Title	Kernel principal component analysis residual diagnosis (KPCARD): an automated method to remove cosmic ray artefacts in Raman spectra.
Author(s)	Li, Boyan; Calvet, Amandine; Casamayou-Boucau, Yannick; Ryder, Alan G.
Publication Date	2016
Publication Information	Li, B., Calvet, A., Casamayou-Boucau, Y., Ryder, A.G. (2016) 'Kernel principal component analysis residual diagnosis (KPCARD): an automated method to remove cosmic ray artefacts in Raman spectra'. Analytica Chimica Acta, 913 :111-120.
Link to publisher's version	http://dx.doi.org/10.1016/j.aca.2016.01.042
Item record	http://hdl.handle.net/10379/5598
DOI	http://dx.doi.org/10.1016/j.aca.2016.01.042

Kernel principal component analysis residual diagnosis (KPCARD): an automated method to remove cosmic ray artefacts in Raman spectra. B. Li, A. Calvet, Y. Casamayou-Boucau, A.G. Ryder, *Analytica Chimica Acta*., 913, (111-120), (2016). DOI: 10.1016/j.aca.2016.01.042

Kernel principal component analysis residual diagnosis (KPCARD): *an automated method for cosmic ray artifact removal in Raman spectra.*

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

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

* Corresponding author. **Tel:** +353 9149 2943 **Email:** alan.ryder@nuigalway.ie

Abstract:

A new, fully automated, rapid method, referred to as *kernel principal component analysis residual diagnosis (KPCARD)*, is proposed for removing with cosmic ray artifacts (CRAs) in Raman spectra, and in particular for large Raman imaging datasets. KPCARD identifies CRAs via a statistical analysis of the residuals obtained at each wavenumber in the spectra. The method utilizes the stochastic nature of CRAs; therefore, the most significant components in principal component analysis (PCA) of large numbers of Raman spectra should not contain any CRAs. The process worked by first implementing kernel PCA (kPCA) on all the Raman mapping data and second accurately estimating the inter- and intra-spectrum noise to generate two threshold values. CRA identification was then achieved by using the threshold values to evaluate the residuals for each spectrum and assess if a CRA was present.

CRA correction was achieved by spectral replacement where, the nearest neighbor (NN) spectrum, most spectroscopically similar to the CRA contaminated spectrum and principal components (PCs) obtained by kPCA were both used to generate a robust, best curve fit to the CRA contaminated spectrum. This best fit spectrum then replaced the CRA contaminated spectrum in the dataset. KPCARD efficacy was demonstrated by using simulated data and real Raman spectra collected from solid-state materials. The results show that KPCARD is fast (<1

min per 8400 spectra), accurate, and precise and suitable for the automated correction of very large (>1 million) Raman datasets.

KEYWORDS: Cosmic ray artifacts; Raman spectroscopy, Correction, Kernel PCA, quantitative.

1. Introduction

Raman spectroscopy is widely used spectroscopic for the analysis of chemically complex materials in many areas, and in particular the pharmaceutical industry [1-6]. The use of Raman imaging/mapping spectroscopy provides micron-scale spatial information about materials, and is popular for the identification of active pharmaceutical ingredients (APIs), detection of contaminants/impurities, and mapping the distribution of components in solid-state materials [3, 5, 7-14]. However, charge-coupled device (CCD) detectors used in Raman spectroscopy are sensitive to cosmic ray events, which generate occasional, positive, unidirectional, erroneous spikes in Raman spectra. The frequency and location of these cosmic ray artifacts (CRAs) is random, and peak intensity and width can vary very significantly. CRAs increase unwanted signal variance, distort spectra when overlapped with Raman bands, thus complicate signal interpretation, and degrade chemometric modelling accuracy [7, 12, 15].

There are many methods available for dealing with CRAs, and in general, one of four approaches is used. The first is a replicate measurement approach, acquiring additional spectra for each sample (or grid point in a map) and then discarding CRA contaminated spectra via manual or automated assessment [7]. This approach is based on the fact that the probability of CRAs appearing at the same positions in two sequential spectra is low and thus using spectral comparison the CRA can be identified and eliminated [16, 17]. However, when dealing with large numbers of spectra in Raman mapping this would impose an unsustainable increase in spectral acquisition time. The second category involves optical hardware design, for example by having a straight slit image on the CCD one can compare the spectra from multiple strips to identify and then automatically correct for CRAs in a single acquisition [18]. However, this may not be available on all models of Raman spectrometer. User set parameter-based methods are the third category, which include both threshold- and filter-based methods [12, 19-27].

