Automatic Background Removal and Correction of Systematic Error Caused by Noise Expecting Bio-Raman Big Data Analysis

Akunna Frances UJUAGU, Ziteng WANG, and Shin-ichi MORITA†

Graduate School of Science, Tohoku University, 6-3 Aramaki-Aza-Aoba, Aoba, Sendai 980-8578, Japan

Keywords background removal, subtraction, systematic error, noise, correction, Raman, bio-Raman, analysis, big data

† To whom correspondence should be addressed.
E-mail: shinichi.morita.c6@tohoku.ac.jp

Abstract
Spectral pre-treatments such as background removal from Raman big data are crucial to have a smooth link to advanced spectral analysis. Recently, we developed an automated background removal method, where we considered the shortest length of a spectrum by changing the scaling factor of the background spectrum. Here, we proposed a practical way to correct the systematic error caused by noise from measurement. This correction realized more effective and accurate automatic background removal.
Research on single live cells and biological tissues using a Raman microscope (bio-Raman research) has attracted attention, forming an active and competing research field, mainly because of its non-invasive, non-labeling, and less water influence features as well as simultaneous detection of many chemical species. Bio-Raman data is very informative but large and complicated. To disentangle the complicated big data, advanced spectral analysis such as principal component analysis and non-negative matrix factorization has been applied. Two-dimensional correlation analysis is also one of the candidates for bio-Raman spectral analysis.

Spectral pre-treatments such as background removal, “trend” correction, and normalization of Raman big data are crucial to have a smooth and effective link to the advanced spectral analysis. The spectral pre-treatments have to be automated because of the large number of bio-Raman data that frequently reaches more than a thousand. Here we recognize that, in this paper, the word “background” refers to a practically obtained spectrum (e.g., a spectrum component of glass or quartz solid substrate that supports a live adherent cell), while the word “trend” is used otherwise (suppose, e.g., a globally lying spectrum component of auto fluorescence that a live cell gives off when the pure spectrum of the auto fluorescence is not measured).

Under this situation, we developed an automated background removal method, in which we considered the shortest length of a spectrum by changing the scaling factor of the background spectrum. The proposed method was effective and versatile for big data analysis. For instance, background components of quartz substrate were automatically removed from bio-Raman spectra of single live cells; binary background components of glass and glue were removed automatically and simultaneously from imaging Raman spectra of gray hair fixed on glass using glue. But there existed a few weak points. One of the problems was systematic error of scaling factor caused by noises of measured spectra.

In this paper, (i) we describe how the method generates systematic error originating from noise of spectral measurements when we estimate a scaling factor; (ii) we propose a practical way to correct the systematic error; (iii) simulation study was added expecting practical application. This method serves effectively not only for background removal, but also quantitative estimation of a target molecular species. The way of correcting systematic error caused by noise will realize a more accurate quantitative spectral analysis as well as more effective automatic background removal.

Let us first briefly describe to the conventional theoretical framework. We suppose that
multiple Raman spectra \( p(w_j) \) and a single background spectrum \( q(w_j) \) are experimentally measured, where \( w_j \) is Raman shift and \( j \) is a counting integer, that is, \( j = 1, 2, 3 \ldots n \). If the intervals of \( w_j \) are not evenly spaced, \( p(w_j) \) and \( q(w_j) \) are transformed to be evenly spaced by using an interpolation technique. Using scaling factors \( k \), we define the multiple subtracted spectra \( r(w_j) \) as:

\[
 r(w_j) = p(w_j) - k q(w_j). \tag{1}
\]

We optimize the scaling factors \( k \) so that multiple \( r(w_j) \) become independent of \( q(w_j) \). To do so, by considering the shortest lengths of spectra \( r(w_j) \), the scaling factors \( k \) can be automatically and approximately determined by the following simple relation,\(^{11}\) that is,

\[
 k = \frac{\sum_{j=1}^{n} p'(w_j) q'(w_j)}{\sum_{j=1}^{n} q'(w_j) q'(w_j)}. \tag{2}
\]

The symbol prime ( ‘ \() \) indicates the first derivative of a spectrum. Savitzky Golay method is used to calculate the first derivative spectra and to yield the smoothing.\(^{13}\) For the Savitzky Golay method, including the first derivative and the smoothing, “the degree of a polynomial”, \( d \), and “the number of data points”, \( p \), should be determined to give a sufficient approximation. This Eq. (2) serves properly when the magnitude of noise from measurement is negligible.

