Analytical Methods

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2021, **13**, 2037

An improved PD-AsLS method for baseline estimation in EDXRF analysis

Qingxian Zhang,[ab] Hui Li,[*ab] Hongfei Xiao,[ab] Jian Zhang,[ab] Xiaozhe Li[ab] and Rui Yang[ab]

Received 22nd January 2021
Accepted 15th March 2021

DOI: 10.1039/d1ay00122a

rsc.li/methods

Baseline correction is an important step in energy-dispersive X-ray fluorescence analysis. The asymmetric least squares method (AsLS), adaptive iteratively reweighted penalized least squares method (airPLS), and asymmetrically reweighted penalized least squares method (arPLS) are widely used to automatically select the data points for the baseline. Considering the parametric sensitivity of the aforementioned methods and the statistical characteristics of the X-ray energy spectrum, this paper proposes an asymmetrically reweighted penalized least squares method based on the Poisson distribution (PD-AsLS) to automatically correct the baseline of X-ray spectra. Monte Carlo (MC) simulation is used to obtain the background spectrum, and PD-AsLS is used to estimate the baseline of the background. The relative error and the absolute error between the simulated background and PD-AsLS estimated background are used to determine the accuracy of PD-AsLS. The correlation coefficient (COR) and the root mean square error (RMSE) between the estmated baseline and the real baseline are calculated, and results of PD-AsLS are compared with results of three other classical methods (arPLS, airPLS and AsLS) to evaluate the reliability of PD-AsLS. The results of PD-AsLS show that the COR is above 0.95 and RMSE is less than 6. The stability and the practicability of PD-AsLS are also evaluated in experiments. A sample is measured five time to get its X-ray energy spectra, and the coefficient of variation (CV) of the estimated baseline is smaller than that of measured spectra. Experiments show that PD-AsLS can estimate baselines better than arPLS without any overestimation. Those results indicate that PD-AsLS can reliably estimate the baselines of X-ray spectra and effectively suppress the statistical fluctuation.

1. Introduction

In energy-dispersive X-ray fluorescence (EDXRF) analysis, baseline correction – the first step in spectrum analysis – is crucial.[1] It significantly affects the accuracy of the subsequent steps, such as multi-peak analysis and content calculation. The baseline can be estimated correctly by appropriately selecting data points of the baseline to fit the spectrum.[2]

In the baseline correction, the data points can be selected manually, semi-automatically, or automatically. The most frequently used methods for this purpose are the sensitive nonlinear iterative peak-clipping (SNIP) algorithm,[3–5] fitting,[6] and the iterative method.[2,7] However, SNIP relies on the obtainment of the average full width, but sometimes it is difficult to obtain the average full width, which limits the use of SNIP, and the iterative method also depends on low pass filters[8–10] to distinguish peaks from the baseline. Thus, Eilers and Boelens developed the asymmetric least squares method (AsLS) to automatically select data points for the baseline. However, the estimation of the AsLS depends on selected parameters.[11,12] Further, Zhang introduced the adaptive iteratively reweighted penalized least squares method (airPLS) to use in nuclear magnetic resonance (NMR); in this method, the weight vector $\mathbf{w}$ is obtained adaptively by using an iterative method.[13] Based on the generalized logistic function, Baek proposed the asymmetrically reweighted penalized least squares method (arPLS), which can estimate the noise level iteratively and adjust the weight factor accordingly.[14] However, for each channel on the whole spectral data, airPLS and arPLS adopt the same mean and standard deviation of the region where the measured signal is larger than the smoothed signal. Both methods do not take the statistical characteristics of counts at each channel into account.

In light of previous research studies, this paper proposed a new method, referred to as the asymmetrically reweighted penalized least squares method based on the Poisson distribution (PD-AsLS). In this method, based on the characteristic statistical model of the X-ray spectrum and the Poisson distribution, the weight is calculated iteratively by the preceding estimated result. Monte Carlo (MC) simulation and experiments are performed to evaluate the accuracy of PD-AsLS. The

[a]Chengdu University of Technology, Chengdu, Sichuan 610000, China. E-mail: lynn@stu.cdut.edu.cn

[b]Applied Nuclear Techniques in Geosciences Key Laboratory of Sichuan Province, Chengdu, Sichuan 610000, China

correlation coefficient (COR) and the root mean square error (RMSE) of PD-AsLS are calculated and compared with three other classical methods (arPLS, airPLS and AsLS) to evaluate the reliability of PD-AsLS. The stability and the practicability of PD-AsLS are also evaluated in experiments.

2. Methods

The Whittaker smoother has been successfully applied in several fields of science and works as follows.[10] Suppose that **y** and **z** are column vectors (y_i is a series of length m, and z_i is another series corresponding to y_i). In this smoothing method, **y** is the original signal and **z** is the smoothed signal. **z** can be obtained by minimizing the penalized least square function:

$$S = \sum_i^m (y_i - z_i)^2 + \lambda \sum_i^m \left(\Delta^2 z_i\right)^2, \quad (1)$$

where $\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2}$. The first term in S expresses the fidelity of **z** to **y**, the second term in S measures the roughness of data **z**, and λ is an adjustable parameter to balance the two terms.