Implementation requires user intervention, and method performance in terms of CRA removal depends significantly on the specification of critical thresholds and/or filter size. Improper parameter choice may result either in incomplete CRA removal or spectral signal distortion and there is a time penalty associated with the optimization process. The fourth category follows a fully automated approach, which aims to remove subjective user input and improve productivity when dealing with large datasets. The only existing method was two-dimensionally coincident second difference cosmic ray spike removal (2DCDR) [15].

The need for a simple, fully automated method that could generate minimal spectral distortion motivated us to develop this novel CRA removal method. KPCARD first used the stochastic nature of CRAs coupled with accurate noise measurements available from the large Raman mapping datasets and the residuals from kPCA for identification. Then the most similar NN spectrum and the principal components from kPCA were used to fit and replace CRA contaminated spectra. Here we quantify the benefits of CRA removal by KPCARD in terms of effects on spectral variance and model accuracy.

2. Materials and Methods

2.1 Materials: Piracetam (2-oxo-1-pyrrolidineacetamide, polymorphic form III), L-proline (≥99%), L-tyrosine (≥98%) and L-cysteine·HCl·H_2O (≥98%) were purchased from Sigma-Aldrich (Ireland) and used as received. 50 binary powder mixtures of piracetam and proline were prepared with 0 to 100% (w/w%) piracetam content as previously described [7].

2.2 Raman instrumentation: Raman spectra (180–1896 cm^{-1}/2 cm^{-1} resolution) of tyrosine and cysteine were measured using a 785 nm excitation Avalon Instruments RamanStation spectrometer with a laser power of ~70 mW at the sample. Each sample was scanned over a 5×5 grid (0.2 mm spacing), and at each grid-point a 3×10 second exposure was accumulated. These were then averaged (3×25 spectra) to give the low noise spectra for building the simulated data. Raman spectra of the piracetam/proline mixtures were obtained with a RAMAN WORKSTATION™ Analyzer (Kaiser Optical Systems, Inc.), with 785 nm excitation and PhAT imaging capability [7]. Raman spectra (200–1896 cm^{-1}) were collected from on each of 10 channels using a 29×29 pixel grid with 1 mm spacing and a 1-second exposure. Each sample map dataset consisted of 8410 spectra generated by the 10 channels.

2.3 Simulated data & data analysis: To test efficacy, synthetic Raman datasets were created as follows: *(1)* Raman spectra of pure L-tyrosine and L-cysteine·HCl·H_2O were superimposed in random proportions to create 500 simulated mixture spectra with randomly generated concentrations. Each simulated spectrum had 859 wavenumber channels (180 to 1896 cm^{-1}) with a 2 cm^{-1} interval; *(2)* Gaussian high-frequency, white noise at five levels of 0, 0.001, 0.005, 0.01 and 0.02 (*defined as the ratio of the standard deviation of the white noise to the highest Raman spectral signal*), were randomly generated and then added into the data; *(3)* Finally, a subset of 30 CRA contaminated spectra was generated with 54 spikes, the remaining 470 spectra were CRA-free. These spikes were randomly assigned different wavenumbers and intensities, and given a random width of between 1 and 8 pixels, which was either smaller than or comparable to the bandwidths of the Raman bands of tyrosine and/or cysteine (the MATLAB code and process description is provided in the *supplemental information, SI*, S4). Then, CRAs were superimposed on the simulated spectra, with varying levels of added noise. MATLAB R2014b (The MathWorks Inc., Natick, MA) with in-house written code for data processing which was carried out on a standard desktop computer: Microsoft Windows 7 OS, Xeon 2.8 GHz CPU, and 6 GB RAM.

