Analyst

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2017, **142**, 2460

Simultaneous spectrum fitting and baseline correction using sparse representation

Quanjie Han, [a,b] Qiong Xie,*[a,b] Silong Peng[a,b] and Baokui Guo[a,b]

Received 24th October 2016,
Accepted 5th May 2017

DOI: 10.1039/c6an02341j

rsc.li/analyst

Sparse representation has been applied in many domains, such as signal processing, image processing and machine learning. In this paper, a redundant dictionary with each column composed of a Voigt-like lineshape is constructed to represent the pure spectrum of the sample. With the prior knowledge that the baseline is smooth and sparse representation coefficient for a pure spectrum, a method simultaneously fitting the pure spectrum and baseline is proposed. Since the pure spectrum is nonnegative, the representation coefficients are also made to be nonnegative. Then through alternating optimization, a surrogate function based algorithm is used to obtain the sparse coefficients. Finally, we adopt one simulated data set and two real data sets to evaluate our method. The results of quantitative analysis show that our method successfully estimates the baseline and pure spectrum and is superior compared to other baseline correction and preprocessing methods.

1 Introduction

Infrared spectroscopy is a simple and reliable technique widely used in general chemistry and specifically in organic chemistry. It can be used to detect the structure of measured samples and is particularly useful to measure the concentration of various compounds in different food products.[1] According to the Beer–Lambert Law, absorbance of the infrared spectrum of the sample is proportional to the concentration of the sample with the thickness of the sample being controlled, thus resulting in the foundation for inverse calibration models such as Partial Least Squares (PLS) regression for the prediction of the concentration of analytes of interest. However, because of instrumental measurement effects, a baseline is usually superimposed on the pure spectrum of the investigated sample. In general, the collected spectroscopic data consist of the pure spectrum of the sample, a baseline and associated noise. From the frequency domain perspective, a baseline is a slow varying background belonging to the low frequency domain and noise is in the high frequency domain. Since the baseline will deteriorate the quantitative calibration results of the samples, one of the fundamental problems in the analysis of spectroscopic data is the separation of useful information contained in the peaks of the pure spectrum from unnecessary background and noise.[2]

In the past few years, numerous techniques for baseline estimation have been proposed.[3] For a signal composed of a low-frequency component and a sparse-derivative component, a low pass filter has been designed which has successfully estimated the baseline of the signal.[4] One disadvantage is that the derivative of the signal must be sparse. Since the baseline is a smooth slow varying background, polynomials and splines have been used to fit the baseline. An improved iterative polynomial fitting with automatic threshold (IIPFAT) has been proposed to estimate the baseline,[5] but it does not perform sufficiently in low signal to-noise and signal-to-background ratio signals. Recently, a baseline correction method based on asymmetric least squares (asLS) has been proposed,[6] which has shown effectiveness for both simulated and real data sets. Subsequently, two modified methods with the ability of updating the asymmetric weight automatically have also been designed, which are termed adaptive iteratively reweighed penalized least squares (airPLS)[7] and asymmetrically reweighed penalized least squares (asPLS),[8] respectively. Furthermore, a multiple spectral baseline correction algorithm using asymmetric least squares was also proposed[9] to exploit the common information in the spectra. According to the statistical learning theory, two different probability distributions were firstly imposed on the baseline points and peak points, respectively. Then an EM algorithm was used to separate the baseline points and peak points. Finally the P-splines were used to fit the baseline.[10] The Corner-Cutting (CC) method was derived from the techniques used in computer aided design and provided an efficient baseline calculation through an iterative process.[11] Also, a method based on exponential smoothing (ATEB) was proposed recently to correct the base-

[a]*Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China. E-mail: qiong.xie@ia.ac.cn; Tel: +86 8254 4571*
[b]*University of Chinese Academy of Sciences, China*

line of a high resolution spectral profile.[12] These methods can all estimate the baseline successfully, but they still cannot separate the pure spectrum from noise. Using the peak information, a method known as Statistics-sensitive Nonlinear Iterative Peak-clipping (SNIP) was proposed.[13] Although the peak height was crucial to the calibration result, sometimes the estimated baseline was not accurate. In order to directly evaluate the calibration performance of the baseline correction method, a baseline correction combined partial least squares (BCC-PLS) algorithm was introduced,[14] but it is known only to be effective for a polynomial type baseline.

Theoretically, the final lineshape of the spectrum can be attributed to three factors: Doppler broadening, radiation damping and collision broadening. Doppler broadening gives the Gaussian lineshape, while radiation damping and collision broadening result in the Lorentzian lineshape.[15] Since these effects act only on molecules, the true lineshape called Voigt lineshape is the convolution of these two lineshapes. In order to avoid the cumbersome computation, the linear combination of these two lineshapes has been used in practice. An alternative approach for obtaining the Voigt-like lineshape was suggested[16] and adopted[17] for hyperspectral curve fitting. Instead of formulating the curve fitting method as a nonlinear least squares problem and solved by the Levenberg–Marquardt scheme,[18] we constructed a redundant dictionary with each column consisting of a Voigt-like lineshape to represent the pure spectrum. Considering that the number of peaks is much smaller than the columns of the dictionary, the representation of spectroscopic data under this dictionary is underdetermined. To surmount the uncertainty of the representation coefficient, the sparsest one will be considered. Sparse representation has been extensively used in recent years and a significant amount of research focus has been made on sparse models and their applications.[19] Moreover, the baseline estimation using sparsity has been proposed elsewhere.[20]

In the former methods, the baseline estimation and pure spectrum fitting were conducted separately. In this paper, with smooth constraint of the baseline and the sparse representation coefficient of the pure spectrum under the constructed dictionary, we can estimate the baseline and pure spectrum simultaneously. The outline of this paper is organized as follows. In section two, we propose our method and an algorithm to solve it. Experiments on one simulated data set and two real data sets are conducted to justify the performance of our method in section three, and we draw some conclusions in the final part.

