B&W TEK
A Metrohm Group Company

Technical Note

Spectral Preprocessing for Raman Quantitative Analysis

Introduction

Raman spectroscopy has become an increasingly common technique for process analytics in the pharmaceutical and chemical industries due to its nondestructive measurements, fast analysis times, and ability to do both qualitative and quantitative analysis. Spectral preprocessing algorithms are routinely applied to quantitative spectroscopic data in order to enhance spectral features while minimizing variability unrelated to the analyte in question. Understanding the possible preprocessing steps and how to apply them correctly may be daunting for those without a formal background in chemometrics. The goal of this document is to discuss the main preprocessing options pertinent to Raman spectroscopy with real applications examples, and to review the algorithms available in B&W Tek and Metrohm software so that the reader becomes comfortable applying them to build Raman quantitative models.

19 Shea Way Newark, DE 19713, USA • Tel :1-855-MY-BWTEK • Fax +1-302-368-7830 • Web: www.bwtek.com
Doc Rev: 410000057-A (05/25/2021)

Spectral Preprocessing of Raman Data

Spectral preprocessing is used to remove or minimize effects in spectral data that are not directly related to the spectral changes associated with the system under study. Preprocessing is also applied to enhance the ability to distinguish subtle spectral differences such as small peak intensity or spectral shifts.

Let's explore some of the spectral preprocessing steps particularly relevant to Raman data.

Baseline Removal

Baseline removal or baseline correction (referred to as de-trend in Vision software) is especially useful in removing variable backgrounds like fluorescence or interfering ambient light from Raman data when clear Raman peaks are still present in a spectrum. There are many subtle mathematical approaches to baseline removal, but it generally involves calculating a least squares fit of a polynomial to describe the baseline of the spectrum, and then subtracting that function from the spectrum. Figure 1 shows an example of a baseline correction in BWSpec software applied to a spectrum of carbon black powder. Carbon samples have varying backgrounds and should be baseline corrected before additional spectral analysis, such as calculation of the intensity ratio of the D and G bands.