3. Methodology

Conventional notation was adopted throughout this paper: Uppercase boldface letters for matrices (as **X**), lowercase boldface for vectors (as **x**), italicized subscript characters for vector index (as $\mathbf{x}_i$ or $\mathbf{e}_j$), and lowercase italicized letters for scalars (as e_{ij} or n_{ij_orig}). Superscripts are assigned as follows: T, vector or matrix transpose; and –1, matrix inverse. Principal component analysis (PCA) is one of the most important techniques in multivariate data analysis [28]. PCA can be expressed in terms of a product of two matrices **U** and **V**, and a residual matrix **E**:

$$\mathbf{X} = \mathbf{U}\mathbf{V}^{T} + \mathbf{E} \qquad (1)$$

where the column eigenvector matrix **V** (loading matrix) of the cross product matrix $\mathbf{X}^{T}\mathbf{X}$ holds the latent variables (or PC factors), along the wavenumber/wavelength axis. **U** (Score matrix) represents the sample/object distribution pattern in **X**, which corresponds to the eigenvector matrix of the cross product matrix $\mathbf{X}\mathbf{X}^{T}$. **V** provides chemically or physically meaningful interpretation to the patterns observed in **U**. The residual matrix **E** is the portion of the data not explained by the PC factors used for the data representation, and this comprises elements such as

noise, experimental errors, and uncertainties. **V** can be obtained by a variety of PCA algorithms: singular value decomposition (SVD), eigenvalue decomposition (EVD), NIPALS, and POWER methods [29]. When **X** is very large, the EVD algorithm is preferred as it is much faster. Obtaining **V** is crucial because **V** can be used to obtain **U**, and calculate the scores of new objects:

$$\mathbf{U} = \mathbf{XV} \quad (2)$$

$$\mathbf{U}_{new} = \mathbf{X}_{new}\mathbf{V} \quad (3)$$

However, when *n* (variables) >>*m* (spectra) in **X**, the eigenvector is first estimated through **U** from $\mathbf{XX}^T$ instead of $\mathbf{X}^T\mathbf{X}$, and then **V** is indirectly obtained.

$$\mathbf{V} = \mathbf{X}^T\mathbf{U}\mathbf{\Gamma}^{-1/2} \quad (4)$$

where, $\mathbf{\Gamma}$ is the diagonal matrix of eigenvalues of $\mathbf{XX}^T$ and $\mathbf{XX}^T$ is termed a kernel matrix. This approach is much faster than trying to directly obtain **V** from $\mathbf{X}^T\mathbf{X}$ and thus less computational effort is needed [30, 31].

If *r* factors adequately represent the spectral features in **X**, then the residual **E** should consist only of white Gaussian noise ($\mathbf{E}_{noise\_w}$), heteroscedastic noise ($\mathbf{E}_{noise\_hetero}$), and CRAs:

$$\mathbf{E} = \mathbf{E}_{noise\_w} + \mathbf{E}_{noise\_hetero} + \mathbf{E}_{CRA} \quad (5)$$

If *m* spectra are measured under identical conditions it is then often assumed that the high frequency noise is normally or approximately normally distributed, uncorrelated, and random. In contrast, CRAs by their nature are stochastic in terms of frequency of occurrence, location, intensity, and peak width. If noise (*i.e.*, $\mathbf{E}_{noise\_w} + \mathbf{E}_{noise\_hetero}$) can be accurately estimated, then CRAs can be discriminated, and the challenge was therefore how to accurately measure the true noise level. This involves first estimating the noise at a given wavenumber across all the different spectra, and second recognizing that noise varies with spectrum wavelength due to instrumental effects. Once this was done, two threshold values (t_1, t_2) were generated for each point in the spectrum, and used to identify CRAs.

The upper limit of the confidence level of the distributed residue $\boldsymbol{e}_j$ of *m* objects at the *j*th wavenumber was calculated using accepted statistical procedures (α=0.01 or 0.05), according to:

$$P\,(cl_{lower\_j} \le e_{ij} \le cl_{upper\_j}) = 1-\alpha,\ (i=1, 2, \ldots, m, \text{ and } j=1, 2, \ldots, n) \quad (6)$$

where e_{ij} is the element of $\mathbf{e}_j$, cl_{upper_j} and cl_{lower_j} were the upper and lower limits, respectively, when the confidence interval (1−α) was applied. The first threshold (t_1) was obtained from:

$$t_{1\_j} = \lambda_{1\_j} \cdot cl_{upper\_j} \quad (7)$$

where λ_{1_j} was a penalty parameter required to ensure CRA detection sensitivity yet preserved the real Raman band intensity variation and thus avoid misidentifying real Raman bands as CRAs. This was analogous to the definition of limit of detection (LOD) where the analyte signal was at least three times greater than background noise, and in that simple case, λ_{1_j} =3. However, for Raman spectroscopy where heteroscedastic noise is present (due to varying wavelength sensitivity of CCD detectors), the λ_{1_j} value has to be adjustable, and here we used a value of 60. This value is instrumentation and spectral noise level dependent.

