Electronic Supplementary Material (ESI) for Analyst.
This journal is © The Royal Society of Chemistry 2016

Please do not adjust margins

Journal Name

ARTICLE

# Optimization of Raman-Spectrum Baseline Correction in Biological Application

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Shuxia Guo,[a] Thomas Bocklitz,[a,b,*] and Jürgen Popp[a,b,c]

In the last decade Raman-spectroscopy has become an invaluable tool for bio-medical diagnostics. However, a manual rating of the subtle spectral differences between normal and abnormal disease states is not possible or practical. Thus it is necessary to combine Raman-spectroscopy with chemometrics in order to build statistical models predicting the disease states directly without manual intervention. Within chemometrical analysis a number of corrections have to be applied to receive robust models. Baseline correction is an important step of the pre-processing, which should remove spectral contributions of fluorescence effects and improve the performance and robustness of statistical models. However, it is demanding, time-consuming, and depends on expert knowledge to select an optimal baseline correction method and its parameters every time working with a new dataset. To circumvent this issue we proposed a genetic algorithm based method to automatically optimize the baseline correction. The investigation was carried out in three main steps. Firstly, a numerical quantitative marker was defined to evaluate the baseline estimation quality. Secondly, a genetic algorithm based methodology was established to search the optimal baseline estimation with the defined quantitative marker as evaluation function. Finally, classification models were utilized to benchmark the performance of the optimized baseline. For comparison, model based baseline optimization was carried out applying the same classifiers. It was proven that our method could provide a semi-optimal and stable baseline estimation without any chemical knowledge required or any additional spectral information used.



Please do not adjust margins

## Simulation

**Spectra construction**

To allow an understanding of the working of our proposed method artificial spectra were constructed within the wavenumber range from 300 to 3000 $cm^{-1}$.

Firstly, we combined eight Gaussian peaks at arbitrary wavenumber positions 500, 700, 900, 970, 1050, 1500, 2000 and 2600 $cm^{-1}$. The maximum intensities of these peaks varied within the range 800-2200, while the full width at half maximum (FWHM) varied within the interval 50-120 $cm^{-1}$. The three peaks at 900, 970, and 1050 $cm^{-1}$ were chosen so close that they overlapped to generate a complex structure (See Fig. S1 (a)).

Secondly, three series of curves were prepared to construct baseline profiles. Each series contains a second-order polynomial and five Gaussian peaks with large bandwidth. Details about the related parameters are listed in Tab. S1. Accordingly, three baseline profiles were created by adding up all compositions within each curve series, which are plotted in Fig. S1 (b).

Finally, the pure spectrum (Fig. S1(a)) was added up with the three baseline profiles (Fig. S1(b)), generating three spectra. Additionally, Poisson distributed noise was generated and added up to these three spectra. In this way, three simulated spectra with different baseline patterns were created, as shown in Fig. S1 (c).

Tab. S1. Parameters of the three curve series for constructing baseline profiles

| | | Gaussian 1 | Gaussian 2 | Gaussian 3 | Gaussian 4 | Gaussian 5 | 2nd order Polynomial |
|---|---|---|---|---|---|---|---|
| **Series 1** | A | 1500 | -750 | 1125 | 1125 | 1875 | $0.00015*(x-600)^2$ |
| | $\mu$ | 500 | 1000 | 1500 | 2300 | 2800 | |
| | $\sigma$ | 900 | 900 | 900 | 900 | 900 | |
| **Series 2** | A | 2250 | -375 | -1125 | 1875 | 1500 | $0.00015*(x-1500)^2$ |
| | $\mu$ | 400 | 800 | 1200 | 1800 | 2800 | |
| | $\sigma$ | 1000 | 1000 | 1000 | 1000 | 1000 | |
| **Series 3** | A | 1875 | -1125 | 1500 | -1125 | 1875 | $0.00015*(x-2500)^2$ |
| | $\mu$ | 600 | 1200 | 1700 | 2000 | 2700 | |
| | $\sigma$ | 1100 | 1100 | 1100 | 1100 | 1100 | |

**Grid search**

All simulated spectra were smoothed by a Savitzky-Golay filtering with a window width of 11 and an order of 2. Reference spectra were obtained by subtracting the true baseline profiles from the smoothed spectra. The subtracted spectra were used as reference spectra instead of the pure spectrum shown in Fig. S1(a). This was done to make the reference spectra comparable to baseline corrected spectra, which were also generated from the smoothed spectra.



