A Model-Free, Fully Automated Baseline-Removal Method for Raman Spectra

H. GEORG SCHULZE, ROD B. FOIST, KADEK OKUDA, ANDRÉ IVANOV, and ROBIN F. B. TURNER*

Michael Smith Laboratories, The University of British Columbia, 2185 East Mall, Vancouver, BC, Canada, V6T 1Z4 (H.G.S., R.B.F., K.O., R.F.B.T.); Department of Electrical and Computer Engineering, The University of British Columbia, 2332 Main Mall, Vancouver, BC, Canada, V6T 1Z4 (R.B.F., A.I., R.F.B.T.); and Department of Chemistry, The University of British Columbia, 2036 Main Mall, Vancouver, BC, Canada, V6T 1Z1 (K.O., R.F.B.T.)

We present here a fully automated spectral baseline-removal procedure. The method uses a large-window moving average to estimate the baseline; thus, it is a model-free approach with a peak-stripping method to remove spectral peaks. After processing, the baseline-corrected spectrum should yield a flat baseline and this endpoint can be verified with the χ^2-statistic. The approach provides for multiple passes or iterations, based on a given χ^2-statistic for convergence. If the baseline is acceptably flat given the χ^2-statistic after the first pass at correction, the problem is solved. If not, the non-flat baseline (i.e., after the first effort or first pass at correction) should provide an indication of where the first pass caused too much or too little baseline to be subtracted. The second pass thus permits one to compensate for the errors incurred on the first pass. Thus, one can use a very large window so as to avoid affecting spectral peaks—even if the window is so large that the baseline is inaccurately removed—because baseline-correction errors can be assessed and compensated for on subsequent passes. We start with the largest possible window and gradually reduce it until acceptable baseline correction based on the χ^2-statistic is achieved. Results, obtained on both simulated and measured Raman data, are presented and discussed.

Index Headings: **Automated baseline subtraction; Model-free baseline removal; Peak stripping; Moving average; Raman spectroscopy; Vibrational spectroscopy; Baseline-free spectra.**

INTRODUCTION

Due to an increase in high-throughput instrumentation and otherwise improved and accelerated collection efficiencies, the automated processing of large spectral collections of data is increasingly sought after.[1–9] One of the common processing needs—and one of the most difficult to automate—is the removal of spurious background signals,[9–11] which, in general, may interfere with further processing,[12–14] complicate quantification,[9,14,15] or hinder the presentation and visualization of relevant data.[16] For example, when raster scanning a plant leaf to obtain information about the distribution of its chemical constituents using Raman microspectroscopy[17,18] over an area of 70 μm × 100 μm with 5 μm resolution, over 300 spectra with steeply sloping baselines are generated. In the absence of an automated method, these spectra need to be corrected by hand. Clearly, this can require a considerable amount of time and may lead to fatigue accompanied by increased likelihood of operator error/bias. For example, based on a variety of spectra, we have found the average time to correct a baseline manually to be about 100 s.[11] Thus, correcting 300 spectra would take in excess of eight hours to correct manually. In addition to combating fatigue and maintaining consistency, it is furthermore desirable to perform automated baseline correction with the same or better speed and accuracy than attainable manually by appropriately trained operators.[11,19]

Despite a number of near-fully automated methods being devised,[9,11,14,20–23] with the exception of a second-derivative-based method by Rowlands and Elliott[24] requiring low-noise spectra, a truly model-free, fully automated procedure remains elusive. Extant methods often require filter coefficient settings, wavelet selection, polynomial specification, stopping criterion specification, and so on. One possible approach to this problem is to use a method that can be implemented with its parameters initially at their extremes and to iterate subsequently to their optima or to acceptable values. The latter thus requires a suitable stopping criterion. We have previously used this approach to fully automate a Savitzky–Golay smoothing filter with initial parameters of a zero order and a three-point window and iterating until meeting the widely used χ^2 stopping criterion.[25] We follow a similar approach here to create a model-free baseline-removal method that can be implemented in a fully automated and "parameter-free" manner, the latter in the sense that user intervention is not required.

Savitzky–Golay filters are widely known filters and have become the standard against which other smoothing methods are compared;[26–28] they are easy to implement, fast to execute, and provide generally good results.[28] The zero-order Savitzky–Golay filter is also known as a moving average filter. A general characteristic of smoothing filters is that lower frequencies are passed and that the passband of low frequencies narrows as the window size increases.[26] Thus, a very large window zero-order Savitzky–Golay filter could be used to pass the baseline but attenuate high-frequency noise and, to a lesser extent, the signals of interest. When combining this approach iteratively with peak stripping, effective baseline estimation can be obtained. We have therefore investigated the potential of implementing the zero-order Savitzky–Golay filter, in conjunction with peak stripping and the Pearson χ^2 stopping criterion as further explained below, to produce a truly model-free and fully automated baseline removal method.

THEORY

The Baseline-Estimation Problem. Measured data obtained on N detector channels can be represented as

$$\mathbf{m} = (\mathbf{b} + \mathbf{x}) * \mathbf{p} + \mathbf{n} = \mathbf{b} * \mathbf{p} + \mathbf{x} * \mathbf{p} + \mathbf{n} \qquad (1)$$

where the underlying signal vector, **x**, plus the baseline, **b**, is convolved (*) with an instrumental blurring function, **p**, and measurement noise, **n**, is added. The purpose of baseline estimation is to produce an estimate of **b** from **m**. We use a

Received 21 May 2010; accepted 28 September 2010.
* Author to whom correspondence should be sent. E-mail: turner@msl.ubc.ca.
DOI: 10.1366/10-06010

0003-7028/11/6501-0075$2.00/0

large-window zero-order Savitzky–Golay filter (i.e., a moving-average filter) to estimate the baseline, where the estimated baseline is given by the filter output. Baseline estimation is improved by stripping signals above the baseline from the spectrum. This process is iterated until no more signal is removed, that is, until the signal-to-noise ratio (SNR) of the remaining signals reaches a specified value, or until some termination threshold is reached.[11]

The Automation Problem. Filter design requires the selection of a number of different filter parameters as mentioned above. In the case of Savitzky–Golay filters, this is particularly uncomplicated since there are only two parameters to select, these being the number of points in the local modeling window and the polynomial order used for local modeling of the data. In the present case, this is further simplified by using the zero-order filter because it has the best low-pass characteristics among polynomial filters.[26,29,30] Nevertheless, selection of the window size is required, making automation difficult.[30] In general, a bigger window leads to more noise rejection and also to more signal distortion.[30] However, signal distortion is not problematic because we employ signal stripping as part of the baseline-estimation procedure. Thus, we overcome the problem of window size selection by starting with the largest possible window and decrementing its size as needed.

Instead of the criteria above regarding when to terminate peak stripping, one could take an alternative approach by asking when the baseline is estimated well enough. The answer to this is simple: the baseline should be, within reason, flat. When baseline correction is applied *to the* "*corrected*" *spectrum*, the estimated baseline should be flat after peak stripping. Moreover, flatness can be assessed with the χ^2-statistic. Importantly, if the baseline is not flat, the estimate obtained on the second correction (i.e., on the "corrected" spectrum) will reflect regions of over-correction or under-correction of the first correction. Thus, a better estimate of the baseline is arrived at by summation of the successive estimates on prior results. As soon as an estimate is flat (i.e., the baseline is zero), further estimates will all be flat (disregarding accumulation of small errors due to the peak-stripping operations) and no further improvements accrue in the overall estimate. Estimation should therefore be terminated when an acceptably flat baseline is obtained. Thus,

$$\hat{\mathbf{b}} = \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2 + \ldots + \hat{\mathbf{b}}_i, \qquad \hat{\mathbf{b}}_i \neq 0 \tag{2}$$

where $\hat{\mathbf{b}}_i$ indicates the ith successive estimate of the baseline $\hat{\mathbf{b}}$, not *de novo*, but on the spectrum after termination of the previous correction sequence.

