ISSN:2231-0711

Kya Abraham Berthe et al | IJCSET(www.ijcset.net) | 2017 | Volume 7 Issue 12, 84-91

# Automatic Baseline extraction based on PCA (Principal Component Analysis) method

Dr Kya Abraham Berthe[1*], [2]Pr Lamissa Diabate, [3]Pr Stephen Reichenbach

[1] *Département de Physique, Facultés des Sciences et Techniques (FST), Université des Sciences Techniques et Technologies BE 3206 Bamako Mali ;*
Email: berthemoni@yahoo.com
[2] *Pr Lamissa Diabate, Deparetement de Genie Industrie, Ecole Nationale d'Ingenieurs-Abderamane Baba Toure Bamako, Mali,*
[3] *Department of computer Ccience and Engineering, University of Nebraska-Lincoln, Nebraska, United States.,*

**Abstract: Recent baseline extraction and correction techniques are based on Penalized Least Square method; which is focused on two mains parameters: weight vector and smooth parameters estimation. Weight vector is computed iteratively based on the difference between original signal and the i$^{th}$ extracted baseline, when the smooth parameters are generally computed empirically. The drawback of these techniques is that the algorithm associated for baseline optimization; which mainly overestimated if the signal is below a fitted baseline and under estimated when the signal is above a fitted baseline. In this paper, we proposed an efficient algorithm for robust baseline extraction; in which the optimal weight vector is computed based on logic distribution function; and, the smooth parameters using PCA method. The new algorithm has been extended to existing extraction methods. Simulations results have shown the effectiveness of the propose algorithm, and the advantage of using multi-smooth parameters for automatic baseline removal.**

Keywords: **Baseline extraction method, Spectral analysis, logic distribution function, Penalized Least Square method, PCA, Optimization**

## I. INTRODUCTION

Data processing techniques dealing with various wavelengths is still a challenge. Mainly all of them share a common pre-processing step; which is the removal of extraneous baseline (background) signal from the data of interest. Baseline can be caused by a large number of factors depending on the type of spectrum [1, 2, 3, 4, 5]. Baseline correction problem is common to many areas of spectrum analysis, a large variety of techniques have been proposed [6, 10, 7, 8, 9]. Accurate algorithm for automatic removal of baseline (RBL) signal [6, 7, 11, 12] is important.

The nature of background noise and additive noise make it is hard to correct or extracted the baseline. Existence of the baseline and random noises can negatively affect qualitatively or quantitatively peaks alignment in spectrum analysis. If we consider the baseline always appears as a sample-independent smooth curve; it should be fitted and corrected routinely to mitigate the negative influence. Efficient peak detection algorithm needs accurate baseline extraction method. Several peak detection algorithms have been proposed [3, 6, 13, 14]. However, different drawbacks have been pointed out as: (i) true signal could be removed during the process; (ii) baseline removal step may get rid of true peaks or created new false peaks [14, 15]. In theory, two approaches are mainly used to eliminate the baseline: (a) first approach need a prior knowledge of the type of noise or signal to be extracted, and (b) the second approach do not need any information of the type of noise, which mostly reflect the reality. Several methods were proposed for baseline elimination using second approach [14, 16]; most of them fit a baseline used polynomial function by cutting out signal peaks iteratively or by using linear constraints. For baseline extraction and optimization, numerous algorithms have been proposed [17, 18, 19, 20]. Example Bivariate shrinkage estimator in stationary domain to avoid removing true peaks in demising step, and zero-crossing lines in multi-scale of derivative Gaussian wavelet is investigated with mixture of Gaussian to estimate discriminative parameters of peaks is recently proposed by Nha et al [15]. To avoid removing true peaks, CWT-based pattern-matching algorithm was also introduced in study by Du et al [21] using Mexican Hat wavelet.

Baseline extraction is also necessary to improve low concentration chemical signal processing; in this approach, Jakob et al [22] introduced the roughness penalty method to reduce the influence of measurement noise. Shao et al [23] proposed wavelet transform for baseline correction. To correct the measured spectra during elution for the background contribution Boeleans et al [24] applied asymmetric least squares regression; Cheung et al [25] proposed the Asymmetric Least Square in order to remove any unavoidable noise in gas chromatography, also a modify least-squares polynomial curve fitting to avoid shortcomings for simple curve fitting have been proposed by Lieber et al [19].

Most of above existing methods need the prior knowledge of the type of noise or compute noise empirically; few of them are based on automatic estimation. In this approach, Zhi-Min et al[1] proposed adaptive iterative reweighted Penalized Least Squares (arPLS) which method does not require any intervention and prior information about noise. He works by iteratively changing weighted of sum squares errors (SSE) between the fitted baseline and original signals, and the weights of the SSE are obtained adaptively using the difference between the previously fitted baseline and the original signals.

Efficient baseline extraction method depends on two parameters: the weight vector, which is computed iteratively;

and the smooth parameters mainly estimated empirically. Based on an extensive review of existing literature, we propose a full algorithm, which allowed an automatic computation of optimal smooth parameters based on PCA method. PCA is commonly used for dimensionality reduction, and optimization [22, 26, 28, 29]. The proposed algorithm for smooth parameters computation is fast. The proposed smooth parameters computation algorithm was been extended to the existing baseline extraction method. We also proposed a new extraction method in which weight vector is computed iteratively based on logic distribution function. Most importantly, this framework supports automated construction of BLR techniques.

The paper is organized as the follows: in next section, we presented a review of recent baseline extraction methods; the optimization problem has been also analyzed. In section three is focused on the framework of the propose method for weight and smooth parameters computation, the simulation and discussion are presented in section four, following the conclusion in section five..