2 Problem formulation

Throughout this paper, we use $\|x\|_0$ and $\|x\|_1$ to denote the number of non-zeros and sum of absolute values for elements in x respectively, $\|x\|_2$ being the Euclidean norm. Since the baseline is discrete in manner, in order to describe the smoothness of the baseline, a difference matrix is employed. Assume that the length of the spectrum is N, then the k-th order difference matrix will be of size $N \times (N - k)$. The second order difference matrix is displayed as follows:

$$\begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & \cdots & \cdots \\ 0 & 1 & -2 & -1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}. \tag{1}$$

Effectively, each measured spectrum x consists of the pure spectrum s of the sample, a baseline z and noise n. Assuming that they satisfy additive property:

$$x = s + z + n. \tag{2}$$

Since each pure spectrum can be approximated by a linear combination of Gaussian, Lorentzian or Voigt-like lineshapes, we constructed a dictionary Φ with each column composed of a Voigt-like lineshape which can fully represent the spectrum. Analogous to ref. 16, each Voigt-like lineshape can be represented as

$$s(i) = \frac{c}{(1 + \beta^2\psi^2(i))^{\frac{1}{\beta^2}}}, \tag{3}$$

where

$$\psi(i) = \frac{i - i_0}{\sqrt{2}\sigma} \tag{4}$$

$i = 1, 2,\ldots,N$, i_0 is the peak location and σ is the width of the lineshape, c denotes the normalization constant, and β is the tunable parameter for achieving a more Gaussian-like or a Lorentzian-like lineshape. When $\beta = 1$, we obtain a Lorentzian shape:

$$s(i) = \frac{c}{1 + \frac{(i - i_0)^2}{2\sigma^2}}. \tag{5}$$

Taking the logarithm of s and by the Taylor's expansion of $\ln(1 + x)$, we can show that s will approximate a Gaussian shape as β approximates 0:

$$s(i) = ce^{-\frac{(i-i_0)^2}{2\sigma^2}}. \tag{6}$$

Matching pursuit[21] was introduced to decompose a signal as a linear combination of members in a specified family of basis. For the infrared spectroscopic data, the family of basis is the Gaussian, Lorentzian or Voigt-like lineshapes. In order to obtain different redundant dictionaries, different sampling methods can be employed for the peak location i_0 and the width of the lineshape σ. Analogous to ref. 21, assume that each spectrum is of length N, let $\sigma = a^j$, $\Delta u = a^{-1}$ and $i_0 = pa^j\Delta u$, where $0 < j < \log_2 N$ and $0 \le p < N \cdot 2^{-j+1}$ to generate a redundant dictionary. For the real data sets, we set $\beta = 0.5$ to obtain an equally mixed Lorentzian–Gaussian curve to represent the pure spectrum.

Since each pure spectrum can be represented by the dictionary Φ, there is a representation coefficient α such that:

$$s = \Phi\alpha. \tag{7}$$

Moreover, the baseline is smooth and the pure spectrum is positive, so we can use the formula

$$\arg\min_{z,\alpha} \|x - \Phi\alpha - z\|_2^2 + \lambda_1 \|Dz\|_2^2, \quad \alpha \geq 0 \tag{8}$$

to measure the fidelity of the fitting of the measured spectrum and the roughness of the baseline, where D denotes the difference matrix whose order is usually set as two or three. Since the dictionary is redundant, the solution to α is not unique. One alternative is to add a constraint $\|\alpha\|_0$ in (8) to find the sparsest one. However, in most cases, the l_0-norm problem is N–P hard, and the convex relax: l_1-norm is adopted. Finally our method is described as follows:

$$\arg\min_{z,\alpha} \|x - \Phi\alpha - z\|_2^2 + \lambda_1 \|Dz\|_2^2 + \lambda_2 \|\alpha\|_1, \quad \alpha \geq 0. \tag{9}$$

As we can see, when λ_2 approaches infinity, the representation coefficient α becomes zero, this is just the Whittaker Smoother[22] used for de-trending. To solve (9), we optimize z and α alternatively. Taking the partial derivative of (9) with respect to z and setting it to zero, we obtain

$$z = (I + \lambda_1 D^T D)^{-1}(x - \Phi\alpha). \tag{10}$$

Since α is regularized by the l_1-norm, several methods can be adopted to solve it, such as ADMM (Alternating Direction Method of Multipliers),[23] IRLS (Iteratively Reweighted Least Squares),[24] surrogate function[25] and so on. Note that the proximity operator of $\lambda\|\alpha\|_1$ is $\frac{1}{2}(\alpha - v)^2 + \lambda\|\alpha\|_1$, whose solution is the soft threshold function: $S_\lambda(v) = \max(v - \lambda, 0) - \max(-v - \lambda, 0)$. In the following, we use the surrogate function method to obtain the solution for α.

$\varphi(x,x_0)$ is called a surrogate function of $f(x)$ if the following holds:

(i) $\varphi(x,x_0) \geq f(x)$ for all x;

(ii) $\varphi(x_0,x_0) = f(x_0)$.

