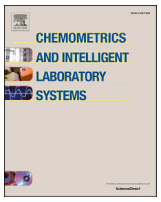
Chemometrics and Intelligent Laboratory Systems 227 (2022) 104606

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



Two-stage iteratively reweighted smoothing splines for baseline correction

Jiajin Wei [a,1], Chen Zhu [b,1], Zhi-Min Zhang [c], Ping He [d,*]

[a] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
[b] Department of Technology Management for Innovation, School of Engineering, The University of Tokyo, Tokyo, Japan
[c] College of Chemistry and Chemical Engineering, Central South University, Changsha, Hunan, China
[d] Division of Science and Technology, BNU-HKBU United International College, Zhuhai, Guangdong, China

ARTICLE INFO

Keywords:
Baseline correction
Chromatograms
Raman spectra
Robust weight
Smoothing spline

ABSTRACT

This paper reviewed several iteratively reweighted baseline correction methods. We note in the literature that the estimated baselines are susceptible to random noises in a low signal-to-noise signal. When the acquired signals are complex-structured, the estimated baselines may still contain the peak information. This paper proposes a new approach named two-stage iteratively reweighted smoothing splines (RWSS) to cope with those situations. The proposed method estimates the baselines by applying weighted smoothing splines in two stages. The first stage applies the smoothing splines with Tukey's Bisquare weights to estimate the baselines, while the second stage is designed to fine-tune the first stage's result. Specifically, the weighted smoothing splines are applied again to remove the remaining peak information, where the weights for the peak regions are inversely proportional to the error variances. By simulation studies, the performance of the two-stage RWSS algorithm is among the best in terms of the root mean square error. Finally, we conducted three real data studies, i.e., chromatograms, infrared spectra, and Raman spectra, to verify the reliability of our new algorithm in practical tasks by evaluating principal components and classification accuracy. The new algorithm is implemented in R language, where the source code is available at https://github.com/rwss2021/rwss.

1. Introduction

Signals of analytical instruments consist of true signal information (peaks), a baseline, and random noises. The baseline could be either a straight line with a certain slope (linear) or a flat curve (non-linear), but it is assumed to be continuous and slowly varying [1]. The detection and estimation of the baseline information from the acquired signals is a prerequisite step in spectroscopic analysis. This is because the existence of the baseline will adversely affect the subsequent quantitative analysis, such as peak area and height estimation.

Previously, a manual approach has been proposed to remove the baseline information from spectra [2]. It involves manually selecting representative points of a baseline, and then the baseline is estimated by linear approximation, polynomial approximation, or spline interpolation based on the selected points. However, selecting good representative points is not always easy because the accuracy largely depends on personal experience. Also, since the baselines are generally varied among different spectra, the process would be time-consuming if one has to deal with a large number of spectra. In view of this, many automatic baseline correction approaches have been proposed. This paper aims to contribute a new approach to estimate baselines automatically and compare different automatic baseline correction methods with signals of different analytical instruments. Before proceeding further, we give a brief review of the leading automatic correction approaches.

Many approaches have been proposed for automatically removing baselines from spectra. These approaches are mainly based on wavelet transform, robust local regression, spline fitting, and penalized least squares [3–7]. Wavelet transform assumes that the baseline information should be well-separable from the signal in the frequency domain [4,8,9]. It estimates the baseline by separating the low-frequency baseline part from the high-frequency true signal part. Ruckstuhl et al. (2001) proposed a robust baseline correction technique by the local regression with Tukey's Bisquare weights, where one needs to specify an appropriate bandwidth to yield a satisfactory baseline estimation [10]. Besides the local regression, some spline fitting approaches have also been applied to estimate the baseline, which require specifying the number of knots being used [6,11].

In addition, there are other approaches based on the Whittaker smoother or the penalized least squares. Eilers and Boelens (2005)

* Corresponding author.
E-mail address: heping@uic.edu.cn (P. He).
[1] Co-first authors.

https://doi.org/10.1016/j.chemolab.2022.104606
Received 21 September 2021; Received in revised form 25 March 2022; Accepted 9 June 2022
Available online 15 June 2022

proposed an effective baseline estimator by combining asymmetric weights with the Whittaker smoother, which is referred to as the asymmetric least squares smoothing (AsLS) [12]. The authors assigned a very small weight $0.001 \leq p \leq 0.1$ to the wave regions with positive residuals, while the other regions with negative residuals were given a much larger weight $1 - p$. In order to improve the performance of AsLS, He et al. (2014) suggested an improved asymmetric least squares (IAsLS) approach for measuring the fit to the first derivative of the signals [13]. Moreover, Zhang, Chen, and Liang (2010) proposed an adaptive algorithm to determine the weights of penalized least squares in each iteration (airPLS), where the weight for the wave regions with positive residuals was zero, and an adaptive weight was computed for the other regions [5].

More recently, a penalized least squares method with multiple new constrains was proposed under the condition of symmetric characteristic peaks [14]. These approaches required no prior information and are easy to implement. However, it may be challenging to tune the key parameters to achieve the optimal baseline estimation. More importantly, numerical results indicate that the estimation results may be suboptimal when the structure of an acquired signal is complex, i.e., the signal with multiple and wide peaks or when the signal-to-noise ratio (SNR) is low. Besides the iteratively reweighted methods, there are other novel methods based on modifying the residuals iteratively. For example, Xu et al. (2021) proposed the iterative smoothing splines with root error adjustment (ISREA) [7]. Not limited to the methods associated with the penalized least squares, Li et al. (2020) developed a sparse Bayesian learning (SBL) model that assumes the second derivative of the baseline is normally distributed with a zero mean and a variance determined by a prior distribution [15].

In this paper, from another perspective, we consider the weighted smoothing splines with the robust weights and provide some theoretical derivation for the weights and spline fitting. Inspired by Ruckstuhl et al. (2001), who have applied Tukey's Bisquare weights in local regression [10], we note that combining Tukey's Bisquare weights with the weighted smoothing splines may yield more accurate baseline estimates in most low SNR circumstances. However, the estimated baseline may still contain peak information when the original signal greatly fluctuates. Thus, a new baseline estimation method is developed in this paper to cope with the problems, which is referred to as the two-stage iteratively reweighted smoothing splines (RWSS). The proposed method can be divided into two stages. In the first stage, we apply Tukey's Bisquare weights to the iteratively weighted smoothing splines. In the second stage, to further eliminate the remaining peak information, we apply the weighted smoothing splines iteratively again to the resulting signals from the first stage, where the weights for the peak regions are inversely proportional to the error variances in each iteration.

The remainder of this paper is organized as follows. Statistical theory and algorithm description for the two-stage RWSS are demonstrated in Section 'Methodology.' In Sections 'Simulation studies' and 'Simulation results,' we conduct simulation studies to compare the proposed algorithm to three existing methods including, for example, airPLS, ISREA, and SBL. In Section 'Real data studies', we apply the new algorithm to three real datasets, i.e., chromatograms, infrared spectra, and Raman spectra, and evaluate the performance on several classification and regression tasks. Finally, we conclude this paper in Section 'Conclusion' and provide additional technical or simulation results in the Appendix.

2. Methodology

2.1. Weighted smoothing splines

In this paper, the baseline of a signal is captured by the smoothing splines in two stages. The first stage applies the smoothing splines with robust weights to yield the general baseline information from an acquired signal. We note that there is a high possibility that the baseline information derived from the first stage is susceptible to the peak information when the structure of a raw signal is complex. The second stage solves this problem by using the weighted smoothing splines again, where the weights are computed in a different way to the first stage (see an illustration in Appendix A.1).

Compared to other spline methods, the smoothing splines less concern the knot selection problem by using the maximal set of knots [16]. The smoothing splines can be derived by minimizing the residual sum of squares with a penalty term:

$$\text{Minimize } \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(z)\}^2 dz, \tag{1}$$

where x_i represents the spectral wavenumber i or the x-coordinate for a certain analytical signal, y_i represents the intensity of the analytical signal at x_i, with $i = 1, \ldots, n$, and λ is a smoothing parameter. The first term of (1) measures the goodness of fit, while the second term penalizes the curvature of the derived spline. The smoothing parameter λ balances the tradeoff between the goodness of fit and the curvature of the resulting spline. The unique solution of formula (1) is a natural cubic spline with knots at the unique values of x_i [16].

To estimate the baseline from a given signal, in this paper we consider the weighted smoothing splines, where the objective function is given as

$$\text{Minimize } \sum_{i=1}^{n} w_i\{y_i - f(x_i)\}^2 + \lambda \int \{f''(z)\}^2 dz. \tag{2}$$

The advantage of the weighted smoothing splines allows us to be more focused on the wavenumbers that should be included in the estimation procedure. The wavenumbers, where the observed values of the signal are smaller than those of the estimated baseline, are defined as *baseline regions*. On the other side, the wavenumbers outside the baseline regions are defined as *peak regions*, which are typically regarded as outliers by (2). Hence, large weights are assigned to the wavenumbers that belong to the baseline regions while wavenumbers in the peak regions are given by small weights.

Some baseline estimation methods simply assign a zero weight to the wavenumbers inside the peak regions. However, this may lead to information loss in the baseline regions as well, especially when SNR is very low, because it may be difficult to distinguish the noises and the peaks correctly. To solve this problem, we apply Tukey's Bisquare weight [17]:

$$w(x_i) = \begin{cases} 1 & r_i < 0, \\ \{\max[1 - (r_i/k)^2, 0]\}^2 & \text{otherwise}, \end{cases} \tag{3}$$

where $r_i = \{y_i - \hat{f}(x_i)\}/\sigma$ represent the scaled residuals, σ is the scale parameter, and $\hat{f}(x_i)$ are the estimates of $f(x_i)$. The hyperparameter k controls the robustness of the estimation procedure. If the value of k is large, most of the wavenumbers will be preserved, and only those with high intensity will be ignored. By contrast, if the value of k is small, most of the wavenumbers will be ignored, and only those with low intensity will be preserved.

