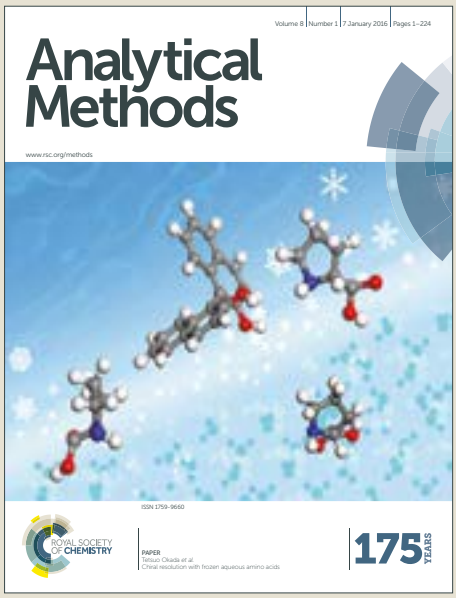
Chemometric approaches to low-content quantification (LCQ) in solid-state mixtures using Raman mapping spectroscopy.

Analytical Methods

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: B. Li, Y. Casamayou-Boucau, A. Calvet and A. G. Ryder, *Anal. Methods*, 2017, DOI: 10.1039/C7AY01778B.



ROYAL SOCIETY OF CHEMISTRY

rsc.li/methods

View Article Online
DOI: 10.1039/C7AY01778B

Journal Name

ARTICLE

Chemometric approaches to low-content quantification (LCQ) in solid-state mixtures using Raman mapping spectroscopy.

Received 00th March 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Boyan Li,[a] Yannick Casamayou-Boucau,[a] Amandine Calvet,[a] and Alan G. Ryder[a]*

Abstract The low-content quantification (LCQ) of active pharmaceutical ingredients or impurities in solid mixtures is important in pharmaceutical manufacturing and analysis. We previously demonstrated the feasibility of using Raman mapping of micro-scale heterogeneity of solid-state samples combined with partial least squares (PLS) regression for LCQ in a binary system.[1] However, PLS is limited by the need for relatively high calibration sample numbers to attain high accuracy, and a rather significant computational time requirement for the large Raman maps. Here we evaluated alternative chemometric methods which might overcome these issues. The methods were: net analyte signal coupled with classical least squares (NAS-CLS), multivariate curve resolution (MCR), principal component analysis with CLS (PCA-CLS), and the ratio of characteristic analyte/matrix bands combined with shape-preserving piecewise cubic polynomial interpolation curve fitting (BR-PCHIP). For high (>1.0%) piracetam analyte content, all methods were accurate with relative errors of prediction (REP) of: <1.1%. For LCQ (0.05−1.0% w/w), three methods were able to predict piracetam content with reasonable levels of accuracy: 6.97% (PCA-CLS), 9.13% (MCR), and 12.8% (NAS-CLS). MCR offered the best potential as a semi-quantitative screening method as it was ~40% quicker than PLS, but was less accurate due to being more sensitive to spectral noise factors.

Introduction

Chemical imaging by Raman spectroscopy has great potential for measuring the chemical composition of solid-state materials providing high spatial resolution and molecular specificity, and thus in recent years has gained ever-increasing attention for both academic and industrial applications.[2-7] Chemical images, particularly those based on vibrational spectroscopies, contain information concerning the sample composition, which is essential for microscopic identification, location, distribution of chemical species. This spatial and composition knowledge can help monitor manufacturing processes, for instance, in the mass production of solid-state pharmaceutical formulations such as tablets.[5-12]

Raman spectroscopy is advantageous in terms of being fast, non-contact, non-destructive, inexpensive, and often requires minimal sample preparation. Moreover, the often sharp, well-defined Raman spectra and high spatial resolution frequently enable easier chemical discrimination of distinct sub-micro-sized zones or layers within samples.[2, 6, 7, 13] As a result, Raman spectroscopy has been widely applied to the detection of low-content polymorphs, impurities, and contaminants, and the characterization of active pharmaceutical ingredients (APIs), excipient distribution in solid materials.[6, 8, 11, 14-16] This is in contrast with conventional methods such as high performance liquid chromatography (HPLC) for impurity analysis which requires extensive sample preparation, or powder X-ray diffraction (PXRD) used for polymorph analysis which has poor low-content quantification (LCQ) and spatial resolution. The combination of spectroscopic imaging with multivariate statistical analysis can further dramatically enhance the information outcome from Raman imaging.[7, 17, 18] Multivariate exploratory analysis like principal component analysis (PCA) can be used to determine the specific sources of subtle and gross sample variance and curve resolution methods have been used frequently to show component distributions.[4, 7, 8, 19, 20]

We previously demonstrated that Raman mapping of micro-scale heterogeneity in a solid sample mixture, combined with chemometric and statistical analysis can be used for accurate LCQ.[1] This first attempt used partial least squares (PLS) regression which is problematic in terms of the requirement for large numbers of calibration samples in order to achieve high accuracy. Furthermore, in many cases a sufficiently large calibration sample set may not be always available because of sample availability and/or complexity, or the costs associated with preparation of calibration samples. Here we report the competitive assessment of four different chemometric techniques for LCQ of binary solid-state mixtures by Raman mapping spectroscopy. The methods were compared to determine which offered the optimal quantitative performance, LCQ accuracy, and potential for smaller numbers of calibration samples. The methods were: net analyte signal coupled with classical least squares (NAS-

[a] Nanoscale Biophotonics Laboratory, School of Chemistry, National University of Ireland, Galway, Galway, Ireland.

† Footnotes relating to the title and/or authors should appear here.
Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

View Article Online
DOI: 10.1039/C7AY01778B

CLS), multivariate curve resolution (MCR), PCA with CLS (PCA-CLS), and analyte/matrix band ratio combined with shape-preserving piecewise cubic polynomial interpolation curve fitting (BR-PCHIP).

Experimental

Materials and Methods

Solid mixtures of piracetam (2-oxo-1-pyrrolidineacetamide, polymorphic form III) and L-proline (≥99%) were prepared in triplicate and Raman data collected as described previously.[1] However, here we only used data from 50 of the 61 samples (20 mixtures between 0.05% and 0.95% and 30 mixtures between 1.0% and 100%), and the data set thus comprised of 150 Raman maps (50 samples × 3 replicates). This was done to see if we could reduce the overall sample number and still obtain acceptable accuracy. Piracetam and proline were selected because they have approximately equal Raman scattering coefficients and ~50% overlap in the spectral region measured (see S-10, *SI* for more details). Raman spectra (200–1896 cm^{-1}) were collected using a RAMAN WORKSTATION™ Analyzer (Kaiser Optical Systems, Inc.), with 785 nm excitation. Raman maps of 29×29 points were recorded with 1 mm spacing using 1 second exposure. At each point, there were 10 probe channels (Raman map channel) which generated a total of 8410 spectra per map.

