Bertinetto, Carlo G.; Vuorinen, Tapani

Automatic Baseline Recognition for the Correction of Large Sets of Spectra Using Continuous Wavelet Transform and Iterative Fitting

Published in:
Applied Spectroscopy

DOI:
10.1366/13-07018

Published: 01/01/2014

Document Version
Peer reviewed version

Please cite the original version:
Bertinetto, C. G., & Vuorinen, T. (2014). Automatic Baseline Recognition for the Correction of Large Sets of Spectra Using Continuous Wavelet Transform and Iterative Fitting. *Applied Spectroscopy*, *68*(2), 155-164. https://doi.org/10.1366/13-07018

Automatic Baseline Recognition for the Correction of Large Sets of Spectra Using Continuous Wavelet Transform and Iterative Fitting

Carlo G. Bertinetto,* Tapani Vuorinen

Department of Forest Products Technology, School of Chemical Technology, Aalto University, P.O. Box 16300, 00076 Aalto, Finland

A new algorithm for the automatic recognition of peak and baseline regions in spectra is presented. It is part of a study to devise a baseline correction method that is particularly suitable for the simple and fast treatment of large amounts of data of the same type, such as those coming from high-throughput instruments, images, process monitoring, etc. This algorithm is based on the continuous wavelet transform, and its parameters are automatically determined using the criteria of Shannon entropy and the statistical distribution of noise, requiring virtually no user intervention. It was assessed on simulated spectra with different noise levels and baseline amplitudes, successfully recognizing the baseline points in all cases but for a few extremely weak and noisy signals. It can be combined with various fitting methods for baseline estimation and correction. In this work, it was used together with an iterative polynomial fitting to successfully process a real Raman image of 40 000 pixels in about 2.5 h.

Index Headings: **Baseline correction; Continuous wavelet transform; CWT; Iterative fitting; High-throughput spectroscopy; Raman imaging.**

INTRODUCTION

During the past decades, the evolution of techniques for multivariate analysis, spectral resolution, and machine learning led to an increasingly sophisticated level of information that can be extracted from spectroscopic measurements. At the same time, these techniques require an ever higher quality of the data in order to produce reliable results. For this reason, experimental measurements often undergo a preprocessing phase to separate indeterministic signal components from the spectral features of interest. Frequently, spectra are contaminated by what is commonly referred to as "baseline", i.e., wide fluctuations of the measured signal that are unrelated to the phenomenon under investigation. Whereas these fluctuations do not always hamper qualitative analysis, as spectral features may still be recognizable with the baseline embedded,[1] they have adverse effects on quantitative analysis, reducing the simplicity and robustness of any mathematical model based on these spectra.[2]

Several methods were developed for baseline correction. Roughly, they can be classified as: (a) methods that estimate the baseline from a set of spectra, and (b) methods that process spectra individually. The first category derives the baseline from the relationships among the spectra of an entire set, usually employing multivariate analysis techniques. The simplest methods consist in identifying a spectral component that is approximately constant throughout the set and separating it from the rest of the signal. Depending on the situation, this component may be attributed to the baseline[3] or the spectral peaks.[4] This approach is simple and often effective, but the constant-component assumption limits its use to very specific cases, such as calibrations. Other methods make use of additional data associated to the spectra, e.g., the concentration of a particular substance, and identify as baseline the component that has no correlation to these data. The mathematical techniques used for this purpose include orthogonal signal correction,[5] partial least squares,[6] and methods based on information theory.[7,8] This approach is generally very accurate and reliable, but can be used only when some associated information is available. Moreover, its outcome is strongly related to this specific information and may not be suitable in different investigations or experimental setups.

The second category of baseline correction methods processes spectra individually using criteria, such as shape and characteristics of bands, to separate the signal originated by a particular phenomenon of interest from other effects occurring in the analyte or in the instrument (e.g., distinguishing Raman scattering from fluorescence emission or variations in laser intensity). In general, the baseline is expected to have a low curvature, whereas the most rapid oscillations are considered random noise, and the "true" signal lies somewhere in the middle. Although these methods do not make use of precise numerical data associated to the spectra, they do exploit some form of background knowledge, such as the type of function and/or parameters to approximate the baseline curve or the signal peaks. The background knowledge for baseline correction may also be provided by the scientist's experience, as in manual methods. These consist in selecting a series of points representative of the baseline by visual inspection and then interpolating or fitting these points with a suitable function, e.g., linear, polynomial, or spline functions. A more sophisticated variant has manual correction performed on the principle components of a set of spectra.[9] Although the manual approach is still widely used, it is subjective, not perfectly reproducible, and time consuming.[10]

Many different methodologies were developed to obtain an objective and reproducible baseline correction. A comprehensive description of most of them can be found in the work of Schulze et al.[11] and a more recent

Received 25 January 2013; accepted 10 September 2013.

* Author to whom correspondence should be sent. E-mail: carlo.bertinetto@aalto.fi.

DOI: 10.1366/13-07018

0003-7028/14/6802-000000/0

one in the paper by Rowlands and Elliott.[12] These methods differ in the algorithms and mathematical criteria, e.g., derivatives, entropy, or frequency, used to separate the "true" signal from the baseline and other kinds of noise. They all require the user to specify certain parameters and/or stopping criteria, though the number of these parameters varies for each method.

The increasing use of high-throughput measurement techniques that produce large amounts of data brought about the need for fast spectral processing. Various authors tackled this issue by proposing automated or semi-automated algorithms for baseline correction that reduce human intervention. Effective solutions were devised using iterative polynomial fitting,[13] penalized quantile spline regression,[14] adaptive least squares/Whittaker smoother,[15–17] moving average-peak stripping,[18–20] local second derivative,[12] and morphological or geometrical approaches.[21,22] The performance of these methods differ in terms of accuracy, computational speed, amount of human intervention, and types of spectra to which they can be applied; these goals are usually conflicting. Some methods are designed, at least in principle, to be able to process any spectrum using absolutely no human intervention or knowledge. However, these fully automated methods often require computational times that are too long for certain applications. For example, the algorithm proposed by Schulze et al.[19] is reported to correct a single baseline in at least 20 s. For a typical spectral image of 100 × 100 pixels, the corresponding overall correction time is in the order of days. Moreover, the applicability of such methods is limited by the actual difficulty in defining a universal criterion to identify the "true" signal that works for any spectrum. This even caused some authors to introduce user-defined knowledge back into algorithms that were initially devised as parameter free, e.g., to distinguish broad spectral bands from baseline sections with a high curvature.[15,23] In our opinion, getting completely rid of instance-related background knowledge may not even be achievable, not least because the very definition of what constitutes "true" signal and noise depends on the particular investigation.

In this framework, it would be highly useful to devise a method that is as general, automatic, and reproducible as possible, but at the same time, fast enough to handle large sets of spectra. To achieve such a compromise, we explore the capabilities of an approach based on the selection and fitting of points representative of the baseline, as in manual methods, but introducing some mathematical tools that make this process automatic. In particular, the present paper focuses on the issue of baseline recognition and proposes an algorithm that employs the continuous wavelet transform (CWT) together with Shannon entropy and statistical distribution of noise.

The wavelet transform (WT) is a mathematical technique that can extract information on the frequency and position of a signal through decomposition into appropriate basis functions.[24] The discrete form of WT has been used to correct baselines in many works, but was not deemed very suitable for automated methods because its outcome is highly dependent on several manually selected parameters.[11] The CWT was used more seldom for baseline correction. In the most notable application,[25] it was employed together with an iterative thresholding on the power spectrum to recognize baseline points, which were subsequently fitted with a Whittaker smoother. This method was later improved[17] to better recognize broad peaks and baseline points in low and congested regions. The reported spectra were corrected almost completely automatically, but the user was still required to manually indicate the presence of broad peaks and set a smoothness parameter for the fitting function.

The algorithm presented in this paper is first assessed by performing baseline recognition on simulated spectra with different curve types, noise levels, and signal intensities. As this study is still in an initial stage, not all cases are taken into account, and some approximations are introduced. However, the shown results are significant for a practical use of this algorithm in several types of spectroscopy and in very frequent cases, such as the treatment of a large set of spectra with similar characteristics, in which at least a part of the spectra have a good signal. Typical examples of these situations are the analysis of spectral images or the monitoring of industrial processes. To show that practical applications are already possible at this stage, the algorithm is also employed in combination with an iterative polynomial fitting to carry out a full baseline correction of real Raman spectra from an image of 200 × 200 pixels.