Note that, in Tukey's Bisquare weight, two parameters k and σ should be determined in advance. Cleveland (1981) [3] chose the value of the robustness parameter $k = 4.05$, and Ruckstuhl et al. (2001) [10] applied a slightly smaller value of $k = 3$. In simulation studies, we note that an extremely small k may yield a very flat estimated baseline and may thus deteriorate the results of the weighted smoothing splines. In contrast, with $k \to \infty$, the Tukey's Bisquare weight (3) tends to be one for all the wavenumbers, and so model (2) is nearly identical to the unweighted smoothing splines. In Section 'Simulation results', we will illustrate in detail the effect of k on the baseline estimation. Next, to calculate the scaled residuals r_i, we apply the median absolute values (MAV) of the residuals to estimate σ, where

$$\hat{\sigma}_{\text{MAV}} = \frac{\text{median}\{|y_i - \hat{f}(x_i)|\}}{0.6745}. \tag{4}$$

In the literature, estimator (4) has been applied to compute Tukey's Bisquare weight in a local regression model for baseline correction [10].

In this paper, the simulation and real data studies will also show that estimator (4), together with the hyperparameter k, can assist our new algorithm in estimating the baseline robustly. For more details about MAV and estimator (4), one may refer to Rousseeuw and Croux (1993) [18] and Leys et al. (2013) [19], and the references therein.

2.2. Iteratively reweighted smoothing splines algorithm — Stage 1

The proposed algorithm estimates the baseline by applying weighted smoothing splines iteratively. In each iteration, it involves solving a weighted smoothing spline in the following form:

$$\text{Minimize } \sum_{i=1}^{n} w_{1i}^{t}\{y_i - f^{t}(x_i)\}^2 + \lambda_1 \int \{f^{t''}(z)\}^2 dz, \tag{5}$$

where w_{1i}^{t} is the weight assigned to the ith wavenumber at the tth iteration, for $i = 1, \ldots, n$. The setting of initial weights for all the wavenumbers is $\mathbf{w}_1^0 = \mathbf{1}$, where $\mathbf{w}_1^0 = \{w_{11}^0, \ldots w_{1n}^0\} = \{1, \ldots, 1\}$. After that, the weights in next iterations ($t > 0$) are computed by Tukey's Bisquare weights:

$$w_{1i}^{t} = \begin{cases} 1 & r_i^{t-1} < 0, \\ \left[\max\{1 - (r_i^{t-1}/k)^2, 0\}\right]^2 & \text{otherwise}, \end{cases} \tag{6}$$

where $r_i^{t-1} = \{y_i - \hat{f}^{t-1}(x_i)\}/\sigma$, and σ is estimated by $\hat{\sigma}_{\text{MAV}} = \text{median}\{|y_i - \hat{f}^{t-1}(x_i)|\}/0.6745$. Note that the estimate $\hat{f}^{t}$ is a candidate baseline at the tth iteration. Unit weights will be assigned to the wavenumbers inside the baseline regions. The weights for the non-baseline regions will be derived based on Tukey's Bisquare weights. In this way, only the wavenumbers with large digressions will be considered as peaks, and the wavenumbers with fewer digressions will still be involved in the next iteration with the shrinkage weights.

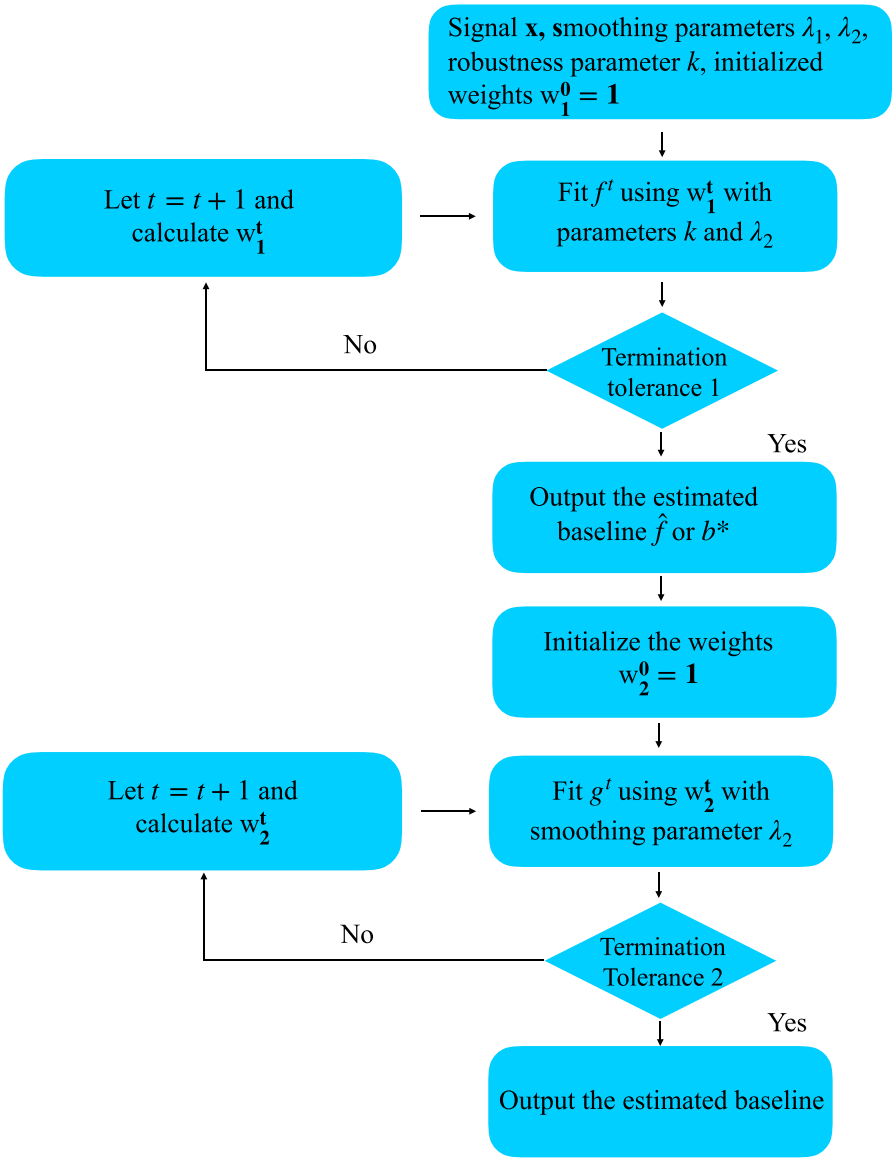


Fig. 1. Flowchart for the framework of the two-stage RWSS algorithm.

The iteration process will stop when the termination tolerance is met. The termination tolerance is given as

$$\left|\frac{\sum_{j\in\mathscr{R}_b}|r_j^{t-1}|}{\sum_{j\in\mathscr{R}_b}|r_j^{t}|}-1\right|<\text{tolerance},$$

where $\mathscr{R}_b=\{r|r<0\}$ indicates the scaled residuals measured in the baseline regions. The algorithm will converge when the estimated baselines are almost unchanged.

2.3. Iteratively reweighted smoothing splines algorithm — Stage 2

By simulation studies, we note that the first stage estimation is good enough for baseline correction in most cases. Hence, we assume most peak information has been removed in the first stage. In practice, however, when the overlapped and/or sharp peaks exist, some peak information may still remain in the estimated baseline. To solve this problem, we assume the association between the estimated baseline from the first stage and the true baseline is

$$b^*=b+\epsilon, \tag{7}$$

where b^* is the estimated baseline from the first stage, b represents the true baseline, and ϵ is the error composed of random noises and/or remaining peak information. Based on this, we assume that ϵ is a random variable from the normal distribution with non-constant variances. The assumption of non-constant variation accounts for the remaining peak information embedded in the error term.

Consequently, we replace the true baseline b by a functional curve g in the following process, and then it yields that

$$b_i^*=g(x_i)+\epsilon_i, \tag{8}$$

where $g(x_i)$ expresses the true value of the baseline at wavenumber i, ϵ_i are independent and normally distributed with zero mean and non-constant variances $\sigma^2_{x_i}$, or equivalently, $\epsilon_i\sim^{\text{ind}}N(0,\sigma^2_{x_i})$, $i=1,\ldots,n$. In fact, we figure out that deriving the maximum likelihood estimator of g can be regarded as solving a weighted smoothing spline, where the weights are inversely proportional to error variances $1/\sigma^2_{x_i}$. One may find the proof in Appendix A.2. With the above results, we apply the weighted smoothing splines with the following objective function:

$$\text{Minimize }\sum_{i=1}^{n}w_{2i}^{t}\{b_i^*-g^t(x_i)\}^2+\lambda_2\int\{g^{t''}(z)\}^2\,dz. \tag{9}$$

However, we note that the algorithm may be divergent when most weight estimates for $1/\sigma^2_{x_i}$ tend to zero during the iterations. To solve this problem, we proposed an adjusted weight for (9) such that w_{2i}^t for $t>0$ are computed by

$$w_{2i}^t=\begin{cases}1-\dfrac{\left(\sigma_{x_i}^{t-1}\right)^{-2}}{\sum_{i=1}^{n}\left(\sigma_{x_i}^{t-1}\right)^{-2}} & b_i^*-\hat{g}^{t-1}(x_i)<0,\\ \dfrac{\left(\sigma_{x_i}^{t-1}\right)^{-2}}{\sum_{i=1}^{n}\left(\sigma_{x_i}^{t-1}\right)^{-2}} & \text{otherwise.}\end{cases} \tag{10}$$

For $t=0$, the setting of initial weights for all wavenumbers is 1. When $b_i^*-\hat{g}^{t-1}(x_i)<0$, we assume the wavenumber with a larger residual contains more baseline information, so a larger weight is assigned. Moreover, we terminate the iteration when the number of wavenumbers

