SPD/TN/106 Issue 1.0 September 2006
Application note from the Spectroscopy Products Division

Chemometric analysis and pre-processing techniques applied to Raman mapping experiments on pharmaceutical samples

Overview

The role of chemometrics in the analysis of Raman spectroscopy data is becoming increasingly important for many different application areas. The main reasons for this are:

- **it allows the qualification and quantification of very complex systems such as biological materials[1]**
- **chemometric algorithms are constantly being developed, so there is a constant supply of new and improved methods**
- **chemometric methods are multivariate and can analyse the whole data set simultaneously. This allows spectral information to be accurately and directly related to chemical properties unlike univariate methods. Fast imaging techniques such as rapid line focus imaging are ideal, as more data is available to improve the model quality[2].**

This technology note details some of the different chemometric analysis methods currently available, why different algorithms are used to gain specific information, and how data pre-processing can aid data interpretation. Finally the application of chemometrics techniques to data from a pharmaceutical tablet is illustrated.

Introduction to chemometrics

What is chemometrics?
Chemometrics is defined by the International Chemometrics Society (ICS) as 'the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods'.
Raman spectroscopy is used to obtain, either directly or indirectly, information relating to the properties of the analyte. This information can take the form of component concentration, local stresses, and sample crystallinity, for example.

Univariate statistical methods are the simplest way of extracting such information from Raman data. Direct Raman spectral information, such as relative band intensities and band positional changes, can be used in a univariate method to evaluate component concentration and local sample stresses. Multivariate techniques use more than one spectral variable to extract information about a spectrum. Perhaps the main advantage of using such a multivariate approach is the lower inherent error of the final information. The whole spectrum can be used to derive the information, rather than a single Raman band for example, so all changes within the spectrum are considered, including those not immediately obvious.

Chemometrics is a group of multivariate data analysis methods. It allows a large number of Raman spectra to be analysed simultaneously and useful trends to be extracted from complex variables. To achieve this, the spectra are placed in the rows of a matrix (**X**), the columns of this matrix representing the intensity change at a specific spectral wavenumber.

Different chemometric methods
Grouping of chemometric methods can be based on the information derived from the method. Initial grouping is broadly based on whether the derived information is qualitative or quantitative.
Typical qualitative methods used are cluster analysis and classification. In these methods, spectra are assigned to groups based on either prior knowledge (supervised pattern recognition) or due to their similarity (unsupervised pattern recognition). These methods therefore allow qualitative answers to questions such as "Is the sample A or B?" This question has a whole multitude of different applications. "Is the tissue cancerous or not?", "Is my polymorph monohydrate or dihydrate?"

Quantitative methods are used to build calibration sets from the spectral data. Real quantitative information can be gained from this such as: "What is the concentration of compound A and B in my sample?" Typical quantitative methods are multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS). In such methods a calibration model is constructed from a set of spectra of known concentrations. A second known set is used to validate this model and allow unknown concentrations to be predicted.

What is PCA?

Perhaps the most commonly known chemometric method is principal component analysis (PCA). This method can be used directly to gain qualitative information, or as a generic pre-processing step. Quantitative information can only be inferred indirectly from the data. PCA rotates the matrix onto a new co-ordinate system, $\mathbf{X}_{\mathbf{new}}$**,** where maximum variance is explained using orthogonal axes for each principle component.

$$X_{new} = T*V + E$$

This co-ordinate system is decomposed into two sets of matrices that are abstract representations:

- Scores (T) - concentration-like
- Loadings (V) - spectrum-like

PCA is therefore used to show data variance, but the scores and loading do not necessarily directly represent chemical information.

Multivariate curve resolution methods

A secondary group of chemometric methods is used to gain useful chemical information. It includes multivariate curve resolution (MCR) techniques. MCR methods allow principal component information to be rotated into physically meaningful components.

X is therefore decomposed into concentration profiles (C) and (pure) spectra (S), rather than into abstract vectors:

$$X = C*S^T + E$$

Different methods are used to obtain C and S. This is important as the quality of information will vary depending on:

- purity of spectra in the dataset
- specific spectral forms of the components

Renishaw's Empty Modelling™ is a form of MCR using an alternating least squares algorithm (MCR-ALS) to solve the above equation[3]. Component spectra and concentration profiles are iteratively resolved in this technique.

Orthogonal projection analysis (OPA) is used to gain concentration and spectral estimates for each principle component. MCR-ALS is then performed on the data to gain the desired information.

Direct classical least squares (DCLS) analysis is similar to the curve resolution techniques in that **X** is related to useful concentration information (C). Here, reference spectra are required to calculate concentrations using a least square fit.

MCR methods allow pseudo-quantitative information to be derived (concentration profiles) as the spectra are seen as linear combinations of pure component spectra. Real, and highly accurate, concentration values can only be gained through use of calibration systems. Curve resolution techniques are therefore an ideal method for analysing unknown mixtures, as they require no calibration information, and in several cases no prior-knowledge of components.