METHOD

Baseline Recognition Algorithm. The vector $\mathbf{y}$ of spectral intensities is modeled as $\mathbf{y} = \mathbf{s} + \mathbf{g} + \mathbf{n}$, where $\mathbf{s}$ is the spectral signal of interest, $\mathbf{g}$ is the baseline, and $\mathbf{n}$ is zero-mean random noise. The spectral correction consists in estimating $\mathbf{g}$ and subtracting it from $\mathbf{y}$. Two initial assumptions are made: (i) the baseline has a lower curvature (i.e., is smoother) than the rest of the signal; (ii) there are segments of the spectrum that do not contain spectral bands, i.e., these segments of $\mathbf{y}$ are composed of $\mathbf{g}$ and $\mathbf{n}$ only. These conditions are commonly met in many types of spectroscopy, e.g., Raman, infrared (IR), nuclear magnetic resonance (NMR), or mass spectrometry. The baseline is estimated by fitting the band-free points with an appropriate function, depending on the type of spectrum. These points are automatically selected by a baseline recognition algorithm, denoted in this paper as CWT-BR, based on the continuous wavelet transform (CWT).

The wavelet transform is a mathematical tool for time-frequency analysis that involves decomposing a signal into a set of appropriately defined basis functions named wavelets. A detailed explanation of the theory can be found in the work of Meyer.[24] Very briefly, wavelets are obtained by linear transformation (scaling and translation) of a locally oscillating curve ψ called mother wavelet. The scaling and translation are quantified by two parameters a and b, respectively; their allowed values discriminate between discrete wavelet transform, in which a and b are quantized, and CWT, in which a and b can assume any value. The scaling parameter a can be associated with the curvature of the signal, whereas b refers to its original domain, e.g., spectral units. Being

$f(x)$ the signal under consideration and $\psi_{a,b}$ a wavelet derived from a particular choice of ψ, a, and b, the formal expression of the CWT is

$$CWT(a,b) = \int_{-\infty}^{\infty} \psi_{a,b}^{*} \times f(x)dx = \langle \psi_{a,b} | f(x) \rangle \tag{1}$$

with $\psi_{a,b}^{*}$ indicating the complex conjugate of $\psi_{a,b}$ and $\langle \psi_{a,b} | f(x)$ being a notation used for the inner product. In other words, it is a convolution of the wavelet with the signal over a continuous range of scales (a). It expresses the similarity coefficient between the wavelet and the signal at each point b, and it may thus be used to extract the part of the signal that is most similar to the employed wavelet and scale. By choosing a wavelet and scale similar to the peaks, one can highlight the spectroscopic bands and suppress the rest, i.e., background and random noise.

The basic scheme of the CWT-BR method is illustrated in Fig. 1. Given a spectrum (Fig. 1a), its CWT is calculated for several scales (Fig. 1b), and a particular value of a, denoted as $\hat{a}$, is selected (Fig. 1c). The spectral regions with low CWT($\hat{a}$) values, i.e., lower than a threshold θ (Fig. 1d), are considered as peak free and belonging to the baseline (Fig. 1e). The mother wavelet employed in this algorithm is the inverse of the second derivative of the Gaussian curve, also called Mexican hat function, depicted in Fig. 1f. This function has already been used in other studies[26,27] and is known to effectively approximate most types of spectroscopic signals. The parameters $\hat{a}$ and θ are determined by automatic series of operations explained below, which work on a few additional assumptions: (i) random noise is considered to have a normal distribution; (ii) when analyzing a set of several spectra, at least some of them have a good signal-to-noise ratio (SNR); and (iii) the spectra within the set are not very different in shape and minimum distance among peaks (meaning the distance between the closest peaks in a spectrum).

Selection of Scale. Let γ_a be the CWT of the spectrum vector $\mathbf{y}$ for a given scale parameter a. Let H be the Shannon entropy of γ_a, defined as

$$H = -\sum_i \left(\frac{c_i}{C} \cdot \ln \frac{c_i}{C} \right) \tag{2}$$

where c_i are the elements of γ_a and $C = \sum_i c_i$; the term inside parentheses is assigned zero for every $c_i = 0$.[*] This function expresses how much a signal curve is concentrated in fewer and narrower peaks: its values range from $\ln N$ (with N the number of elements of the signal vector) for a flat signal to zero for a signal that is null everywhere but for one single element. Here, H is used to find the scale parameter for which the CWT mostly highlights the peaks of interest over background fluctuations, such as baseline and random noise. From the example in Fig. 1b, it can be observed that, as a increases, the random noise is gradually suppressed, but the CWT gets broadened in the peak regions and tends to be more affected by the baseline curvature. The best compromise is reached at the scale corresponding to a minimum of $H(a)$, denoted as $\hat{a}$, see Fig. 1c. As illustrated in Figs. 2a, 2b, and 2c, $\hat{a}$ decreases when the peaks in a spectrum are more near each other; increasing the number of peaks without changing the distance between the two closest ones (not shown here) was not observed to have any effect. On the other hand, higher random noise pushes $\hat{a}$ to greater values, see Fig. 2d, and usually plays a larger role than peak distance. The influence of the baseline is more complex and less clear, but it can be ignored if the considered spectra have a sufficiently good signal. In spectra with several peaks with various width and intensity, there might be more than one local minimum of the $H(a)$ function, see Fig. 2e. In this case, $\hat{a}$ is chosen on the first local minimum, i.e., the one with lowest a, because the other minima are likely to derive from the suppression of the smallest peaks by the CWT at higher scales.

The Shannon entropy criterion may become less reliable with very noisy signals. Therefore, if a set of several spectra with somewhat similar peak shape and distance between closest peaks is to be analyzed, it is preferable to select $\hat{a}$ based on the best spectra of the set and use an average value for all of them. When the number of spectra is very large, i.e., of the order of thousands or more, it is convenient to do this through an automatic statistical procedure. In this study we employed the following:

1. A sample of k spectra is randomly picked from the initial group; $k \approx 1/10$ of the total.
2. For each of the k picked spectra, an approximate SNR, denoted as $\widetilde{\mathrm{SNR}}$, is evaluated as:

$$\widetilde{\mathrm{SNR}} = \frac{\max(\mathbf{y}) - \min(\mathbf{y})}{\sigma(d\mathbf{y})} \tag{3}$$

where $\mathbf{y}$ is the spectrum, σ means standard deviation, and $d\mathbf{y}$ is the differential spectrum, i.e., the vector $[y_2 - y_1, y_2, \ldots, y_N - y_{N-1}]$, with N the length of $\mathbf{y}$.

3. The $k/3$ spectra with highest $\widetilde{\mathrm{SNR}}$ are retained, while the others are discarded.
4. For the $k/3$ retained spectra, $\mathbf{y}$ is extended by repeating the first and the last element 400 times to avoid artifacts caused by border effects. The extended spectra $\mathbf{y}^{+}$ have the form:

$$\mathbf{y}^{+} = [\overbrace{y_1, \rightleftharpoons, y_1}^{400}, y_1, y_2, \rightleftharpoons, y_{N-1}, y_N, \overbrace{y_N, \rightleftharpoons, y_N}^{400}] \tag{4}$$

5. The CWT of $\mathbf{y}^{+}$, denoted γ_a^{+}, is computed using Mexican hat wavelet for several values of a ranging from 1 to M, with M a scale corresponding to a wavelet much broader than the peaks (in this instance $M = 60$).
6. The Shannon entropy H is calculated for each γ_a^{+} vector, ignoring the first 400 and the last 400 elements that derive from the artificial spectrum extension. Before calculating H, it is important to remove any spike noise, e.g., cosmic ray peaks in Raman spectra.

[*] Note: this definition for the entropy of a signal is not to be confused with the one often found in other works, which uses the probabilities of possible values instead of the values themselves.