## II. RECENT BASELINE EXTRACTION METHODS REVIEW

### A. Baseline extraction theory

Recent years have seen the effectiveness of least square method for background (called baseline) noise extraction. Background noise signal degrades the accuracy and precision of analysis; it also reduces the detection limit of the instrumental technique. Baseline extraction involve the computation of noise signal from input (original) signal, the extraction produces a heavily biased approximation that does not fit peaks in the input. The goal of spectra baseline correction algorithm is to remove noise from original signal without deteriorated the useful signal. Recent approaches for signal correction or baseline extraction have associated a nonlinear function with OLS (optimality Least Square) and WLS (weighted least squares). Baseline estimation formula can be written as (reference [1, 2, 29]):

$$Q = S + R \tag{1}$$

Where S is the sum of squares residual difference between the original signal or spectrum, and R is the associated penalized function all computed iteratively. The residual difference is definite as:

$$S = \sum_i w_i \left( y - z_i \right)^2 \tag{2}$$

Where $y$ is the original signal, $z_i$ the extracted baseline, and $w_i$ the weight vector at the $i^{th}$ iteration respectively. And, $R$ is the penalized function characterizing the roughness of $z_i$. Computed as:

$$R = \lambda_d \sum_i \left( \Delta^d z_i \right)^2 \tag{3}$$

Where $\Delta$ is the differential matrix, and $d$ the order of differential matrix ( $d = 1, 2, \cdots, n$ ), $\Delta^d z = D_d z = \Delta\left(\Delta^{d-1} z\right)$ in this paper n=2), and $\lambda_d$ the smooth parameter (constant) associated with the differential matrix. In this paper, by selecting n=2 equ 3 is written as:

$$R = \lambda_1 \sum_i \left( \Delta^1 z_i \right)^2 + \lambda_2 \sum_i \left( \Delta^2 z_i \right)^2 \tag{4}$$

Where $\lambda_1$ , $\lambda_2$ are the smoothness factors for the first and second order variation of R. In equa 4 $\Delta^1 z = D_1 z = z_i - z_{i-1}$ and $\Delta^2 z = D_2 z = z_i - 2z_{i-1} + z_{i-2}$ represented the first and second order differential matrix associated with the smooth parameter respectively. The matrix $D_1$ and $D_2$ can be selected as [1, 2]:

$$D_1 = \begin{bmatrix} 1 & -1 & 0 & & 0 \\ 0 & 1 & -1 & & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & & 0 & 1 & -1 \end{bmatrix}, D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & 1 & -2 & 1 & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{bmatrix} \tag{5}$$

The optimization of Q can be resumed as to find the optimal $z$ with a fixe smooth parameters ( $\lambda_1$ , $\lambda_2$ ) written as:

$$z = (w + \lambda_1 D_1^T D_1 + \lambda_2 D_1^T D_1)^{-1} wy \tag{6}$$

Up to a constant factors or coefficients ( $\lambda_1$, and $\lambda_2$ ), the optimization Q should be equivalent to the minimization of following equation [1, 2]:

$$Q = \underset{i}{optimal} \left\{ \sum_i w_i (y - z_i)^2 + \lambda_1 \sum_i \left( \Delta^1 z_i \right)^2 + \lambda_2 \sum_i \left( \Delta^2 z_i \right)^2 \right\} \tag{7}$$

For a number of iterations $I$ ( $I > 1$) select the $z_i$ which minimized equ (7); the optimal $z_i$ must verified equ (6). From equ (7), three cases can be observed according the value of the smooth parameter: i) if $\lambda_1 \neq 0$ and $\lambda_2 = 0$ , the first smooth parameter is only to extracted the baseline, ii) if $\lambda_1 = 0$ and $\lambda_2 \neq 0$ , the secondly smooth parameter is only to extracted the baseline, and iii) if $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$ , the two smooth parameters are used for baseline extraction.

### B. Optimization Framework

The optimization framework is focused on two main parameters: (a) weight vector $w_i$ computation, which element are the diagonal matrix of $w$ and, (b) the choice of smooth parameters ( $\lambda_1$ , $\lambda_2$ ) which should control the roughness of the baseline signal.

1) *Weight vector computation:* Efficient algorithm for weight vector computation is the main issue for efficient baseline extraction. Since 1994, Eilers et al [29], have proposed ALS (Asymmetric Least squares) method using probability approach to compute the weight, where the weight component $w_i$ is definite according the sign of the

different between the original y signal and the estimated noise signal $z_i$. Mostly the recent weight value attributions are based on analysis of the sign $d_i$ from the following equation:

$$d_i = y - z_i \tag{8}$$

Which can be resume as: $w_i$ is $p$ if $y - z_i \le 0$ and $1-p$ if $y - z_i > 0$ for each iteration, where $p$ (optimal value $p = 0.025$) is a constant belonging to interval [0,1]; for ALS (Asymmetric Least Square) method. The constancy of $w_i$ components. In this respect, Zhang et al[1] proposed arPLS (adaptive iterated re-weighted Penalized Least Square) method. The weight assignment can be resumed as the follows: for original signal ( $y$ ) greater than the candidate of the baseline ( $z_i$ ), noise can be regarded as a part of the peak; thus weight is set to zero; otherwise the weight vector $w_i$ is obtained adaptively using exponential function [1]. The problem of the method is the final baseline is underestimated in the no peak region and the height of peaks might be overestimated in single signal baseline extraction. To resolve this drawback Sung-June et al [2] 2015, proposes a partially balanced weighting called arPLS (asymmetric re-weighted Penalized Least Square) method, they computed the weight vector by introducing the so call logistic function which is an exponential function [2]. The main drawback of above methods is that the smooth parameters mostly are estimated empirically.

