


RESEARCH ARTICLE

# Vector casting for noise reduction

Medhanie Tesfay Gebrekidan<sup>1,3</sup>  | Christian Knipfer<sup>3,4</sup> |  
Andreas Siegfried Braeuer<sup>1,2</sup>

<sup>1</sup>Institute of Thermal, Environmental, and Resources' Process Engineering (ITUN), Technische Universität Bergakademie Freiberg (TUBAF), Freiberg, Germany

<sup>2</sup>Zentrum für effiziente Hochtemperatur-Stoffwandlung (ZeHS), Technische Universität Bergakademie Freiberg (TUBAF), Freiberg, Germany

<sup>3</sup>Erlangen Graduate School in Advanced Optical Technologies (SAOT), Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany

<sup>4</sup>Department of Oral and Maxillofacial Surgery, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

## Correspondence

Andreas Siegfried Braeuer, Institute of Thermal, Environmental, and Resources' Process Engineering (ITUN), Technische Universität Bergakademie Freiberg (TUBAF), Leipziger Straße 28, 09599 Freiberg, Germany.  
Email: andreas.braeuer@tu-freiberg.de

## Funding information

European Union's Horizon 2020 research and innovation programme under ERC Starting Grant agreement, Grant/Award Number: 637654 (Inhomogeneities); Wilhelm Sander-Stiftung, Grant/Award Number: 2017.111.1

## Abstract

We report a new method for the reduction of noise from spectra. This method is based on casting vectors from one data point to the following data points of the noisy spectrum. The noise-reduced spectrum is computed from the casted vectors within a margin that is identified by an envelope-finder algorithm. We compared here the presented method with the Savitzky–Golay and the wavelet transform approaches for noise reduction using simulated Raman spectra of various signal-to-noise ratios between 1 and 25 dB and experimentally acquired Raman spectra. The method presented here performs well compared with the Savitzky–Golay and the wavelets-based denoising method, especially at small signal-to-noise ratios and furthermore relies on a minimum of human input requirements.

## KEYWORDS

envelope detection, noise reduction, peak detection, Raman spectra, vector casting

## 1 | INTRODUCTION

Spectral analysis involves processing of spectroscopic data or patterns for quantification and/or identification of samples or processes.<sup>[1–7]</sup> The spectroscopic raw data usually contain contributions originating from the desired signal itself, the noise and from the

background or interferences from undesired signals.<sup>[8]</sup> One of the first processing steps (often the first one) of spectroscopic raw data is the elimination of the noise or the reduction of the noise level. This is especially challenging when the signal-to-noise ratio (SNR) is small, meaning when the differentiation between noise and signal based on solely intensity or peak height is

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. Journal of Raman Spectroscopy published by John Wiley & Sons Ltd

not straightforward. Existing noise reduction algorithms can reduce the noise level on the one hand, but on the other hand—especially at small SNRs—can also manipulate and with this falsify the desired signal contribution.

The origin of small SNRs can be manifold. Small available or realizable excitation powers<sup>[9]</sup> in combination with small interaction probabilities (cross sections) between the excitation and the matter under investigation often result in small SNRs.<sup>[10–12]</sup> Also, a less efficient signal detection or short acceptable signal integration time can result in small SNRs.<sup>[13,14]</sup> On the contrary, even the long integration of low signal levels can lead to small SNRs, as together with the signal also the thermal background together with its thermal noise is accumulated. The ineluctable contamination of spectroscopic data with noise therefore limits the performance of spectroscopic techniques.<sup>[15–17]</sup>

Many postprocessing techniques have been used to denoise spectroscopic data, such as the Savitzky–Golay (SG) filter,<sup>[18,19]</sup> smoothing based on the wavelet transform method,<sup>[20–22]</sup> the “perfect smoother” method,<sup>[23]</sup> the finite impulse response (FIR) smoother,<sup>[24]</sup> and smoothing based on the “Wiener estimation.”<sup>[25]</sup> The SG smoother is the most popular and frequently used method to denoise spectroscopic data.<sup>[8,25,26]</sup> It is based on the least-squares fitting of polynomials of specified order to connected data points contained in a moving window of specified size. The larger the size of the window is chosen, meaning the more data points are considered for the polynomial fit, the more the raw spectral data are smoothed.<sup>[24]</sup> Not only the noise but also the sharp signal features can potentially be smoothed out, like it is the case for all smoothing algorithms. Thus, a compromise needs to be made between smoothing out the noise and a loss of spectral information by carefully adjusting the window size and polynomial order of the filter.

Smoothing based on wavelets is simple to use, while adapting well to the form of the signal being smoothed.<sup>[22,27]</sup> Here, the noisy raw spectral data are transformed into a wavelet domain by decomposing it into a set of orthonormal wavelet basis functions. The major signal trends of the spectrum are assignable to large wavelet coefficients, whereas the noise is assignable to only small coefficients.<sup>[20,28]</sup> Hereupon, the noise is suppressed by thresholding the wavelet coefficients. Then the not-suppressed coefficients are reverse transformed to obtain the noise-reduced spectra. However, the selection of the wavelet basis functions and the threshold value have a great impact on the performance of the method and are strongly problem dependent.<sup>[25]</sup> Moreover, the application of this approach to spectra with small

SNR can reduce, remove, or manipulate also signal contributions.<sup>[29]</sup>

Člupek et al.<sup>[24]</sup> tested the FIR smoother to suppress noise in spectra. They reported that this technique offers better preservation of the real signal contribution compared with the SG smoother. However, it is demanding in computation.<sup>[24,25]</sup> Using the “Wiener estimation,” Chen et al.<sup>[25]</sup> developed a method on the basis of spectral reconstruction to recover spectra with small SNR. In comparison with other denoising methods such as the SG method, the FIR smoother, and the wavelet transform method, their method showed excellent performance. However, a calibration data set that relies on input spectra with large SNRs is required for the successful denoising of spectra with small SNR.

We here introduce a vector casting method for noise reduction. We compared its performance with the frequently used SG and wavelets denoising methods. The performance comparison considers the extractability of the real signal contribution. To the best of our knowledge, vector casting has never been applied to denoise spectra.

## 2 | MATERIAL AND METHODS

### 2.1 | Samples

We used two sets of samples to validate the vector casting method. The two sets comprise simulated Raman spectra and experimentally acquired Raman spectra. At this point, it should be underlined that the vector casting method is not limited to the treatment of Raman spectra. Therefore, the descriptions provided in the sections that follow are provided in a general context and can be transferred to any kind of spectral data. We only consider contributions to the spectroscopic data coming from the real signal and from the noise. We neglect the potentially occurring contributions of a background, as the background is usually subtracted from the spectroscopic data using baseline correction methods.<sup>[11,30,31]</sup> These baseline correction methods can still be applied after the noise reduction method.

The simulated spectra

$$R(x_i) = S_{\text{sig}}(x_i) + N_{\text{sim}}(x_i) \quad (1)$$

are the summation of a pure signal spectrum  $S_{\text{sig}}(x_i)$  and a noise spectrum  $N_{\text{sim}}(x_i)$ , where  $x_i$  is the variable (Raman shift in the case of Raman spectra, wavelength in the case of fluorescence spectra, wavenumber in the case of absorption spectra, temperature in the case of differential scanning calorimetry spectra, theta in the case

of X-ray diffraction spectra, etc.). Due to Doppler broadening, collisional broadening and optical effects spectral lines, peaks, or bands are never strictly monochromatic but feature a distribution around their centre.<sup>[32]</sup> Spectral signal profiles can be fitted by Lorentzian, Gaussian, or Voigt profiles.<sup>[33]</sup>

The simulated signal spectrum

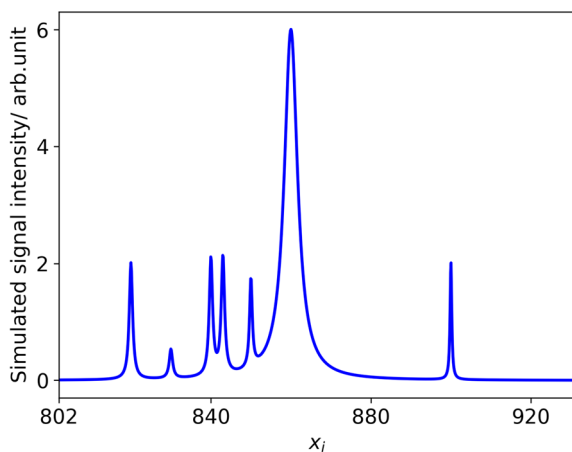
$$S_{sig}(x_i) = \sum_{n=1}^7 A_n \left[ 1 + \left( \frac{x_i - x_n}{\sigma_n} \right)^2 \right]^{-1} \quad (2)$$

is the summation of seven Lorentzian peaks having different amplitudes  $A_n$ , widths  $\sigma_n$ , and being centered at different variables  $x_n$ . We chose Lorentzian peaks as they best reflect theoretical Raman signal lines.<sup>[34]</sup> The usage of other peak shapes or a different number of peaks would not influence the vector casting method.