If the window is too large and baseline correction cannot be effected, there should be a failure of convergence for the χ^2-statistic (between a flat line and $\hat{\mathbf{b}}_i$) with increasing i. Thus, in the absence of a decreasing χ^2-statistic, the window size is reduced

$$\text{window size} = N - n, \qquad n = 1, 2, 3, \ldots, N-3 \tag{3}$$

and the process is started anew. In the results reported here, though, we reduced the window size to speed computations according to

$$\text{window size} = \text{integer}(N/n), \qquad n = 1, 2, 3, \ldots, N/2 \tag{4}$$

The χ^2-statistic is widely used as a measure of similarity between different distributions and often an acceptable value for this statistic is taken as N.[29] When the χ^2-statistic exceeds this value, the baseline of the ith corrected spectrum is not acceptably flat and further iterations are required if there is evidence of convergence. Thus, as is common in regularization methods,[29] we employ the χ^2-statistic as stopping criterion here. Since calculating the latter is based on a very simple automated noise estimation procedure,[31] user intervention is not required.

METHODS

The automated baseline estimation (ABE) method was tested on synthetic data, simulated to represent Raman spectra, and real Raman spectra. Briefly, we generated seven Lorentzian peaks, the first with a full width at half-maximum (FWHM) of approximately ten channels and the remainder with FWHM of six channels each, resulting in an average FWHM of seven channels. Spaced at reduced intervals, these peaks then created uncongested and congested spectral regions. They were convolved with an instrumental point spread function approximated by a Gaussian distribution (five channels to 1σ) to simulate instrumental blurring and were added to a baseline. Three types of baseline were generated: (1) an exponential baseline, (2) a Gaussian baseline, and (3) a sigmoidal baseline (cumulative Gaussian) to investigate the method's performance on a variety of dissimilar baselines. Baselines had signal-to-baseline ratios (SBR, defined as the intensity of the tallest peak [signal maximum − signal minimum] relative to the baseline height [baseline maximum − baseline minimum]) of 1, 0.1, and 0.01. Synthetic spectra of 1001 points each were generated with a signal-to-noise ratio (SNR, defined as the intensity of the tallest peak [signal maximum − signal minimum] relative to the baseline noise) of 10 by adding a constant level of Gaussian (white) noise with an original standard deviation of 1.0 to the spectrum consisting of seven Lorentzian peaks. We simulated spectra with a constant but reasonably challenging level of noise, irrespective of baseline intensities, in order to keep the number of variables tractable. Each of the nine sets (three baseline types with three SBRs each) consisted of ten identical spectra but with unique noise distributions. Spectral ends were extrapolated before baseline correction with a second-order polynomial to provide padding (of window size) in order to mitigate potential edge effects from baseline correction. Typical examples of spectra from these sets are shown in Fig. 1.

Overall, the baseline-correction approach consists of three nested sets of iterations. The innermost routine performs peak stripping, the middle routine checks for baseline flatness and invokes the inner routine using the current window size as long as further improvements in baseline flatness (of the already processed spectrum) can be obtained, and the outer routine applies the stopping criterion or adjusts the window size with subsequent invocation of the middle routine. Details follow below.

Starting with the largest possible window (i.e., of the same size as the spectrum), a baseline is estimated by convolving the spectrum with a zero-order Savitzky–Golay filter. After baseline estimation, all spectral regions above the baseline are removed. More specifically, peaks above the estimated baseline plus twice the noise standard deviation (as determined from the starting spectrum) are reduced to the level of the

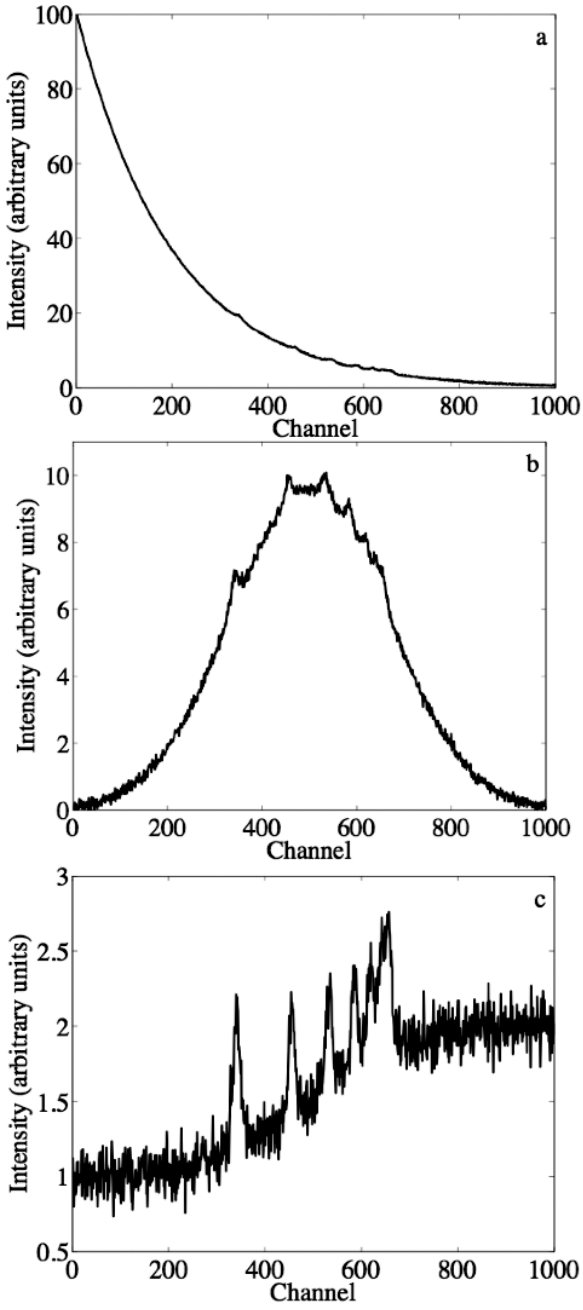

FIG. 1. Sample simulated Raman spectra superimposed on three baseline types: (**a**) an exponential baseline; (**b**) a Gaussian distribution baseline; and (**c**) a sigmoidal baseline. Baselines of all three types also had three signal-to-baseline ratios each, for example approximately (**a**) 0.01, (**b**) 0.1, and (**c**) 1. All nine sets of spectra contained 10 identical spectra but with independent, normally distributed noise to give a signal-to-noise ratio (tallest peak/baseline noise) of ~10.

baseline. Thereafter, zero-mean random Gaussian noise, estimated from the starting spectrum,[31] is optionally added (discussed below) to those positions where peak stripping occurred. Thus, although noise in the Raman spectra may be heteroscedastic for high SNR spectra, the spectrum after progressive stripping of its peaks retains noise characteristics generally similar to those of the original spectrum. Noise injection is aimed at diminishing possible distortions arising from sharp discontinuities between stripped and intact regions. The baseline estimation followed by stripping is then repeated on the resulting partially stripped spectrum. It is terminated when the maximum number of iterations is exceeded, no more stripping occurs, or aggressive stripping causes the estimated baseline to become negative. The estimated baseline, final stripped spectrum, corrected spectrum (i.e., the starting spectrum − estimated baseline), and number of stripping iterations required are returned as outputs from the peak stripping (i.e., the innermost) routine.