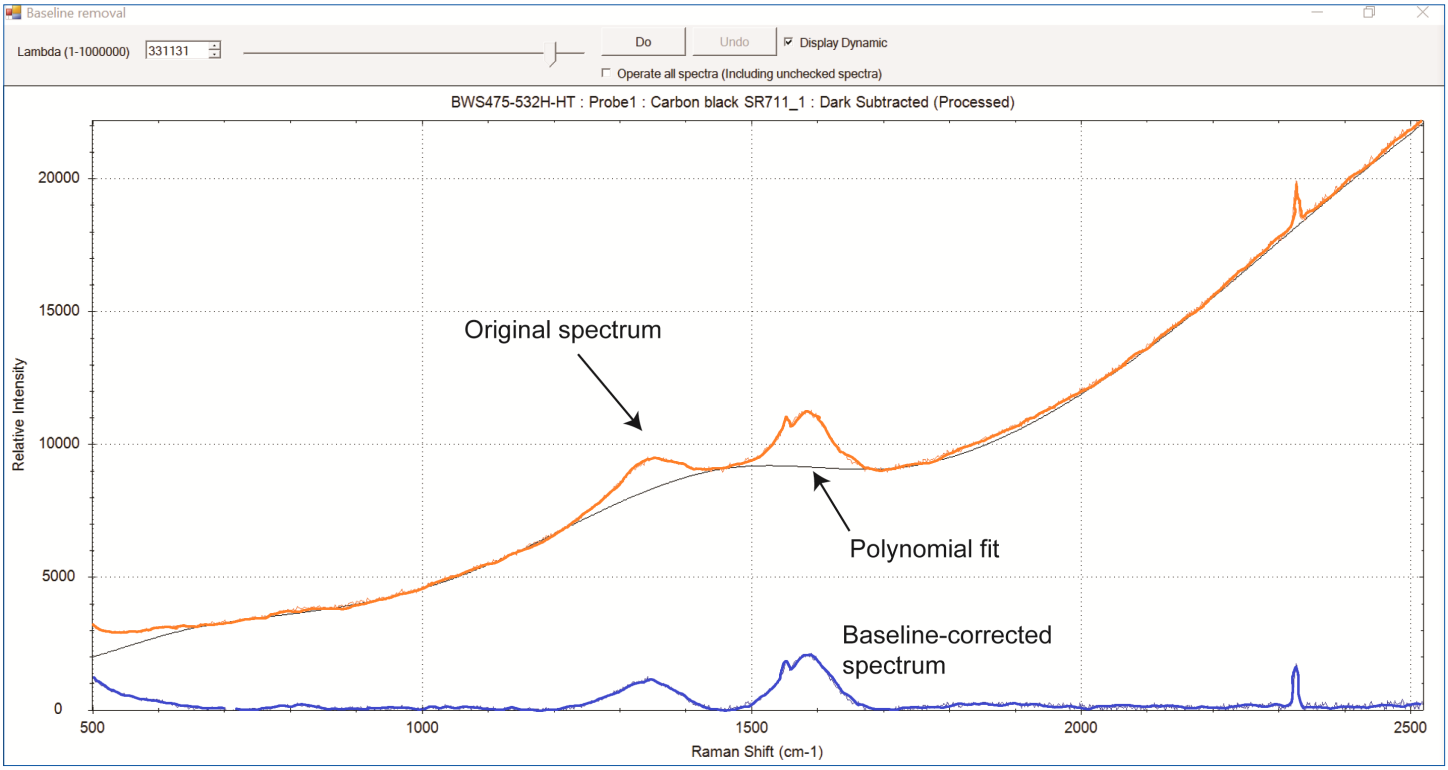

Figure 1. Baseline correction of a carbon black sample. Manipulating the slide bar changes the polynomial fit (spectra have been artificially bolded for clarity)

Baseline removal algorithms are available in BWSpec, BWIQ and Vision software. Caution should be taken when applying baseline corrections for large datasets with varying baselines, as one fit may not be optimized for all of the spectra in the set. Instead, derivatives are recommended for removing baseline effects from quantitative data sets.

Derivatives

Derivatives are common preprocessing steps for both Raman and NIR data. Derivatives are applied to spectral data to enhance spectral features and eliminate baseline effects. First and second-order derivatives are typically used, as higher-order derivatives can amplify unwanted noise. There are multiple approaches to derivatives in B&W Tek and Vision software, but by far the most commonly used for Raman data is the Savitzky-Golay derivative.

Savtizky-Golay derivatives are performed by fitting a polynomial to a short segment of data points, and then the derivative of this function is calculated at the center point of the segment. The segment is often referred to as a "window" size. It is typically best to use a large window size, as small window sizes generate noisier data that is more sensitive to slight variations in Raman shift. Figure 2 shows the comparison of two Savitzky-Golay first derivatives with different window sizes applied to the same set of data; the spectrum with the larger window size shows considerably less noise than the spectrum with the smaller window.