Here we incorporate the concept of noise to the conventional theoretical framework. When noise is considered, \( p(w_j) \) and \( q(w_j) \) are denoted \( p_{NP}(w_j) \) and \( q_{NP}(w_j) \). The spectra, \( p_{NP}(w_j) \) and \( q_{NP}(w_j) \), can be decomposed into signal and noise conceptually, that is,

\[
p_{NP}(w_j) = p_{TRUE}(w_j) + n_{pTRUE}(w_j)
\]
\[
q_{NP}(w_j) = q_{TRUE}(w_j) + n_{qTRUE}(w_j), \tag{3}
\]

where \( p_{TRUE}(w_j) \) and \( q_{TRUE}(w_j) \) correspond to signal, while \( n_{pTRUE}(w_j) \) and \( n_{qTRUE}(w_j) \) correspond to noise. For noise, Gaussian distribution is assumed. By substituting Eq. (3) into Eq. (2), we update the scaling factors \( k \) to \( k_N \) by considering noise, that is, we obtain Eq. (4) and (5):

\[
k_N = \frac{\sum_{j=1}^{n} p'_{NP}(w_j) q'_{NP}(w_j)}{\sum_{j=1}^{n} q'_{NP}(w_j) q'_{NP}(w_j)}. \tag{4}
\]

\[
k_N =
\]

\[

\]
\[
\sum_{j=1}^{n} \left\{ p'_{\text{TRUE}}(w_j) q'_{\text{TRUE}}(w_j) + p'_{\text{TRUE}}(w_j) n'_{q\text{TRUE}}(w_j) + n'_{q\text{TRUE}}(w_j) q'_{\text{TRUE}}(w_j) + n'_{q\text{TRUE}}(w_j) n'_{q\text{TRUE}}(w_j) \right\}
\]
\[
\sum_{j=1}^{n} \left\{ q'_{\text{TRUE}}(w_j)^2 + 2 q'_{\text{TRUE}}(w_j) n'_{q\text{TRUE}}(w_j) + n'_{q\text{TRUE}}(w_j)^2 \right\}
\]

(5)

Noises lying on a spectrum (the first derivatives of noises return noises) statistically give positive or negative values equally. The summations of noises \( n'_{p\text{TRUE}}(w_j) \) or \( n'_{q\text{TRUE}}(w_j) \) are therefore statistically reduced to zero. Utilizing this feature, many of the terms related to noise in Eq. (5) can be approximated to zero, that is, \( \Sigma n'_{p\text{TRUE}} \) \( n'_{q\text{TRUE}} \) statistically goes to zero because the term is based on multiplications of different noises to result noises; \( \Sigma p'_{\text{TRUE}} n'_{q\text{TRUE}} \), \( \Sigma n'_{p\text{TRUE}} q'_{\text{TRUE}} \), and \( \Sigma q'_{\text{TRUE}} n'_{q\text{TRUE}} \) statistically goes to zero because the terms are based on signals multiplied with noises to yield weighted noises. The exception is the term of \( \Sigma (n'_{q\text{TRUE}}^2) \) in the denominator. \( \Sigma (n'_{q\text{TRUE}}^2) \) yields a certain positive value. This is because \( n'_{q\text{TRUE}}(w_j) \) is multiplied by itself always giving positive values. Eventually, \( k_N \) is simplified to

\[
k_N \approx \frac{\sum_{j=1}^{n} \left\{ p'_{\text{TRUE}}(w_j) q'_{\text{TRUE}}(w_j) \right\}}{\sum_{j=1}^{n} \left\{ q'_{\text{TRUE}}(w_j)^2 + n'_{q\text{TRUE}}(w_j)^2 \right\}}. \tag{6}
\]

By this relation, we clearly find how the scaling factors \( k_N \) are deviated from the true values. Only the background spectrum, \( q_{Nq}(w_j) \), causes systematic error in \( k_N \) and the \( k_N \) giving smaller values systematically.

