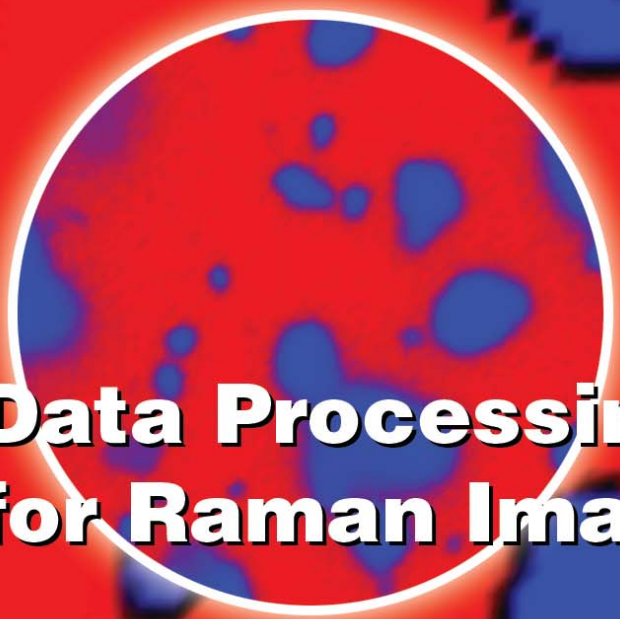
Global Edition

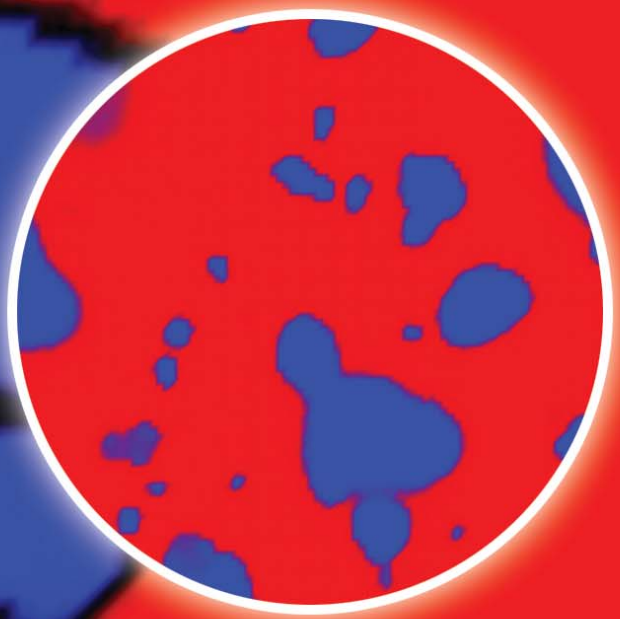
Spectroscopy®

Solutions for Materials Analysis



Data Processing for Raman Imaging

FT-IR for Analysis of Bacteria in Food



August 2007

www.spectroscopyonline.com

AN ADVANSTAR PUBLICATION

Multivariate Data Processing of Spectral Images: The Ugly, the Bad, and the True

The results of various multivariate data-processing methods of Raman maps recorded with a dispersive Raman microscope are reported. Criteria for "good" data-processing strategies are suggested.

Eunah Lee, Fran Adar, and Andrew Whitley

The latest trend of hybrid instrumentation — combining two or more techniques in one instrument for enhanced material characterization — reflects the demand for "total characterization" in research by academia and industry alike. Among these hybrid techniques, hyperspectral imaging — combining spectroscopy and microscopy — has gathered a rapidly growing following. The power of simultaneous molecular and morphological analysis providing the complete picture of the sample is highly attractive.

What enhances the power of the hyperspectral imaging analysis is the vast amount and high information content of the data that hyperspectral imaging generates. Subtle material inhomogeneities and differences only observable through spectroscopy can be found and highlighted when the data are mapped spatially and compared. It is important to note that these characteristics set hyperspectral imaging as a field on its own merit, instead of an extension or expansion of spectroscopy or microscopy.

To condense the vast amount of hyperspectral imaging data into scientifically meaningful results, or chemical images, many mathematical algorithms are being used. Because of its highly versatile nature, a single algorithm or a fixed procedure does not work for all data sets. For example, a typical data processing procedure for near-infrared (NIR) hyperspectral imaging is quite different from that for Raman hyperspectral imaging.