2) *Smooth parameters:* Research focused on data smoothing (noise data) in baseline extraction for noisily data is based on penalized regression splines smoothing methods [7] characterized by the introduction of smooth parameters. The idea of using differences in a penalty goes back at least to Whittaker [30] in 1923. Recently some authors combine Gaussian Mixture Model (extreme value) and HP filtering to compute an efficient weight vector [31].

Several methods or techniques are used to estimate the smooth factors $\lambda_1$ , $\lambda_2$ ; but the common approach is the empirical estimation or computation, based on [32]. In this approach, Gianluca [12] in 2013 used two differential matrix (first and second) and chooses $\lambda_1 = \lambda_2 = \lambda$ in case of non-isotropic smoothing data, where $\lambda$ a positive constant selected empirically. Also in 1994 Eilers et al used both a first ( $\lambda_1$ )- and second ( $\lambda_2$ )-order penalty to control the smoothness to analysis data by using the following formula:

$$\begin{cases} \lambda_2 = \lambda^2 \\ \lambda_1 = \alpha\lambda \end{cases} \tag{9}$$

Where $\alpha$ is called the pleasant factor depending on the type of signal, and $\lambda$ is a constant selected in order to keep the impulse response from becoming non-positive [29]. Some authors [9, 12] use cross-validation method to determine $\lambda_2$ and $\lambda_1$. or find some relation between them.

## III. PROPOSED METHOD

### A. Weight vector estimation

The propose method uses logic distribution function for weight vector $w_i$ optimization, which function is widely applied in signal processing [33]. Let resume our proposed weight vector estimation method, we denoted by $d_i^-$, $d_i^+$ be the set of data for $d_i \le 0$ and $d_i > 0$ respectively. We used the partial balance asymmetric weights logic distribution function defines as.

$$w_i = \begin{cases} \mathrm{f}(d_i^-, m_-, \sigma_-) & if \quad d_i \le 0 \\ 1 \quad if \quad d_i > 0 \end{cases} \tag{10}$$

Where $m_- = mean(d < 0)$ , $\sigma_- = std(d < 0)$ the mean and standard deviation of $d_i^-$; and $m_-$, $\sigma_-$ , and the function f is definite as:

$$w_i = \begin{cases} \mathrm{f}(d_i^-, m_-, \sigma_-) = \dfrac{1}{1 + e^{(d(d \le d) + \sigma_-)/m_-}}, & if \ d \le 0 \\ 1, \ if \ d > 0 \end{cases} \tag{11}$$

For each $w_i$ , the smoothness of baseline function $z_i$ depend on smooth parameter, computed used PCA method

### B. PCA for Smooth factors computation

1) *Introduction:* Principal component analysis is a prevalent data reduction tool that transforms the data orthogonally and reduces its dimensionality. It is an important well-studied subject in statistic and signal processing. PCA is a well-known statistical technique that has been widely applied to solve important signal-processing problems like future extraction, signal estimation [34, 35]. We proposed a new approach of smooth parameters computation using PCA. From equ (7) two steps are used; firstly we used PCA to estimate the eigenvalues; then the optimal smooth parameters are computed based on some appropriated formula. The detail of the method and simulation is presented in the next section.

2) *Mathematical Approach:* As mentioned above, from equ (6), according smooth parameters cases, we have can write the following equation.

$$\begin{cases} z = \left(w + D_1 D_1^T\right)^{-1} \mathrm{wy} \qquad (\lambda_1 = 1, \ \lambda_2 = 0) & (12.\,a) \\ z = \left(w + D_2 D_2^T\right)^{-1} \mathrm{wy} \qquad (\lambda_1 = 0, \ \lambda_2 = 1) & (12.\,b) \\ z = \left(w + D_1 D_1^T + D_2 D_2^T\right)^{-1} \mathrm{wy} \ \ (\lambda_1 = 1, \ \lambda_2 = 1) & (12.\,\mathrm{c}) \end{cases}$$

For each case (equ (12, 1), (12, b) and (12, c)), we denote by $z_i$ the value of smooth vector at iteration $i^{th}$. To estimate the eigenvalues, we used the following approach: firstly, let $z_1, z_2, \cdots, z_i$ be the set of $z_i$ ( $i > 1$ ); where $z_i$ is a $1 \times l$ dimension ( $l$ the length of $z_i$ ) vector at i[th] iteration obtaining using equ 12. For $i > 1$, we compute the mean vector as a single vector by:

$$z_{aver} = \frac{1}{i}\left(z_1 + z_2 + \cdots + z_i\right) \quad (13)$$

In order to re-center the data, we subtracted $v_{aver}$ from each vector $z_i$. Secondly, we definite the matrix $B$ as a $i \times l$ dimension matrix whose $i^{th}$ column is definite by $z_i - z_{aver}$, so $B$ can be written as:

$$z = [z_1 - z_{aver}, \cdots, z_i - z_{aver}] \quad (14)$$

than, we definite the covariance matrix S as:

$$S = \frac{1}{i-1} B \times B^T \quad (15)$$

Where $B^T$ is the transposed matrix of B; the dimension of S is $i \times i$.