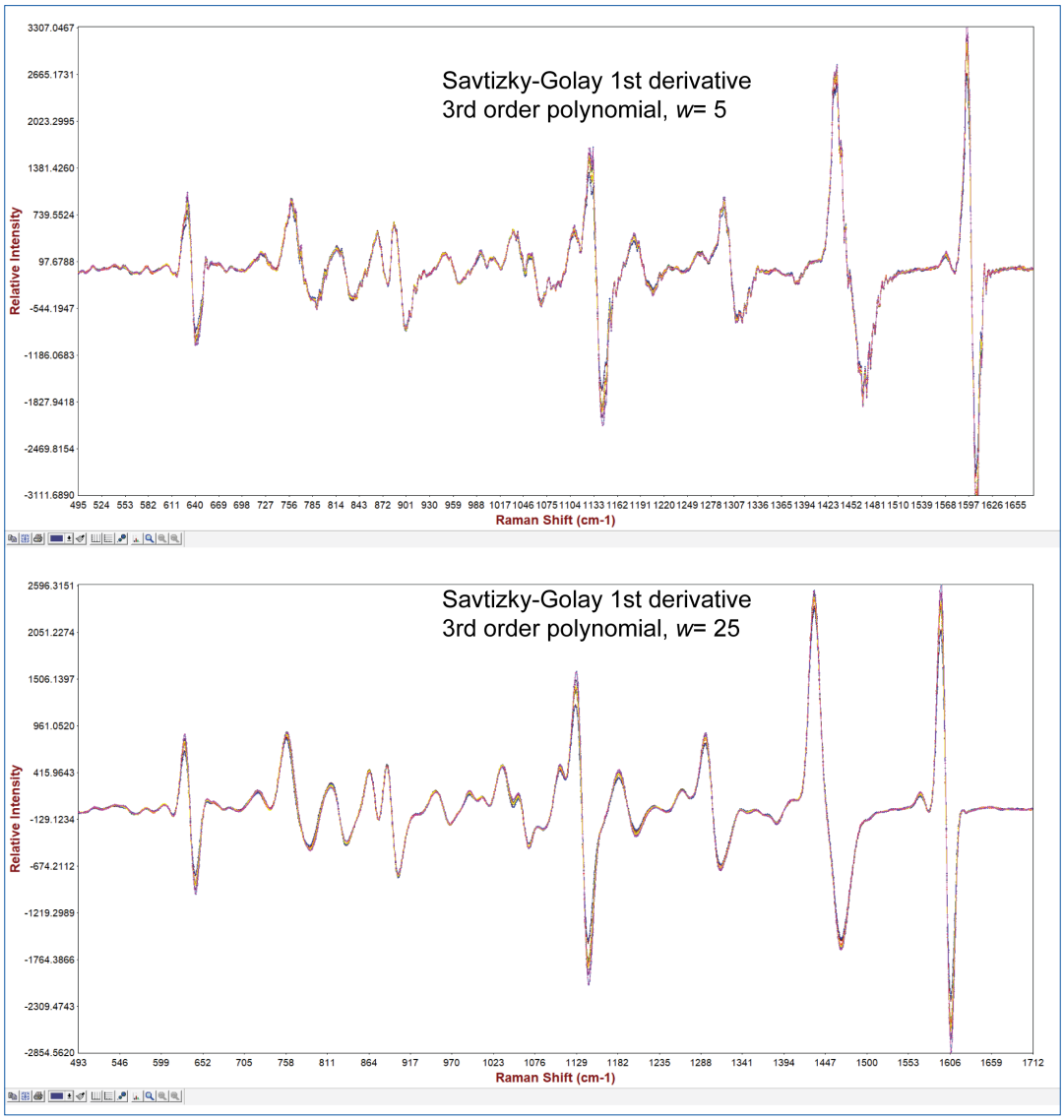

Figure 2: (Top) Data processed with SG 1st derivative with a window= 5, and (bottom) data processed with SG 1st derivative with a window=25. The spectra are samples of an emulsifier used in pesticides with increasing amounts of analyte.

Region Selection

Models can be built with specific spectral regions in order to exclude regions where there is little information or unrelated variability; this results in simpler models with fewer latent variables. In BWIQ and Vision software, spectral regions can be manually selected. Typically, selecting the whole fingerprint region (~200-1800 cm^{-1}) is sufficient for building a Raman model. An experienced spectroscopist may choose more tailored regions, but a common mistake when using multivariate regression is to select too narrow a region that contains only the features corresponding to the analyte of interest. Determination of analyte concentration requires quantification of both the analyte and the reference (everything else), so if only the analyte features are included, the model lacks a reference and may become unstable. This is especially true if a normalization step is also applied.

As an example, consider a simple partial least squares (PLS) model quantifying benzonitrile in a mixture with cyclohexane. The model consists of 12 spectra from 6 samples, with benzonitrile concentration ranging from 10% to 35% v/v. As shown in Figure 3, benzonitrile has a strong peak at 2232 cm^{-1} that is attributed to the nitrile CN stretch.