There is a scientific foundation underlying the variety of method developments to differentiate hyperspectral imaging as science from art. In this article, using Raman hyperspectral imaging data as an example, the method development of processing and interpreting hyperspectral imaging data will be demonstrated for qualitative analysis. The importance of verification and validation in interpreting chemical images is highlighted.

Experimental

A few pharmaceutical crystals were spread on a microscope slide. Raman maps were recorded with a dispersive Raman microscope (LabRAM HORIBA Jobin Yvon, Edison, New Jersey). Data processing was performed using LabSPEC 5 (HORIBA Jobin Yvon) and ISys 4 (Malvern Instruments).

The instrument conditions were: 500 μm wide confocal hole, 150 μm wide entrance slit, 600 gr/mm grating, and 100× objective lens (NA = 0.9). Various exposure times (100, 300, and 900 ms) were used to achieve a range of signal-to-noise ratios (S/N). A mechanical mapping stage (Marzhauzer) and a HeNe laser (633 nm) were used. The spectral range measured was 377 − 1540 cm^{-1}, and the mapping area 30.6 × 30 μm^2 was generated with 0.3 μm step (103 × 101 points).

Data were pretreated to remove the cosmic spikes, normalize to unit variance, and offset minimum intensity to zero. Two algorithms — K Harmonic Mean Clustering and Score segregation — were applied and the classification quality of the results compared.

Results and Discussion

A Raman map is composed of hundreds to thousands of spectra, and manual examination of every spectrum is impossible. Therefore, the first step of qualitative analysis of a Raman map is to achieve "good" classification. Good classification separates all spectra into groups so the spectra within a group are similar to each other, while spectra from different groups are different from each other, significantly and systematically. Once the good classification is achieved, representative spectra of each group can be examined and identified. And then the characteristics of spatial distribution of each chemical component can be investigated.

The typical data processing strategy of a Raman map is pretreatment, univariate or

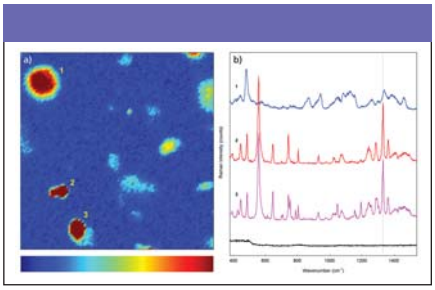


Figure 1: (a) An intensity map at 1331 cm^{-1} with the color scale bar from the raw data. (b) Single point spectra from the top (area 1, blue spectrum), the middle (area 2, red spectrum), and the bottom (area 3, pink spectrum) red areas. Note that red and pink spectra are similar (of the same material), while the blue spectrum is quite different (of a different material). Black spectrum is from the substrate, blue areas in the image. The dotted line indicates the spectral position of the intensity map.

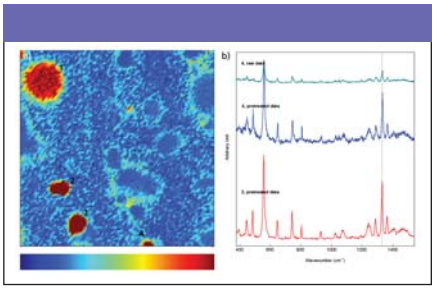


Figure 2: (a) An intensity map at 1331 cm^{-1} with the color scale bar from the re-treated data. (b) Single point spectra from the middle (area 2, red spectrum) and the new (area 4, blue spectrum) red areas. Teal spectrum is the same spectrum as the blue one, scaled with respect to the red spectrum, reflecting the relative overall intensity in the raw data.

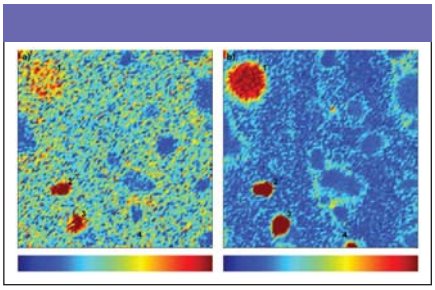


Figure 3: Comparison between low and high S/N data. (a) An intensity map at 1331 cm^{-1} of a Raman map measured with 100 ms/spectrum accumulation time and at low S/N. (b) An intensity map at 1331 cm^{-1} of a Raman map measured from the same sample under the identical conditions except for the accumulation time of 900 ms/spectrum, resulting in a high S/N.

multivariate analysis, verification with the original (either raw or pretreated) data, image processing, and extracting information.

