Design of Infrared Spectrum Information Processing Algorithm for Fourier Infrared Spectrometer

Tuo Rui[1], Ren Wanjie[1], Hu Guoxing[1(✉)], Cai Chen[1], Lin Shuai[1], and Zhao Huan[2]

[1] Shandong Institute of Nonmetallic Materials, Jinan 250031, Shandong, China
sdjnhgx@163.com

[2] Army Armament Department Military Representative Office in Ji'nan, Jinan 250031, Shandong, China

Abstract. Fourier transform infrared spectrometer has a wide range of applications in many fields. In order to ensure the spectral quality of the spectrometer output, it is necessary to perform certain processing on the original spectrum. After having developed the Fourier transform infrared spectrometer, in this paper we design the infrared spectrum information processing algorithm. The basic transformations of the spectrum, such as spectral derivation, spectral normalization, centralization, and normalization, are realized. The methods of wavelet transform and S-G smoothing filtering are used to filter out the noise. By means of multivariate scattering correction method, the baseline shift and offset phenomenon of the infrared spectrum of the sample are corrected. Combining principal component analysis and Mahalanobis distance, a detection method of abnormal samples is proposed. Through the combination of multiple data processing algorithms, the processed spectra can play a better role in subsequent spectral analysis.

Keywords: Fourier transform infrared spectrometer · Spectral information processing · Filtering and denoising · Baseline correction

1 Introduction

Fourier transform infrared spectrometer can carry out qualitative and quantitative analysis of samples, and has been widely used in many fields such as medicine, chemical industry, geologic mining and so on [1, 2]. In the process of infrared spectrum signal acquisition, it may be affected by factors such as the state of the spectrometer, acquisition background, detection conditions, etc., resulting in interference in the measured spectrum [3, 4], such as noise interference. Since the background is collected every time while collecting a sample spectrum, the change of the background causes the spectrum to have a baseline drift phenomenon. Other factors such as abnormal sample interference and light scattering will also reduce the accuracy and stability of the model. Therefore, preprocessing the spectral data is a key step to ensure the output performance of the Fourier transform infrared spectrometer. In this paper, the processing of spectral

S.-H. Wang and Y.-D. Zhang (Eds.): ICMTEL 2022, LNICST 446, pp. 711–723, 2022.
https://doi.org/10.1007/978-3-031-18123-8_56

information mainly includes the basic transformation of the spectrum, the filtering of redundant noise interference and other irregular influencing factors, such as baseline drift caused by background interference during acquisition, noise interference of instruments and detection environments, and abnormal detection in spectra.

In order to remove the redundant noise interference of the spectrum, McClure et al. made a detailed study on the influence of the random noise of the spectrum on the model [5]. They confirmed that the random noise superimposed on the spectral signal will deteriorate the accuracy of the model. For the influence of baseline drift, baseline shift and uneven particle distribution on the spectrum, the commonly used solutions are the first derivative, the second derivative and the multivariate scattering correction [6]. The existence of abnormal sample data will affect the predictive ability of the model and cause deviations in the prediction. Commonly used methods for identifying abnormal samples include Mahalanobis distance method and principal component analysis, and the combination of partial least squares principal component score and Mahalanobis distance. The Mahalanobis distance method was used to identify the abnormal value of the leaf spectrum of Junzao [7]. During the sample spectrum collection process, the collected spectrum inevitably has interference due to the instrument, the sample itself or other reasons. Using the original spectrum directly will lead to poor model accuracy and instability. The sample spectrum information can be processed according to the research experience and the characteristics of the sample. This research will carry out the design of infrared spectrum information processing algorithm for our developed Fourier transform infrared spectrometer. The methods of spectral preprocessing include derivation, standard normal transformation, smoothing and filtering, multivariate scattering correction, etc. In actual processing, various methods will be combined in a certain order according to specific conditions.

2 Basic Transformation of Infrared Spectrum

2.1 Spectral Derivation

Spectral derivation is one of the commonly used preprocessing methods in infrared spectroscopy, which can eliminate baseline drift and improve spectral resolution. The direct difference method is used for the derivation of the spectrum. As a discrete spectrum derivation method, for the discrete spectrum x_i, $i = 1,\ldots, n$, the first order derivative and second derivative spectra at the wavelength i and the difference width g are calculated according to the following methods. The formula for the first derivative is

$$x_{i,1st} = \frac{x_i - x_{i+g}}{g} \tag{1}$$

And the second order derivative formula is

$$x_{i,2nd} = \frac{x_i + x_{i+2g} - 2x_{i+g}}{g^2} \tag{2}$$

Taking $Mn_3Al_2(SiO_4)_3$ material as an example, Fig. 1 shows the original spectrum and its first derivative, and Fig. 2 shows the original spectrum and its second derivative.