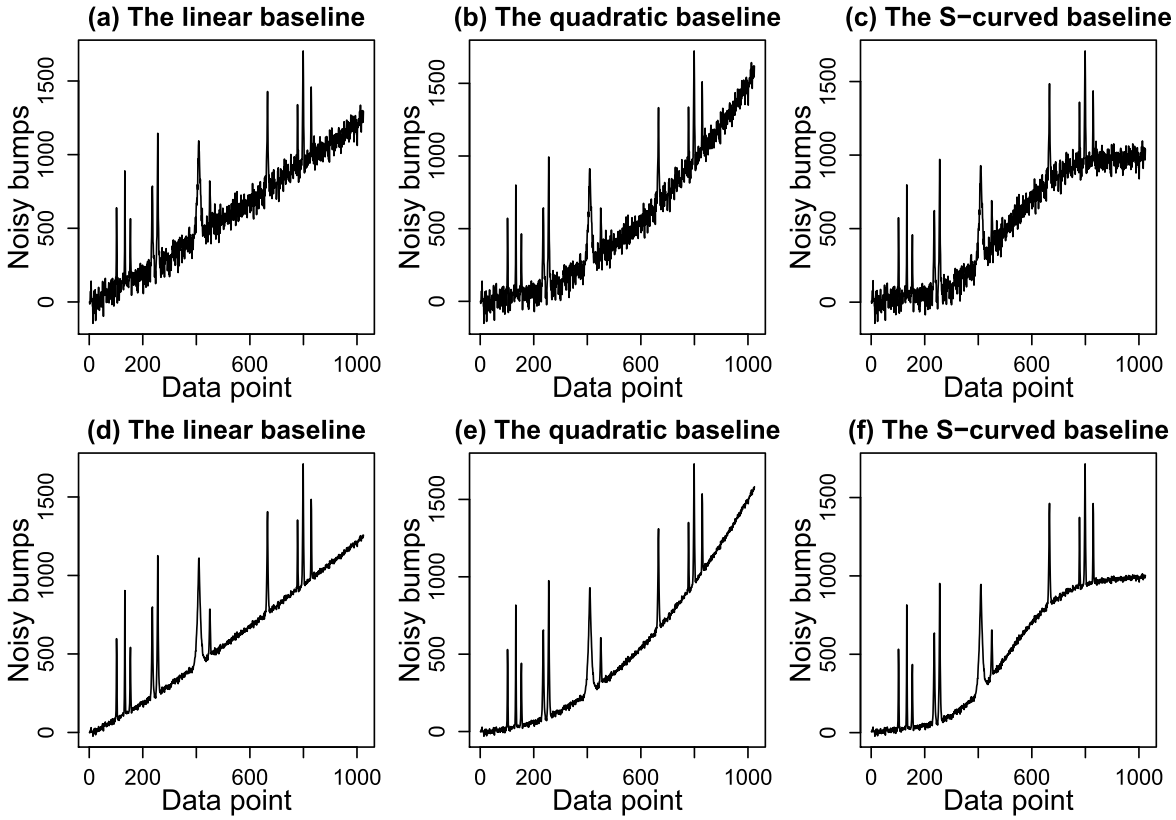


Fig. 2. Uncorrected signals with the linear, quadratic and S-shape curved baselines, where the top three panels (a) to (c) represent SNR = 2, and the bottom three panels (d) to (f) represent SNR = 10.

with $b_i^* - \hat{g}^{t-1}(x_i) < 0$ is less than 1/10 of the total wavenumbers. This approach avoids the divergence of the algorithm when too many weights of the wavenumbers tend to zero.

To estimate the unknown non-constant variances $\sigma^2_{x_i}$, it has been suggested using the log squared residual method in the literature. Recall that model (8) is given as $b_i^* = g(x_i) + \epsilon_i$, where $\epsilon_i \sim^{\text{ind}} N(0, \sigma^2_{x_i})$. Following the method by Wasserman (2006), we let $\sigma(x_i) = \sigma_{x_i}$ be the non-constant standard deviations and let $\epsilon_i = \sigma(x_i)e_i$, where e_i are independent and identically distributed random variables from $N(0, 1)$ [20]. Accordingly, model (8) can be rewritten as

$$b_i^* = g(x_i) + \sigma(x_i)e_i, \tag{11}$$

$$\log[\{b_i^* - g(x_i)\}^2] = \log\{\sigma^2(x_i)\} + \log(e_i^2), \tag{12}$$

where $\log\{\sigma^2(x_i)\}$ can be estimated by the smoothing splines, and $\log(\cdot)$ represents the natural logarithm. Accordingly, the procedure for estimating the non-constant variances $\sigma^2_{x_i}$ at the tth iteration is given as.

(i) Compute the log squared residuals $r_i^{*t} = \log[\{b_i^* - \hat{g}^t(x_i)\}^2]$;
(ii) Use the smoothing splines to regress r_i^{*t} on x_i and obtain the estimates $\hat{r}_i^{*t}(x_i)$;
(iii) Estimate the non-constant variances by $(\hat{\sigma}^t_{x_i})^2 = \exp\{\hat{r}_i^{*t}(x_i)\}$, where $\exp(\cdot)$ represent the exponential function.

The algorithm for stage 2 will stop when the following termination is met:

$$\left| \frac{\sum_j |R_j^t|}{\sum_j |R_j^{t-1}|} - 1 \right| < \text{tolerance}, \tag{13}$$

where $R_j^t = b_j^* - \hat{g}^t(x_j)$ in the tth iteration. For illustration, we show in Fig. 1 the framework of the proposed two-stage algorithm.

3. Simulation studies

3.1. Simulated data

Simulated data consist of true signal information, baselines, and random noises. The true signal is generated by the Donoho and Johnstone functions, which can be obtained by the DJ.EX function from the R package *WaveThresh* [21]. Specifically, three types of artificial baselines, including linear, quadratic, and S-shape curved, are added to the true signal, respectively. In addition, the random noises are generated with SNR = 2 or 10. For illustration, we present the simulated noisy data with the three baselines in Fig. 2.

3.2. Simulated complex-structured data

To further evaluate the proposed method, we generate more complex signals by simulation, where the peaks are more wide-ranging. The top two panels of Fig. 3 illustrate the pure peak information. In particular, we provide a signal with nine peaks in the top-right panel including two close to the boundaries. We also add the quadratic baseline to the peaks, as well as random noises with SNR = 2 or 10, and then present the noisy signals in the middle and bottom panels of Fig. 3.

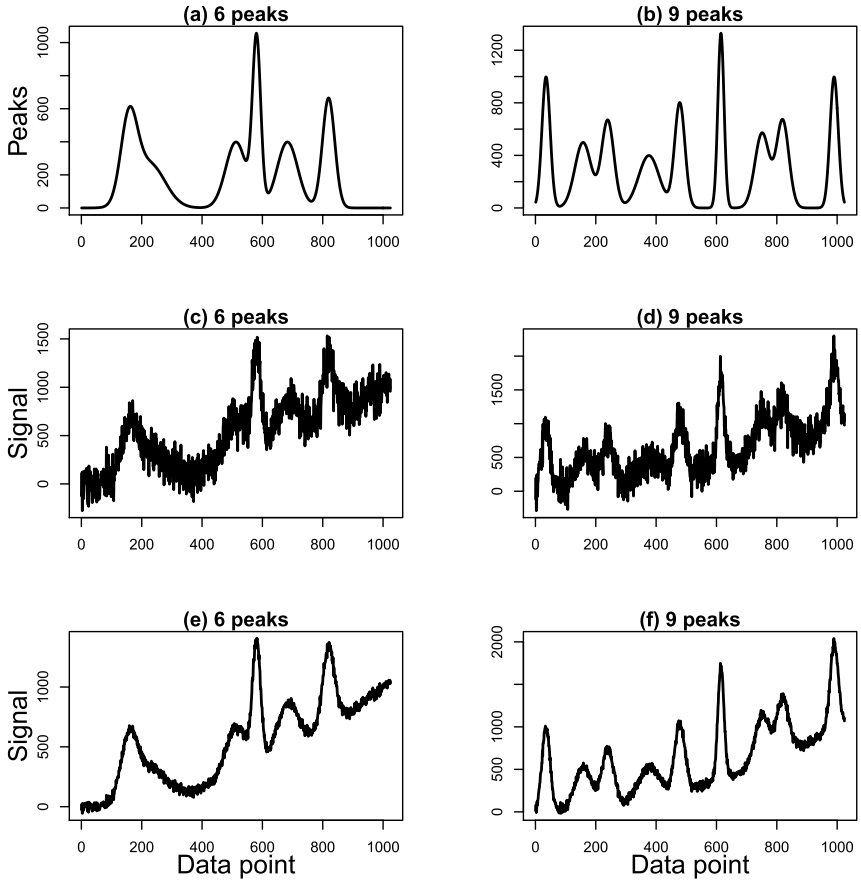


Fig. 3. Simulated complex-structured signals with six or nine peaks. The top two panels (a) and (b) represent the pure peaks, the middle two panels (c) and (d) represent SNR = 2, and the bottom two panels (e) and (f) represent SNR = 10.

