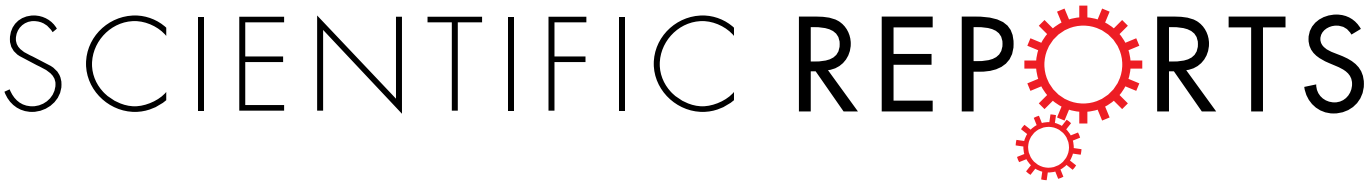

OPEN

Method for Removing Spectral Contaminants to Improve Analysis of Raman Imaging Data

Received: 29 July 2016
Accepted: 29 November 2016
Published: 05 January 2017

Xun Zhang[1], Sheng Chen[1], Zhe Ling[1], Xia Zhou[1], Da-Yong Ding[1], Yoon Soo Kim[2] & Feng Xu[1]

The spectral contaminants are inevitable during micro-Raman measurements. A key challenge is how to remove them from the original imaging data, since they can distort further results of data analysis. Here, we propose a method named "automatic pre-processing method for Raman imaging data set (APRI)", which includes the adaptive iteratively reweighted penalized least-squares (airPLS) algorithm and the principal component analysis (PCA). It eliminates the baseline drifts and cosmic spikes by using the spectral features themselves. The utility of APRI is illustrated by removing the spectral contaminants from a Raman imaging data set of a wood sample. In addition, APRI is computationally efficient, conceptually simple and potential to be extended to other methods of spectroscopy, such as infrared (IR), nuclear magnetic resonance (NMR), X-Ray Diffraction (XRD). With the help of our approach, a typical spectral analysis can be performed by a non-specialist user to obtain useful information from a spectroscopic imaging data set.

Using Raman imaging technique to obtain colourful chemical images is just a beginning for studying the chemical properties of a sample[1]. Chemists are more interested in interpreting the secrets locked within the spectral imaging data. Further data analysis, e.g. multivariate methods, allows the convoluted information content to be sorted according to the hypothesis that the original data is reconstructed from a limited number of significant factors[2]. However, such analysis is particularly sensitive to the presence of outliers, so that the original spectra involving spectral contaminants cannot be analyzed directly[3].

Generally, there are two major contaminants that spill over into the Raman channels along with the actual signals: (1) sample and background fluorescence, as well as thermal fluctuations of the charge coupled device (CCD), can markedly affect the spectral baseline resulting in baseline drifts[4]; (2) cosmic rays are sporadic background artifacts detected by sensitive detectors, which manifest in spectra as narrow-bandwidth spikes[5]. These samples or instrument dependent contaminants are inevitable during the measurement. Therefore, it is essential that the spectral data should be pre-processed by commonly used methods, which are available in instrumentation software or analysis programs before other algorithms are implemented (Table S-1).

Methods for handling baseline drifts and spikes are two independent subjects. The diverse sources of background and additive noise make it hard to correct baseline for experimental spectral data[6]. Wavelet transform[7], derivative[8], robust local regression[9] and polynomial fitting[10] were introduced to eliminate the varying background. Some drawbacks, however, such as poor performances in low signal-to-noise ratio environments and dependence on a given user's experience, have to be eliminated because they may lead to poor reproducibility of the calibration results. The approaches for removing spikes commonly fall into two categories depending on how the algorithm is designed. Methods in first category try to exclude the spikes on a single-scan spectrum via filtering algorithms such as wavelet processing, median filters and polynomial filters[11]. These methods suffer from serious limitations since they rely on an assumption of maximum spike bandwidth. Spectral distortion occurs when the bandwidth of spikes is comparable to spectral features of interest[12]. The alternative category suppresses spikes by comparing measured similar spectrum (referential spectrum). A typical example is "nearest neighbor correlation algorithm", which firstly confirms a referential spectrum by comparing the cross-correlation coefficient between spectra of adjacent and secondly eliminates the spikes by setting a threshold value on the basis of the user's experience[13]. The core of this category is how to confirm the referential spectrum, while the cross-correlation coefficient is inadequate to describe the features of numerical value. Moreover, the threshold

[1]Beijing Key Laboratory of Lignocellulosic Chemistry, Beijing Forestry University, Beijing, 100083, China. [2]Department of Wood Science and Engineering, Chonnam National University, Gwangju 500757, South Korea. Correspondence and requests for materials should be addressed to F.X. (email: xfx315@bjfu.edu.cn)

Wavenumbers (cm^{-1})	Components	Assignments
1095	C, H	heavy atom CC and CO stretching vibration
1123	C, H	heavy atom CC and CO stretching vibration
1163	C, H	heavy atom CC and CO stretching vibration plus HCC and HCO bending vibration
1275	L	aryl-O of aryl OH and aryl O−CH3; guaiacyl ring with C=O group
1331	L, C, H	HCC and HCO bending vibration
1378	C, H	HCC, HCO, and HOC bending vibration
1460	L, C, H	HCH and HOC bending vibration
1603	L	aryl ring stretching vibration, symmetrical vibration
1656	L	ring conjugated C=C stretching vibration of coniferyl alcohol; C=O stretching vibration of coniferaldehyde
2889	C, H	CH and CH_2 stretching vibration
2940	L, C, H	CH stretching vibration in OCH_3 asymmetric vibration

Table 1. Raman peak positions and bands assignments for major structures of poplar. C: Cellulose; H: Hemicellulose; L: Lignin.

value set by experience will also lead to unstable results. As a result, although this algorithm provides a new idea to resolve this issue, most of the applications are still limited to simulation data. Practical examples of this method are particularly scarce.