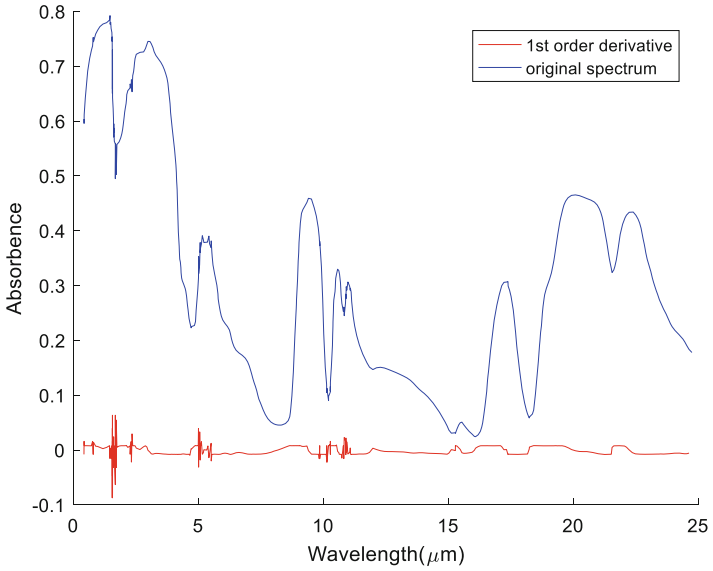

Fig. 1. Infrared spectrum of $Mn_3Al_2(SiO_4)_3$ and its first order derivative.

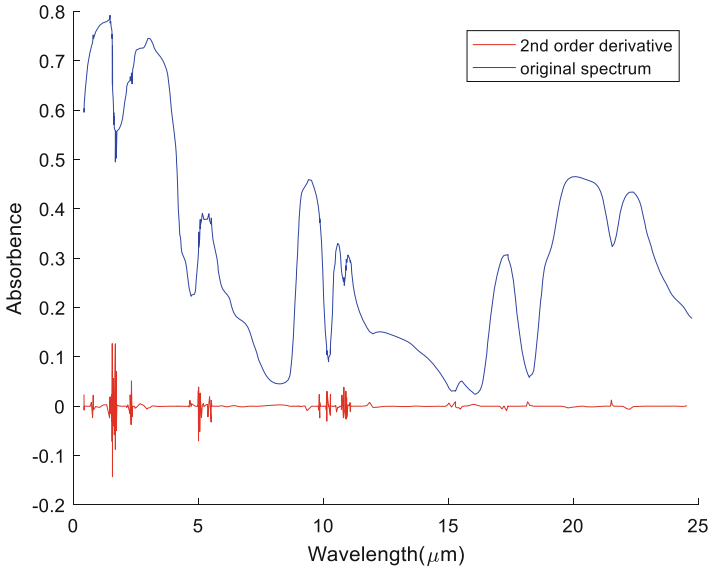

Fig. 2. Infrared spectrum of $Mn_3Al_2(SiO_4)_3$ and its second order derivative.

It can be seen that for the spectrum with high resolution and many wavelength sampling points, the derivative spectrum obtained by the direct difference method can meet the requirements. However, for the spectrum of sparse wavelength sampling points,

the derivative obtained by this method will have a large error. In this case, the Savitzky-Golay convolution derivation method can be used for calculation.

2.2 Normalization

When using infrared spectroscopy, it is necessary to correlate the characteristics of the spectrum with the properties or structural characteristics of the sample to be tested. Therefore, it is often necessary to use data enhancement algorithms to reduce or eliminate some redundant information. Commonly used algorithms include centralization, standardization and normalization. The main function of normalization is to normalize the ordinate of the spectrum, which is convenient for quantitative analysis of infrared spectrum. For absorbance spectra, the absorbance of the maximum absorption peak after normalization was normalized to 1 and the baseline was normalized to 0.

The specific calculation formula is shown in formula (3), where x is the absorbance corresponding to a certain wavelength, $x_{\min}$ is the minimum value of absorbance among all absorbance values in the spectrum, $x_{\max}$ is the maximum value of absorbance among all absorbance values in the spectrum, and x^* is the normalized absorbance value after processing, being between [0,1].