This journal is © The Royal Society of Chemistry 20xx

The same grid search procedure as for the real Raman spectra was performed on the smoothed spectra. The Euclidean distance

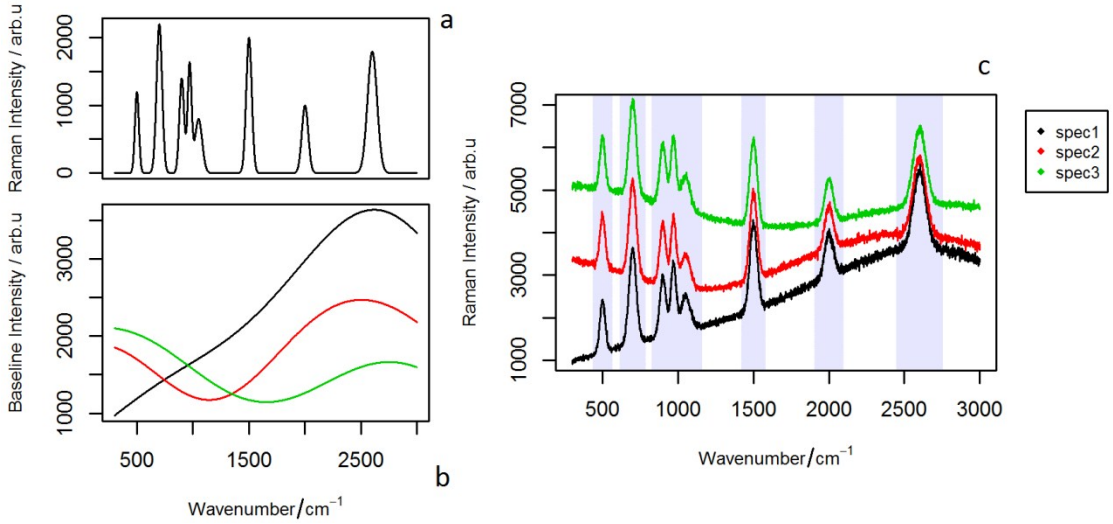


Fig. S1 Constructed Spectra for simulation. (a) Pure spectrum containing eight peaks with various intensities and FWHMs. A wide peak was constructed by an overlap of the three peaks at 900, 970, and 1050 cm$^{-1}$. (b) Three baseline profiles were generated by combination of one second order polynomial and five Gaussian peaks with large bandwidth. (c) Simulated spectra, created from three parts, pure spectra, baseline, and Poisson distributed noise.

between the baseline corrected spectra and the reference spectra ($\|S_c - S_r\|_2$) was computed and employed as the benchmark of baseline correction. Here we assume that a 'correct' baseline correction provide a minimal Euclidean distance. Meanwhile, the $R^{12}$ values were calculated according to eq. (1). The grey shaded regions shown in Fig. S1(c) were used as peak region, while the rest was utilized as no-Raman-information region. All three Raman spectra were vector normalized within the whole wavenumber range. According to our purposed method, an optimal baseline correction can be expected at the minimum of $R^{12}$. Afterwards, the optimal baseline corrections were selected according to two mechanisms, i.e., the minimal $R^{12}$ value and the minimal Euclidean distance. The results were plotted in Fig. S2, where the region with the overlapped peak was highlighted in a zoomed image. As shown for the first two simulated spectra, the two mechanisms yielded an identical baseline correction. While a slight difference was observed for the third simulated spectrum. Besides this neither of mechanism can exactly eliminate the baseline, demonstrated by the mismatch between the reference and the baseline corrected spectra. This also indicates the impossibility to exactly separate baselines and Raman signals. Nevertheless, a high consistency was observed between the baseline corrections optimized by these two mechanisms. That is to say, the $R^{12}$ value can indeed reflect the goodness of baseline correction without knowing the reference spectra, which means a feasibility to optimize baseline correction according to $R^{12}$ values.

