Decomposition Method for Raman Spectra of Dentine

Oleg O. Frolov[1,2*], Pavel E. Timchenko[1,2], Elena V. Timchenko[1,2], and Larisa T. Volova[2]

[1] Samara National Research University, 34 Moskovskoe shosse, Samara 443086, Russian Federation
[2] Samara State Medical University, 89 Chapaevskaya str., Samara 443099, Russian Federation
*e-mail: frolovaleh@gmail.com

Abstract. This article presents a developed two-stage method for decomposing a spectral contour with a high degree of overlap of Raman scattering (RS) lines. The algorithm allows you to take into account the error in the position of the Raman bands and other parameters, and work with asymmetric lines. By determining the final model based on multiple spectra, it allows the Raman lines to be correctly identified. A model experiment was carried out to reconstruct the composition and parameters of elementary lines based on synthetic spectra. The error in determining the line parameters was characterized by MAPE (mean absolute percentage error) for peak intensity being 0.3%, MAPE for half-width dx being 0.3%, and MAE (mean absolute error) for the peak position x_0 being 0.1 cm^{-1}. The algorithm was applied to the real problem of analyzing the spectra of demineralized dentin. © 2024 Journal of Biomedical Photonics & Engineering.

Keywords: Raman spectroscopy; decomposition; clustering; biomaterials.

Paper #9074 received 3 Mar 2024; revised manuscript received 6 Aug 2024; accepted for publication 10 Aug 2024; published online 7 Sep 2024. doi: 10.18287/JBPE24.10.030303.

1 Introduction

Raman spectral data in a number of problems have a number of common problems, such as band overlap, noise, line broadening and other factors. These problems create difficulties for subsequent spectral analysis. Therefore, the decomposition of Raman spectra is necessary and important, and also makes it possible to reduce the dimension of the feature space in further statistical analysis [1−4].

Despite the advantages of Raman spectroscopy, there is no single universal approach to the problem of decomposition of Raman spectra. To resolve overlapping bands, two conventionally separated fundamental approaches are used: deconvolution methods [5−7] and indirect methods of hard modeling [1, 2, 4]. There are blind and semi-blind deconvolution methods.

Typically, blind and semi-blind deconvolution methods use a priori knowledge of the instrumental broadening function to build a parametric model [5], and some of them use various regularizers [6, 7]. Blind and semi-blind deconvolution methods allow, to a certain extent, to reconstruct the original spectra and estimate the instrumental broadening function. However, overlapping bands cannot be completely resolved and the influence of noise cannot be avoided.

Alsmeyer's indirect hard modeling method [1, 2] establishes a peaked model of Raman spectra to resolve overlapping bands. The authors also believe that the actual Raman spectrum is a set of Lorentz lines, and the Raman spectrum measured by the spectrometer is a set of the Voigt function. It is also indicated that the correct selection of the number of lines, their width and position is extremely important and is a key factor in the successful modeling of the spectral contour.

The Voigt function is the result of the convolution of the Cauchy-Lorentz distribution and the Gaussian distribution. Spectrum deconvolution methods are precisely aimed at selecting the broadening function, which in the general case is described by a Gaussian distribution. Broadening effects of any origin lead to a convolution of the Lorentz line profile and the profile characteristic of the broadening [3]. The algorithm described in article [4] is positioned as a method for modeling the Raman spectrum using Voigt functions, which also allows taking into account the broadening of the Lorentz line.

The decomposition method implemented in the article [8] effectively combines evolutionary algorithms with gradient-based optimization to solve the complex problem of spectral decomposition. The innovative use of Gender Genetic Algorithm (GGA) enhances the

performance by introducing diversity in the genetic operations, thereby improving the robustness of the solution.

Traditional methods, such as Savitzky-Golay (SG) smoothing and discrete wavelet transformation (DWT), have been widely used for spectral denoising and decomposition. However, these techniques sometimes fall short in accurately recovering the true signal, especially in the presence of strong noise or complex spectral features [9, 10].

Recent advancements have introduced more sophisticated approaches, such as the Hilbert Vibration Decomposition (HVD) and deep learning-based methods. The HVD, for example, decomposes the Raman spectra into components, enabling more precise peak identification and reconstruction from noisy data. This method has shown significant improvements over traditional techniques in terms of signal preservation and noise reduction [9].

The advanced methods for Raman spectral decomposition, particularly those integrating machine learning and deep learning frameworks, offer several advantages. These include improved accuracy in peak identification, enhanced noise reduction, and the ability to handle large datasets with complex spectral features. For instance, the application of convolutional neural networks (CNNs) has shown promise in refining the initial estimates provided by traditional methods, leading to more accurate spectral reconstructions. According to a study presented in Refs. [9−11] the CNN model was trained on a diverse dataset of Raman spectra. This approach demonstrated remarkable accuracy in decomposing complex spectral data with overlapping peaks and varying noise levels. The CNN's ability to generalize from the training data allowed it to effectively handle new, unseen spectra with minimal manual intervention.

However, these advanced methods also come with limitations. The requirement for extensive training datasets and computational resources can be a significant barrier, especially for researchers with limited access to such resources. Additionally, the complexity of these methods may limit their accessibility to users without specialized knowledge in machine learning and data processing.

The development of decomposition methods is not limited to Raman spectroscopy alone. Techniques such as nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry, and chromatography also rely heavily on effective spectral decomposition for accurate analysis. For instance, in NMR spectroscopy, methods like Maximum Entropy and Bayesian approaches have been utilized to decompose and interpret complex spectra. Similarly, in mass spectrometry, algorithms such as the deconvolution of overlapping peaks using Gaussian and Lorentzian functions are common practices [12].