The pretreatment procedure is as important as, if not more important than, the data processing itself. Raw count intensity of a spectrum originates from many sources: Raman scattering, as well as cosmic ray events, dark current and random noise, and fluorescence, all of which contribute to the baseline and other interferences. The goal of the pretreatment procedure is to separate these baseline and other interferences from Raman scattering and eliminate them. When successful, the pretreatment procedure can accomplish adequate classification.

Using the data set from the experiment described earlier, an intensity map (Figure 1a) at 1331 cm^{-1} was created using the raw data measured with 900 ms/spectrum acquisition time. The map is in a pseudo-color scale ranging from red (high intensity) to blue (low intensity). There are three discrete red areas (noted as 1, 2, and 3) indicating significantly higher intensity areas than surrounding areas, which might imply that they are of the same material. Single point spectra (Figure 1b) from each area indicate, however, that there are two different spectral species, and therefore, two different chemical compounds.

Figure 2a is the intensity map at the same spectral position, created using the pretreated data. The area 1 (top left corner) is still colored in red but in reduced intensity compared to the image from the raw data (Figure 1a). On the other hand, a new area (noted as 4) has appeared, colored in red. The spectrum from area 4 is similar to those from areas 2 and 3 (Figure 2b). The "classification quality" of the pretreated data is better than that of the raw data, even though it has not reached the satisfactory level, because it failed to exclude area 1 completely. The reason for the difference in classification quality in this particular case is that the overall intensities of the spectra within area 4 are weak, capturing probably an edge of a crystal rather than the bulk of it. Other reasons include baseline levels, fluorescence levels, and random noise levels.

It is worth noting that the intensity map of the pretreated data appears "ugly" even though it is closer to the scientific truth than that of the raw data.

Data processing methods can be categorized into two groups — univariate and multivariate analysis methods. Univariate analysis methods monitor one variable per spectrum (such as the intensity at a certain spectral position). The intensity maps shown in Figures 1 through 3 are results of univariate analysis. In this case, monitoring the intensity maximum or the integrated intensity (the sum of the intensity from baseline to the height and back to the baseline) of a characteristic band belongs to the univariate method. Other variables often monitored include peak positions and bandwidths.

Univariate analysis is intuitive to apply and straightforward to interpret. No method development is required, so real-time monitoring of various spectral bands is possible. On the other hand, for good classification, at least one spectral species for each component within the Raman map must be unique to that species so that a marker band for each component can be found.

In addition, the impact of the S/N level is relatively high. Figure 3 shows intensity maps at 1331 cm^{-1} from two Raman maps measured from the same sample under the identical conditions except for the accumulation times to achieve different S/N in the spectra, low and high, relatively and respectively. Pretreated data were used.

Note that area 4 is absent in the low S/N data while present in the high S/N data. Also note that, in this case, the "pretty" image (high S/N, high contrast) is closer to the scientific truth, compared to the "ugly" image (low S/N, low contrast).

In contrast, multivariate analysis takes the entire spectrum into account instead of a single value per spectrum. Its goal is to recognize and isolate unique chemical species from the entire spectral data set. In chemometrics, the resulting pseudospectra are called *loadings*. In clustering, they are called

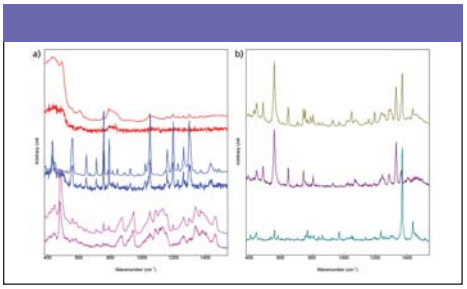


Figure 5: KHMC results verification. (a) Centroids 1, 2, and 5 in comparison with single point spectra from high scores areas, respectively. (b) Centroid 4 in comparison with single point spectra from two discrete high scores areas.