Figure 1 shows the simulated spectrum  $S_{sig}(x_i)$ . The number seven of Lorentzian peaks and the parameters of these peaks were chosen to imitate overlapping peaks ( $x_n = 840$ ), narrow peaks ( $x_n = 848, x_n = 900$ ), small peaks ( $x_n = 830$ ), and broad peaks ( $x_n = 820, x_n = 860$ ).

Noise in spectroscopic data acquired with a non-intensified charge coupled device consists of Poisson noise (shot and thermal noise) and Gaussian noise (read-out noise). However, above certain noise levels, Gaussian noise is a good approximation for Poisson noise.<sup>[21,35]</sup> The sum of a mutually independent zero-mean Gaussian noise is still a zero-mean Gaussian noise with variance equal to the sum of the variances of the independent Gaussian noises.<sup>[36]</sup> Thus, spectroscopic noise can be modelled as the summation

$$N_{sim}(x_i) = e(x_i) \sqrt{n_{ph}^2(x_i) + n_{th}^2 + n_{rd}^2} \quad (3)$$



**FIGURE 1** Simulated spectrum (intensity as a function of a variable  $x_i$ ) consisting of seven Lorentzian peaks [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

of shot noise  $n_{ph}$  (also referred to as photon noise), thermal noise  $n_{th}$ , and readout noise  $n_{rd}$ .<sup>[35,37]</sup>  $e(x_i)$  is Gaussian noise having a standard deviation of one and mean of zero.<sup>[36,38]</sup> The shot noise

$$n_{ph}(x_i) = \sqrt{S_{sig}(x_i)} \quad (4)$$

is the square root of the signal  $S_{sig}(x_i)$  and with this is a function of the variable  $x_i$ . Variables  $x_i$  with large signal feature a large shot noise, whereas variables without signal feature no shot noise.

The thermal noise

$$n_{th} = \sqrt{B} \quad (5)$$

is approximated by the square root of the thermal background  $B$ . The thermal background is supposed to be a constant over  $x_i$ . Also, the read out noise is considered as a constant  $c$  over  $x_i$ .

$$n_{rd} = c \quad (6)$$

The SNR in decibel is computed according to

$$SNR = 10 \log_{10} \left( \frac{\sum_{i=1}^N |S_{sig}(x_i)|^2}{\sum_{i=1}^N |N_{sim}(x_i)|^2} \right) \quad (7)$$

from the integral of the squared signal and the integral of the squared noise.<sup>[39]</sup> The SNR can be changed by changing the constant  $B$  in Equation (5), the constant  $c$  in Equation (6), or by scaling the signal  $S_{sig}(x_i)$  in Equation (4).

For the acquisition of the experimental Raman spectra,<sup>[11]</sup> we used as excitation source a diode laser (Toptica DLpro) emitting 785-nm radiation and a spectrometer (Ventana from Ocean Optics) for signal detection between 800 and 940 nm, which corresponds to Raman shifts between 200 and 2,000  $\text{cm}^{-1}$ . With an excitation laser power of 10 mW, we collected Raman spectra of liquid ethanol at various integration times between 20 and 1,000 ms. From the different integration times, experimental spectra  $R(x_i)$  with various SNRs resulted. Also, the experimentally acquired spectra are composed of a signal and a noise contribution. Additionally, a quasi-noise-free (low-noise) Raman spectrum of ethanol was acquired with an excitation power of 300 mW and 1,000 ms of signal integration time. This quasi-noise-free spectrum can be considered as a reference spectrum or as a quasi-pure signal spectrum  $S_{sig}(x_i)$ . We chose ethanol for the acquisition of the experimental spectra, as the

Raman spectrum of ethanol also shows narrow, broad, and overlapping peaks.

The noise-reduced spectrum  $r(x_i)$  results after noise reduction from either the simulated or the experimental spectra  $R(x_i)$  with either the vector casting method, the envelope-finder method, the SG method, or the wavelet transform method. In order to compare the performance of the different noise reduction methods, we quantified the deviation between the real signal spectrum  $S_{sig}(x_i)$  and the noise-reduced spectra  $r(x_i)$  derived with the different methods according to Equations (8a) to (8d) as proposed by Barton et al.<sup>[35]</sup> These equations quantify (a) the mean improvement of the signal quality across the entire spectral range (Equation (8a)), (b) monitor whether or not the algorithm interacted negatively with the spectral peaks (Equation (8c)), and (c) quantify the signal-to-noise improvement (Equation (8d)) relative to SNR of the original noisy spectrum ( $R(x_i)$ ).

$$SNR_{global}(r(x_i)) = \frac{\max(r(x_i))}{RMSE(r(x_i), S_{sig}(x_i))}, \quad (8a)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (r(x_i) - S_{sig}(x_i))^2}{N}}, \quad (8b)$$

$$SNR_{peak}(r(x_i)) = \frac{\max(r(x_i))}{RMSE(r(x_i)_{peak}, S_{sig}(x_i)_{peak})}, \quad (8c)$$

$$SNR_{prod}(r(x_i)) = \frac{SNR_{global}(r(x_i))}{SNR_{global}(R(x_i))} \cdot \frac{SNR_{peak}(r(x_i))}{SNR_{peak}(R(x_i))}. \quad (8d)$$

### 3 | RESULTS AND DISCUSSION

The vector casting method requires preprocessing of the raw spectra. In the first step, the top and bottom envelopes of the noisy spectra have to be identified using an envelope-finder algorithm that is described in detail below. Afterwards, the vectors are casted within the margin of the before identified envelopes for the derivation of the noise-reduced spectrum  $r(x_i)$ . We want to emphasize here that already the envelope-finder algorithm alone provides a significant noise reduction.

#### 3.1 | Envelope-finder algorithm

The envelope-finder algorithm aims at identifying a smooth top  $E_{top}(x_i)$  and a bottom  $E_{bottom}(x_i)$  envelope of the noisy spectrum  $R(x_i)$ . In the first level (Level 1), all data points of  $R(x_i)$  are classified as either peak or valley,

irrespectively of whether the peak is due to noise or due to a real signal. On this account, a forward  $R'_f$  and a backward  $R'_b$  differentiation is made.

$$R'_f(x_i) = R(x_i) - R(x_{i+1}), \quad (9)$$

$$R'_b(x_i) = R(x_i) - R(x_{i-1}). \quad (10)$$

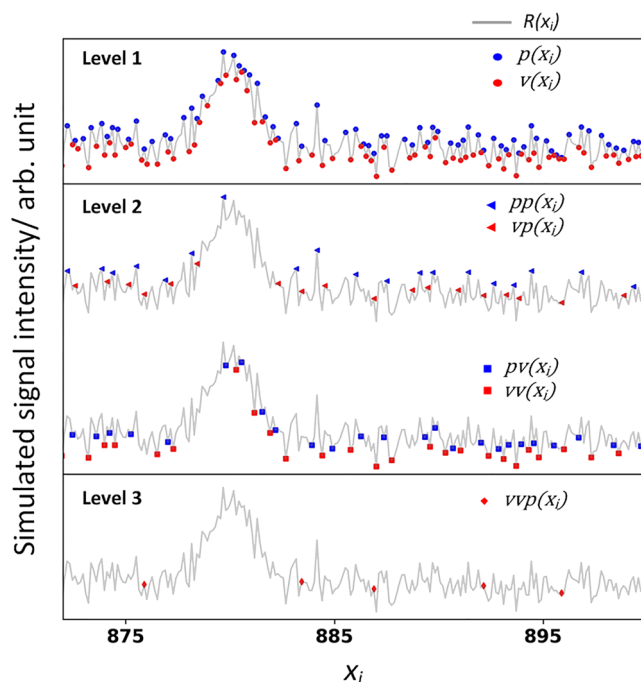
A data point is a peak  $p(x_i)$  or a valley  $v(x_i)$  if its forward and backward differentiation are both positive or negative, respectively.

$$p(x_i) = R(x_i), \text{ if } (R'_b(x_i) > 0) \&\& (R'_f(x_i) > 0), \quad (11)$$

$$v(x_i) = R(x_i), \text{ if } (R'_b(x_i) < 0) \&\& (R'_f(x_i) < 0). \quad (12)$$

Figure 2 top shows the Level 1 peaks and valleys.