Here, we propose a novel approach for intelligently removing the spectral contaminants to improve analysis of Raman imaging data named "automatic pre-processing method for Raman imaging data set (APRI)" (Fig. 1). APRI consists of two complementary algorithms: (a) the adaptive iteratively reweighted penalized least-squares (airPLS algorithm) for baseline correction (Fig. 2a); (b) despiking algorithm on the basis of principal component analysis (PCA-despiking algorithm, Fig. 2b). Our method is neither dependent on the sample characteristics nor the measurement conditions. With the use of APRI, a typical spectral analysis can be performed by a non-specialist user to obtain useful information from a spectroscopic imaging data set.

Methods

Materials. An inclined 10-year-old poplar tree (*Populus nigra* L.) was provided by the arboretum of Beijing Forestry University, China. A small sample block was cut out from the seventh annual ring of the xylem. Without any embedding routing, a 10-μm-thick transverse section was prepared on a sliding microtome (Leica 2010R). It was then placed on a glass slide with a drop of D_2O and sealed with a coverslip for micro-Raman measurement. In Raman spectrum, D_2O can reduce the fluorescence of lignin and has a marked peak at 2490 cm^{-1}.

Micro-Raman system. The Raman imaging data set was acquired by using a micro-Raman system (LabRam Xplora, Horiba Jobin Yvon) equipped with a confocal microscope (Olympus BX51) and a motorized stage. The measurement was performed at room temperature (25 ± 3 °C). The entire optical system is diffraction-limited up to an object-side numerical aperture (NA) of 1.40 (Olympus, 100×, oil) and, consequently, is provided with a high spatial resolution in axial and lateral dimensions. Linear polarized laser ($\lambda = 532$ nm) was focused with a small spot size (theoretical 1.22λ/NA). Its power on the sample surface was around 8 mW. The Raman light was collected by a semiconductor-cooled charge coupled device (CCD) detector behind a grating spectrometer (1200 groves/mm). We selected 0.5 μm as the step for mapping and each pixel corresponds to one scan. The spectrum of each scan was obtained by averaging 4 s cycles. The confocal aperture was set at 100 μm. The reported depth resolution for the 100 μm confocal hole, based on the silicon (standard) phonon band at 520 cm^{-1}, was ~ 4 μm.

Data processing. The instrumentation software LabSpec (Horiba Jobin Yvon) was utilised to setup and control the micro-Raman system. The .ngc file, which is the format of the original Raman imaging data, was converted to .mat file for further data processing. Our software, APRI, is available as a Matlab script. The open-source code for APRI is available in Supplementary Materials.

Baseline correction method: airPLS algorithm. Conventional notation was adopted throughout this paper: uppercase bold face letter for matrices (as **X**), lowercase boldface for vectors (as **X**), italicized subscript characters for vector index (as x_i or z_i), and lowercase italicized letters for scalars (as $x_i(j)$). Superscripts are assigned as follows: T, vector or matrix transpose; and −1, matrix inverse. A Raman imaging data can be regarded as a matrix **X** of dimension *m* by *n*, in which the digitized Raman spectrum of each recorded position corresponds to a row vector in the data table:

$$\mathbf{X} = \begin{bmatrix} x_1(1) & x_1(2) & \dots & x_1(n) \\ x_2(1) & x_2(2) & \dots & x_2(n) \\ \vdots & \vdots & \vdots & \vdots \\ x_m(1) & x_m(2) & \dots & x_m(n) \end{bmatrix} \quad (1)$$

where *m* is the number of spectra traces in this data set and *n* is the number of data points per spectrum along the wavenumber (or any other spectral variable) axis, respectively.

