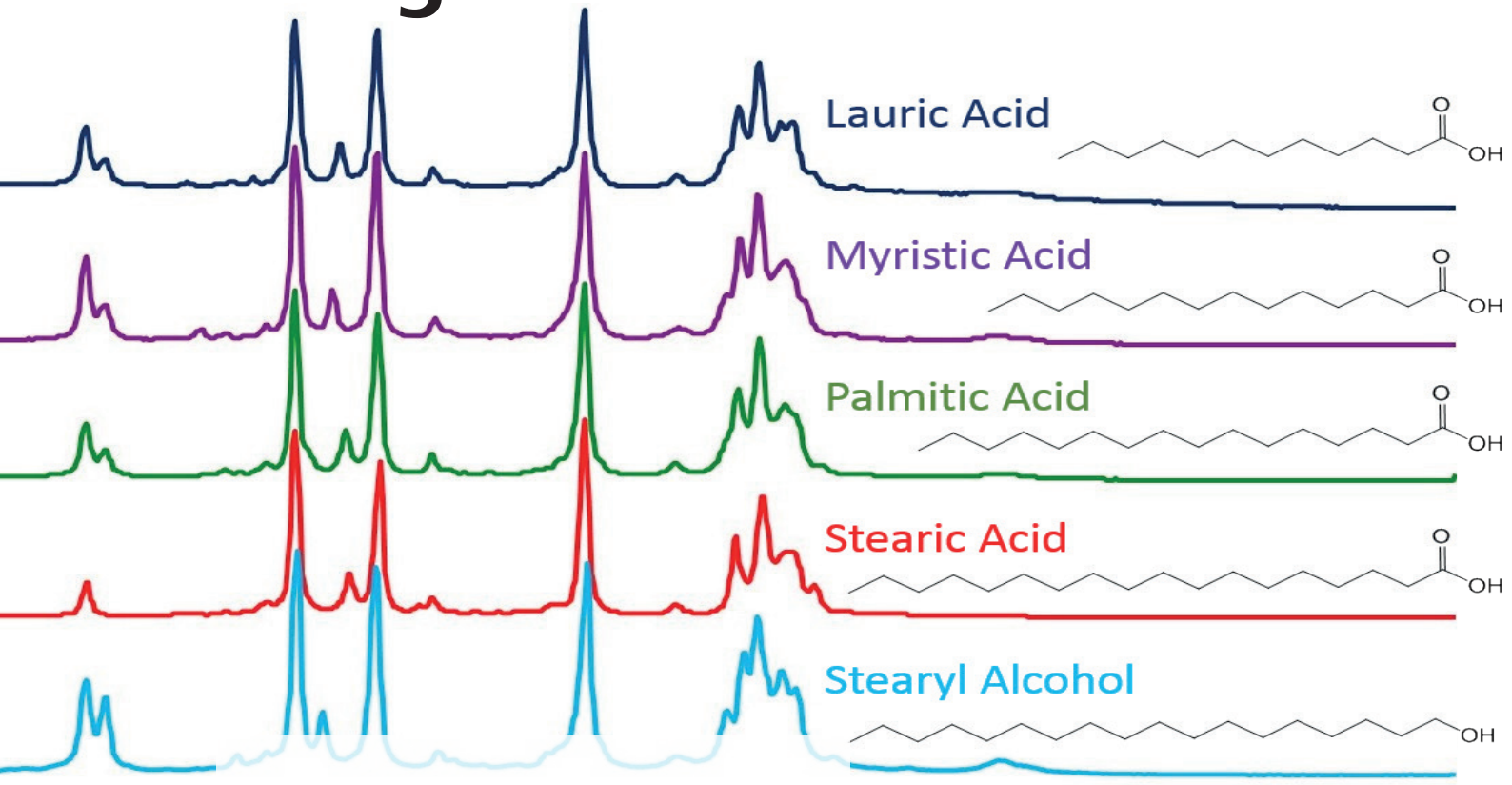
Verification, p-values, and Training Sets for the Mira P



Melissa Gelwicks

This white paper differentiates between methods for identification of unknowns and verification of known materials. The goal of this publication is, ultimately, to inform the user of the capabilities of the handheld Metrohm Raman Mira P system. Best practices for building robust training sets for materials verification with Mira P can also be found here.