These methods, while effective in their respective domains, highlight the versatility required in spectral decomposition approaches. Each technique presents unique challenges and benefits, often requiring tailored solutions to achieve optimal results.

The goal of the work is to develop an algorithm for automatic decomposition of Raman spectra of biomaterials.

2 General Scheme of the Algorithm

The input data is an array of analyzed Raman spectra after preprocessing and parameters set by the user.

Preprocessing consisted of normalizing the spectra using the standard normal variation (SNV) method [13]. The spectra were smoothed using the Maximum Likelihood Estimation Savitzky-Golay filter method (MLE-SG) [14] with the σ parameter equaled to 4. To eliminate the contribution of autofluorescence in the Raman spectrum, various methods of baseline correction can be applied.

The only requirement for the baseline correction algorithm is to determine a baseline close to the true one, so that it does not lower the spectrum contour into the negative region or raise it. If the Raman spectrum is defined as Curve 0 in the Fig. 1, then in the process of spectrum decomposition, lines may appear that describe elementary lines that do not exist in reality in range ~300 cm^{-1}. And in the case of Curve 2, part of the Raman spectrum is in the negative region. The ideal case is shown as Curve 1.

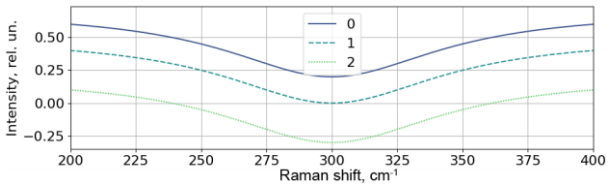


Fig. 1 The result of determining the baseline corrected Raman spectrum in the region of the local minimum.

Fig. 2 shows a general block diagram of the newly developed method for decomposing Raman spectra, which consists of two main stages: iterative selection of line composition for each Raman spectrum and searching for a general model that satisfies all analyzed spectra.

Optional, the preparatory stage is dividing the spectrum into independent spectral regions, which allows reducing the number of necessary operations and optimizing the algorithm in terms of execution time. An example of dividing the spectrum into regions is shown in Fig. 3. Such a delimitation of regions is possible only under the condition that the spectral regions are independent, that is, a peak at the boundary of one region does not affect the peaks of another spectral region. Which is applicable in case of using Gaussian profile. The possibility of separation into independent contours with Lorentz or Voigt functions remains questionable, because although they have a significant part of the intensity in the tails, there is no certainty that after baseline correction these tails remained.

When approximating the model and iteratively adding new lines, the algorithm will complete the

calculation faster when processing each part of the spectrum independently than when processing the entire spectrum. When testing the algorithm on the spectra of demineralized and mineralized dentin, the algorithm performed 34 times faster when dividing the spectra into independent spectral contours than when processing the entire Raman spectrum.

The result of the algorithm is a model consisting of a set of elementary lines with ranges of parameter values that describe the shape of these lines. The stages of the algorithm are described in detail below:

1. iterative search for line composition;
2. cluster analysis of models.

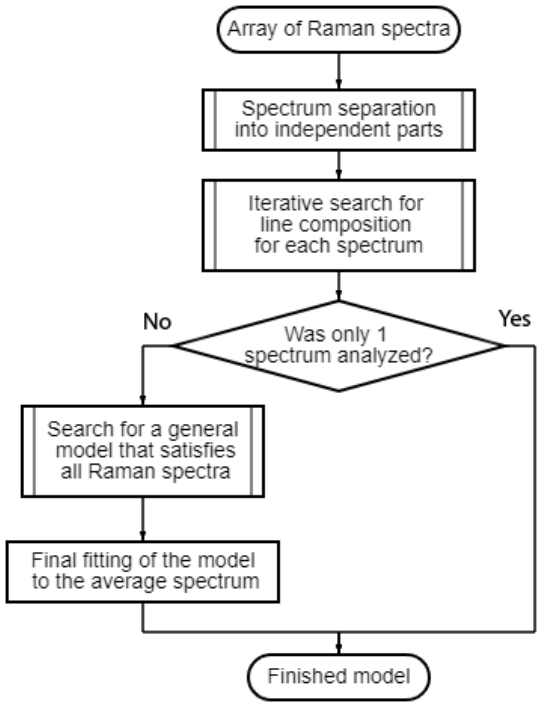


Fig. 2 Block diagram of the proposed spectrum decomposition algorithm.

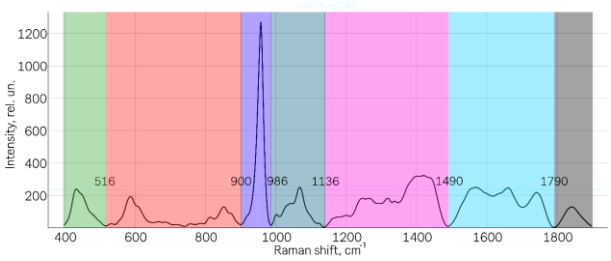


Fig. 3 Splitting the spectrum into 7 independent spectral contours.

3 Description of the Algorithm for Iterative Search for Line Composition

The algorithm is based on the sequential addition of elementary lines and approximation of the resulting model until the addition of new lines significantly improves the model.

The block diagram of the iterative search algorithm for a model describing the spectral contours of the Raman spectrum is shown in Fig. 4.

Let us choose a Gaussian as an elementary line. The optimization method is the least squares method. Lorentzian, PreudoVoigt, or Voigt can also be used as an elementary line.