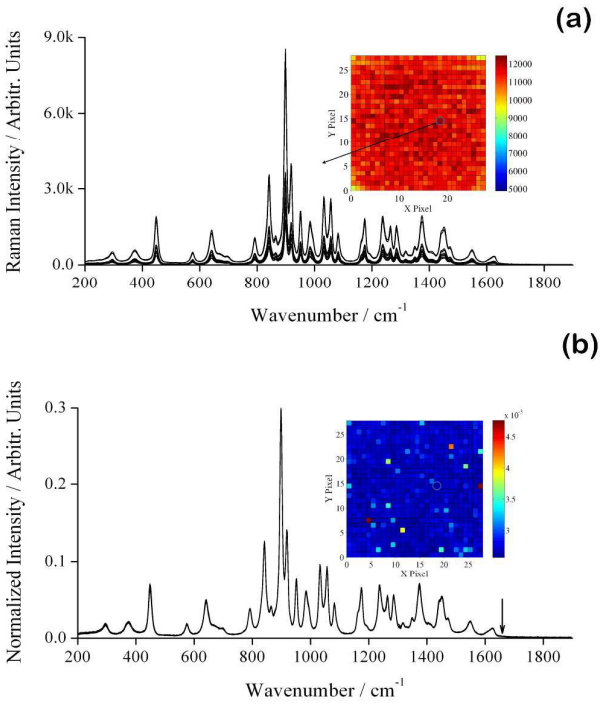
MATLAB R2014b (MathWorks Inc., Natick, MA) was used to implement in-house-written routines for data processing and computation (Windows 7 desktop PC, Xeon 2.8 GHz CPU, 6 GB RAM). MATLAB codes for ant colony optimization (ACO)[21, 22] were generously provided by Prof. A.C. Olivieri (Universidad Nacional de Rosario, Argentina) while MCR-BANDS was downloaded from www.mcrals.info.

The Raman data arrays were first unfolded into two-way matrices prior to spectral pre-processing and implementation of chemometric methods (see *Supplemental Information*). Appropriate data pre-processing was performed to maximize the extraction of chemical information from the Raman spectra to improve model accuracy. The pre-processing methods were: (*i*) baseline correction using morphological weighted penalized least-squares (MPLS)[23] to mitigate baseline artefacts arising from particle size effects and sample surface roughness; (*ii*) cosmic ray artefacts (CRAs) removal with kernel principal component analysis residual diagnosis (KPCARD);[24] (*iii*) exclusion of abnormally weak Raman spectra (generally caused by map edge effects), threshold set at 30% of the average total integrated spectral intensity for each Raman mapping measurement; (*iv*) multiplicative scatter correction (MSC)[25-27] and/or standard normal variate (SNV) [25-27] to reduce scattering variations between spectra or measurements for the PLS modelling procedure (SNV was found to be the best method); (*v*) spectrum normalization to scale each individual spectrum in the predefined spectral region to unit area under the spectrum curve; and (*vi*) variable selection by ACO using Olivieri's method.[1, 22]

Model quality and prediction performance was assessed using multiple parameters as previously described[1]: root mean square error: of calibration (RMSEC), of cross validation (RMSECV) and, of prediction (RMSEP), as well as the relative error of prediction (REP%=100×RMSEP/$\bar{y}_{pred}$, where $\bar{y}_{pred}$ = mean value of measured piracetam content of prediction set), of calibration (REC%=100×RMSEC/$\bar{y}_{cal}$, where $\bar{y}_{cal}$ = mean value of measured piracetam content of calibration set), of cross validation (RECV%=100×RMSECV/$\bar{y}_{cv}$, where $\bar{y}_{cv}$ = mean value of measured piracetam content of cross validation set), and the square of the correlation coefficient (R^2) between predicted and measured values.[28, 29] For all calibration models reported, a leave-one-out (LOO) cross validation was used.

Methodology

The key study goal was to evaluate model efficacy in terms of accuracy, robustness, and computational efficiency for the various approaches. The simplest BR-PCHIP method, used the ratio of bands at 1652 cm^{-1} (piracetam) and 448 cm^{-1} (proline). However, in more complex samples, constituent-specific spectral bands free of interferences may not often be available and therefore, multivariate methods are required to better extract the quantitative information from Raman mapping data which can be either two-way matrix or three-way data arrays.[4, 7, 19, 30] Multivariate methods include CLS,[8, 17, 31, 32] PCA,[8, 30] MCR,[8, 32] and PLS.[1, 33, 34] Generally, each method yielded a score or parameter for each spectrum, and then these scores or parameters were mapped as a function of spatial coordinates so as to quantitate the molecular components at each pixel and assess their distribution on the measured sample surface. Here, MCR-alternating least squares (MCR-ALS)[35] was used to quantify piracetam content and MCR-BANDS[36, 37] was used to evaluate the degree of rotational ambiguities associated with the MCR solutions (*SI*). Unlike PLS which requires a sufficiently large set of representative samples, MCR-ALS could potentially require fewer calibration samples for quantification (*i.e.* to scale the MCR scores in the estimated concentration profiles). This is of practical importance in that it reduces the workload involved in producing quantitative models.



View Article Online
DOI: 10.1039/C7AY01778B

Figure 1: (a) *As acquired* spectra from the 10 spectrometer channels at a specific pixel from one typical Raman map of a 0.103% piracetam mixture. The inset shows the integrated intensity Raman map and the colour bar indicates Raman intensity. (b) Normalized spectra from plot (a). Inset map was constructed using the 1652 cm^{-1} piracetam band intensity from the normalized spectra. The grey open circle specifies the example pixel used in prediction (*vide infra*).

Results and discussion

Raman mapping data analysis:

Figure 1 shows a single Raman integrated intensity map acquired from a piracetam-proline mixture (0.103% w/w piracetam). At each map pixel the 10 probe channels produced different spectra and since each channel binned different numbers of detector rows, spectral intensities and channel signal-to-noise ratios (SNRs)[1] varied significantly. The integrated intensity was computed by: first averaging all 10 channel spectra, and then calculating the Frobenius norm value of this resultant spectrum. The variation in map colour (Figure 1, inset) thus provided a simple visual representation of gross sample heterogeneity. However, it does not provide specific information about the source of this variation which could be caused by either instrument derived spectral intensity variation, variations in sample physical properties, or variations in the chemical compositions across the measured sample surface, or a combination of all these factors. The only identifiable outliers caused by mapping edge effect were the three corner pixels with very low intensity spectra. Overall sample macro-homogeneity was very good, and this was substantiated by the small variation of the integrated intensity across the image: the *mean* and *standard deviation* of the integrated intensity values for the 841 pixels were $(11.40\pm0.55)\times10^3$ in arbitrary units (<5%). Weak spectra were excluded from data analysis and modelling, if their integrated intensity was less than 30% of the average spectral intensity of the entire Raman map.