Metrohm International Headquarters; Herisau, Switzerland

Metrohm Raman

Introduction

The Metrohm Instant Raman Analyzer Pharmaceutical (Mira P) is a handheld Raman spectrometer designed for rapid, nondestructive identification and verification of chemicals, materials, and pharmaceuticals. Raman spectroscopy is an established technique for identification of unknowns by comparison of sample spectra within reference libraries; however the Mira P is uniquely capable of material verification within the Metrohm Raman product line. This white paper describes how statistical analyses relate to experimental design and how both can help the user create robust models for verification.

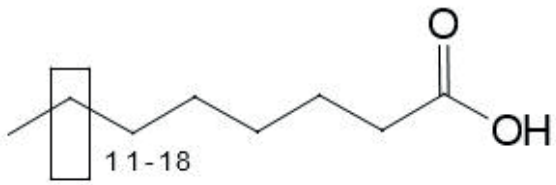


Figure 1. Structure of fatty acids sampled below

HQI for Identification of Unknowns

For *identification* with the Mira P, a Pearson correlation technique generates a Hit Quality Index (HQI) or r^2, which is a measure of spectral similarity between unknown and library spectra. The displayed match score varies between 0-1, where larger values indicating greater spectral similarity. The instrument generates a list of compounds with HQI scores above a specified threshold, usually 0.85. This method of identification is a) easy to implement, b) fast, and c) suitable for use with extensive chemical libraries.

While this method is widely used and reliable in some applications, it does not adequately account for very minor differences between similar molecules in Raman spectra. For instance, Figure 2 illustrates the spectra of a family of four fatty acids and a similar alcohol, which differ primarily in the length of the saturated carbon chain, Figure 1. The spectral similarity is undeniable, and it reflects the similarity of the compounds.