In the second level (Level 2), data points of  $p(x_i)$  and  $v(x_i)$  are searched for peaks and valleys by forward and backward differentiation. Figure 2 middle shows these computed peak and valley data points of the peaks and valleys obtained in Level 1. The notations  $pp(x_i)$  and  $vp(x_i)$  indicate the peaks of  $p(x_i)$  and valleys of  $p(x_i)$ , respectively. Similarly,  $p v(x_i)$  and  $v v(x_i)$  means, respectively, peaks of  $v(x_i)$  and valleys of  $v(x_i)$ . Computing the peaks and valleys recursively, in a third level, the peaks



**FIGURE 2** Illustration of cascaded local Extrema search at level 1, level 2, and level 3 [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

and valley of  $pp(x_i)$ ,  $vp(x_i)$ ,  $pv(x_i)$ , and  $vp(x_i)$  can be computed by forward and backward differentiation as well. Figure 2 bottom shows the Level 3 valleys of the Level 2 valleys of the Level 1 peaks, which is referred to as  $vvp(x_i)$ . It also can be seen that the first two red diamonds ( $vvp(x_i)$ ) in Figure 2 bottom can be considered as left and right border of a signal peak.

Figure 3a again shows that these  $vvp(x_i)$  data points (red diamonds) are suitable as indicators for the left and the right border of potential peaks. Figure 3b,c shows upon which criteria a decision is made whether or not a signal peak is situated between two  $vvp(x_i)$  data points, here the left border (lb) and the right border (rb) diamond. The procedure for the classification of peak and peak-free regions is as follows:

### 1. The maximum value of the difference

$$\Delta I_{\max} = \max(|R(x_i) - R(x_{i+1})|) \Big|_w \quad (13)$$

between two consecutive  $vvp(x_i)$  data points (lb and rb) is computed. The window size  $w$  is defined automatically by the  $vvp(x_i)$  and does not have to be chosen by the user. In Figure 3b, the maximum difference  $\Delta I_{\max}$  is shown as green line and can be defined as the maximum deviation between consecutive data points within the window.

### 2. The height $P_h$ of a potential signal peak which in Figure 3b is shown as a blue line

$$P_h = \max(R(x_i) - \text{Baseline}) \Big|_w \quad (14)$$

is computed as the maximum difference between a data point and the baseline within the window. The baseline

is the direct connection between the two  $vvp(x_i)$  data points.

### 3. The window is classified as a peak region $w^p$ if

$$P_h > \Delta I_{\max} \quad (15)$$

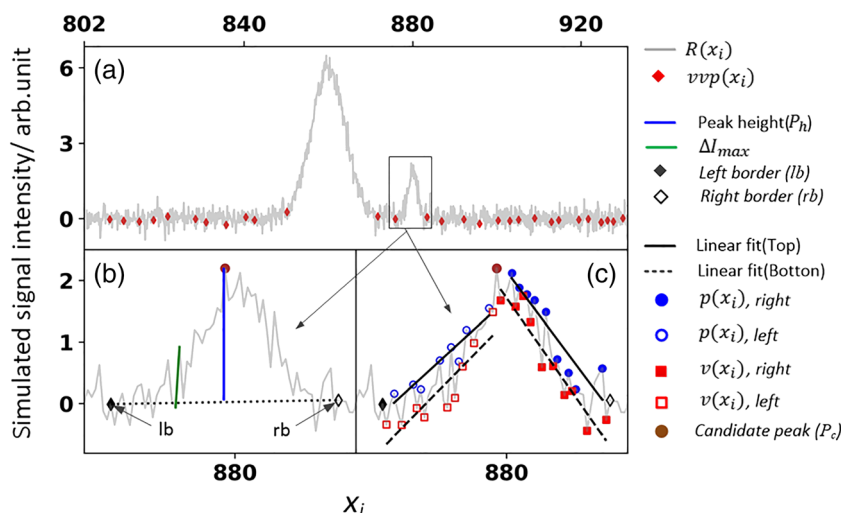
and if according to Figure 3c, the slopes of the linear fits of  $p(x_i)$  and  $v(x_i)$  left of the maximum of the potential peak are both positive and right of the potential peak are both negative.

$$\begin{aligned} \text{left side slopes} &> 0, \\ \text{right side slopes} &< 0. \end{aligned} \quad (16)$$

Otherwise, the window is classified as peak-free  $w^{pf}$  region.

The Level 1 peak  $p(x_i)$  and valley  $v(x_i)$  (Figure 2 top) data points are a good first estimate of the top and bottom envelopes of the noisy signal  $R(x_i)$ . For smoothing these envelopes  $p(x_i)$  and  $v(x_i)$ , we considered a moving window  $w^m$  consisting of  $2n+1$  peak ( $P_i$ ) or valley ( $V_i$ ) data points. Here, the window size has to be set manually. Later, the influence of the chosen moving window size  $w^m$  onto the noise reduction performance will be discussed. Here, we assigned  $P_i$  to the data points of  $p(x_i)$  confined within  $w^m$ , and  $C_i^{\text{top}}$  is one of the data points of  $P_i$ , which is located at center of the  $w^m$ . Similarly,  $V_i$  is assigned to the data points of  $v(x_i)$  included in  $w^m$ , and  $C_i^{\text{bottom}}$  is the central data point of  $V_i$ . The central peak  $C_i^{\text{top}}$  or valley  $C_i^{\text{bottom}}$  data point within the moving window  $w^m$  is updated by a new value. The updating procedure depends on whether  $C_i^{\text{top}}$  or  $C_i^{\text{bottom}}$  is within a peak region  $w^p$  or within a peak-free region  $w^{pf}$ . If they are part

**FIGURE 3** (a) Simulated noisy spectrum (grey line) and detected level 3 minima of maxima of  $R(x_i)$  (red filled diamonds). (b) a window margined between left border (lb) and right border (rb) showing the candidate peak ( $P_c$ ) as brown point, maximum difference between consecutive data points in green line and the peak height ( $P_h$ ) in blue line. (c) Illustration of how linear regression lines are fitted to the level one peak and valley data points within the window margined by lb and rb [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]





of a peak-free region, they are substituted by the moving window  $w^m$  average.

$$C_i^{top} = \frac{\sum P_i}{2n+1}, \text{Top envelope} \quad (17)$$

$$C_i^{bottom} = \frac{\sum V_i}{2n+1}, \text{Bottom envelope}$$

If they are identified as part of peak region  $w^p$ , they are substituted by

$$C_i^{top} = \frac{\sum P_i |P_i - C_i^{top}| \leq \Delta I_{max}}{j}, \text{Top envelope}$$

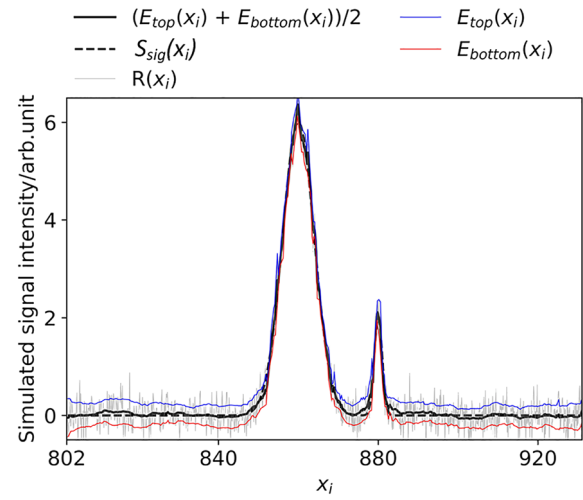
$$C_i^{bottom} = \frac{\sum V_i |V_i - C_i^{bottom}| \leq \Delta I_{max}}{j}, \text{Bottom envelope} \quad (18)$$

the average of all peak/valley data points in the window for which the value of the difference of the central window point  $C_i^{top}$  or  $C_i^{bottom}$  and the respected peak or valley data point is smaller than or equal to  $\Delta I_{max}$ , where  $\Delta I_{max}$  is computed according to Equation (13).  $j$  is the number of peak ( $P_i$ ) or valley ( $V_i$ ) data points contained within the moving window  $w^m$  where  $|P_i - C_i^{top}| \leq \Delta I_{max}$  or  $|V_i - C_i^{bottom}| \leq \Delta I_{max}$ .