A well-corrected baseline is expected to be flat and of zero intensity. If the χ^2-statistic between a flat line and the estimated baseline continues to decrease, the corrected spectrum is again subjected to the peak-stripping baseline-estimation procedure. These baseline-testing iterations provide successive intermediate baseline estimates and are recorded along with their χ^2-values and their sequence number. If not, the process is terminated and a concluding baseline estimate is made by autosmoothing[25] the final stripped spectrum. An overall baseline estimate is then made by summing the successive intermediate estimates obtained from χ^2-based testing. This baseline is subtracted from the original spectrum to obtain a baseline-corrected spectrum.

If the minimum χ^2-value recorded above exceeds the χ^2 stopping criterion, the window size is decreased according to Eq. 4 and the entire procedure is restarted using the original spectrum. The process is terminated when the χ^2 stopping criterion is met and when repeated subtractions no longer provide an improvement in χ^2-values. At this point, a good guess of the window size and the number of estimates required for baseline flattening is available. However, due to noise in the spectrum and noise added when stripping peaks, there is some variability in the parameters obtained: repeated applications of the procedure to exactly the same spectrum leads to slightly different parameter outcomes. Therefore, parameter space is then doubled and a search is made for the parameters corresponding to the best χ^2-value. These are then used to obtain the final result. Along with χ^2-values, the root-mean-square (rms) error for final estimated baselines, vis-à-vis synthetic baselines, is determined as a figure of merit. A flow chart is given in Fig. 2.

To assess the performance of the ABE procedure on real data, Raman spectra were obtained from samples of solid triacontanol and triacontanoic acid as well as a tomato skin. The tomato skin was mechanically removed from the fruit. Raman spectra were collected on a Raman microscope with a charge-coupled device (CCD) detector (RM 1000, Renishaw, Gloucestershire, UK). A 50×/0.75 NA (numerical aperture) objective (Leica Microsystems, Wetzlar, Germany) was used, resulting in power at the focal point from the 785 nm diode laser of ~35–45 mW. Spectra were collected over a 350 to 2000 cm^{-1} range with collection times of 10–30 s, one to two accumulations, and using a 100 μm slit width.

Matlab 7.0 (The MathWorks, Natick, MA) was used to implement the ABE procedure. The computation platform was a 2.5 GHz dual processor PowerPC G5 running under Mac OS X 10.4.1 (Apple Computer, Santa Clara, CA). Code for the ABE procedure is available from the authors upon request.

RESULTS AND DISCUSSION

Simulated Spectra. The sample spectra in Fig. 1 suggest that manual correction of such spectra may be difficult. Where the SBR is very low (e.g., Fig. 1a) signal peaks are hardly noticeable. In other cases, even with moderate SBR (e.g., Fig.

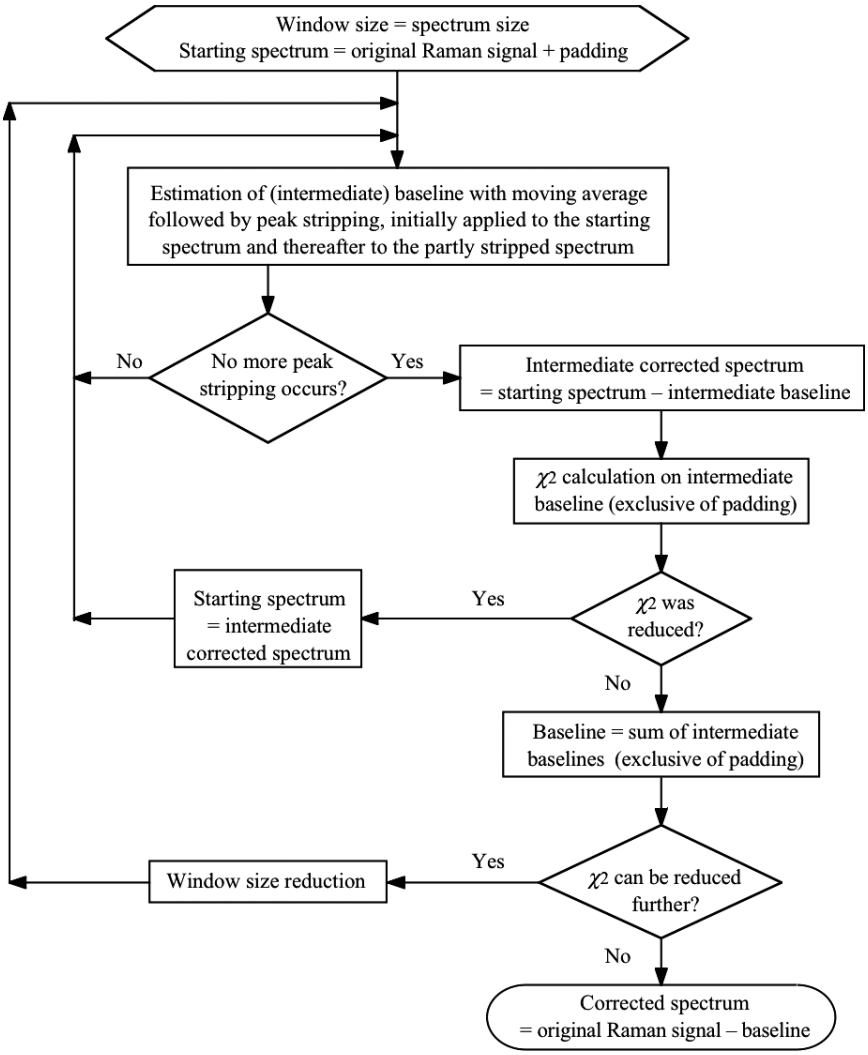

FIG. 2. A flow chart of the operations used to implement the automated baseline-correction method.

1b), strong baseline curvatures and overlapping peaks make the estimation of the baseline problematic. With high SBR (e.g., Fig. 1c), the relatively low SNR complicates baseline location. We therefore consider the simulated data non-trivial from the perspective of performing baseline correction on them.

Figures 3, 4, and 5 show results for flattening the exponential, Gaussian, and sigmoidal baselines, respectively. Statistical results and parameter values are reported in Table I. Spectra recovered from baselines with SBRs of 1, 0.1, and 0.01 are shown in panels a, c, and e, respectively (gray traces, superimposed black traces are those of the set of blurred Lorentzians) while panels b, d, and f show the corresponding estimated (gray) and true (black) baselines. Remarkably, even with low SBRs, a spectrum can still be recovered, albeit at the cost of some distortion.

Smoothing the concluding baseline estimate (i.e., the remnant spectrum after peak stripping) has the advantage of reducing correlated noise in the final flattened baseline. Therefore, if the final corrected spectrum is also smoothed, some degree of denoising of the baseline is accomplished because both high- and low-frequency noise in the flat baseline is then reduced. For example, the corrected spectrum in Fig. 3a shows the flattened baseline with low-frequency correlated noise largely absent while the presence of such noise in a baseline would appear similar to the slight artifact near channel 150 in Fig. 3e.

Baselines with greater curvature are harder for the ABE procedure to deal with; thus, the Gaussian and sigmoidal baselines generally require smaller windows than the exponential baselines to be estimated effectively (e.g., Table I). Because the method starts with the largest possible window, relatively more computational time is needed to reach the required smaller window sizes. Furthermore, using smaller windows for baseline estimation results in the erosion of peaks in areas of peak congestion (e.g., channel 500 to channel 700 in Fig. 1c). Thus, more of the peak intensities are partitioned into the baseline, as is evident when comparing Figs. 4e and 5e to Fig. 3e where the differences between actual and estimated baselines are shown as black traces offset below the recovered spectra.

