How Using Raman Spectroscopy and SIMPLISMA Can Accelerate the Study of Polymorphs: A Case Study Using Carbamazepine

The authors show how a multivariate curve resolution algorithm, called SIMPLe-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA), can facilitate the quantitative and qualitative analysis of difficult samples, and apply the algorithm to a technically challenging Raman spectra series for carbamazepine polymorphs.

Michel R.J. Hachey, Andrey Bogomolov, Keith C. Gordon, and Thomas Rades

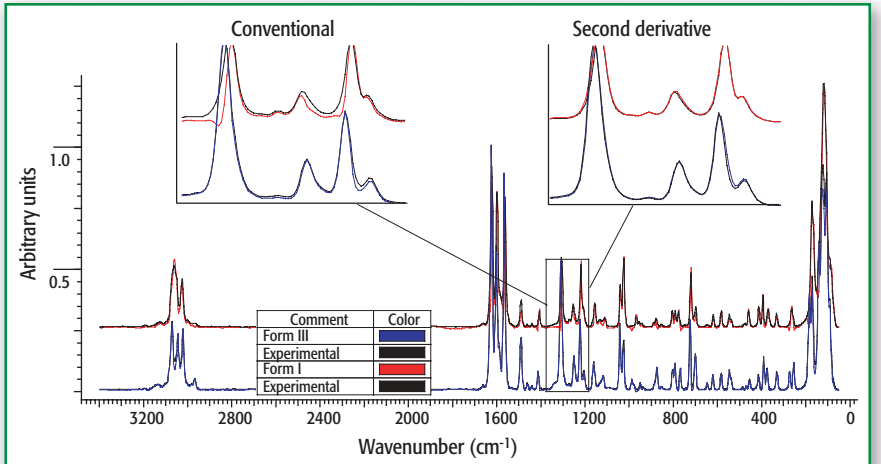

Figure 1. Comparison of spectra extracted with SIMPLISMA versus corresponding experimental spectra in black.

Raman spectroscopy has been described as an "awakening giant" in the analytical world. Advances on the instrumental side have overcome most of the difficulties that have held back this technique in the past, and have provided new advantages. As a result, it probably is fair to say that the biggest hurdle facing Raman users lies in the interpretation of their data and not its acquisition. For example, it often is nontrivial and time consuming to interpret Raman spectra of compounds exhibiting polymorphism (different crystalline phases of the same compound). As a general estimate, a third of all pharmaceutical drugs and 3% of all organic compounds exhibit polymorphism (1). Commonly, polymorphic samples require complex factor-based calibration methods to help quantify and identify the components present (2–4), which, despite efforts to simplify their use, still require a high-level of expertise.

In this study, we will show how a multivariate curve resolution algorithm — described by some as the "sleeping giant" of the chemometrics field — can facilitate the quantitative and qualitative analysis of difficult cases such as polymorphs. Curve resolution techniques belong to the family of qualitative analytical methods and are especially useful for cases where the chemistry and spectroscopy of a sample are inadequately known. Curve resolution is aimed at extracting spectral and curve shapes mostly for identification or interpretation purposes. In contrast, calibration techniques are focused on quantitative information, for example, predicting component concentrations, and might not produce pure spectra of mixture components (for example, principle component regression [PCR]). In this paper we will present an original approach utilizing purity function, a main component of the SIMPLe-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA) algorithm, as a variable selection routine for calibration. Variable selection is a necessary stage of the univariate (single-wavelength) technique, but also can be successfully applied as a preliminary stage in multiple linear regression (MLR). As well, in some specific cases, variable selection is relevant to prepare data for factor-based calibration such as PCR or PLS.

As a case study, the SIMPLISMA curve resolution algorithm will be applied to a technically challenging Raman spectra series for carbamazepine (CBZ) polymorphs. Carbamazepine is a drug for controlling epileptic seizures, which can exist in different polymorph forms. The forms show distinct dissolution rates affecting their effectiveness as a drug (5).