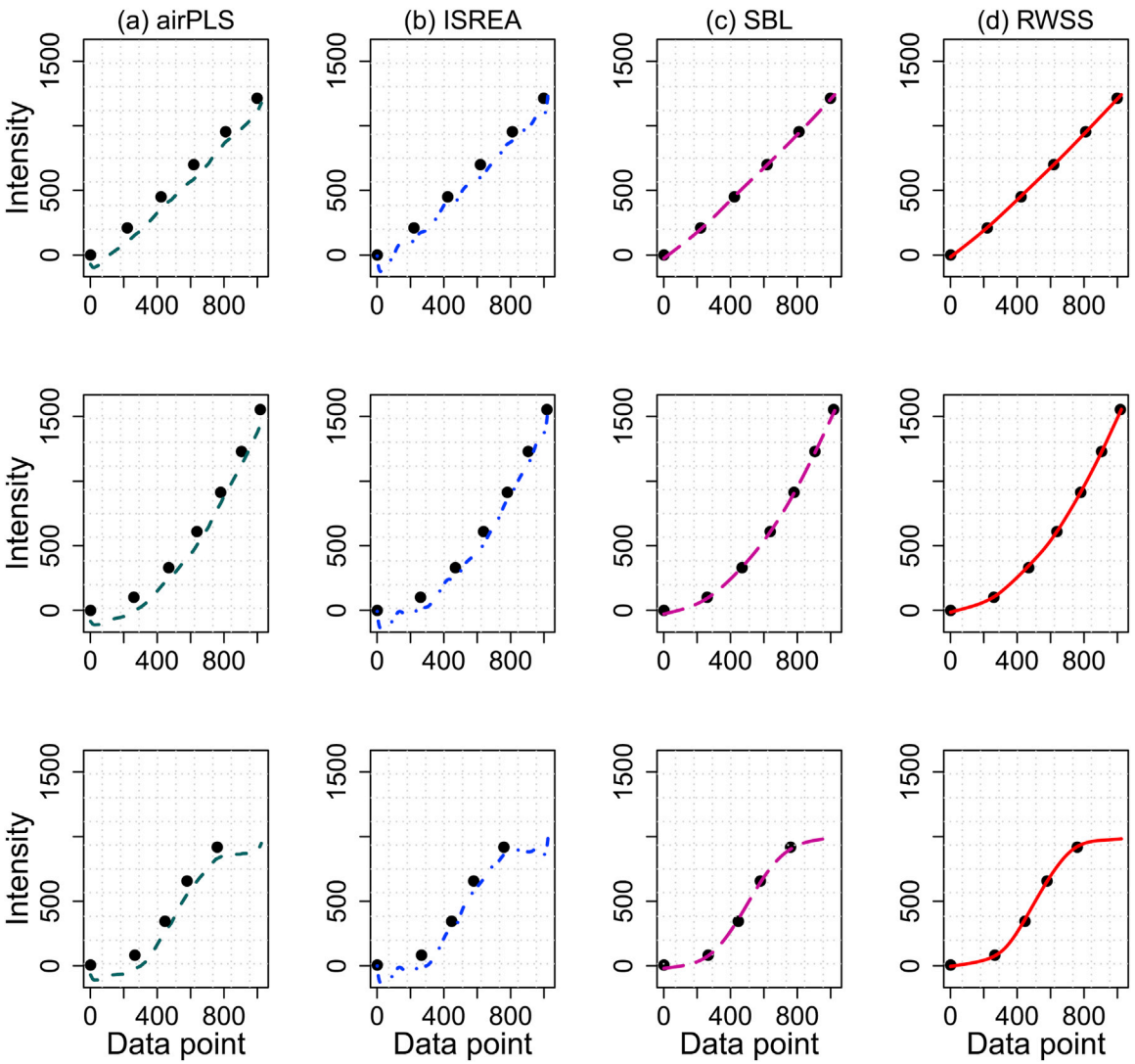


Fig. 4. Comparison of the four methods using simulated data with SNR = 2. The dotted lines represent the true baselines, column (a) with the short dashed lines represents airPLS, column (b) with the dot-dashed lines represents ISREA, column (c) with the long dashed lines represents SBL, and column (d) with the solid lines represents the two-stage RWSS algorithm.

4. Simulation results

4.1. Tuning parameters

The proposed two-stage RWSS algorithm involves three parameters, including the smoothing parameters λ_1 and λ_2 for the two stages, and the robustness parameter k. In practice, if λ_1 and λ_2 are large, the estimated baseline will be smooth and flat; in contrast, if λ_1 and λ_2 are very small, the estimated baseline will be flexible and easily susceptible to the peak information and random noises.

In the literature, grid searching is a commonly used strategy to select multiple parameters [15,22]. For the parameter settings, we consider k from $\{4, 6, 8, 10, 20, 40, 80, 160, 320\}$. Then, for each k we select λ_1 and λ_2 from $\{10^{-5}, 5\times10^{-5}, 10^{-4}, 5\times10^{-4}, \ldots, 10^{-1}, 5\times10^{-1}\}$, independently. In next sections, we will show that the optimal λ_1 and λ_2 under different settings of k are able to yield reliable performance, which suggests that the two-stage RWSS algorithm is robust for different values of k.

4.2. Comparison to the existing methods using simulated data

In this section, we conduct comparative studies to evaluate the performance of the two-stage RWSS algorithm. Three existing methods are also considered for comparison, including the adaptive iteratively reweighted penalized least squares (airPLS), the iterative smoothing splines with root error adjustment (ISREA), and the sparse Bayesian learning model (SBL). For the setting of airPLS, we select the optimal smoothing parameter from $\{10, 20, \ldots, 10000\}$, which yields the lowest root mean square error (RMSE):

$$\mathrm{RMSE}(\hat{b}) = \sqrt{\frac{\sum_{i=1}^{n}(\hat{b}_i - b_i)^2}{n}}, \tag{14}$$

where $\hat{b}$ is a generic form of the estimated baseline, $\hat{b}_i$ is the baseline estimate at the ith data point, and b_i is the true baseline value at the ith data point. For ISREA, we choose the optimal number of knots from 5 to 25. Moreover, the initial parameters of SBL remain the same as Li et al. (2020) [15]. Finally, we report the simulation results in Fig. 4.

From Fig. 4, the two-stage RWSS algorithm and SBL are comparable and both outperform the other two methods when SNR = 2. In contrast, the estimated baselines by airPLS and ISREA are slightly lower than the true baseline. When SNR = 10, Fig. 5 shows that all the four methods can yield satisfactory estimates for the true baseline in most settings. For a clearer presentation, the RMSEs of the four methods are also reported in Table 1. The two-stage RWSS algorithm provides the most accurate estimates. ISREA and SBL both outperform airPLS.

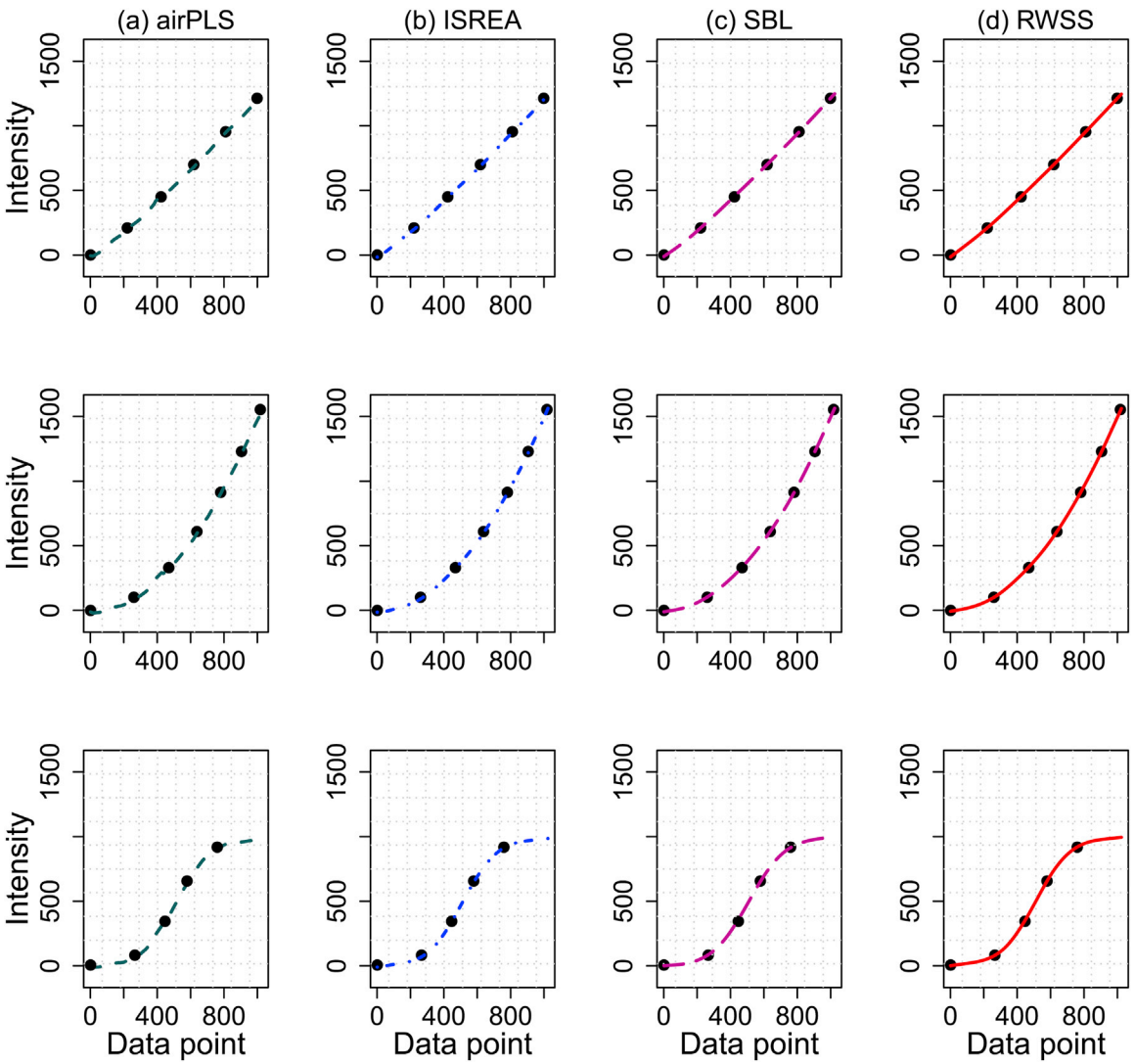


Fig. 5. Comparison of the four methods using simulated data with SNR = 10. The dotted lines represent the true baselines, column (a) with the short dashed lines represents airPLS, column (b) with the dot-dashed lines represents ISREA, column (c) with the long dashed lines represents SBL, and column (d) with the solid lines represents the two-stage RWSS algorithm.

Table 1
The RMSEs of the four methods by using simulated data.

Baseline type	SNR	airPLS	ISREA	SBL	RWSS
linear	2	101.23	82.67	8.37	**4.42**
quadratic	2	99.94	82.68	9.48	**7.76**
S-curved	2	102.86	82.68	14.20	**7.08**
linear	10	14.98	9.17	2.63	**1.96**
quadratic	10	15.84	9.16	2.93	**2.66**
S-curved	10	15.32	9.16	6.79	**2.50**

To further evaluate the robustness of the two-stage RWSS algorithm, we provide some additional results with the minimal RMSEs for individual values of k. For the sake of brevity, we postpone the numerical results to Appendix B. From Table A1 in Appendix B, it is evident that, even if we do not select the optimal k, the performance of the two-stage RWSS algorithm is still better than, or at least comparable to, the other methods in most settings.

4.3. Comparison to the existing methods using complex-structured data

In this section, the simulated complex-structured data are used to compare the performance of the four methods, including airPLS, ISREA, SBL, and the two-stage RWSS algorithm. The parameters of the four methods are selected by using the same strategy as before.

As illustrated by Fig. 6, the estimated baselines by airPLS and ISREA are fluctuating when SNR = 2. In fact, the estimated baselines can be smoother by increasing the smoothing parameter of airPLS and decreasing the number of knots of ISREA. By simulation, however, the smoother estimated baselines will be located much lower than the true baseline, which will increase the RMSE. We also note that SBL is able to yield smoother and more accurate estimated baselines than airPLS and ISREA, but the estimated baselines still contain the peak information. It is evident that the two-stage RWSS is among the best methods. When SNR = 10, Fig. 7 shows that all the four methods perform much better than the low SNR case. For a better comparison, we summarize the numerical results in Table 2, from which we note that the two-stage RWSS provides the most accurate estimates.