Sharp curvatures and large windows also generate the opposite problem; here, sharp curves where the baseline is underestimated are subtracted from the spectrum and thus become partitioned into the peak "space". Figure 6a shows examples of repeated baseline correction (more below) on the same spectrum originally with a SBR 0.01 exponential background (e.g., Fig 1a). Pronounced artifacts due to curvature are evident in some of these results centered near

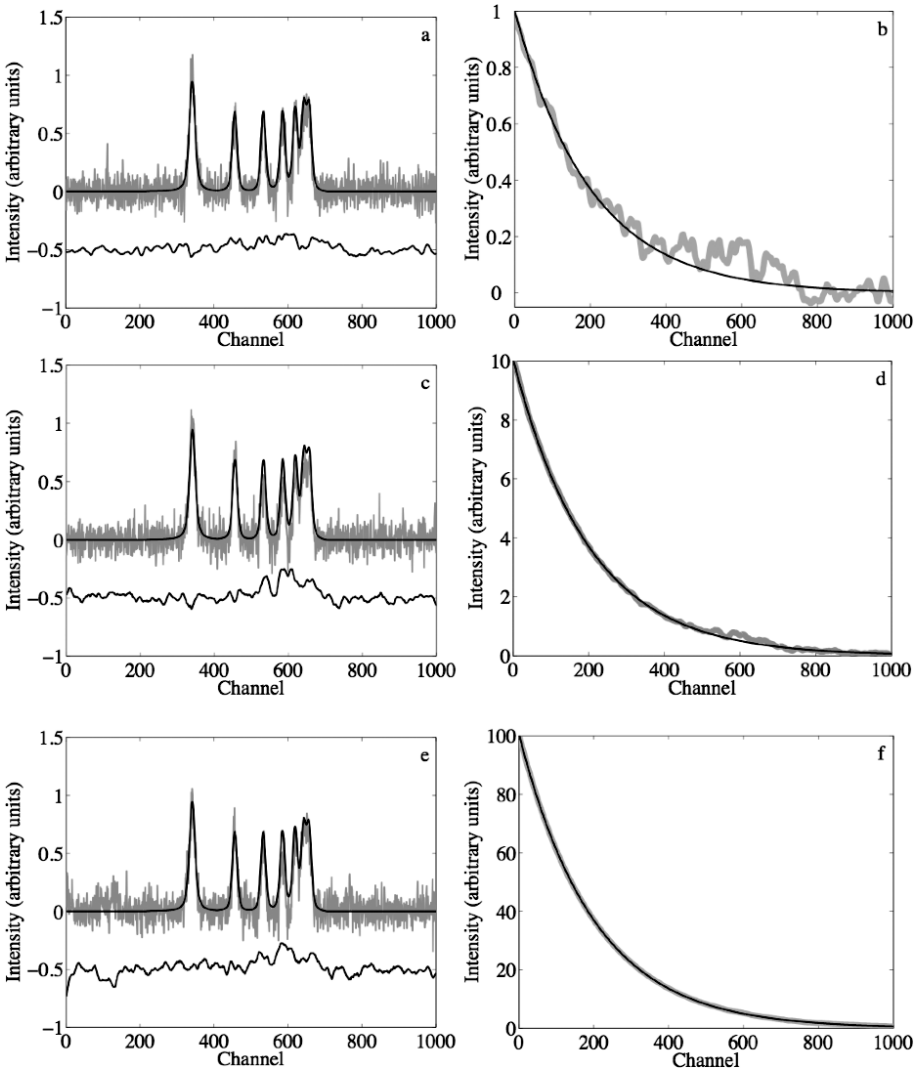

FIG. 3. Spectra (gray traces), after automated baseline correction, for exponential baselines with signal-to-baseline ratios of (**a**) 1, (**c**) 0.1, and (**e**) 0.01 with the "ideal" (noiseless and baseline-free) spectrum superimposed thereon (black). The estimated (gray) and true baselines (black) for the corresponding signal-to-baseline ratios are shown in panels (**b**), (**d**), and (**f**), and the differences between them are shown as black traces negatively offset for clarity in panels (**a**), (**c**), and (**e**).

channels 20 and 180. The greatest errors coincide with the use of the largest windows and with more baseline-correction iterations using these large windows. However, there is no correspondence between the terminal χ^2-values (i.e., terminal flatness) achieved and the presence or severity of artifacts.

The fact that repeated processing of the same spectrum does not yield identical outcomes is due to the peak stripping procedure where noise, estimated from the spectrum, is injected into the stripped spectrum at those positions formerly occupied by peaks. In the absence of this intervention, baseline correction on the same spectrum mentioned above (i.e., originally with a SBR 0.01 exponential background), produces no artifacts and identical results when repeated ten times. Values, corresponding to the first entry in Table I, are 112 (window), 5 (baseline-testing iterations), 5.58 (χ^2_{terminal}), and 3.89 (rms). Where the same approach (i.e., no noise injection during peak stripping) is taken on ten identical spectra but each with independent noise, these values are 106.50 ± 5.80 (window), 5.10 ± 0.32 (baseline-testing iterations), 6.12 ± 0.84 (χ^2_{terminal}), and 3.94 ± 0.17 (rms), giving means ± standard deviations. Likewise, there were no artifacts.

The absence of artifacts, however, comes at the cost of increased peak erosion. Thus, where artifacts do not overlap with peaks, an interesting option arises if one were able to discriminate against or ameliorate the effects of artifacts. One possibility is to do a number of repeated corrections on the same spectrum (e.g., ten times, Fig. 6a) and then to average these (Fig. 6b) before baseline correction (Fig. 6c, light gray trace) and smoothing (Fig. 6c, dark gray trace) of the average spectrum. An alternative, based on previous work,[32] is to multiply all the spectra (corresponding channels). Since the baseline correction leads to a near-flat baseline with approximately zero mean, the baseline of the multiplied spectrum is much reduced compared to the peaks. Spectral intensity values less than a threshold (e.g., three times the standard deviation of the noise in the original spectrum, raised to the power of the number n of spectra multiplied together) are set to zero (Fig. 6d, light gray trace) and the nth root (n here is 10) of the resulting spectrum (Fig. 6d, dark gray trace) is obtained.

