Iterative Reweighted Quantile Regression Using Augmented Lagrangian Optimization for Baseline Correction

Publisher: IEEE

Cite This

PDF

Quanjie Han ; Silong Peng ; Qiong Xie ; Yifan Wu ; Genwei Zhang **All Authors**

73
Full
Text Views

Alerts

Manage Content Alerts

Add to Citation Alerts

Abstract: Based on baseline is a smooth curve and under the collected spectrum, a robust penalized quantile regression with B-spline basis has been proposed to baseline estimation.... **View more**

Metadata

Abstract:

Based on baseline is a smooth curve and under the collected spectrum, a robust penalized quantile regression with B-spline basis has been proposed to baseline estimation. Then an iterative reweighted method has been adopted for quantile regression optimization. Instead of man tuning the hyperparameter in penalized quantile regression, augmented Lagrangian method is applied to hyperparameter optimization. Experiments on simulated and real data sets show that our method is more effective in baseline correction than other baseline estimation methods in simulated data set. For real data set, the calibration results after the baseline correction step are better than other preprocessing and baseline correction methods.

Published in: 2018 5th International Conference on Information Science and Control Engineering (ICISCE)

Date of Conference: 20-22 July 2018

Date Added to IEEE *Xplore*: 17 January 2019

ISBN Information:

INSPEC Accession Number: 18410561

DOI: 10.1109/ICISCE.2018.00066

Publisher: IEEE

Conference Location: Zhengzhou, China

Downl
PDF

Contents

SECTION I. Introduction

Since Fourier transform spectrometer is rapid and nondestructive, Fourier transform infrared spectroscopy has been widely used in Chemometrics, food, wine and other related fields for sample components analysis [1]. Generally speaking, the obtained Fourier transform infrared absorption spectroscopy consists of the true sample spectrum, baseline and noise. Baseline together with noise will significantly deteriorate the performance of chemometric calibration algorithms, so baseline correction and spectrum denoising is an important preprocessing step for spectrum quantitative analysis.

Baseline estimation can be dated back to late 1970s [2]. Up to now, there have several assumptions been imposed on baseline: Firstly, from the frequency prospective, baseline is in low frequency part while noise generally lives in the high frequency part, low-pass filter has been constructed to correct the baseline [3], [4]. Secondly, baseline is a smooth curve which underlies the collected spectrum: it was fitted by polynomials [5] and Bernstein polynomials were proposed for extraction of baseline of NMR signals [6]. Last but not the least, baseline points and spectrum peak points belong to different clusters, which can be separated. In order to separate the baseline points and peak points, Rooi proposed a mixture model for baseline estimation. The baseline points were characterized by a Gaussian distribution while the peak points were subject to a uniform distribution. Using EM algorithm, after the baseline points and peak points having been separated, a penalized B-splines basis was used to fitting the baseline [7]. The Gaussian mixture model was also used for DNA sequence baseline correction in [8]. Since baseline underlies the obtained spectrum, an asymmetrically weighted least squares (asLS) with roughness penalty was proposed for baseline estimation [9]. The weights for the baseline points below the spectrum were set manually by a constant which usually will overestimate the peak and there were two parameters need to be optimized, [10] proposed the adaptive iteratively reweighted Penalized Least Squares (airPLS) and a partially balanced weighting scheme was also proposed in [11] for baseline smoothing (arPLS). Besides, based on the spectrum of the sample can be approximated by Voight lineshape, a method simultaneously fitting the pure spectrum and baseline using sparse representation (SSFBCSP) was proposed in [12] and a multiple spectral baseline correction method which combined the information of several spectral was used for Guotai wine baseline correction [13]. Simple Least squares regression corresponding to the conditional mean value regression and the least squares regression is sensitive to outliers and noise. Besides, in order to obtain the regression equation for other quantiles, quantile regression was proposed by Koenker and Bassett [14] and it was first used for baseline correction in [15].

