Electronic Supplementary Information (ESI)

Baseline correction for Raman spectra using penalized spline smoothing based on vector transformation

Yaoyi Cai[a,b], Chunhua Yang[a], Degang Xu[a,*], Weihua Gui[a]

[a] College of Information Science and Engineering, Central South University, Changsha 410083
[b] College of Engineering and Design, Hunan Normal University, Changsha, Hunan, 410083, China

*Corresponding author, E-mail address: dgxu@csu.edu.cn

Comparison among airPLS, IAsLS, VTPspline baseline correction methods for simulated Raman spectra. The estimated baselines and performance indices among VTPspline, airPLS and IAsLS are shown in Fig. S1. It is found that the baselines of the original Raman spectra are over-fitted by the airPLS method at 50-150 cm^{-1} (baseline b1) and 500-650 cm^{-1} (baseline b2). Obviously, the intensities of Raman characteristic peaks are weakened in these areas. In addition, the IAsLS method did not fit the baselines well at 480-650 cm^{-1} (baseline b1), 500-650 cm^{-1} (baseline b1) and 700-950 cm^{-1} (baseline b2). By contrast, based on the proposed VTPspline method, the baselines in simulated spectra y1 and y2 are fitted well at the whole range of Raman spectrum.

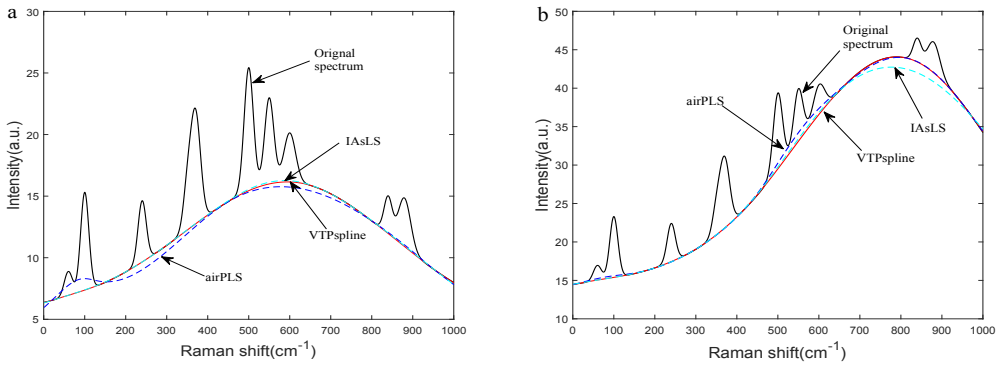


Fig. S1 Comparison among airPLS, IAsLS and VTPspline baseline correction methods with simulated Raman spectra y1(a) and y2(b) respectively.

Evaluating the performance of the proposed VTPspline method. The SNR (Signal to Noise Ratio) of the low noise spectrum was set to 35 dB and that of the high noise spectrum was set to 25 dB, as shown is Fig. S2. Then estimated baselines which are treated by the proposed VTPspline baseline correction method for the simulated Raman spectra with low and high noises are shown in Fig. S3. As shown in Fig. S3, under the condition of noise with different intensities, the suspected background is distributed around the real baseline and the estimated baseline which is smoothed by penalized spline smoothing decrease the int erference of noises. It is obviously that owing to the penalized spline smoothing, the noises in the Raman background region have little effect on the Raman spectral baseline estimation process. The simulated spectra with low and high noise are treated by above three methods and the estimated baselines are shown in Fig. S4. Meanwhile the

RMSEs for the above three methods are displayed in Fig. S5 for easy comparison. Let us see the estimated baselines in detail. In Fig. S3 and S4, the baselines for spectrum y_1 and y_2 in low and high noise are well-estimated by VTPspline. However, there is a certain degree of deviation of spectral baseline fitted by the other two methods (500-650 cm^{-1} in Fig. S4(a), 480-670 cm^{-1} in Fig. S4(b), 800-950 cm^{-1} in Fig. S4(c), 500-700 cm^{-1} in Fig. S4(d) and etc.). The simulation shows that, the RMSE of VTPspline is much less than the other methods for the Raman spectrum in low and high noise, which means that the baselines are more accurately estimated by VTPspline method. It is important to note that the performance of the VTPspline for spectrum y_2, which is composed of multiple overlapped peaks and strong fluorescence background, is far better than the other two methods.

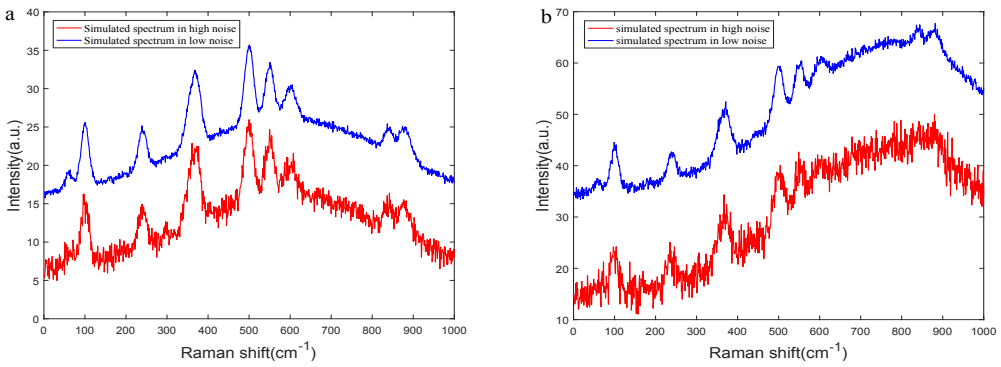


Fig. S2. Simulated spectra y_1(a) and y_2(b) in high and low noise.