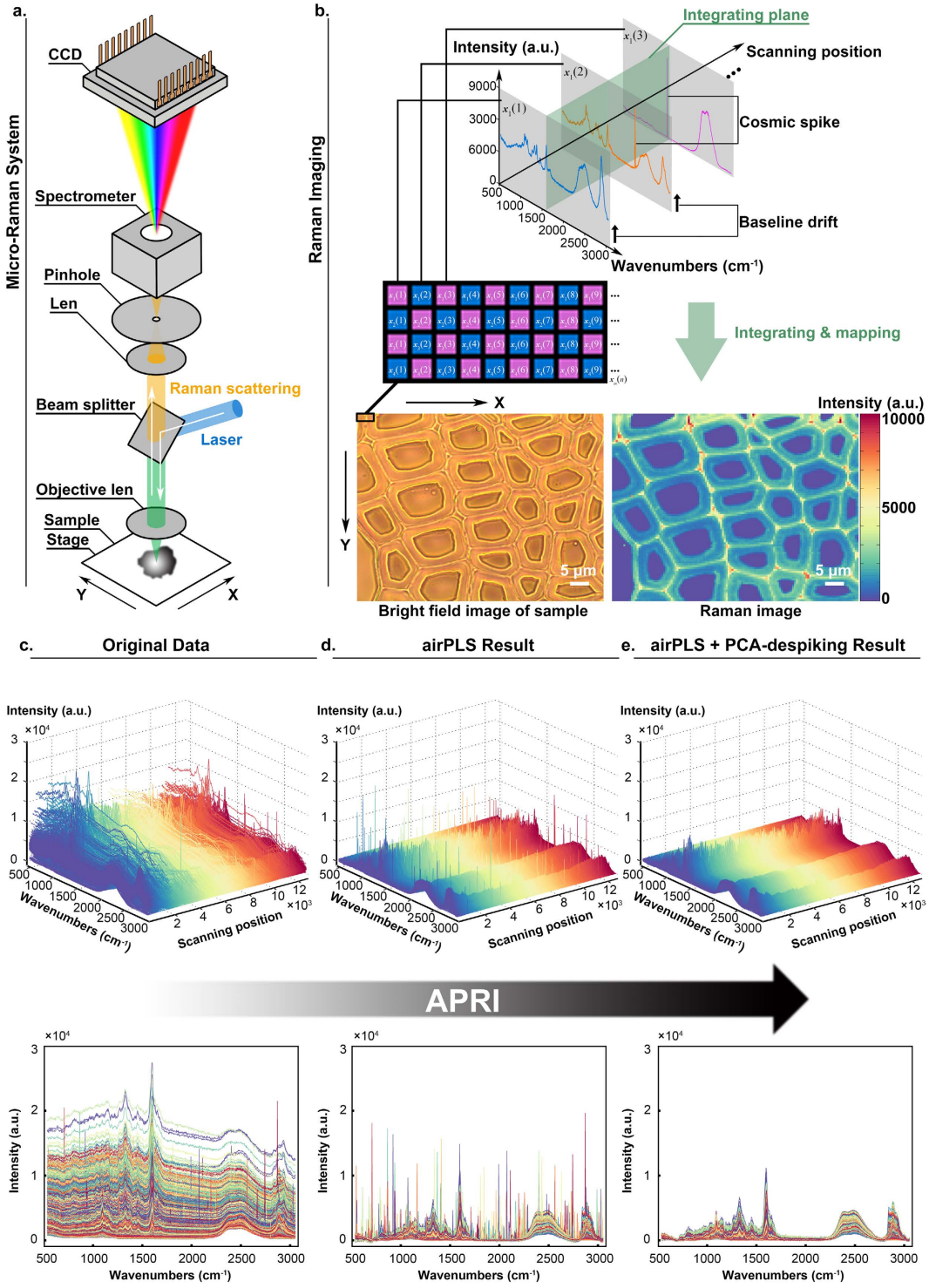

Figure 1. An overview of Raman imaging and APRI: The sample is measured by a micro-Raman system which couples an optical microscope to a Raman high-resolution spectrometer with a charge coupled device (CCD) detector. The bright field image obtained by the optical microscope is used to record the spatial information of the sample. The spectra are recorded as a matrix and corrupted by the contaminants from sample and CCD. The Raman image is achieved by integrating a specific Raman peak. Here we set a integrating plane around 1600 cm^{-1} to generate the Raman image of lignin. It can be used to locate the lignin semi-quantitatively.

The adaptive iteratively reweighted penalized least-squares (airPLS) algorithm was first proposed by Zhang Z.M. in 2010[14]. It works by treating the Raman imaging data set **X** as vectors in Hilbert space. We assume that

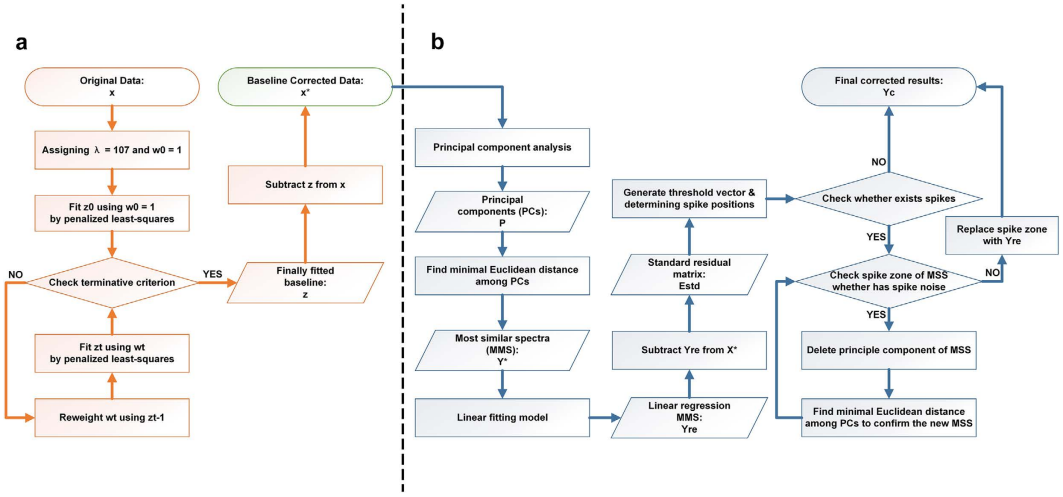

Figure 2. The flow chart of APRI: (**a**) adaptive iteratively reweighted penalized least-squares (airPLS) algorithm for baseline correction; (**b**) a despiking algorithm on the basis of principal component analysis (PCA-based despiking). Each step is corresponding to the equation in the Theory section.

x is the row vector of initial spectral data and **z** is the fitted vector. The fitted vector **z** is calculated by balancing the fidelity of **z** to **x** and the roughness of **z**:

$$Q = F + \lambda R = ||\mathbf{x} - \mathbf{z}||^2 + \lambda||\mathbf{D}\mathbf{z}||^2 \tag{2}$$

where Q is the balancing result, F is the fidelity expressed as the sum of squared errors between **x** and **z**, R is the roughness computed as the first derivative (matrix **D**) of **z**, λ is a parameter for controlling smoothness of the fitted vector (balancing coefficient), respectively. By finding the vector of partial derivatives and equating it to zero, i.e. $\partial Q/\partial \mathbf{z} = 0$, the solution of minimization problems of equation (2) is given as follows:

$$\mathbf{z}^{\mathrm{T}} = (\mathbf{I} + \lambda \mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{x}^{\mathrm{T}} \tag{3}$$

A weight vector of fidelity **W**, which is a diagonal matrix with w_i on its diagonal, is introduced for baseline correction. Thus equation (2) and equation (3) change to:

$$\mathbf{z}^{\mathrm{T}} = (\mathbf{W} + \lambda \mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\mathbf{W}\mathbf{x}^{\mathrm{T}} \tag{4}$$

Without setting zeros to the weight vector at positions corresponding to peak segments, the equation (4) can be categorized as a smoothing algorithm. Subsequently, the adaptive iteratively reweighted procedure is performed to calculate the weights and adds a penalty item to control the smoothness of the fitted baseline. Each step of the adaptive iteratively reweighted procedure involves solving a weighted penalized least squares problem of the following form:

$$Q^t = \sum_{i=1}^{n} w_i^t (x_i - z_i^t)^2 + \lambda \sum_{j=2}^{n} (z_j^t - z_{j-1}^t)^2 \tag{5}$$

