New Perspectives on Preprocessing of Spectroscopic Data: Effective preprocessing of spectroscopic information for multivariate data analysis

C. Cornett[1,*], J. Rantanen[1], J.X. Wu[1], T.P. Munk[1], V. Koradia[1], M. Savolainen[1], and F. Tian[1]

OBJECTIVES

To emphasize rational preprocessing of spectra for multi-variate data analysis and introduce a simple and com-putationally efficient multiplicative scatter correction and some surprising uses of an older algorithm

METHODS

The problem. Raman, NIR etc. spectra often show significant effects from the physical setup of the measurements. If the distance between e.g. a fiber-optic probe and the sample changes a different proportion of the light will reach the sample and the detector (figure 1a). If this is the only phenomenon influencing the measurement a purely multiplicative effect will be observed (figure 1b). For Raman spectra fluorescence can also be a major problem.

These effects can be extremely detrimental for quantitative measurements and are often reduced by employing a suitable preprocessing method before further data analysis. Popular choices are Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV), and 1st or 2nd derivatives. MSC and SNV provide a measure of correction for both multiplicative and additive effects, while derivatives corrects additive effects

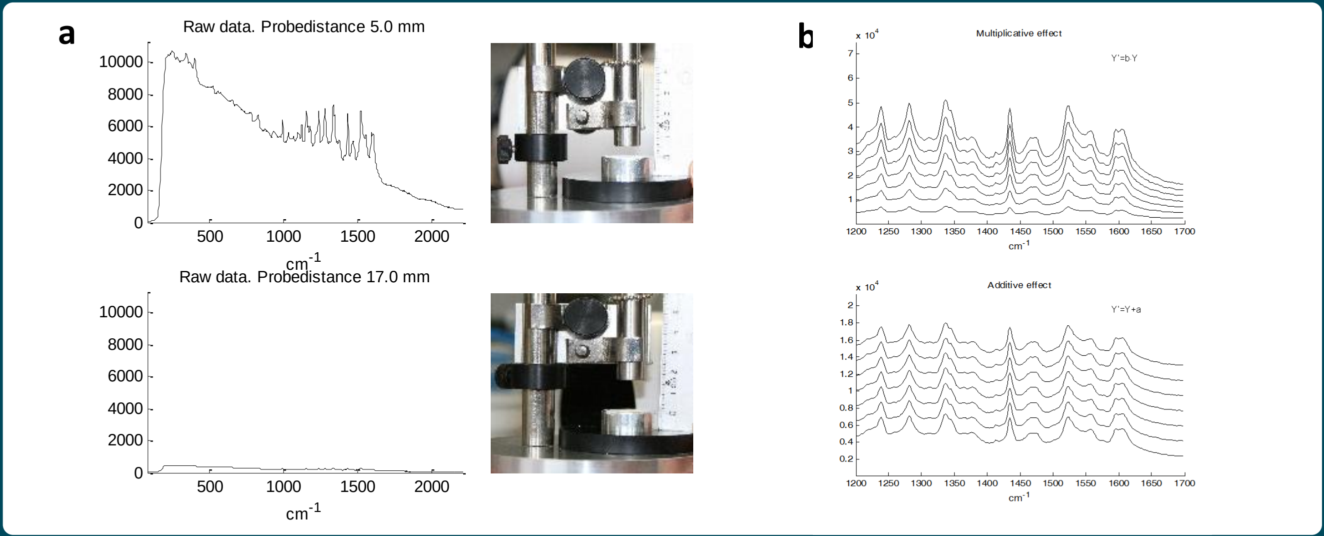


Figure 1. a) Effect of changed probe distance (and focus of the LASER).
b) Definition of multiplicative and additive effects.

However, in some cases MSC and SNV can compromise linearity (figure 2), if the effect to be corrected for is not dominant, that is if a third component in the sample (apart from physical effects and chemical information of interest). Recently MSC has been developed to extended Multiplicative Scatter Correction (eMSC), but for some applications separation of scaling (multiplicative correction) and correction for additive effects would be preferred.