In addition, we also compute the RMSE of the two-stage RWSS for each value of k with the results given in Table A2 in Appendix B. The results with varying k are stable if we can select the optimal λ_1 and λ_2. In most settings, the two-stage RWSS performs better than, or almost as well as, the other three methods.

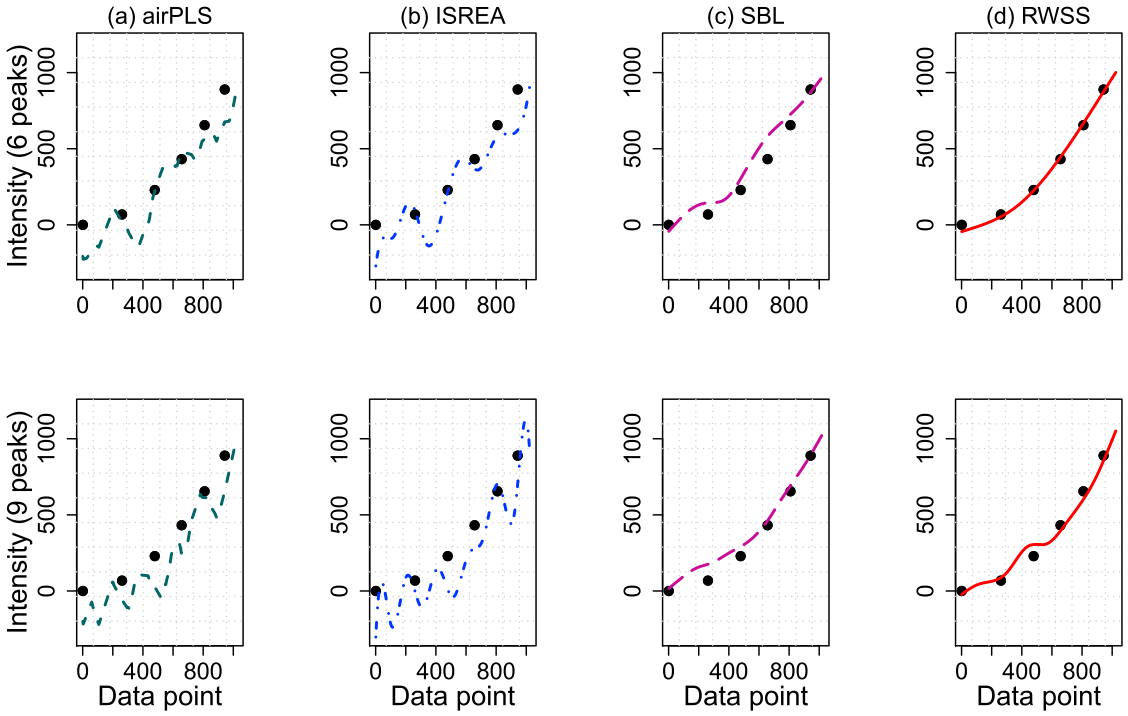


Fig. 6. Comparison of the four methods using complex-structured data with SNR = 2. The dotted lines represent the true baselines, column (a) with the short dashed lines represents airPLS, column (b) with the dot-dashed lines represents ISREA, column (c) with the long dashed lines represents SBL, and column (d) with the solid lines represents the two-stage RWSS algorithm.

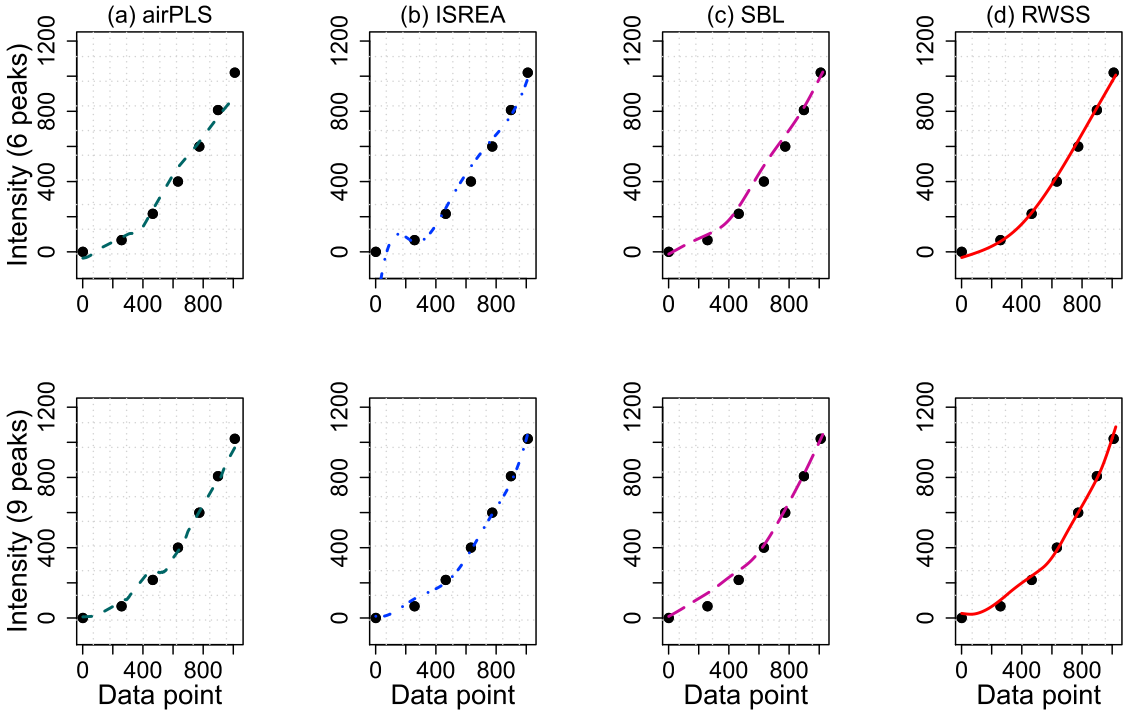


Fig. 7. Comparison of the four methods using complex-structured data with SNR = 10. The dotted lines represent the true baselines, column (a) with the short dashed lines represents airPLS, column (b) with the dot-dashed lines represents ISREA, column (c) with the long dashed lines represents SBL, and column (d) with the solid lines represents the two-stage RWSS algorithm.

4.4. Determining parameters by the generalized cross-validation

Besides the grid searching, we also provide a variation of the original two-stage RWSS algorithm by which we can simplify the selection of the two smoothing parameters λ_1 and λ_2 to some extent. The generalized cross-validation (GCV) is an efficient tool to determine the optimal smoothing parameter λ of the smoothing splines that can yield an accurate curve estimate, minimizing the squared-error loss [16]. However, in

Table 2
The RMSEs of the four methods by using complex-structured data.

Number of peaks	SNR	airPLS	ISREA	SBL	RWSS
6	2	145.72	142.17	83.24	**17.91**
9	2	157.80	168.82	67.38	**46.32**
6	10	49.22	62.00	47.66	**17.64**
9	10	29.81	23.48	45.06	**22.28**

Table 3
The results of PLSR on the uncorrected spectra and the corrected spectra by airPLS and the two-stage RWSS.

	RMSECV(PCs)	
Method	Moisture	Sugar
uncorrected	0.56(9)	2.03(8)
airPLS	0.53(9)	**1.65(6)**
two-stage RWSS	**0.48**(8)	1.89(8)

the context of baseline estimation, we need to opt for the smoothing parameters that can (i) automatically detect the peak regions (e.g., assigning small weights to the wavenumbers within the peak regions) and (ii) well fit the baseline regions. To satisfy these two requirements, a modified version RWSS-GCV is proposed with two initial values λ_1^0 and λ_2^0. The selection for the two initial values is expected to be much easier in this case than the grid searching since they will be further iteratively updated. Specifically, we compute the GCVs for the two stages as follows:

$$\mathrm{GCV}(\lambda_1^t) = \frac{1}{n}\frac{\sum_{i=1}^{n}\{f^{t-1}(x_i) - f^t(x_i)\}^2}{\{1 - n^{-1}\mathrm{trace}(S_{\lambda_1^t})\}^2}, \tag{15}$$

$$\mathrm{GCV}(\lambda_2^t) = \frac{1}{n}\frac{\sum_{i=1}^{n}\{g^{t-1}(x_i) - g^t(x_i)\}^2}{\{1 - n^{-1}\mathrm{trace}(S_{\lambda_2^t})\}^2}, \tag{16}$$

where $f^t(x_i)$ and $g^t(x_i)$ are the baseline estimates in the tth iteration of the first and second stages, respectively; $S_{\lambda_1^t}$ and $S_{\lambda_2^t}$ are smoother matrices for the two stages; and the function trace(·) represents the trace of a matrix. Throughout this paper, we set the initial values as $\lambda_1^0 = 10^{-5}$ and $\lambda_2^0 = 10^{-5}$. Moreover, thanks to the robustness of the two-stage RWSS with respect to k, we let $k = 4$ for RWSS-GCV as it yields accurate estimates in most settings of our simulation studies.

The simulation results are provided in Table A3 (Appendix B), where it can be shown that RWSS-GCV performs reasonably well on the simulated data without complex structure, despite the fact that it gives results that are somewhat less accurate than the original RWSS. The R codes for RWSS-GCV are available online at https://github.com/rwss2021/rwss/blob/main/R/RWSS-GCV.R. Note that RWSS-GCV is designed to increase productivity in real-world situations when the user is confronted with a large number of analytical signals. However, if the number of analytical signals is not too large, we still recommend using the two-stage RWSS algorithm with the grid searching, which can achieve the most reliable estimates. We expect that future work will be needed to improve RWSS-GCV, especially for the complex-structured data.