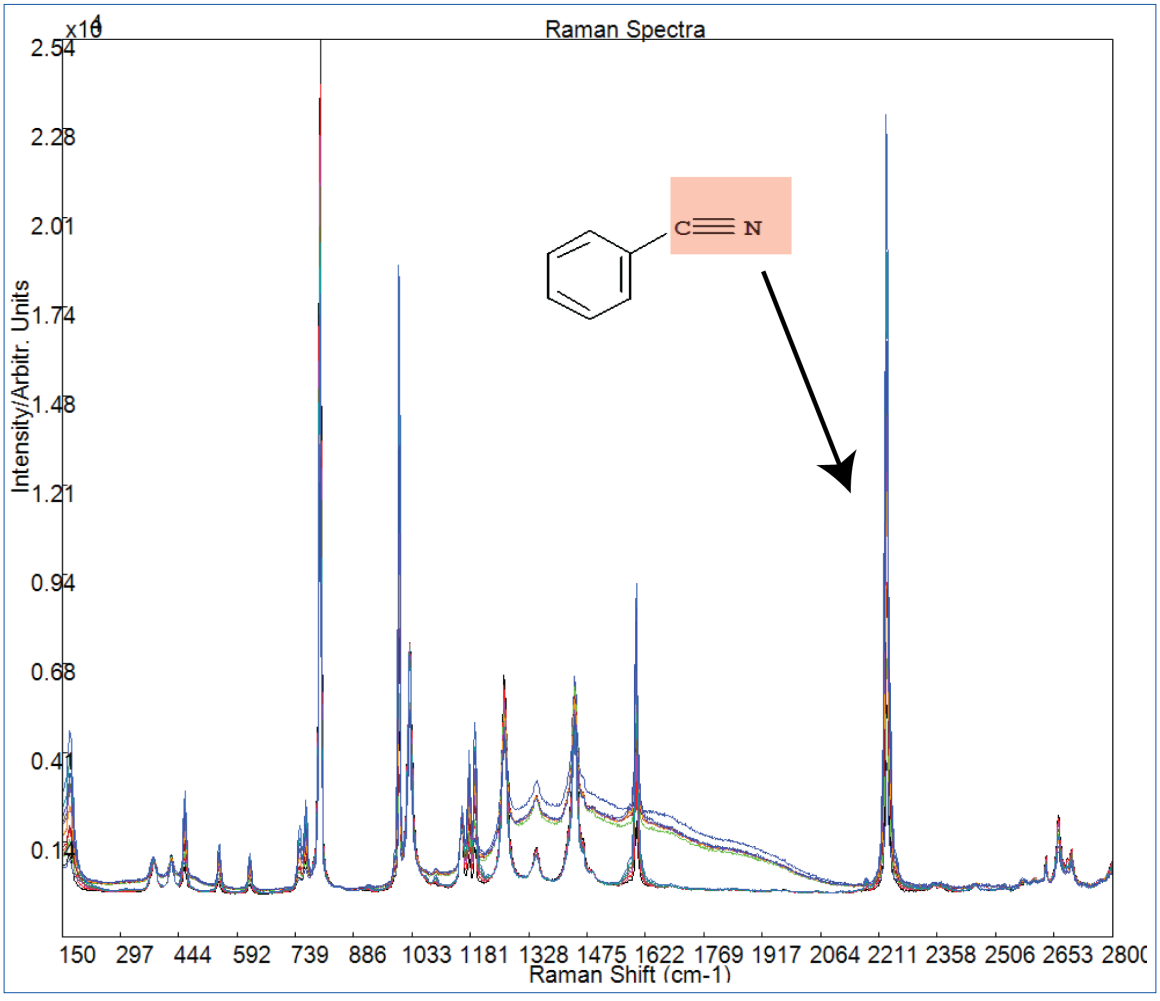

Figure 3: Raw spectra of benzonitrile and cyclohexane mixture.

Figure 4 shows the predicted vs. measured plots from two PLS models created from the data. If the region from 300-2300 cm^{-1} is selected (Figure 4a), a model with good linearity is obtained with a very low RMSE of 0.21%. On the other hand, if a narrow region of 2150-2300 cm^{-1} is selected (Figure 4b), the model has a poor linearity and a high RMSE of 1.64%. The poor performance of the latter model is due to the lack of reference, which results in identical spectral intensity of all spectra after normalization.

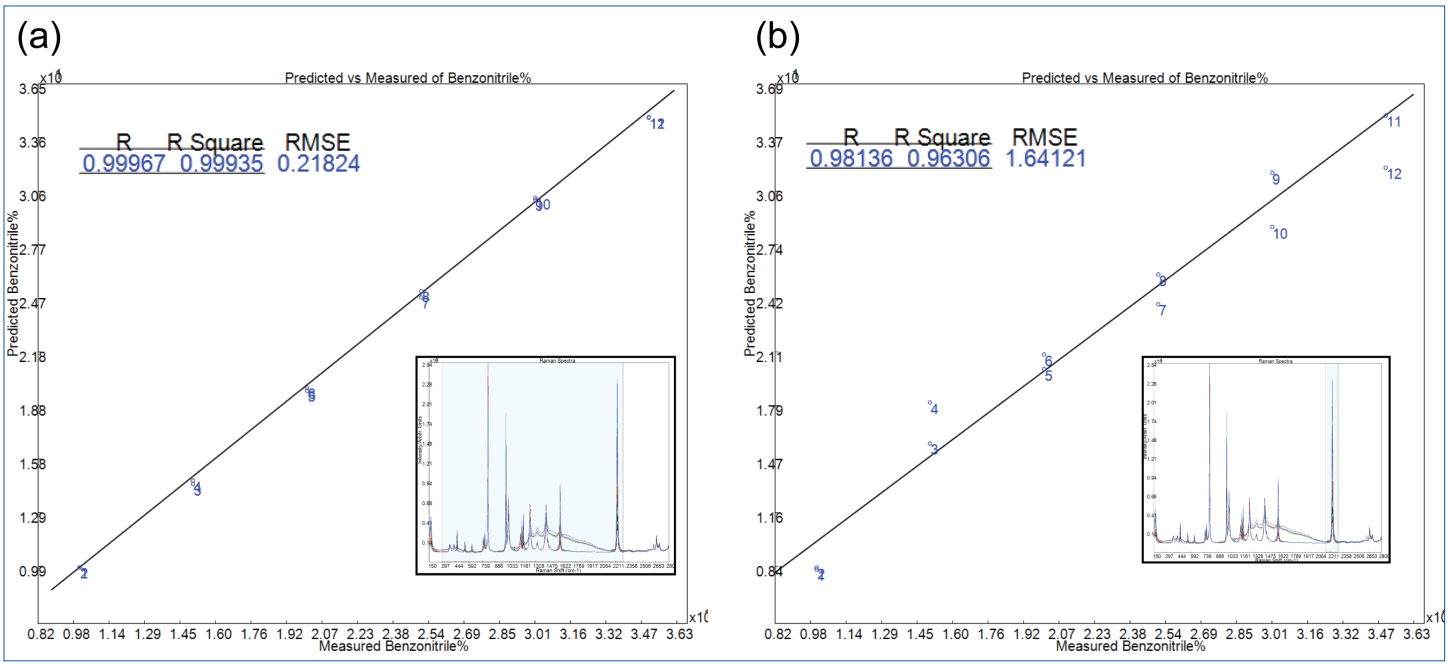

Figure 4: (a) PLS model with 300-2300 cm^{-1} selected and (b) PLS model with 2150-2300 cm^{-1} selected.