Real Spectra. Real Raman spectra obtained from solid triacontanol and triacontanoic acid, plant growth promot-

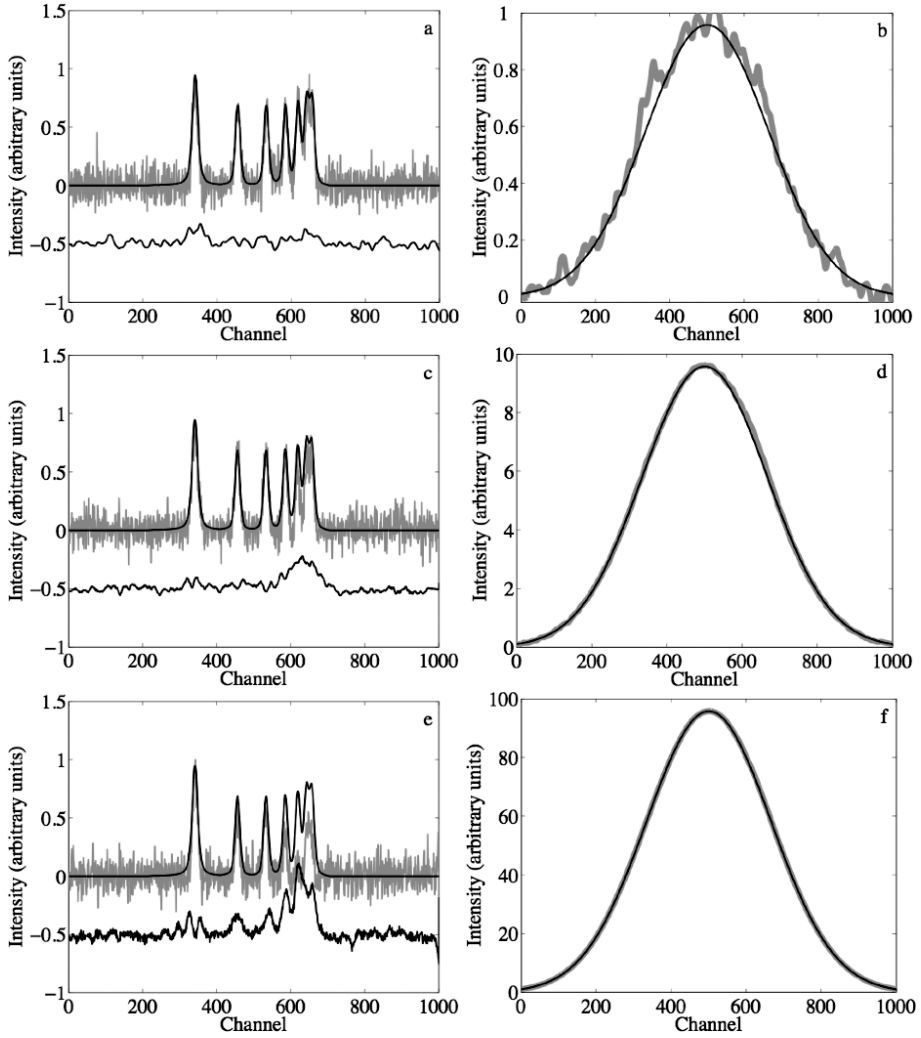

FIG. 4. Spectra (gray traces), after automated baseline correction, for Gaussian baselines with signal-to-baseline ratios of (**a**) 1, (**c**) 0.1, and (**e**) 0.01 with the "ideal" (noiseless and baseline-free) spectrum superimposed thereon (black). The estimated (gray) and true baselines (black) for the corresponding signal-to-baseline ratios are shown in panels (**b**), (**d**), and (**f**), and the differences between them are shown as black traces negatively offset for clarity in panels (**a**), (**c**), and (**e**).

ers,[33,34] and from the skin of the tomato fruit have moderate to pronounced sloping baselines as shown by the gray traces (scaled and offset for ease of viewing) in Fig. 7. The black traces in Fig. 7 show the results after correction. Application of the ABE method to the tomato-skin spectrum produced the result shown in Fig. 7a where a maximum window size of 1651 wavenumbers and 12 baseline-testing iterations were used. This result was not completely satisfactory. However, when started with an initial window half the size (i.e., $n = 2$ in Eq. 4), convergence occurred to a window with a size of 826 wavenumbers (i.e., $n = 2$ in Eq. 4) and nine baseline-testing iterations at this window size were used. The result is shown in Fig. 7b. A similar problem arose for the other two spectra. The triacontanol spectrum, after baseline correction, is shown in Fig. 7c, where a maximum window size of 1651 wavenumbers and 11 baseline-testing iterations were used. When started with an initial window half the size (i.e., $n = 2$ in Eq. 4), convergence occurred to a window size of 332 wavenumbers (i.e., $n = 5$ in Eq. 4) and four baseline-testing iterations at this window size were used (Fig. 7d). Fully automated baseline correction of the triacontanoic acid produced the spectrum shown in Fig. 7e, where convergence to the maximum window size of 1651 wavenumbers occurred and nine baseline-testing iterations were needed. Likewise, when started with an initial window half the size, convergence occurred to a window size of 332 wavenumbers (i.e., $n = 5$ in Eq. 4) and three baseline-testing iterations at this window size were used and the spectrum in Fig. 7f was obtained.

We also observed some artifacts as a consequence of ABE, especially at spectral edges (e.g., right edge, Fig. 7b). Thus, attempts to pad the spectra by extrapolating beyond their edges were not entirely successful and require additional attention. Furthermore, the trade-off with "denoising" of correlated noise is the loss of small peaks from corrected spectra (cf. Fig. 7c to Fig. 7d near 1400 cm^{-1}). In addition to these problems, we noted that repeated processing resulted in nearly identical results for both triacontanol and triacontanoic acid; thus, the methods proposed above (i.e., averaging, multiplication) for discriminating against artifacts induced by ABE could not be applied to them. Nevertheless, despite these difficulties, good performance was obtained on real Raman spectra where peaks of varying width as well as different baseline features, such as exponential- and linear-like features, were present within the same spectrum.

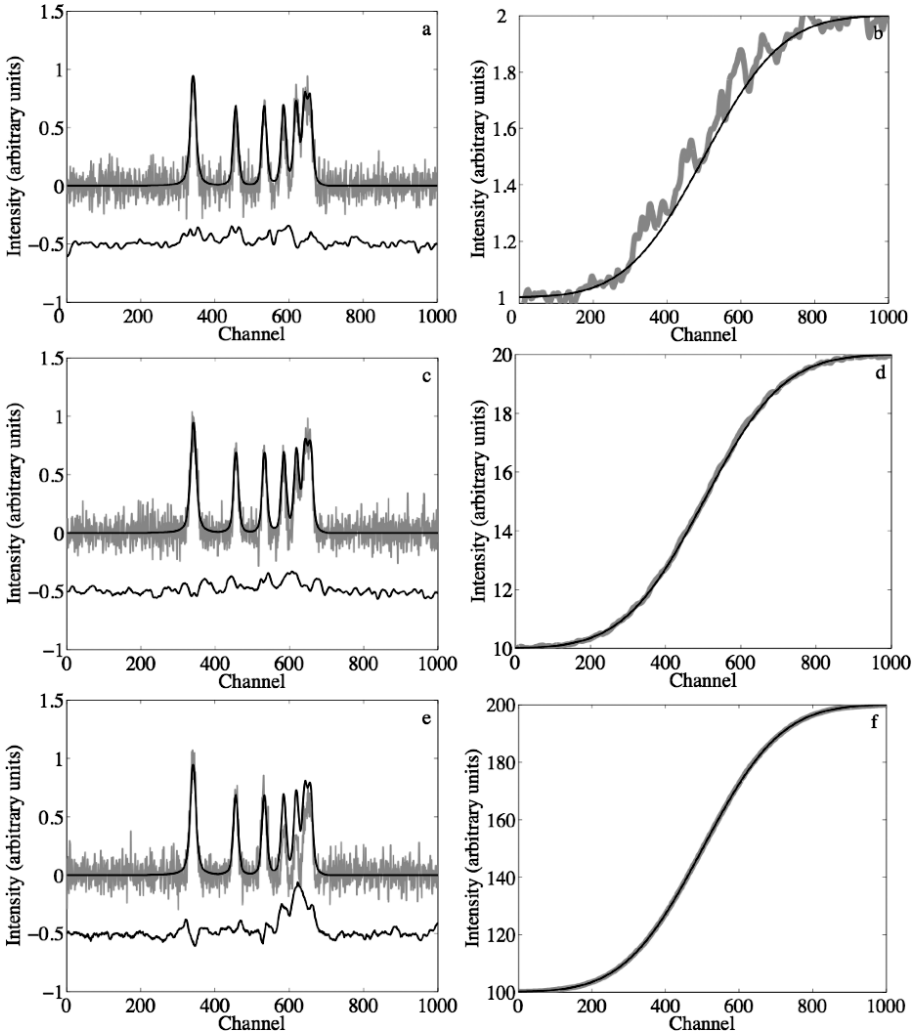

FIG. 5. Spectra (gray traces), after automated baseline correction, for sigmoidal baselines with signal-to-baseline ratios of (**a**) 1, (**c**) 0.1, and (**e**) 0.01 with the "ideal" (noiseless and baseline-free) spectrum superimposed thereon (black). The estimated (gray) and true baselines (black) for the corresponding signal-to-baseline ratios are shown in panels (**b**), (**d**), and (**f**), and the differences between them are shown as black traces negatively offset for clarity in panels (**a**), (**c**), and (**e**).