*3) Proposed Algorithm:* Let $S_i$ be the symmetric and positive matrix S with dimension $i \times i$, obtained at $i^{th}$ iteration. We denote by $D_i$ the eigenvalue matrix and $E_i$ the eigenvalue vector ($1 \times i$ dimension) composed of $D_i$ diagonal elements; $E_i$ can be written as $E_i = \left(\eta_1, \eta_2, \eta_3, \cdots, \eta_r\right)$ where $r = 1, 2, \cdots, i$. Each eigenvalue $\eta_r$ can be viewed as an estimation of noise variance [19, 23, 35, 36]. To select the $i^{th}$ iteration $E_i$, we used the decreasing rule; i.e. each $E_i$ vector component elements should satisfied to the following relation $\eta_r > \eta_{r-1} > \cdots > \eta_1 > 0$. We denoted by $E_j$ the eigenvalue vector which satisfied to the decreasing rule. For each $E_j$, we denote by $D_j$, $Z_j$ and $S_j$ the associate diagonal, baseline and covariance matrix respectively; with $Z_j = [z_1, \cdots, z_i]$ and $j = 1, 2, \cdots, m$ where $m$ is the number of matrix satisfying the decreasing rule. To estimate the optimal eigenvalue vector $E_j$, we computed the eigenspread coefficient definite as $C_{spr}(j) = \eta_i / \eta_1$, which characterize the convergence speed . The smaller eigenspread coefficient will be; faster and smoother the extracted baseline is [26]; so the minimum $C_{spr}(j)$ value of coefficient correspond to the optimal $j_{op}$, $E_{j_{op}}$, $D_{j_{op}}$, $Z_{j_{op}}$ and $S_{j_{op}}$ respectively.

*4) Optimal smooth parameters estimation:* Using equ 12 and based on the propose algorithm (section III.2.3), we computed the smooth parameters for each case (12, a; 12, b; 12, c ) according the value of $j_{op}$.

- *Smooth curve based on first differential matrix only:*

In this case only $\lambda_1$ is used as smooth parameter; for $j_{op}$ we computed $\lambda_{op}$ using the following formula:

$$\lambda_{op} = \text{med}^2 \times \frac{N_{z_{\min}}}{N_{z_{\max}}} \quad (16)$$

Where

$$\begin{cases} \text{med} = \text{median}\left(E_{j_{op}}\right) \\ N_{z_{\min}} = \min(Z_{j_{op}}) = \min\limits_{\substack{u=1 \text{ to } i \\ n=1 \text{ to } l}} \left( \sum_{l=1}^{L} \left| z_{j_{op}}^{u,n} \right| \right) \\ N_{z_{\max}} = \max(Z_{j_{op}}) = \max\limits_{\substack{j=1 \text{ to } i \\ n=1 \text{ to } l}} \left( \sum_{l=1}^{L} \left| z_{j_{op}}^{u,n} \right| \right) \end{cases} \quad (17)$$

With $l = length\left(z_{j_{op}}^{u}\right)$

- *Smooth parameter based on second differential matrix only:* For this case only $\lambda_2$ is used as smooth parameter; the optimal smooth parameter is computed by the following formula:

$$\lambda_{op} = \eta_{\max}^2 \times \frac{N_{z_{\min}}}{N_{z_{\max}}}, \ with \ \ \eta_{\max} = \max_{\eta_i}\left(E_{j_{op}}\right) \quad (18)$$

Where $N_{z_{\max}}$ and $N_{z_{\min}}$ are estimated used equ 17.

- *Smooth parameter estimation based on first and the second differential matrix***:** The two smooth parameters ( $\lambda_1$, $\lambda_2$ ) are computed, the relation between them is defined as:

$$\begin{cases} \lambda_2 = \lambda_{op} \\ \lambda_1 = \alpha \lambda_{op} \end{cases} \quad (19)$$

Where the so call pleasant coefficient [1] $\alpha$ is compute as:

$$\alpha = \frac{\text{median}(E_{j_{op}})}{\left\| S_{j_{op}} \right\|_F} \quad (20)$$

and $\| \ \|_F$ is the frobenius norm of $S_{j_{op}}$.

## IV . Simulation Results and Discussions

### A. Smooth parameters

In this paper, we used the same data that have been using by Wong et al in 2005 [37]. Equ 15-18 are used for optimal smooth parameters estimation; which result has been presented in the table 1. Table 1 contains the optimal value of $j_{op}$ and the smooth parameters $\lambda_{op}$ for the ASL, arPLS, asPLS and the proposed method.

From table 1, we find that for the same method, the optimal iteration $j_{op}$ "corresponding to the maximum spreadvalue" and smooth parameter $\lambda$ may be different and depends on the case. The smooth parameters $\lambda_1$ computed in case one from equ 15-16 are 1668.7, 176, 534.7 and 384.5

for ALS, arPLS asPLS and the proposed method respectively, with the corresponding $j_{op}$ (1, 2, 3 and 5).

The optimal smooth parameters according the two other cases and their optimal iteration are also included in table 1. These values will be used to determine the optimal baseline in next section.