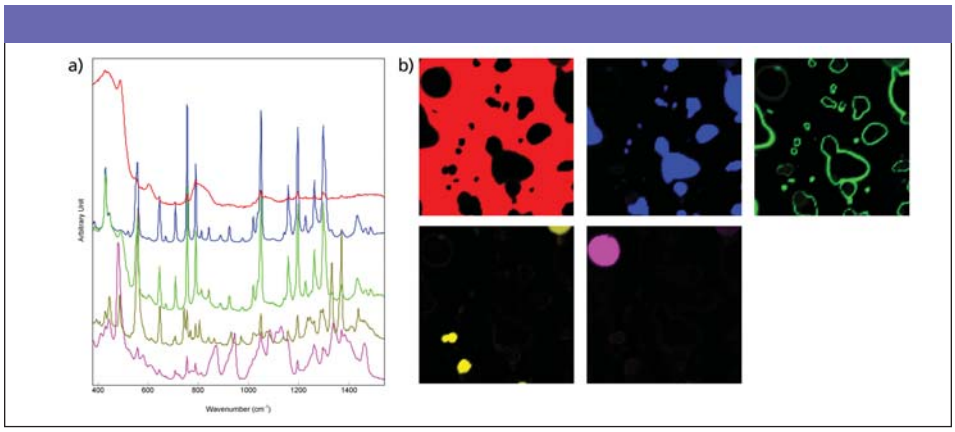


Figure 4: KHMC analysis results. (a) Centroids. (b) (Membership) Scores images for each centroid. Centroids and scores images are color-coded.

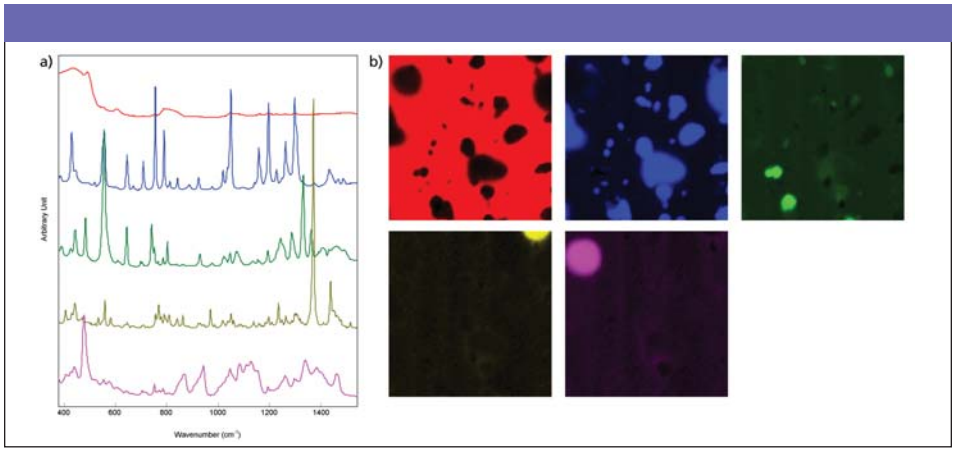


Figure 6: Score segregation analysis results. (a) Loadings. (b) Scores images for each loading. Loadings and scores images are color-coded.

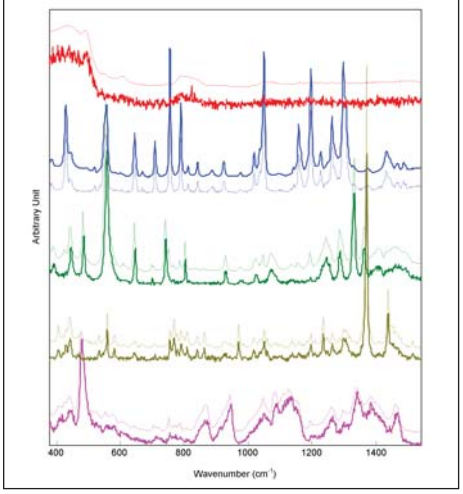


Figure 7: Score segregation results verification. Single point spectra from high score areas are compared to the corresponding loadings. They show very good agreement.

centroids. As soon as loadings or centroids are determined, the spectral contributions of each loading or centroid to all individual spectra within the Raman map are calculated. These contributions are called *scores*. In a spectroscopist's terminology, a high score indicates high similarity or contribution of the loading to the spectrum. In a chemist's terminology, a high score might imply a high concentration of the chemical component that the loading represents. A scores image shows the spatial distribution of the loadings. Successful multivariate analysis isolates all unique spectral species without duplicates or missing components as loadings or centroids, and scores images highlight areas whose spectra contain high distribution of the designated loading or centroid.