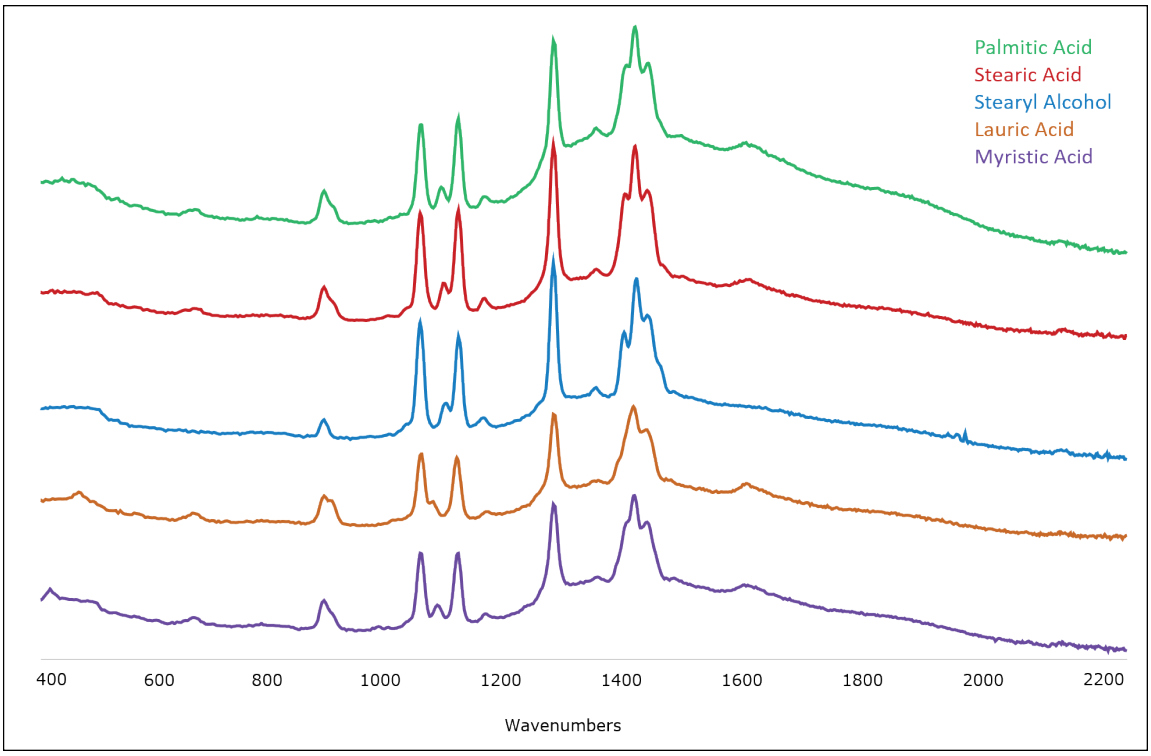


Figure 2. Raman spectra of very similar fatty acids/alcohol

HQI Values

Because HQI is a measurement of correlation between reference and sample spectra, it can result in misidentification of very similar materials. In other words, false positives can be a result of analysis with HQI. The values in Table 1 reflect the similarity of the molecules and their spectra shown in Figures 1 and 2.

Table 1. HQI match scores for fatty acid family

Library Match (HQI)					
	Lauric Acid	Myristic Acid	Palmitic Acid	Stearic Acid	Stearyl Alcohol
Lauric Acid	1.00	0.98	0.95	0.95	0.88
Myristic Acid	0.98	1.00	0.98	0.96	0.91
Palmitic Acid	0.96	0.98	1.00	0.98	0.94
Stearic Acid	0.94	0.96	0.97	1.00	0.96
Stearyl Alcohol	0.87	0.91	0.93	0.97	1.00

When each compound in this family is compared to the others, the reported HQI values are all above the assigned threshold of 0.85. As a result, there is poor differentiation between these materials.

Verification of Samples with p-values

A verification method should successfully address this issue. Unlike identification techniques based on similarities between spectra, the verification method reflects spectral differences. This method is based on Principal Component Analysis (PCA), a statistical analysis that reduces a complex data set down to basic features that best describe the data, its "principle components."[2] This method transforms highly correlated spectra into a small set of orthogonal variables, which can be visualized as scatter or score plots. Thus, the spectra shown in Figure 2 can be modeled in a way that describes variances within and between compounds rather than similarities, Figure 3:

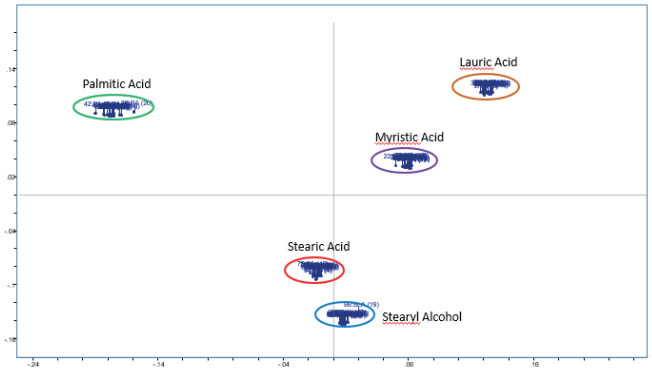


Figure 3. PCA scatter plots depicting fatty acid family

PCA scatter plots, along with a defined confidence interval, become the reference models against which future samples are measured[3]. Each sample spectrum is projected onto the PCA model to see how well it fits into the model limits, which are determined by the confidence interval.