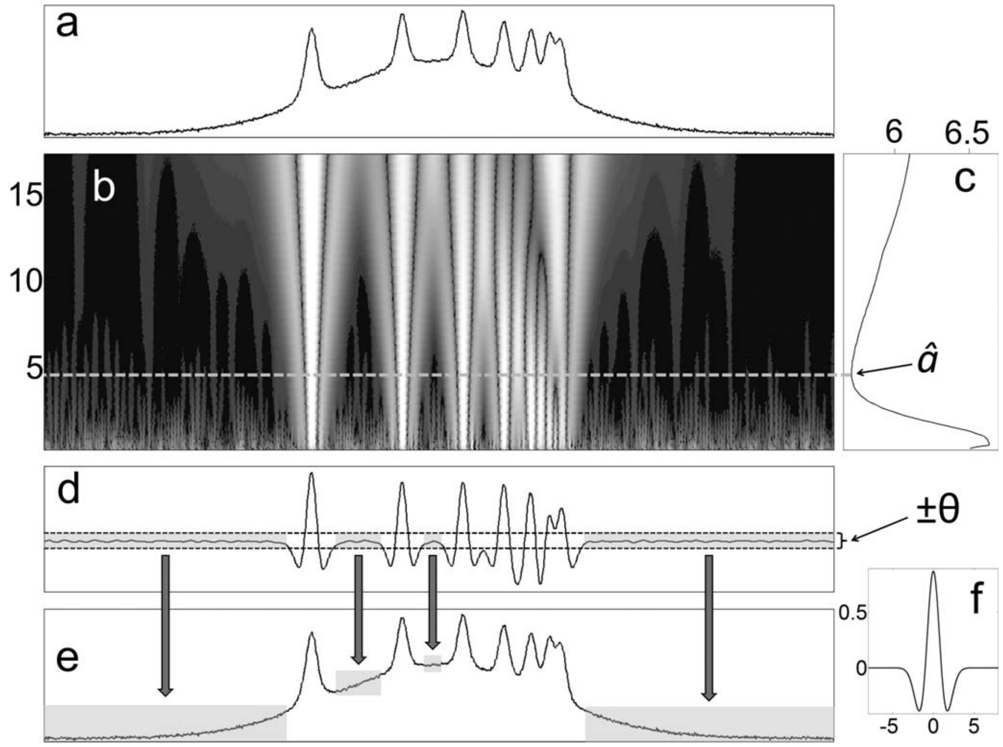

FIG. 1. Basic scheme of the CWT-BR algorithm: (**a**) Example spectrum (conventional units); (**b**) continuous wavelet transform (CWT): lighter shade corresponds to higher values and the vertical axis indicates the scale parameter *a*; (**c**) Shannon entropy of the CWT: the entropy value and the scale parameter are on the horizontal and vertical axes, respectively; (**d**) CWT for the chosen scale parameter *â*, corresponding to the minimum of the Shannon entropy. A threshold θ is defined and is indicated by the horizontal dashed lines; (**e**) the regions for which |CWT| < θ, indicated by the darker shade, are assigned to the background; (**f**) Mexican hat mother wavelet.

7. For each considered spectrum, the smallest *a* corresponding to a local minimum of *H*(*a*) is selected.
8. The mean among all the selected *a* is chosen as the value of *â* and is used in the next calculations.

Selection of the Threshold on the CWT. Because of possible deviations caused by random noise, the most appropriate threshold θ to discriminate CWT values corresponding to regions with or without peaks needs to be defined for each spectrum. This is done by the following operations:

1. The **y**+ and $\gamma_{\hat{a}}^{+}$ vectors are computed as in steps 4 and 5 of scale selection, using Mexican hat and scale parameter â.
2. A histogram *P* of the probability density function of $\gamma_{\hat{a}}^{+}$ is plotted, ignoring the elements corresponding to the extended points (i.e., the first and last 400 points).
3. The contributions to the CWT deriving from random noise are assumed to have a normal distribution around zero. The portion of *P* with values $\delta : \max(P) \geq \delta > 1/3 \max(P)$, i.e., the central part corresponding to the smallest elements of $\gamma_{\hat{a}}^{+}$, is fitted with a Gaussian curve. The number of bins in *P* is initially set as 200, then corrected by assigning it the value (approximated to the greater integer) $n = 8 \cdot R/\tilde{\sigma}$, with $R = [\max(\gamma_{\hat{a}}^{+}) - \min(\gamma_{\hat{a}}^{+})]$ and $\tilde{\sigma}$ the standard deviation of the Gaussian fit. The fit is performed again with the new binning, and this correction is repeated until the variation of *n* between consecutive iterations is less than 5% or it changes sign (never more than three iterations were needed for any spectrum processed in this work).
4. The value of θ is defined in relation to the Gaussian fit according to the following function:

$$\theta = \tilde{\sigma}\left(0.6 + 10 \cdot \frac{N_D}{N_G}\right) \tag{5}$$

where N_G is the area of *P* under the Gaussian curve, defined as the portion corresponding to the $\gamma_{\hat{a}}^{+}$ elements $\gamma_i : -3\tilde{\sigma} < \gamma_i < +3\tilde{\sigma}$, and N_D is the area of the remaining histogram. The N_D/N_G ratio expresses the weight of the randomly distributed $\gamma_{\hat{a}}^{+}$ elements compared to the rest of the signal. Here, θ typically takes the value of $\approx 5\tilde{\sigma}$ for spectra with high SNR and $0.6\tilde{\sigma}$ for very noisy spectra, in which the Gaussian distribution covers the whole range of *P*, see Fig. 3.

For every sequence of five spectral points for which $|\gamma_{\hat{a}}^{+}| < \theta$, the middle point of the sequence is assigned as baseline. Examining sequences instead of single points avoids mistaking for baseline all the zero crossings that frequently occur in CWT.

Iterative Polynomial Fit for the Refinement of Baseline Correction in Raman Spectra. For the experimental Raman spectra presented in this paper, the baseline points identified with CWT-BR were refined and fitted with an iterative polynomial fitting (IPF) algorithm, inspired by other methods known in the literature.[13] A fifth-order polynomial was used, which is known to effectively approximate the baseline in Raman spectra mainly caused by fluorescence of the analyte.[28,2]

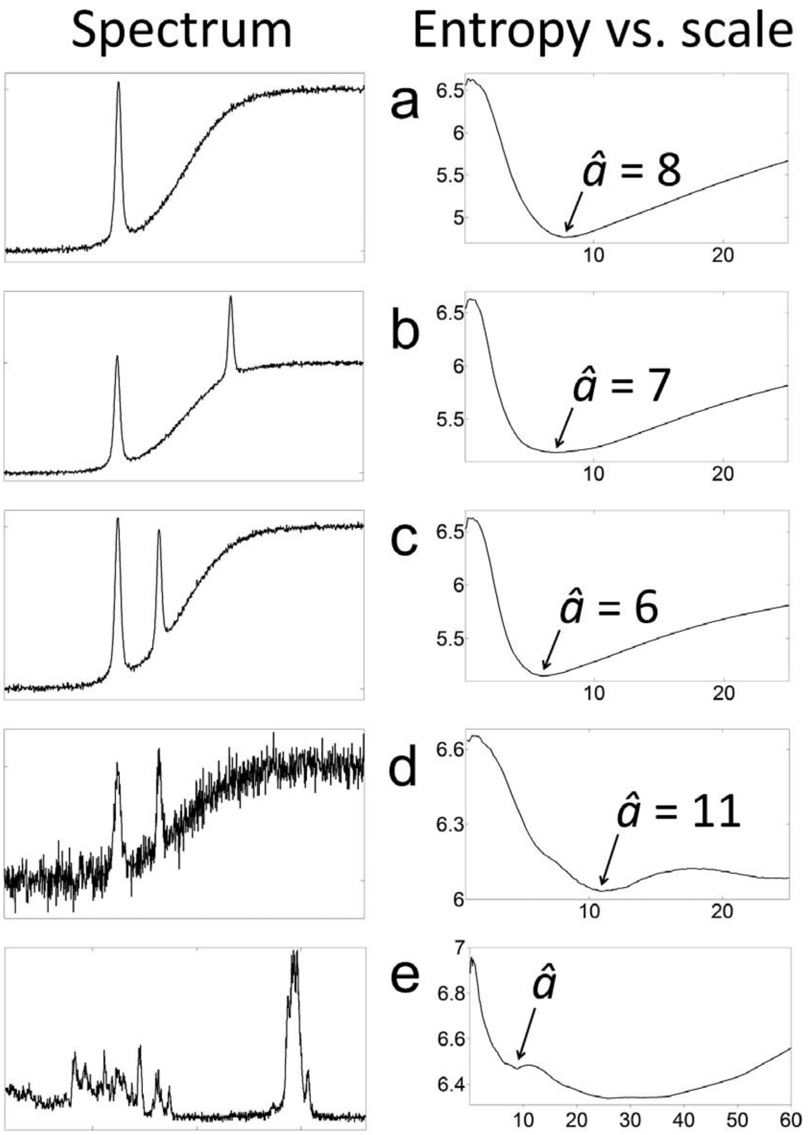

FIG. 2. Examples of spectra and their corresponding plot of the Shannon entropy of the CWT (vertical axis) as a function of the scale parameter (horizontal axis). The chosen $\hat{a}$ value is indicated on each plot. For more a detailed explanation, see text.