Correction of the systematic error caused in \( k_N \) is possible. The corrected scaling factor \( k_{Ncorr} \) is obtained by modifying the denominator part of Eq. (4), given by

\[
k_{Ncorr} = \frac{\sum_{j=1}^{n} p'_{NP}(w_j) q'_{Nq}(w_j)}{\sum_{j=1}^{n} \left\{ q'_{Nq}(w_j)^2 \right\} - \sum_{j=1}^{n} \left\{ n'_{q\text{TRUE}}(w_j)^2 \right\}}. \tag{7}
\]

\( n'_{q\text{TRUE}}(w_j) \) is the first derivative of \( n_{q\text{TRUE}}(w_j) \) defined in Eq. (3). Of note is that \( n_{q\text{TRUE}}(w_j) \) is solely of conceptual existence. The true values are impossible to obtain directly by practical measurements.

Our strategy is to statistically estimate the expected value of \( \Sigma (n'_{q\text{TRUE}}^2) \) in Eq. (7). For this, we use Savitzky Golay smoothing filter\(^1\) on the experimentally measured \( q_{Nq}(w_j) \) to yield a signal rich and noise reduced component, \( q_{SGS}(w_j) \). Using the practically obtained signal \( q_{SGS}(w_j) \), noise \( n_{qSGS}(w_j) \) is practically estimated by
\[ n_{qSGS}(w_j) = q_{Nq}(w_j) - q_{SGS}(w_j). \]  

(8)

Please note that the Savitzky Golay smoothing method does not return the signal component perfectly and some of noise component is mixed into the signal component of \( q_{SGS}(w_j) \). Thus, \( \Sigma (n'_{qSGS})^2 \) becomes smaller than \( \Sigma (n'_{TRUE})^2 \). To calibrate this, using the factor \( f \) and \( \Sigma (n'_{qSGS})^2 \), Eq. (7) is expressed by

\[
k_{Ncorr} = \frac{\sum_{j=1}^{n} p_{Np}(w_j) q_{Nq}(w_j)}{\sum_{j=1}^{n} (q_{Nq}(w_j))^2 - f \sum_{j=1}^{n} (n'_{qSGS}(w_j))^2}.
\]  

(9)

The factor \( f \) can be estimated by developing a calibration line between multiple \( \Sigma (n'_{TRUE})^2 \) values versus multiple \( \Sigma (n'_{qSGS})^2 \) values by changing magnitude of noise. To develop the calibration line, we use simulation spectra imitating the background spectrum of \( q_{Nq}(w_j) \), that is,

\[ q_{NqIMIT}(w_j) = q_{IMIT}(w_j) + n_{qTRUE}(w_j). \]  

(10)

\( q_{IMIT}(w_j) \) is a single spectrum that imitates the signal component of the experimentally obtained background spectrum of \( q_{Nq}(w_j) \). Several Raman peaks are assumed using Gaussian function (one can also use Lorentzian function or Voigt function depending on an experimental situation. We consider that the use of Gaussian function gives enough approximation to statistically estimate a \( \Sigma (n'_{TRUE})^2 \) value in a typical bio-Raman research). The noise can be considered \( n_{qTRUE}(w_j) \) based on Eq. (3). Generating \( n_{qTRUE}(w_j) \) repeatedly by changing the magnitude, we obtain multiple spectra \( q_{NqIMIT}(w_j) \). For calibration, we assume the relation of \( \Sigma (n'_{TRUE})^2 = f \Sigma (n'_{qSGS})^2 \), and using the least squares we statistically estimate the factor \( f \). It seems no other way to determine the factor \( f \), without preparing the \( n_{qTRUE}(w_j) \) by Eq. (10).