The weight vector **w** is obtained adaptively using an iterative method. One should give an initial value $\mathbf{w}^0 = 1$ at the starting step. After initialization, the **w** of each iterative step t can be obtained using the following expressions:

$$w_i^t = \begin{cases} 0 & x_i \geq z_i^{t-1} \\ te^{\frac{z_i^{t-1} - x_i}{d^t}} & x_i < z_i^{t-1} \end{cases} \tag{6}$$

Vector d^t consists of negative elements of the differences between **x** and $\mathbf{z}^{t-1}$ in the t^{th} iteration step. The fitted value $\mathbf{z}^{t-1}$ in the previous $(t-1)^{\mathrm{th}}$ iteration is a candidate of baseline. If the value of the i^{th} point is greater than the candidate of baseline, it can be regarded as part of a peak. Thus its weight is set to zero to ignore this point at the next iteration of fitting. In airPLS algorithm, the iterative and reweight approaches are utilized to gradually and automatically eliminate the points of peaks and preserve the baseline points in the weight vector **w**. Iteration will stop either with the maximal iteration times or when the terminative criterion is reached. The terminative criterion is defined by:

$$d^t < 0.001 \sum_{i=1}^{n} |x_i| \tag{7}$$

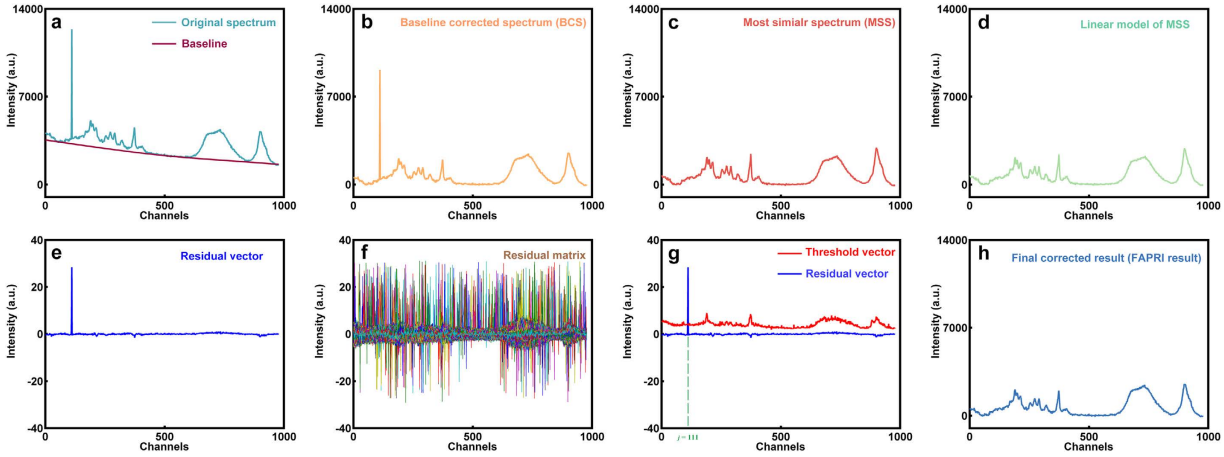

Figure 3. The procedure of APRI for a single spectrum. (**a**) The two spectral contaminants (baseline drift and spike) are observed in original spectrum. The baseline is calculated by using airPLS algorithm. (**b**) The baseline corrected spectrum (BCS) is corrected by subtracting the baseline from the original spectrum. (**c**) The most similar spectrum (MMS) is confirmed by PCA. (**d**) Linear fitting operation is necessary to reduce the large differences between BCS and MSS. Here, the spectrum of linear model is similar to that of MSS since the differences among BCS and MSS are not so apparent. (**e**) The standard residual vector is available by subtracting linear model of MSS from BCS. (**f**) The residual matrix is obtained by stacking the entire residual vectors together. This matrix is applied to determine the threshold vector. (**g**) The position of spike is located based on the residual vector and the threshold vector. (**h**) The two spectral contaminants are suppressed in the final pre-processed result.

The vector d^t also consists of negative elements of the differences between $\mathbf{x}$ and $\mathbf{z}^{t-1}$ (Fig. 3a). The corrected data $\mathbf{x}^*$ is achieved by subtracting finally fitted baseline $\mathbf{z}$ from original data $\mathbf{x}$ (Fig. 3b). Here, the maximum iterative time and λ are respectively set to 20 and 10^7 by considering the computation time and the required smoothness of the corrected result. The original algorithm is open source software available in https://code.google.com/archive/p/airpls/downloads.

Spikes removal method: PCA-despiking algorithm. Principle component analysis (PCA) is one of the multivariate methods and has been widely used with large multidimensional data sets[15]. The use of PCA allows the number of variables in a multivariate data set to be reduced, while retaining variation present in the data. The correlation coefficient array $R_j(l)$ of the baseline corrected data $\mathbf{X}^*$ is calculated as follows:

$$R_j(l) = \frac{\mathrm{cov}(x^*(j), x^*(l))}{\sigma_{x^*(j)} \times \sigma_{x^*(l)}}$$
$$j, l = 1, 2, \ldots, n \quad (8)$$

where $\mathrm{cov}(x^*(j), x^*(l))$ is the covariance of column vectors $x^*(j)$ and $x^*(l)$, $\sigma_{x^*(j)}$ is the standard deviation of $x^*(j)$, and $\sigma_{x^*(l)}$ is the standard deviation of $x^*(l)$, respectively.

Then the eigenvalues and eigenvectors are determined from the correlation coefficient array,

$$(R - \lambda_k I_m) V_k = 0 \quad (9)$$

where λk—eigenvalues and $\sum_{k=1}^{n} \lambda_k = n$; $V_k = [v_k(1), v_k(2), \ldots, v_k(n)]^{\mathrm{T}}$—eigenvectors corresponding to the eigenvalues λ_k.