Normalization

Quantitative Raman models are highly influenced by fluctuations in overall spectral intensity. Intensity fluctuations can arise from many factors, such as spectrometer throughput drift, excitation power instability, pathlength differences, and physical differences of sampling positioning. Although within a short period these fluctuations can be prevented, in the long term they are inevitable. Spectral normalization can effectively remove the effects of overall intensity variation, and therefore is a necessary step for construction of a robust regression model.

There are many different mathematical approaches to normalization in spectroscopy. Standard normal variate (SNV) and multiplicative scattering correction (MSC) are the two most common normalization algorithms for vibrational spectroscopy and are available for use in BWIQ and Vision software. Spectroscopists tend to favor SNV over MSC, because MSC is a scatter correction based on the mean of the entire data set, while SNV is based on the standard deviation of an individual sample spectrum, and not dependent on the entire data set.

Figure 5 shows spectra of aqueous solutions containing varying amounts of glucose and lactate. SNV is applied and compared to the non-normalized data with baseline variations (insert). Region selection should be selected prior to normalization so that the excluded regions are not taken into account.

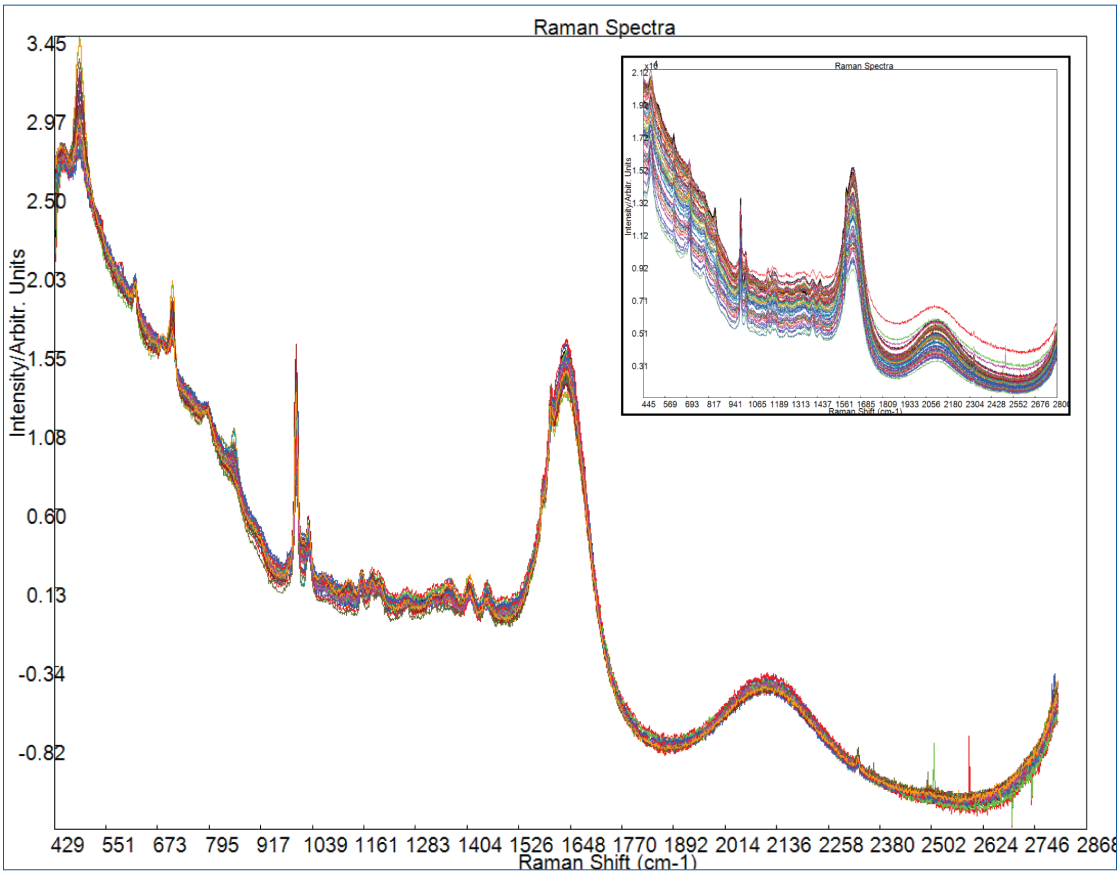

Figure 5: Raman data set of glucose and lactate in water with SNV applied (insert shows raw non-normalized data)

Mean Centering

Mean centering subtracts the mean spectrum of the dataset from each spectrum. This is a necessary step for PLS and PCA based models, as both techniques analyze the variance of the data set. BWIQ has a separate step for mean centering and therefore must be included explicitly as a preprocessing step, while Vision software mean centers the training spectra implicitly.