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

Figure 3 shows the result of normalizing the original infrared spectrum of $Mn_3Al_2(SiO_4)_3$ material.

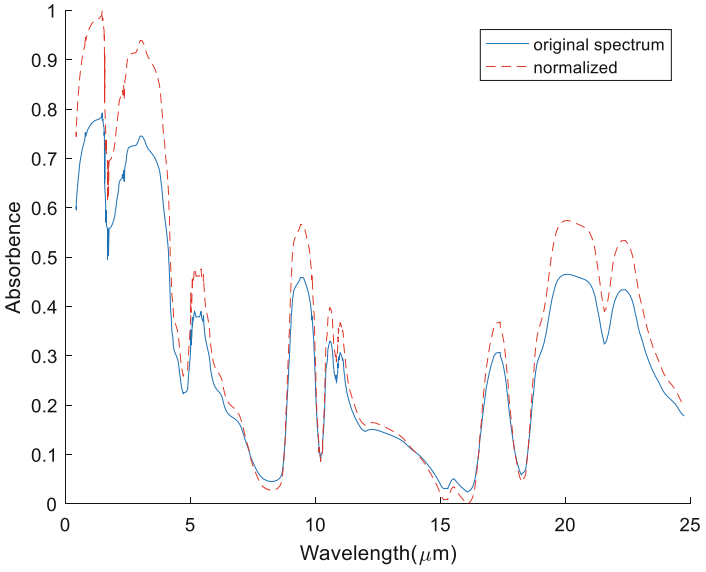

Fig. 3. Infrared spectrum and its normalization.

2.3 Centralization

Centralization, also known as zero-average, is the basic transformation of the infrared spectrum. It mainly completes the translation of the spectrum and moves it to the position with 0 as the center point. By subtracting the average value of all the data from each spectral data, the average value of the spectral data after centering is 0, and the variance is not limited. The centralization enables all spectral data to be distributed on both sides of the zero point, fully reflecting the change information, and effectively removing the impact of changes caused by objective factors such as temperature or human operation on the spectral data. The specific calculation formula is shown in formula (4), where x is the ordinate value corresponding to a certain wavelength, and μ is the average value of the ordinate corresponding to all wavelengths of the spectrum.

$$x^* = x - \mu \tag{4}$$

Still taking the $Mn_3Al_2(SiO_4)_3$ material in Fig. 1 as an example, after centering the original infrared spectrum, the spectrum is shown in Fig. 4.

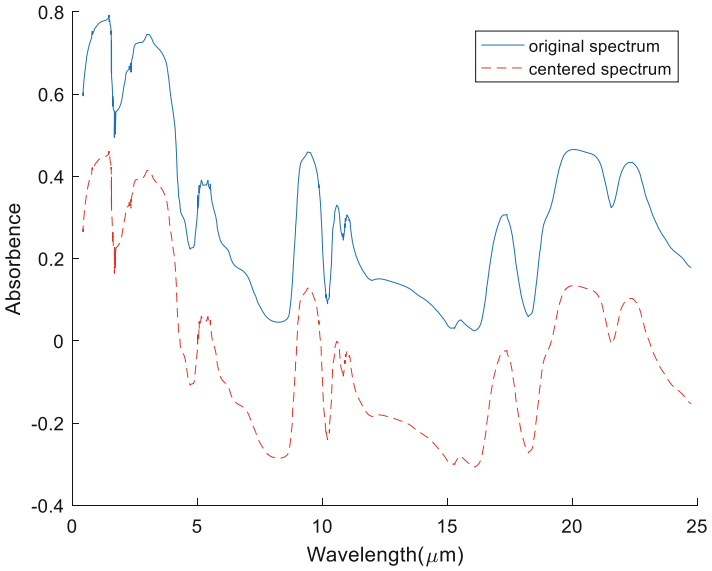

Fig. 4. Infrared spectrum and its centralization.

2.4 Standardization

As one of the basic transformations of infrared spectroscopy, standardization maps the data to a standard normal distribution with a mean of 0 and a standard deviation of 1. On the basis of data centralization, the data is divided by the standard deviation of all spectral data to make it satisfying the standard normal distribution.