Method and Material

SIMPLISMA is discussed in detail elsewhere (6, 7), so we will only give a quick summary here. The algorithm assumes the existence of a wavelength that experiences substantial intensity contribution from only one of the components in the mixture. In an environment where Beer's law is obeyed, the intensity of the pure variable can be related to a concentration profile. If pure variables can be found for every component in a mixture, then the intensity at these wavelengths can be used to resolve the corresponding spectra through a least squares fit. SIMPLISMA finds pure variables by taking the ratio of the standard deviation to

Michel R.J. Hachey is a technical marketing specialist at ACD/Labs (Toronto, Ontario, Canada). E-mail: michel@acdlabs.com. **Andrey Bogomolov** is a project leader at ACD/Labs for optical spectroscopy. **Keith C. Gordon** is an associate professor with the University of Otago's Department of Chemistry and **Thomas Rades** is a professor at the university's School of Pharmacy (Dunedin, New Zealand).

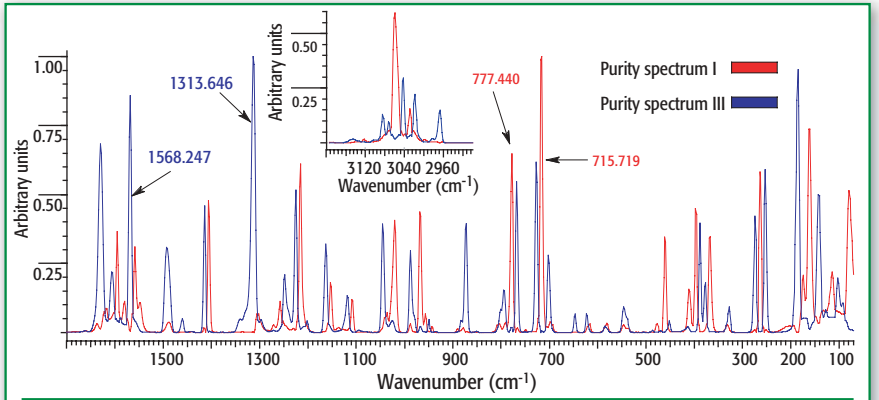

Figure 2. Purity spectrum for carbamazepine in form I and III. The analysis order was rearranged so that the purity spectra relate to only one form and residual noise.

the mean at each frequency, yielding a purity spectrum where the intensity generally is correlated to the purity of the variable. This purity spectrum is calculated stepwise. When going to the next step, the modeled variance relates to the selected pure variables whose effects are eliminated mathematically so that the purity spectrum reflects only residual variance of the data. We used a software implementation of SIMPLISMA executed by ACD/Labs (8) under license from Eastman Kodak.

The carbamazepine calibration data consisted of a binary mixture of CBZ I and III (at 20% [w/w] intervals) in duplicates. The data was baseline corrected. For more details, see reference 4.

Identifying the Components in a Raman Series. To see how multivariate mixture analysis can be used to remedy some of the knowledge gaps about a sample set, let us assume the worst case faced by an analyst, which is that of no prior information. SIMPLISMA therefore was applied to a series of 12 Raman spectra about which we will presume no knowledge. Visual inspection of the residuals, resolved curves, as well as other diagnostic tools (6) at different number of components revealed intuitively and statistically that only two components were present in the mixture. Figure 1 shows the two resolved spectra in red and blue.

The resolved pure component spectra for the two-component model were used to query a spectral database containing the Raman spectra of CBZ form I and III. The experimental hits are overlaid in black against the corresponding resolved spectra in blue and red in Figure 1. The fit is good overall, except for slight intensity dips in the resolved spectra relative to the experimental spectra (especially in form I) as seen in the zoomed view (on the left). This occurs because the selected wavenumbers for the "pure variable" of form I or III were not as pure as was hoped for, and actually included some minor but perceptible extraneous intensity contribution, which caused a distortion of the spectra. Unexpected dips (especially if yielding negative intensities) in the resolved spectrum often are indicators of the presence of interferents at the suggested pure variable. Because the resolution quality problem is small it could be ignored, but another option is to run SIMPLISMA in the second derivative mode to help better resolve the pure variables. The right zoom box in Figure 1 shows that the second derivative mode results in a better pure variable resolution as evidenced by the better overlap of resolved spectra with the experimental ones. Using the resolved spectra as references, Table I gives the hit quality index (%HQI) for a Euclidean distance search against the pure spectra of the CBZ I and III polymorphs. This provides quantitative evidence for the identity of the components.