This article proposes a quantile regression with penalized B-splines for baseline correction, where the B-splines are used to represent the baseline. Instead of the primal-dual interior or simplex method for solving quantile regression problem, we propose the iterative reweighted least squares to tackle the quantile regression, which has several advantages: it is easy to implement than the linear programming methods; iterative

Downl
PDF

Contents

reweighted least squares usually gives more accurate result. In order to avoid the optimization of the regularization parameter in penalized B-splines, the augmented Lagrangian method is also proposed for hyperparameter optimization. This paper is divided into the following parts: Section II provides the detailed introduction for our algorithms; Section III displays our experiments setting; The experiments results and discussion are shown in Section IV; The final part is devoted to Conclusion.

SECTION II. Problem Formulation

A. P-Splines and Quantile Regression

Roughness penalty approach to problems in regression has gained much popularity in recent years, especially in functional data analysis (FDA) [16]. Considering the nonparametric regression problem: given n data pairs $(x_i,\ y_i), i=1,2,\ldots,n$, find a function g such that

$$y_i=g(x_i)+e_i \tag{1}$$

View Source

Where e_i is the error term in i-th sample, which is usually assumed normally distributed. Without constraint on g, then the error term can be zero just by interpolating the points using piecewise linear function. In order to find a compromise between the fidelity of curve fitting and avoiding of rapidly fluctuating curve, a penalty is posed on the curvature of g, then the smooth penalty based regression becomes

$$\sum_{i=1}^{n}(y_i-g(x_i))^2+\lambda\int g''(x)dx \tag{2}$$

View Source

the second derivative can be replaced by other higher order derivatives. For computation convenience, the function g is represented by some basis functions usually. In functional data analysis, the Fourier basis is used for periodic functions, while B-spline basis is adopted for nonperiodic functions. Assume that g can be represented by B-spline basis

$$g=B\boldsymbol{\alpha}=\sum_{j}B_{i,j}\alpha_j \tag{3}$$

View Source

By the recurrence relation and formula for derivatives of B-splines given by de Boor [17], the continuous penalty $\int g''(x)dx$ is equivalent to the smooth of the representation coefficient. Then (2) becomes

$$\Vert\boldsymbol{y}-B\boldsymbol{\alpha}\Vert_2^2+\lambda\Vert D\boldsymbol{\alpha}\Vert_2^2 \tag{4}$$

View Source

where D is the difference matrix, the difference order is usually set to and three. The B-splines with penalty is called P-splines [18].

Download PDF

Contents

From the maximal likelihood point of view, the least squared term is corresponding to the noise is Gaussian and what we obtain is the conditional expectation of y given x, which is sensitive to outliers. To get robust estimator, the absolute deviation has been adopted which corresponding to the conditional median. In order to get the information of other quantiles, Konerker proposed the quantile regression, which can be formulated as an optimization problem:

$$arg\min_{\boldsymbol{z}} \sum_i \rho_\tau(y_i - z_i) \tag{5}$$

View Source

where $\rho_\tau = u(\tau - 1(u < 0)) = \tau u + +(1-\tau)u_-, u_+ = \max\{u,\ 0\}$ is the positive part of u, while $u_- = \max\{-u,\ 0\}$ is the negative part of u. The median regression is corresponding to $\tau = 0.5$.

B. Iterative Reweighted Quantile Regression with Augmented Lagrangian Optimization

Since baseline is running below the spectrum, we should impose asymmetrically penalty for estimated points. For points above the original spectrum, a large penalty should be set, while for points under it, a small penalty should imposed. From quantile point of view, the baseline is at the low quantile part of the original spectrum.