For practical use, we demonstrated a procedure to correct systematic error caused by noise using spectra simulated practical observations in bio-Raman research (here, the word simulated was used solely for expecting practical use; the meaning is conceptually different from the simulation spectra used in the calibration). In Fig. 1(a), 11 simulated Raman spectra \( p_{Np}(w_j) \) were prepared. In a single spectrum, 15 Raman peaks and a trend component due to auto-fluorescence were assumed. Please note that the peak positions and the heights in a simulated Raman spectrum are arbitrarily determined; thus, there is no physical / chemical meaning on them. The number of spectral variable was 667 and the data spacing for the horizontal axis, \( \Delta w \), was set \( \Delta w = 1.5 \). One of the spectra \( p_{Np} \) was arbitrarily selected, indicated by the bold blue line. Separately, a background spectrum \( q_{Nq}(w_j) \) was given, simulated by the

[...]

Analytical Sciences
Advance Publication by J-STAGE
Received April 2, 2020; Accepted April 9, 2020; Published online on April 17, 2020
DOI: 10.2116/analsci.20C005
bold red line. The scaling factors of the simulated background components of $q_{\text{TRUE}}(w_j)$ in observed spectra $p_{Np}(w_j)$ were always put $k_{\text{TRUE}} = 1.00$. In Fig. 1(b), the background removed spectra $r(w_j)$ were given using Eq. (1) using the conventional background removal method. For the Savitzky Golay method, the degree of a polynomial, $d$, and the number of data points, $p$, were chosen $d = 6$ and $p = 17$. We recognized visually that ca. 2% of the background component still remains slightly as indicated in the region of the asterisk. This is nothing but the systematic error of the scaling factor $k = k_N$. For the single spectrum $p_{Np}(w_j)$ indicated by blue, the scaling factor was estimated that $k = k_N = 0.98$. The value was smaller than the initial setting value of $k_{\text{TRUE}} = 1.00$.

We correct the systematic error using Eq. (9). Fig. 2(a) plots multiple $\sum (n'_{q_{\text{SGS}}})$ values versus multiple $\sum (n'_{q_{\text{TRUE}}})$ values by changing magnitude of noise. For the Savitzky Golay smoothing, the degree of a polynomial, $d$, and the number of data points, $p$, were chosen $d = 6$ and $p = 17$. The black straight line is the calibration line with the factor $f = 6.3$ that was determined by the least squares method based on the plots of $\sum (n'_{q_{\text{SGS}}}) = f \sum (n'_{q_{\text{TRUE}}})$.

To estimate $\sum (n'_{q_{\text{SGS}}})$, $n_{q_{\text{SGS}}}(w_j)$ was obtained by Eq. (8); and the first derivative $n'_{q_{\text{SGS}}}(w_j)$ was squared then summed up. To estimate $\sum (n'_{q_{\text{TRUE}}})$, $n_{q_{\text{TRUE}}}(w_j)$ was generated in Eq. (10); and the first derivative $n'_{q_{\text{TRUE}}}(w_j)$ was squared then summed up.

For the single spectrum $q_{Nq}(w_j)$ indicated by blue in Fig. 1(a) and (b), using Eq. (9), the corrected scaling factor $k_{\text{Ncorr}}$ became $k_{\text{Ncorr}} = 1.00$ (before the correction, $k = k_N = 0.98$.) Specifically the correction served effectively.

We confirm that whether the correction of the systematic error serves statistically in a proper manner. Fig. 2(b) summarizes the scaling factors before / after the correction, $k_N$ or $k_{\text{Ncorr}}$, as a function of the variance of $n'_{q_{\text{SGS}}}$, var($n'_{q_{\text{SGS}}}$). The var($n'_{q_{\text{SGS}}}$) corresponds to magnitude of noise. The scaling factors before the correction showed the systematic errors. By increasing noise, the scaling factor became smaller. The scaling factor values after the correction were located around 1.00, indicating that our proposed correction of the systematic error caused by noise served statistically in a proper manner. The arrow in Fig. 2(b) indicates the group for the magnitude of noise of the simulated spectra $p_{Np}(w_j)$ in Fig. 1(a) and (b).