One of the input parameters is the maximal possible width (HWHM) for all peaks. This eliminates the disadvantage inherent in many deconvolution methods, when the result of adding a too broad line was beyond the original spectral contour and differed from the physically based line composition and distribution.

In practical terms, when performing spectral decomposition or fitting Raman spectra, it is common to allow the peak widths to vary rather than constraining them to be identical. This flexibility often leads to more accurate and meaningful fits.

The next important feature of the proposed algorithm is that if, during the process of adding, a new line ends up in the place of one already added at previous iterations, then this new line is skipped with the area reset to zero at ± ½ HWHM from the position of the peak.

An example of the final result of selecting a model that describes the spectral contour is shown in Fig. 5. The process of adding new elementary lines stops when the intensity of the new peak is less than the threshold. An intensity less than this threshold is considered noise.

On different spectrometers, working with different media and laser radiation sources, the contour and width of the Raman line can change significantly, and the noise level can vary greatly, which must be taken into account when processing spectra.

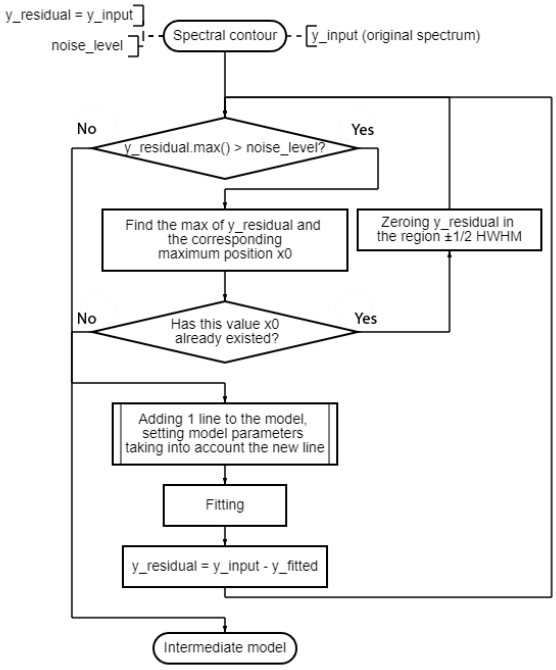


Fig. 4 Block diagram of the iterative algorithm for selecting a model describing the spectral contour.

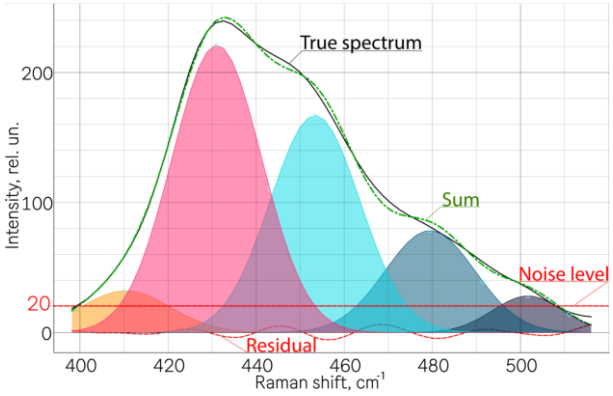


Fig. 5 Final result of adding lines for the independent spectral contour with intensity threshold equaled to 20.0.

In total, the spectral contour (*x_input*, *y_input*) and parameters set by the user are used as input parameters:

1) model optimization method; the proposed decomposition method is not tied to a specific optimization method. It can be selected one of the following methods:

– Least-Squares, Trust Region Reflective method [15];
– Levenberg-Marquardt [16, 17];
– Nelder-Mead [18];
– Limited-memory BFGS - L-BFGS-B [19];
– Powell [20];
– Conjugate-Gradient [21];
– Constrained optimization by linear approximation Cobyla;
– BFGS [22];
– Truncated Newton [23];
– other methods;

2) half width at half maximum (HWHM) *max_dx*;

3) the threshold value of the peak intensity *noise_level*, upon reaching which the addition of new lines stops (can be calculated from spectra based on noise);

4) optional for Voigt and Pseudo-Voigt lines. Limitation of the maximum share of the Lorentz distribution with respect to the Gaussian distribution.

Further "Least-Squares, Trust Region Reflective" method was used with *max_dx* equaled to 12.0, and *noise_level* equaled to 0.0074493.

In the cycle, new lines are added to the model, the model is optimized, and the remainder *y_residual* is updated. If, when the algorithm starts, lines have already been added by the user, then the analysis takes place taking into account these initial parameters. The number of lines and line parameters are unknown in advance.

A number of restrictions are imposed on the line parameters. The peak intensity *a* is not greater than the maximum value in the interval $x_0 \pm$ HWHM and not less than 0.

HWHM *dx*, limited above by user choice *max_dx*, below by line half-width at 0 cm^{-1} (at the exciting laser wavelength).

Fluctuations in the position of the band position x_0 are calculated using the following Eq.:

$$\Delta x_0 = \frac{min\_fwhm}{4} + 1. \quad (1)$$

These restrictions do not allow the line to go beyond the boundaries of the original spectral contour. Also, if the determination of the added peak position occurs again, the area adjacent to this coordinate is reset to zero to prevent an infinite loop.

The result of the analysis by the algorithm is a sum-of-lines model that describes each spectral region. Since for different Raman spectra the resulting line composition may differ and analysis is required to find a model suitable for all analyzed Raman spectra.

If there is a priori information about the parameters of the spectral line, then the method will supplement this information by adding new lines.

4 Algorithm for Analyzing Models after Iterative Search for Line Composition

The result of the analysis at the first stage is a set of peaks coordinates for each spectral region of each analyzed spectrum. Variants of decomposition models are converted into arrays of x_0 coordinates of peaks positions.