Due to the smooth of baseline, it can be represented by B-splines

$$\boldsymbol{z} = B\boldsymbol{\alpha} \tag{6}$$

View Source

Then the quantile regression with P-splines for baseline correction can be described as follows:

$$arg\min_{\boldsymbol{\alpha}} \sum_i \rho_\tau(y_i - \sum_j B_{i,j}\alpha_j) + \lambda \|D\boldsymbol{\alpha}\|^2 \tag{7}$$

View Source

After the representation coefficient $\boldsymbol{\alpha}$ is obtained, then the baseline is estimated by $z = B\boldsymbol{\alpha}$.

Considering that $\rho_\tau = u(\tau - 1(u < 0)) = \tau u_+ + (1-\tau)u_-$ is not differentiate at zero. By using iterative reweighted least squares and noting that $|u| = \frac{u^2}{|u|}$, we can set the asymmetric weight as

$$w_i = \begin{cases} \frac{\tau}{|y_i - z_i| + \epsilon} & y_i \geq z_i \\ \frac{1-\tau}{|y_i - z_i| + \epsilon} & y_i < z_i \end{cases} \tag{8}$$

View Source

where $\epsilon > 0$ is added by avoiding the divided by zero problem. Then the quantile regression can be optimized by

$$arg\min_{\boldsymbol{\alpha}} \sum_i w_i (y_i - \sum_j B_{i,j}\alpha_j))^2 + \lambda \|D\boldsymbol{\alpha}\|_2^2 \tag{9}$$

View Source

Download PDF

Contents

In reality, the success of baseline estimation depends on choosing the hyper-parameter λ properly. To avoid the tuning of λ, we propose to paraphrase (9) as

$$arg\min_{\boldsymbol{\alpha}}\sum_i w_i(y_i-\sum_j B_{i,j}\alpha_j))^2, -\epsilon\leq D\boldsymbol{\alpha}\leq\epsilon \qquad (10)$$

View Source

where the inequality is applied element-wise. With augmented Lagrangian optimization, (10) becomes

$$arg\min_{\boldsymbol{\alpha}}\sum_i w_i(y_i-\sum_i B_{i,j}\alpha_j))^2+\boldsymbol{v}^T D\alpha+\frac{\rho}{2}\|D\boldsymbol{\alpha}\|_2^2$$

View Source

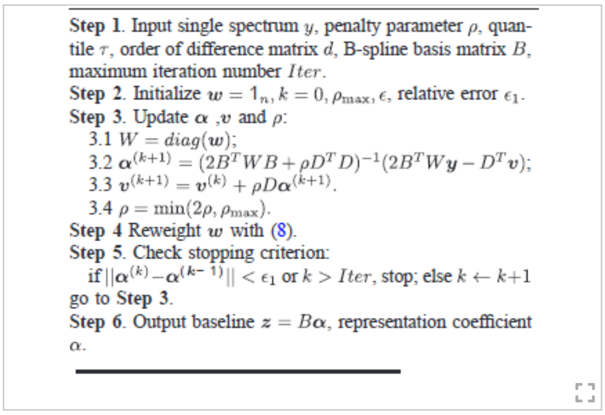
Let W denote the diagonal matrix with $\boldsymbol{w}=(w_i)$ on its diagonal, (11) can be described as

$$arg\min_{\boldsymbol{\alpha}}(\boldsymbol{y}-B\boldsymbol{\alpha})^T W(\boldsymbol{y}-B\boldsymbol{\alpha})+\boldsymbol{v}^T D\boldsymbol{\alpha}+\frac{\rho}{2}\|D\boldsymbol{\alpha}\|_2^2 \qquad (1$$

View Source

where $\boldsymbol{v}$ is the Lagrangian multipliers and ρ is a penalty parameter. We summarize the iterative reweighted quantile regression with augmented Lagrangian Optimization method (IRQRAL) as follows:

Algorithm 1: Irqral



```
Step 1. Input single spectrum $y$, penalty parameter $\rho$, quantile $\tau$, order of difference matrix $d$, B-spline basis matrix $B$, maximum iteration number $Iter$.
Step 2. Initialize $\boldsymbol{w}=1_n, k=0, \rho_{\max}, \epsilon$, relative error $\epsilon_1$.
Step 3. Update $\boldsymbol{\alpha}$ ,$\boldsymbol{v}$ and $\rho$:
    3.1 $W=diag(\boldsymbol{w})$;
    3.2 $\boldsymbol{\alpha}^{(k+1)}=(2B^TWB+\rho D^TD)^{-1}(2B^TW\boldsymbol{y}-D^T\boldsymbol{v})$;
    3.3 $\boldsymbol{v}^{(k+1)}=\boldsymbol{v}^{(k)}+\rho D\boldsymbol{\alpha}^{(k+1)}$.
    3.4 $\rho=\min(2\rho,\rho_{\max})$.
Step 4 Reweight $\boldsymbol{w}$ with (8).
Step 5. Check stopping criterion:
    if $||\boldsymbol{\alpha}^{(k)}-\boldsymbol{\alpha}^{(k-1)}||<\epsilon_1$ or $k>Iter$, stop; else $k\leftarrow k+1$
go to Step 3.
Step 6. Output baseline $\boldsymbol{z}=B\boldsymbol{\alpha}$, representation coefficient $\alpha$.
```

SECTION III. Experiments

A. Dataset Description

To evaluate the performance of the proposed method, one simulated data and one real data set are used for quantitative analysis. The simulated data consists of six Gaussian peaks and a sinusoidal baseline and an exponential baseline are added to the peaks respectively. Besides, a uniform random noise is generated whose amplitude doesn't exceed 0.01 of the maximal height of the peaks. The real data is the corn data set which consists of 80 NIR spectra of corn measured on spectrometers mp5 and mp6 respectively and the spectra were collected in the region of 1100-2498 nm. There are four components have been measured: moisture, oil, protein, starch. In our experiment, the mp5 data set is adopted to compare the calibration result of our method with other preprocessing methods after the baselines being corrected.

B. Model Evaluation

For simulated data set, the true baseline is known, we can use root mean squared error (RMSE) to find the optimal parameters. In our experience, the quantile $\tau = 0.01$ and the order of difference matrix $d = 3$ can always give desired results. Since whatever we choose ρ, the IRQRAL algorithm will converge, so we fix $\rho = 1$. The RMSE is computed by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{n}} \tag{13}$$

View Source

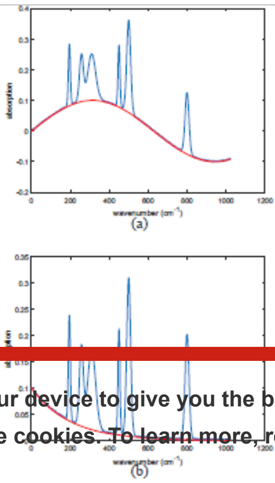
where z is the true baseline, $\hat{z}$ is the estimated baseline and n is the length of simulated data.

Since the true baselines for real data set are unavailable, the quantitative results of spectra after the baselines having been corrected are used to metric the success of baselines estimation. We split the data set into training set and test set. Firstly, each response component is sorted, then the second of every four is taken as test set, the others are treated as training set; then leaving one out cross validation is used for calibration. In order to avoid over fitting, a criterion based on testing the significance of incremental changes in PRESS with an F-test [19] is used for the choice of the number of latent variables. In this work, a 95% confidence interval is employed. Finally, the root mean squared error of prediction (RMSEP) is used to evaluate the performance of each method.

SECTION IV.
Results and Discussions

A. Simulated Data Sets

With respect to simulated data with sinusoidal baseline and exponential baseline, the estimated baselines by our algorithm are shown in Figure 1. The asLS, airPLS, arPLS, SSFBCSP baseline correction methods and the iterative reweighted quantile regression for baseline estimation without augmented Lagrangian optimization (IRQR) are used to compared with our method, the parameters of each method are optimized by grid search. The RMSE for each baseline correction method are detailed in Table I and Table II respectively.