If we define B as the set of points selected after applying CWT-BR, the algorithm is schematized as follows:

1. Fit all the points in B with a fifth-order polynomial.
2. If $y_b - p_b > 1.5\ S$ (where y and p are the spectral intensity and the polynomial approximation, respectively, point $b \in B$, and S is the standard error of estimate), remove b from B.
3. Fit again the remaining points in B.
4. If $y_i < p_i$ for $i = l, l+1, l+2, \ldots, l+k$ (i.e., the spectrum is smaller than the polynomial approximation for a sequence of k consecutive points, with $k \geq N/100$), add the point $[i + (k/2)]$ to B.
5. Repeat steps 1–4 until convergence of B.

This procedure removes incorrect assignments, such as points in congested peaks that are identified as baseline and overestimations of the baseline that produce negative portions in the final corrected spectrum.

All the algorithms described in this paper were written in-house using MATLAB version 8.0 R2012b (The Mathworks, Natick, MA). The calculations were performed on a Fujitsu Esprimo E910 computer enabled with an Intel Core i5-3470 CPU of 3.20 GHz, 8 GB RAM and running on Windows Vista operating system. The codes of the algorithms are available on request.

Experiments. The performance of the CWT-BR algorithm was tested on simulated spectra taken or slightly modified from another baseline correction paper.[18] They consist of a vector of length 1001 containing seven Lorentzian peaks convoluted with Gaussian curves; three peaks are well separated, while the other four are partially overlapped, see Fig. 4a. These peaks are added a baseline and normally distributed random noise. The baseline can be either a sigmoidal, Gaussian, or exponential curve, with different signal-to-baseline ratio (SBR), defined as the height of the tallest peak ([signal maximum] − [signal minimum], before adding baseline or noise) relative to the baseline amplitude ([baseline maximum] − [baseline minimum]). Five

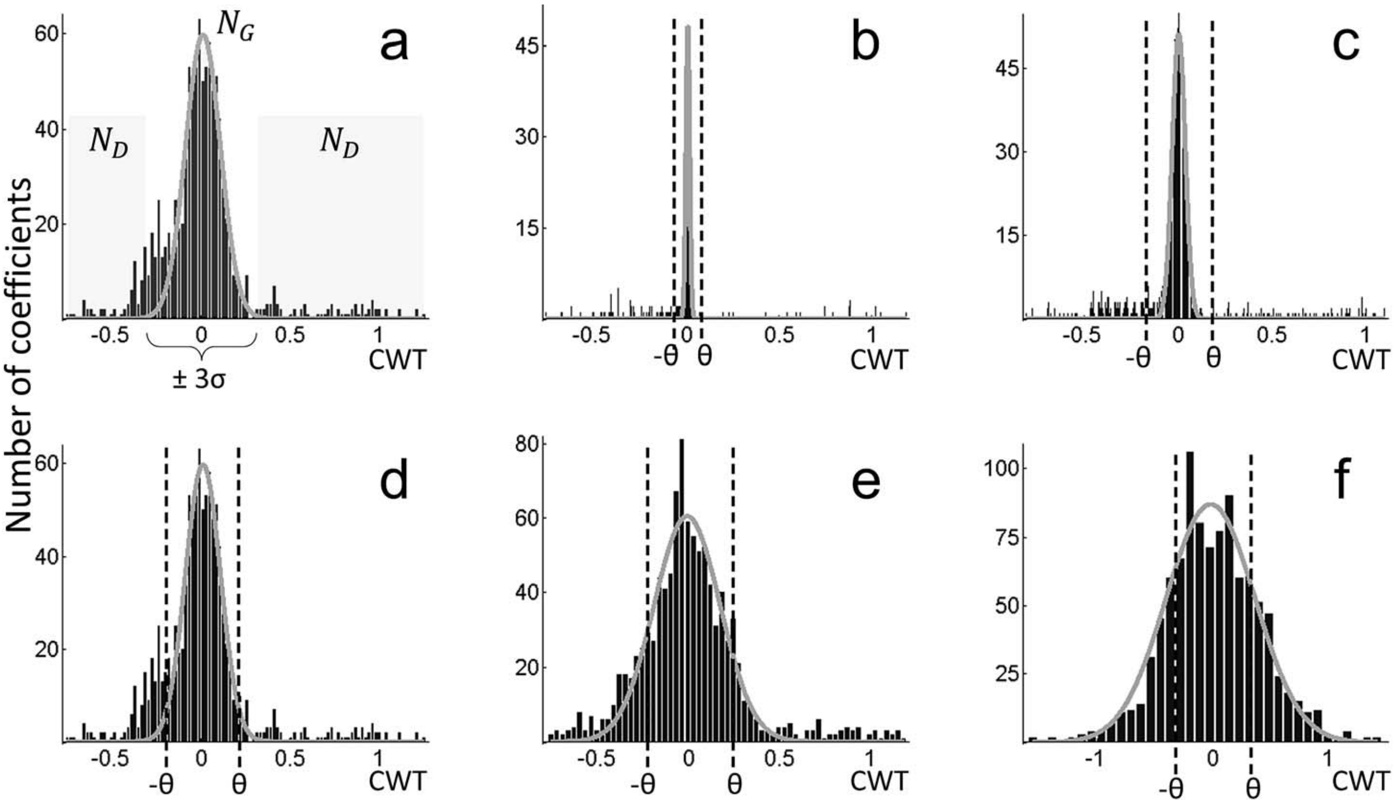

FIG. 3. Histograms of the probability distribution function of CWT values at scale $\hat{a}$ for the simulated spectra with sigmoidal baseline SBR = 1 and various SNRs (see text for acronyms). The Gaussian fit of the central part of the histogram is drawn as a gray line. (**a**) Spectrum with SNR = 10. The shaded rectangles indicate the values that fall within 3 times the standard deviation of the Gaussian fit (N_G) and those that are outside this range (N_D). (**b**)–(**f**) Spectra for SNR = 100, 30, 10, 6, 3, respectively. The vertical dashed lines indicate the threshold (θ) on the CWT. Note how θ is distanced from the bell-shaped curve for high SNR and gradually intersects it as noise increases.

spectra have a sigmoidal baseline with SBR = 1 and signal-to-noise ratio (SNR, defined as the height of the tallest peak relative to the standard deviation of the random noise) of 100, 30, 10, 6, and 3, respectively. Six spectra were generated with the sigmoidal baseline combining SBR of 0.1 and 0.01 with SNR of 100, 10, and 3, respectively. Eighteen spectra were generated combining SBR of 1, 0.1, and 0.01 with SNR of 100, 10, and 3 for the Gaussian and exponential baseline, respectively. The total number of simulated spectra is 29.

The CWT-BR combined with IPF was employed to correct the baseline of a large set of real spectra. In particular, we applied it to a Raman image of 200 × 200 pixels taken from a sample of Scots pine (Pinus sylvestris) using an Alpha300 R Confocal Raman microscope (Witec GmbH, Germany).[29] The considered wavenumber range was 160–3620 cm^{-1}.

RESULTS AND DISCUSSION

Simulated Spectra. Because the number of simulated data was small, the spectra on which to perform the entropy-based scale selection were picked manually instead of automatically. The Shannon entropy of the CWT vectors for scale parameters ranging from 1 to 60 was calculated for the three spectra with highest SBR and SNR. The minimum entropy was found for a scale parameter $\hat{a} = 5$, which was used for all simulated spectra.

As a quantitative measure of the quality of the baseline recognition, the specificity and sensitivity parameters[30] were calculated and are reported in Table I. Specificity is defined as the percentage of spectral points recognized as baseline among the "true" baseline points. Similarly, sensitivity is the percentage of points recognized as peak (i.e., not recognized as baseline) among the "true" peak points.† In this calculation, peak points were considered those with intensity greater than 4% of the maximum of the nearest peak, before adding baseline and random noise. It must be pointed out that this definition of peak regions is conventional; therefore, these quality parameters must be taken as indicative information rather than absolute measurement. Some visual examples of the results are shown in Fig. 4. In all the figures of spectra presented in this paper, the baseline points are marked by gray vertical lines.