In Equation (17), the central peak  $C_i^{top}$  or valley  $C_i^{bottom}$  data points are updated by averaging all the peak  $P_i$  or the valley  $V_i$  data points within  $w^m$ , respectively. Contrary, Equation (18) updates the central peak  $C_i^{top}$  or valley  $C_i^{bottom}$  data points by averaging peak  $P_i$  or valley  $V_i$  data points, which fulfill a condition  $|P_i - C_i^{top}| \leq \Delta I_{max}$  or  $|V_i - C_i^{bottom}| \leq \Delta I_{max}$ . This condition makes sure that only peak  $P_i$  or valley  $V_i$  data points that are not far from  $C_i^{top}$  or valley  $C_i^{bottom}$  are considered to update  $C_i^{top}$  or valley  $C_i^{bottom}$ , respectively.

By linear interpolation between all updated peak  $C_i^{top}$  and valley  $C_i^{bottom}$  data points for all variables  $x_i$  that according to Equations (11) and (12) have neither been assigned to a valley nor a peak point, the noise-reduced top envelope  $E_{top}(x_i)$  and bottom envelope  $E_{bottom}(x_i)$  are generated. Figure 4 shows both of them computed for a moving window with the size  $n = 9$ . Scheme 1 presents the flow chart of the envelope-finder algorithm where  $m$  is total number of peak/valley data points of the noisy spectrum.

Figure 4 shows the top and the bottom envelope as blue and red line. The raw spectral data  $R(x_i)$  are shown as grey line. The real signal contribution  $S_{sig}(x_i)$  behind



**FIGURE 4** Simulated noisy spectrum (grey line), linearly interpolated top envelope (blue line) and bottom envelope (red line), the mean of the top and bottom envelope (black line), and the real signal spectrum as dashed line [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

the spectral data is shown as dashed line. The solid black line shows the

$$E_{mean}(x_i) = \frac{E_{bottom}(x_i) + E_{top}(x_i)}{2} \quad (19)$$

mean of the top and the bottom envelopes  $E_{mean}(x_i)$ . Apparently,  $E_{mean}(x_i)$  is already close to  $S_{sig}(x_i)$ . This indicates that the envelope-finder algorithm alone has a great potential of noise reduction. Nevertheless, the noise level can be even more reduced if in a next step vectors are casted within the margins of the top and the bottom envelopes.

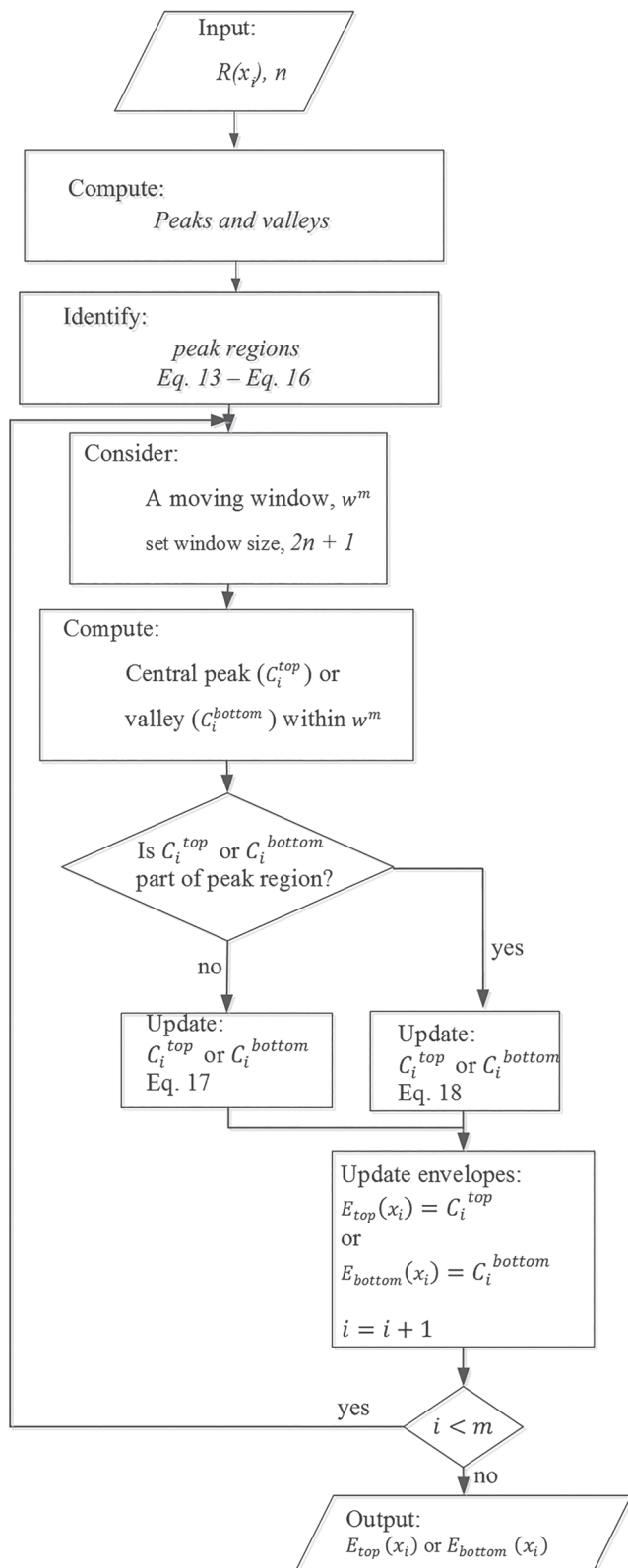
### 3.2 | Vector casting based smoothing

Vectors are casted from a starting already noise-reduced point  $r(x_k)$  to subsequent not yet noise-reduced data points  $R(x_{i > k})$ , as it is illustrated in Figure 5. The starting already noise-reduced point  $r(x_k)$

$$r(x_k) = E_{mean}(x_{i=0}) \quad (20)$$

is the mean of the bottom and top envelope at  $x_i = 0$ . From this already noise-reduced point  $r(x_k)$ , vectors

$$\vec{a}_i = \begin{pmatrix} R(x_{i > k}) - r(x_k) \\ x_i - x_k \end{pmatrix} \quad (21)$$



**SCHEME 1** Flow chart of envelope-finder algorithm

can be casted to all subsequent data points available with  $i > k$ .

Second, all vectors  $\vec{a}_i$  that cross either the top or the bottom envelope are deleted from the set of vectors  $\vec{a}_i$ .

Deleted vectors are highlighted in red in Figure 5, whereas remaining vectors are highlighted in green.

Third, for each of the remaining vectors the slope

$$b_i = \frac{R(x_{i+k}) - r(x_k)}{x_i - x_k} \quad (22)$$

is computed from which the average slope

$$\bar{b}_k = \frac{\sum b_i}{l} \quad (23)$$

of all the remaining vectors is computed, where  $l$  is the number of remaining vectors.