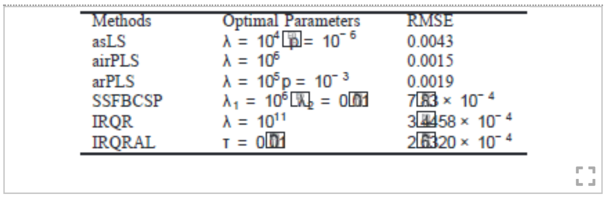
(a)

(b)

Figure 1.
(a) Original spectrum (blue) and estimated baseline (red) for peaks with sinusoidal baseline. (b) Original spectrum (blue) and estimated baseline (red) for peaks with exponential baseline.

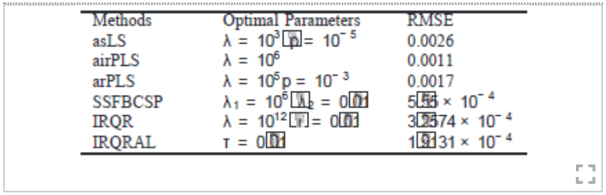
We can see that our method is better than the other methods and quantile regression with augmented Lagrangian optimization is generally outperforms the one without it.

Table I Rmse for each estimation method of sinusoidal baseline



Methods	Optimal Parameters	RMSE
asLS	$\lambda = 10^4$, $p = 10^{-6}$	0.0043
airPLS	$\lambda = 10^6$	0.0015
arPLS	$\lambda = 10^5 p = 10^{-3}$	0.0019
SSFBCSP	$\lambda_1 = 10^6$, $\lambda_2 = 0.01$	7.83×10^{-4}
IRQR	$\lambda = 10^{11}$	3.4458×10^{-4}
IRQRAL	$\tau = 0.01$	2.6320×10^{-4}

Table II Rmse for each estimation method of exponential baseline



Methods	Optimal Parameters	RMSE
asLS	$\lambda = 10^3$, $p = 10^{-5}$	0.0026
airPLS	$\lambda = 10^6$	0.0011
arPLS	$\lambda = 10^5 p = 10^{-3}$	0.0017
SSFBCSP	$\lambda_1 = 10^6$, $\lambda_2 = 0.01$	5.55×10^{-4}
IRQR	$\lambda = 10^{12}$, $\tau = 0.01$	3.2574×10^{-4}
IRQRAL	$\tau = 0.01$	1.9131×10^{-4}

B. Corn Data Set

The original spectral and the estimated baselines, the baseline corrected spectral by our algorithm are displayed in Figure 2. In Figure 2(a), we can see that corn data set spectral have severe baseline drift, which will adversely influence the calibration and prediction results conducted on it. While seeing from Figure 2(b), our method has successfully corrected the spectral to zero baseline.

To evaluate the performance of our method, spectral preprocessing methods include multiplicative scatter correction (MSC) [20], standard normal variate (SNV) [21], extended inverse scatter correction (EISC) [22], extended multiplicative signal correction (EMSC) [23] and baseline correction methods consist of asymmetrically least squares (asLS), multiple spectral baseline correction (MSBC) and simultaneous spectrum fitting and baseline correction using sparse representation (SSFBCSP) are used to get transformed spectral. The transformed spectra are mean centered before calibration. The RMSEP for moisture, oil, protein, starch are detailed in Table III. For oil component, no methods give desired results. But our algorithm gets better results in other three components and is better than the method without augmented Lagrangian optimization.

C. Discussion

Since our quantile regression problem is just an iterative weighted least squares problem, the Augmented Lagrangian method which can update the dual variables and penalty parameter gradually and the experiments show the desired result. Besides, for the generation of B-splines matrix, the location of knots of B-splines can be set the same as wave number of spectrum, but it may generate many redundant B-spline basis to represent the baseline. Set the location of knots optimally, which can improve the computation efficiency.