For all spectra with SBR of 1 and 0.1, when the SNR was equal to 100, the baseline points were recognized almost perfectly, and their quality parameters are very close to 100%. A visual example of this outcome is shown in Fig. 4b. With SNR equal to 30 and 10, a slight decrease of the parameters was observed, especially for sensitivity. However, it must be pointed out that most of this decrease derives from a different recognition of

† These definitions conceive the problem as a peak recognition rather than a baseline recognition. Nevertheless, it was here preferred to leave these names unchanged.

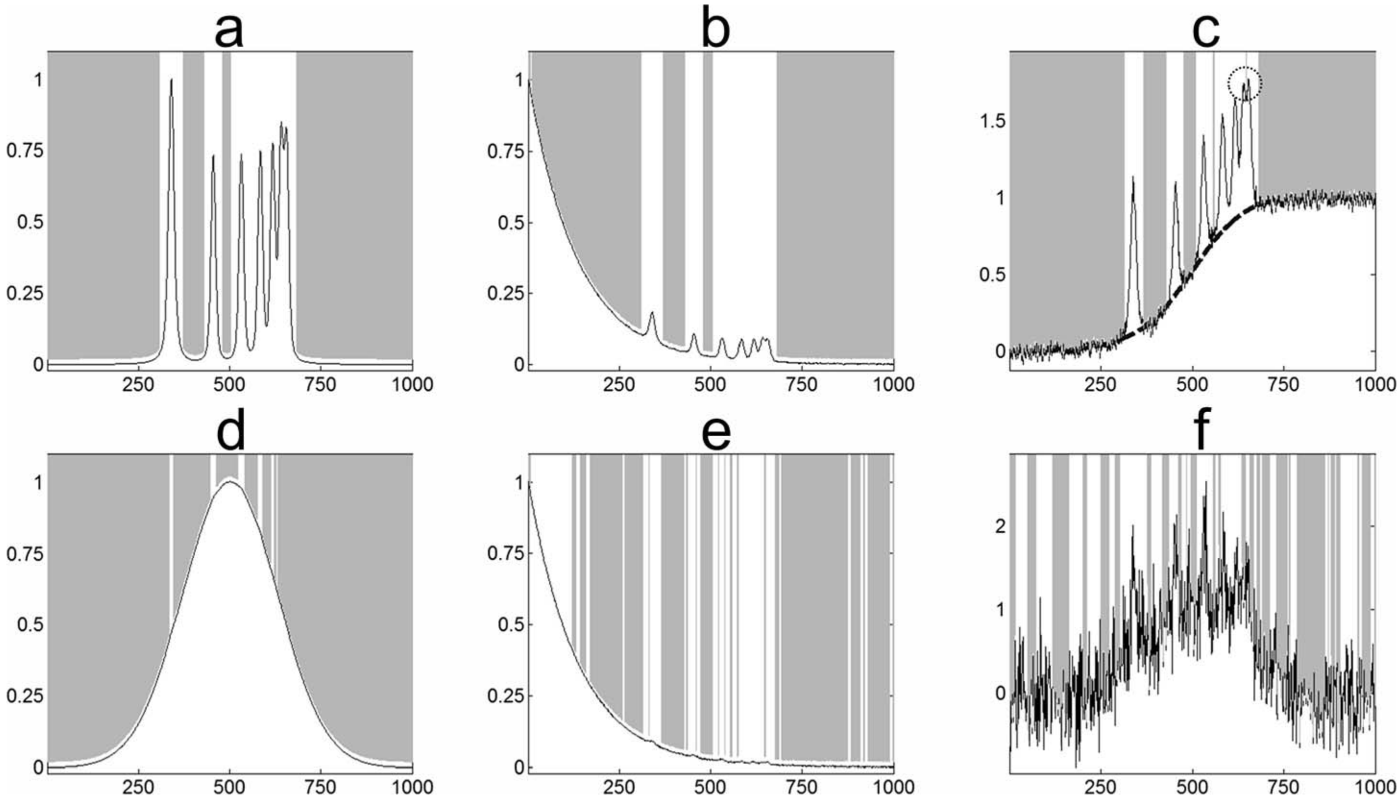

FIG. 4. Examples of simulated spectra and baseline regions, indicated by the gray shade. The abscissa indicates the vector element number; spectral intensities are in conventional units. (**a**) Simulated peaks before the addition of the baseline and random noise. Gray areas indicate the regions that are considered baseline for the calculation of sensitivity and specificity parameters. (**b**)–(**f**) Gray vertical lines indicate the points selected by the CWT-BR algorithm. Characteristics of the spectra: (**b**) exponential baseline, SBR = 0.1, SNR = 100; (**c**) sigmoidal baseline, SBR = 1, SNR = 30; (**d**) Gaussian baseline, SBR = 0.01, SNR = 100; (**e**) exponential baseline, SBR = 0.01, SNR = 10; (**f**) Gaussian baseline, SBR = 1, SNR = 3. (**c**) The dotted circle highlights a point between two congested peaks, which is interpreted as baseline by CWT-BR. The thick dashed line is a fit of the baseline points by an iteratively corrected Whittaker smoother (see text).

peak tail points. Because the exact baseline/peak distinction in these regions is somewhat conventional, this result cannot really be considered an error. Very few points unquestionably belonging to the baseline were identified as peaks. Provided that enough baseline points are recognized elsewhere, this kind of error is not detrimental for a subsequent fit. On the other hand, points at the top of congested peaks could sometimes be recognized as baseline, as can be seen in Fig. 4c for point 649 on the abscissa (highlighted by a circle). Although this assignment is not correct, it can be easily removed by iterative refinement methods. To provide an example of this possibility, the baseline points recognized in the spectrum in Fig. 4c (SBR = 1, SNR = 30) were fitted with a Whittaker smoother (WS),[31] setting a smoothing parameter $\lambda = 10^6$, removing all baseline points greater than the curve fit by more than three times the standard error of estimate and fitting again until the sequence of baseline points did not vary. Point 649 was removed after one iteration, and the resulting baseline approximation is plotted in Fig. 4c as a thick dashed line. It is not in the scope of this paper to carry out an extensive study on baseline estimation using WS with CWT-BR; therefore, this calculation was not repeated for the rest of the data. It is in our future plans, though, to devise an automated selection of the optimal value of λ for the WS fit, based on the noise level and information derived from multi-scale CWT.

For the spectra with SBR = 0.01, the baseline recognition was not always sufficiently correct because the CWT was sometimes more affected by the curvature of the baseline than by the tiny peaks. When noise was low (SNR = 100), the consequently small value of θ caused only the top of the peaks to be recognized as such, as can be observed in Fig. 4d. In the case of the exponential baseline, the leftmost side is interpreted as

TABLE I. Sensitivity and specificity parameters for the baseline recognition of simulated spectra.

Baseline curve type	SNR[a]	SBR[b] = 1		SBR = 0.1		SBR = 0.01	
		Sens.[c]	Spec.[d]	Sens.	Spec.	Sens.	Spec.
Sigmoidal	100	99	99	98	99	86	92
	30	90	100	–	–	–	–
	10	90	89	82	97	82	93
	6	82	66	–	–	–	–
	3	75	44	80	46	77	39
Gaussian	100	99	99	93	100	25	100
	10	84	96	84	98	84	70
	3	78	49	90	42	77	50
Exponential	100	99	99	97	99	83	84
	10	86	96	84	92	83	74
	3	80	48	85	46	81	50

[a] Signal-to-noise ratio.
[b] Signal-to-baseline ratio.
[c] Sensitivity.
[d] Specificity.

a peak because of the high CWT generated by the sharp interception between the steepest part of the curve and the constant 400 point extension, see Fig. 4e. This issue could probably be resolved by using another extension in which the curve reaches a constant value smoothly rather than instantly. It is worth noting that a mirror-image extension, which is a more common way to deal with border effects,[32] in this case, would enhance this artifact even more. It must also be pointed out that these spectra are intended as a test on the performance of the CWT-BR algorithm in cases of extremely weak signals.