Which methods should I use on my mapping dataset? Many multivariate methods for gaining similar information are used, depending on the exact information desired, and the type of data that is present. The main MCR considerations are:
OPA spectral estimates - pure spectra required
OPA concentration estimates - pure bands but not necessarily pure spectra required
Empty modelling™ - cases where one main component exists
The DCLS method is ideally used in cases where reference spectra are available for all chemical components comprising the mixed dataset.

A calibration-driven algorithm, such as PLS, is preferred where highly accurate concentration and other quantitative information is desired. A large known sample set is required for the model to be created and tested appropriately.

Data pre-processing

Data pre-processing prior to deriving quantitative or qualitative information is a useful and often necessary step when considering high quantity Raman mapping data. The different pre-processing steps can be defined as follows:

- Cosmic ray removal (nearest neighbour method)
- Noise filtering
- Normalisation

Cosmic rays (CR) are high-energy particles originating beyond the Earth. They randomly impact upon the CCD detector of the Raman instrument. When this occurs the spectrum will have an additional unwanted sharp feature unrelated to the Raman information.
CR features are usually removed by performing additional spectral accumulations, subsequently taking the median of each recorded pixel value (spectral data point). Such methods effectively guarantee CR free data, however they require additional time. If cosmic ray features are not removed they can influence the chemometric model, adversely affecting its validity.

An effective way to remove CR features from a Raman mapping data set, whilst still maintaining Raman band integrity, is to use the nearest neighbour comparison method.

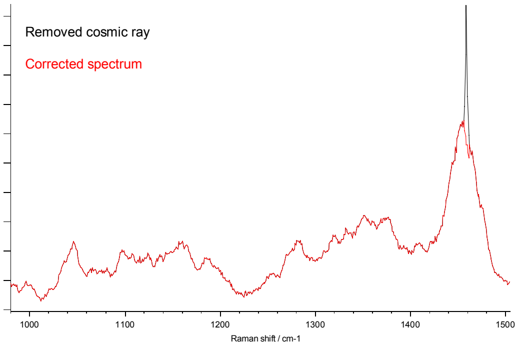

Figure 1
Example of cosmic ray removal pre-processing on a complex Raman band system

The nearest neighbour method compares the intensity at the same wavenumber with neighbouring spectra. The spectra from which the neighbours can originate are collected separately (at different times) and must represent similar component locations (i.e. mapping data must be spatially over-sampled). Using this method, the most similar CR feature-free nearest neighbour is chosen to identify CR signals in the original spectrum. Data normalisation allows the CR intensity values to be replaced by those of the most similar nearest neighbour. This ensures that band structure integrity is maintained.

Noise filtering uses PCA reconstruction to remove data variance not attributable to real or significant data (i.e. noise).

Data filtered in this way has significantly lower noise and further chemometric processing benefits as a result. The improvement will depend on the original data quality and the data set size. Low signal-to-noise data within a very large dataset will benefit from significant improvements to the noise level.

This is often the case for pharmaceutical tablet mapping.

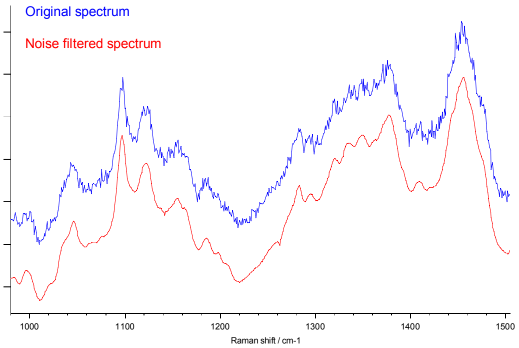

Figure 2
Example of noise filtering using PCA

Normalisation is a necessary step due to the local variations and perturbations that can result from the Raman experiment. The exact normalisation technique used will largely depend on the type of chemometric analysis to be performed. Perhaps of most significance is the variation in laser focus on the sample. In some cases changes in focus of < 1 µm can significantly change the intensity but not the form of the spectrum. The aim of this step is to remove intensity variations not arising from compositional changes in the sample.

Chemometric analysis example on pharmaceutical tablet

Tablet mapping is an established application area for Raman spectroscopy, where active pharmaceutical ingredients (API's) and excipients can be identified and differentiated. The recent addition of Renishaw's rapid line focus imaging technique enables larger areas and more data to be collected from tablets at high spectral and spatial resolution. Quality Raman data representative of the tablet component contents and their distribution can now be collected on a realistic timescale.

The following example shows how a multivariate curve resolution technique can be used to extract valuable chemical information from a multi-component pharmaceutical tablet.

The distribution of an API within a formulation is thought to have a crucial effect on the bio-effectiveness of the drug, primarily, by modifying the release rate. In addition to this, the ability to identify very low concentration components is becoming increasingly important for the following reasons:

- high potency API's, where the natural concentration will be very low within the formulation.
- to verify the presence or absence of unwanted, often very low concentration, polymorphs of the API.