Fourth, for the not yet noise-reduced data point  $x_{i=k+1}$  that is situated one increment right of the already noise-reduced data point  $x_k$ , the new noise-reduced value

$$r(x_{i=k+1}) = r(x_k) + \bar{b}_k \cdot (x_{i=k+1} - x_{i=k}) \quad (24)$$

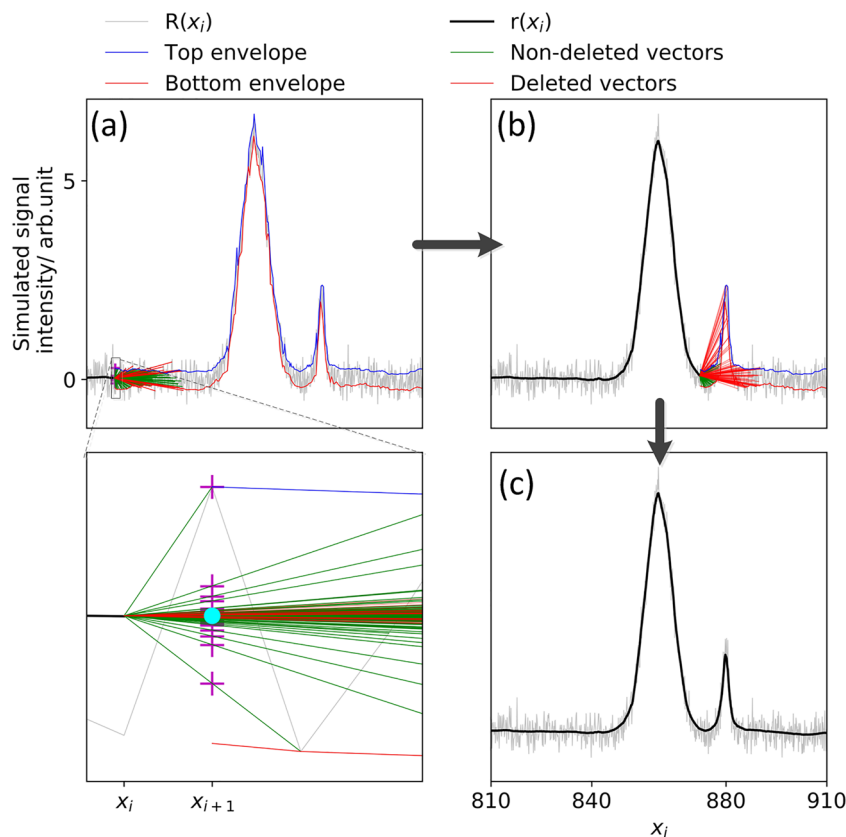
is computed from the average slope of the remaining vectors and the distance between the two neighbouring variables  $x_{i=k}$  and  $x_{i=k+1}$ . This procedure is repeated as long as all raw data points  $R(x_i)$  have been replaced by noise-reduced data points  $r(x_i)$ .

Figure 5a,b shows the noise reduction due to the vector casting method in a spectral region that does not contain a signal peak and in a spectral region that does contain a signal peak respectively. Figure 5a (zoomed plot) shows the details of the computation of the next noise-reduced data point starting from the previous one and Figure 5c shows as solid black line the computed noise-reduced spectrum  $r(x_i)$ .

In Figure 5, vectors are not casted from the previously noise-reduced data point to all of the subsequent data points but only to subsequent data points contained in a certain window  $w^{vector}$ . Casting the vectors not to all subsequent data points but only to data points contained in a certain window reduces the computation demand significantly. In Figure 5, the size of the window  $w^{vector}$  in which the vectors are casted is  $M = 150$ , meaning, that vectors are casted to the subsequent 150 data points. Scheme 2 shows the flow chart of the vector casting method.

### 3.3 | Parameter tuning effect

The algorithms outlined in the previous section requires two input parameters: the size  $n$  of the moving window  $w^m$  and the size  $M$  of the window in which vectors are casted  $w^{vector}$ . In order to



**FIGURE 5** The noisy raw spectrum is represented by the grey line. The smooth top and bottom envelope are highlighted in blue and red. The noise-reduced data points are represented by the black solid line. (a) Illustration of the vector casting method in a spectral range that does not contain a signal peak. Details of computation of the denoised spectrum data points, interceptions of selected vector lines (magenta +), and mean of interceptions of selected vector lines (cyan circle). Green vectors remain as they do not intersect the top or the bottom envelope. Red vectors are deleted as they intersect the top or the bottom envelope. (b) Illustration of the vector casting method in a spectral region that does contain a signal peak. (c) Illustration of the complete noise-reduced spectrum in the respective section [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

investigate the effect of these parameters, we applied the vector casting method at different values of  $n$  and  $M$ . In Figure 6, we showed the results at  $n = 1, 5, 9, 11$  keeping  $M = 150$ .

Increasing  $n$  initially from  $n = 1$  to  $n = 5$  improves the smoothness of the noisy signal especially for small SNR (peak-free region Figure 6). However, the vector casting method is rather insensitive to further increase of the size of the smoothing window from  $n = 9$  to  $n = 11$ . The peak regions are also less sensitive to the change in  $n$  as compared with the peak-free regions because the data points for averaging are determined automatically (Equation (18)) where only small number of nearby data points are involved.

The effect of varying the number of vectors to be casted is shown in Figure 7 and was tested by setting  $M = 50, 100$ , and  $150$  keeping  $n = 9$ . Compared with the mean of envelopes (black line in Figure 7), casting vectors show significant improvement. However, increasing  $M$  further than  $M = 50$  did not show significant improvement as the noise-reduced spectra look rather similar. This can be justified by the circumstance that the larger the distance between  $x_i$  and  $x_k$  is, the less is the probability of the corresponding vector to be included in the computation of the new noise-reduced data point  $r(x_{k+1})$  in Equation (24).

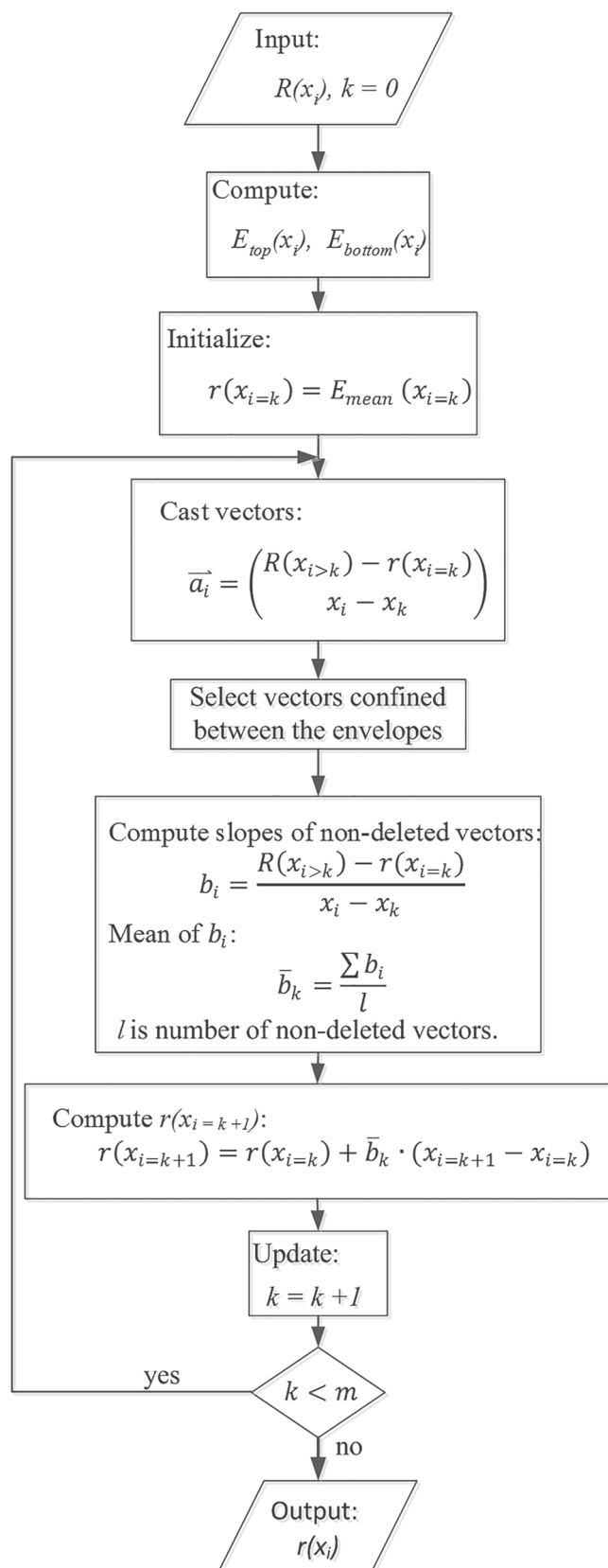
### 3.4 | Comparison with Savitzky–Golay and wavelet transform smoothing techniques

Figure 8 shows the simulated signal spectrum as solid black line and as grey simulated raw spectra with noise levels between 1 and 25 dB. At each SNR, 10 samples were simulated. The raw spectra are noise reduced using the presented vector casting method, the presented envelope-finder algorithm, the SG method, and the wavelet transform method. For the SG and the wavelet transform method, the input parameters were optimized with respect to a maximum overall SNR performance between the obtained noise-reduced spectrum and the pure signal spectrum according to Equation (8d). Figure 9a,b shows the parameters selected to give optimal denoised spectra for the wavelets and SG methods, respectively.