As expected, for the spectra with SNR = 3, in which the peaks are almost completely swamped into the random noise fluctuations, the worst baseline recognition was obtained. Like the aforementioned cases with very low SBR, these spectra are to be considered another test on the CWT-BR with a particularly difficult signal. Figure 3f shows how a high noise level produces a Gaussian distribution so broad that it almost entirely mixes with the CWT values originated from the real signal. Nevertheless, the recognition of the peak points was rather good, as indicated by the sensitivity value of 75% or greater. On the other hand, several gaps in the baseline recognition were observed in the areas where noise gave rise to a profile similar to a peak, as in Fig. 4f. It must be stressed that these spots can be easily mistaken for a peak also by visual inspection. Considering that the assumptions and criteria used by CWT-BR are substantially not very different from those of manual methods, these kinds of errors could reasonably be expected.

For the reported simulated spectra, the CWT-BR algorithm was able to provide a satisfactory recognition of the baseline, except for extremely weak or noisy signals, in a few seconds and virtually without any human intervention (the only one being the indication of which were the best spectra). Although this method often does not outperform the human eye and is therefore not the most suitable for treating very tricky baselines, it appears to be a very good choice for the quick and reproducible processing of a large number of not particularly difficult signals. It is also more automatic than methods that recognize the baseline using CWT and iterative thresholding,[25,17] in which the user must indicate the scale parameter or at least the presence of broad peaks. Moreover, these methods need a rather large difference in curvature between peaks and baseline, or they may not converge (although the user may toggle a different recognition strategy in the presence of broad peaks). The CWT-BR yields results even when the peak-baseline curvature difference is small, although a large difference is surely beneficial.

Experimental Raman Spectra. Cosmic ray peaks were removed from the Raman image of a wood cross section by the "Cosmic Ray Removal" function of WITec Project 2.10 (Witec GmbH, Germany). A scale parameter $\hat{a} = 8$ was selected by applying the procedure described in the Methods section. Each spectrum was then processed by CWT-BR and the iterative polynomial fitting (IPF) algorithm. Because in this case it is not possible to objectively know the true position of peaks and shape of baseline, the quality of the processing can be evaluated only by visual inspection. Four spectra were picked from very different areas of the image and are shown in Fig. 5 as representatives of the overall results. In particular, the spectra from the first three rows of Fig. 5 are taken from the middle lamella, secondary cell wall, and hollow lumen (filled with epoxy resin), respectively.[29] The spectrum in the fourth row is also from the secondary cell wall, but has a lower SNR as compared to the one in the second row. The baseline points and polynomial baseline approximation after CWT-BR and IPF, respectively, as well as the final baseline-subtracted curve are depicted for each spectrum.

As can be observed in the left-side column of Fig. 5, after applying only the CWT-BR algorithm all the major spectral features were recognized. However, several points were assigned as baseline in congested peak regions, and broad peaks were recognized only partially. As a consequence, the polynomial fit (dotted line) of these recognized points overestimates the spectra in some areas and, supposedly, the true baseline. These results show that, although CWT-BR was able to deal with congested peaks in the simulated spectra of the previous section, it still needs to be improved to analyze spectra with a greater variability in peak breadth, probably by taking more than one CWT scale into consideration.

The middle column of Fig. 5 shows the baseline approximation and fitting points after using the IPF algorithm. This operation removed all the spurious assignments, except for few noisy low-value points, and the final fit appears substantially correct. The results can be even more appreciated in the right-side column, which shows the final corrected spectra after baseline subtraction. Within the precision allowed by noise, these spectra are all aligned on the horizontal zero line, and no negative regions are observed. It can be noticed that the baseline correction does not perform nor require any prior treatment, such as smoothing, which could risk removing or distorting parts of the signal. Although the use of a fifth-degree polynomial plays a big role in the final outcome and is not applicable to any type of spectrum, this approach that combines CWT-BR and IPF has a potential advantage over methods based only on iterative fitting[13] because it does not depend entirely on the choice of the interpolating curve.

A simple way to perform a statistical evaluation of the effect of the performed baseline correction on the whole image is through principal component analysis (PCA).[33] Figure 6 plots the first four principal components (denoted PC1–4) for the raw Raman image (left column) and the baseline-corrected one (right column), which explain 98% and 85% of the total variance, respectively. Here, PC1 and PC2 of the raw spectra show clear features of the baseline curvature, whereas the corresponding PCs after baseline correction are predominantly flat on the zero line except in peak regions. Also, PC4 shows a slope for both the raw and corrected spectra, but it is considerably less steep in the latter case.

The computation of the CWT-BR and IPF on our platform proceeded at a rate of 4–5 spectra per second, and the treatment of all the 40 000 spectra in the image took about 2.5 h. It is difficult to compare this speed with that of other methods in the literature because most of

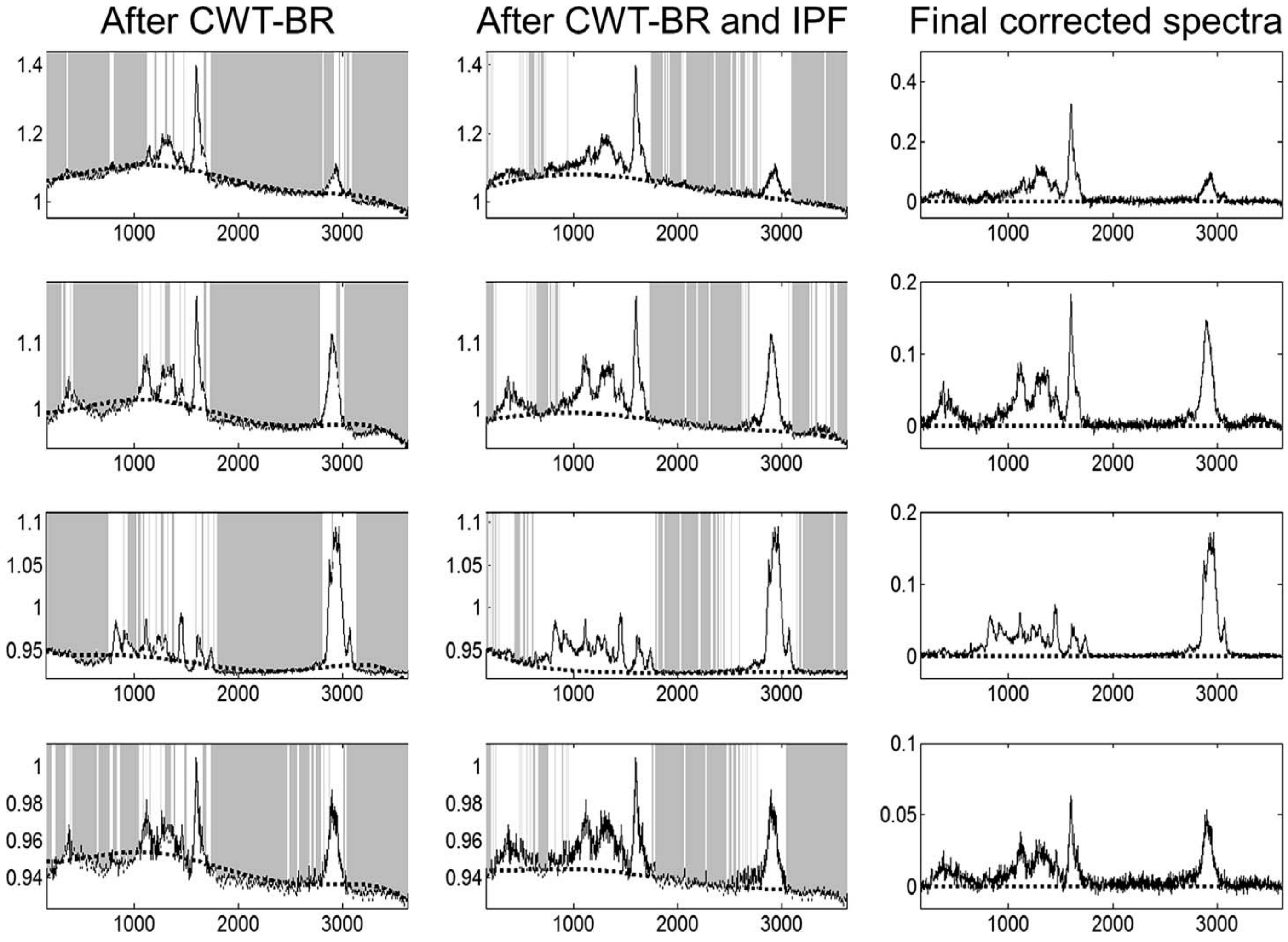

FIG. 5. Real Raman spectra taken from an image of *Pinus sylvestris* and outcome of different preprocessing steps. The abscissa is in wavenumber units (cm^{-1}), spectral intensities are in arbitrary units. The polynomial fit of baseline points, indicated by gray vertical lines, is drawn as a dotted curve. Left column: spectra after application of CWT-BR and simple polynomial fit. Middle column: spectra after application of CWT-BR and IPF. Right column: final corrected spectra after subtraction of the baseline estimation. First row (from top): spectrum from middle lamella. Second row: spectrum from secondary cell wall. Third row: spectrum from hollow lumen. Fourth row: spectrum from secondary cell wall with lower SNR than the spectrum in the second row.

the authors do not report a precise value, if they report any at all, and very few carry out such large calculations. However, it is among the fastest methods reported. Perhaps more importantly, this duration is of the same order as the duration of the corresponding experimental measurement. Since in nearly all laboratories there is enough computing power to carry out a measurement and some calculations simultaneously, this preprocessing algorithm can be used by researchers without slowing down their work.