Table 1: optimal iteration value $j_{op}$ and smooth parameters $\lambda_{op}$ obtaining from the equ 15-18.

| Methods | $\lambda_1 \neq 0, \lambda_2 = 0$ | |
|---|---|---|
| | $j_{op}$ | $\lambda_1$ |
| ***ALS*** | 1 | 1668.7 |
| ***arPLS*** | 2 | 176 |
| ***asPLS*** | 3 | 534.7 |
| ***Proposed*** | 5 | 384.5 |

*Table 1 (a): used only first differential matrix to extract baseline*

| Methods | $\lambda_1 = 0, \lambda_2 \neq 0$ | |
|---|---|---|
| | $j_{op}$ | $\lambda_2$ |
| ***ALS*** | 2 | 948731.8 |
| ***arPLS*** | 1 | 408038.5 |
| ***asPLS*** | 2 | 1264547 |
| ***Proposed*** | 2 | 4643134 |

*Table 1 (b): used only second differential matrix to extract baseline*

| Methods | $\lambda_1 \neq 0, \lambda_2 \neq 0$ | | |
|---|---|---|---|
| | $j_{op}$ | $\lambda_2$ | $\lambda_2$ |
| ***ALS*** | 1 | 2466140 | 785.2 |
| ***arPLS*** | 1 | 364151,4 | 201.75 |
| ***asPLS*** | 3 | 625927.3 | 375.65 |
| ***Proposed*** | 1 | 2039706 | 793.917 |

*Table 1 (a): used first and second differential matrix to extract baseline*

B. Baseline optimization

***1)*** *Introduction:* Optimal baseline is extracted using the data (smooth parameters) of table 1. For example in case if first differential matrix ($\lambda_1 \neq 0, \lambda_2 = 0$) is only used to extracted the baseline, the smooth parameters $\lambda_1$ values are the following value: 1668.7, 176, 534.7 and 384.5 for ALS, arPLS, asPLS and Proposed method respectively (table 1 column 3). The same approach is used if the second differential matrix is only used to smooth the extracted baseline ($\lambda_1 = 0, \lambda_2 \neq 0$) the data of column 5 of table 1 should be used as the smooth parameters $\lambda_2$ according each method. Than if in this case the combine smooth parameters ($\lambda_1 \neq 0, \lambda_2 \neq 0$) for optimal baseline extraction, data of column 6 and 7of table 1 as $\lambda_1$ and $\lambda_2$ respectively.

***2)*** *Method estimation:* As mentioned in the introduction for each case and method we associated the corresponding smooth parameters from equ 10. To estimate the optimal extracted baseline: (i) firstly, we fixed the same iteration number for each case and method. (ii) secondly, we estimated the degree of smoothness by comparing the extracted wave (baseline) to the fundamental. We denoted $v_u$ the extracted baseline vector at $u^{th}$ iteration ($u = 1, 2, \cdots, U$); and $\delta_u$ the fundamental wave at the $u^{th}$ iteration, definite as:

$$\delta_u = \sin(\varpi_u t + \varphi_u) \tag{21}$$

where $\varpi_u = 2\pi / T_u; T_u = 2(t^u_{\max} - t^u_{\min})$ with $t^u_{\max}$ and $t^u_{\min}$ values of t corresponding to the maximum and minimum value of $v_u$ ; and $\varphi_u$ the initial phase (value of $v_u$ for $u = 1$).

To improve our analysis, we definite the similarity coefficient as:

$$sim_u = (\delta_u - v^u) / \delta_u \tag{22}$$

The smaller is $sim_u$, the smoother the extracted baseline $v^u$ will be. (ii) thirdly, to strength our estimation, we introduce other statistical approaches called Contrast Noise Ratio (CNR) which are efficiently used for complex noise baseline extraction [38]; two different concepts are used; the first is focused on amplitude of the activation signal (amplitude), by definite CNR as the amplitude measurement to the extracted baseline variance [39, 40] defined by:

$$CNR_{u,1} = 10\log_{10}\left(\frac{A^2}{\sigma^2_{v^u}}\right) \tag{23}$$

Where $A$ is the absolute value of amplitude of the original signal with baseline which is the difference between the baseline of the signal and the signal peak. While the second definition incorporates the standard deviation of the activation as the signal of interest; based on the ratio of variance in dB scale [39, 40] as:

$$CNR_{u,2} = 10\log_{10}\left(\frac{\sigma^2_X}{\sigma^2_{v^u}}\right) \tag{24}$$

Where $\sigma_X$ is the variance of the original signal (signal with noise).

In theory, the smaller is the $SNR_u$ ratio the better will be the proposed method, but as mentioned on above the drawback is that some real peak can be eliminate. This approach will be more efficient if we have a prior knowledge of noise of baseline to be extracted; which is not the case in practice. In case of multiple events causing several peaks in the signal and the timing of the stimuli will have an effect on the height of the peak [41].

For complex and multiple peak signal baseline extraction, the amplitude of the signal could be either the difference between the baseline and the maximal height of the signal, or the mean amplitude over all peaks [42].

To selected the optimal extracted baseline, we mainly focused on two feature the smoothness characterized by $sim_u$ and the two features ratios which are $CNR_{u,1}$ and $CNR_{u,2}$; by using the $CNR_u / sim_u$ ratio which simulation is represented on table 2.

| Method | $\lambda_1 \neq 0 \quad \lambda_2 = 0$ | |
|---|---|---|
| | CNR2/sim | CNR1/sim |
| Als | 310.8367 | 514.6228 |
| arPLS | 307.3595 | 367.1587 |
| asPLS | **582.3476** | **674.352** |
| Proposed | 574.838 | 666.4418 |

Table 2 (a): simulation ration ($CNR_u / sim_u$) if we used only first matrix differential matrix for baseline extraction and smooth.

| Method | $\lambda_2 \neq 0 \quad \lambda_1 = 0$ | |
|---|---|---|
| | CNR2/sim | CNR1/sim |
| Als | 251.6472149 | 256.287798 |
| arPLS | 1044.03966 | 1155.04533 |
| asPLS | **1661.690079** | **1731.2072** |
| Proposed | 1050.298343 | 1217.6105 |

Table 2 (b): simulation ration ($CNR_u / sim_u$)used only second matrix differential matrix for baseline extraction and smooth.

| Method | $\lambda_1 \neq 0 \quad \lambda_2 \neq 0$ | |
|---|---|---|
| | CNR2/sim | CNR1/sim |
| Als | 349.5602094 | 412.8795812 |
| arPLS | 512.4619423 | 637.9120735 |
| asPLS | 1365.536232 | 1406.717391 |
| Proposed | **1617.45339** | **1750.877119** |

Table 2 (a): simulation ration ($CNR_u / sim_u$) using only first and second differential matrix for baseline extraction and smooth.