If an internal standard (IS) is present it is fairly simple to obtain this separation. However, if no IS is present, for Raman spectroscopy the baseline (in the absence of fluorescence) can be used as this represents the “foot” of the Rayleigh line and for physical effects like probe distance is sufficiently similar to an IS.

a

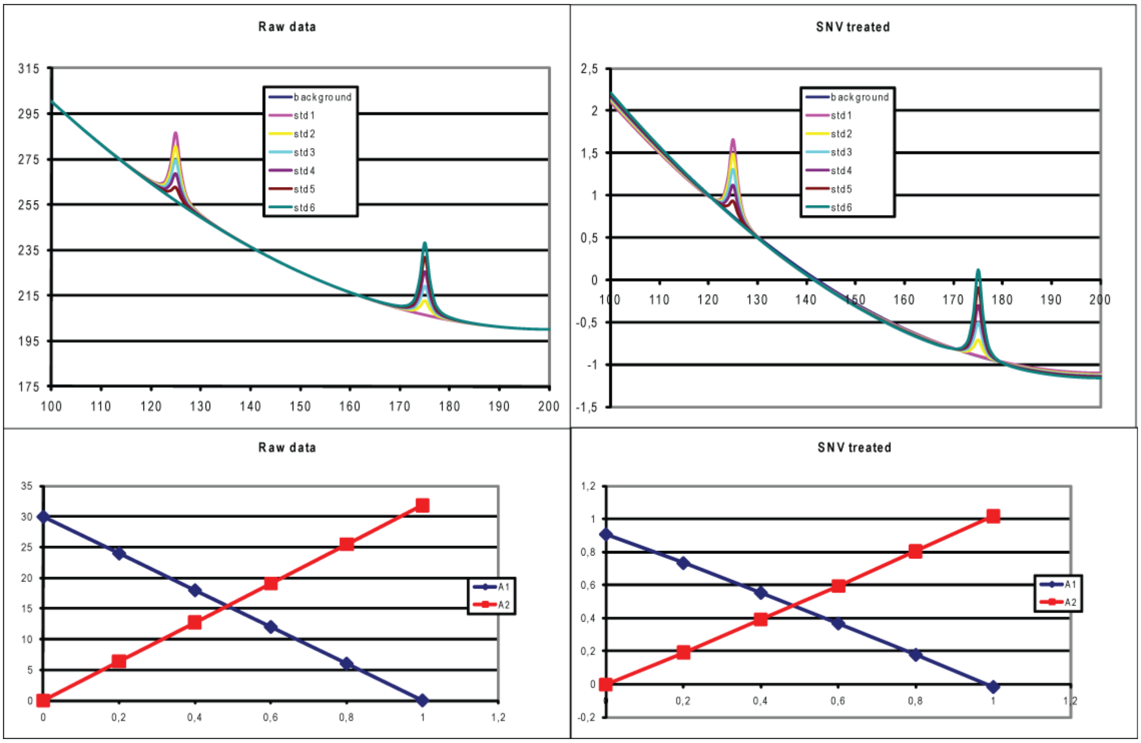


Figure 2. SNV on simulated data (linear mixtures of two compounds with slightly different molar absorption coefficient/scattering efficiencies.

Each: upper left untreated data, upper right SNV treated data. Lower graphs: maximum of peak as a function of concentration.

a) Large background, note the linear “standard curve”.