Table I. Hit quality index for the pure spectra of CBZ form I and III versus resolved spectra.[a,b]

Resolved Spectrum[c]	%HQI Exptl. Form I	%HQI Exptl. Form III
Form I (second derivative)	93.65[d]	77.14
Form I (conventional)	94.65[d]	66.15
Form III (second derivative)	75.42	95.21
Form III (conventional)	59.42	87.79

a. %HQI was generated from a standard Euclidean distance search implemented in ACD/UV-IR Manager (8).
b. Pure samples spectra were excluded from the data set before generating the resolved spectra.
c. Second derivative mode and conventional SIMPLISMA mode results are presented.
d. The %HQI for these two numbers should be considered equivalent, because it is likely due to noise which is unfortunately accentuated in the second derivative case.

Calibration. Picking wavenumbers for calibrations can be one of the most challenging steps in spectroscopic analysis. Interestingly, the pure variable spectrum calculated by SIMPLISMA offers a natural tool for guiding wavenumber selection. Figure 2 shows essentially the purity spectra for CBZ form I (in red) and III (in blue). Two of the purest wavenumbers — having the highest intensity and occurring in regions judged to be less convoluted — were identified in each spectrum for a subsequent calibration analysis.

Because SIMPLISMA can be applied just as easily in second derivative mode, pure wavenumbers also were selected for those cases as indicated in Figure 3. This figure shows the inverted second derivative plot truncated at zero and the corresponding purity spectrum in second derivative mode for the region between 1540–1640 cm^{-1}. By changing the sign of the second derivative to obtain the inverted view, we ensure that the positive peak in the derivative plot matches peaks in the original spectra, which makes comparisons more intuitive. The negative region of the inverted second derivative series is truncated because, as shown by Windig and Stephenson (7), it is more likely to contain contributions from overlapping components, and therefore, ignoring it generally provides a better estimate of the selected pure component's concentrations. The second derivative calculations were obtained with the simple difference method using a smoothing window of three points.

Table II shows some straightforward multiple linear regressions (MLR) that were performed using some of the pure

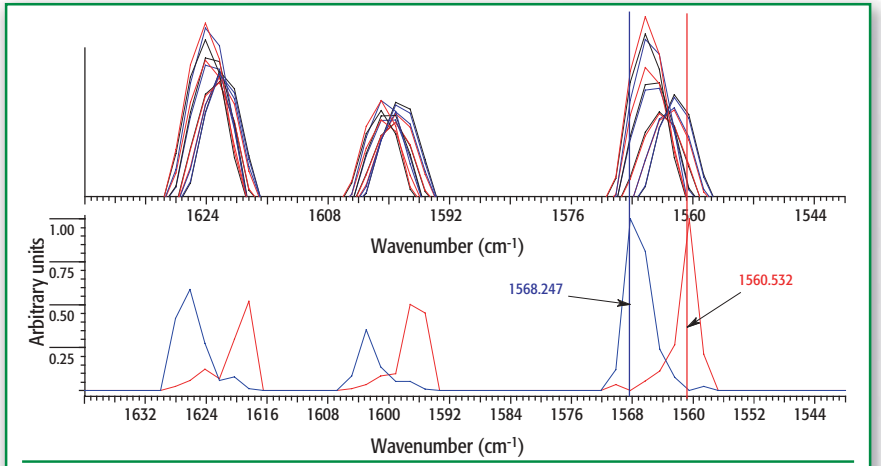

Figure 3. Purity spectrum in second derivative mode and an inverted second derivative spectrum.

variable wavenumbers from SIMPLISMA. The correlation coefficient (R^2), standard error estimate (SEE), and F for regression provide an indication of the model quality as well as robustness. Not surprisingly, the single wavenumber calibration (inverse Beer's law) plots show some of the worst performances. Using multiple "pure" wavenumbers of the same polymorphic form does not improve things significantly. The best results in conventional mode occur when using the purest wavenumber for each form present so as to benefit from interferent corrections. However, the best overall calibration results are provided by the second derivative calibrations. The single wavenumber calibration results using the inverted second derivative are comparable to the two wavenumber MLR calibrations in conventional mode, with the statistics being only slightly worse in the case of CBZ III and slightly better for CBZ I. Finally, we note that the second derivative using two pure variable for form I and III, respectively, provides the best overall calibration statistics.