Eilers and Boelens proposed the AsLS method, applying the Whittaker smoother to estimate the baseline (let **y** be the X-ray spectrum recorded at equal energy intervals, and let **z** be the smoothed baseline), and added the weight vector w_i to eqn (1). It then follows that

$$S = \sum_i^m w_i (y_i - z_i)^2 + \lambda \sum_i^m \left(\Delta^2 z_i\right)^2. \quad (2)$$

The minimization problems lead to the following system of equations:

$$(\mathbf{W} + \lambda \mathbf{D}'\mathbf{D})\mathbf{z} = \mathbf{W}\mathbf{y}, \quad (3)$$

where $\mathbf{W} = diag(w_i)$ and **D** is the second-order difference matrix: $\mathbf{Dz} = \Delta^2 \mathbf{z}$. AsLS also introduces an asymmetry factor p ($0 < p < 1$). If $y_i > z_i$, the weight is set to $w_i = p$; otherwise, $w_i = 1 - p$. In AsLS, the asymmetry factor p is given the same value for the entire baseline region. However, considering the difference between the smoothed baseline and the measured spectrum, the weight of different baseline regions should be given different values. Therefore, to achieve a satisfactory result, the asymmetry factor and smoothing parameters must be optimized.

To mitigate this problem, this paper takes the characteristic distribution of X-ray spectra into account. The sample is measured 30 times to obtain the energy spectrum, and the standard deviations are calculated as $\boldsymbol{\sigma}$ − SD. The standard deviation of the Poisson distribution is calculated and recorded as $\boldsymbol{\sigma}$ − PD ($\sigma_i = \sqrt{\bar{y}_i}$, where y is the energy spectrum count, i is the channel of the spectrum, and y_i represents the mean of measuring spectra). Fig. 1 shows a graph of $\boldsymbol{\sigma}$ − SD and $\boldsymbol{\sigma}$ − PD. The curve of $\boldsymbol{\sigma}$ − SD is similar to that of σ − PD, but the distribution of σ − SD is more dispersed.

Because the counts of the X-ray energy spectrum conform to the Poisson distribution, the baseline should also approximate

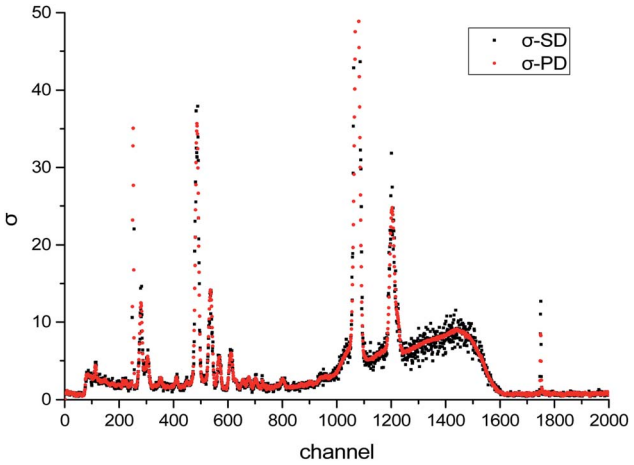

Fig. 1 Standard deviations of the sample and Poisson distribution.

the Poisson distribution. Based on the results above, the probability, $P(y_i,z_i)$, for each measured count to be the real count can be described as $P(y_i,z_i)$, which is a Poisson distribution varying randomly with the expectation. Then, the equation for $P(y_i,z_i)$ becomes

$$P(y_i, z_i) = \frac{{z_i}^{y_i}}{y_i!} \mathrm{e}^{-z_i}, \quad (4)$$

where z_i is the estimated baseline count, as the expectation in Poisson distribution, and y_i is the measured spectrum count. According to AsLS, we assume that when $y_i < z_i$, $w_i = 1$, and when $y_i \geq z_i$, w_i is the probability distribution function. The equation for w_i becomes

$$w_i = \begin{cases} K(y_i, z_i), & y_i \geq z_i \\ 1, & y_i < z_i \end{cases}. \quad (5)$$

The function $K(y_i,z_i)$ in eqn (5) is the normalization function to increase the estimation accuracy, represented as follows:

$$K(y_i, z_i) = \frac{P(y_i, z_i)}{P(z_i, z_i)} = \frac{\frac{{z_i}^{y_i}}{y_i!} \mathrm{e}^{-z_i}}{\frac{{z_i}^{z_i}}{z_i!} \mathrm{e}^{-z_i}} = \frac{z_i!}{y_i!} {z_i}^{y_i - z_i}. \quad (6)$$

When y_i is far large than z_i, the probability for it to be the baseline is small, so it can be assumed that $w_i = 0$. When $y_i \geq z_i$ and the difference is not large, it can be considered that the difference between y_i and z_i is caused by statistical fluctuations, and assumed that $0 < w_i < 1$. As shown in Fig. 2, when $y_i < z_i$, $w_i = 1$; further, when $y_i \geq z_i$, w_i becomes smaller. In comparison to the curve $z_i = 5$, when z_i is set to a larger value, such as $z_i = 30$, the curve exhibits a gentler decline.