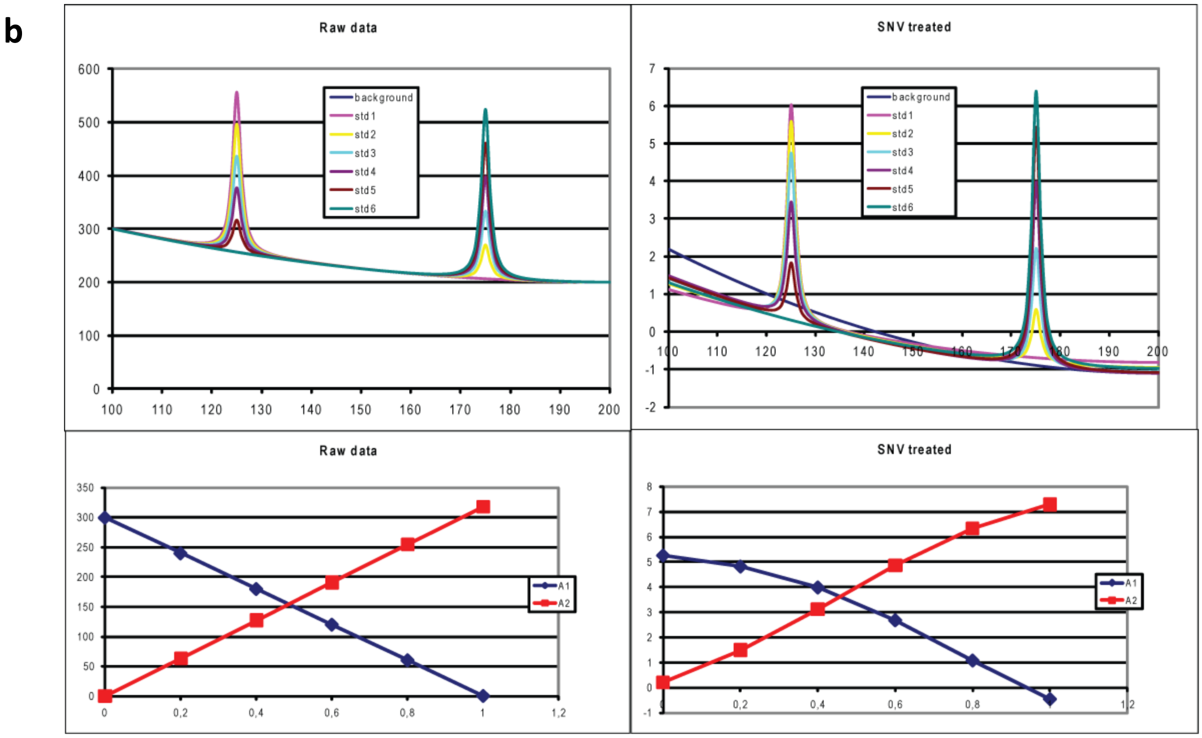


Figure 2. SNV on simulated data (linear mixtures of two compounds with slightly different molar absorption coefficient/scattering efficiencies.

Each: upper left untreated data, upper right SNV treated data. Lower graphs: maximum of peak as a function of concentration.

b) Same background, signals 10 times stronger, note the non-linear “standard curve”.

$$I(\bar{\nu}_i)_{SNV} = \frac{I(\bar{\nu}_i) - \langle I(\bar{\nu}_i)\rangle}{\sqrt{{}^{obs}\sigma^2}} = \frac{I(\bar{\nu}_i) - \langle I(\bar{\nu}_i)\rangle}{\sqrt{{}^{obs}\sigma^2_{Interferent} + {}^{obs}\sigma^2_{Desired} + {}^{obs}\sigma^2_{Physical} + \ldots}}$$

⬆

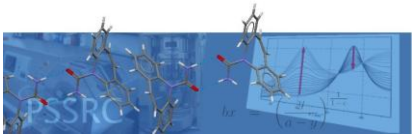
Equation 1. SNV.

$I(\bar{\nu}_i)$ i’th point of untreated spectrum.

$I(\bar{\nu}_i)_{SNV}$ i’th point of treated spectrum.

$\langle I(\bar{\nu}_i)\rangle$ mean/average of untreated spectrum.

$\sqrt{{}^{obs}\sigma^2}$ standard deviation of untreated spectrum.



RESULTS

Simple Multiplicative Correction (SIMPC)

We have chosen to use a simple similarity measure stating that selected baseline intervals should be similar to e.g. the average of the intervals of all the spectra. This works surprisingly well when using the scalar product of the averaged baseline intervals with the baseline intervals of the i’th spectrum. We have also implemented a version with additive correction included (Simple Multiplicative and Additive Correction -SIMAPC), which is very similar to what could be called “interval MSC”.