At this stage, the total number of lines, the coordinates of the cluster center and the boundaries of coordinate fluctuations are still unknown.

To form the final model from many models, an algorithm is used, the block diagram of which is presented in Fig. 6.

The first operation is clustering using randomly selected two spectra as the first clusters.

The block diagram of the clustering algorithm is shown in Fig. 7. The input of the algorithm is an array of coordinates x_0 of the peak's positions $x_0\_lines = [x_1, x_2, x_3, \ldots x_n]$, the maximal threshold value of the half-width of the HWHM line. The clustering function is called *N* times (number of Raman spectra).

Initial clusters are formed based on the first two arrays x_1 and x_2 using the DBSCAN clustering method [24] with parameters ε equaled to hwhm, *min_samples* = 2.

Next, the points of the remaining spectra are added to these initial clusters closest to each point. If for a point the distances to all cluster centers exceed HWHM, then this point forms a new cluster. The cycle ends after adding points from all sets x_i.

As many options as possible should be used as sets for forming initial clusters. Since with this clustering algorithm it is necessary to take into account that if the first clusters are formed on the basis of two deviant sets of coordinates x_0, then the number of clusters as a result may change, as well as the position of certain peak.

Therefore, the clustering procedure is repeated *N* times and each time the first 2 sets of data are selected randomly to determine the initial clusters to which points will be added. Then 86.46% of the entire sample participates in the formation of the first clusters.

The resulting set of average cluster values is a set of size *N* and contains possible variants of the final model. Subsequent steps are shown in Fig. 5, after the first cycle.

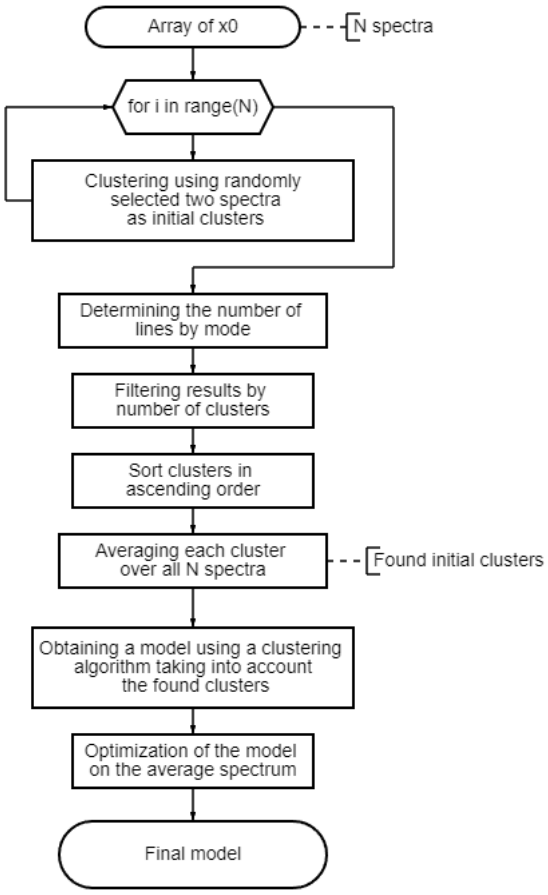


Fig. 6 Block diagram of the algorithm for finding the decomposition model of Raman spectra.

Next comes the determination of the number of lines by mode.

All results are filtered by this number of lines. The remaining arrays are sorted in ascending order.

Next, all the values of each cluster are averaged to obtain the initial positions of the line.

The resulting set is again fed into the clustering algorithm (Fig. 6) as initial clusters, without their formation based on those randomly selected using using Density-based spatial clustering of applications with noise (DBSCAN) algorithm.

The resulting averages of the clusters are the center of the coordinate region of the Raman line, and the standard deviation of all cluster points is the deviation of the center x_0. Based on these data (number of lines, coordinate positions of peak), the final optimization of the model is carried out on the average Raman spectrum. And the result obtained is a ready-made model for all analyzed Raman spectra.

The second stage algorithm essentially analyzes the distribution density of all possible variants of the positions of spectral lines in the Raman spectrum, determined at the first stage, and returns the final model as a set of elementary lines.

The described algorithm deals with these issues by employing methods that allow for flexible peak widths and iterative addition of elementary lines until a satisfactory model is achieved. This flexibility helps in accurately modeling broad features and overlapping bands commonly found in spectra of amorphous materials and combination bands.

Fig. 8 shows the distribution density of the found coordinates of peak positions for each spectral region based on the analysis of 344 dentin Raman spectra and the found positions x_0 with standard deviation Δx_0. It can be seen that certain positions of the elementary lines coincide with the position of the distribution density and are located in the locations of the local maxima of the spectral contour.

Fig. 8 shows a decomposition model consisting of 87 Gaussians optimized for an averaged Raman spectrum. To interpret the spectrum in Fig. 9, we should consider the presence of regions that might originate from amorphous material, which are often very broad. Amorphous materials typically produce broad peaks in a Raman spectrum due to their lack of long-range order, which results in a wide range of vibrational environments for the scattering centers.

In Raman spectra, combination bands may also appear, which can have different inherent widths compared to the primary vibrational modes. These combination bands result from the simultaneous excitation of two or more vibrational modes and often appear at frequencies corresponding to the sum or difference of the individual mode frequencies. They can be broader due to the combined broadening effects of the individual modes.