The simulation result is presented in fig 1 represented (a) the extracted baseline in case of only first differential matrix ($\lambda_1 \neq 0, \lambda_2 = 0$); (b) ; represented the extracted baseline if we used only second differential matrix to extracted the baseline ($\lambda_1 = 0, \lambda_2 \neq 0$), and (c) the extracted baseline in case of first and second differential matrix ($\lambda_1 \neq 0, \lambda_2 \neq 0$)

The simulation results using the data from table 2 have been presented in fig 1.

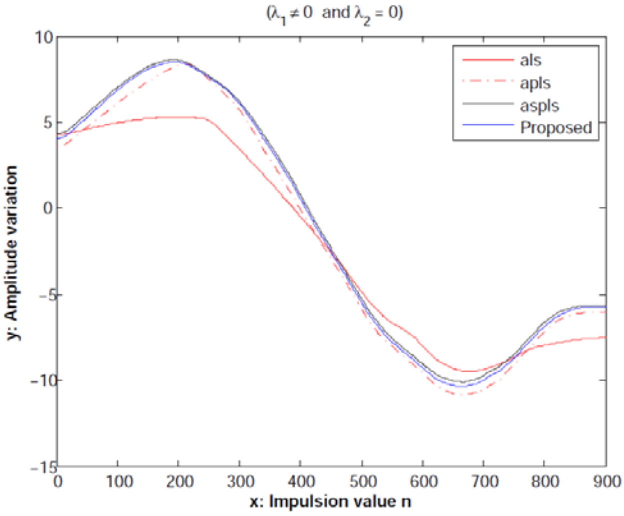


*Fig 1 (a): used only the first differential matrix*

From table 2 (a), the maximum ratio $CNR/sim$ values are 582.34757 and 674.352 for $CNR_2/sim$ and $CNR_1/sim$ respectively asPLS method; in case of using only the first differential matrix to extract the baseline. In conclusion using only first differential matrix for baseline extraction asPLS method is more efficient than ALS, asPLS and our propose extraction method. Which is confirmed fig 1, a

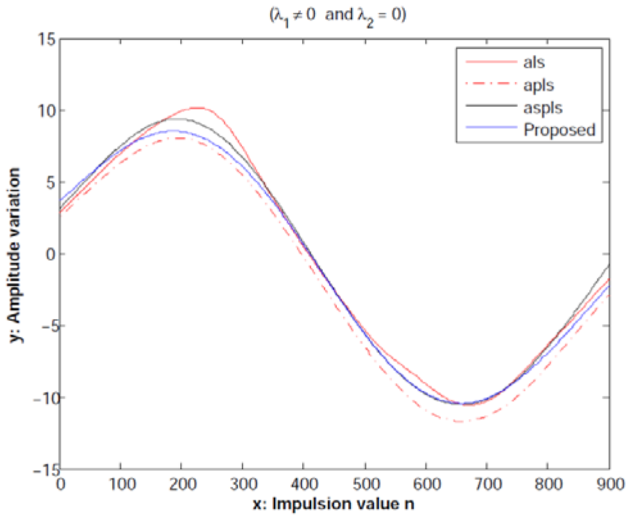


*Fig 1 (b): used only second first differential matrix*

Fig 1, b; represented the extracted baseline if we used only second differential matrix to extracted the baseline ($\lambda_1 = 0, \lambda_2 \neq 0$). The smoothness of the extracted baseline confirm the result of table 2 (b). The extracted baseline based on asPLS is butter (ratio $CNR_2/sim$, $CNR_1/sim$ is

1661.69008, 1731.2072 respectively) than other existing method and the proposed method.

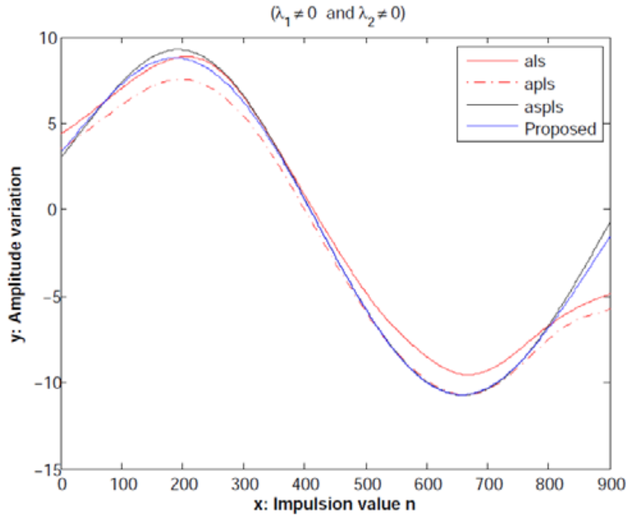


*Fig 1 (c): used the first and second differential matrix*

Fig 1, c; shown the extracted baseline in case of $\lambda_1 \neq 0, \lambda_2 \neq 0$ ; that is mean the first and second differential matrix have been used to extract the baseline. From the table 2 the optimal extracted baseline method is the proposed method (ratio $CNR_2/sim$ , $CNR_1/sim$ is 1617.45339, 1750.877119 respectively). So, the proposed method is a butter method to extract baseline extraction in case of using two smooth parameters. Fig 1. d; represented the fig 1 c with the original signal y.