The baseline estimation process is shown in Fig. 3.

(1) At the start, **D** is a difference matrix and λ is defined by the user. At the first iteration, $\boldsymbol{W}$ is the identity matrix.

(2) Calculate **z** from the equation $(\mathbf{W} + \lambda \mathbf{D}'\mathbf{D})\mathbf{z} = \mathbf{W}\mathbf{y}$ (**z** is the baseline at the current step).

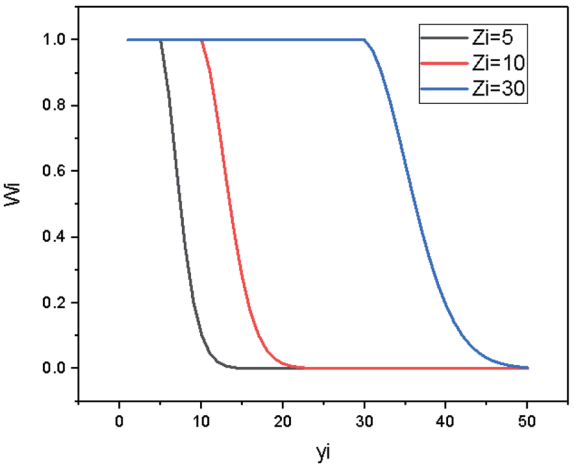

Fig. 2 Relationship between w_i and y_i with varying z_i values.

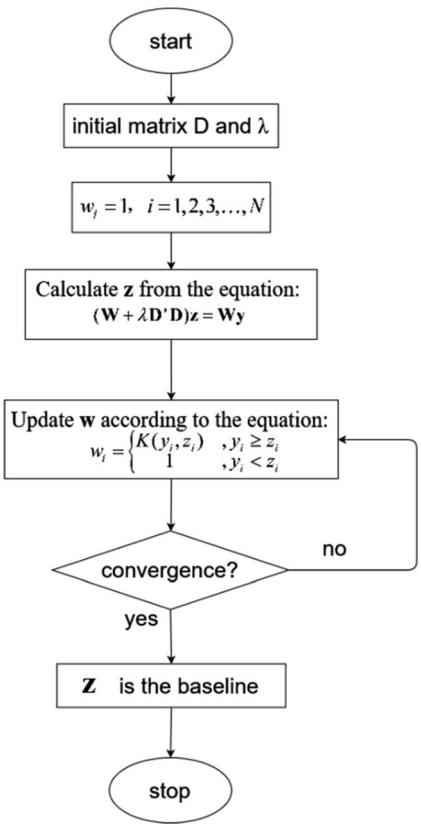

Fig. 3 Baseline estimation process.

(3) Update $\mathbf{w}$ according to the equation: $w_i = \begin{cases} K(y_i, z_i) & , y_i \geq z_i \\ 1 & , y_i < z_i \end{cases}$

(4) Check the stop condition, as follows:

$$(z_i^{\mathrm{j}} - z_i^{\mathrm{j}-1}) < 0.1, \quad i = 1, 2, 3 \ldots m,$$

where j is the j^{th} iteration. If the stop condition is reached, the iteration is halted; else, steps (2) to (4) are repeated.

3. Simulation and experiments

3.1 Simulation data analysis

According to ref. 15, the MC simulation is used to evaluate the baseline estimation algorithm and calculate the relative error. The simulation structure is set according to the actual measurement condition, using a Si-PIN detector. The working voltage of the X-ray tube is 35 keV and the working current is 1 μA. In Fig. 4(a), the X-ray energy spectrum is recorded by the detector – as the energy deposition spectrum – and the broadening spectrum is calculated using the Gaussian energy broadening formula. By eliminating the characteristic peaks of the energy deposition spectrum, the continuous actual background can be obtained.[15] Then, the proposed baseline correction method, PD-AsLS, is used to calculate the estimated background. The absolute error and the relative error between the actual background and the estimated background is used to evaluate the effectiveness of PD-AsLS.

As shown in Fig. 4(a), the energy deposition spectrum and broadening spectrum are very close between the 200 and 1024 channel, but greatly differ between the 80 and 200 channel. Fig. 4(b) and (c) show different estimated baselines of the energy deposition spectra with different values of λ; the result shows that the shape of the estimated baseline is consistent with that of the real baseline. The baseline can be easily obtained by removing the characteristic X-ray counts. The absolute error and the relative error between the estimated baseline and the real spectrum are shown in Fig. 4(d); the blue line is the relative error. As seen in Fig. 4(d), the error is large between the 60 and 180 channel. In the corresponding part of Fig. 4(a), the counts are small, the characteristic peaks overlap, and the difference between the estimated baseline and the real spectrum is significant. In the remaining channels of the spectrum, the relative error is less than 10%. Based on calculations, the average relative error of the rest of the channels is 5.69%.