Confidence Intervals

The confidence interval is defined by a Hotelling T^2 ellipsoid, the ovals in the figures above and below, and is a very important designation of how much variance is acceptable within each group.[4] For example, confidence levels of 90 and 95% have been projected onto the plots shown in Figure 4; both are very good representations of the data set, but they differ in the acceptance level of the model. In example A, the 90% confidence level means that fewer samples will be accepted as belonging to the training set, but the model produces greater confidence in the accuracy of the results. A 95% confidence level shown in B illustrates that samples with a greater level of variance (distance from the center=Mahalanobis distance) may be verified as part of the model.

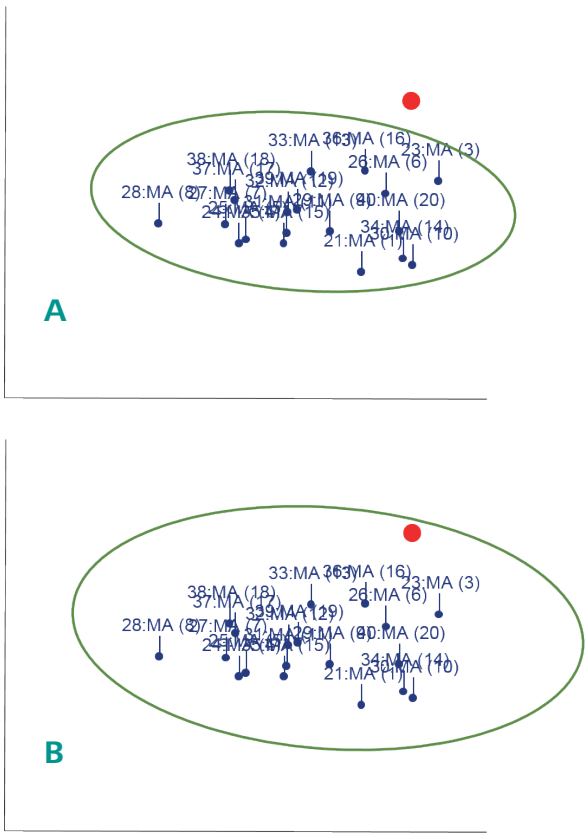


Figure 4. Confidence levels represented by Hotelling T^2 ellipsoids. A= 90%, B= 95%

Hypothesis Testing and p-values

When a sample spectrum is projected onto the model space, the result is a p-value: an indication of how well the sample fits within the model limits at a given confidence level.[4] In other words, the p-value addresses the significance of results when a statistical hypothesis test is performed. For PCA analysis and verification of Raman spectra, the null hypothesis (Ho) reads, "The measured spectrum belongs to the training set used to build the model." A small p-value (<0.05) indicates strong evidence against Ho, and so the null hypothesis is rejected and the sample FAILS to belong to the model. A large p-value results in a PASS, indicating that the sample belongs to the model population, and higher p-values are accepted with greater confidence. PCA and subsequent p-values provide a very different portrait of the fatty acid family, as compared to analysis with HQI. Table 2 indicates infinite distinction between the fatty acids.

Table 2. p-values and validation results for fatty acid family

	Training Sets				
	Lauric Acid	Myristic Acid	Palmitic Acid	Stearic Acid	Stearyl Alcohol
Lauric Acid	PASS 0.127	FAIL 0.00	FAIL 0.00	FAIL 0.00	FAIL 0.00
Myristic Acid	FAIL 0.00	PASS 0.494	FAIL 0.00	FAIL 0.00	FAIL 0.00
Palmitic Acid	FAIL 0.00	FAIL 0.00	PASS 0.331	FAIL 0.00	FAIL 0.00
Stearic Acid	FAIL 0.00	FAIL 0.00	FAIL 0.00	PASS 0.365	FAIL 0.00
Stearyl Alcohol	FAIL 0.00	FAIL 0.00	FAIL 0.00	FAIL 0.00	PASS 0.628

Building Models through Training Sets

The effectiveness of a PCA model depends entirely on the *training set*, which is the library of highly correlated spectra represented in the model population. Using the Mira P Raman spectrometer and MiraCal software, the user builds a training set by collecting a minimum of 20 spectra of a single substance with some allowed variation.