Downl
PDF

Contents

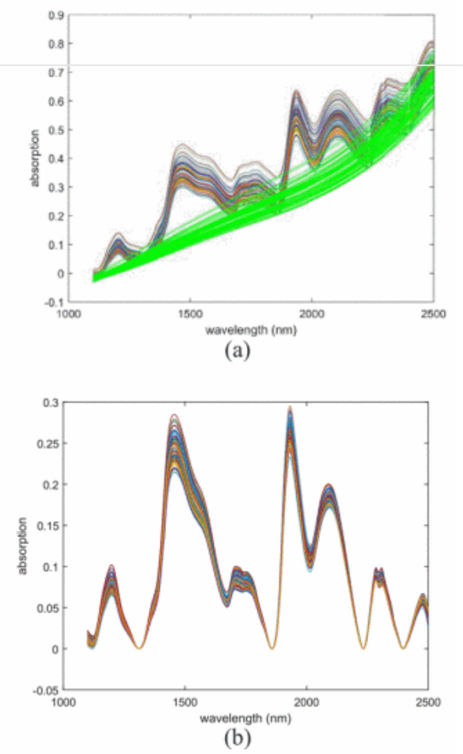
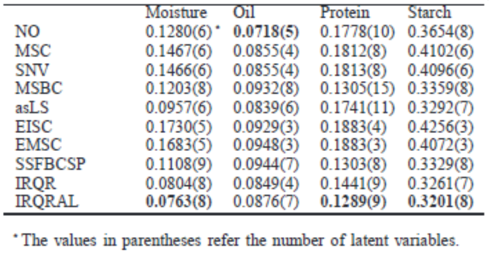


Figure 2.
(a) The estimated baselines (green) and original spectra. (b) Baselines corrected spectra.

Table III Rmsep for each method



	Moisture	Oil	Protein	Starch
NO	0.1280(6)*	**0.0718(5)**	0.1778(10)	0.3654(8)
MSC	0.1467(6)	0.0855(4)	0.1812(8)	0.4102(6)
SNV	0.1466(6)	0.0855(4)	0.1813(8)	0.4096(6)
MSBC	0.1203(8)	0.0932(8)	0.1305(15)	0.3359(8)
asLS	0.0957(6)	0.0839(6)	0.1741(11)	0.3292(7)
EISC	0.1730(5)	0.0929(3)	0.1883(4)	0.4256(3)
EMSC	0.1683(5)	0.0948(3)	0.1883(3)	0.4072(3)
SSFBCSP	0.1108(9)	0.0944(7)	0.1303(8)	0.3329(8)
IRQR	0.0804(8)	0.0849(4)	0.1441(9)	0.3261(7)
IRQRAL	**0.0763(8)**	0.0876(7)	**0.1289(9)**	**0.3201(8)**

* The values in parentheses refer the number of latent variables.

SECTION V.
Conclusion

The proposed method uses the iterative reweighted quantile regression and augmented Lagrangian optimization for the baseline estimation, which can free us of the parameter optimization used by other baseline estimation methods. And since the quantile regression is more robust than the least squares, this baseline estimation method can not only be used for the Gaussian noise situation, but also for other heterogeneous and dependent data error circumstance.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China (Grant NO.61571438) and Science Foundation Research Project of Beijing, China (Grant NO.1152001).

Contents

Downl
PDF

Authors

Figures

References

Keywords

Metrics

Footnotes

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | Privacy & Opting Out of Cookies

IEEE Account

» Change Username/Password

» Update Address

Purchase Details

» Payment Options

» Order History

» View Purchased Documents

Profile Information

» Communications Preferences

» Profession and Education

» Technical Interests

Need Help?

» **US & Canada:** +1 800 678 4333

» **Worldwide:** +1 732 981 0060

» Contact & Support

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | Sitemap | Privacy & Opting Out of Cookies