The idea of indirectly estimating noise in a spectrum is not new, but the ability to deal automatically and efficiently with very large numbers of Raman spectra is innovative [15, 32]. The method principle is briefly described here. The noise-only variance of an original spectrum was equal to the variance of its second difference spectrum divided by six, if no band residuals or other artefacts were present in the second difference spectrum:

$$n_{i\_orig}^2 = n_{i\_2diff}^2 / 6, \qquad (i = 1, 2, ..., m) \qquad (8)$$

where n_{i_orig} and n_{i_2diff} denoted the noise level in the ith original spectrum and its second difference spectrum, respectively. The second difference spectrum was obtained by first shifting the original spectrum by one channel to a greater index number and wraparound of the end, second by subtracting the original from the shifted spectrum to produce the first difference spectrum, and then repeating the first and second steps on the first difference spectrum to produce the second difference spectrum.

To determine the noise, n_{i_2diff}, the second difference spectrum was segmented into many windows with a defined window size (30 channels were used here). Note that the n_{i_2diff} noise estimate was not related to the window size chosen, however, the window size should be smaller than a spectral region (SR_{no}) where there were no signal bands present in the original spectrum, *i.e*. noise only region. Then, the noise was individually calculated for each window and the window with the minimum noise level ($noise_{ref}$) was found, and this should correspond to SR_{no}. Starting with the SR_{no} window and moving only one channel (*called adjacent channel*) to a higher index number along the wavenumber axis, a new window of the same size was created and the noise level, $noise_{new}$, estimated for this new window. If $noise_{new}$ was $> \sqrt{6}$ times larger than $noise_{ref}$, then it indicated that either a CRA or Raman band residual was present at the adjacent channel in the second difference spectrum, and this fact was recorded. Then, the value

of the adjacent channel was replaced with the mean value of this new window and the noise level recalculated, giving a new $noise_{ref}$. The window was advanced to the next channel and the procedure repeated, until the last channel had been analyzed.

Once completed, the procedure was repeated in the opposite direction, moving from the minimum noise window to a lower index number. The combination of both runs generated a continuous noise estimate along the wavenumber axis (j=1, 2, …, n) for each spectrum, *i.e.*, n_{ij_2diff} for the ith (i=1, 2, …, m) second difference spectrum and n_{ij_orig} for the original spectrum according to Equation (8). This noise estimate explicitly accounted for any heteroscedastic noise related such as wavelength-dependent detector performance, and was used to define the second threshold:

$$t_{2\_ij} = \lambda_{2\_ij} \cdot n_{ij\_orig} \quad (9)$$

The penalty factor λ_{2_ij} was set to 3, again analogous to the LOD definition, but the value can be varied where necessary. Finally, when the two thresholds were applied to each element of the residual **E** for the spectra in Equations (1) and (5), and if:

$$e_{ij} > \min(t_{1\_j}, t_{2\_ij}) \quad (10)$$

then a CRA was present, otherwise e_{ij} represented noise. The noise estimate and CRA identification were implemented automatically using in-house written MATLAB codes. When this was applied to Raman data, the output was a list of CRA contaminated spectra and the wavenumber locations of the artefacts. CRA contaminated spectra can be discarded and this may be acceptable if there are very large datasets and the target analytes are present in relatively high concentrations. However, for low-concentration quantification applications [7], this may not always be ideal and it is more advisable to use a correction procedure, here we implemented a chemometric curve fitting and approximation based approach.

The CRA contaminated spectrum $\mathbf{x}_i$ was first compared with spectra in the adjoining pixels (here a 3×3 pixel neighborhood) using the NN comparison method [12, 33] and the most similar spectrum ($\mathbf{x}_{nn}$) was selected. The rationale for this was based on the fact that the probability of a CRA appearing at the exact same wavenumber in a spatially adjacent spectrum is low. The NN spectrum ($\mathbf{x}_{nn}$) and the r PC factors were then combined to fit the CRA contaminated spectrum $\mathbf{x}_i$ using a linear model:

$$\mathbf{x}_i = a\mathbf{x}_{nn} + b\mathbf{1} + \sum_{k=1}^{r} c_k \mathbf{v}_k + \mathbf{e}_i \qquad (11)$$