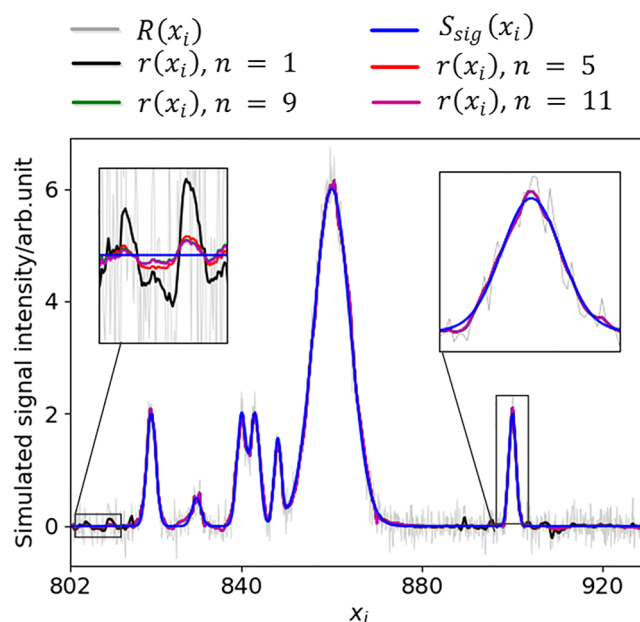
With respect to the SG method, the window size was varied from three to the maximum odd number that was smaller than or equal to the number of data points of the spectrum, and the polynomial order was varied between one and nine. During denoising of the simulated noisy signals, as it can be seen in Figure 9b, polynomial order of three and window size of nine were more frequently selected.

With respect to the wavelet transform method, a wavelet denoising function (*wdenoise*) using the software

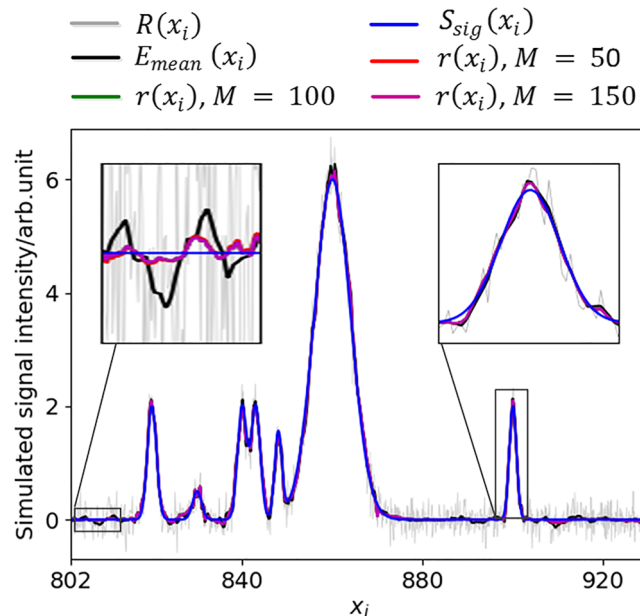




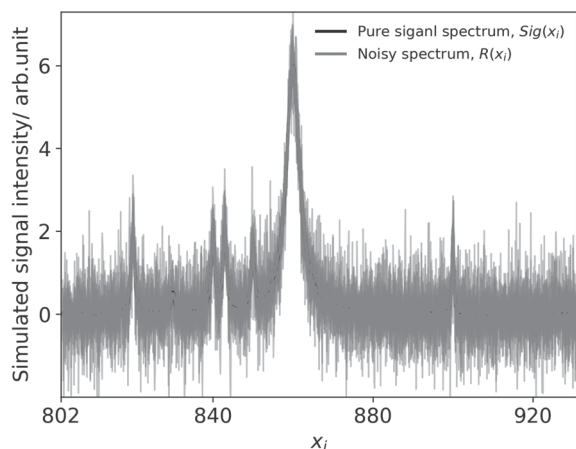
**SCHEME 2** Flow chart of vector casting method for denoising of noisy spectrum



**FIGURE 6** Vector casting method applied at different smoothing window size of  $n = 1$  (black solid line),  $n = 5$  (red solid line),  $n = 9$  (green solid line), and  $n = 11$  (magenta solid line). Grey solid line shows the simulated noisy spectra, and blue solid line shows pure reference signal [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**FIGURE 7** Vector casting applied at different values of number of casted vectors  $M = 50$  (red solid line),  $M = 100$  (green solid line), and  $M = 150$  (magenta solid line). The mean of the envelopes is shown in black solid line. Grey solid line shows the simulated noisy spectra, and blue solid line shows pure reference signal [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**FIGURE 8** Synthetic noisy spectra (grey lines) simulated at different SNR levels (1, 2, ..., 20, 25 dB) and reference spectrum (black line)

package “Wavelet Toolbox” in MATLAB (by MathWorks Inc.) was used. Improved implementation versions of the wavelet denoising technique<sup>[40,41]</sup> can exist; however, the relevant codes are not available and thus could not be applied. Therefore, using the wavelets denoising built in MATLAB, we varied the level of decomposition between 1 and 10. Four different threshold selection rules<sup>[42]</sup> were tested. For the selection of the suppression coefficients, mean, median, soft and hard thresholding<sup>[43]</sup> approaches were evaluated. Moreover, two different wavelet families (symlets and Daubechies) were tested. Figure 9a shows the frequency of usage of these parameters while optimally denoising the simulated noisy signals with wavelets method. With respect to the envelope-finder approach, we used a size of the moving window of  $n = 9$ , and for the vector casting method, we casted the vectors in a window containing 150 data points.

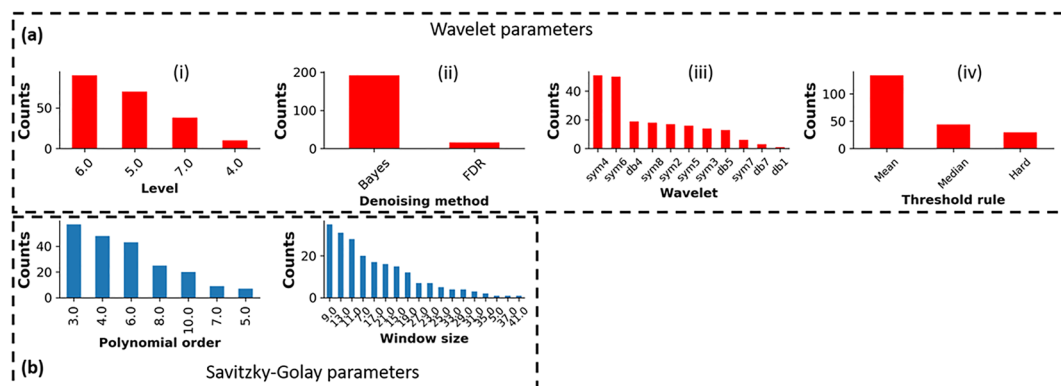
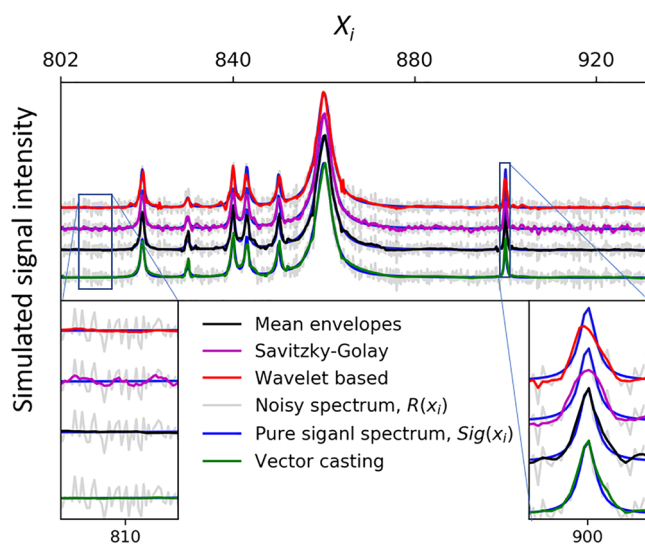