The k-th iteration of z and α is denoted as $z^{(k)}$ and $\alpha^{(k)}$, respectively, and let

$$L(\alpha) = \|x - \Phi\alpha - z^{(k)}\|_2^2 + \lambda_1 \|Dz^{(k)}\|_2^2 + \lambda_2 \|\alpha\|_1. \tag{11}$$

Since the second term on the right hand side in (11) is independent of α, the optimal value of α is the same as $L(\alpha)$ without it. For simplicity, considering $t^{(k)} = x - z^{(k)}$, we use the equation:

$$L(\alpha) = \|\Phi\alpha - t^{(k)}\|_2^2 + \lambda_2 \|\alpha\|_1. \tag{12}$$

For constructing a surrogate function for $L(\alpha)$, we use the equation:

$$\begin{aligned} Q(\alpha, \alpha^{(k-1)}) = {} & \|\Phi\alpha - t^{(k)}\|_2^2 + \lambda_2 \|\alpha\|_1 + c\|\alpha - \alpha^{(k-1)}\|_2^2 \\ & - \|\Phi\alpha - \Phi\alpha^{(k-1)}\|_2^2, \end{aligned} \tag{13}$$

where c is larger than the squared spectral radius of Φ. Opening the various terms in $Q(\alpha,\alpha^{(k-1)})$ and re-organizing them, we can obtain a new formula,

$$\begin{aligned} Q(\alpha, \alpha^{(k-1)}) = {} & c\left\|\alpha - \alpha^{(k-1)} + \frac{\Phi^T\Phi\alpha^{(k-1)} - \Phi^T t^{(k)}}{c}\right\|_2^2 \\ & + \lambda_2 \|\alpha\|_1 + \text{const.} \end{aligned} \tag{14}$$

The constant in the above expression is independent of α and

$$\alpha^{(k)} := \arg\min_{\alpha} Q(\alpha, \alpha^{(k-1)}). \tag{15}$$

Using soft thresholding, we can obtain

$$\alpha^{(k)} = S_{\frac{\lambda_2}{2c}}\left(\alpha^{(k-1)} + \frac{\Phi^T t^{(k)} - \Phi^T\Phi\alpha^{(k-1)}}{c}\right). \tag{16}$$

Since $\alpha^{(k)}$ is positive, it is projected on its positive part. It is known that the l_2-norm and l_1-norm are separable element-wise, and therefore, our method can process multiple spectra simultaneously. With the alternating optimization of z and α, the final estimated z is used as the estimated baseline and $s = \Phi\alpha$ is the estimation of the pure spectrum. In our algorithm, the terminal criterion will be a maximum iteration number achieved or the relative error of the baseline is below some specified threshold; we can thus summarize the algorithm as follows:

Algorithm 1: Simultaneous spectrum fitting and baseline correction using sparse representation (SSFBCSP)

Step 1. Input single spectrum or spectral matrix x, dictionary Φ, regularizers λ_1 and λ_2, order of difference matrix d, and maximum iteration number *Iter*;

Step 2. Initialize $\alpha = 0$, relative error ε and Preprocessing:

2.1 Mat = $I + \lambda_1 D^T D$;

2.2 Cholesky decomposition of Mat: L = chol(Mat);

2.3 $A = \Phi^T\Phi$.

Step 3. Update z and α:

3.1 $z^{(k)} = L^{-1}L^{-T}(x - \Phi\alpha^{(k-1)})$;

3.2 $\alpha^{(k)} = S_{\frac{\lambda_2}{2c}}\left(\alpha^{(k-1)} + \frac{\Phi^T(x - z^{(k)}) - A\alpha^{(k-1)}}{c}\right)$;

3.3 Project $\alpha^{(k)}$ on its positive part.

Step 4. Check stopping criterion:

if $\|z^{(k)} - z^{(k-1)}\|/\|z^{(k-1)}\| < \varepsilon$ or $k > Iter$, then stop;

else $k \leftarrow k + 1$ and go to **Step 3**.

Step 5. Output baseline z, representation coefficient α.

3 Experiments

One simulated and two real spectral data sets were used to evaluate the performance of the proposed method. The programs were written in house in Matlab Version R2014a (The MathWorks, Inc.) and run on a personal computer with a 3.60 GHz Intel Core on a Windows 7 operating system.

3.1 Simultaneous spectrum fitting and baseline correction on simulated data

The simulated data consists of six Gaussian peaks, one type of baseline (sinusoidal baseline or exponential baseline) and one kind of noise (Gaussian noise or a uniform random noise). The mean and the standard deviation of Gaussian noise is taken as zero and 1% of the intensity of the spectrum, respectively, while the uniform noise does not fluctuate greater than 1% of intensity of the spectrum. Since we have known that the pure spectrum consists of Gaussian lineshapes, a dictionary with a column consisting of a Gaussian lineshape has been constructed where $a = 3$. When $d = 2$, $\lambda_1 = 10^6$ and $\lambda_2 = 0.01$, the outcomes of sinusoidal and exponential baselines with Gaussian noise or uniform noise are shown in Fig. 1–4 respectively.

In order to carefully view the performance for the baseline correction and spectrum estimation of our algorithm, the comparison between the estimated baseline and the true baseline and the comparison between the estimated spectrum and the true spectrum for the spectrum with sinusoidal baseline and Gaussian noise are shown in Fig. 5 and 6, respectively. The results for other cases are similar and are omitted here for briefness.

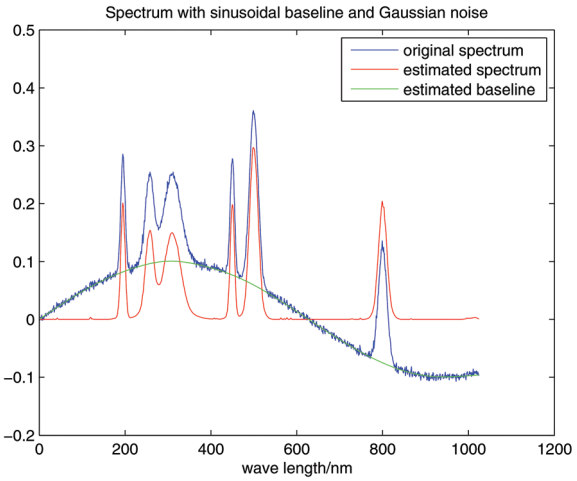

Fig. 1 The estimated baseline and estimated spectrum for the spectrum with sinusoidal baseline and Gaussian noise by SSFBCSP.