The final results on the presented Raman image show that the CWT-BR algorithm, despite its current limitations when used alone, can already have successful practical applications if combined with iterative refinement methods. The required background knowledge consisted of the fitting function and the notion that all the considered spectra have similar types of signals. These requirements are suitable for many typical experimental setups, in which the approximate characteristics of the peaks are known in advance, though their exact shape and position vary.

The overestimation of the baseline after the first step, i.e., CWT-BR and simple polynomial fit, was not a significant problem after employing IPF. Nevertheless, a more accurate recognition is desirable as it would allow for using CWT-BR with many other fitting methods. As a future development, we expect to improve the algorithm by using information from more than one CWT scale, for example, exploiting several local minima of the entropy-scale curve, and by making it work with other types of noise distribution, such as the Poissonian. The possibility of combining CWT-BR with different fitting techniques may lead to a very flexible approach, in which the user could opt for a more general baseline correction or a more constrained (and supposedly faster) one, depending on the available background knowledge. Our ultimate purpose is to devise, with the help of other mathematical tools, a baseline correction methodology in which the only information provided by the user is grouping the data into sets of analogous measurements. These sets are often defined by the structure of the instrumental output, e.g., images, time sequences, etc.

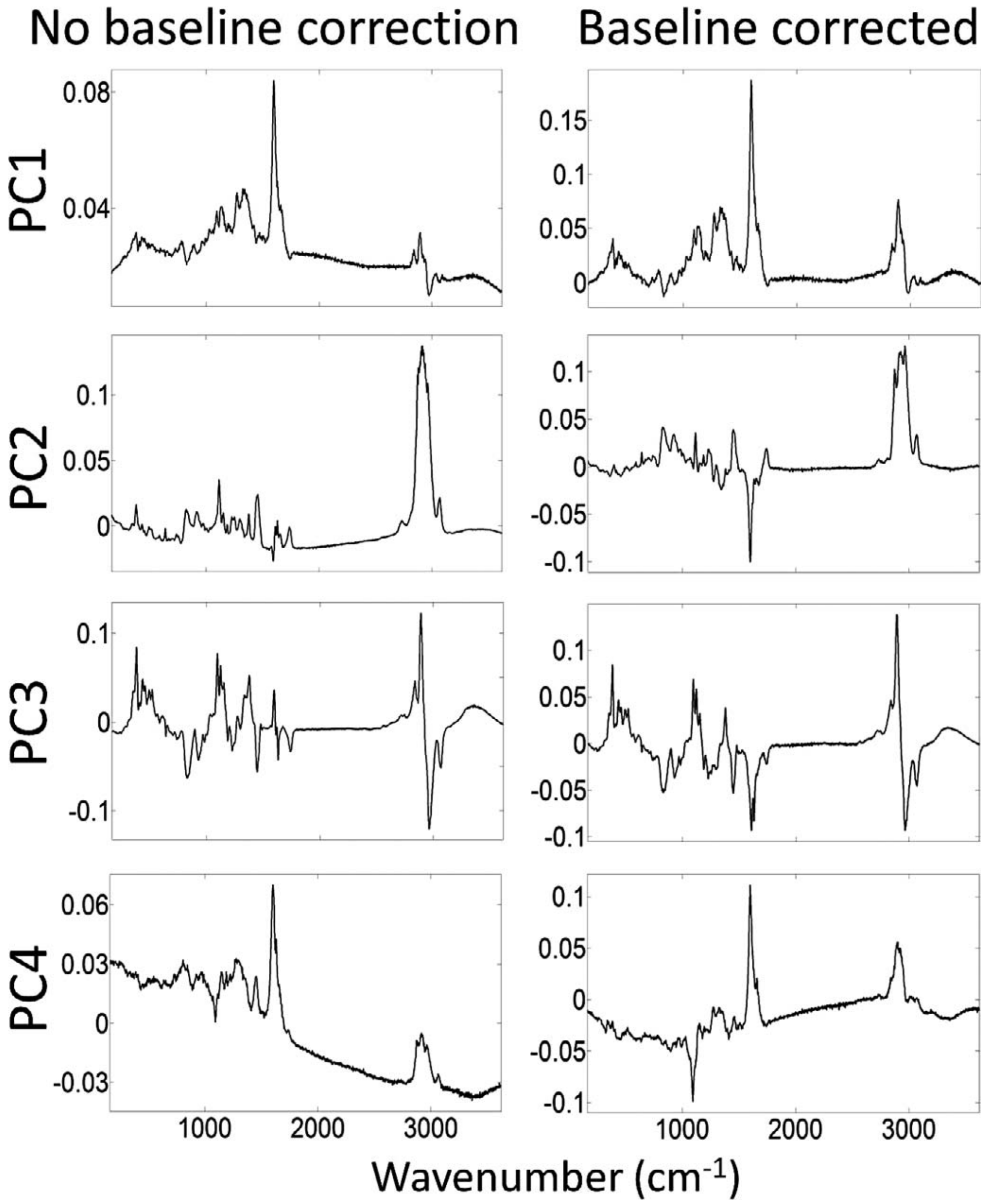

FIG. 6. First four principal components (PCs) of the Raman image of Pinus sylvestris using raw spectra (left column) and baseline corrected spectra (right column), respectively.

The CWT-BR algorithm, which functions needing only the definition of the measurement dataset, is a step towards the realization of such a methodology.

CONCLUSIONS

A new algorithm, denoted as CWT-BR, to recognize baseline points in spectra was presented in this paper. It is based on the continuous wavelet transform and uses the criteria of Shannon entropy and statistical noise distribution to automatically determine its parameters. It works under certain assumptions that are frequently found when analyzing large sets of spectra of the same type, in which at least a part of them are of good quality.

On the reported simulated spectra, the algorithm was able to recognize the baseline points correctly, except for extremely weak or noisy signals. It was shown that the wrong assignments can be easily corrected in a subsequent phase by appropriate iterative fitting methods.

On the spectra of a real Raman microscopy image, the use of CWT-BR alone did not produce a fully satisfactory result, but a very good one was obtained after combining it with an iterative polynomial fitting (IPF) algorithm. The quality of the baseline correction was evaluated by observing a few sample spectra and by principle component analysis. The whole calculation required no significant human intervention and was carried out in a time short enough for practical application on sets containing thousands of spectra.

The presented experiments show that CWT-BR is a promising tool for baseline correction and can already be successfully applied in combination with appropriate fitting algorithms. Its characteristics make it particularly suitable for a fast, simple, and automatic treatment of large series of spectra, such as the ones produced by high-throughput instruments. Further development of the methodology is prospected, with the aim of making it applicable to more general cases.

ACKNOWLEDGMENTS

The financial contribution by the Academy of Finland is gratefully acknowledged.