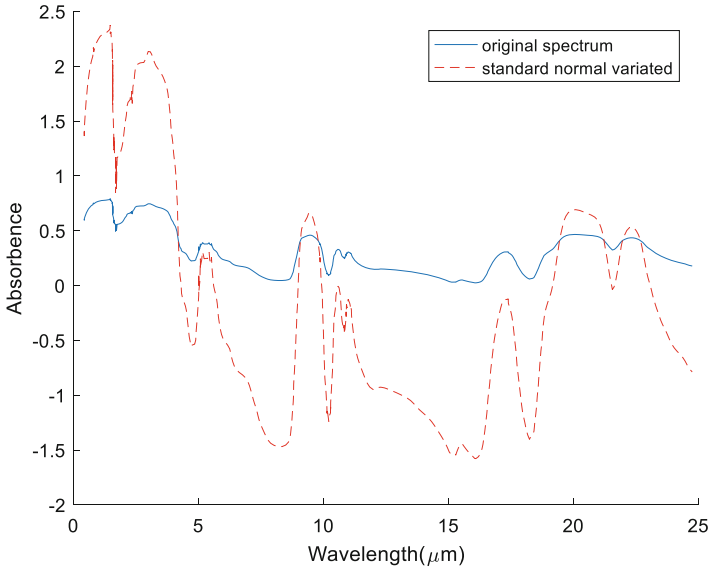

Fig. 5. Infrared spectrum and its normalization.

$$x^* = \frac{x - \mu}{\sigma} \tag{5}$$

The specific calculation of standardization is shown in formula (5), where x is the original spectral data, μ is the average value of all spectral data, σ is the standard deviation of all spectral data, x^* is the value after normalization, which obeys the standard normal distribution $x^* \sim N(0, 1)$. For the $Mn_3Al_2(SiO_4)_3$ material in Fig. 1, the normalized spectrum is shown in Fig. 5.

3 Filtering and Denoising of Infrared Spectrum

The output of infrared spectrometer not only contains useful information, but also superimposes random errors, such as noise. The methods of noise filtering include Kalman filter, wavelet analysis, wavelet packet transform, smooth noise filtering and so on. In this paper, wavelet packet transform and S-G convolution smoothing are used to realize the filtering and denoising of infrared spectrum.

3.1 Wavelet Packet Filter Denoising

Wavelet Packet Transform (WPT) has higher accuracy and flexibility in signal analysis than wavelet transform, and has finer local analysis capabilities. Wavelet transform is mainly used for signal noise filtering, data compression and model transfer, while wavelet packet analysis is mainly used for signal noise removal [8]. As shown in Fig. 6,

the wavelet packet transform can not only decompose the low-frequency part of the signal, but also the high-frequency part. This decomposition has neither redundancy nor omissions, so it contains a lot of medium and high frequency information. Signals are able to perform better time-frequency localized analysis.

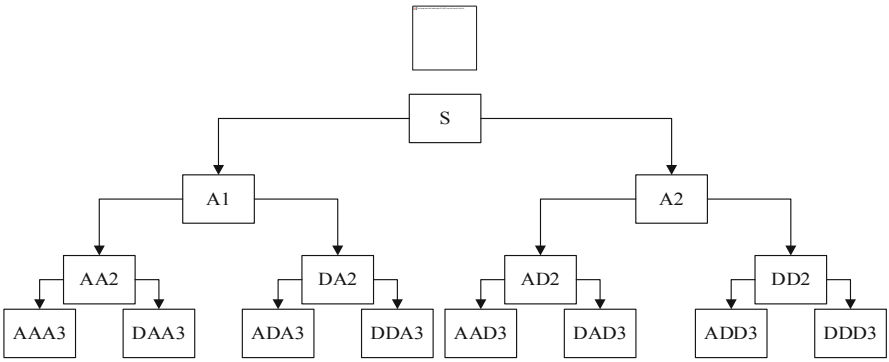

Fig. 6. Three-layer wavelet packet decomposition tree.

Wavelet packet transform can decompose spectral information into background information, component information and noise. The basic steps of wavelet packet threshold denoising are as follows [9]:

(1) Wavelet packet decomposition of the signal. Select the appropriate wavelet function and decomposition scale according to the signal, and calculate the optimal wavelet basis given the entropy criterion;
(2) Threshold quantization of wavelet packet decomposition coefficients. Choose the appropriate threshold rule according to experience, select the appropriate threshold, and process the decomposed wavelet packet coefficients;
(3) Wavelet packet reconstruction. The original spectral signal is reconstructed from the wavelet packet decomposition coefficients of the N-th layer and the processed coefficients.