After normalization, the spectra at each map pixel for all channels were virtually identical in terms of both intensity and band shape (Figure 1b), apart from differences in baseline and SNR. The second map (Figure 1b inset) produced using the 1652 cm^{-1} piracetam band intensity after normalization and averaging to a single spectrum per pixel showed a relatively weak *mean* intensity and relatively high *standard deviation* of $(27.44\pm1.72)\times10^{-4}$. The relatively large ~6.3% deviation indicated that spectral normalization and the use of an analyte-specific spectral band more accurately represented the true chemical heterogeneity of the sample caused by the minor component. Some of the other spectral variance such as changes in baseline, and unwanted scattering variation could also be related to compositional heterogeneity and thus contain useful information. However, these effects can also be caused by physical changes in the bulk matrix (*i.e.* proline) and in general they degrade quantitative model accuracy. Therefore, it was necessary to eliminate these effects and CRAs, prior to chemometric modelling. This has been described in detail elsewhere.[1, 24]

Since each spectral channel had a different SNR, this also necessitated calculation of separate, channel specific calibration models. For all multivariate calibration modelling the averaged spectra (generated from each Raman map channel) for all samples were used to build the calibration models. These were then deployed to predict the local concentrations at each pixel [1] (Figure S-1 in *SI* for scheme).

BR-PCHIP quantification:

This involved several steps: the pre-processed spectra from each Raman map channel were averaged into a single spectrum. Second, the 1652/448 cm^{-1} intensity ratios (Figure 2a) were calculated for the average spectra (*SI*). Third, using these band ratios, univariate calibration models using leave-one-out cross-validation[38] were built for the 0–99.9%, 0–4.0%, and 4.0–99.9% piracetam concentration ranges (Figure 2b, Table 1). The 100% piracetam content samples were excluded from modelling because no proline was present and thus any band ratios would be noise determined. For BR-PCHIP all RMSEC and REC% values were equal to, or approached zero, due to the use of a spline (interpolation) curve fitting method.

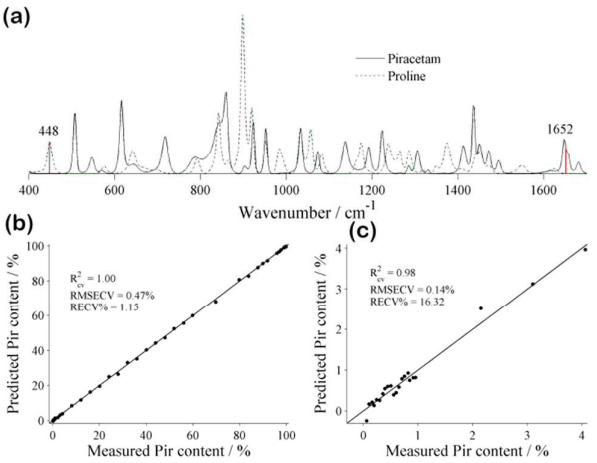


Figure 2: (a) Raman spectra of pure piracetam (solid) and proline (dashed); (b/c) The piracetam quantification BR-PCHIP models (spectrometer channel 5 data) with different piracetam content ranges, and only the cross-validation results shown.

When piracetam content was high (4.0–99.9%), BR-PCHIP quantification was accurate, yielding the *mean* and *standard deviation* values of 0.62 ± 0.09% for all ten channels. However, both full- (*Model1*) and low concentration (*Model2*) range models had average RECV values which were too large for LCQ (1.19% and 18.67% respectively). Ultimately, BR-PCHIP failed because it was intrinsically a univariate method and did not make full use of the multivariate spectral information.

Table 1: RMSECV values (in w/w%) obtained for BR-PCHIP piracetam quantification models at each spectrometer channel. Model accuracy was assessed by RECV% for cross-validation. The greyed-out rows show the mean and standard deviation calculated for all 10 channels.

BR-PCHIP model	***Model1***	***Model2***	***Model3***

View Article Online
DOI: 10.1039/C7AY01778B

Piracetam in %	0–99.9	0–4.0	4.0–99.9
Channel 1	0.465	0.139	0.582
Channel 2	0.541	0.249	0.673
Channel 3	0.372	0.115	0.449
Channel 4	0.461	0.147	0.602
Channel 5	**0.469**	**0.138**	**0.602**
Channel 6	0.448	0.165	0.583
Channel 7	0.480	0.144	0.611
Channel 8	0.573	0.108	0.735
Channel 9	0.491	0.133	0.623
Channel 10	0.555	0.224	0.777
mean value	0.49	0.16	0.62
standard dev.	0.06	0.05	0.09
RECV%	1.19	18.67	0.94

PLS quantification:

PLS was implemented using similar methods to those described in an earlier publication and are details are included in the supplemental information. One reason why PLS generated accurate LCQ was that there were a sufficiently large number of calibration samples available. Since predictive error is directly dependent on calibration set size, more and more samples are required to reduce prediction error,[39] particularly for LCQ and this is not always feasible. There are also many practical difficulties with preparing calibration samples with precisely known low levels of analytes or contaminants. Therefore, we needed to examine the feasibility of using NAS and MCR based approaches where smaller sized calibration sample sets might be employed, particularly for more complex mixtures with multiple low content components.

NAS-CLS quantification:

NAS-CLS models used three specific Raman spectral ranges (480–830 cm^{-1}, 1040–1510 cm^{-1} and 1628–1740 cm^{-1}, *See S-9, SI*) and a different set of segmented analyte concentration ranges (0–1.0%, 1.0–21.5%, 21.5–85.0%, 85.0–100%). The resultant RMSECV and RMSEC values (Table 2) were low, and similar in magnitude, indicating reasonably fitted models. This was due to the mutual correlation (0.95) between the NAS factor ($\mathbf{S}_{nas}$) and piracetam spectrum (Table 3 & Figure 3) where it was clear that $\mathbf{S}_{nas}$ contained most of the piracetam bands in these three spectral ranges, even though the profile was not identical to the piracetam spectrum. The low, negative correlation (−0.32) with the proline spectrum meant that the NAS-CLS method was nearly able to filter out the background matrix signal, without the requirement of any *a priori* knowledge/input. Importantly, the higher correlation to the piracetam spectrum and lower correlation to the proline spectrum explained why the NAS-CLS *Model1* had better accuracy (*i.e.*, REC=3.03%, RECV=3.53%) than the PLS *Model1* (REC=5.66%, RECV=6.47%). However, in *Model2* the use of low piracetam content (0–1.0%) and the fact that quite small variations between the sample spectra were present resulted in a second NAS factor ($\mathbf{S}_{nas\_M2}$, Figure 3). The correlation coefficients between $\mathbf{S}_{nas\_M2}$ and $\mathbf{S}_{nas}$ (0.92), piracetam (0.92), and proline (−0.18), indicated that there were certain differences between *Model2* $\mathbf{S}_{nas\_M2}$ and *Model1* $\mathbf{S}_{nas}$; $\mathbf{S}_{nas\_M2}$ was not able to carry over some contribution from piracetam to which small spectral variations of the low-content samples should have been ascribed. This negatively impacted the model accuracy (Table 2) compared to PLS, with NAS-CLS *Model2* having an REC=8.13% (RECV=9.26%).