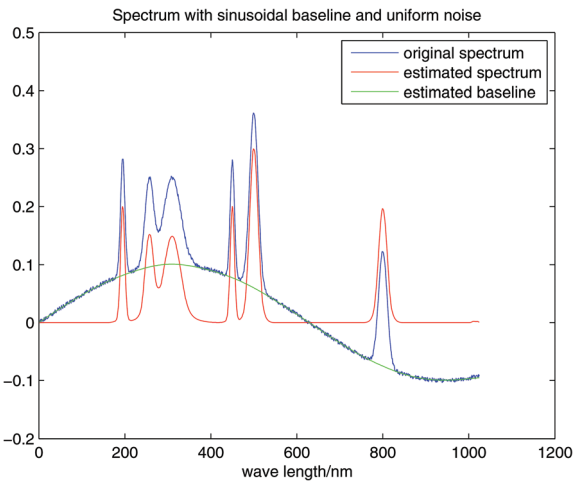

Fig. 2 The estimated baseline and estimated spectrum for the spectrum with sinusoidal baseline and uniform noise by SSFBCSP.

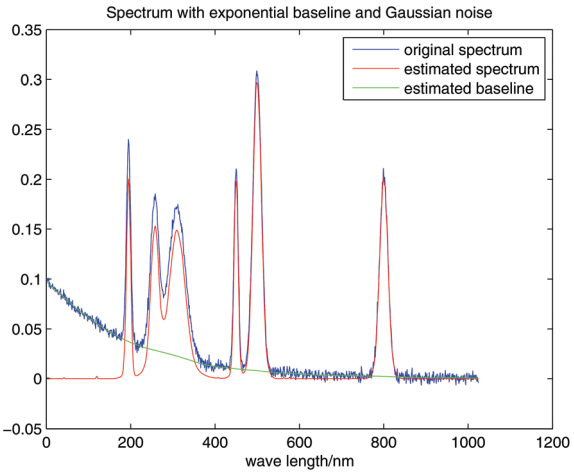

Fig. 3 The estimated baseline and estimated spectrum for the spectrum with exponential baseline and Gaussian noise by SSFBCSP.

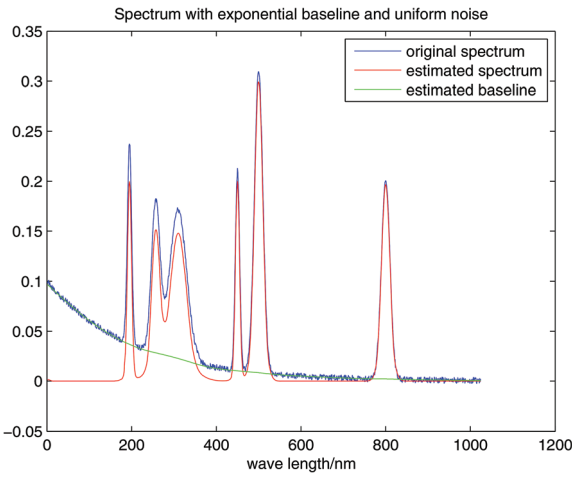

Fig. 4 The estimated baseline and estimated spectrum for the spectrum with exponential baseline and uniform noise by SSFBCSP.

To show the sparsity of the representation coefficient, the case of the sinusoidal baseline with uniform noise is shown in Fig. 7. Considering that the length of the representation coefficient is 1531 and the number of non-zeros is 204, it shows that the representation is sparse. When the parameters $\lambda_1 = 2 \times 10^7$ and $\lambda_2 = 0.1$, the representation coefficient displayed in Fig. 8 is much sparser but almost does not influence the fitting result significantly.

From the figures displayed above, we can see that our algorithm has successfully estimated the baseline and spectrum, and even the representation coefficient as being sparse. For

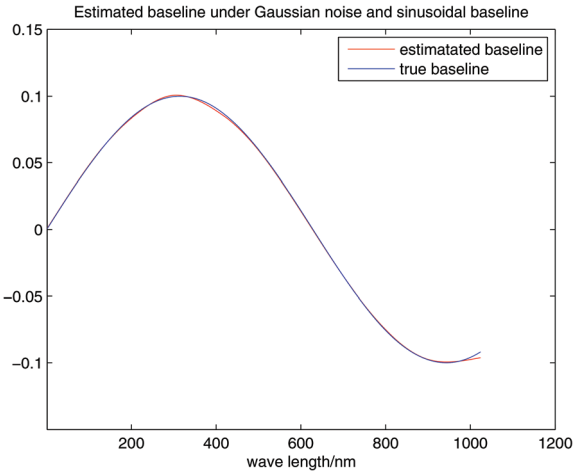

Fig. 5 Comparison of the true baseline and estimated baseline for the spectrum with sinusoidal baseline and Gaussian noise.

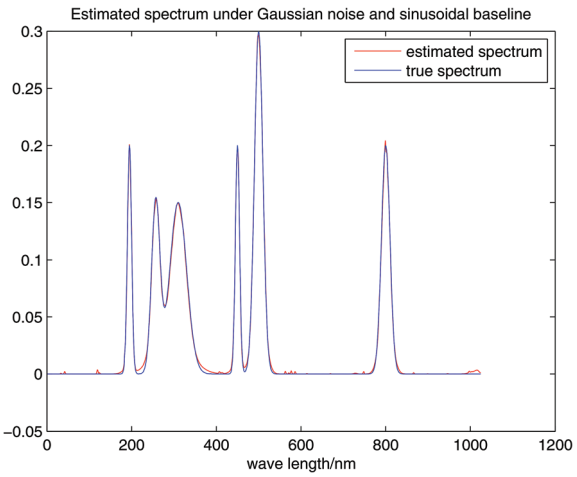

Fig. 6 Comparison of the true spectrum and estimated spectrum for the spectrum with sinusoidal baseline and Gaussian noise.