In this paper, to ensure the effectiveness of PD-AsLS, we adopted the COR and RMSE to evaluate the accuracy of the baseline estimating method. The COR indicates the consistency between the estimated baseline and the real baseline, while the RMSE expresses the difference between them. These terms can be expressed as follows:

$$\mathrm{COR}(Z, B) = \frac{\sum (z - \bar{z})\left(b - \bar{b}\right)}{\sqrt{\sum (z - \bar{z})^2 \left(b - \bar{b}\right)^2}}, \tag{7}$$

$$\mathrm{RMSE}(Z, B) = \sqrt{\frac{\sum (z - b)^2}{N - 1}}, \tag{8}$$

where Z is the estimated baseline, B is the baseline from the energy deposition spectrum, N is the channel count, $\bar{z}$ is the mean of the estimated baseline, and $\bar{b}$ is the mean of the energy deposition spectrum.

Various values of λ are set and the results are recorded in Table 1. It can be seen from the table that, when $\lambda > 1 \times 10^4$, the RMSE and COR change only slightly with changes in λ. This

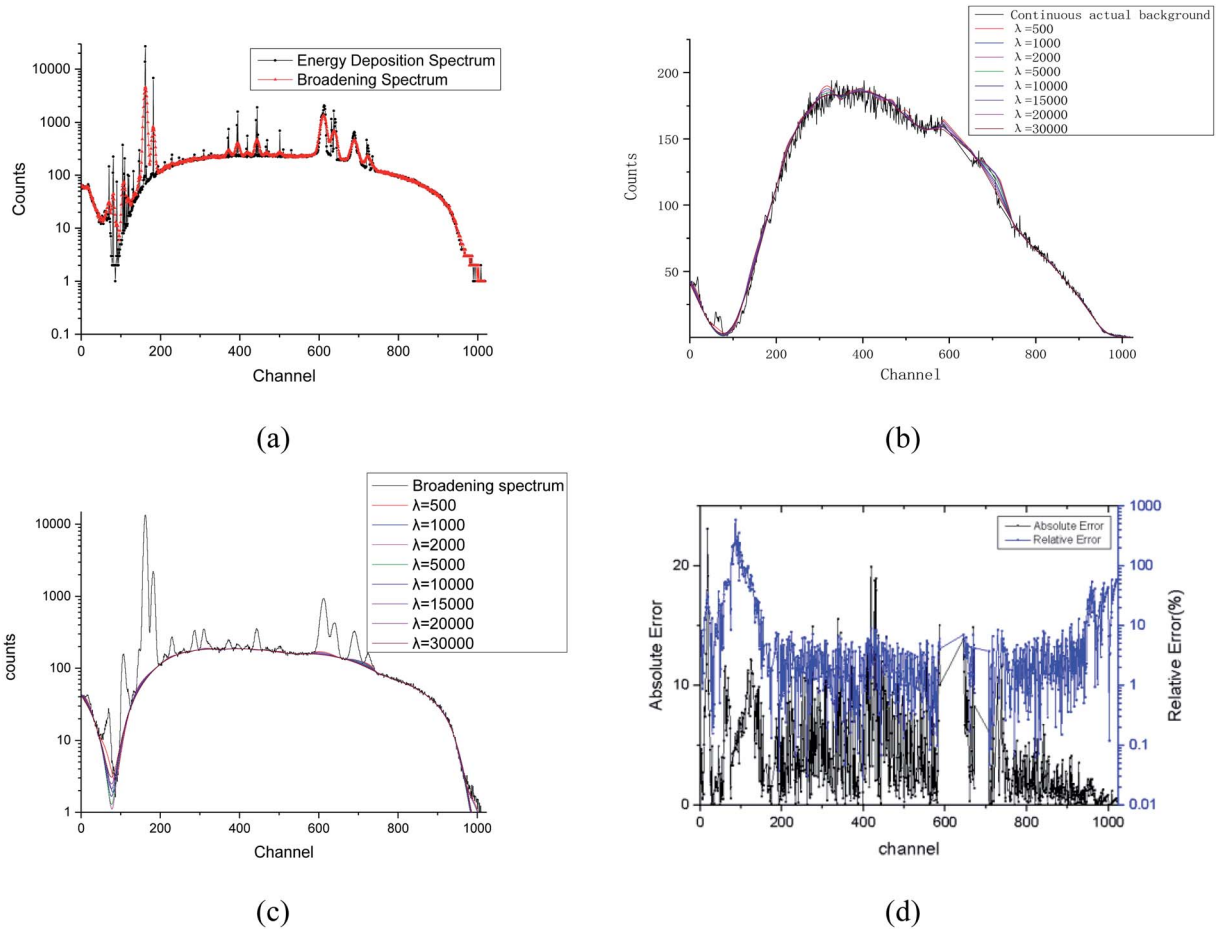

Fig. 4 Energy deposition spectrum and broadening spectrum (a); comparison of the continuous actual background and the baseline estimated by PD-AsLS with different values of λ (b); the broadening spectrum and the baseline estimated by PD-AsLS with different values of λ (c); and the absolute error and the relative error of PD-AsLS (d).