For wavelet packet transformation, the selection of wavelet basis is very critical. Generally, a suitable wavelet basis function is selected from the four aspects of compactness, regularity, vanishing moment and symmetry. The commonly used wavelet basis functions are Daubechies wavelet system, SymletsA function system, Meyer wavelet, Coiflet wavelet system, Biorthogonal wavelet, and the commonly used wavelet functions in spectral denoising are db2, db4, sym6, boir2.4 [10]. After experimental comparison, this project selects a db4 wavelet for spectral denoising.

Once the signal undergoes wavelet packet transformation, the information is distributed in each frequency band. The effective spectral signal is usually concentrated in the low frequency band. On the larger wavelet packet coefficient, the noise energy is generally distributed on the entire coefficient axis, so it can be considered that the signal is generally concentrated in the amplitude value. The larger wavelet packet coefficients

and the noise are distributed on the smaller amplitude wavelet packet coefficients, so the threshold method can be used to extract useful signals. Because the threshold selection is too large, the details of the useful signal will be filtered out, and the threshold selection is too small, the denoising effect is not ideal, so it is necessary to select an appropriate threshold to quantize the wavelet packet decomposition coefficients. In this paper, the Sqtwolog length logarithm criterion is selected to set the threshold, and the specific calculation formula is as follows.

$$H = \sqrt{2 * \log(L(s))} \tag{6}$$

where H is the set threshold, and $L(s)$ is the length of the signal. Continuing the example of Fig. 1, Fig. 7 shows the actual effect of wavelet packet denoising of infrared spectra.

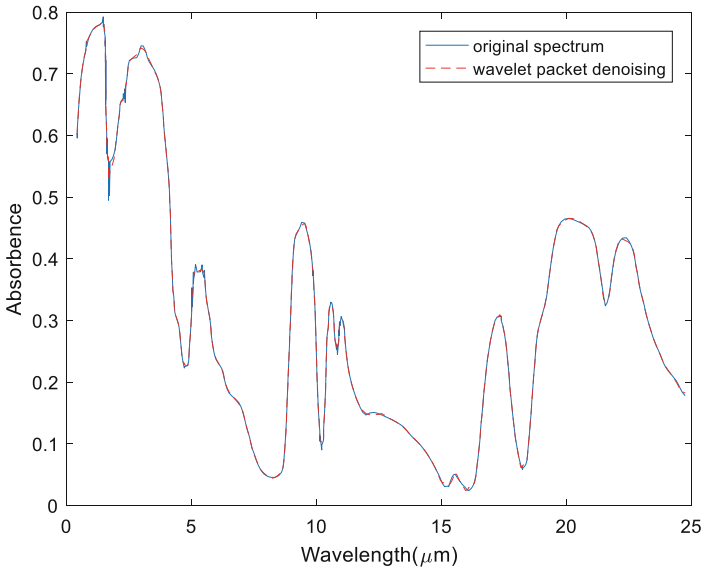

Fig. 7. Comparison before and after wavelet packet denoising of infrared spectrum.

3.2 Savitzky-Golay Convolution Smoothing

Savitzky-Golay convolution smoothing method, also called S-G smoothing, uses polynomials for data smoothing. S-G smoothing uses polynomials to perform polynomial least squares fitting on the data in the moving window, and its essence is a weighted average method. S-G smoothing can retain useful information in spectral signals, eliminate random noise, and make the curve smoother. It is a widely used denoising method at present [11].

A subset of the original spectral data is selected as the window instead of the entire spectrum. The width of the smoothing window is set to $2m + 1$, that is, the window

width $n = 2m + 1$. Assuming that the original data points within the window can be fitted with a k-1 polynomial, i.e.

$$y_i = a_0 + a_1 i + a_2 i^2 + \cdots + a_{k-1} i^{k-1} \tag{7}$$

where $i = (-m, -m + 1, \cdots, 0, 1, \cdots, m-1, m)$. Therefore, the above-mentioned polynomial can be obtained for each of the n original data points in the window, and n such polynomials constitute a k -element linear equation system, and the k fitting parameters a_j need to be solved. Generally, the selected filter window width n should be greater than or at least equal to k. When $n = k$, the fitting parameters can be solved by linear algebra; if $n > k$, the least squares method can be used to solve.