The eigenvectors are aligned in descending order with respect to their eigenvalues, i.e. $\lambda_k > \lambda_{k+1}$. The uncorrelated principal component is formulated as:

$$p_i(k) = \sum_{j=1}^{n} x_i^*(j) \times v_k(j) \quad (10)$$

$$P_k = [p_1(k), p_2(k), \ldots, p_m(k)]^{\mathrm{T}} \quad (11)$$

where P_k is termed as the principal component. When applying PCA to the data, a significant factor is how many principal components to keep. This is determined by using the proportion of total population variance. The proportion (PR) of k^{th} principal component (P_k) is shown as follows:

$$PR_{P_k} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_n} \quad (12)$$

where λ_k is the eigenvalue (variance) of k^{th} eigenvector. This represents the explanation of k^{th} principal component for the spectral data set. In APRI, the first q principal components are retained when the cumulative PR exceeds 85%, which means that 85% of the data variance was explained. The selected principal components can reflect the original data very well and thus the dimension of the data is reduced to save the running time in subsequent calculation of the distances. A principal components matrix PC (dimension m by q) is then achieved. In this matrix, each row of PC is treated as the feature of corresponding Raman spectrum.

The squared Euclidean distance array $D_i(j)$ between each two features is calculated as:

$$D_i(j) = \sum_{k=1}^{q} [PC_i(k) - PC_j(k)]^2 \tag{13}$$

where $PC_i(k)$ and $PC_j(k)$ are the features of i^{th} and j^{th} spectrum in k^{th} principal components. The most similar spectrum (MSS, $\mathbf{y}^* = [y_1^*, y_2^*, \ldots, y_n^*]$) of each spectrum ($\mathbf{y} = [y_1, y_2, \ldots, y_n]$) is determined by finding the minimal Euclidean distance between the features of baseline corrected spectrum (BCS) and referential spectra (Fig. 3c). Since all of the spectra are taken into consideration, it should be noted that the procedure of confirming MSS will cost a lot of time if the size of original data set is too large.

A linear model is proposed by the least-squares algorithm to further reduce the variation between $\mathbf{y}$ and $\mathbf{y}^*$ (Fig. 3d). The linear regression spectrum $\mathbf{y}_{re}$ is generated by $\mathbf{y}^*$, and the sum squares errors S between $\mathbf{y}$ and $\mathbf{y}_{re}$ is given as follows:

$$\mathbf{y}_{re} = a\mathbf{y}^* + b \tag{14}$$

$$S = \|\mathbf{y} - \mathbf{y}_{re}\|^2 = \|\mathbf{y} - a\mathbf{y}^* - b\|^2 \tag{15}$$

where a is the scaling factor and b is the offset, respectively. By calculating the vector of partial derivatives and equating it to zero, i.e. $\partial S/\partial a = 0$, $\partial S/\partial b = 0$, the solution of a and b is achieved,

$$a = \frac{n\sum_{i=1}^{n} y_i y_i^* - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} y_i^*}{n\sum_{i=1}^{n} (y_i^*)^2 - (\sum_{i=1}^{n} y_i)^2} \tag{16}$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} y_i^*\right) \tag{17}$$

where y_i and y_i^* are elements of $\mathbf{y}$ and $\mathbf{y}^*$, respectively. The residual vector $\mathbf{e}$ is calculated as follows:

$$\mathbf{e} = \mathbf{y} - \mathbf{y}_{re} \tag{18}$$

This procedure is applied to magnify the intensity of existing spikes and makes them more obvious targets for locking their positions (Fig. 3e). The residual vector $\mathbf{e}$ is normalized by:

$$\mathbf{e}_{std} = \mathbf{e}/\sigma_{\mathbf{e}} \tag{19}$$

where $\sigma_{\mathbf{e}}$ is the standard deviation of $\mathbf{e}$. The normalization can ensure that the residual vector is independent of the numerical range of the spectral intensities. It is particularly important for extending our method to other samples. The determination of spike positions is based on a residual matrix $\mathbf{E}_0$ achieved by stacking all of $\mathbf{e}_{std}$ together (Fig. 3f). A threshold vector is required for filtering the spike signals in E_0. Since the spike has features including narrow-bandwidth, relative high intensity, positive (unidirection) and random occurrences in spectral data, the spike signals in each channel should be positive numbers and far away from normal signals. We first obtain a new residual matrix $\mathbf{E}$ by ranking the values of each column in $\mathbf{E}_0$ are ranked from largest to smallest. The threshold vector $\mathbf{t}$ is then automatically generated at the point of abrupt change in each channel based on the derivative of $\mathbf{E}$ (Figs 3g and 4). The shape and width of spikes depend heavily on instrument/measurement mode and manufacturer software. In our experiment, a spike zone generally contains 20 to 40 variables in spectrum. Thus we select the spike range as 41 variables to ensure that all affected variables are included. Without consideration of the shape of the spike, the PCA-despiking algorithm is finished by replacing the spike zone with the corresponding $\mathbf{y}_{re}$ (Fig. 3h). It should be noted that a wider spike range means more variables are corrected, while introducing MSS can guarantee the shape of corrected spectrum shape won't be distorted much. According to the principle of the algorithm, bad replacement may appear in final result if the spike zone of corresponding MSS also has spike noise. In this case, instead of the first referential spectrum, the spectrum with next minimal distance as the MSS is selected.

Our algorithm relies on the differences between BCS and MSS, which may lead to unsatisfactory results when the spikes of them appear at the same position. Although the probability of this event is low, it limits the performance of the algorithm to a certain extent. Since the MSS of each spectrum changes dynamically during the iterative procedure so that the appearances of spikes at the same position are evitable, our solution is to perform the APRI once again.