Table 1 Results for the COR and RMSE with varying values of λ

λ	5×10^2	1×10^3	2×10^3	5×10^3	1×10^4	1.5×10^4	2×10^4	3×10^4
COR	0.997	0.997	0.997	0.997	0.997	0.997	0.998	0.998
RMSE	5.825	5.670	5.558	5.350	5.167	5.091	4.979	5.021

indicates that PD-AsLS is not sensitive to the value of λ, proving that PD-AsLS can estimate the baseline stably and efficiently.

At $\lambda = 10\,000$, $p = 0.001$ (another parameter of AsLS), Table 2 lists the results of the COR and RMSE obtained from five samples respectively estimated by four methods (PD-AsLS, arPLS, airPLS and AsLS). It is clear from the table that values for the COR of all four methods are close to 1, which implies that the curve of the estimated baseline is close to that of the real baseline, and those methods can estimate the baseline effectively. The value of RMSE estimated by PD-AsLS is less than that estimated by the other three methods (arPLS, airPLS and AsLS), which verifies that the baseline estimated by PD-AsLS is closer to the actual baseline.

Table 2 Comparison of the COR and RMSE values estimated by PD-AsLS, arPLS, airPLS and AsLS

	COR				RMSE			
No.	PD-AsLS	arPLS	airPLS	AsLS	PD-AsLS	arPLS	airPLS	AsLS
1	0.998	0.998	0.994	0.989	5.469	6.078	11.292	15.939
2	0.998	0.997	0.971	0.964	4.842	6.433	18.177	22.833
3	0.998	0.997	0.968	0.958	4.720	5.768	18.318	23.995
4	0.998	0.998	0.993	0.975	5.262	5.821	9.752	21.034
5	0.998	0.998	0.994	0.993	5.591	6.704	11.887	13.961

3.2 Experimental data analysis

The instrument used for experimentation was a micro X-ray tube working at a high voltage of 35 kV and a current of 2 μA. It consists of a Ag target as the anode, a shield that has an output window, and an aluminium filer, with a thickness of