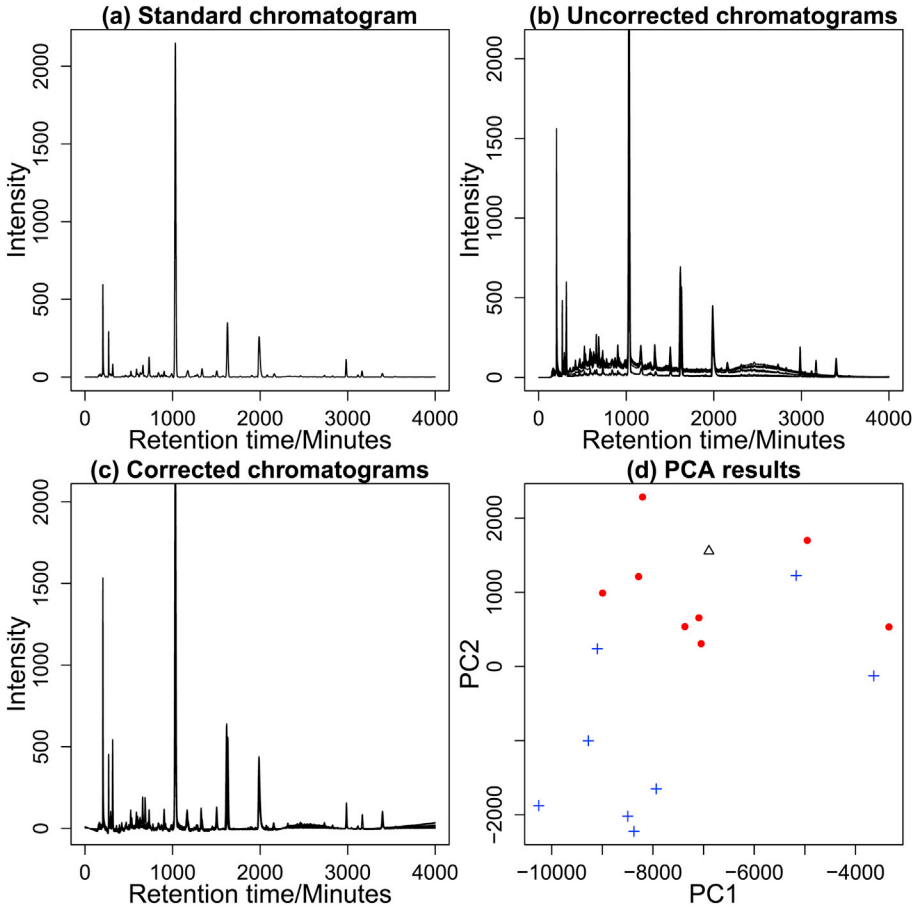


Fig. 8. Panels (a), (b) and (c) show the standard, uncorrected and corrected chromatograms, respectively. Panel (d) shows the first two principal components of the standard, uncorrected and corrected chromatograms, where the triangle represents the standard chromatogram, the plus signs represent the uncorrected chromatograms, and the solid points represent the corrected chromatograms.

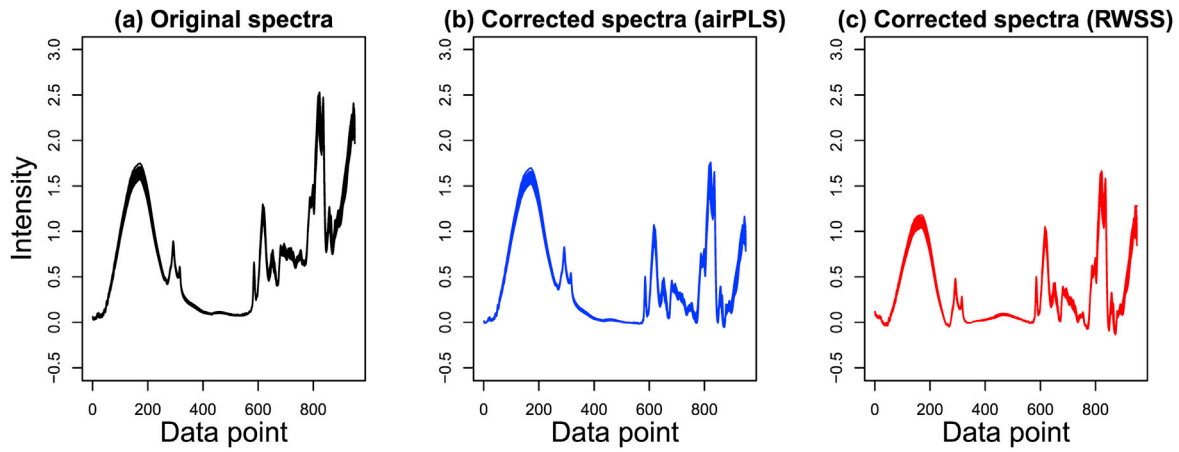


Fig. 9. The original spectra of marzipan and the corrected spectra by airPLS and the two-stage RWSS.

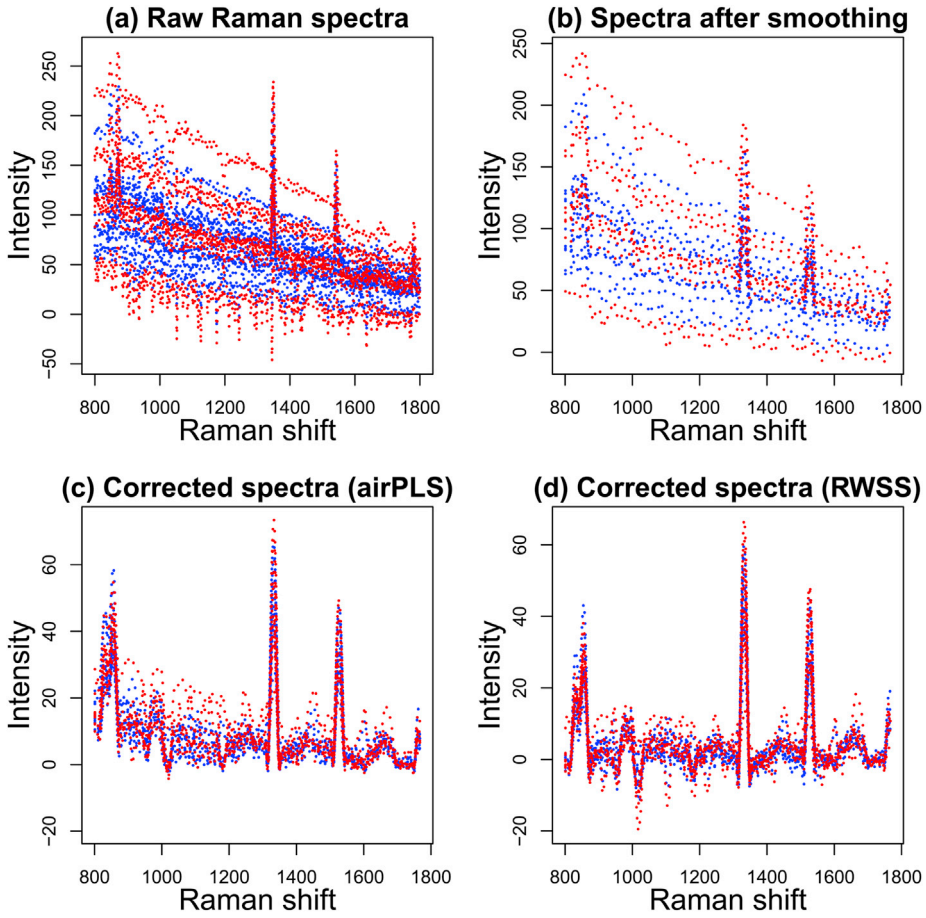


Fig. 10. The top two panels (a) and (b) illustrate the raw Raman spectra from the thumbnail location and the spectra by Savitzky-Golay smoothing. The bottom two panels (c) and (d) illustrate the corrected spectra by airPLS and the two-stage RWSS, respectively.

Table 4
The results of SVM with 2–5 principal components of the uncorrected spectra and the corrected spectra.

Method/Accuracy	2 PCs	3 PCs	4 PCs	5 PCs
uncorrected	0.35	0.35	0.25	0.30
airPLS	0.55	0.50	0.50	0.60
ISREA	0.55	0.50	0.55	0.55
SBL	0.40	0.30	0.30	0.30
two-stage RWSS	0.70	0.70	0.60	0.65

5. Real data studies

5.1. Chromatograms

For baseline correction, Zhang, Chen, and Liang (2010) [5] applied airPLS to 8 HPLC chromatograms of Red Peony Root. By comparing the first two principal components of the uncorrected and corrected chromatograms to those of a standard chromatogram, the author found that the chromatograms corrected by airPLS became closer to the standard chromatogram. Inspired by this, we revisit the Red Peony Root

Chromatograms to evaluate our two-stage RWSS algorithm. For more details about the Red Peony Root data, one may refer to Zhang, Chen, and Liang (2010) [5] and the references therein. From the top two panels and the bottom-left panel of Fig. 8, it is evident that the corrected chromatograms are more compact compared to the uncorrected ones. Moreover, the bottom-right panel of Fig. 8 shows that the first two principal components (PCs) of the corrected chromatograms move closer to the standard chromatogram, which indicates the reliability of our algorithm.

5.2. Infrared spectra

In this section, we assess the usefulness of the two-stage RWSS algorithm by considering a data set with 32 infrared spectra of marzipan ranging from 6500 to 650 cm^{-1}, which is available at http://www.models.life.ku.dk/marzipan [22,23]. For illustration, the original uncorrected spectra are shown in the left panel of Fig. 9. By applying airPLS with $\lambda_{airPLS} = 200$, the corrected spectra are shown in the middle panel of Fig. 9. Then, the right panel of Fig. 9 exhibits the corrected spectra by the two-stage RWSS algorithm with $k = 320$, $\lambda_1 = 0.00001$, and $\lambda_2 = 0.001$.

The marzipan data set contains two response variables: moisture and sugar. In the literature, Han et al. (2017) used the marzipan data to test 10 baseline correction methods, applied the partial least squares regression (PLSR) to the uncorrected and corrected spectra, and then compared the baseline correction methods by RMSEs of cross-validation (RMSECVs) [22]. In the light of this, we first remove the baselines of the marzipan spectra using airPLS and the two-stage RWSS algorithm, respectively. After that, we compare the performance of the two methods based on RMSECV. The leave-one-out cross-validation (CV) is applied to select the optimal number of PCs. Finally, we report the RMSECVs associated with the two response variables in Table 3. It is evident that both two methods can improve the RMSECVs of the PLSR model. The two-stage RWSS method, in particular, works well for the response variable moisture. In addition, by comparing the results from Table 3 to those from Han et al. (2017), we note that the two-stage RWSS algorithm is among the best baseline correction methods.