The scaling a and offset b factors accounted for spectral intensity variations between $\mathbf{x}_i$ and $\mathbf{x}_{nn}$. Vector $\mathbf{v}_k$ and scaling factor c_k corresponded to the kth PC of the loading $\mathbf{V}$ and its weight in the fit respectively. The error $\mathbf{e}_i$ included the deviation of the linear fitting from the true relationship and contribution from spectral noise sources. Finally, a best-fit spectrum ($\mathbf{x}_{ri}$) was generated by:

$$\mathbf{x}_{ri} = a\mathbf{x}_{nn} + b\mathbf{1} + \sum_{k=1}^{r} c_k \mathbf{v}_k \qquad (12)$$

It then replaced the CRA contaminated spectrum in the dataset used for chemometric modelling. KPCARD can be automatically implemented using MATLAB and the only prerequisite was that all the spectral features of Raman data were extracted into the lowest number of PCA factors and that no spectral information was retained in the residuals. A flowchart illustrating all the steps in the KPCARD method is shown in Figure 1a.

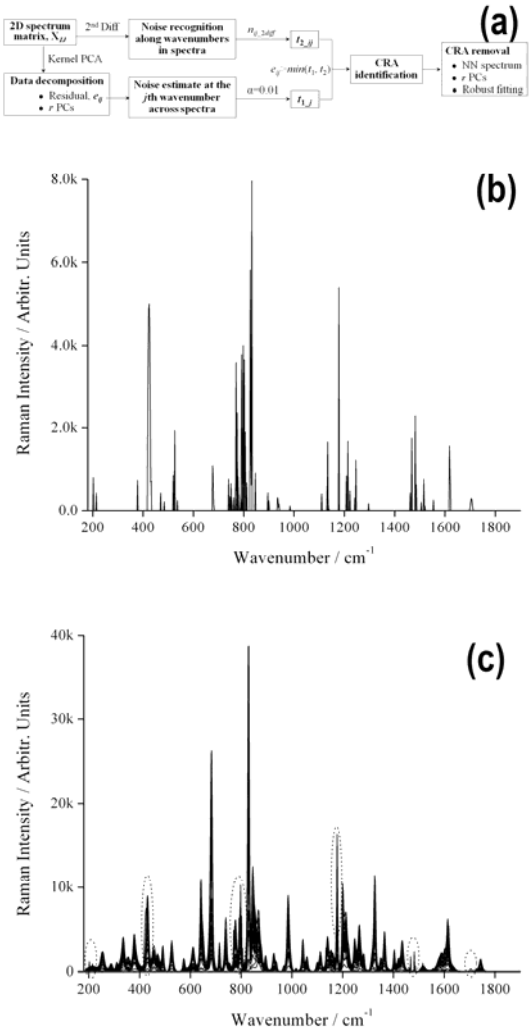

Figure 1: (**a**) Flowchart showing the implementation steps for the KPCARD method; (**b**) The synthetic CRAs that were added to the simulated Raman spectra, and (**c**) Superimposed spectra of tyrosine and cysteine without any additional noise. Dotted ellipses highlight specific CRAs.

4. Results and Discussion

To demonstrate KPCARD method efficacy, the simulated data with known, added synthetic CRAs, was first used to quantify the negative effects of CRA contamination on partial least-squares (PLS) based quantification, second to test method efficacy on datasets with varying levels of noise, and third to compare KPCARD with other CRA correction methods.

4.1 CRA contamination effect: To investigate the detrimental effect of CRAs on model accuracy, the simulated spectra with five different added levels of noise were used for tyrosine quantification. The 54 synthetic CRAs (Figure 1b) had variable locations, bandwidths, and intensities. Most were one pixel wide, but two were eight pixels wide (Table S-1, *SI*). These were added randomly into 30 simulated spectra (Figure 1c) and the spectra with and without CRAs, were individually subjected to PLS regression using one factor, mean-centering pretreatment, and leave-one-out cross-validation [34]. Model accuracy was assessed using root mean square error of calibration (RMSEC), root mean square error from cross-validation (RMSECV), square of the correlation coefficient (R^2) between predicted and measured tyrosine content, and the percent variance of spectral data (%**X**) and tyrosine content (%**y**) captured by the model (Table 1). CRA contamination degraded PLS model accuracy whereas noise had little effect because PLS *per se* could filter noise to some extent. For example in the 0.005 noise data, RMSEC/RMSECV values for the spectra in the absence of CRAs were both rather small, 0.013/0.019. After CRA superimposition, RMSEC and RMSECV deteriorated, increasing by 31% and 35%, while R^2 and data percent variance captured by the model decreased (74.4% to 74.1%). This shows that CRAs have a significant effect on quantification accuracy and the issue becomes more significant as the target analyte concentration ranges decrease to the low-content regimes below ~0.5% w/w [7].