API and excipient distribution using Empty Modelling™ multivariate curve resolution (EM-MCR)

Rapid line focus imaging experiments were performed on a sectioned Piriton® allergy tablet. The tablet contains a mixture of standard excipients and an API (chlorphenamine maleate). The tablet was sectioned using a mechanical cutting technique to ensure sample flatness.

Raman spectra from a map area 0.6 mm x 0.5 mm were collected using an inherent spatial resolution of 6 µm (generating 12,000 Raman spectra). The collected spectra were pre-processed using the cosmic ray removal, noise filtering and normalisation techniques previously described.

The EM-MCR method was then applied to the data using MATLAB (The Mathworks Inc).

For the Piriton® tablet, images representative of map spectra containing specific spectral components are created. This method does not require spectral or component information prior to performing the analysis.

Each image represents the distribution of a single chemical component contained within the tablet. An individual image is shown in Figure 3.

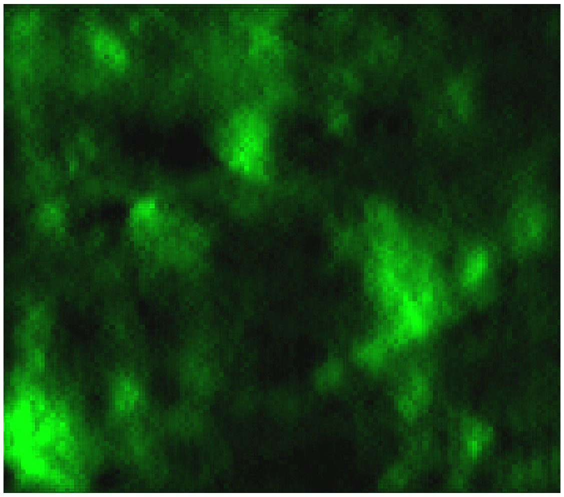

Figure 3

Lactose monohydrate component tablet image

The identity of this component is confirmed through comparison with a reference spectrum of lactose monohydrate (Figure 4). Good agreement between the derived component spectrum and the reference spectrum confirms the effectiveness of the EM-MCR technique.

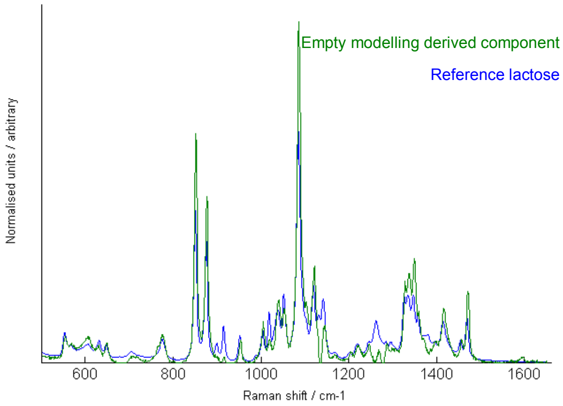

Figure 4
Spectral comparison of a single Empty Modelling™ derived component with that of reference lactose monohydrate

Figure 5 shows overlayed Raman images (red - API, green - lactose, blue - maize starch) derived from the Empty modelling™ principle components.

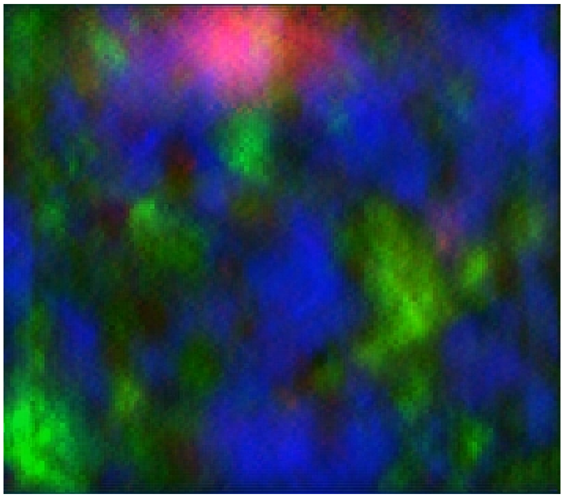

Figure 5
Overlayed multi-component tablet image

The look up table (LUT) for each image gives information on the approximate proportion of each spectral component to each sample location. However, as a result of the normalisation processes required, it is extremely difficult to gain accurate concentration information without using calibration sets. Nevertheless, EM-MCR allows detailed chemical images to be created where no reference information is available.

Conclusion

Chemometrics is a broad field comprising a huge array of different algorithmic methods. The most suitable approach for a given application depends upon the details of the constituent component spectra and the purity with which they occur. The quantity of Raman data and speed at which this can be collected, when using rapid line focus imaging, has made the application of multivariate data analysis techniques, such as chemometrics, an invaluable and powerful tool.

References

1. Lopez-Diez, E.C, Goodacre, R. *Anal. Chem.*, **2004**, *76 (3)*, p585-591.
2. Raman imaging. *Technology note (TN091)* Renishaw plc, UK.
3. UK Patent Application No. GB 0611981.2
4. The Mathworks, Inc. MA, USA.

Renishaw is continually improving its products and reserves the right to change specifications without notice.