Types of Variance

Variance in the training set is necessary to create a robust model that accurately represents the identity of the sample. Variables can be defined as deterministic, which are the known sources of variation inherent in the identity of the sample or the instrument. Stochastic or probabilistic sources of variance include experimental factors that should be accounted for so that they do not interfere with the accuracy of the model in different circumstances.[6]

Sources of Variance

Deterministic variation must be included in a training set in order to create a representative model. For example, the user might build a training set using samples from multiple sources. If specific instrument acquisition parameters, such as laser power, temperature, integration time, and number of scans, are to be used for experiments the training set must be built using those parameters. These parameters are stored as an Operating Procedure to be used for future measurements to maintain the consistency of the model. However, the training set should be built over several days during which the instrument goes through a number of OFF/ON cycles in order to incorporate instrument variability.

Stochastic variation must be included to create a model that corrects for random variation that might interfere with verification of the sample. These are "field conditions" such as ambient light and temperature, container material, heterogeneity of a sample, perhaps even variation in the thickness of a container. They must be accommodated to create a robust and continuous training set that improves the confidence level of the reported p-value. A recap of possible sources of variation is seen in table 3.

Table 3. Possible types of variations

Types of Variation	Variables
Deterministic	Sample Source/ Producer
	Attachment
	Laser Power
	Laser Temperature
	Integration Time
	Raster ON/OFF
	Number of Scans/ Averages
Stochastic	Ambient Light
	Sample Temperature
	Container Material
	Container Thickness
	Sample Homogeneity
	Contaminants

Editing for Robust Training Sets

As a reminder, a robust training set incorporates some variation between spectra, and it must also inherently represent the unique fingerprint of the material of interest. To ensure both of these qualities, visual inspection of the spectra included in a training set, followed by careful editing can improve-materials verification with the Mira P. Unique spectra that are obviously different than others in the set can be removed, so long as care is taken to leave a healthy number of representative spectra.

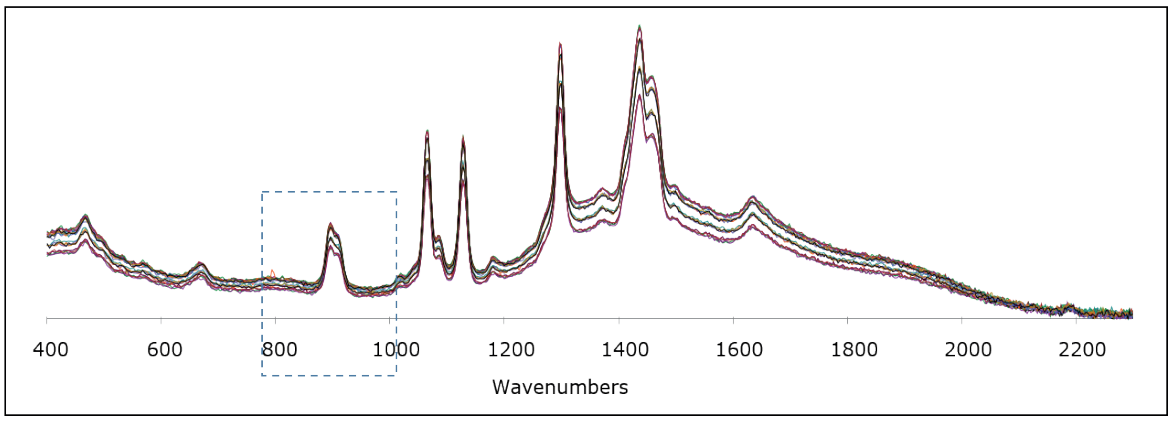


Figure 5. 20 Lauric Acid spectra

As an example, figure 5 is a sampling of the spectra included in the lauric acid training set, which was used to collect the data described in tables 1 and 2. 20 spectra were selected from a total of 60, in order that we might see them better. An acceptable level of variance in intensity (height of the peaks) can be seen, and this represents natural variation encountered during the course of experimentation. If we zoom into the highlighted region of this figure, we can see other examples of how these spectra influence the training set.