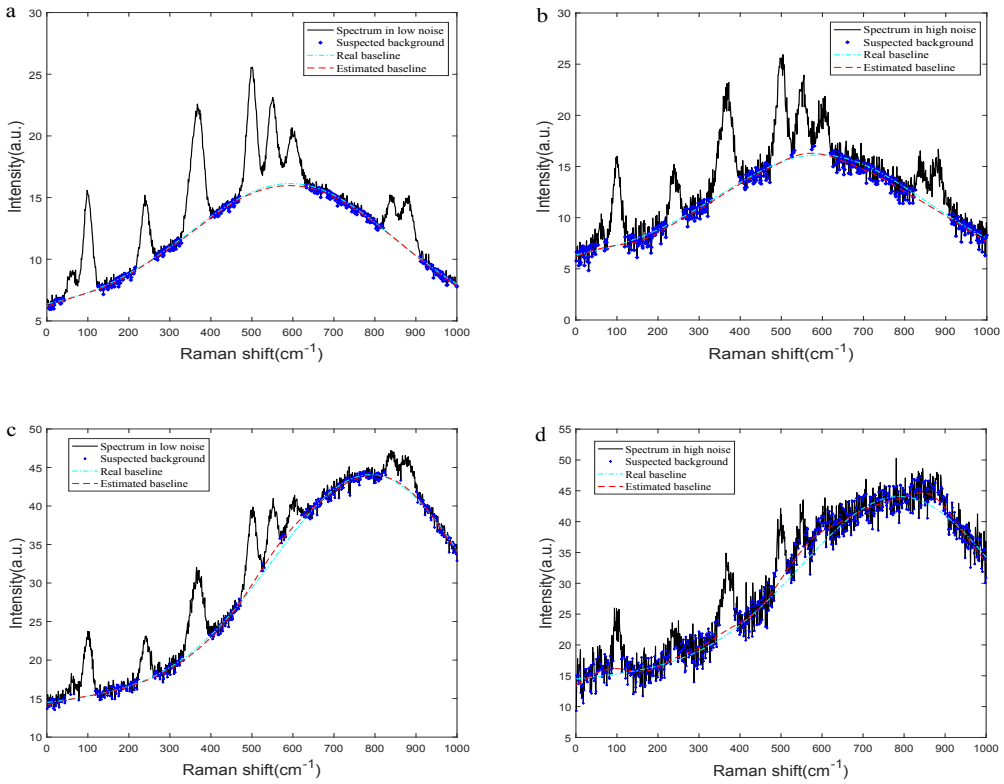


Fig. S3. The optimal estimated baseline for simulated Raman spectra with low and high noise (a) y1 in low noise; (b) y1 in high noise; (c) y2 in low noise; (d) y2 in high noise.