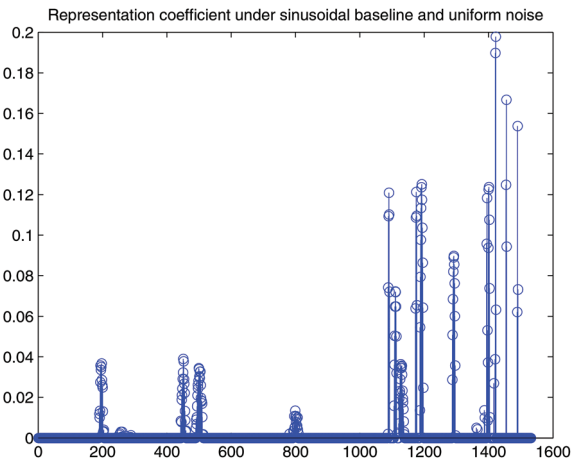

Fig. 7 Representation coefficient of the estimated spectrum for the spectrum with sinusoidal baseline and Gaussian noise. The parameters are $\lambda_1 = 10^6$ and $\lambda_2 = 0.01$.

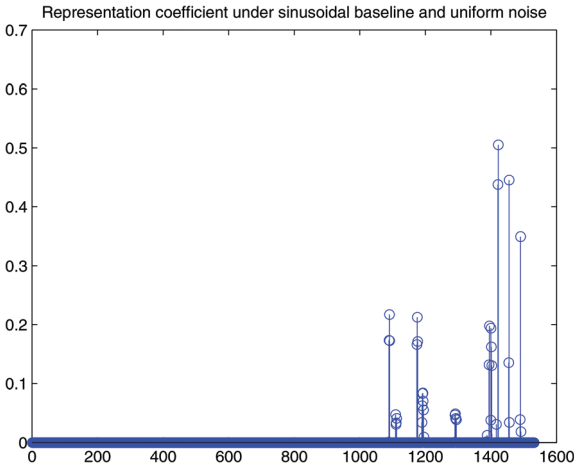

Fig. 8 Representation coefficient of estimated spectrum for the spectrum with sinusoidal baseline and Gaussian noise. The parameters are $\lambda_1 = 2 \times 10^7$ and $\lambda_2 = 0.1$.

the simulated data set, the true baseline and spectrum are known. In order to evaluate our method quantitatively, the root mean square error (RMSE) of the estimated baseline has been adopted. Six other baseline correction methods (asLS, airPLS, SNIP, IIPFAT, CC and ATEB) are compared with our method and the optimal parameters for these methods are selected by grid search. The results are listed in Table 1.

From Table 1, it can be observed that our method outperforms other methods in all cases, while airPLS, ATEB and CC are the second better on GNSB and UNEB, UNSB and GNEB, respectively. Moreover, asLS is just slightly inferior compared to airPLS since the latter method can update the asymmetrical weight adaptively. The other two methods don't show any promising results compared with these methods.

In order to thoroughly compare our method with asLS and airPLS, CC and ATEB, the RMSE of the corrected spectrum is also computed as shown in Table 2. It can be seen again that our method is still better than other methods. The reason may be that our method has considered the noise while other methods need an extra step to denoise.

Finally, for the convergence analysis of our algorithm, we can refer to ref. 25. To get a feeling of the convergence of our

Table 1 RMSE of the estimated baseline

	GNSB[a]	UNSB[a]	GNEB[a]	UNEB[a]
asLS	0.0062	0.0054	0.0051	0.0038
airPLS	0.0052	0.0030	0.0051	0.0023
SNIP	0.0071	0.0043	0.0057	0.0042
IIPFAT	0.0104	0.0089	0.0121	0.0094
CC	0.0094	0.0082	0.0024	0.0064
ATEB	0.0056	0.0029	0.0066	0.0031
SSFBCSP	7.83×10^{-4}	6.29×10^{-4}	6.06×10^{-4}	5.55×10^{-4}

[a] GNSB, UNSB, GNEB, and UNEB denote Gaussian noise sinusoidal baseline, uniform noise sinusoidal baseline, Gaussian noise exponential baseline and uniform noise exponential baseline, respectively.

Table 2 RMSE of the estimated spectrum

	UNSB[a]	GNSB[a]	GNEB[a]	UNEB[a]
asLS	0.0072	0.0057	0.0061	0.0042
airPLS	0.0064	0.0037	0.0060	0.0029
CC	0.0043	0.0090	0.0030	0.0017
ATEB	0.0068	0.0046	0.0062	0.0038
SSFBCSP	0.0018	0.0015	0.0017	0.0015

[a] GNSB, UNSB, GNEB, and UNEB denote Gaussian noise sinusoidal baseline, uniform noise sinusoidal baseline, Gaussian noise exponential baseline and uniform noise exponential baseline, respectively.

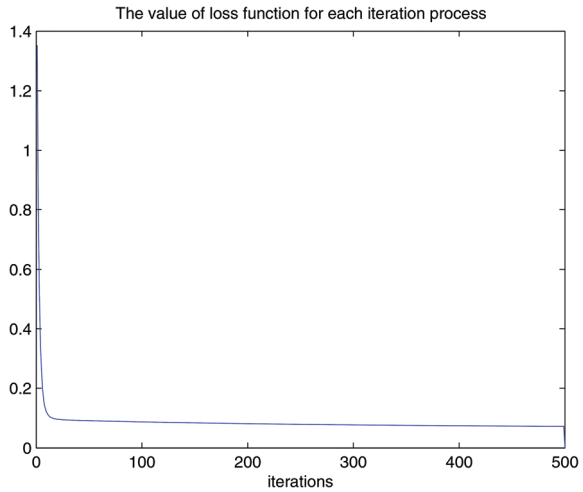

Fig. 9 The value of the loss function L in each iteration.

algorithm, the loss function for the case of the spectrum with sinusoidal baseline and Gaussian noise is shown in Fig. 9. The other cases are similar and are omitted here. We can see that our algorithm can converge in a few iterations.