The vertical dashed line in Figure 6 demonstrates that peak shift alignment is consistent between all spectra. This is a crucial example of the information built into a PCA model, as the unique peaks in any Raman spectra are the fingerprint that makes Raman such a sensitive verification technique. In contrast, the arrows indicate acceptable variance encountered during sampling, which are retained for a robust training set.

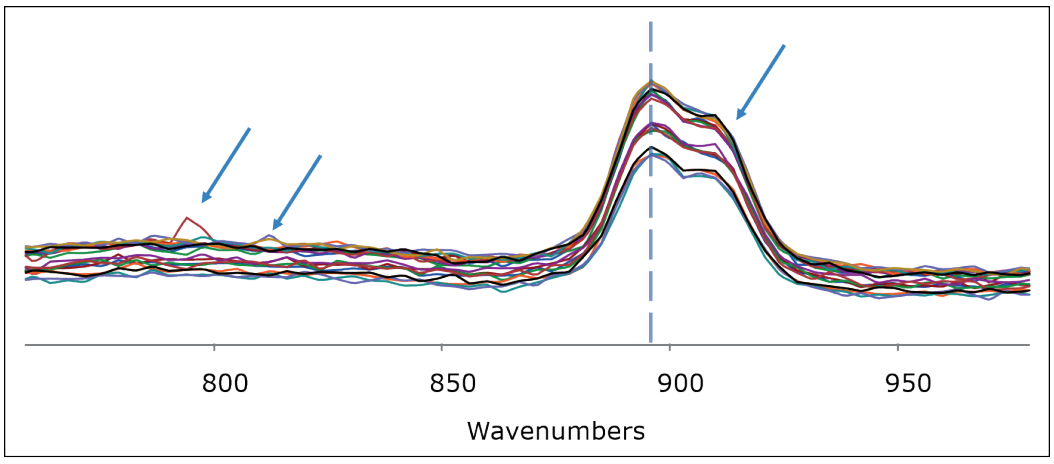


Figure 6. Zoom of Lauric Acid spectra

Conclusion

This paper contrasts identification and verification methods for the Mira P handheld spectrometer, which are distinct analyses for different applications. Identification is used when the identity of a sample is unknown, and verification is used for confirmation of a known sample. Included in this paper are user guidelines for building robust training sets that will optimize the accuracy of the verification method.

References

[1] Bakeev, Katherine A. and Robert V. Chimenti, "Pros and cons of using correlation versus multivariate algorithms for material identification via handheld spectroscopy," Eur Pharm Rev, 2013. https://www.americanpharmaceuticalreview.com/1504-White-Papers-Application-Notes/147135-Pros-and-Cons-of-Using-Correlation-Versus-Multivariate-Algorithms-for-Material-Identification-via-Handheld-Spectroscopy/

[2] Dahiru, Tukur, "p-value, a true test of statistical significance? A cautionary note." Ann of Ibadan Postgrad Med, 2008. (6) 1, pps 21-26.

[3] Varmuza, Kurt and Peter Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*, (2009). CRC Press, Boca Raton, Fl. pps: 321.

[4] O'Connell, Marie-Louise, et al., "Qualitative Analysis Using Raman Spectroscopy and Chemometrics: A Comprehensive Model System for Narcotics Analysis," Applied Spectroscopy, 2010. (64) 10, pps, 1109-1121.

[5] Papoulis, A., & Pillai, U. (2001). *Probability, random variables and stochastic processes.* (4th ed.) New York: McGraw-Hill.

WP-0244EN, published February 2018