Real-Use Example

We will review a real applications example to apply the information we learned in the previous section. This data was collected using the B&W Tek QTRam portable transmission Raman setup. The samples are a set of 3.0 mm thick tablets containing a low dose of acetaminophen (a.k.a. paracetamol, APAP), as well as cellulose, mannitol, croscarmellose, and magnesium stearate excipients. The concentration of acetaminophen ranges from 0-1.5% (w/w) and the target concentration is 0.5%, corresponding to a target dosage of 1.5 mg of acetaminophen per ~300 mg pill. To build a model for prediction of new samples on the QTRam, calibration spectra were

collected using an integration time of 3 seconds and 10 spectral averages. Figure 6 shows the raw data; no other preprocessing is applied other than a dark subtraction and a relative intensity correction (see technical note 410000055-A The Importance of Relative Intensity Correction of Raman Data and How to Utilize it for i-Raman Series Instruments in BWSpec for more information). The spectra were imported into BWIQ for processing and to create a PLS model.

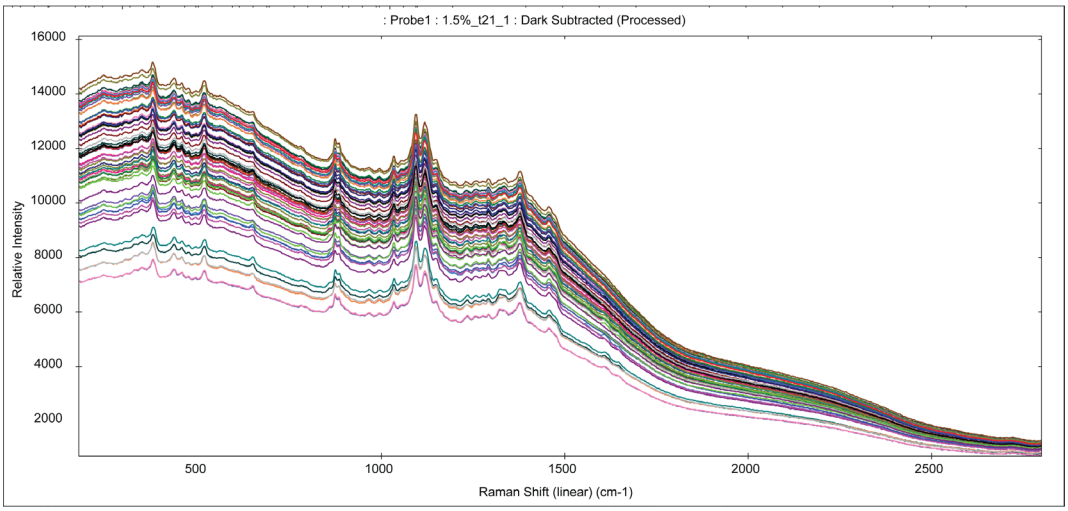

Figure 6: Unprocessed raw spectra of tablets containing 0-1.5% w/w acetaminophen

After data collection, it is useful to compare the sample spectra to the spectra of the individual pure components that make up the sample. Figure 7 shows the comparison of a sample containing 1.5% acetaminophen to spectra of pure acetaminophen, cellulose, and mannitol (croscarmellose and magnesium stearate features are too broad or too weak to distinguish visually). The peaks of the sample spectrum labeled in green are contributed from the acetaminophen in the sample.