Stopping Criteria and Parameter Settings. Our aim is to devise a model-free baseline-removal procedure that can track arbitrary baselines adequately as well as one that is "parameter-free" and can be implemented without recourse to user intervention. In practice, however, whenever the choice of a parameter is circumvented, conditional statements seem to emerge in compensation. Thus the question is: what tradeoff—between parameter selection, parameter space search, and the invocation of thresholds—can be implemented that is effective and reasonable?

On theoretical grounds, as explained above, the search of parameter space starts at the one extreme of employing the

TABLE I. Comparative results (mean ± standard deviation) for the automated baseline-correction method on three sets of simulated Raman spectra with very high, moderately high, and low baselines as indicated by the signal-to-baseline ratios (SBR). Every spectrum in each set of 10 spectra had a unique distribution of Gaussian noise giving a signal-to-noise ratio (SNR) of 10. Upon automated termination, the termination parameters (zero-order Savitzky–Golay smoothing window size, number of successive baseline-testing iterations needed) for each spectrum were recorded. The termination χ^2-value and the root mean square (rms) error for every estimated baseline were also recorded.

Baseline	SBR (approximate)	Window, mean ± sd	Iterations, mean ± sd	$\chi^2_{terminal}$, mean ± sd	rms, mean ± sd
Exponential	0.01	254 ± 64	9 ± 2	13.62 ± 8.95	4.84 ± 2.34
Exponential	0.1	451 ± 209	7 ± 1	8.63 ± 5.23	3.28 ± 0.87
Exponential	1	801 ± 258	7 ± 1	2.93 ± 1.19	1.41 ± 0.15
Gaussian	0.01	159 ± 53	9 ± 2	19.61 ± 12.68	4.29 ± 0.51
Gaussian	0.1	449 ± 383	7 ± 2	79.86 ± 112.01	24.27 ± 33.90
Gaussian	1	688 ± 341	6 ± 2	12.14 ± 6.05	4.29 ± 2.93
Sigmoidal	0.01	434 ± 391	9 ± 3	277.53 ± 436.93	29.52 ± 44.30
Sigmoidal	0.1	214 ± 34	7 ± 1	6.43 ± 3.48	2.46 ± 0.91
Sigmoidal	1	518 ± 183	6 ± 2	3.58 ± 1.60	1.87 ± 0.59

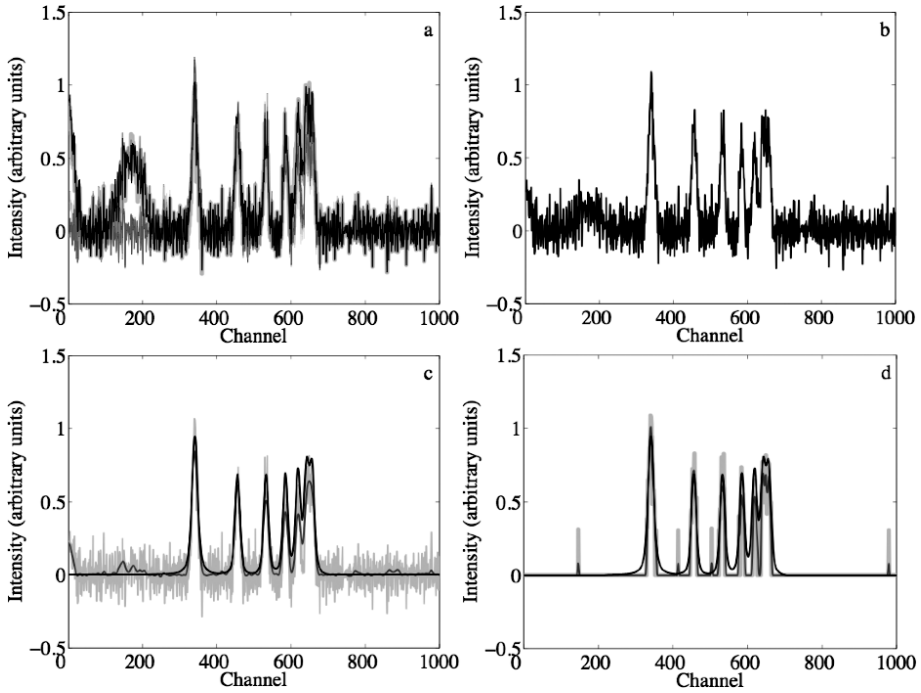

FIG. 6. A spectrum with exponential baseline and a signal-to-baseline ratio of 0.01 repeatedly (10 times) subjected to baseline correction: (**a**, 10 grayscale traces) shows that automated baseline estimation with noise injection at the position of stripped peaks can produce artifacts (e.g., near channels 20 and 200) but often with reduced peak-base erosion. When these spectra are averaged, the average (**b**) can be baseline-corrected to reduce the presence of artifacts and erosion in peak bases (**c**, light gray trace) and optionally smoothed with an automated procedure (**c**, dark gray trace). An alternative is to multiply the 10 spectra, set all baseline values below a threshold to zero, and get the 10th root of the product spectrum (**d**, light gray trace) with optional subsequent smoothing bases (**d**, dark gray trace). In both (**c**) and (**d**) the "ideal" (noiseless and baseline-free) spectrum is superimposed in black for comparison.

largest possible window and using it in a zero-order Savitzky–Golay filter to estimate the baseline repeatedly. If the filter window is too large, the success in estimating the baseline is limited and repeated applications can begin to generate distortions in the baseline, causing it to become progressively less flat. The trend is captured by testing with the χ^2-statistic (latest baseline estimate compared to a flat line using the noise level of the original spectrum) and when it starts to increase, it suggests that no further improvements could be obtained (Fig. 8, bottom rows). This establishes the optimal number of successive baseline-removal iterations and the minimum χ^2-value for a window of given size. Furthermore, if the minimum χ^2-statistic obtained is in excess of the number of channels, the result is not deemed acceptable.[29]

Thus, the window size is reduced and the search continues with a smaller window that is better able to fit the baseline. Even though, at some point, the χ^2-statistic falls below the threshold for acceptability, further improvements can be obtained and should be sought. This is because, if baseline correction is completely successful, an estimated baseline on a corrected spectrum should be completely flat (the success condition). In that case, the χ^2-statistic would be zero. Therefore, the lower the χ^2-statistic, the better the result (Fig. 8, middle rows).