Juxtaposing the above polynomials, the following matrix operations can be obtained:

$$\begin{pmatrix} y_{-m} \\ y_{-m-1} \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & -m & \cdots & (-m)^{k-1} \\ 1 & -m+1 & \cdots & (-m+1)^{k-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & m & \cdots & (m)^{k-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{k-1} \end{pmatrix} + \begin{pmatrix} e_{-m} \\ e_{-m-1} \\ \vdots \\ e_m \end{pmatrix} \tag{8}$$

It can be simplified to the following system of overdetermined equations

$$Y_{(2m+1)\times 1} = X_{(2m+1)\times k} \cdot A_{k\times 1} + E_{(2m+1)\times 1} \tag{9}$$

The calculation formula of the final solution of the filter value Y is as follows

$$\hat{Y} = XA = X(X^T X)^{-1} X^T Y = BY \tag{10}$$

Among them $B = X(X^T X)^{-1} X^T$ is the filter coefficient matrix, which is determined by and only by the X matrix, and the B matrix is a $(2m + 1) \times (2m + 1)$ matrix. According to the coefficient matrix, the S-G smooth fitting equation can be obtained.

Selecting the window width as 5 and the order of the fitting polynomial as 2, the result of S-G smoothing filtering on the infrared spectrum of $Mn_3Al_2(SiO_4)_3$ is shown in Fig. 8.

4 Multivariate Scattering Correction

Due to the influence of instrument background, sample particle size and other factors, baseline drift often occur in infrared analysis, and baseline correction can effectively eliminate these effects [12]. Methods such as peak-valley point leveling, offset deduction, differential processing, and baseline tilt can be used, and the most commonly used method is multivariate scattering correction.

Multiple Scattering Correction (MSC) was proposed by Martens et al. It is a commonly used method in spectral data preprocessing, mainly used to correct the shift and offset of the infrared spectral baseline of the sample. The resulting scattering effects improve the signal-to-noise ratio of the original absorbance spectrum. The method is based on the spectral array of a set of samples, and the basic idea is to effectively separate the absorption information of chemical substances from the scattered light signal

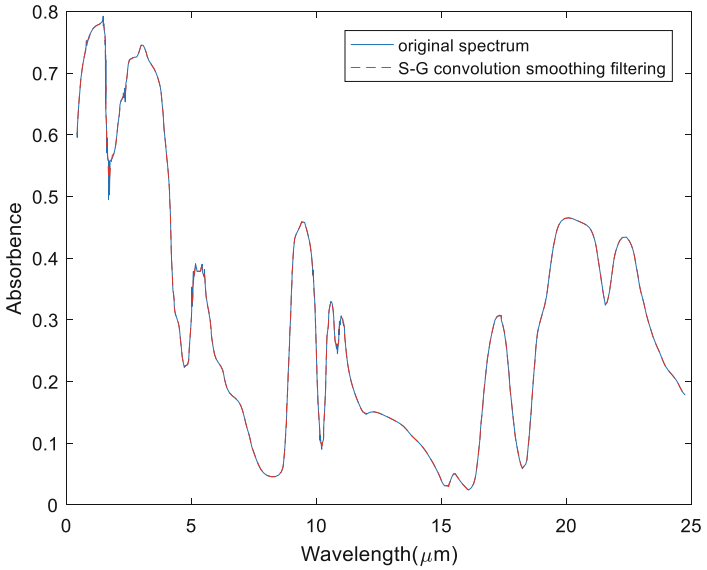

Fig. 8. SG smoothing filter results.

in the spectrum, assuming that the scattering coefficient is the same at all wavelengths. The specific steps of the multivariate scattering correction algorithm to process infrared spectral data are as follows:

Find the average spectrum of all sample spectra. The formula for calculating the average spectrum is:

$$\overline{S_{i,j}} = \frac{\sum_{i=1}^{n} S_{i,j}}{n} \tag{11}$$

where S is the $n \times p$ dimensional spectral matrix, that is, there are n groups of sample spectra, and each spectrum contains p wavelength data.

The spectrum of each sample was subjected to a linear regression operation with the average spectrum, and the average spectrum was regarded as the standard spectrum of the entire spectrum matrix. The regression coefficients and regression constants b_i of each spectrum relative to the standard spectrum were obtained through a linear regression operation m_i. The formula for calculating the univariate linear regression is:

$$S_i = m_i \overline{S} + b_i \tag{12}$$

Among them S_i is the spectral data of the i-th sample, and $\overline{S}$ is the average spectrum calculated in step one, the linear offset m_i and tilt translation b_i are obtained after performing a single linear regression operation.