This journal is © The Royal Society of Chemistry 20xx

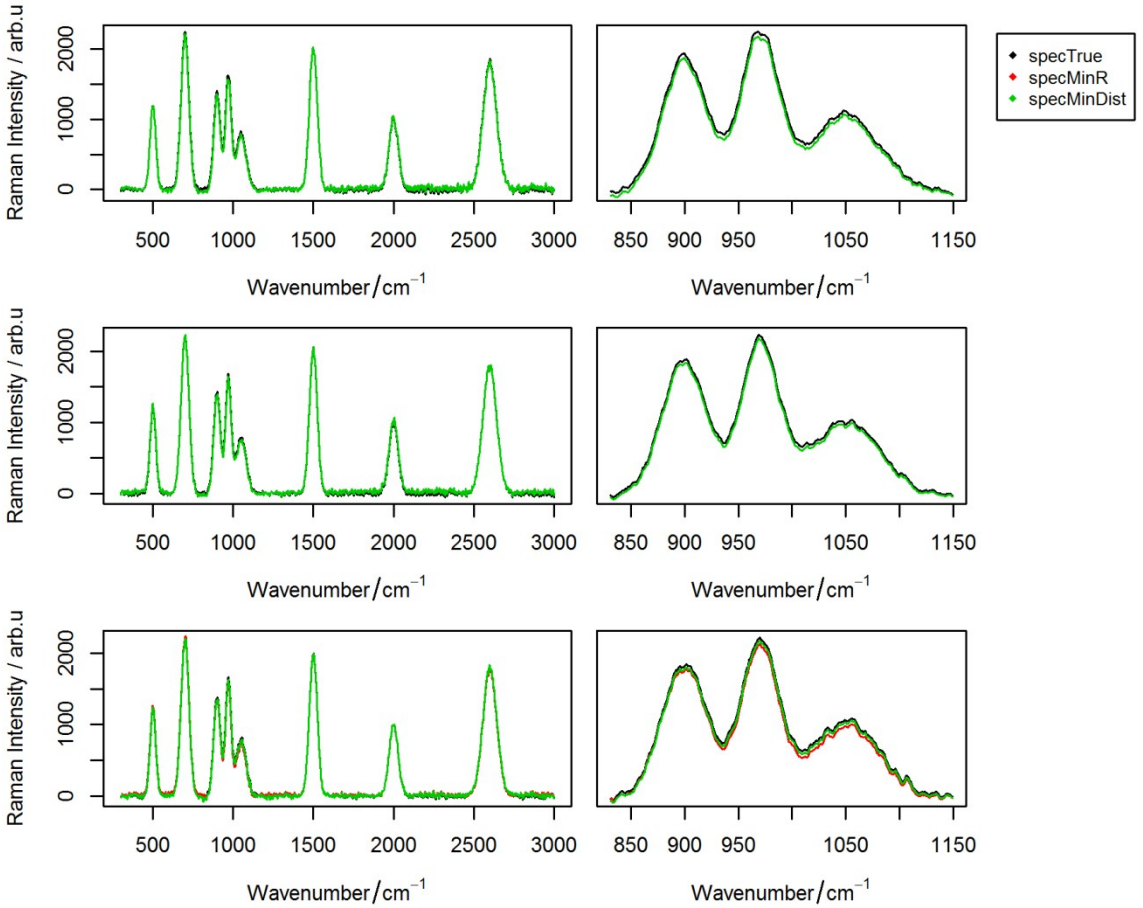


Fig. S2. The reference spectra (black) and optimal baseline corrections obtained by the minimal $R^{12}$ value (red) and minimal Euclidean distance (green). The three overlapped peaks are highlighted in a zoomed version on the right. For the first two simulated spectra, the two mechanisms yielded the identical optimal baseline correction. A slight difference was observed for the third simulated spectrum. Nevertheless, a high consistency was observed between the baseline corrections optimized by the two mechanisms. Besides, a mismatch between the reference and the baseline corrected spectra is observed, indicating the impossibility to exactly separate baselines and Raman signals.



This journal is © The Royal Society of Chemistry 20xx