With smaller windows, however, some amount of peak erosion begins to occur and continued baseline removal appears to aggravate this. Consequently, the χ^2-statistic again starts to increase. With even smaller windows, peak erosion is more aggressive and less-small χ^2-values are obtained (Fig. 8, top rows). Taken together then, the objective is to find a local χ^2-value minimum with the largest possible window size. We mention here a local minimum because the global minimum is likely to occur with the smallest window size, i.e., where both peaks and baseline are removed. The results shown in Fig. 8 were generated when correcting the simulated spectrum with exponential SBR 0.01 baseline.

Conceptually, our ABE procedure is based on a parameter space that is progressively simplified: first, by selecting a Savitzky–Golay filter to estimate the baseline since it has only two parameters (order and window size) and can model any arbitrary baseline; second, by selecting a zero-order filter because it has the best low-pass characteristics; third, by initiating a search for the optimal value of the remaining parameter, starting at its large extreme, and iterating to an expected nearby local χ^2-value minimum. If the local χ^2-value minimum is also acceptable from a statistical point of view, the solution is in hand. Note that one could claim that repeated baseline corrections simply introduce a new parameter (number of repeats or iterations needed) to replace filter order, now fixed at a constant value, as a parameter; however, even when retaining the filter order as a full parameter, repeated baseline corrections would have utility.

Even if this method is not used for automated baseline correction, it provides other benefits. For example, baselines estimated with the method presented here can be used for modeling purposes (e.g., with the commonly used polynomial fitting approaches).[9,11,21,23] Additionally, the same peak-stripping and χ^2-testing approach can be used, but based on polynomial modeling instead of the zero-order Savitzky–Golay filter. Even though a polynomial of a given order is then used, an essentially model-free method results because the baseline is repeatedly estimated and corrected. The same approach is likely adaptable to other parameter-sparse methods, such as the median filter method,[35] as well. Finally,

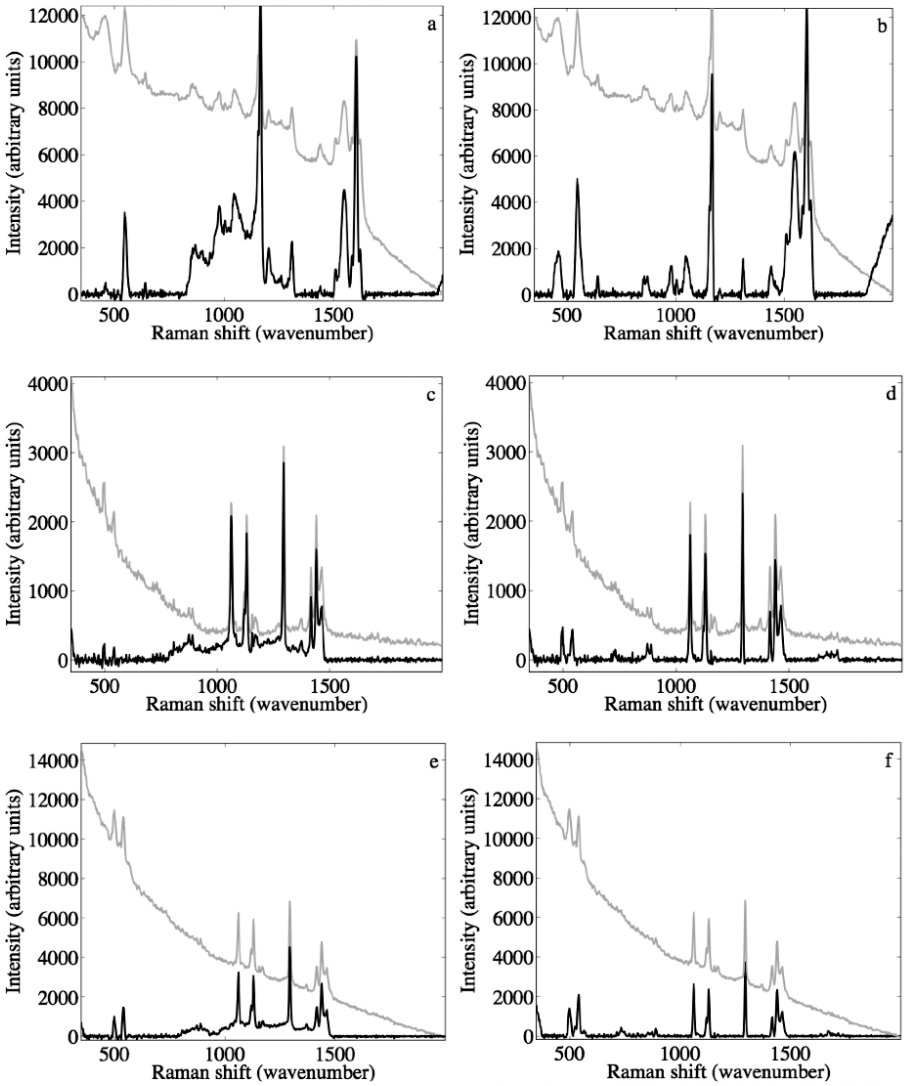

FIG. 7. Real Raman spectra from (**a**, **b**) tomato skin, (**c**, **d**) triacontanol, and (**e**, **f**) triacontanoic acid shown as gray traces and as black traces after baseline correction. Fully automated baseline correction of the tomato skin spectrum starting with the maximum window size leads to incomplete baseline removal (**a**), but a corrected spectrum can be obtained by starting with a smaller (i.e., half the maximum) initial window (**b**). As in (**a**) and (**b**), respectively, but for triacontanol (**c**), (**d**). As in (**a**) and (**b**), respectively, but for triacontanoic acid (**e**), (**f**).

one could use the results (window size, iterations) of the ABE method to guide the subsequent selection of baseline-estimation filter parameters for a variety of other methods.

CONCLUSIONS

High throughput measurement methods are creating vast quantities of data and consequently generate a pressing need for automated methods of analysis. These include the correction of sloping baselines in spectra. We have presented here a fully automated and model-free baseline-estimation procedure capable of baseline correction without, or at most with minimal, user intervention. This ABE method is based on a zero-order Savitzky–Golay filter and the quality of the corrected baseline is assessed with the χ^2-statistic, which thereby serves as a computational stopping criterion. Because we start with the largest possible window, and progressively decrease it, the method is inherently robust with regard to wide and/or overlapping peaks.

Although the proposed ABE method is not restricted to applications involving Raman spectroscopy, simulated and real Raman spectra were used here for evaluation purposes. Our results show that the ABE method consistently produces baseline-flattened spectra of high quality without user intervention. However, the method is not entirely error free. Where baselines are very prominent with strong curvatures, correction requires the use of smaller-than-ideal filter windows, causing peak erosion. In other cases, an undesirable local minimum occurs at the maximum window size, leading to incomplete baseline removal. The latter problem can be overcome by starting the iteration process with a smaller initial window (i.e., half the maximum), as discussed above. Ideally though, we wish to avoid manually setting such "arbitrary" thresholds. Minor problems such as edge artifacts also occur. Nevertheless, we believe that the major benefits of none or minimal user intervention, the ability to handle large data sets with great consistency, and the high quality of baseline-corrected spectra will be of great interest to many spectroscopists with substantial baseline-correction needs.