3.2 Experiments on real data sets

The first real data set is the corn data set which consists of 80 NIR spectra of corn measured on spectrometers mp5 and mp6, and the mp5 data set is used to conduct the experiment. The spectra were recorded in the region of 1100–2498 nm. This data set is available at http://www.eigenvector.com.

The second real data set is the marzipan data set; NIR and IR spectroscopy were applied for a compositional analysis of 32 marzipan samples. Traditional moisture and sugar analysis was performed on all samples. This data set is available at http://www.models.life.ku.dk.

3.2.1 Results of the corn data set. The corn data set is useful for standardization and preprocessing benchmarking. The original spectra of the corn data set, the estimated baselines and the estimated pure spectra by applying our algorithm with $\lambda_1 = 2 \times 10^9$ and $\lambda_2 = 0.02$ are shown in Fig. 10 and 11, respectively. We can see that the original spectra deviate from the zero baseline seriously and our methods have successfully estimated that the baselines and the estimated spectra are on the zero baseline. For the corn data set, the constructed dictionary has 1402 columns. Considering that the length of the representation coefficients is 1402 and the average number of non-zeros for all spectra is about 250, the representation is still sparse. An alternative to obtain a sparser representation is to construct a dictionary which can update the location and width of the peaks for the spectra adaptively. To justify the performance of our algorithm, several standardization and preprocessing methods such as Multiplicative Scatter Correction (MSC),[26] Regularized Multiplicative Scatter Correction (RMSC)[27] and Standard Normal Variate transformation (SNV)[28] have been included. Also, baseline correction methods: Multiple Spectra Baseline Correction (MSBC) algorithm,[9] Corner Cutting (CC)[11] and Two-side Exponential Baseline correction algorithm (ATEB)[12] were used to compare with our method. After the preprocessing process, the corrected spectra were used for quantitative analysis. Just like in ref. 9, each calibration model on corn samples was built on response variables such as moisture, oil, protein and starch, respectively. The corn samples were first sorted by their corresponding

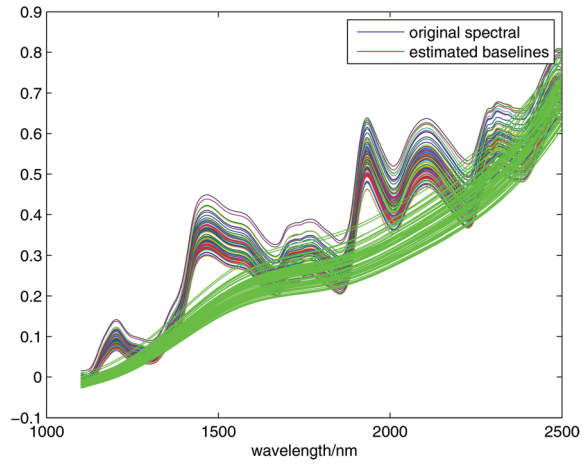

Fig. 10 The green lines are the estimated baselines by SSFBCSP and the spectra above them are the original spectra of the corn data set.

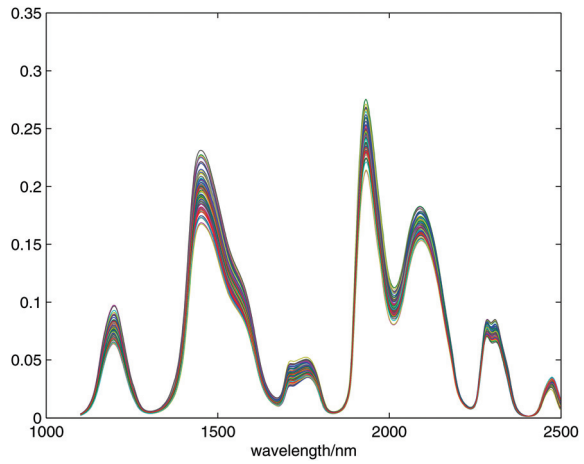

Fig. 11 The estimated spectra of the corn data set by SSFBCSP.

responses and the third one of every fifth sample was assigned to the test set, while the remaining samples were prepared for the calibration set. Partial least squares regression was adopted to build calibration models, the data set was mean centered and the number of latent variables was chosen by leave one out cross validation. To evaluate the performance of each preprocessing method, the accuracy and robustness of the calibration results were measured by root mean square error of prediction (RMSEP) and coefficient of determination (R^2). The optimal parameters in all algorithms were determined by grid research. The results are shown in Tables 3 and 4, respectively. The calibration results with no preprocessing (NO) are served as a benchmark.

From Tables 3 and 4, we can see that all preprocessing methods except ATEB do not gain an improvement on oil response, but our method is slightly inferior to the no preprocessing method. For protein response, CC has the best result, while for the moisture and the starch response, our method is better than other methods. Moreover, the large coefficients of determination in our algorithm show that our method is more robust than other methods and is comparable to the ATEB method.

3.2.2 Results of the marzipan data set. The marzipan data set has been used in ref. 29, where the root mean square error of cross-validation (RMSECV) was recorded to compare various infrared and near infrared set-ups and sampling techniques. By applying our algorithm and setting $\lambda_1 = 2 \times 10^7$ and $\lambda_2 = 0.08$, the estimated baselines and fitted spectra are shown in Fig. 12 and 13, respectively. The situation of the representation coefficient for each spectrum is similar to the corn data set. Besides preprocessing methods: SNV, MSC, Extended inverse scatter correction (EISC) and RMSC which can alleviate the scatter effect of spectra, some other baseline correction methods have been included. The asymmetric least squares (asLS), CC, ATEB and BCCPLS were used for comparison. After the preprocessing procedure, PLS was performed on the corrected spectra and leave one out cross validation was used to obtain the RMSECV. In order to avoid overfitting, the optimal number of latent variables was chosen by the F-test at the 95% confidence level.[30] The final results are listed in Table 5 in detail.