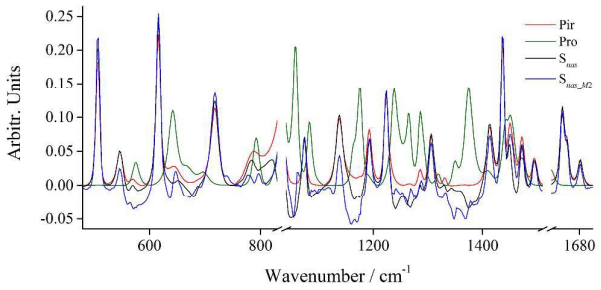


Figure 3: NAS factor profiles from *Model1* ($\mathbf{S}_{nas}$) and *Model2* ($\mathbf{S}_{nas\_M2}$) for the 480–830, 1040–1510, and 1628–1740 cm^{-1} regions, overlaid with the spectra of pure piracetam and proline. Both factors shown were obtained from the spectrometer channel 5 spectra. Similar results were obtained for all the other channels.

Table 2: RMSEC(RMSECV) values (in w/w%) obtained from the NAS-CLS piracetam quantification models for each spectrometer channel. Model accuracy was assessed by REC% and RECV% for calibration and cross-validation respectively. The greyed-out rows show the mean and standard deviation calculated for all 10 channels.

	NAS-CLS				
	Model1	***Model2***	***Model3***	***Model4***	***Model5***
w/w %	0–100	0–1.0	1.0–21.5	21.5–85	85–100
Chan. 1	0.90	0.031	0.097	0.51	0.34
	(1.05)	(0.034)	(0.12)	(0.63)	(0.41)
Chan. 2	0.84	0.044	0.097	0.57	0.34
	(0.98)	(0.05)	(0.13)	(0.69)	(0.43)
Chan. 3	0.81	0.042	0.08	0.44	0.31
	(0.95)	(0.047)	(0.10)	(0.53)	(0.38)
Chan. 4	0.84	0.038	0.099	0.51	0.42
	(0.98)	(0.042)	(0.13)	(0.62)	(0.51)
Chan. 5	**0.86**	**0.034**	**0.11**	**0.54**	**0.41**
	(1.00)	**(0.04)**	**(0.14)**	**(0.65)**	**(0.50)**
Chan. 6	0.83	0.038	0.12	0.53	0.30
	(0.96)	(0.043)	(0.16)	(0.64)	(0.35)
Chan. 7	0.83	0.041	0.10	0.54	0.43
	(0.97)	(0.046)	(0.13)	(0.65)	(0.52)
Chan. 8	0.89	0.04	0.12	0.57	0.28
	(1.03)	(0.05)	(0.16)	(0.70)	(0.35)
Chan. 9	0.86	0.04	0.11	0.56	0.25
	(1.00)	(0.05)	(0.15)	(0.69)	(0.31)
Chan. 10	0.82	0.045	0.08	0.51	0.31
	(0.95)	(0.052)	(0.09)	(0.61)	(0.38)
Mean	0.85	0.040	0.10	0.53	0.34
	(0.99)	(0.045)	(0.13)	(0.64)	(0.41)
Std. dev.	0.028	0.005	0.015	0.038	0.06
	(0.033)	(0.005)	(0.023)	(0.05)	(0.08)
REC%	3.03	8.13	1.12	1.11	0.36

View Article Online
DOI: 10.1039/C7AY01778B

(RECV%)	(3.53)	(9.26)	(1.46)	(1.35)	(0.44)

MCR quantification:

MCR with non-negativity constraints applied to both spectral and concentration values was used to resolve spectra into their contributing pure spectral constituents (*i.e.* loadings). In addition, the spectral values were normalized (each MCR factor normalized to unit length). Two MCR factors (Figure S-9, *SI*) and concentration scores (Figure 4a) were extracted from each channel spectral dataset in the 480–830, 1040–1510, 1628–1740 cm^{-1} regions. The scores show a constant increase/decrease for piracetam/proline with sample number, with very similar values for the replicate measurements. Overall these components explained >99.81% of the spectral variance for all the 10 channels (Table S-2, *SI*). To evaluate the ambiguity of each MCR factor, all the obtained factors were analysed by MCR-BANDS and their relative signal contributions in the mixture spectra were calculated, according to equation (17) (*SI*). For example, the two MCR factors in the channel 5 data gave contribution values of 0.7284, and 0.6531. Moreover, the differences between the maximum and minimum values ($f_{1,2}^{max} - f_{1,2}^{min}$) were all zero which demonstrated that the MCR loadings and concentration profiles obtained were unique, and they did not change during the optimization procedure using these applied constraints. Similar, but different relative contribution values were obtained for all other channels (Table S-3, *SI*).

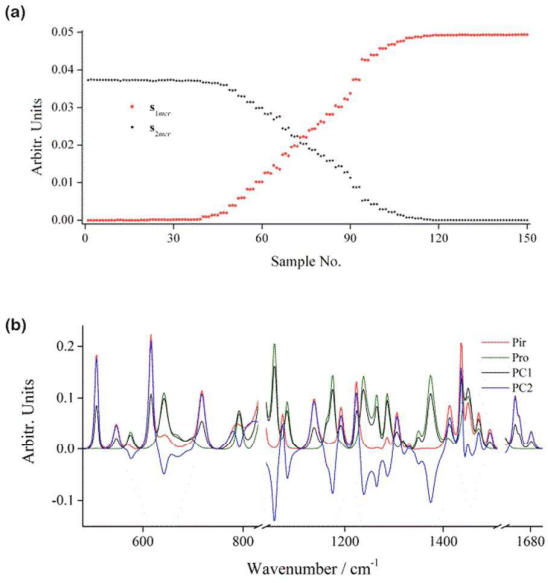


Figure 4: (**a**) MCR scores from channel 5 model (see Fig. S-9, SI for MCR factors); (**b**) Principal component profiles from PCA-CLS modelling, overlaid with Raman spectra of piracetam and proline. Data obtained from *Model1* (channel 5).