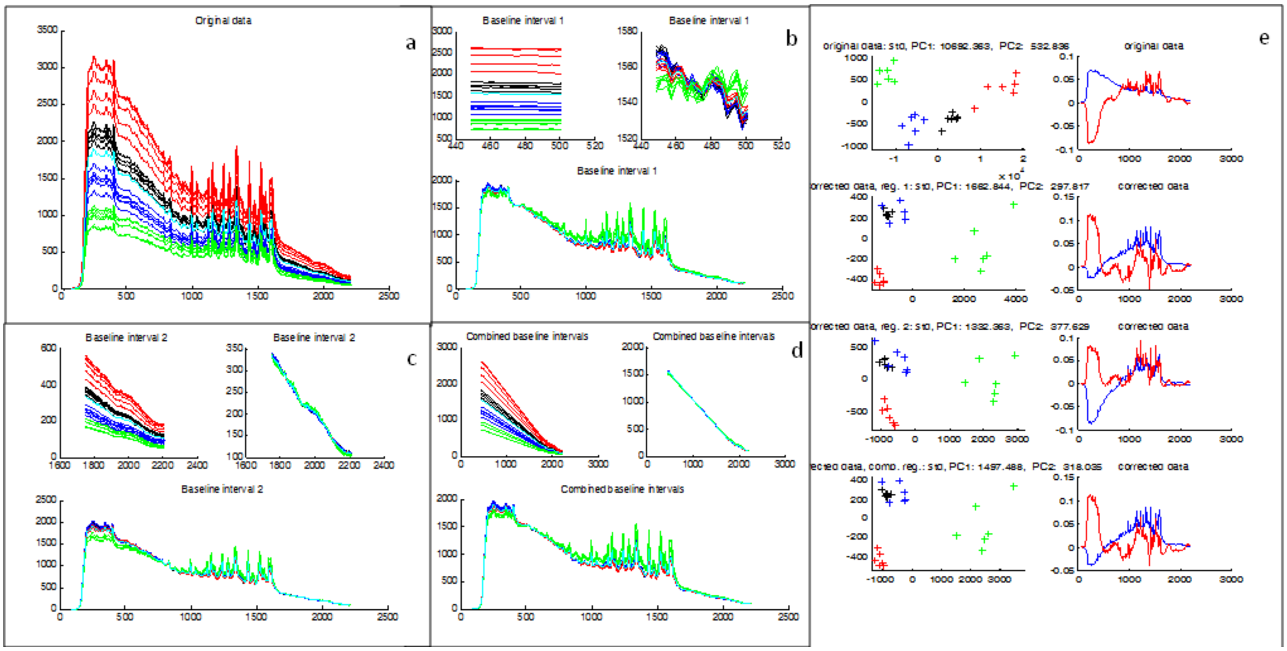


Figure 3. SIMPC applied to calibration data set. A: raw data. b and c: two different baseline intervals. d: baseline intervals b and c combined. e: Principal component analy

RESULTS

Simple Multiplicative Correction (SIMPC)

We have chosen to use a simple similarity measure stating that selected baseline intervals should be similar to e.g. the average of the intervals of all the spectra. This works surprisingly well when using the scalar product of the averaged baseline intervals with the baseline intervals of the i'th spectrum. We have also implemented a version with additive correction included (Simple Multiplicative and Additive Correction -SIMAPC), which is very similar to what could be called “interval MSC”.

Baseline/additive Correction

For this purpose we have implemented an algorithm originally created for removing fluorescence backgrounds from Raman Spectra (Lieber and A Mahadevan-Jansen. Applied Spectroscopy 57 (11), 2003, p1363-1367).

Examples

In figure 3 we show an example of SIMPC applied to a calibration dataset.

In figure 4 we show an example of SIMPC applied to a calibration dataset, followed by baseline removal.