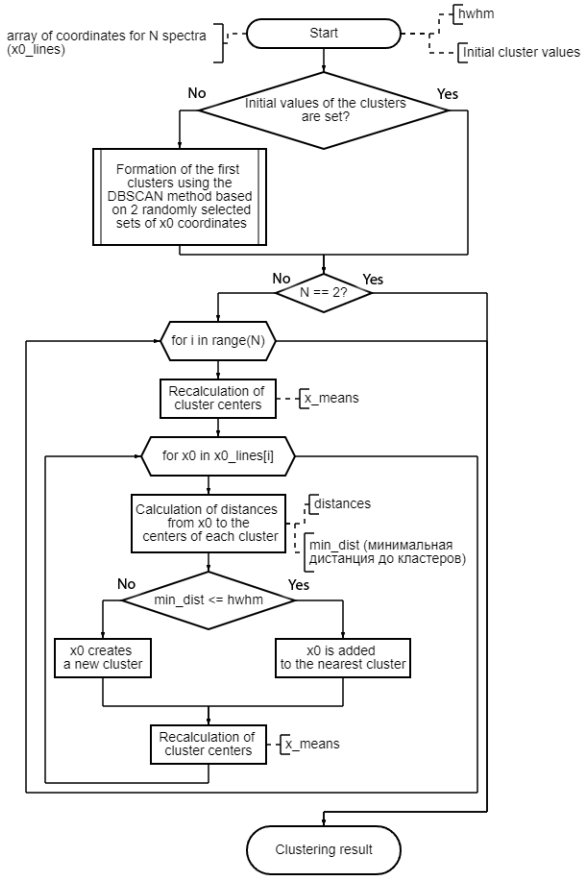


Fig. 7 Block diagram of the clustering algorithm.

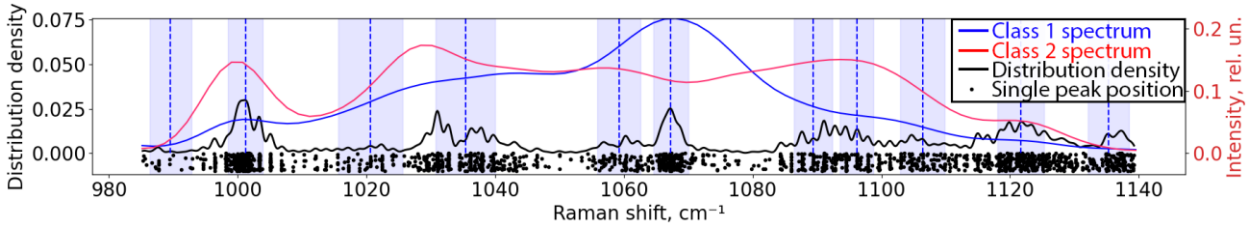
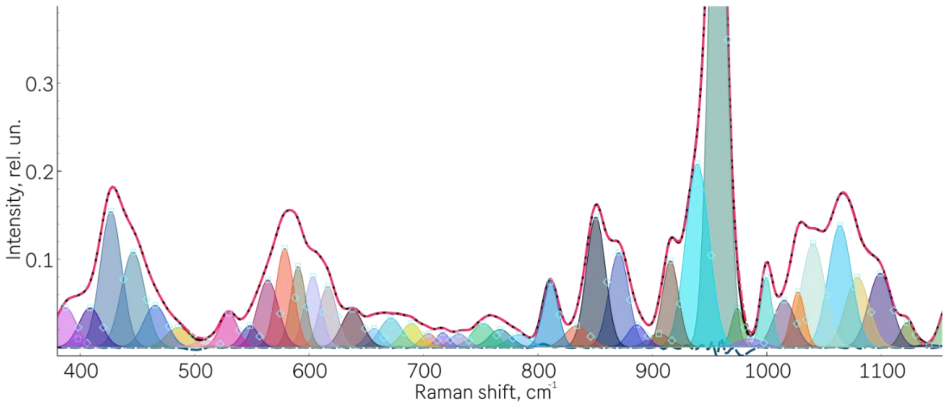
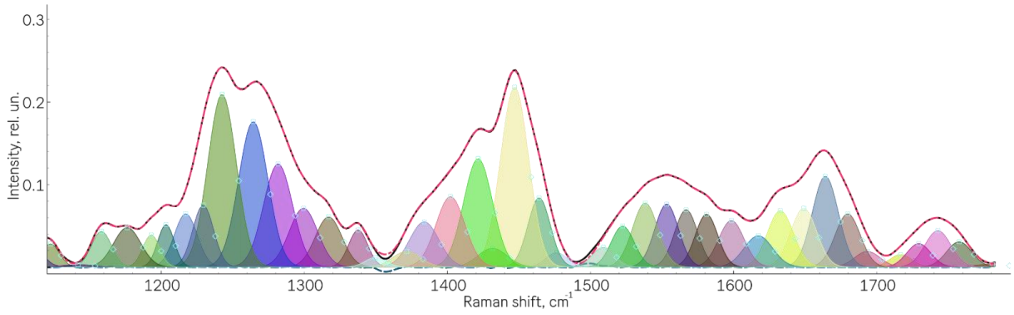


Fig. 8 Distribution density of the found peak positions for each spectral region of 344 dentin Raman spectra.



(a)



(b)

Fig. 9 Final model of Raman spectrum decomposition. (Black solid line is the original spectrum, red dotted line is the simulated spectrum, blue dotted line is the difference between the original and simulated spectrum). (a) Spectrum in the range 400−1150 cm^{-1}, (b) the same spectrum in the range 1150−1800 cm^{-1}.