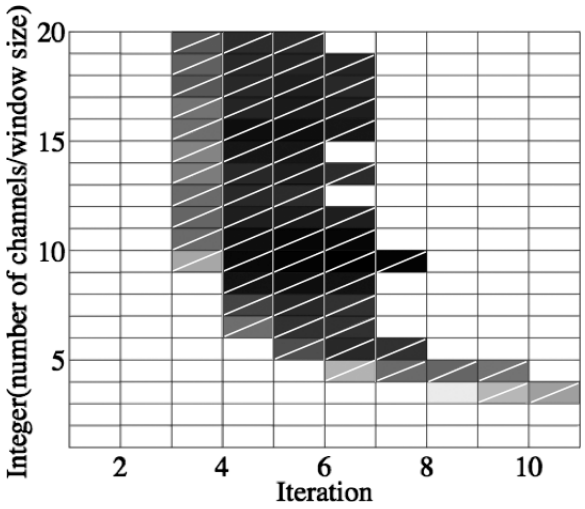

FIG. 8. The χ^2-value "error" surface of baseline correction (no noise injection at the position of stripped peaks) on a spectrum with exponential baseline and a signal-to-baseline ratio of 0.01. Lower χ^2-values are darker; χ^2-values above 100 were truncated to 100 and these are white. The surface indicates that large windows (e.g., number of channels/window size = 1, 2) are not effective in baseline estimation, that smaller windows are more effective, but require more iterations (e.g., number of channels/window size = 3, 4, 5), and that much smaller windows produce aggressive peak erosion with less low χ^2-values (e.g., number of channels/window size = 16+). Thus, the objective is to find a local minimum with the largest window size (e.g., number of channels/window size ~= 8, 9, 10). The automated procedure reported here accomplishes this by starting at the large extreme of the window size parameter and searches until the local χ^2-minimum is reached, i.e., when χ^2-values increase with increasing window size and increasing repeated estimations.

ACKNOWLEDGMENTS

We wish to thank Dr. R. Jetter at the University of British Columbia for tomato, triacontanol, and triacontanoic acid samples. We also thank the Natural Sciences and Engineering Research Council of Canada for financial assistance and the Canadian Foundation for Innovation and the British Columbia Knowledge Development Foundation for funding the resources made available for this work through the UBC Laboratory for Molecular Biophysics. In addition, we thank the Canadian Microelectronics Corporation for providing computer resources through the System-on-Chip Laboratory (Department of Electrical and Computer Engineering).

1. L. Griffiths, Magn. Reson. Chem. **44,** 54 (2006).
2. M. K. Kiymik, I. Güler, A. Dizibüyüka, and M. Akin, Comput. Biol. Med. **35,** 603 (2005).
3. T. Krishnamurthy, U. Rajamani, P. L. Ross, R. Jabbour, H. Nair, J. Eng, J. Yates, M. T. Davis, D. C. Stahl, and T. D. Lee, Toxin Rev. **19,** 95 (2000).
4. J. R. Scott, T. R. McJunkin, and P. L. Tremblay, J. Assoc. Lab. Auto. **8,** 61 (2003).
5. T. Sundin, L. Vanhamme, P. Van Hecke, I. Dologlou, and S. Van Huffel, J. Magn. Reson. **139,** 189 (1999).
6. L. Vanhamme, T. Sundin, P. Van Hecke, S. Van Huffel, and R. Pintelon, J. Magn. Reson. **143,** 1 (2000)
7. H. Waki, K. Katahira, J. W. Polson, S. Kasparov, D. Murphy, and J. F. R. Paton, Exp. Physiol. **91,** 201 (2006)
8. B. Yan, T. R. McJunkin, D. L. Stoner, and J. R. Scott, Appl. Surf. Sci. **253,** 2011 (2006).
9. J. Zhao, H. Lui, D. I. McLean, and H. Zeng, Appl. Spectrosc. **61,** 1225 (2007).
10. P. A. Mosier-Boss, S. H. Lieberman, and R. Newbery, Appl. Spectrosc. **49,** 630 (1995).
11. G. Schulze, A. Jirasek, M. M. L. Yu, A. Lim, R. F. B. Turner, and M. W. Blades, Appl. Spectrosc. **59,** 545 (2005).
12. B. Czarnik-Matusewicz, M. A. Czarnecki, and Y. Ozaki, in *Two-Dimensional Correlation Spectroscopy*, Y. Ozaki and I. Noda, Eds. (American Institute of Physics, New York, 2000), pp. 291–294.
13. P. J. Tandler, P. d. B. Harrington, and H. Richardson, Anal. Chim. Acta **368,** 45 (1998).
14. D. Chang, C. D. Banack, and S. L. Shah, J. Magn. Reson. **187,** 288 (2007).
15. J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi, Bioinformatics **21,** 1764 (2005).
16. Y. V. Karpievitch, E. G. Hill, A. J. Smolka, J. S. Morris, K. R. Coombes, K. A. Baggerly, and J. S. Almeida, Bioinformatics **23,** 264 (2007).
17. M. M. L. Yu, S. O. Konorov, H. G. Schulze, M. W. Blades, R. F. B. Turner, and R. Jetter, Planta **227,** 823 (2008).
18. M. M. L. Yu, H. G. Schulze, R. Jetter, M. W. Blades, and R. F. B. Turner, Appl. Spectrosc. **61,** 32 (2007).
19. A. Jirasek, G. Schulze, M. M. L. Yu, M. W. Blades, and R. F. B. Turner, Appl. Spectrosc. **58,** 1488 (2004).
20. C. Camerlingo, F. Zenone, G. M. Gaeta, R. Riccio, and M. Lepore, Meas. Sci. Technol. **17,** 298 (2006).
21. A. Cao, A. K. Pandya, G. K. Serhatkulu, R. E. Weber, H. Dai, J. S. Thakur, V. M. Naik, R. Naik, G. W. Auner, R. Rabah, and D. C. Freeman, J. Raman Spectrosc. **38,** 1199 (2007).
22. J. C. Cobas, M. A. Bernstein, M. Martín-Pastor, and P. G. Tahoces, J. Magn. Reson. **183,** 145 (2006).
23. T. Lan, Y. Fang, W. Xiong, and C. Kong, Chin. Opt. Lett. **2007,** 613 (2007).
24. C. Rowlands and S. Elliott, J. Raman Spectrosc. (2010), DOI 10.1002/jrs.2691.
25. H. G. Schulze, R. B. Foist, A. Ivanov, and R. F. B. Turner, Appl. Spectrosc. **62,** 1160 (2008).
26. S. E. Bialkowski, Anal. Chem. **60,** 355A (1988).
27. L. S. Greek, H. G. Schulze, M. W. Blades, A. V. Bree, B. B. Gorzalka, and R. F. B. Turner, Appl. Spectrosc. **49,** 425 (1995).
28. H. G. Schulze, R. B. Foist, A. I. Jirasek, A. Ivanov, and R. F. B. Turner, Appl. Spectrosc. **61,** 157 (2007).
29. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C* (Cambridge University Press, New York, 1994), 2nd ed.
30. P. D. Wentzell and C. D. Brown, in *Encyclopedia of Analytical Chemistry*, R. A. Meyers, Ed. (Wiley, Chichester, UK, 2000), pp. 9764–9800.
31. G. Schulze, M. M. L. Yu, C. J. Addison, M. W. Blades, and R. F. B. Turner, Appl. Spectrosc. **60,** 820 (2006)
32. H. G. Schulze, L. S. Greek, C. J. Barbosa, M. W. Blades, and R. F. B. Turner, Appl. Spectrosc. **52,** 621 (1998).
33. A. B. Eriksen, G. Selldén, D. Skogen, and S. Nilsen, Planta **152,** 44 (1981).
34. S. K. Ries, V. Wert, C. C. Sweeley, and R. A. Leavitt, Science (Washington, D.C.) **195,** 1339 (1977).
35. M. S. Friedrichs, J. Biomol. NMR **5,** 147 (1995).