5.3. Raman spectra

5.3.1. Preliminary studies

Raman spectroscopy has been considered as a non-invasive and efficient tool for identifying people with diabetes [24]. The authors conducted a machine learning study on several Raman spectra from different anatomical locations, e.g., the ear lobe, inner arm, thumbnail, and cubital vein, to classify the subjects with type 2 diabetes mellitus (DM2) and the healthy subjects. In this section, we revisit the data from the thumbnail location composed of 11 diabetic patients and 9 healthy subjects, where the region 800–1800 cm^{-1} of each spectrum is included as suggested by the authors. Since the spectra are noisy, we reduce the noise by Savitzky-Golay smoothing, which has been used for Raman spectra pre-processing in the literature [25]. The raw Raman spectra and the spectra after smoothing are shown in the top two panels of Fig. 10.

Now to estimate the baselines, we apply airPLS and the two-stage RWSS, respectively, to the Raman spectra after smoothing. For airPLS, we let $\lambda_{airPLS} = 270$; and meanwhile, we let $k = 320$, $\lambda_1 = 0.0001$ and $\lambda_2 = 0.001$ for the two-stage RWSS algorithm. The corrected spectra are shown in the bottom two panels of Fig. 10. Next, we conduct the principal component analysis (PCA) on the uncorrected and corrected data, respectively, and then apply the support vector machine (SVM) for classification. 10 principal components are used to yields a cumulative proportion of explained variances greater than 90%. By the 10-fold cross-validation, the accuracies of the uncorrected data, the corrected data by airPLS, and the corrected data by the two-stage RWSS are 0.3, 0.35 and 0.4, respectively. Although, airPLS and the two-stage RWSS can slightly improve the accuracy, the manual selection of the parameters may not yield the optimal classification results.

5.3.2. Selecting parameters by cross-validation

In this section, a strategy of selecting parameters automatically by cross-validation (CV) is designed, which will be applicable and effective in a supervised or semi-supervised scenario. We again apply the principal component analysis (PCA) and the support vector machine (SVM) to classify the DM2 subjects and the healthy subjects. For comparison, we include airPLS, ISREA, SBL and the two-stage RWSS in this study. Instead of selecting the parameters manually, we apply the 10-fold CV for each method to find the optimal parameters that achieve the highest accuracy of classification. Specifically, the parameters are selected by following the same strategy as described in the simulation studies. In Table 4, we show the results of SVM yielded by 2–5 principal components of the uncorrected and corrected spectra. It is evident that the corrected spectra by airPLS, ISREA and the two-stage RWSS provide higher accuracies than the uncorrected spectra. In contrast, SBL does not perform well. It is also noteworthy that the two-stage RWSS outperforms airPLS and ISREA in most settings, revealing our new algorithm has the potential to improve the upper limit of classification performance.

6. Conclusion

This paper proposed a two-stage algorithm for baseline correction and verified its performance on various simulated signals and real data. The simulation results showed the reliability of the proposed algorithm. In particular, the two-stage estimation can extract the baseline information accurately no matter whether the SNR of the analytical signal is low or high. Especially in the low SNR cases, the new algorithm performs much better than the existing iteratively reweighted baseline correction methods. For further evaluation, we conducted three real data studies, including, for example, chromatograms, infrared spectra, and Raman spectra, which also show the advancement of our new algorithm. Note that the three data sets are available at https://github.com/zmzhang/airPLS, http://www.models.life.ku.dk/marzipan, and https://www.kaggle.com/codina/raman-spectroscopy-of-diabetes [5,23,24]. With the above evidence and the comparative results, the two-stage RWSS algorithm can serve as a reliable and promising tool for baseline correction. We note that there exists papers in the literature that studied the baseline correction or the noise assessment for a specific analytic signal [7,25,26]. In the future, we expect that our two-stage RWSS algorithm will shed light on new directions on more accurate baseline estimation of a signal from a given instrument.

Author statement

Jiajin Wei: Methodology, Software, Formal analysis, Writing - Original Draft.

Chen Zhu: Methodology, Software, Formal analysis, Writing - Original Draft.

Zhi-Min Zhang: Conceptualization, Writing - Review & Editing.

Ping He: Conceptualization, Writing - Review & Editing.

Funding

Ping He's research was supported by the Internal Research Grant (R202010) of BNU-HKBU United International College, and the National Natural Science Foundation of China (62076029).

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

The authors sincerely thank the Editor and two anonymous reviewers for their constructive comments on this paper.

Appendix A

Appendix A.1

For illustration, we show the estimated baselines from the first and second stages in Fig. A1. It is evident from the figure that the second stage can further remove the remaining peak information.

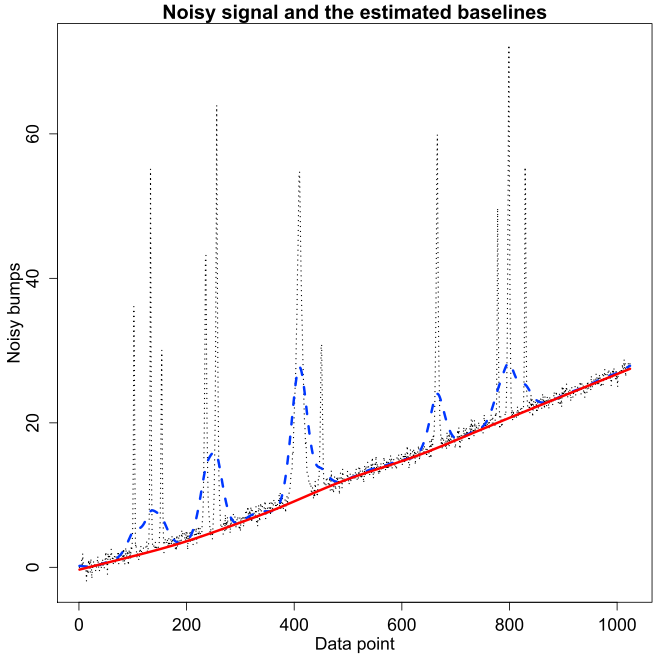


Fig. A.1. The uncorrected signal and the estimated baselines from the first and second stages of RWSS, where the dotted line represents the uncorrected signal, the dashed line represents the estimated baseline from the first stage, and the solid line represents the estimated baseline from the second stage.

Appendix A.2

From the first stage estimation, the algorithm yields an estimate b^* for the true baseline. However, when the structure of the acquired signal is complex, the estimated baseline b^* may still contain the peak information. In view of this, we propose to derive a latent relationship between b^* and the true baseline b, where

$$b^* = b + \epsilon. \tag{A1}$$

The error term ϵ is composed of random noises and some remaining signal information. Since the signal information inside of the random noises may not be too much, we assume the random noises are still from the normal distribution but with non-constant variances.

For ease of notation, let b_i^* be the baseline estimate at wavenumber i from the first stage and $g(x_i)$ be the true value of the baseline at wavenumber i, then we have

$$b_i^* = g(x_i) + \epsilon_i, \tag{A2}$$

where ϵ_i are independent and normally distributed with zero mean and non-constant variances. Namely, $\epsilon_i \sim^{\text{ind}} N(0, \sigma_{x_i}^2)$, $i = 1, \ldots, n$. By definition, the log-likelihood function for g is given as

$$\begin{aligned} l(g; \sigma_{x_i}^2) &= \log\left\{ (2\pi)^{-\frac{n}{2}} \prod_{i=1}^{n} (\sigma_{x_i}^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{[b_i^* - g(x_i)]^2}{\sigma_{x_i}^2} \right\} \right\} \\ &= -\frac{n}{2} \log 2\pi - \frac{n}{2} \sum_{i=1}^{n} \log \sigma_{x_i}^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{[b_i^* - g(x_i)]^2}{\sigma_{x_i}^2}, \end{aligned} \tag{A3}$$

where $\log(\cdot)$ represents the natural logarithm. Meanwhile, to control the roughness of the estimation, we add a penalty term to the log-likelihood function such that

$$
\begin{aligned}
l(\theta;\lambda) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\sum\nolimits_{i=1}^{n}\log\sigma_{x_i}^2 \\
&\quad -\frac{1}{2}\sum\nolimits_{i=1}^{n}\frac{[b_i^* - g(x_i)]^2}{\sigma_{x_i}^2} - \frac{1}{2}\lambda\int\{g''(z)\}^2 dz \\
&= -\frac{n}{2}\log 2\pi - \frac{n}{2}\sum\nolimits_{i=1}^{n}\log\sigma_{x_i}^2 \\
&\quad -\frac{1}{2}\sum\nolimits_{i=1}^{n}\frac{[b_i^* - N_i^T\theta]^2}{\sigma_{x_i}^2} - \frac{1}{2}\lambda\int\left\{\sum\nolimits_{j=1}^{n}N_j''(z)\theta_j\right\}^2 dz \\
&= -\frac{n}{2}\log 2\pi - \frac{n}{2}\sum\nolimits_{i=1}^{n}\log\sigma_{x_i}^2 \\
&\quad -\frac{1}{2}\sum\nolimits_{i=1}^{n}\frac{[b_i^* - N_i^T\theta]^2}{\sigma_{x_i}^2} - \frac{1}{2}\lambda\theta^T\Omega_N\theta,
\end{aligned}
\tag{A4}
$$

where N is an $i \times j$ matrix with $\{N\}_{ij} = N_j(x_i)$, $N_j(x)$ is the n-dimensional set of basis functions for natural splines, θ_j is the coefficient of the corresponding basis function $N_j(x)$, and $\{\Omega_N\}_{jk} = \int N_j''(z)N_k''(z)\,dz$ [16].