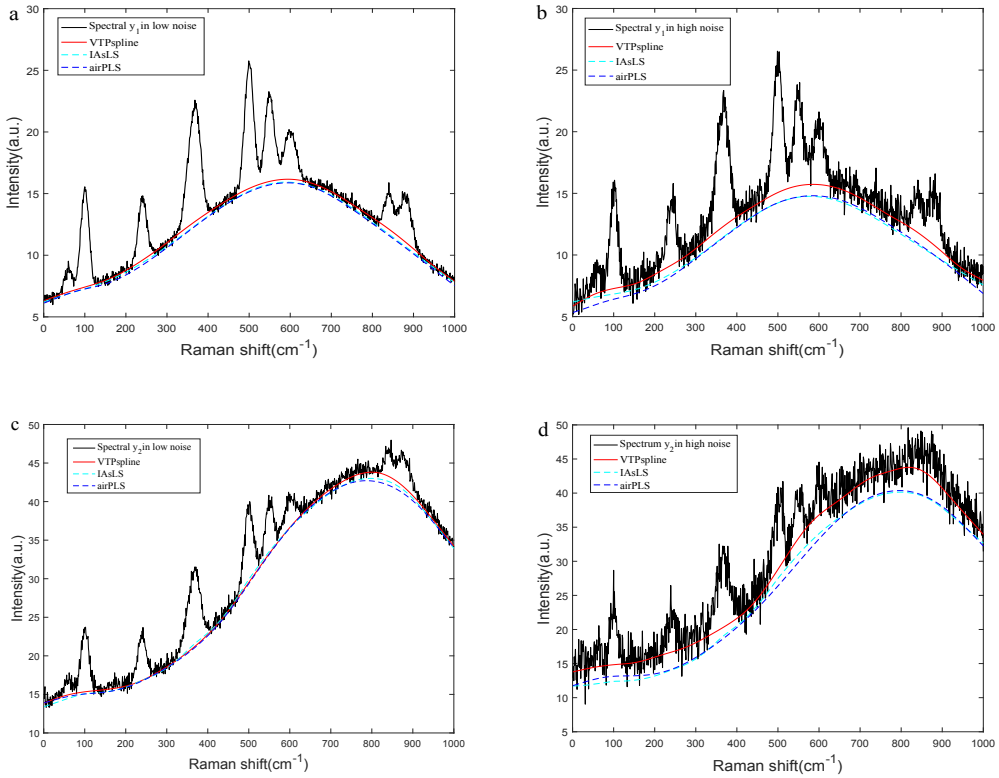


Fig. S4. The estimated baselines based on IAsLS, airPLS and VTPspline methods: (a) y_1 with low noise; (b) y_1 with high noise; (c)y_2 with low noise; (d)y_2 with high noise.

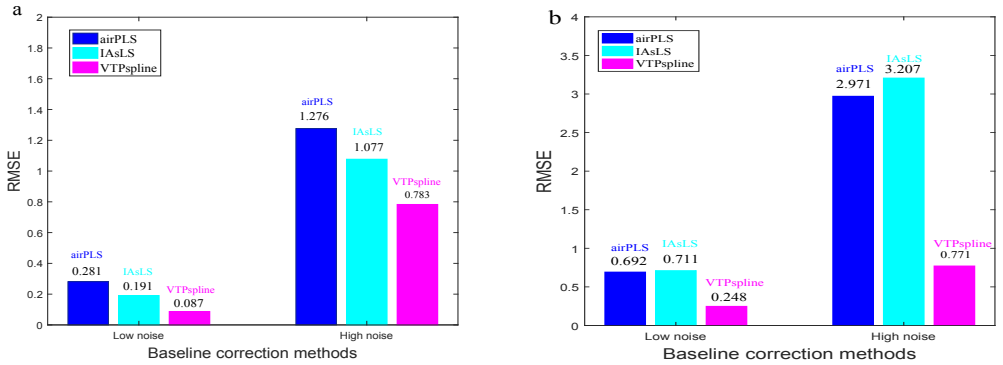


Fig. S5. RMSEs of corrected simulated spectra in low and high noise with simulated spectra y_1 (a) and y_2 (b) respectively.

Choosing the suitable smoothing parameter and number of equidistant knots. The baselines were estimated by VTPspline method with increasing parameters (λ and *k*) and the RMSEs of the proposed method were calculated. Fig. S6 shows the performance of VTPspline while varying the parameter λ and maintaining the value of parameter *k* at 100. It is found that, when $\lambda \leq 2$, the performance of the VTPspline method for spectrum y_1 and y_2 is poor. When $\lambda > 12$, the performance of the proposed method for both two spectra become worse. Meanwhile, it can also be found that from Fig. S7: The larger the curvature of the real baseline, the narrower the range which a value can be taken for the smoothing parameter λ. Fig. S7 shows the estimated and real baselines for spectrum y_1 and y_2 with the value of λ being set to 5. To obtain the optimal performance, the smoothing

parameter λ should be set within the range of 2.5-8. Fig. S8 and Fig. S9 investigated the performance of VTPspline for different numbers of equidistant knots *k* where the optimal smoothing parameter λ has been set as 5. The simulations show that if the number of knots of B-spline curves for the VTPspline method is limited (*k*=16 for spectrum y_1 and y_2), the fitted baselines will not adapt to the dramatic change of fluorescence background in the Raman spectrum. However, too many knots which used to obtain the smooth baseline will increase the amount of calculation and take a relatively long time. To balance the fitting time of baseline and its requirement in precision, the number of the B-spline knots can be set as one twentieth of the total wavenumbers.

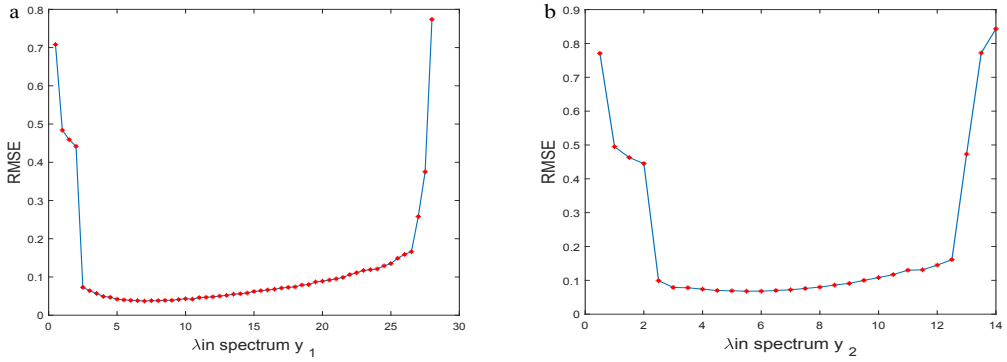


Fig. S6. The RMSEs of the VTPspline method for simulated spectrum y_1 and y_2 when varying the smoothing parameter λ