For channel 5, two MCR factors explained 99.82% of the spectral variance, and Figure 4a shows the MCR scores of the two factors ($\mathbf{S}_{1mcr}$ and $\mathbf{S}_{2mcr}$ respectively) obtained from 150 Raman maps (0–100% w/w piracetam). When the MCR factors were compared with spectra of pure piracetam and proline it was clear that $\mathbf{S}_{1mcr}$ was piracetam, and $\mathbf{S}_{2mcr}$ was proline (Fig. S-9, *SI*). The calculation of correlation coefficients furthermore confirmed their identification as all values were equal to one (Table 3). The resulting MCR scores were scaled to the piracetam content by least squares, and 50 calibration models were built (Table 4). These were all accurate except for the low-content *Model2* (0–1.0% w/w) where large REC% (8.77%) and RECV% (12.95%) were obtained. The model accuracy was comparable to the NAS-CLS models (Table 2). Particular attention must be paid to the MCR decomposition of all the 10 channel spectra (Table S-2, *SI*), and the %LOF values (that is the amount of spectral variance not explained by the MCR models) were between 0.1693% and 0.1913%. According to the variance principle: $(var(c_1\boldsymbol{s}_1 + c_2\boldsymbol{s}_2) = c_1^2\sigma_{11} + c_2^2\sigma_{22} + 2c_1c_2\sigma_{12})$,[40] the low-content piracetam (c_1) in the binary mixture spectrum ($\boldsymbol{x}_i = c_1\boldsymbol{s}_1 + c_2\boldsymbol{s}_2$) could generate at least c_1^2 of the spectral variance. For example, a piracetam content of 0.42% in a binary mixture could represent ~0.18% of the overall spectral variance, which was similar in magnitude to the %LOF values measured here for example in the channel 5 MCR models. This can explain why MCR may not detect concentration changes when the piracetam content was very low, and thus LCQ became difficult. Therefore, to improve MCR *Model2* (0–1.0% piracetam content) accuracy, the model residual (**E**) in Equation (15) (*SI*) was considered as a *pseudo-component*, and the corresponding scores were obtained and then used for compensating LCQ in MCR quantification models together with the MCR scores of piracetam ($\mathbf{S}_{1mcr}$) by least squares. The use of a correlation constraint[41, 42] to improve the MCR modelling here is probably not appropriate as we do not have any significant, extrinsic spectral interferences, and the test system is a simple binary mixture. Furthermore, the accuracy of the high concentration range models, (Table 4) indicated that spectral noise (both dark & shot) was the primary factor affecting model error. In the future however, the use of correlation constraints in more complex multi-component mixtures may be required for accurate LCQ.

PCA-CLS quantification:

PCA-CLS analysis on the same datasets required four PCs for data decomposition. Figure 4b shows PC1 (83.20% of the variance explained) and PC2 (16.52%) used in *Model1* (0–100% w/w piracetam) for the channel 5 data. The PC3 (0.15%) and PC4 (0.07%) plots were given in Figure S-5 (*SI*). The correlation coefficients with pure piracetam and proline spectra (Table 3) indicated that PC1, PC2 and PC3 were all the composites of both species. PC2 was more representative of piracetam, but still contained proline spectral features. This is unsurprising because PCs have to be orthogonal to best explain data variance but this does not correspond to the PCs being identical to pure component spectra. Therefore, the use of PCA scores of a single PC, to quantify a real chemical species may not always be valid, and this can lead to inaccuracies. Finally, PC1, PC2, and PC4 scores were combined to develop the PCA-CLS piracetam content calibration models for five concentration ranges (Table 4). All model errors were quite small, and PCA-CLS was the second best performing of the methods investigated. This good performance was explicitly derived from

View Article Online
DOI: 10.1039/C7AY01778B

the use of the combination of scores of the three most significant components, rather than a single PC.

For each method, the combination of all the individual channel calibration models was used for prediction of the bulk sample concentrations. Since the calibration models for each channel were built using the data averaged from all pixels in the Raman map (29×29), the prediction of the individual (n≈8410) spectra from each sample map could be considered as being quasi-independent. The first step was to use *Model1* (the wide concentration range) to provide an initial estimate of the local piracetam concentration at each pixel in the Raman map. Next, the appropriate segmented concentration range model (Table 1, Table 2, and Table 4) was selected, and then piracetam content was re-estimated more accurately. This was done for all 10 data channels, at each pixel, generating ~8400 piracetam concentration predictions per sample. The final predicted sample piracetam content was the mean value of all of these prediction scores. A comparison of the output histograms for the different methods and a single sample prediction is provided in Figure 5a.

The same procedure was applied to the five quantification methods and the method prediction performances were compared (Table 5). For the high-content range (>1.0%) all methods were good, yielding good prediction accuracies with relatively low REP% values (Figure S-4, *SI*). PCA-CLS was best at quantitating piracetam content, with a REP% value of 0.62. The BR-PCHIP method was limited to <97.0% w/w piracetam content but was the fastest along with MCR requiring ~60% of the time required by PLS. However, for the low-content (0.05−1.0%) range, BR-PCHIP was the worst (REP%=35.74), NAS-CLS was marginal, MCR and PCA-CLS gave acceptable accuracy (REP%<10). PLS was however, the most accurate (REP%=2.81) with a limit of detection (LOD) spanning a range of 0.033−0.056%.[1]

For NAS-CLS, MCR, and PCA-CLS (Figure 5b-d), there was considerably more scatter about the regression line compared to the PLS model (Figure 5e). This was due to the fact that these methods are less well able to discriminate clearly the analyte signal from the noise contribution. For NAS-CLS there were also two significant outliers present (0.197% and 0.357% samples) had overestimated piracetam content by 80% and 41% respectively (Figure 5d), which caused a large REP (12.80%). When these two samples were excluded and the 0.05−1.0% range model recalculated, REP decreased to 7.06%.