Table 1: Summary of PLS models obtained using 30 simulated spectra for investigating the CRA contamination and noise effect on model accuracy for the quantification of tyrosine.

Noise level	**CRAs in spectra**	**RMSEC**	**RMSECV**	R^2	**Percent variance captured**	
					%X	%y
0.000	absent	0.013	0.019	0.998	74.39	99.81
	present	0.017	0.023	0.997	73.85	99.66
0.001	absent	0.013	0.019	0.998	74.39	99.81
	present	0.017	0.023	0.997	73.86	99.66
0.005	***absent***	***0.013***	***0.019***	***0.998***	***74.37***	***99.81***
	present	***0.017***	***0.023***	***0.997***	***73.84***	***99.66***

0.010	absent	0.013	0.020	0.998	74.32	99.80
	present	0.017	0.023	0.996	73.79	99.65
0.020	absent	0.013	0.019	0.998	74.06	99.81
	present	0.017	0.023	0.997	73.53	99.66

4.2 Method performance: KPCARD performance had to be first evaluated by using the simulated Raman data described above. *Accuracy* and *precision* parameters were defined to quantify the CRA correction procedure while the *ratio of additional noise-to-simulated spectra and noise* (*RNSN)* value gives a measure of % variance of the added noise in the Raman datasets:

$$Accuracy = \frac{\text{number of CRAs correctly identified and removed}}{\text{number of total simulated CRAs}} \times 100\% \quad (13)$$

$$Precision = (1 - \frac{\text{covariance of difference spectra between simulated \& CRA removed data}}{\text{covariance of simulated data with no CRAs}}) \times 100\% \quad (14)$$

$$RNSN = \frac{\text{covariance of additional noise}}{\text{covariance of simulated spectra and additional noise}} \times 100\% \quad (15)$$

This *precision* was a measure of data distortion induced after implementation of CRA correction on contaminated spectra compared to uncontaminated spectra. A larger *precision* value equates to a lower level of spectral distortion caused by the CRA removal process.

KPCARD was implemented on the simulated datasets using two PC factors. In cases where the additional noise level was not high (0–0.005) all CRAs were correctly identified and the contaminated spectra were thus corrected (Table 2). The high *precision* obtained indicated that when CRA contaminated spectra were replaced, the real spectral features were preserved and the curve fitting process induced no significant spectral distortion. However, as noise increased, the method failed to detect all the CRAs (as expected) and correct the CRA contaminated spectra, thus *accuracy* dropped significantly, whereas *precision* did not vary much. This proves that KPCARD is able to generate a robust curve fit for the CRA contaminated spectrum correction, and that this process is not significantly affected by noise.

Table 2: Method performance for CRA identification and removal from simulated spectra with different levels of synthetic noise added.

KPCARD					
Noise level	**Spectra with CRAs #Corrected/total**	**CRAs Identified/total**	**Accuracy (%)**	**Precision (%)**	**RNSN (%)**
0.000	30/30	54/54	100	99.98	0
0.001	30/30	54/54	100	100	0.0003
0.005	30/30	54/54	100	99.99	0.0066
0.010	27/30	50/54	92.59	99.99	0.025
0.020	22/30	41/54	75.93	99.95	0.093
DTCSR (thresholds of 0.098 and 0.3)					
Noise level	**Spectra with CRAs #Corrected/total**	**CRAs Identified/total**	**Accuracy (%)**	**Precision (%)**	***CRA-free spectra modified**
0.000	29/30	28/54	51.85	99.996	0
0.001	28/30	28/54	51.85	99.996	0
0.005	29/30	29/54	53.70	99.996	28
0.010	30/30	29/54	53.70	99.996	109
0.020	30/30	31/54	57.41	99.995	378
2DCDR					
Noise level	**Spectra with CRAs #Corrected/total**	**CRAs Identified/total**	**Accuracy (%)**	**Precision (%)**	***CRA-free spectra clipped (prec.)**
0.000	27/30	28/54	51.85	99.13	348 (99.09)
0.001	27/30	4/54	7.41	99.13	348 (99.09)
0.005	27/30	4/54	7.41	99.13	354 (99.09)
0.010	27/30	3/54	5.56	99.11	391 (99.10)
0.020	30/30	1/54	1.85	99.14	448 (99.13)

* This is the number of CRA-free spectra (from the total of 470 CRA free simulated spectra) that were either modified by DTCSR or clipped by 2DCDR during the CRA removal process. The numbers in parentheses denoted the precision calculated from the clipped spectra.