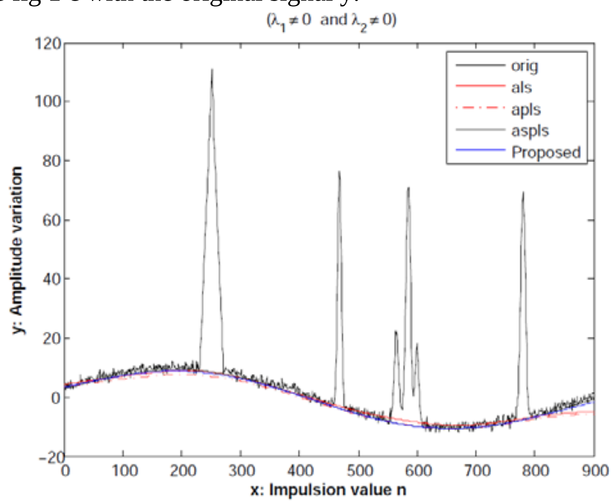


*Fig 1 (d):Optimal baseline extracted and original signal plotted*

## V. Conclusion

In this paper, we firstly make a review of baseline extraction method, and secondly, we the optimization is also be investigated. We proposed and efficient algorithm of signal baseline extraction by computing the smooth parameters based on PCA without using empirical approaches and without any prior knowledge of associate noise. The smooth parameters are calculated using the eigenvalues, several conditions are imposed to compute the optimal smooth parameters values. We also proposed a new method for baseline extraction based on the extreme values. In the proposed method the weight vector is estimated based on the characteristics of the signal. Comparing to the well know existing baseline extraction method, the simulation results show the efficiency of the proposed method in case we used two smooth parameters to extract the baseline.

The new algorithm allowed an automatic baseline extraction without empirical estimation of smooth parameters. This eliminates the need for users to spend valuable time learning the internals of existing approaches in order to facilitate educated choices about which method is best for their application.

## References


[1] Zhi-Min Zhang, Shan Chen and Y-zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. The Royal Society of chemistry Analysis 2010, 135, pp 1138-1146.

[2] Sung-June Baek, Aaron Park, Young Jin Ahn, and Jaebum Choo. Baseline correction using asymmetrically reweighted penalized least squares smoothing. Royal society of chemistry. Analyst 2015, 140 pp 250-257.

[3] Saer Samanipour, Petros Dimitriou-Christidis Jonas Gros Aureline Grange, J. Samuel Arey. Analyte quantification with comprehensive two-dimensional gas chromatography : Assessment of methods for baseline correction, peak delineation, and matrix effect elimination for real samples. Journal of Chromatography A. 1375 2015, pp 123-139

[4] R. Chartrand and W. Yin, “Iteratively reweighted algorithms for compressive sensing,” in IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 2008, pp. 3869– 3872.

[5] Jiangtao Peng, Silong Peng, An Jiang Jiping Weib, Changwen LiJie Tan. Asymmetric least squares for multiple spectra baseline correction. Analytica Chimica Acta 683 2010,pp 63-68.

[6] Paul G. Stevenson, Xavier A. Conlan, Neil W. Barnett. Evaluation of the asymmetric least squares baseline algorithm through the accuracy of statistical peak moments. Journal of chromatography A. 1284,(2013 pp 107-111.

[7] Adele Kuzmiakova, Ann M. Dillner, and Satoshi Takahama. An automated baseline correction propocol for infrared spectra of atmospheric aerosols collected on polytetra-fluoro ethylene (Teflon) filters. Atmospheric Measurement Technique 9, 2016, pp 2615-2631

[8] Mayrim vega-Hermandez, Eduardo Martinez-Montes, Jose M. Sanchez-Bornot, Asustin Lage-Castellanos and Pedro A. Valdes-Sosa. Penalized least squares methods for solving the EEG inverse problem. Statistica Sinica 18, 2008, pp 1535-1551

[9] X. Liu, M. Tanaka, and M. Okutomi. Single –image noise level estimation for blind denoising. Image processing, IEEE Transactions on, 22 (12):2013, pp 5226-5237

[10] Paul H. C. Eilers, Brian D. Marx and Maria Durban. Twenty years of P-splines. SORT 39 (2) July-December 2015, pp 1-38

[11] Bin Wang, Wenzhong Shi, and Zeland Miao. Comparative Analysis for Robust Penalized Spline smoothing Methods. Hindawi Puplishing Corporation, Mathematical Problems in Engineering. Vol 2014, Article ID 642475, pp 11

[12] Frasso Gianluca, Jaeger, Jonathan, Lambert, Philippe. Penalized smoothing approaches for PDEs. 21st meeting of Belgian statistical Sociaty 9-11 October 2013.

[16] P. J. Huber, “Robust estimation of a location parameter” Annals of Mathematical Ststistics, Vol. 35, No. 1, 1964, pp 73-101.

[13] Chaoling Yang, Silong Peng, Qiong Xie, Jiangtao Peng, Jipin Wei, Yong Hu. FTIR spectral subtraction based on asymmetric least squares. 2011 4th International conference on Biomedical Engineering and Informatics (BMEI).