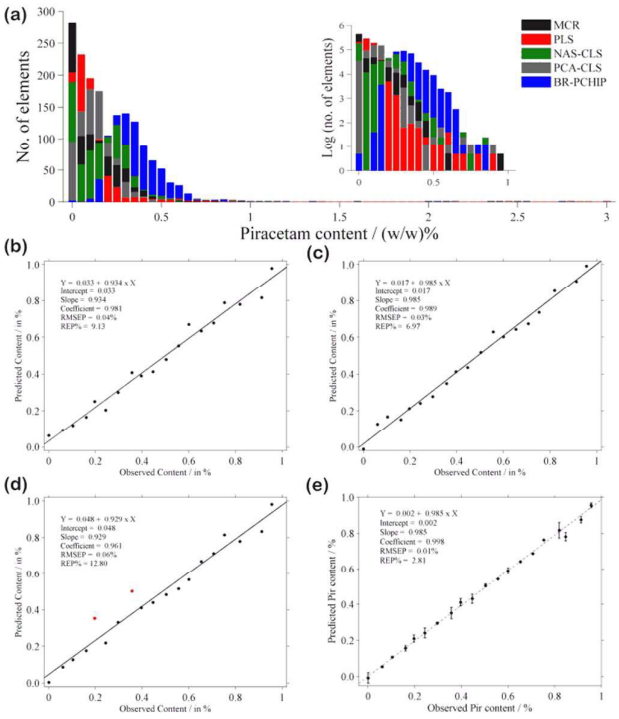


Figure 5: (a) Overlaid histograms of predictions by five methods for the 0.103% piracetam powder mixture (inset shows log plot of the 0.05 to 1% range). Piracetam prediction models in the low-content range (0.05−1.0%) for: (b) MCR; (c) PCA-CLS; (d) NAS-CLS; and (e) PLS. Error bars represent standard error for n=3 replicate samples.

As these methods, all tended to overestimate the minor analyte concentration, probably because spectral noise or proline interference was interpreted as analyte signal. This was due to the fact that most pixels had no piracetam content but spectral noise and variance contributed to each local prediction mostly being returned as a greater value than reality and cumulatively all of the small over predictions led to high errors. This can be seen in the results for the prediction of the overall 0.103% piracetam sample concentration and that at one specified pixel (Figure 5a, Table 5). PLS was the most accurate at 2% high, and MCR was next being 8% higher than the true value. All the other methods overestimated the analyte concentration very significantly (>+10%). Figure 5a shows the prediction histograms where it is clear that the distributions are very different. The most accurate methods (PLS, MCR, and PCA-CLS) had prediction distributions close to or below the real concentration. For BR-PCHIP it was evident that all predictions were overestimated with an almost symmetric distribution centred at ~0.34 w/w%, whereas for the NAS-CLS there was almost a bimodal distribution centred at 0% w/w and ~0.25% w/w respectively.

To assess the effect on reducing sample numbers on the accuracy of the MCR, NAS-CLS and PLS prediction models (3 x 6), a series of models were constructed using the 0.05-1.0% w/w

View Article Online
DOI: 10.1039/C7AY01778B

samples (only for Channel 5 data). The spectra were pre-treated as normal and ACO variable selection used with a simple LOO cross validation. Varying sample numbers from 8 to 18 (supplemental information) were used and the results (Fig. S-8, *SI*) showed that acceptable accuracies (in terms of RMSECV) were obtained for all three methods with 12 samples. However, it must be noted that PLS did outperform the other methods and that this was only a single set of models, implemented on a subset of the data.

Table 3: Correlation coefficients between the pure spectra of piracetam (Pir), proline (Pro), and NAS, MCR, PCA, and PLS factors in *Model1* (channel 5 data). $\mathbf{S}_{nas\_M2}$ is the *Model2* NAS factor.

Factors	$\mathbf{S}_{nas}$	$\mathbf{S}_{nas\_M2}$	$\mathbf{S}_{1mcr}$	$\mathbf{S}_{2mcr}$	PC1	PC2	PC3	LV_{PLS}	Pir	Pro
Pir	0.95	0.92	1	−0.07	0.44	0.81	0.13	0.93	1	−0.08
Pro	−0.32	−0.18	−0.07	1	0.86	−0.64	0.12	−0.40	−0.08	1

Spatial piracetam distribution:

One of the most common uses of chemical imaging is to identify the presence of impurities or chemical heterogeneity on sample surface. For solid powder mixtures when finely milled, the impurity or API should be homogeneously distributed on the macroscale (>mm size). However, on the microscale there can be significant variability and it may be important to know the actual impurity or API concentration at each pixel. For example, when examining homogeneity across the surface of a tablet or medical device.

Prediction scores generated by various methods were used to represent piracetam distribution and Figure 6 shows the score maps for the 0.103% piracetam sample. In the low-content samples, the vast majority of a Raman map will correspond to the matrix (proline here) and the analyte or impurity (piracetam) signal will be relatively weak. The simple use of the integrated signal intensity to represent sample homogeneity (Figure 1a) does not provide any useful information since the evenness of sample surface was a significant factor affecting the signal intensity. It is more common to use a ratio measurement similar to the BR-PCHIP map (Figure 6a), however, the obtainment of a credible ratio is not easy, particularly when weak analyte signal is obscured by noise and/or matrix signal. Overall the heterogeneity patterns of the local score maps of the PLS, PCA-CLS, NAS-CLS (Figure S-6, *SI*), and MCR predictions coincided reasonably well with each other (Table S-4, *SI*) with PCA-CLS being the closest match to the most accurate PLS result.

Table 4: RMSEC(RMSECV) values (in w/w%) obtained for the piracetam quantification models by MCR and PCA-CLS methods for each spectrometer channel. Model accuracy was assessed by REC% and RECV% for calibration and cross-validation respectively. The greyed-out rows show the mean and standard deviation calculated for all 10 channels.

MCR

	Model1	Model2	Model3	Model4	Model5
Piracetam in %	0–100	0–1.0	1.0–21.5	21.5–85	85–100
Chan. 1	0.96	0.044	0.21	0.84	0.38
	(1.11)	(0.060)	(0.30)	(1.08)	(0.52)
Chan. 2	0.90	0.041	0.21	0.84	0.40
	(1.03)	(0.060)	(0.31)	(1.08)	(0.63)
Chan. 3	0.86	0.044	0.20	0.75	0.39
	(1.00)	(0.060)	(0.27)	(0.97)	(0.65)
Chan. 4	0.89	0.04	0.22	0.79	0.43
	(1.03)	(0.07)	(0.30)	(1.02)	(0.66)
Chan. 5	**0.91**	**0.033**	**0.21**	**0.83**	**0.43**
	(1.05)	**(0.045)**	**(0.30)**	**(1.07)**	**(0.65)**
Chan. 6	0.90	0.05	0.22	0.81	0.40
	(1.03)	(0.07)	(0.30)	(1.04)	(0.62)
Chan. 7	0.89	0.04	0.22	0.81	0.42
	(1.03)	(0.060)	(0.30)	(1.03)	(0.66)
Chan. 8	0.93	0.05	0.23	0.86	0.41
	(1.06)	(0.07)	(0.34)	(1.11)	(0.63)
Chan. 9	0.91	0.05	0.23	0.82	0.47
	(1.05)	(0.08)	(0.33)	(1.05)	(0.74)
Chan. 10	0.90	0.044	0.22	0.82	0.44
	(1.03)	(0.063)	(0.32)	(1.07)	(0.71)
mean	0.91	0.043	0.22	0.81	0.42
	(1.04)	(0.063)	(0.31)	(1.05)	(0.65)
Std. dev.	0.03	0.005	0.010	0.03	0.03
	(0.03)	(0.009)	(0.02)	(0.04)	(0.06)
REC%/	3.24	8.77	2.39	1.72	0.44
RECV%	(3.72)	(12.95)	(3.39)	(2.21)	(0.69)