Results and Discussion

Raman spectroscopy can provide chemical information about organic and inorganic substances quickly and nondestructively with little to no sample preparation. The analysis of organic compounds is much more difficult and

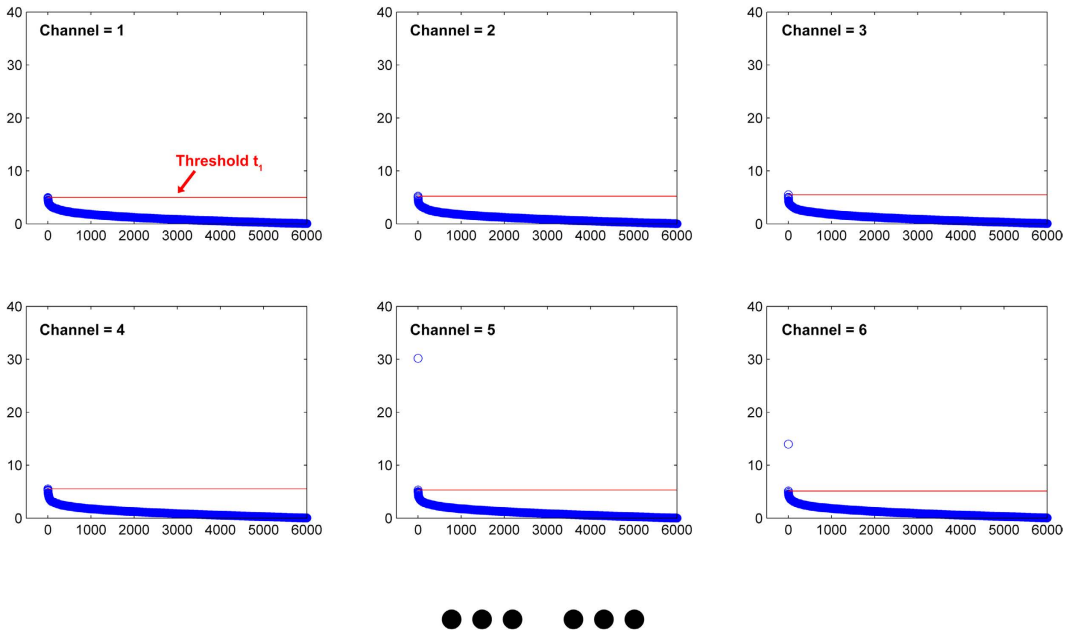

Figure 4. Automatic generation of the threshold vector t. For channel j of the new residual vector $\mathbf{E}$, we just consider the values that are positive based on the positive nature (unidirection) of spikes. Threshold $\mathbf{t}_j$ should satisfy that $\mathbf{t}_i = \mathbf{E}_{ij}$, $(d\mathbf{E}_j < -1)$.

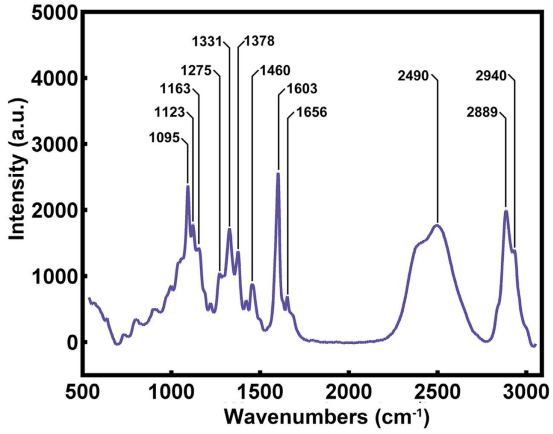

Figure 5. Typical Raman spectrum of poplar in presence of D2O, D2O can reduce the fluorescence of lignin and has a marked peak at 2490 cm^{-1}.

complicated than that of inorganic ones. In this study, the experiments are performed on a transverse section of poplar xylem (a biological sample) to verify the validity of our method. The poplar cell wall is organized in several layers formed at different periods during cell differentiation with different proportions of components, including cellulose, hemicellulose and lignin. Figure 5 shows a typical Raman spectrum of poplar and its band assignments are displayed in Table 1 on the basis of previous literature[16,17]. As a typical biological material, the vibrational spectra of poplar are rather complex with overlapping bands. In particular, cellulose and hemicellulose are hard to discern in spectrum because of their similar chemical bonds[18].

We collected 12,870 spectra from a $65 \times 49.5\,\mu m$ region with a spatial resolution $0.5\,\mu m$/pixel. It can be regarded as a matrix of 12,870 by 977 dimensions. The plots of original Raman spectral data are shown in Fig. 6a. We notice that the original data have significant baseline drifts and spikes. As previously mentioned, these spectral contaminants are caused by sample fluorescence and CCD. In plants, lignin is known to exhibit fluorescence under laser with specific wavelength, which interferes strongly with the determination of spectral peaks[19]. Many efforts have been undertaken by predecessors to eliminate these interferences, such as using near-infrared laser, but the problems are still considerable[20].

The common used instrumentation software, such as LabSpec (Horiba Jobin Yvon) and OPUS (Bruker), provides the polynomial fitting algorithm and the median filtering algorithm to cope with the baseline drifts and the spikes problems, respectively. Polynomial fitting algorithm attempts to estimate the unknown background and