The original spectrum of each sample is based on the average spectrum, and the regression constant and regression coefficient are used to correct the drift and offset of the spectrum. The basic method is to make the difference between the original spectrum

of the sample and its tilt shift, and divide its linear offset at the same time, so as to correct the baseline shift and shift of the spectrum. The spectral absorption information is not affected, so the signal-to-noise ratio of the spectrum is improved. The calculation method is as follows:

$$S_{i(MSC)} = \frac{(S_i - b_i)}{m_i} \tag{13}$$

During the spectrum acquisition process, the phenomenon of baseline drift will occur due to changes in the acquisition environment. Still taking $Mn_3Al_2(SiO_4)_3$ material as an example, the overall spectra of the spectrum repeated fifteen times are shown in Fig. 9, and it can be seen that the spectra has a baseline drift phenomenon. The spectra is processed by the method of multivariate scattering correction, and the result is shown in Fig. 10, which shows the effectively elimination of the baseline drift.

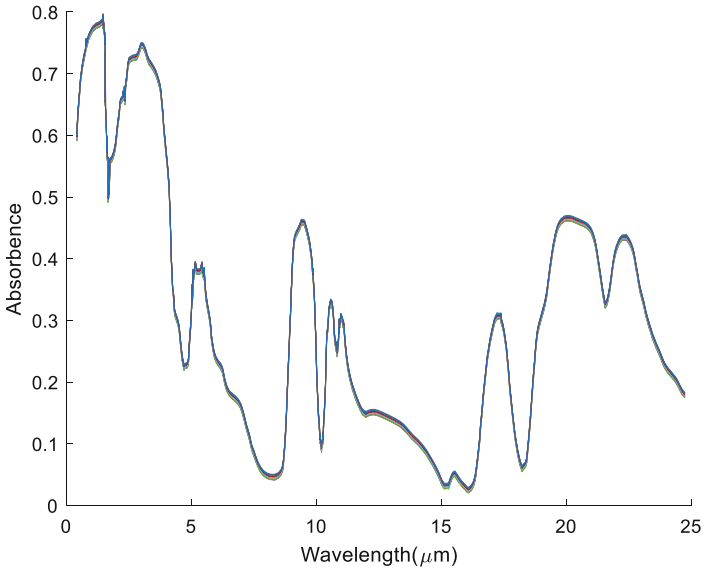

Fig. 9. Original multispectral images.

5 Abnormal Sample Removal

Abnormal samples will have a great impact on the infrared spectral model. The most widely used method is the Mahalanobis distance method [13]. In view of the characteristics of multi-variable spectral wavelength and easy overfitting, this paper combines principal component analysis and Mahalanobis distance to detect abnormal samples.

The principal components of each sample are first calculated, and the principal components of the sample replace the spectral data of the sample. The formula for

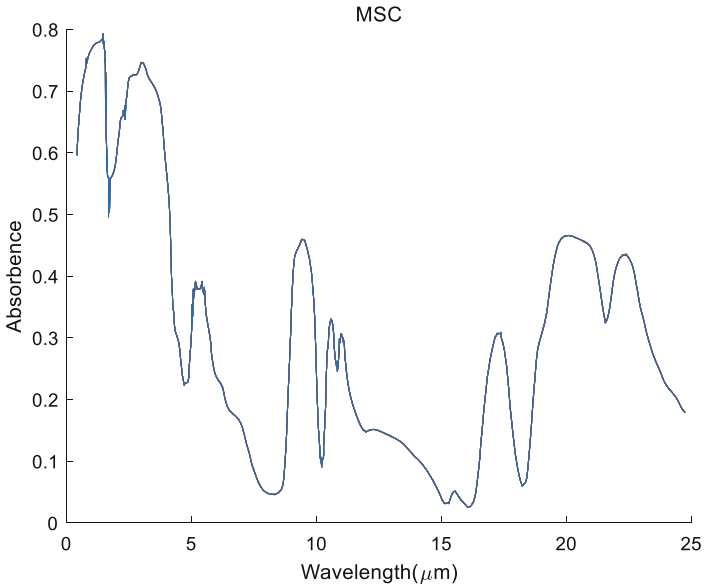

Fig. 10. Spectra after multivariate scattering correction.