The model optimization error was assessed using the indicators R^2 (Eq. (2)), χ^2 (Eq. (3)), reduced χ^2 (Eq. (4)), Akaike Information Criterion (*aic*), and Bayesian Information Criterion (*bic*) (Eq. (5)):

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad (2)$$

$$\chi^2 = \sum_i^N [r_i]^2, \quad (3)$$

$$\chi_v^2 = \frac{\chi^2}{N - N_{\text{varys}}}, \quad (4)$$

$$aic = N \ln\left(\frac{\chi^2}{N}\right) + 2N_{varys}, \quad (5)$$

$$bic = N \ln\left(\frac{\chi^2}{N}\right) + \ln(N)\, N_{varys},$$

where y – original spectrum, f – spectrum obtained using the model by summing a set of lines, r – difference between y and f, N – number of points in the spectrum, N_{Varys} – number of model parameters.

We used the following criteria values for the average spectrum: R_2 was 99.995%, χ_2 was $7.3 \cdot 10^{-5}$, red χ_2 was $1 \cdot 10^{-6}$, *aic* was *−1799, bic* was −1693.

The average value of these criteria for 344 dentin Raman spectra were as follows: R^2 was 99.49%, χ^2 was $1.7 \cdot 10^{-3}$, red χ^2 was $2.2 \cdot 10^{-5}$, *aic* was −1367, *bic* was −1267.

Let us estimate the value of the overlap coefficient (*OVC*) of the Raman lines as the ratio of the area of overlap of one line by all others to the area of this line:

$$OVC = \frac{S_{overlap}}{S_{line}}\ 100\ \%. \quad (6)$$

The median value of the overlap coefficient across all lines in the above model was 79%.

5 Model Experiment, Assessment of the Influence of Noise on the Quality of Decomposition

To test the influence of random Gaussian noise on the result of spectrum decomposition, 72 synthetic Raman spectra with noise superposition were used. The spectrum with the line composition shown in Table 1 and Fig. 10 was used as a model.

A sample of 72 spectra was formed by summing the peak intensities with the parameters from Table 1, so that the peak intensity varied within ±1% randomly to simulate the sample of Raman spectra of bone tissue. The following is a comparison of the results of decomposition of Raman spectra with the superimposition of noise of various intensities for modeling the signal-to-noise ratio (SNR) from 1000 to 1.

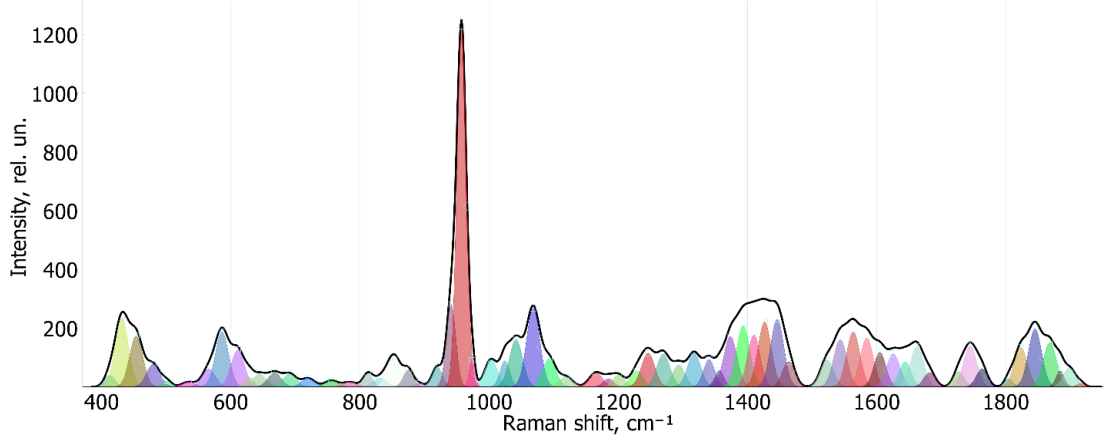


Fig. 10 Synthesized Raman spectrum.

Table 1 Line composition of the simulated Raman spectrum.

Peak position x_0, cm^{-1}	Peak intensity a, rel. units	HWHM dx, cm^{-1}	Peak position x_0, cm^{-1}	Peak intensity a, rel. units	HWHM dx, cm^{-1}
401.0	33.2	7.9	1210.9	20.5	11.8
429.0	220.4	11.4	1229.5	40.9	10.1
449.0	173.2	12.0	1243.2	102.9	12.0
463.3	44.1	10.1	1261.8	75.5	12.0
477.7	63.8	12.0	1274.5	63.8	11.5
495.7	26.0	12.0	1294.4	72.1	12.0
532.9	17.6	9.3	1318.2	114.3	12.0
555.4	31.7	8.3	1338.2	67.8	8.4
570.1	48.3	9.6	1355.0	83.8	12.0
584.6	162.6	12.0	1375.2	168.2	12.0
596.3	57.2	10.1	1394.1	201.5	12.0
610.9	108.4	9.9	1411.2	183.6	12.0
626.8	54.1	9.6	1427.3	198.3	12.0
643.8	36.3	9.7	1445.4	220.7	12.0
663.8	48.2	12.0	1462.5	90.8	11.8
687.0	44.3	12.0	1517.5	44.1	12.0
700.4	6.8	7.1	1531.5	100.7	12.0
714.8	24.8	12.0	1550.1	170.7	12.0
731.3	12.0	11.0	1568.9	163.1	12.0
755.6	22.1	8.9	1586.6	132.0	12.0
769.8	9.5	6.4	1605.4	114.6	12.0
782.8	15.7	7.3	1626.3	98.7	12.0
791.5	6.5	5.2	1644.1	86.9	12.0
811.6	45.3	9.5	1661.6	119.0	12.0

Table 1 Cont.