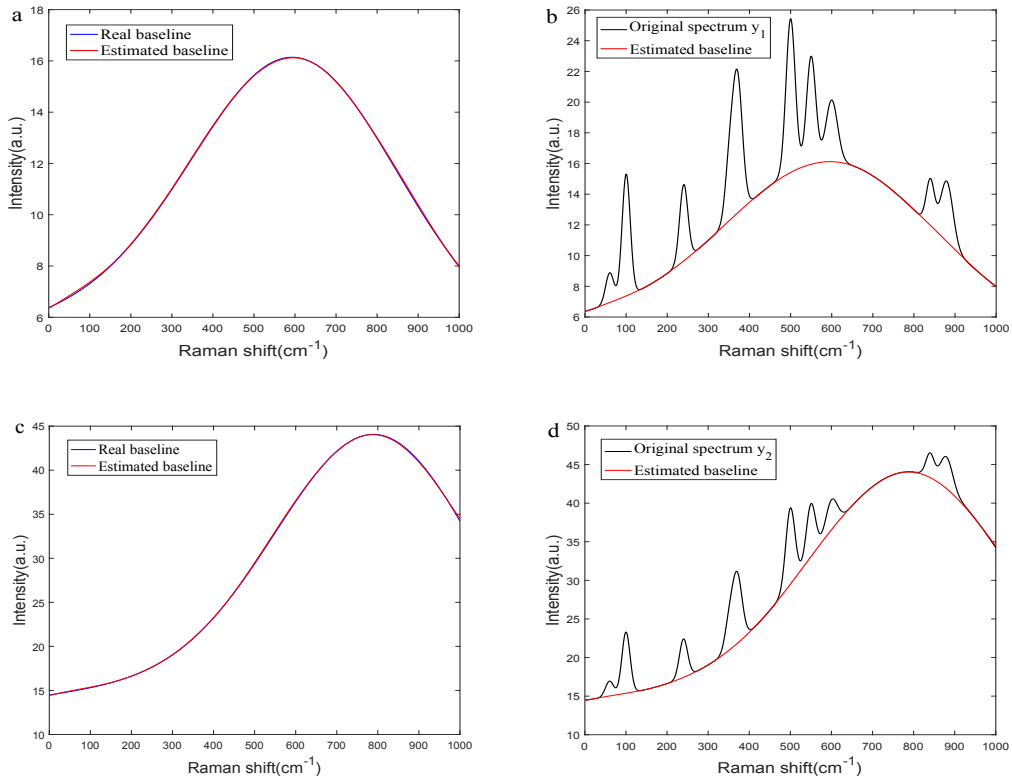


Fig. S7. Baseline correction results for simulated spectrum y_1 and y_2 based on the VTPspline method with suitable parameters (λ=5, $RMSE_1$=0.037, $RMSE_2$=0.068)