PCA-CLS

	Model1	Model2	Model3	Model4	Model5
Pira. w/w %	0–100	0–1.0	1.0–21.5	21.5–85	85–100
Chan. 1	0.55	0.04	0.08	0.51	0.46
	(0.79)	(0.05)	(0.15)	(0.66)	(0.64)
Chan. 2	0.52	0.031	0.08	0.51	0.42
	(0.71)	(0.046)	(0.18)	(0.71)	(0.60)
Chan. 3	0.47	0.026	0.04	0.42	0.38
	(0.68)	(0.041)	(0.08)	(0.64)	(0.53)
Chan. 4	0.52	0.032	0.07	0.46	0.37
	(0.73)	(0.050)	(0.13)	(0.68)	(0.51)
Chan. 5	**0.52**	**0.023**	**0.06**	**0.48**	**0.38**
	(0.72)	**(0.038)**	**(0.12)**	**(0.69)**	**(0.52)**
Chan. 6	0.52	0.030	0.05	0.46	0.35
	(0.70)	(0.058)	(0.10)	(0.65)	(0.49)
Chan. 7	0.52	0.026	0.06	0.46	0.38
	(0.72)	(0.055)	(0.11)	(0.66)	(0.52)
Chan. 8	0.54	0.037	0.07	0.47	0.39
	(0.75)	(0.054)	(0.16)	(0.73)	(0.55)
Chan. 9	0.54	0.042	0.10	0.47	0.35
	(0.76)	(0.061)	(0.18)	(0.70)	(0.61)
Chan. 10	0.47	0.040	0.08	0.32	0.37
	(0.65)	(0.062)	(0.20)	(0.53)	(0.54)

View Article Online
DOI: 10.1039/C7AY01778B

mean	0.52	0.032	0.07	0.46	0.39
	(0.72)	(0.052)	(0.14)	(0.66)	(0.55)
Std. dev.	0.03	0.006	0.02	0.06	0.03
	(0.04)	(0.008)	(0.04)	(0.06)	(0.05)
REC%	1.85	6.66	0.74	0.96	0.41
RECV%	(2.57)	(10.64)	(1.55)	(1.40)	(0.58)

Table 5: Prediction of piracetam content in the full 0.103% piracetam sample and at a single specific pixel, obtained from all ten spectrometer channels by five quantification methods. Statistical prediction of piracetam content in all the powder mixture samples and comparison of computational time required to undertake prediction of 69 mapping experiments. The computational process was fully automated from spectral pre-processing to the production of the predicted values.

	PLS	NAS-CLS	MCR	PCA-CLS	BR-PCHIP*
1.0–100% piracetam					
RMSEP	0.58%	0.40%	0.65%	0.39%	0.37%
REP%	0.93	0.65	1.04	0.62	0.81
R^2	1.00	1.00	1.00	1.00	1.00
0–1.0% piracetam					
RMSEP	***0.01%***	0.06%	0.04%	0.03%	0.17%
REP%	***2.81***	12.80	9.13	6.97	35.74
R^2	1.00	0.96	0.98	0.99	0.93
0.103% piracetam sample					
map	0.105%	0.170%	0.113%	0.160%	0.352%
pixel	0.011%	0.109%	0.146%	0.134%	0.213%
Computational time					
Time (min)	23.5	18.6	14.4	15.3	14.2

* BR-PCHIP could quantify piracetam content in the 1.0–97.0% range.

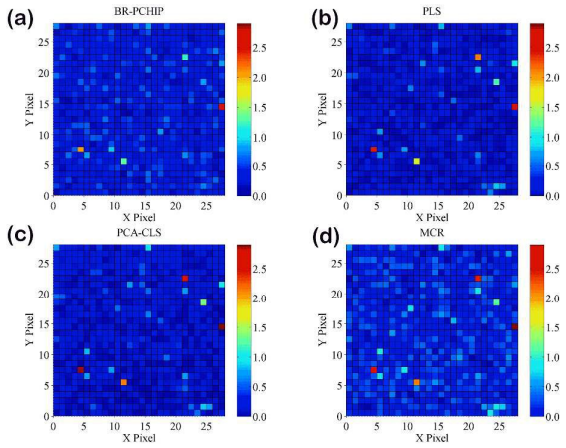


Figure 6: Piracetam distribution maps (0.103% piracetam powder mixture) generated from the prediction scores from: (a) BR-PCHIP, (b) PLS, (c) PCA-CLS, and (d) MCR. Colour bars represent piracetam content in w/w%. Five hotspots clearly have higher local piracetam concentrations, while another 12–20 pixels were identified as having slightly elevated concentrations. Statistical analysis of the predictions at all 841 pixels indicated that no pixel contained pure piracetam (Figure 5a).

Conclusions

For high analyte concentrations (>1% w/w) all methods (PLS, NAS-CLS, MCR-ALS, PCA-CLS, BR-PCHIP) were able to quantify accurately piracetam in this simple binary mixture. However, for LCQ only but both MCR and PCA-CLS showed significant accuracy (REP<10%). However, neither was as good as PLS which had an REP<5%. For practical LCQ implementation by Raman, MCR or NAS-CLS methods need to be used to minimize the requirement for large calibration sample set numbers, particularly for multi-component (n>4) materials which are more representative of pharmaceutical formulations. Here with a simple binary mixture, reducing sample numbers in the models did not show any advantages for MCR/NAS-CLS over PLS methods. Unfortunately, MCR and NAS-CLS accuracy, while sufficient for rapid semi-quantitative screening needs, were more sensitive to spectral/measurement noise than PLS. It was clear that for these methods to attain a similar level of accuracy then we would have to improve the Raman SNR and there are three approaches to achieving this: better spectral data, better hardware, and brute force. To experimentally reduce the noise (shot & dark) and get better data the Raman spectra of low-content samples should be generated from averaged spectra (*n*=4 or more) collected at each pixel in the map. For the Raman hardware increased throughput would enable shorter exposure time while the use of lower dark count, cooled detectors would reduce the dark count contribution. The final approach, is to collect higher resolution maps and increase the number of counts in the prediction histograms. However, advances in photonics technologies are generating various new approaches to Raman imaging some of which can enable very large Raman map data to be acquired in minutes[43] and this will enable these LCQ methodologies to be more efficiently and practically implemented..