Figure 10 shows the SNR achievement of the four denoising methods computed using Equations (8a), (8c), and (8d). As it can be seen in Figure 10a,b, all the denoising methods improved the original SNR across the entire spectral region as well as at sharp peaks. The vector casting and wavelet methods perform better as compared with the other two methods. The vector casting method performs better than wavelet method at lower SNR, whereas the wavelet method exceeds the performance of vector casting method at higher SNRs. Figure 10c depicts the overall performance of the denoising methods in smoothing the noisy signal while at the same time keeping the spectral peaks undistorted. For noisy signal with SNR up to 15 dB, the vector casting method performs better followed by the mean envelopes. For higher SNRs, the wavelets method exceeded the performance of the here proposed methods.

Figure 11 shows the simulated raw spectrum  $R(x_i)$  with a SNR = 10 dB as grey line, the pure signal spectrum  $S_{sig}(x_i)$  as blue line, and the denoised spectra  $r(x_i)$  of vector casting, mean envelope, Savitzky–Golay, and wavelets as green, black, magenta, and red lines, respectively. From the comparison of the pure signal spectrum and denoised signal spectra information can be extracted about the performance of the different noise reduction methods with respect to the level of smoothing at peak-free regions and preservation of spectral shapes at peak regions (zoomed figures in Figure 10). With the vector casting method (green line), the noise-reduced spectrum shows excellent match to the peak locations of the signal

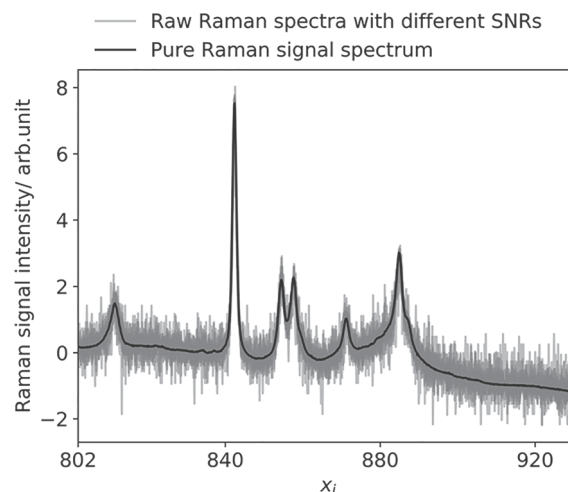


**FIGURE 11** Comparison of noise-reduced spectra  $r(x_i)$  (black line) using (a) vector casting method, (b) envelope-finder algorithm, (c) wavelet based smoothing, and (d) Savitzky–Golay filter with respect to the pure signal spectrum  $R(x_i)$  (blue line). The original noisy spectrum is shown as grey line [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

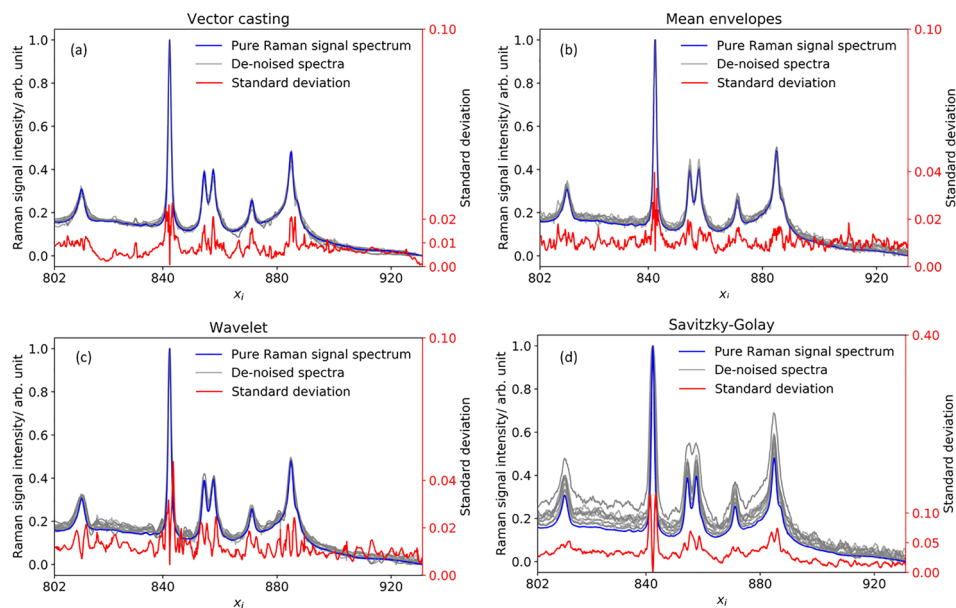
spectrum and preserves the spectral shape information rather well. And the standard deviation of the denoised spectrum is very small compared with the other methods specially at peak-free regions. Denoising applying solely the envelope-finder algorithm (black line) provides an overall noisier noise-reduced spectrum than the vector casting method. Still the peak heights and spectral shape are preserved rather well. The noise-reduced spectrum obtained using the optimized Savitzky–Golay method (magenta) shows a noisier spectrum and that the spectral peak shape information is manipulated compared with the pure signal spectrum. The wavelet transform method (red line) shows a better performance than the Savitzky–Golay method. Nonetheless, the peak positions and the peak shapes are manipulated slightly.

Finally, the performance of the four denoising approaches is compared based on experimentally acquired Raman spectra. Figure 12 shows 14 experimental Raman spectra of ethanol (grey lines) featuring different noise levels. A quasi-pure ethanol signal spectrum (black line) with large SNR is also shown as quasi-pure signal spectrum  $S_{sig}(x_i)$ .

Figure 13a–d shows each 14 Raman spectra (grey lines) of ethanol, normalized to the highest peak at around 845 nm, recovered using vector casting, mean of envelopes, wavelet, and Savitzky–Golay smoothing methods, respectively. The spectrum of ethanol with high SNR is also shown in blue line in the figures for reference. Moreover, to assess the reproducibility of the recovered spectra, the standard deviation of the 14 recovered noise-reduced spectra at each variable (here Raman shift) is computed and depicted alongside the recovered spectra as a red line. The standard deviation is quantified on the



**FIGURE 12** Experimentally measured Raman spectra of ethanol (grey lines) featuring different SNR levels and one quasi-noise-free pure signal Raman spectrum of ethanol (black)



**FIGURE 13** Raman spectra of ethanol (grey lines) denoised by (a) vector casting method, (b) envelope-finder algorithm, (c) wavelet transform method, and (d) Savitzky–Golay smoother. The pure signal spectrum (blue line) and standard deviation of the denoised spectra (red line) are shown [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

right ordinate. As it can be seen from Figure 13, the standard deviation is higher around the peak regions than at peak-free regions. Thus, all techniques affect the peak to some extent. However, with the vector casting method, a better reproducibility of the spectra was obtained. In every peak, the vector casting method achieved the minimum standard deviation. The mean of the envelopes also shows a comparable result with the wavelets method. The standard deviation of the Raman peak at around 812 nm is 0.013, 0.021, and 0.052 for the vector casting, wavelets, and Savitzky–Golay methods, respectively. The peak broadening effect of the Savitzky–Golay technique is highly reflected by the standard deviation of the peak at around 842 nm. Moreover, the standard deviation of the double Raman peaks at 855 nm, shoulder peak at 870 nm, and Raman peak at 884 nm is decreased from 0.07 and 0.025 to 0.021, from 0.04 and 0.019 to 0.016, and from 0.07 and 0.024 to 0.02 with respect to Savitzky–Golay and wavelet methods, respectively.