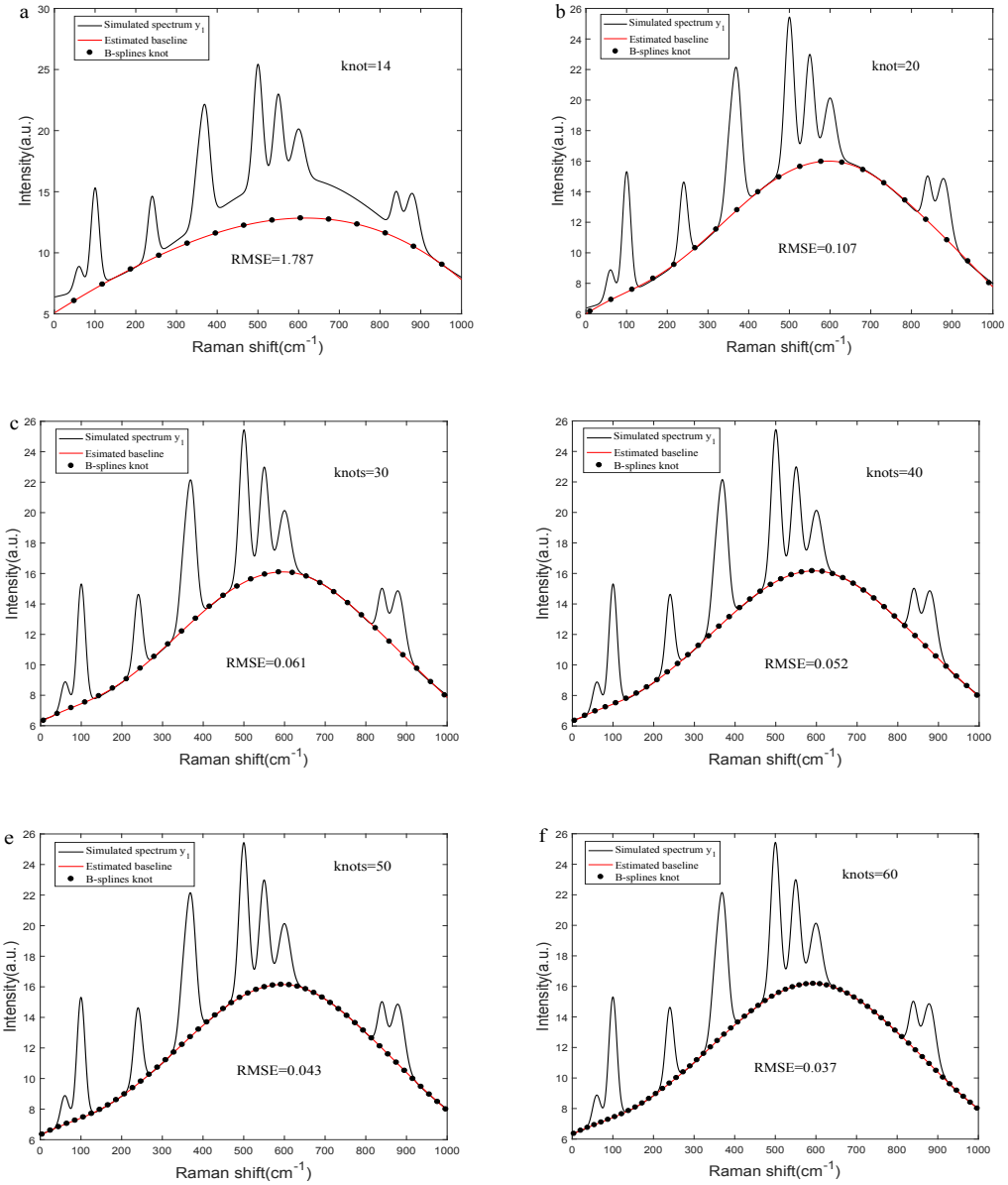
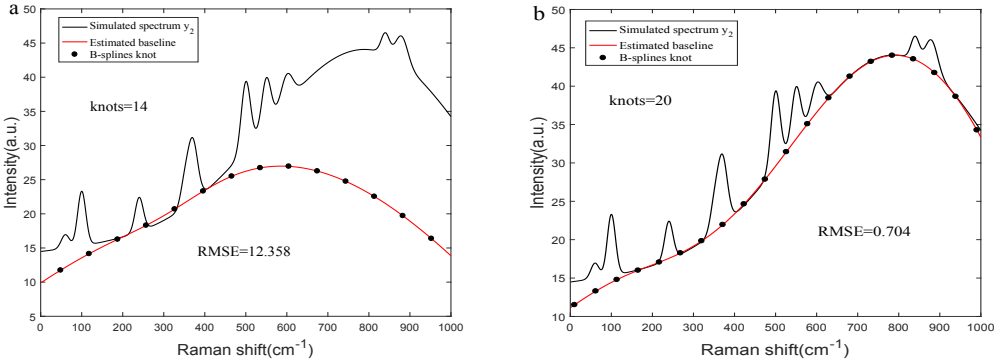


Fig. S8. Baseline correction results for simulated spectrum y_1 based on the VTPspline method with different number of B-splines knots (λ=5)



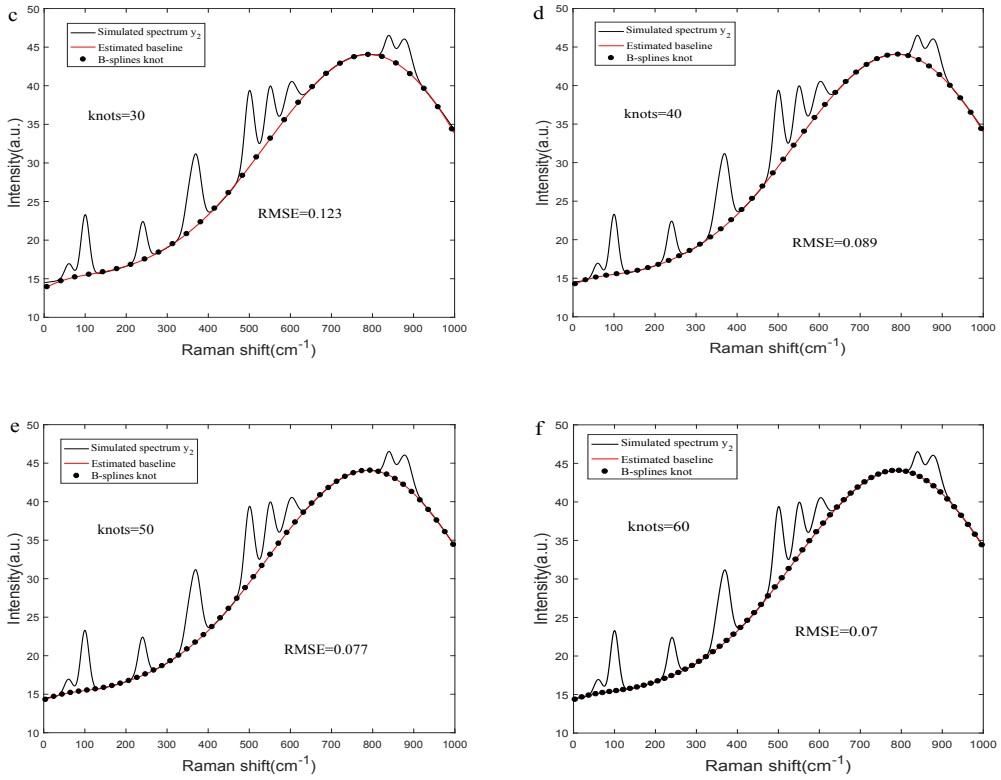


Fig. S9. Baseline correction results for simulated spectrum y_2 based on the VTPspline method with different number of B-splines knots (λ=5)