Peak position x_0, cm^{-1}	Peak intensity a, rel. units	HWHM dx, cm^{-1}	Peak position x_0, cm^{-1}	Peak intensity a, rel. units	HWHM dx, cm^{-1}
832.3	29.8	12.0	1679.0	54.4	12.0
852.5	106.3	12.0	1730.4	62.8	11.7
874.1	55.3	8.5	1744.1	99.6	12.0
888.4	13.6	9.2	1759.5	73.1	12.0
920.8	66.8	10.2	1810.1	51.2	11.5
944.5	364.0	10.0	1827.2	120.8	12.0
957.8	1140.4	7.9	1843.9	148.5	12.0
972.2	117.1	5.9	1860.5	119.0	12.0
1003.0	95.6	11.7	1876.8	87.5	12.0
1018.4	40.4	5.6	1892.0	50.0	12.0
1026.9	72.0	6.9	1907.4	26.5	12.0
1040.8	156.7	12.0			
1066.2	233.5	12.0			
1079.7	78.5	11.2			
1095.1	67.0	9.4			
1107.8	32.2	8.3			
1119.1	19.7	6.1			
1126.7	20.1	6.1			
1158.1	27.5	9.0			
1171.4	32.2	11.1			
1185.5	25.3	12.0			
1198.7	24.7	9.3			

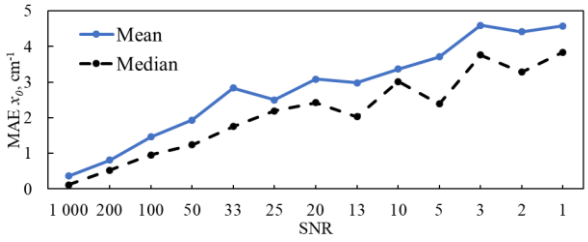


Fig. 11 Error in determining x_0 from the real one.

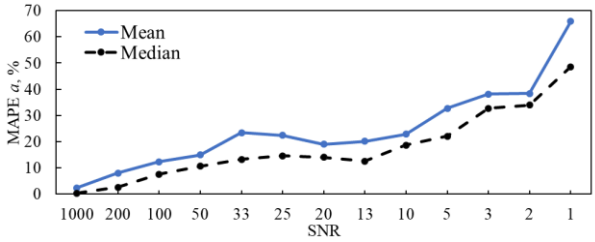


Fig. 12 Error in determining peak intensity.

Fig. 11 shows a graph of the mean absolute error (MAE) of the deviation of the peak position x_0 from the actual one.

For an SNR of ~1000, the algorithm determined the value of x_0 with an average absolute error of 0.11 cm^{-1}. As the noise level increases, the error in determining the line position x_0 also increases.

Fig. 12 shows the mean absolute percentage error (MAPE). The average absolute error in determining the dx half-width of the lines was similarly estimated (Fig. 13).

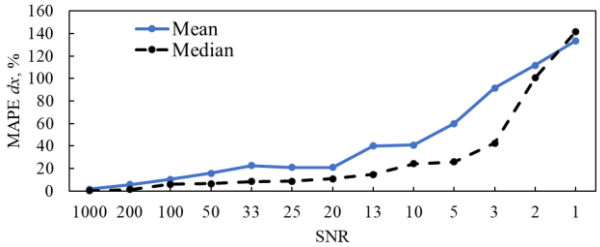


Fig. 13 Error in determining half-width of lines dx.

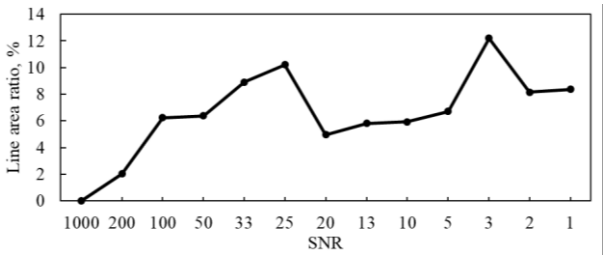


Fig. 14 Ratio of areas of erroneously determined lines to actual ones.

For SNR ~1000, the peak intensity determination error was 0.3%. The error in determining the half-width *dx* was 0.29%.

To assess the quality of the algorithm in terms of minimizing the number of erroneously identified lines, we compare the ratio of the area of erroneously identified lines to the area of valid lines (Fig. 14) using Eq. (7):

$$\frac{S_{noise}}{S} * 100, \quad (7)$$

where S_{noise} is the area of lines erroneously created by the algorithm, approximating noise, S is the area of actually present lines in the spectrum.

The proposed algorithm for decomposing lines of the CR makes it possible to obtain a correct result that corresponds to the actual parameters of the model. The error in determining the peak intensity does not exceed 15% up to an SNR of ~13. The error in determining the peak position x_0 does not exceed 1 cm^{-1} up to an SNR of ~100, and the error in determining the half-width of the lines dx does not exceed 10% up to an SNR of ~25.

Evaluation of the signal-to-noise ratio on real data gave a result of ~340 units. before the smoothing procedure. After filtering the data, the SNR exceeds 1000 units.

6 Conclusion

Two-stage method implementing a novel approach has been developed.

During the model generation stage, a set of models is generated for each spectral contour, providing information about the number and parameters of elementary lines. This stage involves iterative selection and refinement of line compositions for each Raman spectrum.

During the cluster analysis stage, the composition of models is analyzed using cluster analysis to form a final model that describes all the analyzed Raman spectra. This approach allows for the compensation of errors in Raman band positions and other parameters, effectively handling asymmetric lines.