There are many multivariate analysis algorithms available. In this article, two different methods — K Harmonic Means Clustering (KHMC) and Score Segregation (SS) — were selected for comparison.

K-Harmonic Means Clustering (KHMC) (1)

Clustering is designed to group data into "clusters" whose spectra are similar to each other. The representative spectrum of a cluster is the centroid, meaning the center point. K-Harmonic Means Clustering calculates the distance between the centroid and individual spectra using harmonic averages. This distance is what determines if a spectrum belongs to one cluster or the next. KHMC was performed on the pretreated data (900 ms/spectrum acquisition time) using five centroids. The iteration was repeated until the centroid error between one iteration and next reached the threshold of 10^{-3}. Figure 4 shows the results.

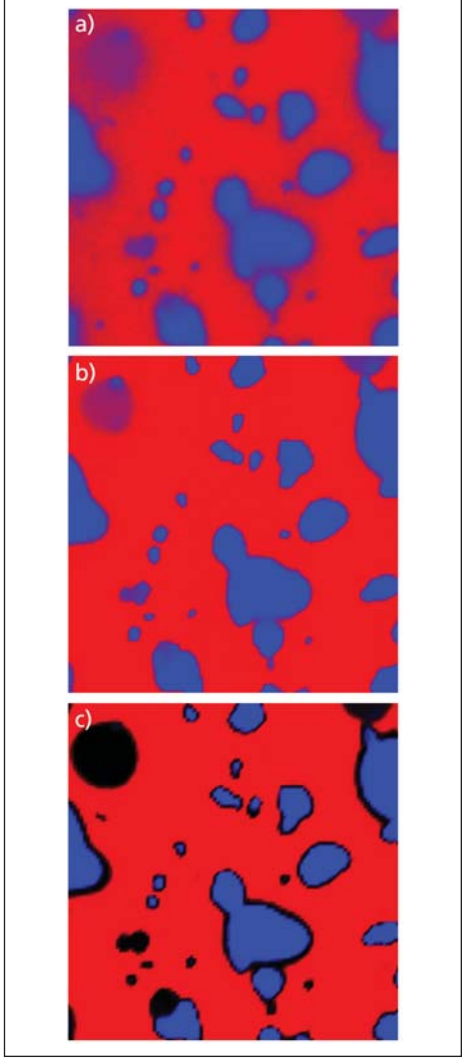


Figure 8: KHMC analysis results of (a) low S/N data, (b) medium S/N data, and (c) high S/N data. Images are created from two scores images each. Black areas represent where scores for both images are low.

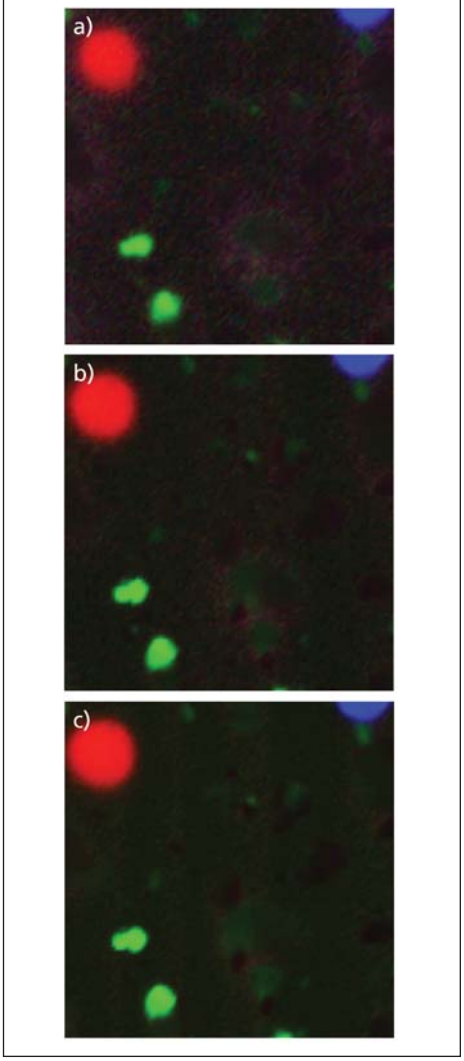


Figure 9: SS analysis results of (a) low S/N data, (b) medium S/N data, and (c) high S/N data. Images are created from three scores images each. Black areas represent where scores for all images are low.