Compared among three methods for the experimental Raman spectra. Three mentioned baseline correction methods were used to estimate the baselines of experimental Raman spectra. The results of the baselines estimation for Raman spectra are shown in Fig. S10. By contrast, the fitting performance of the airPLS and IAsLS methods for the Raman spectra of the Aegirine and Tricyclazole at 200-600 cm-1(Raman spectra of Aegirine), 1000-1500 cm-1(Raman spectra of Aegirine), 2000-2500 cm-1(Raman spectra of Aegirine), 600-800 cm-1(Raman spectra of Tricyclazole), 1300-1500 cm-1(Raman spectra of Tricyclazole), 1900-2400 cm-1(Raman spectra of Tricyclazole) and 2700-3000 cm-1(Raman spectra of Tricyclazole) turned out to be bad. The experimental results obviously show that the proposed VTPspline method can precisely estimate the baseline in full wavenumber range, especially for estimating the background at both begins and ends of the spectra.

Handling the Raman spectra of complex mineral samples which were consist of hundreds of organic and inorganic substances. The proposed VTPspline method was used to handle the Raman spectra of complex mineral samples which were consist of hundreds of organic and inorganic substances. 9 mineral samples have been measured with an Ocean Optics confocal micro-Raman spectrometer, excited with the frequency-doubled Nd:YAG laser (785nm, Pmax of 300mW). The laser light was focused on the sample and the scattered radiation of the sample was collected. Meanwhile, spectra were sampling uniformly at a resolution of 3 cm-1 over a range from 175 to 2210 cm-1. Every sample was measured at several different points on the surface and three accumulations of 10-s integration time was used for each spectrum. Finally, the representative

Raman spectrum of each raw ore sample was calculated by the average of those spectra at different points. The results of the baselines estimation are shown in Fig. S11.

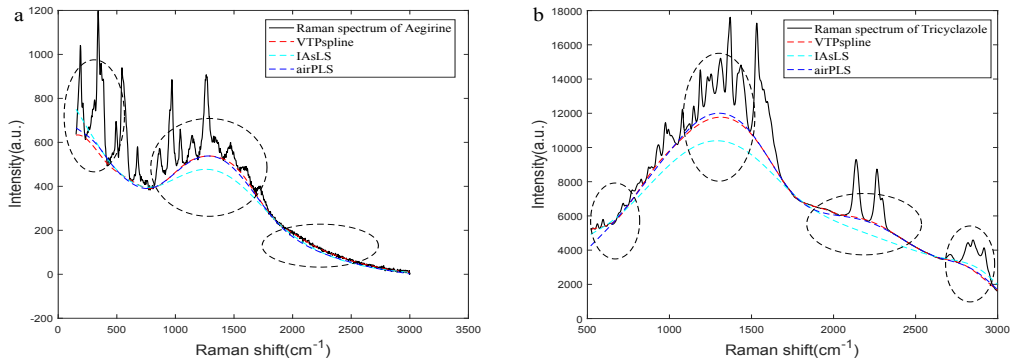


Fig. S10. Comparison among IAsLS, airPLS and VTPspline baseline correction methods for Raman spectrum of Aegirine (a) and Tricyclazole (b) respectively.

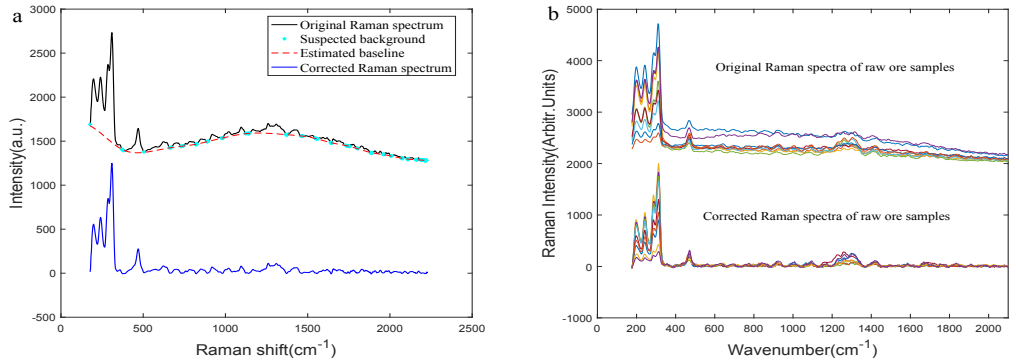


Fig. S11. (a) Original Raman spectra and (b) the preprocessed Raman spectra of the raw ore samples.