These spikes, which could not be identified, were those with intensities nearly equivalent to the magnitude of the added high-frequency noise. Figure 2a shows the four CRAs not identified for the 0.01 noise level dataset (~200 arbitrary units, au, of noise); these were located at 1296, 1506, 1520, and 984 cm^{-1}, with intensities of 175, 208, 110, and 114 au, respectively (Table S-1,

SI). For the 0.02 noise level data, this increased to 13 CRAs in eight spectra, all of which had intensities close to the noise, and were thus not an issue for data analysis. Another important consideration was that the simulated spectra were generated from real Raman measurements of tyrosine and cysteine, which contained additional intrinsic shot and dark noise.

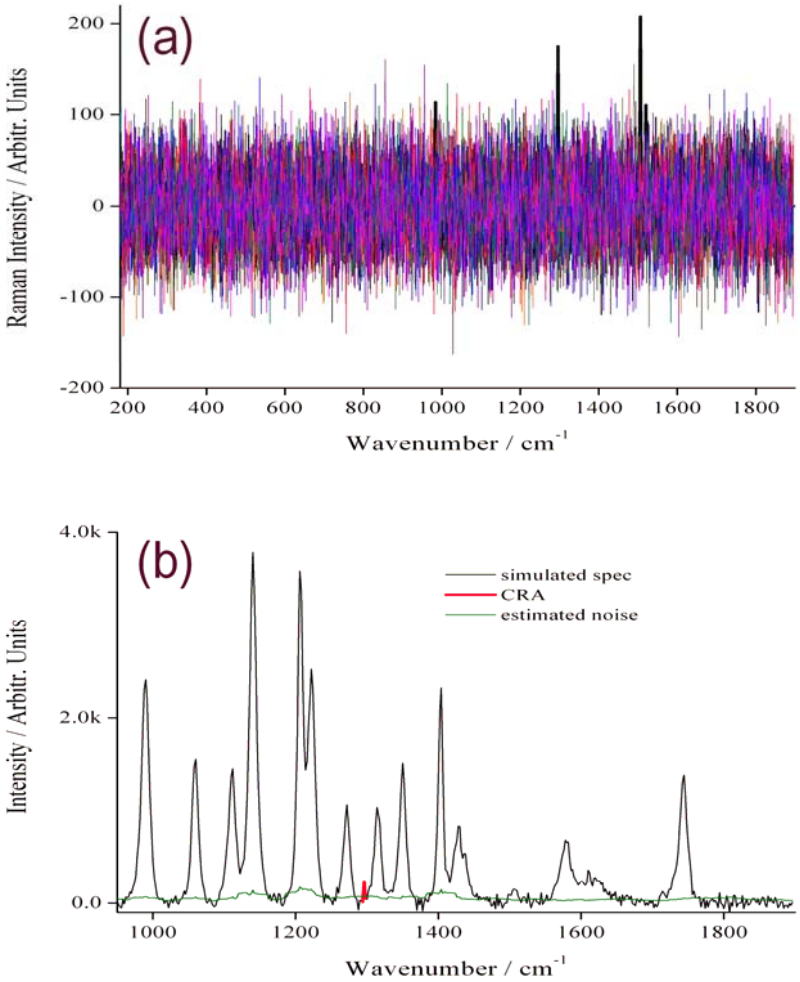

Figure 2: (**a**) Overlay plot of the four CRAs (bold spikes) not identified and the added noise (0.01 level) for the 30 simulated spectra. (**b**) Sample spectrum (black) with the 1296 cm^{-1} CRA (red) and 0.01 added noise, and the estimated noise (green).

Figure 2b shows part of the spectrum in which the undetected 1296 cm^{-1} CRA was located, and the estimated noise was 77.36 au (using a 30-channel window) for the 1296 cm^{-1} location. This comprised of 28.42 au of added noise with the remainder coming from the spectra used for

the simulated data. The CRA intensity (175 au) was less than three times the noise thus accounting for the missed identification. Figure 2b (& Fig. S-1, *SI*) also shows the estimated noise (n_{ij_orig} green curve) which is not completely flat and shows the heteroscedastic nature of the noise, and in particular, the significant shot noise contribution from strong Raman bands.