In conclusion, we proposed the practical way to correct the systematic error and the simulation study expecting practical application. The way of correcting systematic error caused by noise served effectively. In near future, we will study other problems to be solved, for instance, about
“strong curvature of a spectral profile” and “highly overlapped bands”.

References

1) G. J. Puppels, F. F. M. de Mul, C. Otto, J. Greve, M. Robert-Nicoud, D. J. Arndt-Jovin, and T. M. Jovin, *Nature*, **1990**, *347*, 301.
2) Y. Huang, T. Karashima, M. Yamamoto, and H. Hamaguchi, *Biochemistry*, **2005**, *44*, 10009.
3) C. Xie, J. Mace, M. A. Dinno, Y. Q. Li, W. Tang, R. J. Newton, and P. J. Gemperline, *Anal. Chem.*, **2005**, *77*, 4390.
4) D. Zhang, P. Wang, M. N. Slipchenko, and J. Cheng, *Acc. Chem. Res.*, **2014**, *47*, 2282.
5) S. Morita, S. Takanezawa, M. Hiroshima, T. Mitsui, Y. Ozaki, and Y. Sako, *Biophys. J.*, **2014**, *107*, 2221.
6) H. Noothalapan and S. Shigeto, *Anal. Chem.*, **2014**, *86*, 7828.
7) D. D. Lee and H. Seung, *Nature*, **1999**, *401*, 788.
8) I. Noda, *Appl. Spectrosc.*, **1993**, *47*, 1329.
9) T. Sugawara, T. Nakabayashi, and S. Morita, *Anal. Sci.*, **2018**, *34*, 845.
10) M. He, D. Miyata, T. Nakabayashi, and S. Morita, *Anal. Sci.*, **2020**, in press.
11) T. Sugawara, Q. Yang, T. Nakabayashi, and S. Morita, *Anal. Sci.*, **2017**, *33*, 1323.
12) A. F. Ujuagu, M. Furuta, T. Nakabayashi, L. Ito, and S. Morita, *Appl. Phys. Express*, **2020**, *13*, 036501.
13) A. Savitzky and M. J. E. Golay, *Anal. Chem.*, **1964**, *36*, 1627.
Fig. 1. (a) Synthesized bio-Raman spectra simulated for application. The blue color signifies a single bio-Raman spectrum \( p(w_j) = p_N(w_j) \) arbitrarily selected. Please note that the peak positions and the heights in a simulated Raman spectrum are arbitrarily determined; thus, there is no physical / chemical meaning on them. The red color signifies a background spectrum \( q(w_j) = q_N(w_j) \). (b) The subtracted spectra \( r(w_j) \) were calculated by the conventional background removal method. The background was not perfectly removed because of systematic error caused by noise. In the region indicated by the asterisk, the background component slightly remains. For instance, for the specifically chosen single bio-Raman spectrum \( p(w_j) \), the scaling factor was estimated \( k = k_N = 0.98 \).
Fig. 2. (a) plots between $\Sigma(n'_{q\text{SGS}}^2)$ and $\Sigma(n'_{q\text{TRUE}}^2)$. The calibration line, $\Sigma(n'_{q\text{SGS}}^2) = f \Sigma(n'_{q\text{TRUE}}^2)$, was assumed the factor (slope) $f$ was determined $f = 6.3$ using the least squares. (b) plots between variance of $n_{q\text{SGS}}$, var($n_{q\text{SGS}}$), and the scaling factors before / after correction of systematic error caused by noise, $k_N$ or $k_{N\text{corr}}$. The variance, var($n_{q\text{SGS}}$), corresponds to magnitude of noise. Blue solid circles are used for $k_N$ and red open circles are used for $k_{N\text{corr}}$. With spectra more noisy, the scaling factors $k = k_N$ became more deviated from the initial setting value $k_{\text{TRUE}} = 1.00$ giving smaller values. After the correction, the corrected scaling factors $k_{N\text{corr}}$ are indicated around 1.00. The arrow in Fig. 2(b) indicates the group for the magnitude of noise of the simulated spectra $p_{Np}$ exemplified in Fig. 1(a) and (b).