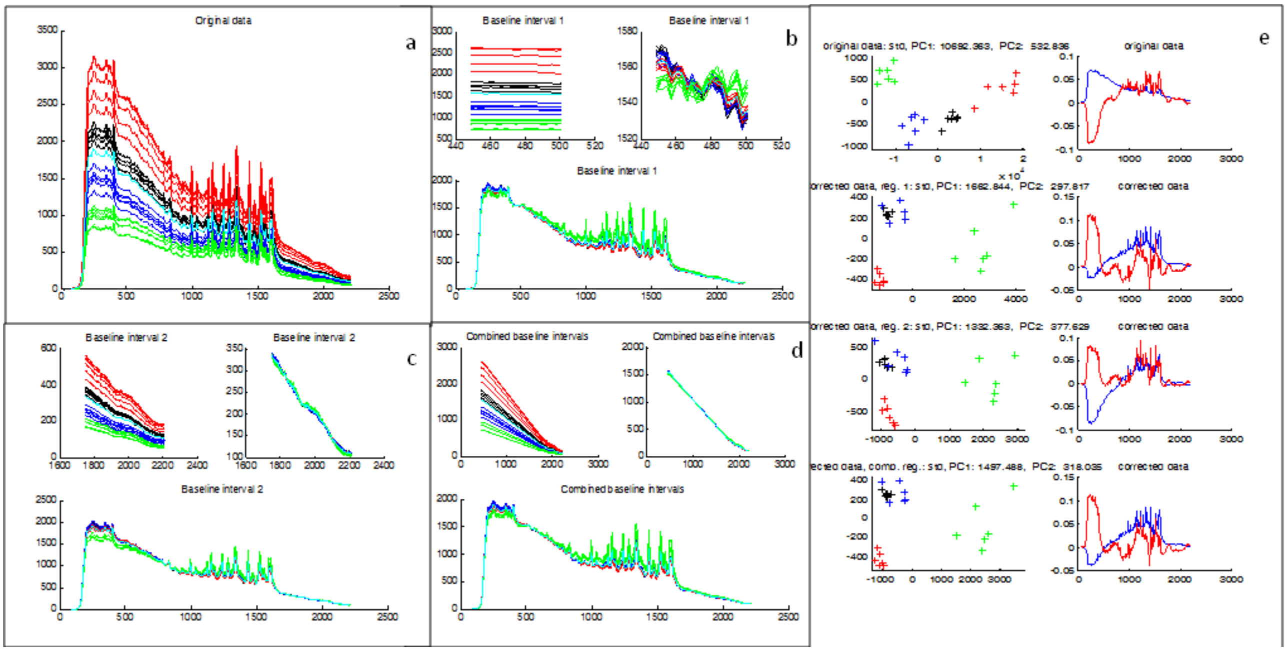


Figure 3. SIMPC applied to calibration data set. A: raw data. b and c: two different baseline intervals. d: baseline intervals b and c combined. e: Principal component analysis of data a to d.

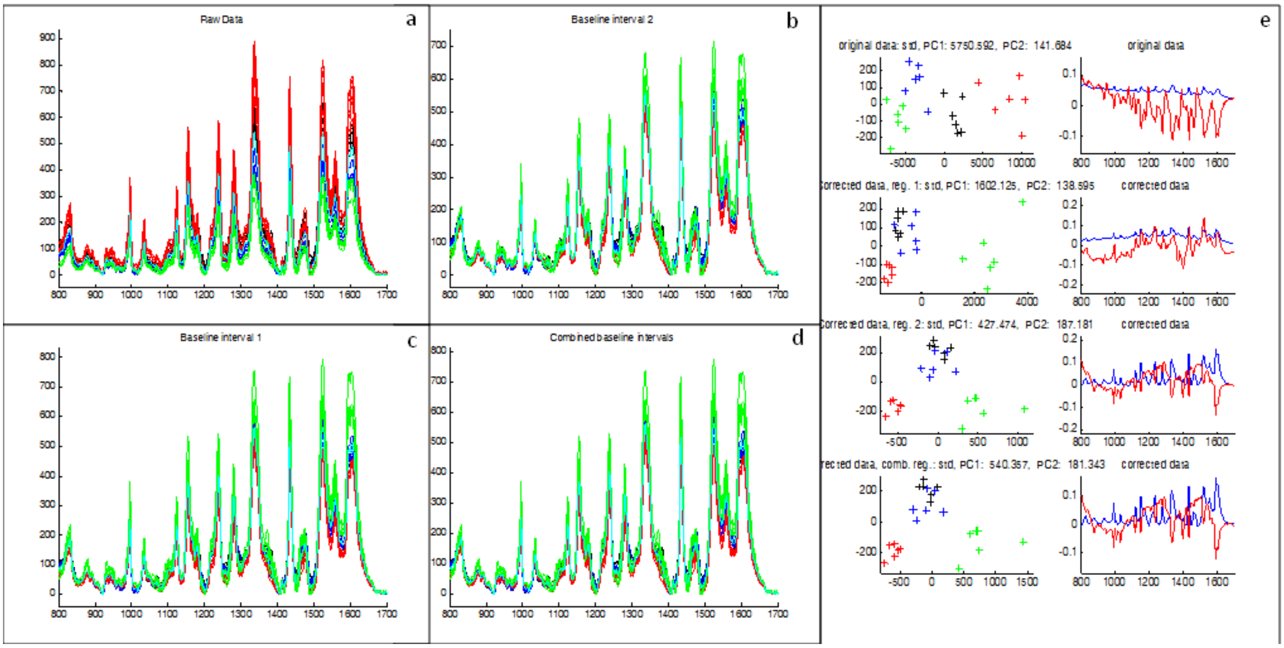


Figure 4. SIMPC applied to calibration data set followed by baseline removal, applied to an interval of the spectra. Otherwise same as Figure 3.

CONCLUSIONS

- We have drawn attention to some possible pitfalls inherent in the MSC and SNV algorithms, and an indication of when these preprocessing methods may be expected to be safe.

- We have pointed out a set of simple, computationally efficient preprocessing method that supplements already known methods