calculating Mahalanobis distance is as follows:

$$D^2(\mathrm{i}) = \left(S_i - \overline{S}\right)\mathbf{C}_S^{-1}\left(S_i - \overline{S}\right)^T \tag{14}$$

where S is the sample data, $\mathbf{C}_S$ is the covariance of the sample data, and $\overline{S}$ is the mean of the data. The threshold is set to distinguish abnormal samples. The threshold setting formula is as follows:

$$D_{\mathrm{t}} = \overline{D} + e\delta_D \tag{15}$$

in which $\overline{D}$ is the mean of the Mahalanobis distance, δ_D is the standard deviation of the Mahalanobis distance, and e is a weight coefficient used to adjust the threshold for judging outliers. The larger the D_t value, the larger the value $D_i - D_t$, and the smaller the value, the less likely an anomaly is. When the average spectrum of a sample is very close to the sample, it is called the neighbor of the average sample.

Using the idea of discriminating spectral outliers combining Mahalanobis distance and principal component analysis, the 15 spectral samples shown in Fig. 9 are analyzed, and the sequence numbers of outliers are 6 and 7, and the abnormal samples are successfully eliminated.

6 Conclusion

Considering our developed Fourier transform infrared spectrometer, the spectral preprocessing algorithm is described in this paper. Firstly, the methods of first-order derivative, second-order derivative, centralization, normalization, and standardization of the

spectrum are discussed, and the spectrum can be transformed according to the detailed requirements. Using wavelet transform, S-G smoothing, principal component analysis-Malanobis distance method, multivariate scattering correction and other methods to process the spectrum, we can filter and denoise the spectrum, analyze the baseline shift and offset phenomenon of the infrared spectrum of the sample, as well as complete the detection of abnormal samples. As the first step of spectral analysis, the design of spectral preprocessing algorithm provides a solid foundation for the qualitative and quantitative analysis of subsequent spectra. Subsequent studies will also be carried out on spectral fitting, peak marking, and spectral feature extraction.

References

1. Dong, X., Guo, L., Wang, F., et al.: Quantitative analysis of microstructure of different coals based on Fourier transform infrared spectroscopy. Chin. Sci. Technol. Pap. **17**(1), 55–61 (2022)
2. Choi, K., Abbott, J., Park, B., Choi, C., et al.: Near-infrared diffuse reflectance for quantitative and qualitative measurement of soluble solids and firmness of delicious and Gala apples. Trans. ASAE **46**(6), 1721–1731 (2003)
3. Peirs, A., Schenk, A., Nicolai, B.: Effect of natural variability among appleson the accuracy of VIS-NIR calibration models for optimal harvest predictions. Postharvest Biol. Technol. **35**, 1–13 (2005)
4. Bessho, H., Kudo, K., Omori, J., et al.: A portable non-destructive quality meter for under standing fruit soluble solids in apple canopies. Acta Hort. **732**, 593–597 (2007)
5. Nicolai, B., Verlinden, B., Desmet, M., et al.: Time-resolved and continuous wave NIR reflectance spectroscopy to predict soluble solids content and firmness of pear. Postharvest Biol. Technol. **47**(1), 68–74 (2008)
6. Xu, G., Yuan, H., Lu, W.: Advances in modern near-infrared spectroscopy and its applications. Spectrosc. Spectral Anal. **02**, 134–142 (2000)
7. Hui, Y.: Study on Spectral Preprocessing Method and Moisture Detection Model of Junzao Leaves. Tarim University (2018)
8. Li, X., Zhang, Y., Liu, Z., et al.: Wavelet analysis and HHT transformation of blasting vibration signal. Explosion & Shock Impact **25**(6), 528–535 (2005)
9. Guo, X., Yang, H.: Wavelet packet denoising based on multi-threshold. In: Proceedings of the 27th China Conference on Control (2008)
10. Xu., X.: Application of enhanced wavelet transform in blasting vibration signal analysis. South China University of Technology (2013)
11. John, A., Sadasivan, J., Seelamantula, C.: Adaptive Savitzky-Golay filtering in non-gaussian noise. IEEE Trans. Signal Process. **69**, 5021–5036 (2021)
12. Zhong, J.: Study on spectral correction method and its application in soil detection. Jinan University (2018)
13. Zhang, L., Wang, W., Gu, Y., et al.: Principal component analysis of near-infrared spectroscopy-Markov distance clustering for authenticity identification of cigarettes. Spectrosc. Spectral Anal. **31**(05), 1254–1257 (2011)