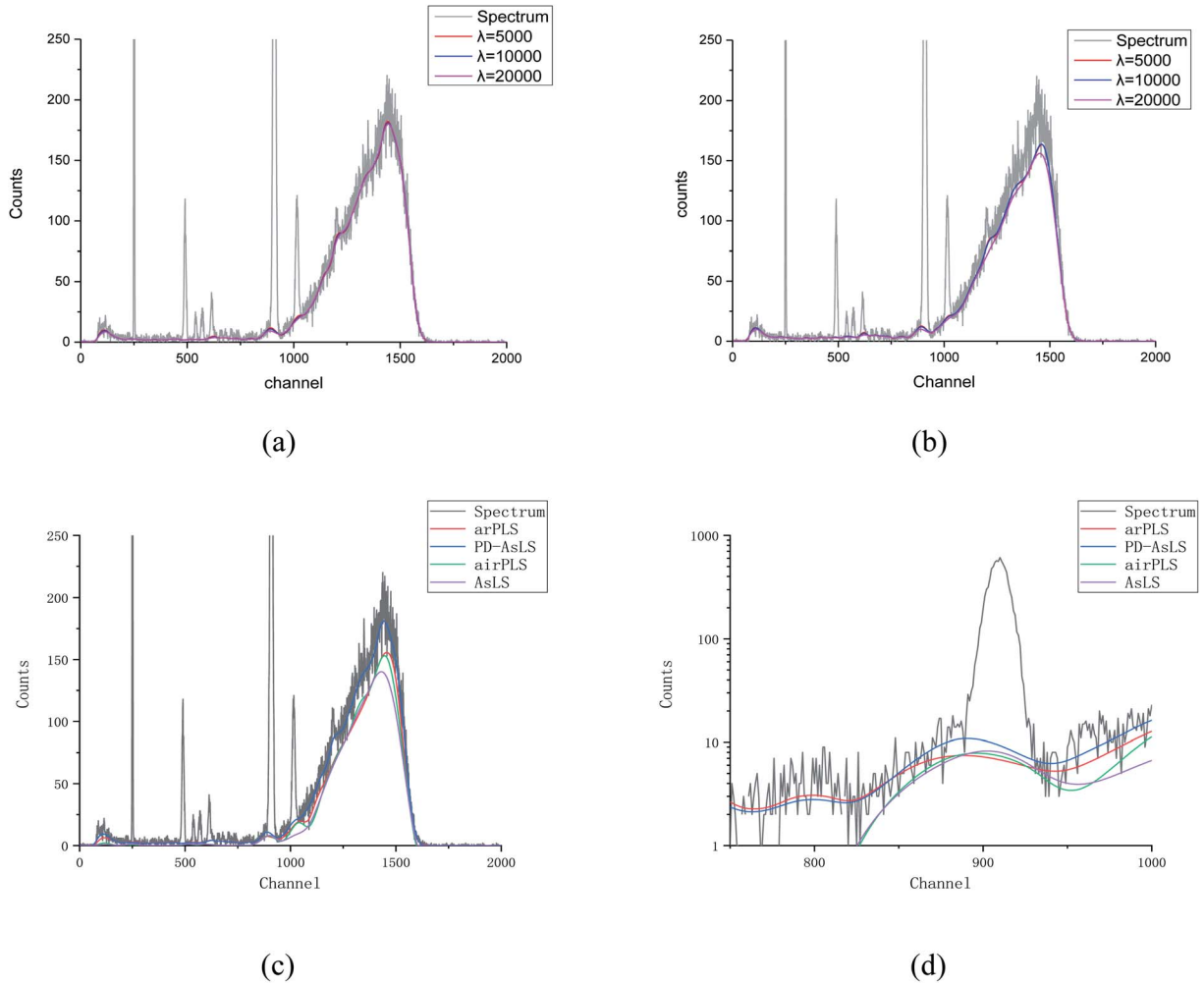

Fig. 5 Actual X-ray baseline spectrum estimated from PD-AsLS (a) and arPLS (b) with different λ; the estimated baseline from arPLS, PD-AsLS airPLS and AsLS when $\lambda = 10\,000$ (c); and details of channels 750 to 1000 (d).

2 mm, in front of the output window. We used a silicon drift detector (SDD), with an FWTH of 152 eV @ 5.8 keV, as the detector. The sample is the national standard sample of soil: GSD8. The measuring time is set to 100 s. PD-AsLS and arPLS are used to estimate the baseline. The results are shown in Fig. 5.

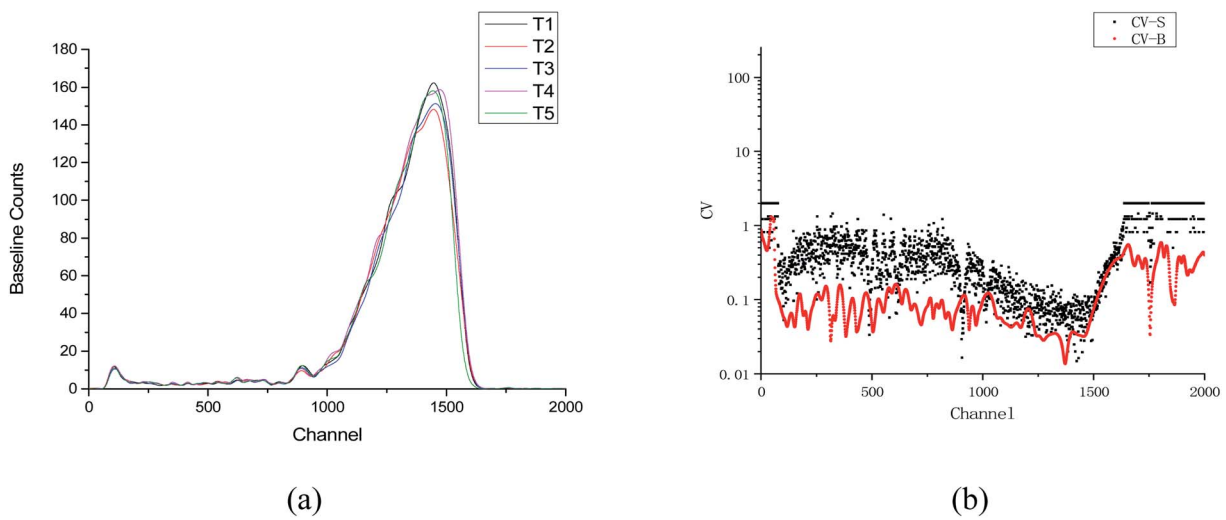

Fig. 6 Baseline of the same sample measured five times (a), and the coefficient of variation (CV) of the real spectrum and the baseline (b).

The results in Fig. 5 are similar to the previous ones. But when $\lambda = 5000$, 10 000, and 20 000, the baseline estimated by arPLS is significantly different. When $\lambda = 5000$, the value is closest to that of the baseline. When $\lambda = 5000$, 10 000, and 20 000, the results obtained from PD-AsLS are close. Obviously, in comparison to the arPLS, PD-AsLS has lower sensitivity to λ.

It can be seen from Fig. 5(c) that the baseline estimated by PD-AsLS (the blue line) fits the tendency of the actual spectrum better, and the baselines estimated by other three methods (arPLS, airPLS and AsLS) are all lower owing to underestimation. From channel 750 to channel 1000, baselines estimated by all four methods have a significant upturn, which indicates that they all have a risk of overestimating the baseline in the peak regions.