From Table 5 we can see that for moisture response, the RMSC method gives the best result, this is probably because the spectra have a severe nonlinear effect with the concentration. BCCPLS which combines the calibration result with baseline correction is slightly more consistent than our suggested method. But for sugar, the RMSECV in our method achieves a significant improvement than other methods. In addition to these methods, asLS obtains improvement com-

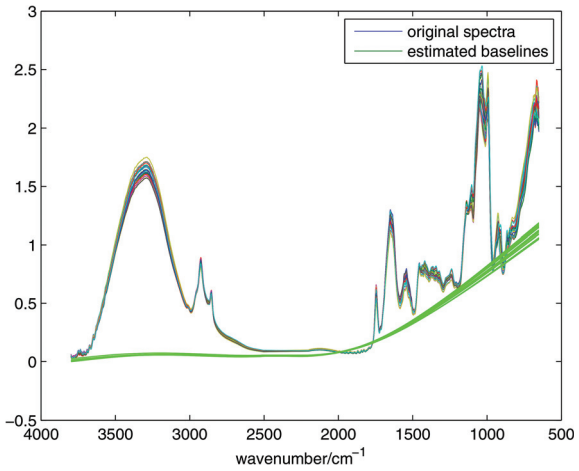

Fig. 12 The green lines are the estimated baselines by SSFBCSP and the spectra above them are the original spectra of the marzipan data set.

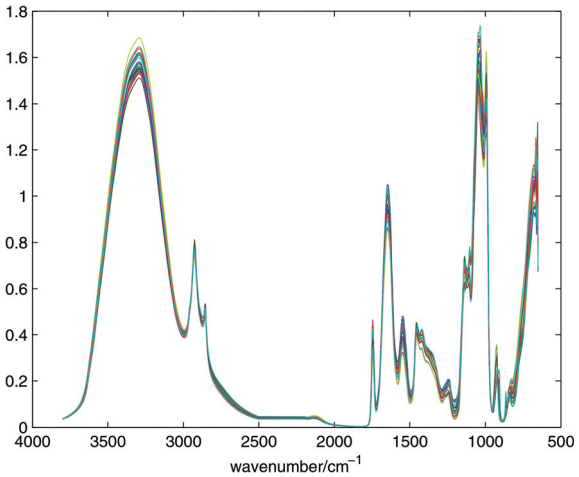

Fig. 13 The estimated spectra of the marzipan data set by SSFBCSP.

Table 3 RMSEP for each preprocessing method for the corn data set

	Moisture	Oil	Protein	Starch
NO	0.1223(9)[a]	0.0868(8)	0.1636(13)	0.4002(11)
MSC	0.1855(9)	0.1024(6)	0.1499(8)	0.3803(9)
RMSC	0.1605(10)	0.0965(7)	0.1658(8)	0.3602(8)
SNV	0.2305(6)	0.1024(6)	0.1501(8)	0.3688(9)
MSBC	0.1203(8)	0.0932(8)	0.1305(15)	0.3359(8)
CC	0.1179(8)	0.0936(9)	**0.1146(10)**	0.3942(10)
ATEB	0.1163(9)	**0.0864(7)**	0.1198(9)	0.3982(9)
SSFBCSP	**0.1108(9)**	0.0944(7)	0.1303(8)	**0.3329(8)**

[a] The values in parentheses refer to the number of latent variables.

Table 4 R^2 for each preprocessing method for the corn data set

	Moisture	Oil	Protein	Starch
NO	0.9556	0.8990	0.9511	0.8974
MSC	0.8881	0.8471	0.9675	0.8956
RMSC	0.9177	0.8710	0.9617	0.9034
SNV	0.8082	0.8466	0.9674	0.8989
MSBC	0.9520	0.8798	0.9668	0.9082
CC	0.9627	0.8641	0.9766	0.8992
ATEB	0.9646	**0.9158**	**0.9797**	0.8915
SSFBCSP	**0.9726**	0.8974	0.9751	**0.9218**

Table 5 RMSECV for each preprocessing method on the marzipan data set

	Moisture	Sugar
NO	0.592(7)[a]	2.242(6)
SNV	0.608(9)	1.944(6)
MSC	0.610(9)	1.943(6)
EISC	0.655(8)	1.889(6)
RMSC	**0.4975(9)**	1.99(6)
asLS	0.573(7)	2.075(7)
CC	0.6238(7)	1.9455(6)
ATEB	0.5954(7)	2.0845(6)
BCCPLS	0.532(7)	2.044(6)
SSFBCSP	0.536(8)	**1.689(7)**

[a] The values in parentheses refer to the number of latent variables.

pared to the no preprocessing method both in moisture and sugar responses. It is known that asLS has problem dealing with spectra with broad peaks, and its performance is inferior compared to our method.

3.3 Discussion of the processing speed and the selection of hyper-parameters

Our method has achieved the desired results on both simulated data set and real data sets; one may wonder how efficient this method is and how its parameters can be tuned. The execution times in the two cases of the simulated data set and two real data sets are shown in Table 6. The execution times in the other two cases of the simulated data set are similar and are omitted here. It can be seen that our method is not efficient compared with other baseline correction methods whose execution times are less than one second. However, one merit of this method is that it can be applied successfully to the noisy data set. Since our method estimates the baseline and pure spectrum simultaneously, the noise can be corrected implicitly. Furthermore, multiple spectra can be processed simultaneously. For instance, for the corn data set containing 80 samples, the average execution time taken for each spectrum is just about 0.25 second. One solution for the bottleneck of the execution time is to design another optimization method which is more expedient than the surrogate function method to solve the l_1-norm problem.