Actually, the proposed VTPspline baseline correction method was used to eliminate the strong fluorescence background for extracting the Raman characteristic peaks of most kinds of minerals which were included in the raw ore samples. And the corrected Raman spectra could be used to identify and determine the mineral composition in raw ore samples. The corrected Raman spectra of raw ore and major minerals are shown in Fig. S12. As shown in Fig. S12, curve a is the Raman signal of stibnite and the highest peak is observed at 312 cm-1, which is associated with the vibrational modes of Sb2S3. The overlapping peak is observed at 288 cm-1, which is attributed to vibration modes of SbS3. The other peaks are at 199 cm-1 and 248 cm-1, which are attributed to the Sb-Sb bonds of the Antimonite crystals. The main Raman bands of sericite (curve b) include 263 cm-1 and 402 cm-1, which are attributed to the cation exchange and longer wavelength lattice vibrations. And the important peak is observed at 699 cm-1, which is attributed to Si-Obr-Si tetrahedral expansion and bending vibration. The Raman bands of quartz (curve c) at 207 cm-1 and 464 cm-1 are attributed to the symmetric Si–O–Si stretching vibration and Si-O-Si bending vibration. The Raman bands of pyrite (curve d) and calcite (curve e) at 374 cm-1, 286 cm-1 and 1087 cm-1 can be attributed to Fe-[S2]2- stretching vibration, external vibration and symmetric

stretching vibration of [CO3]2-group. At last, the spectrum of fluorite (curve f) contains two strong peaks around 1301 cm-1 and 1420 cm-1, which is generated from photoluminescence formed after rare-earth ions of fluorite crystal are irradiated by optimal maser. And the Raman band of fluorite at 321 cm-1 is attributed to the lattice vibration of CaF2 and the spectral intensity is weak. Curve g is the Raman signal of raw ore.

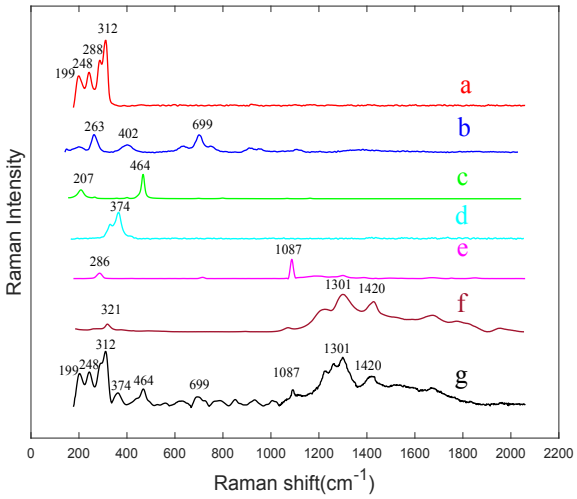
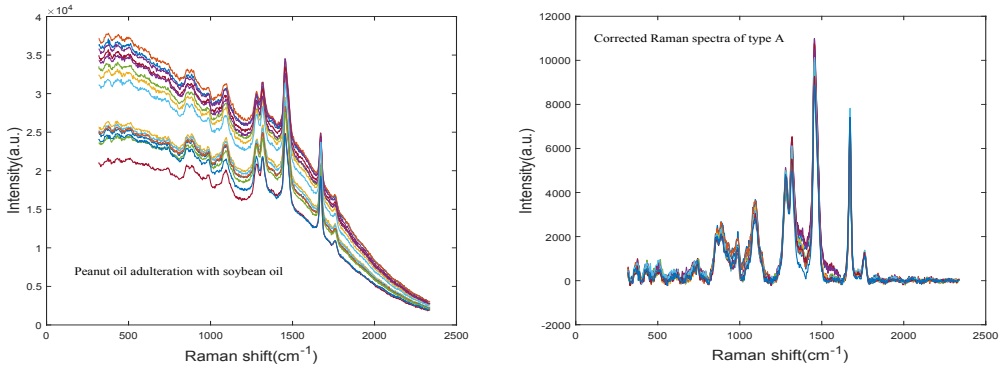


Fig. S12**.** Raman spectra of stibnite (a), sericite (b), quartz (c), pyrite (d), calcite (e), fluorite (f) and raw ore (g).