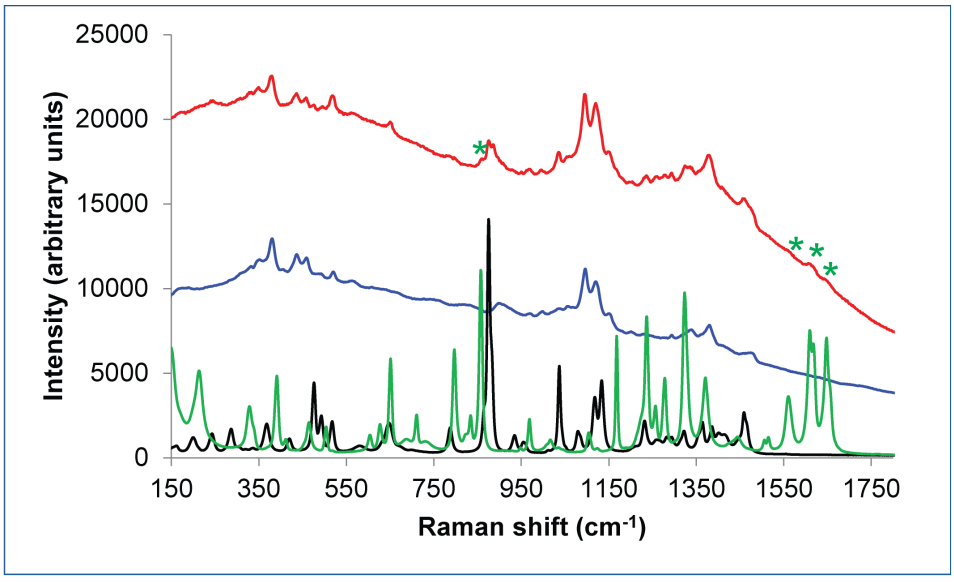

Figure 7: Comparison of spectra of 1.5% APAP tablet sample (red), cellulose (blue), acetaminophen (green), mannitol (black)

Preprocessing Steps

The excipients in the tablets generate a high fluorescence background with 785 nm laser excitation. Baseline correction is not recommend for removing the fluorescence background in this case, as one polynomial fit is not expected to fit every spectrum. Instead, a Savitzky-Golay derivative is preferred as a simpler method. A Savitzky-Golay 1st derivative (cubic order, w=25) is applied to the data to remove the fluorescence background.

When reviewing the spectra, we see that there is significant Raman signal in the fingerprint region, and no significant signal past 1800 cm^{-1}. To exclude the unimportant spectral region from 1800-2800 cm^{-1} in the model, we can apply a manual region selection to only select the fingerprint region (~200-1800 cm^{-1}) for the model. When a manual region selection step is selected in BWIQ, the software will always move it to occur before a normalization step. Figure 8 shows the data processed after Savitzky-Golay 1st derivative, manual region selection, and SNV. The signal at ~860 cm^{-1} and ~1500 cm^{-1} clearly show a change in intensity, which correspond to the increasing concentration of acetaminophen.

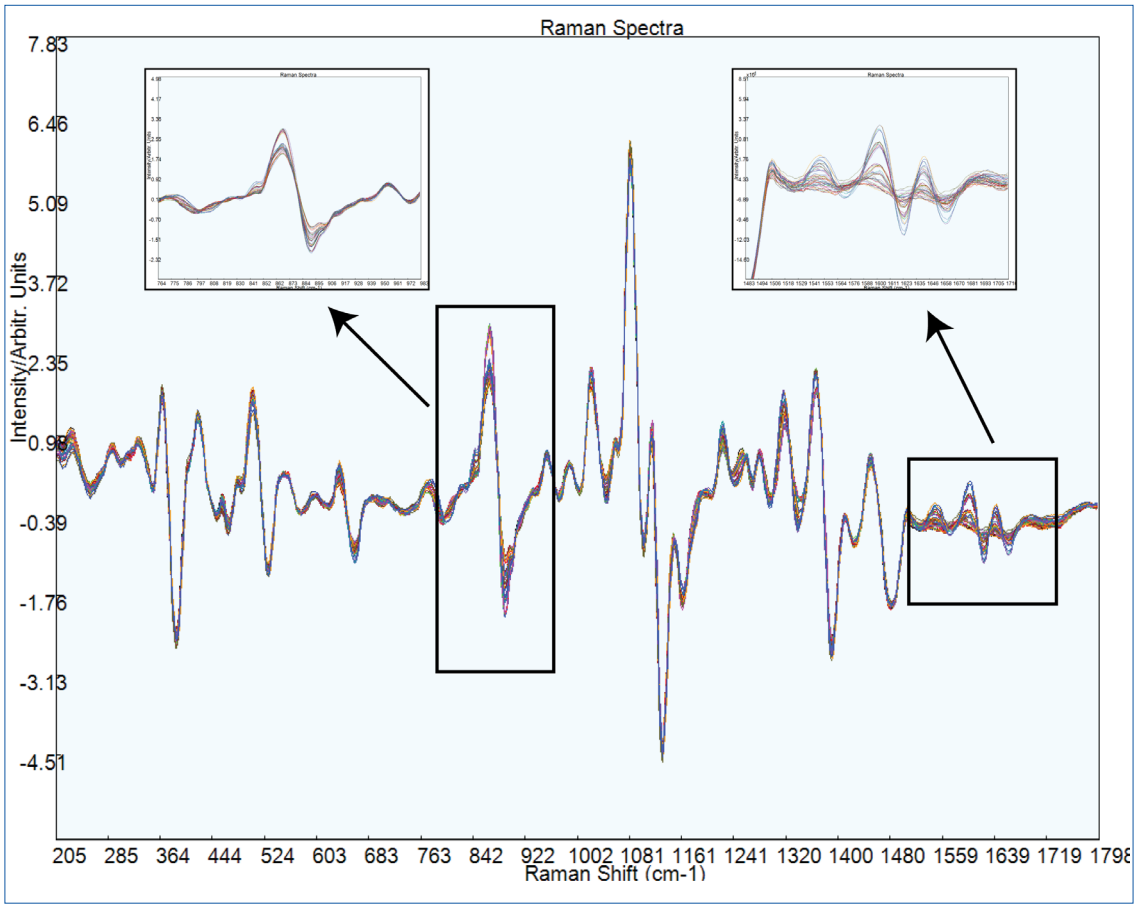

Figure 8: Spectra processed with S-G 1st derivative, manual region selection, and SNV. Spectra show clear changes in intensity with increasing acetaminophen concentration at ~860 cm^{-1} and 1500-1650 cm^{-1}, consistent with Raman peaks of acetaminophen.

