDR after the baseline correction using iBRANN, airPLS, MPLS, iPF and CC algorithms for the GC metabolomic data were 4.18, 3.64, 3.88, 1.88 and 3.08, respectively. The obtained results in this work demonstrated that the developed iterative Bayesian regularized neural network (iBRANN) method in this work thoroughly detects the baselines and is superior over the CC, MPLS, airPLS and iPF techniques. A graphical user interface has been developed for the suggested algorithm and can be used for easy implementation of the iBRANN algorithm for the correction of different chromatography, NMR and Raman data sets.

[*] Corresponding author.
E-mail address: a.mani@modares.ac.ir (A. Mani-Varnosfaderani).

http://dx.doi.org/10.1016/j.aca.2016.08.046
0003-2670/© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The baseline correction is an active research area in the data processing community and has received great deal of attention in recent years for cleaning the raw data acquired from simple and complicated analytical instruments [1–10]. The problem of baseline drift routinely occurs in common quantification and characterization techniques such as NMR, IR, Raman, ion mobility spectroscopy (IMS), cyclic voltammetry (CV), mass spectrometry (MS), gas chromatography (GC) and liquid chromatography (LC) [10–14]. In NMR spectroscopy, the background artifacts occur mainly due to corruption of the first few points in free induction decay [10,11]. In chromatography, the baseline usually comes from column stationary phase bleed, low frequency detectors, and instrumental instabilities [15]. In most spectroscopic techniques, the background signal occurs because of stray light and non-uniform particle/droplet size and distribution. In recent years, promising hyphenated analytical techniques such as GC × GC [16], GC-IR, LC × LC-MS and GC × GC-MS have been developed and used to solve complicated separation and identification problems in chemistry. In many hyphenated techniques, the baseline drift has also been found as a basic problem, which confines the detection limits and sensitivities for determination of the analytes.

The simplest method to remove baseline drift is fitting a straight line using the first and last points of the collected data. The fitted line is then subtracted from the whole data points to produce baseline free signals. This method works under a simple assumption according to which the baseline pattern only changes linearly over all of the channels of the detector. If this simple assumption fails, which usually does, the algorithm returns many negative data points and omits important signals from the raw data. The iterative polynomial fitting (iPF) [17] strategies have been proposed to tackle this problem. In such methods, a polynomial line is fitted on the collected data while the fitted curve is constrained to be noise-free and smooth. Some new strategies have been applied for reducing the weights of the peak segments when estimating the curved background. These methods are based on an assumption which the maximum and near maximum points in an analytical signal are less affected by the baseline drift and should have less weights for deriving the background. This assumption works well for identifying and removing curved and non-uniform baselines in many cases. Some methods based on asymmetric least squares (AsLS) [18,19], exponential smoothing [20] and quintile regression [21] have been proposed for the automatic selection of the weights of the data points. The main strategy for these algorithms is, more or less, the same and is based on the iterative adaptation of the fitted baseline. Generally, the methods change weights iteratively by estimating a baseline. No weight or small weight is given when a signal is above a fitted baseline. In this respect, as the signal goes below a fitted baseline, it gets much more weights and the baseline is re-estimated, iteratively. As a result, the final baseline is underestimated in the no peak region and the height of the peak might be overestimated by the effect. Recently, the asymmetrically reweighted penalized least square (arPLS) [22] has been proposed to solve this problem. In this method, a logistic function is used to handle the weight problem and finally the peak segments in a signal also receive some weights for estimating the baseline drift. Similar to airPLS [23] and AsLS techniques, the arPLS method requires some parameters to be optimized before implementation on data. The parameters of logistic function and the smoothness parameter (usually named as λ) should be tuned and their thorough values are case dependent. If λ is too large, the algorithm would not catch the curved baseline. On the other hand, a fitted baseline would not follow peaks if λ is too small.

Recently, a new method named morphological weighted penalized least squares (MPLS) has been proposed to handle different types of baselines. The most important step in MPLS is background fitting via morphological opening operation [24]. But it will introduce flaws in the peak region that will change the shape. In order to compensate the shortcomings of the opening operation [24], the rough background fitted by the morphological opening operation and the local minimum value are used as weight vectors of penalized least squares, respectively, for background refinement. Another iterative baseline correction method based on the corner cutting (CC) strategy and Bezier smoothing has been proposed by Liu et al. [25]. In this method, the corner points will be omitted from data in an iterative manner and a smooth baseline will be fitted on remaining points using the Bezier method [26]. The methods of arPLS, airPLS, AsLS and MPLS need to define the λ parameter before implementation on the data. The number of corner-cutting steps in CC baseline correction method should also be optimized for thorough detection of the background [25].

In this paper we implemented the Bayesian regularized artificial neural networks (BRANN) [27] for automatic estimation of the curved and smoothed baselines in analytical chemistry for the first time. The proposed algorithm does not require adjusting and optimizing parameters before data analysis. The results of this method have been compared with those of airPLS, MPLS and CC techniques. The proposed method has been tested to estimate the baseline drift for Raman and GC data. A graphical user interface (GUI) has been developed for the proposed algorithm and is available in the supporting material section.

2. Material and methods

2.1. Baseline estimation using penalized least squares

Suppose **x** is the vector of analytical signal and **z** is a fitted baseline vector. The lengths of both vectors are **m**. The appropriateness of the fitted baseline can be expressed as the sum of squares of the differences between **x** and **z**:

$$\mathrm{A} = \sum_{i=1}^{m} (\boldsymbol{x_i} - \boldsymbol{z}_i)^2 \tag{1}$$

The roughness of the fitted baseline can be expressed as its squared and summed differences:

$$\mathrm{R} = \sum_{i=2}^{m} (\boldsymbol{z_i} - \boldsymbol{z}_{i-1})^2 = \sum_{i=2}^{m} (\Delta\boldsymbol{z_i})^2 \tag{2}$$

The balance of the appropriateness and smoothness can be measured by the following equation:

$$Q = A + \lambda R = \|\boldsymbol{x} - \boldsymbol{z}\| + \lambda\|\boldsymbol{D_z}\|^2 \tag{3}$$

Adjusting the λ parameter brings a compromise between the fitting adequacy of the calculated curve and its degree of smoothness. In order to correct the baseline using the penalized least square algorithm, Cobas [28] and Zhang [29] introduced a weight vector of appropriateness, and set zero to the weight vector at a position corresponding to peak segments of **x**. Here the appropriateness of **z** to **x** is modified to:

$$\mathrm{A} = \sum_{i=1}^{m} \boldsymbol{w_i}(\boldsymbol{x_i} - \boldsymbol{z}_i)^2 = (\boldsymbol{x} - \boldsymbol{z})'\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{z}) \tag{4}$$

where, **W** is a diagonal matrix with *wi* on its diagonal. If peak region are known beforehand, $\mathbf{w}_\mathrm{i}$ can be set to zero in those regions and