Acknowledgements

Research was undertaken as part of the Synthesis and Solid State Pharmaceutical Centre, funded by Science Foundation Ireland (Grant No: 12/RC/2275). Kaiser Optical Systems, Inc. (Ann Arbor, MI) and Mr. Harry Owen are thanked for the loan of the Raman instrumentation.

Notes and references

Electronic Supplementary Information (ESI) available, See DOI: 10.1039/x0xx00000x.

1. B. Li, A. Calvet, Y. Casamayou-Boucau, C. Morris and A. G. Ryder, *Anal Chem*, 2015, 87, 3419-3428.
2. P. J. Treado and M. D. Morris, in *Practical Spectroscopy Series; Microscopic and spectroscopic imaging of the*

View Article Online
DOI: 10.1039/C7AY01778B

chemical state, ed. M. D. Morris, 1993, vol. 16, pp. 71-108.
3. M. D. Schaeberle, H. R. Morris, J. F. Turner and P. J. Treado, *Anal. Chem.*, 1999, 71, 175A-181A.
4. A. de Juan, R. Tauler, R. Dyson, C. Marcolli, M. Rault and M. Maeder, *Trac-Trends in Analytical Chemistry*, 2004, 23, 70-79.
5. G. Reich, *Advanced Drug Delivery Reviews*, 2005, 57, 1109-1143.
6. A. A. Gowen, C. P. O'Donnell, P. J. Cullen and S. E. Bell, *European Journal of Pharmaceutics and Biopharmaceutics*, 2008, 69, 10-22.
7. K. C. Gordon and C. M. McGoverin, *Int. J. Pharm.*, 2011, 417, 151-162.
8. L. Zhang, M. J. Henson and S. S. Sekulic, *Anal. Chim. Acta*, 2005, 545, 262-278.
9. M. J. Henson and L. Zhang, *Appl. Spectrosc.*, 2006, 60, 1247-1255.
10. S. Sasic, *Pharm. Res.*, 2007, 24, 58-65.
11. S. Sasic, *Anal. Chim. Acta*, 2008, 611, 73-79.
12. S. Sasic and S. Mehrens, *Anal. Chem.*, 2012, 84, 1019-1025.
13. G. Fevotte, *Chem. Eng. Res. Des.*, 2007, 85, 906-920.
14. S. Sasic and M. Whitlock, *Appl. Spectrosc.*, 2008, 62, 916-921.
15. J. A. Spencer, J. F. Kauffman, J. C. Reepmeyer, C. M. Gryniewicz, W. Ye, D. Y. Toler, L. F. Buhse and B. J. Westenberger, *J. Pharm. Sci.*, 2009, 98, 3540-3547.
16. H. S. Lin, O. Marjanovic, B. Lennox, S. Sasic and I. M. Clegg, *Appl. Spectrosc.*, 2012, 66, 272-281.
17. C. Gendrin, Y. Roggo and C. Collet, *J. Pharm. Biomed. Anal.*, 2008, 48, 533-553.
18. A. Farkas, B. Vajna, P. L. Soti, Z. K. Nagy, H. Pataki, F. Van der Gucht and G. Marosi, *Journal of Raman Spectroscopy*, 2015, 46, 566-576.
19. H. Shinzawa, K. Awa, W. Kanematsu and Y. Ozaki, *J Raman Spectrosc*, 2009, 40, 1720-1725.
20. M. Boiret, N. Gorretta, Y. M. Ginot and J. M. Roger, *J. Pharm. Biomed. Anal.*, 2016, 120, 342-351.
21. M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad and M. Akhond, *J. Chemom.*, 2006, 20, 146-157.
22. F. Allegrini and A. C. Olivieri, *Anal. Chim. Acta*, 2011, 699, 18-25.
23. Z. Li, D. J. Zhan, J. J. Wang, J. Huang, Q. S. Xu, Z. M. Zhang, Y. B. Zheng, Y. Z. Liang and H. Wang, *Analyst*, 2013, 138, 4483-4492.
24. B. Li, A. Calvet, Y. Casamayou-Boucau and A. G. Ryder, *Anal. Chim. Acta*, 2016, 913, 111-120.
25. J. Engel, J. Gerretzen, E. Szymanska, J. J. Jansen, G. Downey, L. Blanchet and L. M. C. Buydens, *Trac-Trends in Analytical Chemistry*, 2013, 50, 96-106.
26. M. Vidal and J. M. Amigo, *Chemom. Intell. Lab. Syst.*, 2012, 117, 138-148.
27. T. Fearn, C. Riccioli, A. Garrido-Varo and J. E. Guerrero-Ginel, *Chemom. Intell. Lab. Syst.*, 2009, 96, 22-26.
28. S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, 58, 109.
29. P. Geladi, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2003, 58, 767-782.
30. H. Grahn and P. Geladi, *Techniques and Applications of Hyperspectral Image Analysis*, John Wiley & Sons Ltd., Chichester, 2007.
31. B. Vajna, I. Farkas, A. Szabo, Z. Zsigmond and G. Marosi, *J. Pharm. Biomed. Anal.*, 2010, 51, 30-38.
32. J. M. Amigo and C. Ravn, *Eur J Pharm Sci*, 2009, 37, 76-82.
33. W. F. de Carvalho Rocha, G. P. Sabin, P. H. Marco and R. J. Poppi, *Chemom. Intell. Lab. Syst.*, 2011, 106, 198-204.
34. T. T. Lied, P. Geladi and K. H. Esbensen, *J. Chemom.*, 2000, 14, 585-598.
35. A. de Juan and R. Tauler, *Critical Reviews in Analytical Chemistry*, 2006, 36, 163-176.
36. R. Tauler, *J. Chemom.*, 2001, 15, 627-646.
37. J. Jaumot and R. Tauler, *Chemom. Intell. Lab. Syst.*, 2010, 103, 96-107.
38. D. M. Haaland and E. V. Thomas, *Anal Chem*, 1988, 60, 1193-1202.
39. H. A. Martens and P. Dardenne, *Chemom. Intell. Lab. Syst.*, 1998, 44, 99-121.
40. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis.* , Pearson Prentice Hall, New Jersey, 6th edn., 2008.
41. M. C. Antunes, J. E. J. Simao, A. C. Duarte and R. Tauler, *Analyst*, 2002, 127, 809-817.
42. L. B. Lyndgaard, F. van den Berg and A. de Juan, *Chemom. Intell. Lab. Syst.*, 2013, 125, 58-66.
43. J. Qi, J. T. Li and W. C. Shih, *Biomedical Optics Express*, 2013, 4, 2376-2382.