[14] Sabine Schnabel. Expectile smoothing[Gianluca Frasso.]: new perspectives on asymmetric least squares. An application to life expectancy Proefschrift Universiteit Utrecht, Utrecht, Printed by Ipskamp Drukkers, Enschede, The Netherlands. 2011

[15] Nha Nguyen, Heng Huang, Soontorn Oraintara and An Vo. Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet. (Bioinformatics, vol 26. ECCB, 2010 pp 659-i665

[16] J. Zhao, H. Lui, D. I. McLean and H. Zeng. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra. Appl. Spectrosc., 2007, 61, pp 1225–1232.

[17] X. G. Shao, A. K. M. Leung and F. T. Chau. Wavelet a new trend in chemistry. Chem. Res., 2003, 36, 276–283.

[18] S. N. Wood Modelling and smoothing parameter estimation with multiple quadratic penalties. J. R. Statist Soc. B. 2000 62, Part 2, pp 413-428.

[19] C. A. Lieber and A. Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological Raman spectra. , Appl. Spectrosc., 2003, 57, pp 1363– 1367.

[20] Paul H.C. Eilers, Brian D. Marx et Maria Durban. Twenty years of P-splines. SORT 39 (2) July— December 2015, pp 1-38

[21] Dubey, S. D. Eliane C). A new derivation of the logistic distribution. Naval Research Logistics Quarterly, 16, 7, 15, 1960, pp 37-40.

[22] Jakob Sigurðsson Smooth noisy PCA Using a First Order Roughness Penalty A Thesis Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in Electrical and Computer Engineering at the University of Iceland 2011, pp 36-46

[23] X. G. Shao, W. S. Cai and Z. X. Pan, Chemom. Intell. Lab. Syst., 1999, 45, pp 249–256.

[24] H. F. M. Boelens, R. J. Dijkstra, P. H. C. Eilers, F. Fitzpatrick and J. A. Westerhuis, J. New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. Chromatogr., A, 2004, 1057, pp 21–30.

[25] W. Cheung, Y. Xu, C. L. P. Thomas and R. Goodacre. Discrimination of bacteria using pyrolysis-gas chromatography-differential mobility spectrometry (Py- GC-DMS) and chemometrics. Analyst, 2009, 134, pp 557–563.

[26] FeiI. T. Jolliffe. Principal Component Analysis Second Edition. Springer 2002, pp 45-100

[27] Robert Reris and J. Paul Brooks. Principal Component Ana;ysis and Optimization A Tutorial. 14th INFORMS Computing Sociaty Conference Richmond, Virginia, January 11-13, 2015; pp. 212-225

[28] Chaman Lal Sabharwal and Bushra Anjum. Principal component analysis as an integral part of data minimg in Health informatics. Proceedings of 31st International society conference on computers and their applications CATA 2016, pp. 251-256 April 05, 2016.

[29] P.H. Eilers, Parametric time warping, Anal. Chem. 76 (2), 2004, pp 404–411

[30] Whittaker, E. T. On a new method of graduation. Proceedings of the Edinburgh Mathematical Association, 78, 1923, pp81-89.

[31] Eilers P H Goeman J. J. Enhancing scartterplots with smoothed densities. Bioinformatics 2004 Mars 22; 20 (5): pp 623-8.

[32] Dinesh Kumar, Umesh Singh and Sanjay Kumar Singh. A Method of Proposing New Distribution and its Application to Bladder Cancer Patients Data. Journal of Statistics Applications & Probability Letters 2, No. 3, 2015, pp 235-245

[33] K. Nidhun, C. Chandran. Importance of Generalized Logistic Distribution in Extreme Value Modeing. Applied Mathematicas, 2013, 4, pp 560-573

[34] J. Mao, A.K. Jain, “Artificial Neural Networks for Feature Extraction and Multivariate Data Projection,” IEEE Transactions on Neural Networks, vol. 6., no. 2, 1995, pp. 296-317.

[35] Y. Cao, S. Sridharan, M. Moody, “Multichannel Speech Separation by Eigendecomposition and its Application to Co-Talker Interference Removal,” IEEE Transactions on Speech and Audio Processing, vol. 5, no. 3, 1997, pp. 209-219.

[36] Guangyong Chen , Fengyuan Zhu, and Pheng Ann Heng An Efficient statistical method for image noise level estimation. ICCV open paper, provided by the Computer Vision Foundation, 2015

[37] Wong, Jason W. H., Durante, Caterina, and Cartwright, Hugh M., Application of Fast Fourier Transform Cross- Correlation for the Alignment of Large Chromatographic and Spectral Datasets. Analytical Chemistry 77 (17), 2005, pp 5655-56.

[38] Robert B Nortrop. Instrumenetation and measurements. Seond edition Taylor & Francis Group 2005 pp 9

[39] Vincent T, Risser L, Ciuciu P Spatially adaptive mixture modeling for analysis of FMRI time series. IEEE Transactions on Medical Imaging 29: 2010 pp1059–1074.

[40] Yousuf M. Soliman. Mean and principal curvature estimation from noisy cloud data of manifolds embedded in $R^n$ Noisy curvature estimation May 10, 2016.

[41] Calhoun V, Adali T, Stevens M, Kiehl K, Pekar J Semi-blind ICA of fMRI: A method for utilizing hypothesis-derived time courses in a spatial ICA analysis. NeuroImage 25: 2005, pp 527– 538.

[42] Marijke Welvaert, Yves Rosseel. On the definition of signal-to-noise ratio and contrast-to- noise ratio for fMRI data. PLOS ONE (www.plosone.org) Novembre 2013 Vol 8 Issue 11, e77089