A sample is measured five times, then, PD-AsLS is adopted to estimate the energy spectrum. The result is recorded in Fig. 6(a), and used to evaluate the stability of PD-AsLS. The coefficient of variation (CV) is used to evaluate the stability of the data. In the graph of CV, the CV of the baseline (CV–*B*) curve is smaller than the CV of the spectrum (CV–*S*) curve. This result indicates that this baseline estimation method (PD-AsLS) can suppress statistical fluctuations effectively. The CV from channel 1200 to channel 1600 is large, possibly owing to the stability of the instrument. These data are approximately the same as the results observed in Fig. 6(a).

High-energy X-ray radiation is used to measure the content of tungsten (W) in the reference material; the high voltage of the X-ray tube is 150 kV and the current is 0.4 mA. The alloy reference material is excited using a collimated X-ray beam and the energy spectrum of the reference material is measured using the CdTe detector. The scattering background is estimated by PD-AsLS and arPLS, with the λ set to 10 000 and 15 000. As shown in Fig. 7(a), in region A, for different λ values, the estimation results for PD-AsLS are similar, indicating that λ has little impact on this method. However, the results for arPLS exhibit obvious differences and run the risk of overestimation. Region B is the characteristic peak of W with the Kα characteristic X-ray energy as 59.3 keV. The gross area is the accumulating counts in region B, and the net area is the accumulating counts after removing the scattering background estimated by PD-AsLS. In region B, the background estimation results of PD-AsLS and arPLS are similar. The gross area is

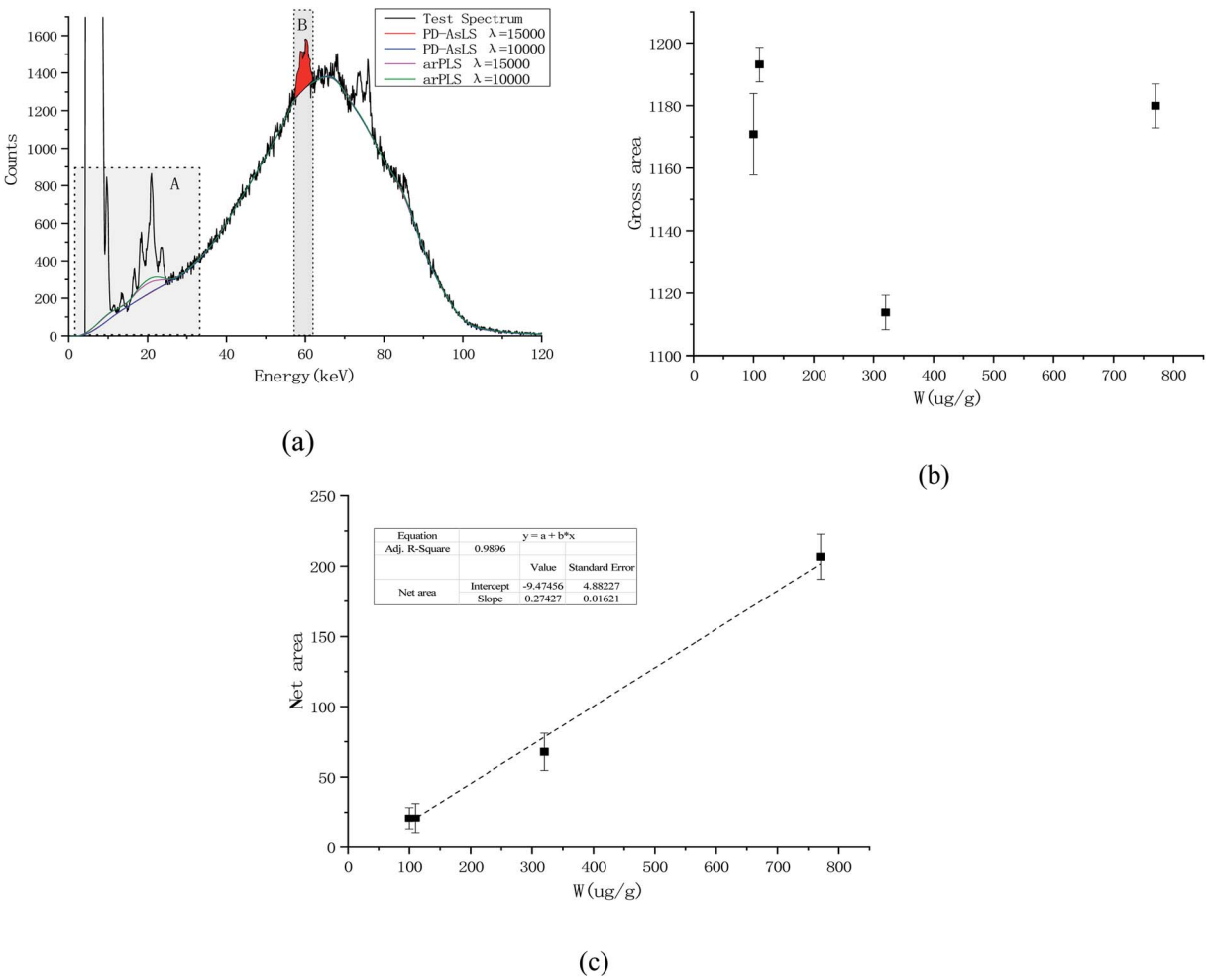

Fig. 7 Test and processing spectra of the standard sample (a), the relationship between gross area and the content of W (b), and the fitting result of the net peak area and the content of W (c).

calculated and shown in Fig. 7(b). As it shows, the counts of the gross area do not have a clear linear relationship with the content of W. The red area is the net area calculated by PD-AsLS. According to the net area, the relationship between the net area and the content of W can be found. As shown in the fitting results of Fig. 7(c), the net area and the content of W have a linear relationship, and the R-square is 0.9896, indicating that PD-AsLS can effectively suppress the effect of the scattering background.