The performance of our method was evaluated on a test sample of 344 Raman spectra. Key metrics include R^2 at 99.49 %, χ^2 at $1.7 \cdot 10^{-3}$, red χ^2 at $2.2 \cdot 10^{-5}$, *aic* at −1367, and *bic* at −1267.

Furthermore, classifiers trained on the peak intensities of the decomposed spectra exhibited classification abilities equal to those trained on full spectra, indicating no loss of significant information during feature space reduction and removal of linearly dependent features. The new method not only retains significant spectral information but also enhances classifier performance by reducing feature space complexity without compromising classification accuracy.

The method allows you to compensate for the error in the position of the Raman bands and other parameters, and work with asymmetric lines. By determining the final model based on the composition of many individual models, it allows for correct detection of lines with a signal-to-noise ratio above 200.

The error in determining the parameters of Raman lines from Raman spectra with a signal/noise ratio of 1000, corresponding to real smoothed Raman spectra, was 0.3% for peak intensity, 0.3% for half-width *dx* and 0.11 cm^{-1} for the peak position x_0.

The source code of the method is provided at the links: https://github.com/DarkMatro/Automatic-decomposition-algorithm-for-raman-spectra and https://github.com/DarkMatro/RS-tool.

Disclosures

The authors declare that they have no conflict of interest.

References

1. F. Alsmeyer, W. Marquardt, "Automatic Generation of Peak-Shaped Models," Applied Spectroscopy 58(8), 986–994 (2004).
2. F. Alsmeyer, H.-J. Koß, and W. Marquardt, "Indirect Spectral Hard Modeling for the Analysis of Reactive and Interacting Mixtures," Applied Spectroscopy 58(8), 975–985 (2004).
3. R. J. Meier, "On art and science in curve-fitting vibrational spectra," Vibrational Spectroscopy 39(2), 266–269 (2005).
4. Y. Chen, L. Dai "Automated decomposition algorithm for Raman spectra based on a Voigt line profile model," Applied Optics 55(15), 4085–4094 (2016).
5. Z. Mou-Yan, R. Unbehauen, "A deconvolution method for spectroscopy," Measurement Science and Technology 6(5), 482–487 (1995).
6. H. Liu, S. Liu, Z. Zhang, J. Sun, and J. Shu, "Adaptive total variation-based spectral deconvolution with the split Bregman method," Applied Optics 53(35), 8240–8248 (2014).
7. H. Liu, T. Zhang, L. Yan, H. Fang, and Y. Chang, "A MAP-based algorithm for spectroscopic semi-blind deconvolution," Analyst 137(16), 3862–3873 (2012).

8. G. A. Kupriyanov, I. V. Isaev, I. V. Plastinin, T. A. Dolenko, and S. A. Dolenko, "Decomposition of Spectral Contour into Gaussian Bands using Gender Genetic Algorithm," Moscow University Physics Bulletin 78(S1), S236–S242 (2023).
9. Q. Zhou, Z. Zou, and L. Han, "Deep Learning-Based Spectrum Reconstruction Method for Raman Spectroscopy," Coatings 12(8), 1229 (2022).
10. X. Bian, Z. Shi, Y. Shao, Y. Chu, and X. Tan, "Variational Mode Decomposition for Raman Spectral Denoising," Molecules 28(17), 6406 (2023).
11. N. Schmid, S. Bruderer, F. Paruzzo, G. Fischetti, G. Toscano, D. Graf, M. Fey, A. Henrici, V. Ziebart, B. Heitmann, H. Grabner, J. D. Wegner, R. K. O. Sigel, and D. Wilhelm, "Deconvolution of 1D NMR spectra: A deep learning-based approach," Journal of Magnetic Resonance 347, 107357 (2023).
12. S. Kaneki, Y. Kouketsu, M. Aoya, Y. Nakamura, S. R. Wallis, Y. Shimura, and K. Yamaoka, "An automatic peak deconvolution code for Raman spectra of carbonaceous material and a revised geothermometer for intermediate- to moderately high-grade metamorphism," Progress in Earth and Planetary Science 11(1), 35 (2024).
13. R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra," Applied Spectroscopy 43(5), 772–777 (1989).
14. S. J. Barton, T. E. Ward, and B. M. Hennelly, "Algorithm for optimal denoising of Raman spectra," Analytical Methods 10(30), 3759–3769 (2018).
15. M. A. Branch, T. F. Coleman, and Y. Li, "A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems," SIAM Journal on Scientific Computing 21(1), 1–23 (1999).
16. K. Levenberg, "A Method for the Solution of Certain Non-Linear Problems in Least Squares," Quarterly of Applied Mathematics 2(2), 164–168 (1944).
17. D. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," Journal of the Society for Industrial and Applied Mathematics 11(2), 431–441 (1963).
18. J. A. Nelder, R. Mead, "A simple method for function minimization," Computer Journal 7(4), 308–313 (1965).
19. C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization," ACM Transactions on Mathematical Software 23(4), 550–560 (1997).
20. M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," Computer Journal 7(2), 155–162 (1964).
21. M. R. Hestenes, E. Stiefel, "Methods of Conjugate Gradients for Solving Linear Systems," Journal of Research of the National Bureau of Standards 49(6), 409 (1952).
22. R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A Limited Memory Algorithm for Bound Constrained Optimization," SIAM Journal on Scientific Computing 16(5), 1190–1208 (1995).
23. R. S. Dembo, S. C. Eisenstat, and T. Steihaug, "Inexact newton methods," SIAM Journal on Numerical Analysis 19(2), 400–408 (1982).
24. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining 226–231 (1996).