Centroids 1 (red), 2 (blue), and 5 (pink) show unique spectral species. Their scores images highlight their spatial distribution. These can be verified by comparing the centroids to single point spectra from high scores areas (Figure 5a). Centroid 3 (green) is a duplicate of centroid 2. Scores image 3 highlights the edges of areas that were highlighted in scores image 2. Centroid 4 (dark yellow) shows mixed spectral features; some are duplicates of the second centroid and some unique. When compared to single point spectra (Figure 5b), it was determined that centroid 4 represents two spectral species.

Although KHMC successfully separates three unique spectral species and generates "pretty" images for all centroids, it failed to separate the other two species. This was not discovered until the verification step, where the original spectra were compared to centroids. As it is shown here, the importance of verification of multivariate analysis with original spectra or alternative analysis cannot be stressed enough. No matter how sophisticated the mathematics, the process is numeric, rather than analytical. Without the verification step, pretty images alone cannot be scientifically meaningful.

Score Segregation (SS) (2)

Score segregation is one of the factor analysis methods. The first step of a factor analysis is to perform Principal Component Analysis, in which orthogonal loadings, or principal components, are extracted. In laymen's terms, principal components represent the spectral

features that show the most variance. They are often mixed, meaning that each principal component most likely contains spectral features from more than one pure component spectrum. Score segregation and other factor analysis methods aim to purify these principal components to yield loadings that are ideally, pure component spectra. The results of SS analysis are shown in Figure 6.

Comparison to single-point spectra determined that all five loadings from SS represent unique spectral species, respectively (Figure 7).

Impact of S/N on Multivariate Analysis

Three Raman maps measured from the same sample under the identical conditions except for the acquisition times (100, 300, and 900 ms/spectrum) to achieve different S/N (low, medium, and high, relatively and respectively) were processed in the identical manner.

Two scores images from KHMC analysis results and three from SS analysis results were combined to create Red-Blue (RB) and Red-Green Blue (RGB) images (Figure 8 and Figure 9, respectively). These images were created by assigning a scores image to red, green, or blue color scale, in which high scores areas were colored in the assigned color and low scores areas in black. Black areas in the combined image, therefore, represent the areas where scores are low for all scores images. In other words, if a scores image represents a chemical component, black areas indicate the absence of chemical components that any of the participating scores images represent. This could mean the presence of unaccounted-for components in the image or no components at all.

For the low and medium S/N data, KHMC can classify two components only partially. For the high S/N data, KHMC successfully classifies three components and generates one duplicate and one mixture. Figure 8 shows the RB images of (membership) two corresponding scores images. Note that high S/N data show high contrast.

For all three data sets, SS successfully classified all five components with classification quality better than the KHMC earlier. RGB images of three corresponding scores images are shown in Figure 9. Again, higher S/N data show higher contrast in the image.

Conclusion

These results demonstrate a number of things. First, and most importantly, the quality of the image (for example, the contrast) does not correlate to the chemical authenticity of the image; verification with original spectra alone can determine that. Second, the proper pretreatment procedure is very important in achieving the chemically correct results. Third, higher S/N spectra yield higher success rate on a wide variety of analysis methods than lower S/N spectra. Therefore, one should always obtain data with as high S/N as time allows. Fourth, for these particular data sets, KHMC requires higher S/N in the raw data than SS to be at least partially successful in classification.

References

(1) *ISys 4.0 Chemical Imaging Software User's Manual*, 76–78 (2006),

(2) *ISys 4.0 Chemical Imaging Software User's Manual*, 96–99 (2006).

Eunah Lee, Fran Adar, and **Andrew Whitley** are with HORIBA Jobin Yvon, Edison, New Jersey.■