4. Results

In this paper, we have modified and improved the AsLS method to the PD-AsLS method, based on the fact that the probability distribution of the counts of the X-ray energy spectrum obeys the Poisson distribution. In the simulation, RMSE and COR are used to compare the efficiency of PD-AsLS and three other classical baseline correction methods (arPLS, airPLS and AsLS), and the result shows that the value of RMSE estimated by PD-AsLS is less than that estimated by the other three methods. According to the experiment results, PD-AsLS can estimate the baseline stably and suppress the scattering background better than arPLS. Results show that PD-AsLS is an effective method for baseline estimation, and the value of λ has a smaller impact on PD-AsLS than arPLS, airPLS and AsLS.

Conflicts of 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

This work was supported by the National Key Research and Development Program of China (grant no. 2017YFC0602100) and National Natural Science Foundation of China (grant no. 41774190).

References

1 L. Ge, *In-situ X radiation sampling technique—applications in mineral resource prospecting,ore mining and mineral processing control*, Sichuan Science and Technology Press, 1997, vol. 224, pp. 131–149.
2 F. Gan, G. Ruan and J. Mo, Chemometrics and Intelligent Laboratory Systems, *Chemom. Intell. Lab. Syst.*, 2006, **82**(1–2), 59–65, DOI: 10.1016/j.chemolab.2005.08.009.
3 M. Morháč, J. Kliman, V. Matoušek, M. Veselský and I. Turzo, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, *Nucl. Instrum. Methods Phys. Res.*, 1997, **401**(1), 113–132, DOI: 10.1016/S0168-9002(97)01023-1.
4 M. Morháč and V. Matoušek, Peak clipping algorithms for background estimation in spectroscopic data, *Appl. Spectrosc.*, 2008, **62**, 92–106, DOI: 10.1366/000370208783412762.
5 M. Morháč and V. Matoušek, *Proceedings of Science*, 2007, **50**.
6 P. Junfeng, *Data processing of gamma-ray spectra*, Shanxi Science and Technology Publishing Company, Shanxi, 1990.
7 M. H. Zhu, L. G. Liu, A. A. Xu and T. Ma, Gamma-ray spectra measured by nai detector, *Chin. Phys. Lett.*, 2008, **25**, 3942–3945, DOI: 10.1088/0256-307X/25/11/029.
8 B. D. Prakash and Y. C. Wei, A fully automated iterative moving averaging (AIMA) technique for baseline correction, *Analyst*, 2011, **136**, 3130–3135, DOI: 10.1039/c0an00778a.
9 Q. Zhang, L. Ge, Y. Gu, Y. Lin, G. Zeng and J. Yang, Background estimation based on Fourier Transform in the energy-dispersive X-ray fluorescence analysis, *X-Ray Spectrom.*, 2012, **41**, 75–79, DOI: 10.1002/xrs.2360.
10 H. G. Schulze, R. B. Foist, K. Okuda, A. Ivanov and R. F. B. Turnera, A small-window moving average-based fully automated baseline estimation method for raman spectra, *Appl. Spectrosc.*, 2012, **66**(7), 757–764, DOI: 10.1366/11-06550.
11 P. H. C. Eilers, A perfect smoother, *Anal. Chem.*, 2003, **75**(14), 3631–3636, DOI: 10.1021/ac034173t.
12 P. H. C. Eilers and H. F. M. Boelens, Baseline Correction with Asymmetric Least Squares Smoothing, *Life Sci.*, 2005, 1–26.
13 Z. Zhang, S. Chen, Y. Liang, Z. Liu, Q. Zhang, L. Ding, F. Ye and H. Zhou, An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy, *J. Raman Spectrosc.*, 2009, **41**(6), 659–669, DOI: 10.1002/jrs.2500.
14 S.-J. Baek, A. Park, Y.-J. Ahn and J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, *Analyst*, 2015, **140**(1), 250–257, DOI: 10.1039/c4an01061b.
15 L. Jia, Y. Gu, Q. Zhang, J. Zhang, X. Yan, Y. Zhang, Y. Wang and L. Ge, The accuracy evaluation method of baseline estimation algorithms in energy dispersive X-ray fluorescence spectrum analysis, *X-Ray Spectrom.*, 2020, DOI: 10.1002/xrs.3180.