Next to the circumstance that the here proposed new method for the denoising of raw spectra outperforms the two most frequently used methods, it has to be mentioned that the newly proposed method also involves a minimum of human interaction. In contrast, our method requires envelope detection that involves peak detection. We also compared the proposed algorithm in terms of computational efficiency. The language used for the implementation was Python.<sup>[44]</sup> The average time taken for the envelope-finder algorithm was comparable with SG on a Dell Latitude E7450 with an Intel Core i7 processor. However, the vector casting method took longer execution time, and the average execution time depends on the number of vectors to be casted.

## 4 | CONCLUSION

In this study, we developed a new method for the processing of spectra that are relevant for the purification of spectral signal from spectra with small SNR. Of course, this technique cannot extract signal peaks that are smaller than the noise level, but it can remove noise, although manipulating the characteristics of the pure signal less than the wavelet transform method or the SG method. Furthermore, the proposed method does only to a minimum extent rely on input parameters that have to be chosen by humans. Summarizing, the proposed method should be considered reliable, robust, and accurate.

## ACKNOWLEDGEMENTS

The project leading to this result has received funding from the Wilhelm Sander-Stiftung, Munich, Germany (Grant 2017.111.1). It also has received funding from the European Union's Horizon 2020 research and innovation programme under ERC Starting Grant agreement 637654 (Inhomogeneities).

## ORCID

Medhanie Tesfay Gebrekidan  <https://orcid.org/0000-0002-0600-7217>

## REFERENCES

- [1] I. Pence, A. Mahadevan-Jansen, *Chem. Soc. Rev.* **2016**, *45*, 1958.
- [2] I. J. Bigio, S. G. Bown, *Cancer Biol. Ther.* **2004**, *3*, 259.
- [3] A. Lorincz, D. Haddad, R. Naik, V. Naik, A. Fung, A. Cao, P. Manda, A. Pandya, G. Auner, R. Rabah, *J. Pediatr. Surg.* **2004**, *39*, 953.

- [4] W. Wang, J. Zhao, M. Short, H. Zeng, *J. Biophotonics* **2015**, 8, 527.
- [5] H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, *Nat. Protoc.* **2016**, 11, 664.
- [6] I. ur Rehman, Z. Movasaghi, S. Rehman, *Vibrational spectroscopy for tissue analysis*, 1st ed., CRC Press, Boca Raton, Florida **2012**.
- [7] F. M. Hoffmann, *Surf. Sci. Rep.* **1983**, 3, 107.
- [8] J. Smulko, M. S. Wróbel, Barman I In **2015**; *IEEE: 2015*, p 1.
- [9] Institute ANS, *American national standard for safe use of lasers*, Laser Institute of America, Orlando, Florida **2007**.
- [10] C. Knipfer, J. Motz, W. Adler, K. Brunner, M. T. Gebrekidan, R. Hankel, A. Agaimy, S. Will, A. Braeuer, F. W. Neukam, *Biomed. Opt. Express* **2014**, 5, 3252.
- [11] M. T. Gebrekidan, C. Knipfer, F. Stelzle, J. Popp, S. Will, A. Braeuer, *Journal of Raman Spectroscopy* **2016**, 47, 198.
- [12] S. Chan, S. Kwon, T. W. Koo, L. P. Lee, A. A. Berlin, *Adv. Mater.* **2003**, 15, 1595.
- [13] E. J. Kostelich, J. A. Yorke, *Phys. Rev. A* **1988**, 38, 1649.
- [14] T. Schreiber, P. Grassberger, *Physics Letters A* **1991**, 160, 411.
- [15] H. Santosa, M. Jiyoun Hong, S.-P. Kim, K.-S. Hong, *Rev. Sci. Instrum.* **2013**, 84, 073106.
- [16] M. Pal, A. Porwal, *IEEE* **2015**, 2015, 752.
- [17] P. Lasch, *Chemom. Intel. Lab. Syst.* **2012**, 117, 100.
- [18] A. Savitzky, M. J. Golay, *Anal. Chem.* **1964**, 36, 1627.
- [19] M. U. Bromba, H. Ziegler, *Anal. Chem.* **1981**, 53, 1583.
- [20] P. M. Ramos, I. Ruisánchez, *Journal of Raman Spectroscopy: An International Journal for Original Work in all Aspects of Raman Spectroscopy, Including Higher Order Processes, and also Brillouin and Rayleigh Scattering* **2005**, 36, 848.
- [21] F. Ehrentreich, L. Sümchen, *Anal. Chem.* **2001**, 73, 4364.
- [22] G. W. Horgan, *Journal of Applied Statistics* **1999**, 26, 923.
- [23] P. H. Eilers, *Anal. Chem.* **2003**, 75, 3631.
- [24] M. Člupek, P. Matějka, K. Volka, *Journal of Raman Spectroscopy: An International Journal for Original Work in all Aspects of Raman Spectroscopy, Including Higher Order Processes, and also Brillouin and Rayleigh Scattering* **2007**, 38, 1174.
- [25] S. Chen, X. Lin, C. Yuen, S. Padmanabhan, R. W. Beuerman, Q. Liu, *Opt. Express* **2014**, 22, 12102.
- [26] J. Fochesatto, J. Sloan In *Spectral Sensing Research For Surface and Air Monitoring In Chemical, Biological and Radiological Defense and Security Applications*; World Scientific: **2009**, p 49.
- [27] V. Barclay, R. Bonner, I. Hamilton, *Anal. Chem.* **1997**, 69, 78.
- [28] L. Quintero, C. Matthäus, S. Hunt, M. Diem, *International Society for Optics and Photonics* **2010**, 7568, 756817.
- [29] X. Fan, X. Wang, X. Wang, Y. Xu, J. Que, H. He, X. Wang, M. Tang, *Journal of Instrumentation* **2016**, 11, P02002.
- [30] M. Kneen, H. Annegarn, *Nucl. Instrum. Methods Phys. Res., Sect. B* **1996**, 109, 209.
- [31] Z.-M. Zhang, S. Chen, Y.-Z. Liang, *Analyst* **2010**, 135, 1138.
- [32] W. Demtröder, *Laser spectroscopy: basic concepts and instrumentation*, 3th ed., Springer-Verlag, Heidelberg **2003**.
- [33] X. Yuan, R. A. Mayanovic, *Appl. Spectrosc.* **2017**, 71, 2325.
- [34] P. Vandenabeele, *Practical Raman spectroscopy: an introduction*, Vol. 2013, Wiley Online Library, Chichester, West Sussex **2013**.
- [35] S. J. Barton, T. E. Ward, B. M. Hennelly, *Anal. Methods* **2018**, 10, 3759.
- [36] L. Azzari, L. R. Borges, A. Foi, *Denoising of Photographic Images and Video*, Springer, Basel, Switzerland **2018** 1.
- [37] R. L. McClain, J. C. Wright, *J. Chem. Educ.* **2014**, 91, 1455.
- [38] L. Alparone, M. Selva, B. Aiazzi, S. Baronti, F. Butera, L. Chiarantini, *IEEE* **2009**, 2009, 1.
- [39] L. Galleani, L. Cohen, D. Nelson Local signal to noise ratio; SPIE, 2006; Vol. 6313, Op.
- [40] M. Srivastava, C. L. Anderson, J. H. Freed, *IEEE Access* **2016**, 4, 3862.
- [41] D. Chen, Z. Chen, E. Grant, *Anal. Bioanal. Chem.* **2011**, 400, 625.
- [42] M. Misiti, Y. Misiti, G. Oppenheim, J. Poggi, *The Mathworks, Inc., Natick, Massachusetts* **2014**.
- [43] B. K. Alsberg, A. M. Woodward, D. B. Kell, *Chemom. Intel. Lab. Syst.* **1997**, 37, 215.
- [44] G. V. Rossum, F. L. Drake, *PythonLabs*, Virginia, USA **2001**.

**How to cite this article:** Gebrekidan MT, Knipfer C, Braeuer AS. Vector casting for noise reduction. *J Raman Spectrosc.* 2020;51:731–743.  
<https://doi.org/10.1002/jrs.5835>