1. L. Shao, P.R. Griffiths. "Automatic Baseline Correction by Wavelet Transform for Quantitative Open-Path Fourier Transform Infrared Spectroscopy". Environ. Sci. Technol. 2007. 41(20): 7054-7059. doi:10.1021/es062188d.
2. M.N. Leger, A.G. Ryder. "Comparison of Derivative Preprocessing and Automated Polynomial Baseline Correction Method for Classification and Quantification of Narcotics in Solid Mixtures". Appl. Spectrosc. 2006. 60(2): 182-193. doi: 10.1366/000370206776023304.
3. Z. Xu, X. Sun, P. de B. Harrington. "Baseline Correction Method Using an Orthogonal Basis for Gas Chromatography/Mass Spectrometry Data". Anal. Chem. 2011. 83(19): 7464-7471. doi:10.1021/ac2016745.
4. J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan. "Asymmetric Least Squares for Multiple Spectra Baseline Correction". Anal. Chim. Acta. 2010. 683(1): 63-68. doi:10.1016/j.aca.2010.08.033.
5. J.A. Westerhuis, S. de Jong, A.K. Smilde. "Direct Orthogonal Signal Correction". Chemom. Intell. Lab. Syst. 2001. 56(1): 13-25. doi:10.1016/S0169-7439(01)00102-2.
6. J. Peng, S. Peng, Q. Xie, J. Wei. "Baseline Correction Combined Partial Least Squares Algorithm and Its Application in On-Line Fourier Transform Infrared Quantitative Analysis". Anal. Chim. Acta. 2011. 690(2): 162-168. doi:10.1016/j.aca.2011.02.001.
7. H.W. Tan, S.D. Brown. "Wavelet Analysis Applied to Removing Non-Constant, Varying Spectroscopic Background in Multivariate Calibration". J. Chemom. 2002. 16(5): 228-240. doi:10.1002/cem.717.
8. Y. Ding, D. Peng. "Quantitative Analysis of Near-Infrared Spectra by Wavelet-Based Interferences Removal and Least Squares Support Vector Regression". J. Comput. 2012. 7(4): 880-889. doi:10.4304/jcp.7.4.880-889.
9. J. Palacký, P. Mojzeš, J. Bok. "SVD-Based Method for Intensity Normalization, Background Correction and Solvent Subtraction in Raman Spectroscopy Exploiting the Properties of Water Stretching Vibrations". J. Raman Spectrosc. 2011. 42(7): 1528-1539. doi:10.1002/jrs.2896.
10. A. Jirasek, G. Schulze, M.M. Yu, M.W. Blades, R.F. Turner. "Accuracy and Precision of Manual Baseline Determination". Appl. Spectrosc. 2004. 58(12): 1488-1499.
11. G. Schulze, A. Jirasek, M.M. Yu, A. Lim, R.F. Turner, M.W. Blades. "Investigation of Selected Baseline Removal Techniques as Candidates for Automated Implementation". Appl. Spectrosc. 2005. 59(5): 545-574.
12. C. Rowlands, S. Elliott. "Automated Algorithm for Baseline Subtraction in Spectra". J. Raman Spectrosc. 2011. 42(3): 363-369. doi:10.1002/jrs.2691.
13. J. Zhao, H. Lui, D.I. McLean, H. Zeng. "Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy". Appl. Spectrosc. 2007. 61(11): 1225-1232.
14. A. Antoniadis, J. Bigot, S. Lambert-Lacroix, F. Letue. "Nonparametric Pre-Processing Methods and Inference Tools for Analyzing Time-of-Flight Mass Spectrometry Data". Curr. Anal. Chem. 2007. 3(2): 127-147. doi:10.2174/157341107780361718.
15. A.T. Weakley, P.R. Griffiths, D.E. Aston. "Automatic Baseline Subtraction of Vibrational Spectra Using Minima Identification and Discrimination via Adaptive, Least-Squares Thresholding". Appl. Spectrosc. 2012. 66(5): 519-529. doi:10.1366/110-06526.
16. Z.M. Zhang, S. Chen, Y.Z. Liang. "Baseline Correction Using Adaptive Iteratively Reweighted Penalized Least Squares". Analyst. 2010. 135(5): 1138-1146. doi:10.1039/b922045c.
17. Q. Bao, J. Feng, F. Chen, W. Mao, Z. Liu, K. Liu, C. Liu. "A New Automatic Baseline Correction Method Based on Iterative Method". J. Magn. Reson. 2012. 218: 35-43. doi:10.1016/j.jmr.2012.03.010.
18. H.G. Schulze, R.B. Foist, K. Okuda, A. Ivanov, R.F. Turner. "A Model-Free, Fully Automated Baseline-Removal Method for Raman Spectra". Appl. Spectrosc. 2011. 65(1): 75-84. doi:10.1366/10-06010.
19. H.G. Schulze, R.B. Foist, K. Okuda, A. Ivanov, R.F. Turner. "A Small-Window Moving Average-Based Fully Automated Baseline Estimation Method for Raman Spectra". Appl. Spectrosc. 2012. 66(7): 757-764. doi:10.1366/11-06550.
20. B.D. Prakash, Y.C. Wei. "A Fully Automated Iterative Moving Averaging (AIMA) Technique for Baseline Correction". Analyst. 2011. 136(15): 3130-3135. doi:10.1039/c0an00778a.
21. R. Perez-Pueyo, M.J. Soneira, S. Ruiz-Moreno. "Morphology-Based Automated Baseline Removal for Raman Spectra of Artistic Pigments". Appl. Spectrosc. 2010. 64(6): 595-600. doi: 10.1366/000370210791414281.
22. N. Kourkoumelis, A. Polymeros, M. Tzaphlidou. "Background Estimation of Biomedical Raman Spectra Using a Geometric Approach". Spectrosc. Int. J. 2012. 27(5–6): 441-447. doi:10.1155/2012/530791.
23. K.H. Liland, E.O. Rukke, E.F. Olsen, T. Isaksson. "Customized Baseline Correction". Chemom. Intell. Lab. Syst. 2011. 109(1): 51-56. doi:10.1016/j.chemolab.2011.07.005.
24. Y. Meyer. "Wavelets—Algorithms and Applications". In: R.D. Ryan, editor. Wavelets: Algorithms and Applications. Philadelphia, USA: Society for Industrial and Applied Mathematics Translation, 1993.
25. J.C. Cobas, M.A. Bernstein, M. Martín-Pastor, P.G. Tahoces. "A New General-Purpose Fully Automatic Baseline-Correction Procedure for 1D and 2D NMR Data". J. Magn. Reson. 2006. 183(1): 145-151. doi:10.1016/j.jmr.2006.07.013.
26. D. Lyder, J. Feng, B. Rivard, A. Gallie, E. Cloutis. "Remote Bitumen Content Estimation of Athabasca Oil Sand from Hyperspectral Infrared Reflectance Spectra Using Gaussian Singlets and Derivative of Gaussian Wavelets". Fuel. 2010. 89(3): 760-767. doi:10.1016/j.fuel.2009.03.027.
27. P. Du, W. Kibbe, S. Lin. "Mass Spectrum Processing by Wavelet-Based Algorithms". 2009. http://www.bioconductor.org/packages/2.11/bioc/html/MassSpecWavelet.html [accessed [Jan 21 2013].
28. C.A. Lieber, A. Mahadevan-Jansen. "Automated Method for Subtraction of Fluorescence from Biological Raman Spectra". Appl. Spectrosc. 2003. 57(11): 1363-1367.
29. T. Hänninen, E. Kontturi, T. Vuorinen. "Distribution of Lignin and Its Coniferyl Alcohol and Coniferyl Aldehyde Groups in Picea Abies and Pinus Sylvestris as Observed by Raman Imaging". Phytochemistry. 2011. 72(14-15): 1889-1895. doi:10.1016/j.phytochem.2011.05.005.
30. S.K. Lau, P. Winlove, J. Moger, O.L. Champion, R.W. Titball, Z.H. Yang, Z.R. Yang. "A Bayesian Whittaker–Henderson Smoother for General-Purpose and Sample-Based Spectral Baseline Estimation and Peak Extraction". J. Raman Spectrosc. 2012. 43(9): 1299-1305. doi:10.1002/jrs.3165.
31. P.H.C. Eilers. "A Perfect Smoother". Anal. Chem. 2003. 75(14): 3631-3636. doi:10.1021/ac034173t.
32. G. Strang, T.Q. Nguyen. Wavelets and Filter Banks. Wellesley, MA, USA: Wellesley-Cambridge Press, 1996.
33. H. Abdi, L.J. Williams. "Principal Component Analysis". WIREs Comput. Stat. 2010. 2(4): 433-459. doi:10.1002/wics.101.

Queries for apls-68-02-02

1. Please verify the references for the sentence beginning, “A fifth-order polynomial was used, which is. . .” as it appears to be missing a number. Editor