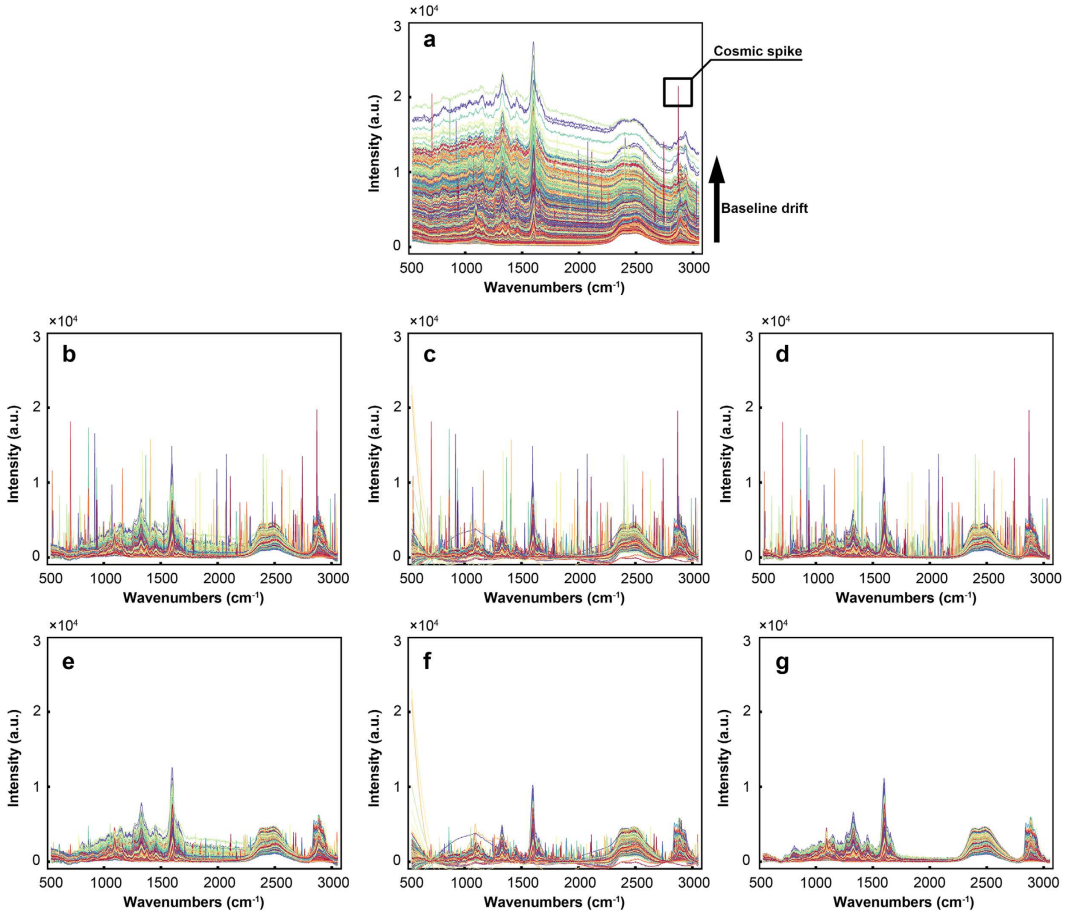

Figure 6. Data pre-processing results achieved by instrumentation software and APRI. (**a**) The plots of the original Raman imaging data. (**b**) Baseline correction by polynomial fitting (degree = 2). (**c**) Baseline correction by polynomial fitting (degree = 2). (**d**) Baseline correction by APRI (degree = 2). (**e**) Despiking operation by median filtering on the basis of (**b**). (**f**) Despiking operation by median filtering on the basis of (**c**). (**g**) Despiking operation by APRI on the basis of (**d**).

abolish the sloped or oscillatory baselines based on user-defined polynomial degree and points[21]. However, it suffers from a very loose baseline if the degree of polynomial is not set properly. In addition, the unknown background of the spectrum is often too complicated to be fitted by a simple polynomial function in practice thereby rendering unsatisfactory corrected results. As shown in Figs 5c and 6b, changing the degree of polynomial will achieve different correction results. When the degree is set to 2, it is found that the correction of baseline is incomplete. If we enlarge the degree number to 8, significantly spectral distortion are observed, such as appearance of negative parts and changes in spectral shape. The median filtering algorithm requires the user to specify the maximum number of spike bandwidth and height as the references for determination of spikes, indicating that the algorithm is effective in detecting the sharp or high spikes but is unable to remove the spikes with low intensity[11]. Figure 5f and 6e show that the median filtering algorithm only can remove parts of the spikes.

The principle of APRI is fundamentally different, allowing automated operation by the applications of airPLS and PCA-despiking algorithm. The airPLS algorithm works by iteratively changing the weights of fitted spectrum to confirm the baseline function. Compared to the conventionally used algorithms of baseline correction, airPLS algorithm achieves a smoother result without adjustable parameters because the baseline is determined by balancing fidelity and roughness of the fitted spectrum. The PCA-despiking algorithm performs PCA to extract the main features (i.e. scores) of the baseline corrected spectra (BCS). For each BCS in the spectral data, the algorithm searches the most similar spectrum (MSS) from all the feature signatures (using the Euclidean distance) and excludes the spikes by comparing the differences between the BCS and MSS. The final results show that the spectral contaminants are perfectly removed (Fig. 6d and g). In order to prove that the results were not idiosyncratic to the particular Raman data set, other examples pre-processed by APRI can be found in Supplementary Information (Figures S-1 to S-5). The core of spike corrections is how to confirm the MSS. Here, we employ PCA and Euclidean distance to find out the MSS. Other judgment methods, such as Mahalanobis distance or Pearson correlation coefficient, also can be applied to determine the MSS. Mahalanobis distance is similar to our method (PCA plus Euclidean-Distance): both of them take into account the correlations of the data. The calculated amount of our method is less than Mahalanobis distance because the original data has been condensed

by PCA. Using Pearson correlation coefficient, by contrast, requires only a low calculation cost. However, it is independent of scaling and thus may distort the final corrected spectra.

To further analyze the Raman spectral data, a standard PCA is performed on original and corrected data. The first 20 row vectors (loadings) of PCA results were selected to generate their mapping images (Figures S-6 to S-25). For the original data results, the baseline drifts exist in the 1–9 spectra with high spatial resolved mapping profiles, while the spikes are visible in the 11–20 spectra with low image resolution. This means that the PCA not only can extract and condense the features of contaminants in the Raman imaging data set, but also can be potentially used as an evaluation of the correction quality. It is found that both the baseline drifts and the spikes disappear in the PCA results of APRI corrected data. Moreover, the mapping profiles show a higher spatial resolution than the original ones, which suggests that the loadings of corrected data are more representative. Although APRI is a powerful tool for removing the spectral contaminants, it may not be suitable in certain applications. Our algorithm for spike correction depends crucially on the availability of a closely similar spectrum without spike at the same position. Spectral distortion may occur when the sample size is too small (generally spectral amounts less than 10, depending on the measurements). Large differences will be found between BCS and MSS thereby providing incorrect determination of the spikes. In such case, filtering algorithms that exclude the spikes on a single-scan spectrum is more appropriate. Additionally, APRI is computationally faster for large-scale Raman imaging data. It takes less than 4 minutes to correct 10,000 spectra on a standard desktop computer (Win7 OS, Intel Core i5-3.2Ghz CPU and 8GB RAM).