By taking partial derivative, we have

$$
\begin{aligned}
\frac{\partial l(\theta;\lambda)}{\partial\theta} &= \sum\nolimits_{i=1}^{n}\frac{N_i(b_i^* - N_i^T\theta)}{\sigma_{x_i}^2} - \lambda\Omega_N\theta \\
&= N^T W(b^* - N\theta) - \lambda\Omega_N\theta,
\end{aligned}
\tag{A5}
$$

$$
\begin{aligned}
\frac{\partial^2 l(\theta;\lambda)}{\partial\theta\partial\theta^T} &= \sum\nolimits_{i=1}^{n}\frac{N_i(b_i^* - N_i^T\theta)}{\sigma_{x_i}^2} - \lambda\Omega_N\theta \\
&= -N^T WN - \lambda\Omega_N,
\end{aligned}
\tag{A6}
$$

where W is a diagonal matrix with the diagonal element $w_i = 1/\sigma_{x_i}^2$. By applying the Newton-Raphson method, we have

$$
\begin{aligned}
\hat{\theta}_{new} &= \hat{\theta}_{old} - \left(\frac{\partial^2 l(\theta;\lambda)}{\partial\theta\partial\theta^T}\right)^{-1}\frac{\partial l(\theta;\lambda)}{\partial\theta} \\
&= \hat{\theta}_{old} + (N^T WN + \lambda\Omega_N)^{-1}(N^T W(b^* - N\hat{\theta}_{old}) - \lambda\Omega_N\hat{\theta}_{old}) \\
&= (N^T WN + \lambda\Omega_N)^{-1}N^T Wb^*,
\end{aligned}
\tag{A7}
$$

$$
\hat{g}_{new} = N\hat{\theta}_{new} = N(N^T WN + \lambda\Omega_N)^{-1}N^T Wb^*. \tag{A8}
$$

In fact, the above procedures are equivalent to solve a weighted smoothing spline [16] such that

$$
\min_g RSS(g,\lambda) = \sum\nolimits_{i=1}^{n} w_i\{b_i^* - g(x_i)\}^2 + \lambda\int\{g''(z)\}^2 dz, \tag{A9}
$$

where $w_i \geq 0$ are the weights. Note that formula (A9) can be rewritten as

$$
\min_g RSS(g,\lambda) = (b^* - N\theta)^T W(b^* - N\theta) + \lambda\theta^T\Omega_N\theta, \tag{A10}
$$

where W is a diagonal matrix with diagonal elements w_i. By taking partial derivative, we have

$$
\begin{aligned}
\frac{\partial}{\partial\theta}RSS(\theta,\lambda) &= \frac{\partial}{\partial\theta}(b^* - N\theta)^T W(b^* - N\theta) + \lambda\theta^T\Omega_N\theta \\
&= -2N^T W(b^* - N\theta) + 2\lambda\Omega_N\theta = 0,
\end{aligned}
\tag{A11}
$$

and solve that

$$
\hat{\theta} = (N^T WN + \lambda\Omega_N)^{-1}N^T Wb^*, \tag{A12}
$$

$$
\hat{g} = N\hat{\theta} = N(N^T WN + \lambda\Omega_N)^{-1}N^T Wb^*. \tag{A13}
$$

We note that these results are equivalent to formulas (A7) and (A8). That is, finding the maximum likelihood estimate for g is to solve the weighted smoothing splines (A9), where the weights are inversely proportional to error variances.

Appendix B. additional simulation results

Table A.1
The RMSEs of the two-stage RWSS algorithm for individual *k* values by using simulated data

k	RMSE					
	SNR = 2			SNR = 10		
	linear	*quadratic*	*S-curved*	*linear*	*quadratic*	*S-curved*
4	4.42	7.76	8.42	1.96	2.66	2.50
6	5.70	10.14	8.42	2.16	2.74	2.77
8	6.15	10.26	8.14	2.18	3.46	2.93
10	6.14	10.35	7.83	2.28	3.88	3.17
20	5.75	12.47	7.36	2.48	4.17	4.25
40	5.74	12.48	7.34	3.74	5.00	4.65
80	5.74	12.48	7.34	3.95	5.24	4.30
160	5.74	12.48	7.15	4.01	5.31	4.39
320	5.74	12.48	7.08	4.04	5.32	4.41

Table A.2
The RMSEs of the two-stage RWSS algorithm for individual *k* values by using simulated complex-structured data

k	RMSE			
	SNR = 2		SNR = 10	
	6 peaks	9 peaks	6 peaks	9 peaks
4	17.91	46.69	17.64	24.71
6	27.55	47.17	24.20	22.28
8	27.47	47.41	24.17	22.78
10	27.38	47.52	24.55	24.07
20	27.22	46.32	19.18	25.81
40	27.21	46.35	19.21	25.98
80	27.20	46.34	19.21	26.02
160	27.20	46.34	19.23	26.03
320	27.19	46.34	19.24	26.03

Table A.3
The RMSEs of RWSS-GCV by using simulated data and simulated complex-structured data

Simulated data	Baseline type		
SNR	*linear*	*quadratic*	*S-curved*
2	13.22	23.59	9.37
10	10.09	11.26	11.47
Simulated complex-structured data	Number of peaks		
SNR	6 peaks	9 peaks	–
2	225.47	178.25	–
10	66.56	86.87	–

References

[1] M. Mecozzi, A polynomial curve fitting method for baseline drift correction in the chromatographic analysis of hydrocarbons in environmental samples, APCBEE Procedia 10 (2014) 2–6, https://doi.org/10.1016/j.apcbee.2014.10.003.

[2] A. Jirasek, G. Schulze, M.M.L. Yu, M.W. Blades, R.F.B. Turner, Accuracy and precision of manual baseline determination, Appl. Spectrosc. 58 (12) (2004) 1488–1499, https://doi.org/10.1366/0003702042641236.

[3] W.S. Cleveland, Lowess: a program for smoothing scatterplots by robust locally weighted regression, Am. Statistician 35 (1) (1981) 54.

[4] C.G. Bertinetto, T. Vuorinen, Automatic baseline recognition for the correction of large sets of spectra using continuous wavelet transform and iterative fitting, Appl. Spectrosc. 68 (2) (2014) 155–164, https://doi.org/10.1366/13-07018.

[5] Z.-M. Zhang, S. Chen, Y.-Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst 135 (5) (2010) 1138–1146, https://doi.org/10.1039/B922045C.

[6] Y. Cai, C. Yang, D. Xu, W. Gui, Baseline correction for Raman spectra using penalized spline smoothing based on vector transformation, Anal. Methods 10 (28) (2018) 3525–3533, https://doi.org/10.1039/C8AY00914G.

[7] Y. Xu, P. Du, R. Senger, J. Robertson, J.L. Pirkle, Isrea: an efficient peak-preserving baseline correction algorithm for Raman spectra, Appl. Spectrosc. 75 (1) (2021) 34–45, https://doi.org/10.1177/0003702820955245.

[8] B. Liu, Y. Sera, N. Matsubara, K. Otsuka, S. Terabe, Signal denoising and baseline correction by discrete wavelet transform for microchip capillary electrophoresis, Electrophoresis 24 (18) (2003) 3260–3265, https://doi.org/10.1002/elps.200305548.

[9] H. Shin, M.P. Sampat, J.M. Koomen, M.K. Markey, Wavelet-based adaptive denoising and baseline correction for maldi tof ms, OMICS A J. Integr. Biol. 14 (3) (2010) 283–295, https://doi.org/10.1089/omi.2009.0119.

[10] A.F. Ruckstuhl, M.P. Jacobson, R.W. Field, J.A. Dodd, Baseline subtraction using robust local regression estimation, J. Quant. Spectrosc. Radiat. Transf. 68 (2) (2001) 179–193, https://doi.org/10.1016/S0022-4073(00)00021-2.

[11] X. Wang, X.-G. Fan, Y.-J. Xu, X.-F. Wang, H. He, Y. Zuo, A baseline correction algorithm for Raman spectroscopy by adaptive knots b-spline, Meas. Sci. Technol. 26 (11) (2015), 115503, https://doi.org/10.1088/0957-0233/26/11/115503.

[12] P.H. Eilers, H.F. Boelens, Baseline correction with asymmetric least squares smoothing, Leiden Univ. Med. Centre Rep. 1 (1) (2005) 5.

[13] S. He, W. Zhang, L. Liu, Y. Huang, J. He, W. Xie, P. Wu, C. Du, Baseline correction for Raman spectra using an improved asymmetric least squares method, Anal. Methods 6 (12) (2014) 4402–4407, https://doi.org/10.1039/C4AY00068D.

[14] G. Yang, J. Dai, X. Liu, M. Chen, X. Wu, Multiple constrained reweighted penalized least squares for spectral baseline correction, Appl. Spectrosc. 74 (12) (2020) 1443–1451, https://doi.org/10.1177/0003702819885002.

[15] H. Li, J. Dai, T. Pan, C. Chang, H.C. So, Sparse bayesian learning approach for baseline correction, Chemometr. Intell. Lab. Syst. 204 (2020), 104088, https://doi.org/10.1016/j.chemolab.2020.104088.

[16] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, first ed., Springer Series in Statistics, New York, 2001.

[17] D.C. Hoaglin, F. Mosteller, J.W. Tukey, Understanding Robust and Exploratory Data Analysis, John Wiley & Sons, New York, 1999.

[18] P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, J. Am. Stat. Assoc. 88 (424) (1993) 1273–1283, https://doi.org/10.2307/2291267.
[19] C. Leys, C. Ley, O. Klein, P. Bernard, L. Licata, Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median, J. Exp. Soc. Psychol. 49 (4) (2013) 764–766, https://doi.org/10.1016/j.jesp.2013.03.013.
[20] L. Wasserman, All of Nonparametric Statistics, Springer Science & Business Media, New York, 2006.
[21] D.L. Donoho, J.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (3) (1994) 425–455, https://doi.org/10.2307/2337118.
[22] Q. Han, Q. Xie, S. Peng, B. Guo, Simultaneous spectrum fitting and baseline correction using sparse representation, Analyst 142 (13) (2017) 2460–2468, https://doi.org/10.1039/C6AN02341J.
[23] J. Christensen, L. Nørgaard, H. Heimdal, J.G. Pedersen, S.B. Engelsen, Rapid spectroscopic analysis of marzipan—comparative instrumentation, J. Near Infrared Spectrosc. 12 (1) (2004) 63–75, https://doi.org/10.1255/jnirs.408.
[24] E. Guevara, J.C. Torres-Galván, M.G. Ramírez-Elías, C. Luevano-Contreras, F.J. González, Use of Raman spectroscopy to screen diabetes mellitus with machine learning tools, Biomed. Opt Express 9 (10) (2018) 4998–5010, https://doi.org/10.1364/BOE.9.004998.
[25] I. Maitra, C.L. Morais, K.M. Lima, K.M. Ashton, D. Bury, R.S. Date, F.L. Martin, Establishing spectrochemical changes in the natural history of oesophageal adenocarcinoma from tissue Raman mapping analysis, Anal. Bioanal. Chem. 412 (17) (2020) 4077–4087, https://doi.org/10.1007/s00216-020-02637-1.
[26] R.J. Combs, Noise assessment for passive ft-ir spectrometer measurements, Electro-Optical Technology for Rem. Chem. Detect. Identification III 3383 (1998) 75–91, https://doi.org/10.1117/12.317638.