In BWIQ, a mean center is applied as a separate step (in Vision software this is done automatically). When a mean center is applied, the spectra are centered around the zero line. Figure 9 shows the dataset with all of the preprocessing algorithms applied including the mean center. The processed data are now suitable to build a robust PLS model.

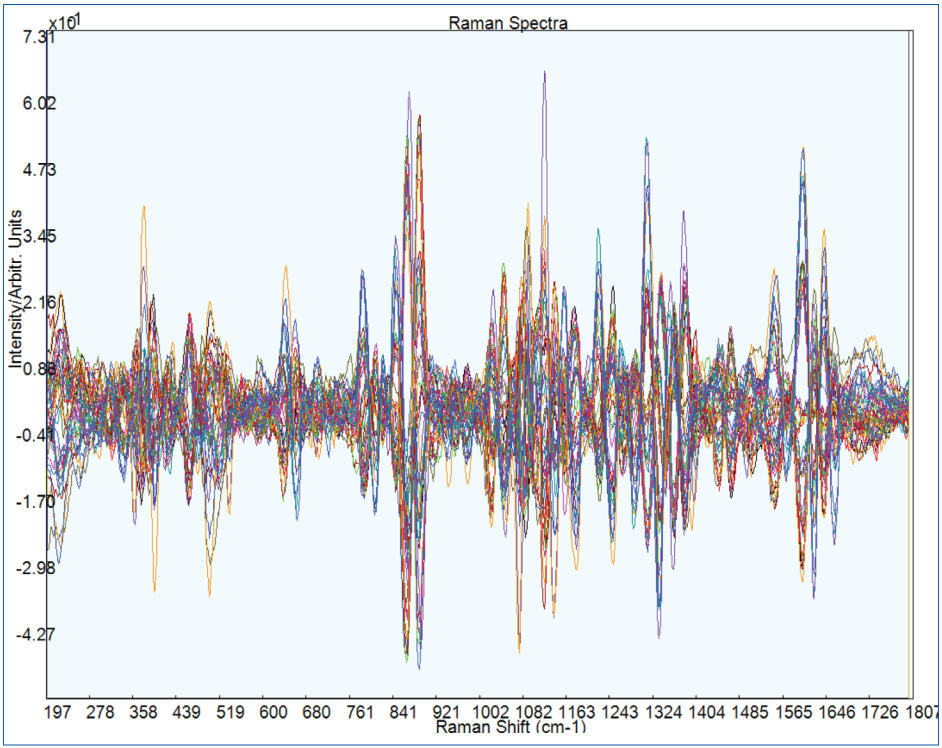

Figure 9: Preprocessed data set including mean center

Table 1 shows the preprocessing steps used in the model, along with their purpose. By applying these preprocessing steps, a robust model can typically be made even by those still new to Raman spectroscopy and without a keen sense of chemometrics.

Table 1

Preprocessing algorithm applied	Purpose
Savitzky-Golay 1st derivative (3rd order, w=25)	Remove fluorescence background
Manual region selection	Isolate spectral region of interest and remove uninformative signal
Standard normal variate	Normalize spectra against intensity fluctuation
Mean center	Remove mean from data set

References

1. J. Huang, S. Romero-Torres and M. Moshgbar. American Pharmaceutical Review. 13, 116-127 (2010). Link
2. J.M. Shaver. Chemometrics for Raman Spectroscopy. In Handbook of Raman Spectroscopy: From the Research Laboratory to the Process Line; I.R. Lewis, H.G.M. Edwards, Eds.; Marcel Dekker, Inc.: New York, 2001; Vol. 28, pp 275-306.
3. M.J. Pelletier. Appl. Spectroscopy. 57, 20A-42A. (2003) https://doi.org/10.1366/000370203321165133
4. Vision software user manual
5. BWIQ software user manual
6. QTRam for Content Uniformity Analysis- A Simple Demonstration. Internal B&W Tek reference document 400000352-B
7. B&W Tek, LLC (2019). QTRam® for Content Uniformity Analysis of Low-Dose Pharmaceutical Tablets (Application Note 410000046), https://bwtek.com/appnotes/qtram-for-content-uniformity-analysis-of-low-dose-pharmaceutical-tablets/