Two parameters in our method need to be set. For the roughness parameter λ_2, there are some intuitions in other baseline correction methods such as asLS. For the sparsity parameter λ_1, one consideration is to approximately set it as the ratio between the standard error of noise and the standard deviation of the expected non-zeros in the solution of the representation coefficient. For our simulated data set, since the standard error of noise and the standard deviation of the expected non-zeros are known, the value of λ_1 is set to about 0.1 firstly, and then the second parameter is adjusted. We found that good fitting results can be obtained. By tuning the parameters finely, we can even achieve better results. For the real data set, an initial guess for λ_1 such as 0.1, the same as for the simulate data set, is fixed firstly and then the rest of the process is the same as that of the simulated data set. Fortunately, there is no dependence on these parameters and there are many pairs of parameters which can give the desired results.

Table 6 Execution time of the SSFBCSP algorithm for each data set

	GNSB	UNSB	Corn	Marzipan
Time	6.66(s)	6.46(s)	22.94(s)	20.64(s)

4 Conclusions

This paper addresses the problem of estimating the baseline and pure spectrum simultaneously. The performance of baseline correction and peak estimation is evaluated on one simulated data set and two real data sets; the experiment results show that our method has some improvement compared to other single baseline correction methods and scatter correction methods. In particular, the large coefficient of determination in the corn data set shows that our method is robust. Since the dictionary must be constructed beforehand in our method, how to obtain an adaptive dictionary which can give a sparser representation coefficient of the pure spectrum is considered in our following studies.

Acknowledgements

The authors would like to gratefully acknowledge the anonymous reviewers for their insightful comments and constructive suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 61571438 and 61601104) and the Beijing Natural Science Foundation (Grant No. 1152001).

References

1 B. G. Osborne and T. Fearn, *Near infrared spectroscopy in food analysis*, Longman, 1986.
2 M. Morháč, *Nucl. Instrum. Methods Phys. Res., Sect. A*, 2009, **600**, 478–487.
3 K. H. Liland, T. Almøy and B.-H. Mevik, *Appl. Spectrosc.*, 2010, **64**, 1007–1016.
4 I. W. Selesnick, H. L. Graber, D. S. Pfeil and R. L. Barbour, *IEEE Trans. Signal Process.*, 2014, **62**, 1109–1124.
5 F. Gan, G. Ruan and J. Mo, *Chemom. Intell. Lab. Syst.*, 2006, **82**, 59–65.
6 P. H. C. Eilers, *Anal. Chem.*, 2004, **76**, 404–411.
7 Z.-M. Zhang, S. Chen and Y.-Z. Liang, *Analyst*, 2010, **135**, 1138–1146.
8 S.-J. Baek, A. Park, Y.-J. Ahn and J. Choo, *Analyst*, 2015, **140**, 250–257.
9 J. Peng, S. Peng, A. Jiang, J. Wei, C. Li and J. Tan, *Anal. Chim. Acta*, 2010, **683**, 63–68.

10 J. J. de Rooi and P. H. C. Eilers, *Chemom. Intell. Lab. Syst.*, 2012, **117**, 56–60.
11 Y. Liu, X. Zhou and Y. Yu, *Analyst*, 2015, **140**, 7984.
12 X. Liu, Z. Zhang, Y. Liang, P. F. M. Sousa, Y. Yun and L. Yu, *Chemom. Intell. Lab. Syst.*, 2014, **139**, 97–108.
13 M. Morháč, J. Kliman, V. Matoušek, M. Veselský and I. Turzo, *Nucl. Instrum. Methods Phys. Res., Sect. A*, 1997, **401**, 113–132.
14 J. Peng, S. Peng, Q. Xie and J. Wei, *Anal. Chim. Acta*, 2011, **690**, 162–168.
15 R. N. Jones, K. S. Seshadri, N. B. W. Jonathan and J. W. Hopkins, *Can. J. Chem.*, 1963, **41**, 750–762.
16 R. D. B. Fraser and E. Suzuki, *Anal. Chem.*, 1969, **41**, 37–39.
17 A. J. Brown, *IEEE Trans. Geosci. Remote Sens.*, 2006, **44**, 1601–1608.
18 P. R. Bevington and D. K. Robinson, *Data reduction and error analysis*, McGraw-Hill, 2003.
19 J. Mairal, F. Bach and J. Ponce, 2014, arXiv Prepr. arXiv1411.3230.
20 X. Ning, I. W. Selesnick and L. Duval, *Chemom. Intell. Lab. Syst.*, 2014, **139**, 156–167.
21 S. G. Mallat and Z. Zhang, *IEEE Trans. Signal Process.*, 1993, **41**, 3397–3415.
22 E. T. Whittaker, *Proc. Edinburgh Math. Soc.*, 1922, **41**, 63–75.
23 S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, *Found. Trends Mach. Learn.*, 2011, **3**, 1–122.
24 P. W. Holland and R. E. Welsch, *Commun. Stat. Methods*, 1977, **6**, 813–827.
25 K. Lange, D. R. Hunter and I. Yang, *J. Comput. Graph. Stat.*, 2000, **9**, 1–20.
26 P. Geladi, D. MacDougall and H. Martens, *Appl. Spectrosc.*, 1985, **39**, 491–500.
27 Y. Mou, X. You, D. Xu, L. Zhou, W. Zeng and S. Yu, *Chemom. Intell. Lab. Syst.*, 2013, **132**, 168–174.
28 R. J. Barnes, M. S. Dhanoa and S. J. Lister, *Appl. Spectrosc.*, 1989, **43**, 772–777.
29 J. Christensen, L. Norgaard, H. Heimdal, J. G. Pedersen and S. B. Engelsen, *J. Near Infrared Spectrosc.*, 2003, **12**, 63–75.
30 D. M. Haaland and E. V. Thomas, *Anal. Chem.*, 1988, **60**, 1193–1202.