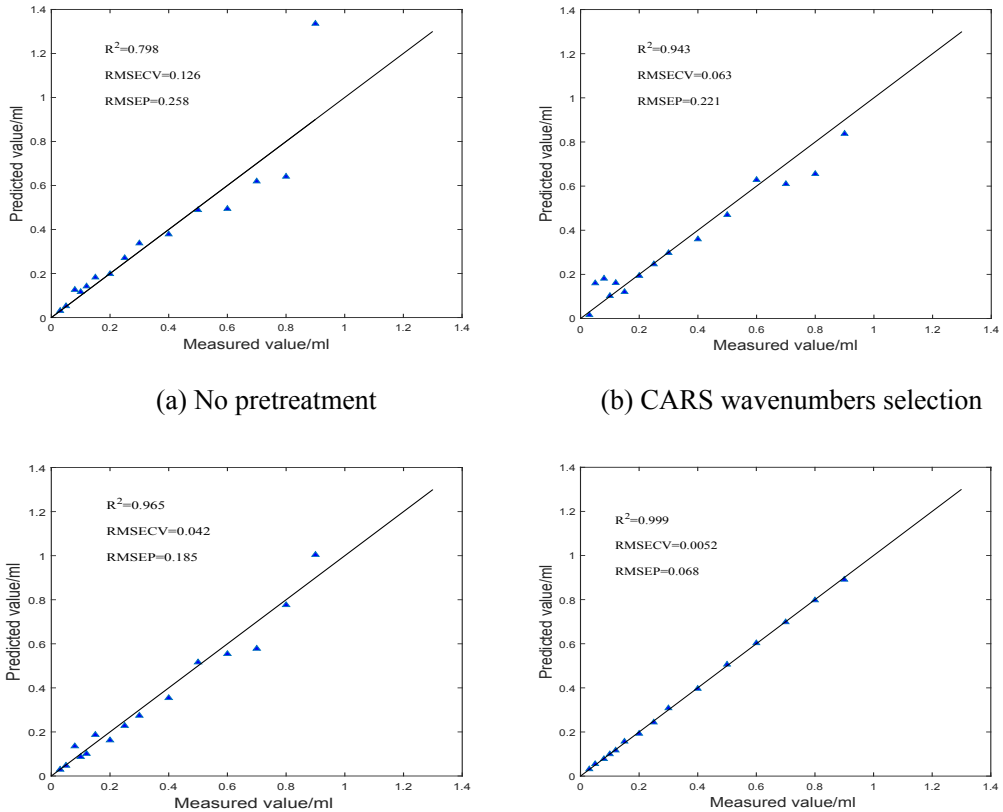
Detection of peanut oil adulteration with soybean oil by Raman spectroscopy and chemometrics. Peanut oil is a kind of edible oil which is relatively easy to be digested. For better composition, its fatty acid is easy to be digested and absorbed by the body. Eating peanut oil can break down cholesterol in the body into bile acid and get it out of the body, thus reducing the amount of cholesterol in plasma. Regular consumption of peanut oil can prevent skin wrinkling and aging, protect blood vessel walls, prevent thrombosis, and help prevent atherosclerosis and coronary heart disease. However, the cost of peanut oil is relatively high, and some immoral sellers often mix other kinds of oil into it, so it is very important to test the purity of peanut oil and the content of the adulterants. Spectra of 15 peanut oil samples with soybean oil (3%, 5%, 8%, 10%, 12%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%) were collected by the portable Raman spectroscopy system. Samples were measured by an Ocean Optics confocal micro-Raman spectrometer with a near-infrared laser at 785 nm. The laser power parameter was set at 300 mW and three accumulations of 5-s integration time was used for each spectrum. The obtained original Raman spectra of peanut oil adulteration with soybean oil were shown in Fig. S13(a). It is obviously that the Raman spectra of the mix samples include strong fluorescence backgrounds. And the fluorescence backgrounds among different oil samples present the big differences. The corrected Raman spectra treated by the proposed VTPspline method are shown in Fig. S13 (b). Both original and corrected Raman spectra were used to make quantitative analysis of the adulterants in the peanut oil samples. Furthermore, competitive adaptive reweighted sampling method (CARS) was used to select the information wavenumbers for improving the performance of the calibration models. Calibration model of quantitative analysis of the content of the adulterant in peanut oil samples is built using partial least square regression (PLSR) method and tested by leave- one-out cross

validation (LOOCV). Root mean square error of cross validation (RMSECV) and prediction (RMSEP), and coefficient of determination (R^2) are used to evaluate the performance of the calibration models. The results are shown in Table S1 and Fig. S14.



(a) Peanut oil adulteration with soybean oil (b) Corrected Raman spectra

Fig. S13**.** The original and corrected Raman spectra of peanut oil adulteration with soybean oil by the proposed VTPspline baseline correction method



(a) No pretreatment (b) CARS wavenumbers selection

(c) VTPspline baseline correction (d) CARS+VTPspline

Fig. S14**.** Correlation between actual and predicted values of peanut oil adulteration set based on PLS and pretreatment methods

Table S1. Prediction results of component in peanut oil samples based on PLS with different pretreatment methods

Adulteration oil	Pretreatment	The optimal factor	R^2	RMSECV	RMSEP
Soybean oil	No	6	0.798	0.126	0.258
	CARS	6	0.943	0.063	0.221
	VTPspline	6	0.965	0.042	0.185
	CARS+VTPspline	6	0.999	0.0052	0.068

It can be seen from Table S1 and Fig. S14 that the proposed VTPspline can eliminate the strong fluorescence background in the Raman spectra and improve the prediction accuracy of the calibration model for the quantitative analysis of the adulterants in the peanut oil.