The spectral contaminants induced by sample or instrumental perturbations are not only common in Raman spectroscopy but also in other methods of spectroscopy, such as infrared (IR), nuclear magnetic resonance (NMR), X-Ray diffraction (XRD). More information can be gleaned from the spectral data by applying various multivariate methods, whilst the unknown background or unwanted peaks may complicate the subsequent analysis of their data. The uniqueness of APRI is the self-adaptive nature of airPLS and PCA-despiking algorithms, which removes the contaminants on the basis of the spectral features themselves. Therefore, APRI is potential to be an effective, freely available tool for pre-processing the data of other spectroscopic techniques to improve the quality of data analysis.

References

1. Butler, H. J. *et al.* Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.* **11,** 664–687 (2016).
2. Shinzawa, H., Awa, K., Kanematsu, W. & Ozaki, Y. Multivariate data analysis for Raman spectroscopic imaging. *J Raman Spectrosc* **40,** 1720–1725, doi: 10.1002/jrs.2525 (2009).
3. Notingher, I. *et al.* Multivariate analysis of Raman spectra for *in vitro* non-invasive studies of living cells. *J. Mol. Struct.* **744,** 179–185 (2005).
4. Barman, I., Kong, C.-R., Singh, G. P. & Dasari, R. R. Effect of photobleaching on calibration model development in biological Raman spectroscopy. *J. Biomed. Opt.* **16** (2011).
5. Zhang, L. & Henson, M. J. A practical algorithm to remove cosmic spikes in Raman imaging data for pharmaceutical applications. *Appl. Spectrosc.* **61,** 1015–1020 (2007).
6. Baek, S. J., Park, A., Ahn, Y. J. & Choo, J. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *Analyst* **140,** 250–257 (2015).
7. Zhang, Z.-M., Chen, S. & Liang, Y.-Z. Peak alignment using wavelet pattern matching and differential evolution. *Talanta* **83,** 1108–1117 (2011).
8. O'Grady, A., Dennis, A. C., Denvir, D., McGarvey, J. J. & Bell, S. E. Quantitative Raman spectroscopy of highly fluorescent samples using pseudosecond derivatives and multivariate analysis. *Anal. Chem.* **73,** 2058–2065 (2001).
9. Leger, M. N. & Ryder, A. G. Comparison of derivative preprocessing and automated polynomial baseline correction method for classification and quantification of narcotics in solid mixtures. *Appl. Spectrosc.* **60,** 182–193 (2006).
10. Cobas, J. C., Bernstein, M. A., Martín-Pastor, M. & Tahoces, P. G. A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *J. Magn. Reson.* **183,** 145–151 (2006).
11. Ehrentreich, F. & Summchen, L. Spike removal and denoising of Raman spectra by wavelet transform methods. *Anal. Chem.* **73** (2001).
12. Li, S. & Dai, L. An improved algorithm to remove cosmic spikes in Raman spectra for online monitoring. *Appl. Spectrosc.* **65** (2011).
13. Behrend, C. J., Tarnowski, C. P. & Morris, M. D. Identification of outliers in hyperspectral Raman image data by nearest neighbor comparison. *Appl. Spectrosc.* **56,** 1458–1461 (2002).
14. Zhang, Z. M., Chen, S. & Liang, Y. Z. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* **135,** 1138–1146 (2010).
15. Geladi, P. & Grahn, H. In *Encyclopedia of Analytical Chemistry* (ed Robert A. Meyers) 13540–13562 (John Wiley & Sons, Ltd, 2000).
16. Agarwal, U. P. Raman imaging to investigate ultrastructure and composition of plant cell walls: distribution of lignin and cellulose in black spruce wood (Picea mariana). *Planta* **224** (2006).
17. Zhang, X. *et al.* Method for automatically identifying spectra of different wood cell wall layers in Raman imaging data set. *Anal. Chem.* **87** (2015).
18. Gierlinger, N. & Schwanninger, M. Chemical imaging of poplar wood cell walls by confocal Raman microscopy. *Plant Physiol.* **140,** 1246–1254 (2006).
19. Donaldson, L. A. & Radotic, K. Fluorescence lifetime imaging of lignin autofluorescence in normal and compression wood. *J. Microsc.* **251,** 178–187 (2013).
20. Agarwal, U. P., Reiner, R. S. & Ralph, S. A. Cellulose I crystallinity determination using FT-Raman spectroscopy: univariate and multivariate methods. *Cellulose* **17,** 721–733 (2010).
21. Mazet, V., Carteret, C., Brie, D., Idier, J. & Humbert, B. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometr. Intell. Lab.* **76,** 121–133 (2005).

Acknowledgements

This work was funded by the Chinese Ministry of Education (113014A), the China National Funds for Distinguished Young Scientists (31225005) and the Fundamental Research Funds for the Central Universities (BLYJ201620).

Author Contributions

F.X. and X.Z. initiated the project. F.X. and X.Z. designed the pre-processing method. X.Z. and S.C., implemented the software. B.W. and X.Z. prepared the poplar sample. Z.J. collected the image data. X.Z. performed the analysis and wrote the manuscript. F.X. and Y.S.K. contributed to suggestions to the experiments, the analysis of the results and their discussion. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zhang, X. *et al.* Method for Removing Spectral Contaminants to Improve Analysis of Raman Imaging Data. *Sci. Rep.* **7**, 39891; doi: 10.1038/srep39891 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.