Discussion

Because calibration of polymorphs tends to be technically challenging, Raman data analysis tends to use factor-based methods, which unfortunately requires considerable training and expertise to use properly. SIMPLISMA multivariate curve resolution is simple enough to use by general chemists and provides tools to facilitate the generation of simpler calibration models. Its advantage is that it is not based upon principle components analysis and it is interactive. The interactivity is important in practical industrial environments where it is not always possible to design experiments, obtain replicate and/or uncontaminated samples when trouble shooting. Failures in the SIMPLISMA model generally are easy to spot due to the unnatural spectral shapes, such as negative intensity peaks, so that it is easy to know when to move to more complex calibration models.

We saw that SIMPLISMA multivariate curve resolution provided us with reference spectra for the pure components in a mixture without a priori knowledge in a self-modeling way that only required ordinary spectroscopic intuition to guide the analysis. Furthermore, the purity spectra generated by SIMPLISMA took some of the guess work out of selecting the best wavenumbers for simple MLR calibration methods. The model suggested that MLR calibration using the second derivative pure variable gives a very good calibration model.

Table II. Calibration statistics for various inverse Beer's law and MLR models that use wavenumbers extracted from the purity spectra.

MLR Calibration for CBZ Form I					
Purity variables from SIMPLISMA					
$\tilde{\nu}$ IIIa	$\tilde{\nu}$ Ia	Form $\tilde{\nu}$ Ib	R^2	SEE (%)	F
–	715.719	–	0.9874	4.2	781
–	715.719	777.440	0.9896	4.0	429
1313.645	715.719	–	0.9973	2.0	1677
Calibration with Second Derivative					
–	1560.532	–	0.9982	1.6	5701
1568.247	**1560.532**	**–**	**0.9986**	**1.5**	**3291**
MLR Calibration for CBZ Form III					
Purity variables from SIMPLISMA					
$\tilde{\nu}$ IIIa	$\tilde{\nu}$ IIIb	$\tilde{\nu}$ Ia	R^2	SEE (%)	F
1313.646	–	–	0.9675	6.7	298
1313.646	1568.247	–	0.9916	3.6	528
1313.646	–	715.719	0.9973	2.0	1677
Calibration with Second Derivative					
1568.247	–	–	0.99167	3.4	1189
1568.247	–	1560.532	0.9986	1.5	3291

12 standard samples were used.

Acknowledgment

We are grateful to Irina Oschepkova for her help obtaining article references, and to Margaret Antler for helpful discussions.

References

1. A.L. Grzesiak, M. Lang, K. Kim, and A.J. Matzger, *J. Pharm. Sci.* **92**, 2260–2271 (2004).
2. T. Head and J. Rydzak, *American Pharmaceutical Review*, 78–84 (Spring 2003).
3. G. Zhou, J. Wang, Z. Ge, Y. Sun, *American Pharmaceutical Review* (Winter 2002).
4. C. Strachan, D. Pratiwi, K.C. Gordon, and T. Rades, *Journal of Raman Spectroscopy* **35** (2004).
5. P.N. Zannikos, W-I. Li, J.K. Drennen, and R.A. Lodder, *Pharm. Research* **8**, 974–978 (1991).
6. W. Windig and J. Guilment, *Anal. Chem.* **63**, 1425–1432 (1991).
7. W. Windig and D.A. Stephenson, *Anal. Chem.* **64**, 2735–2742 (1992).
8. Advanced Chemistry Development, Inc. (ACD/Labs), ver. 8.0, Toronto, ON, Canada, www.acdlabs.com/uvir (accessed April 2003). ■