The number of kPCA factors used is crucial because it affected method sensitivity and accuracy. It is critical that the PCs account for all the spectral information relating to the analytes and that the residuals only contain noise and CRAs. Too many PCs could result in CRAs being removed from the residuals while too few PCs could result in spectral data being incorporated into the residuals leading to Raman bands being misidentified as CRAs.

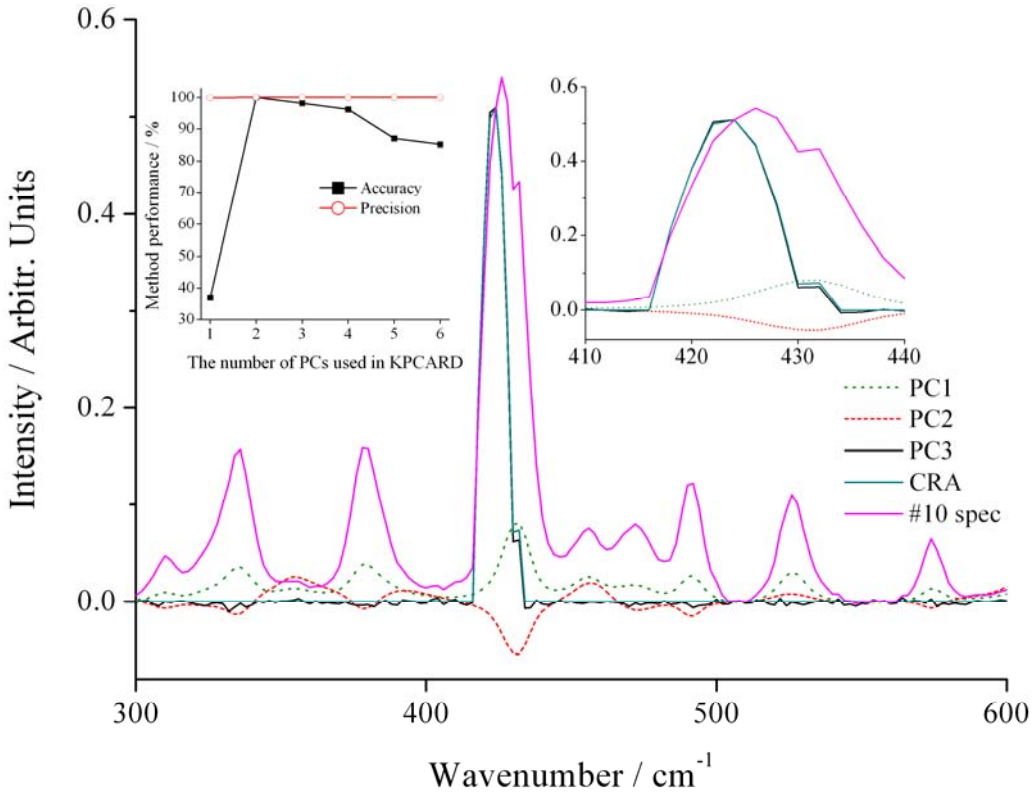

Figure 3: The first three PCs decomposed from the 500 simulated spectra, the #10 CRA contaminated spectrum, and the synthetic CRA superimposed on the spectrum. For clarity, both the #10 spectrum and CRA were scaled. (**left insert**) The effects of the number of principal components used in the KPCARD procedure on the method accuracy and precision for the 0.005 noise level simulated data set. (**right insert**) Expanded view of band.

To assess the impact of PC number we took the 0.005 noise level data (500 simulated spectra, with 30 CRA contaminated spectra having 54 CRAs) and varied kPCA factors from 1 to 6 (Figure 3). It was noted that *accuracy* relied on the number of PCs used, and that two PCs were optimal (all 54 CRAs detected and only the 30 CRA contaminated spectra were corrected). When one PC was used, it was insufficient to represent all the spectral information of two chemical species (tyrosine and cysteine) and as a consequence, some spectral features were retained in the residuals, and therefore were included in the noise estimate. This caused an increase in the two threshold (t_1/t_2) values, which led to the method being less sensitive to discriminating between weak Raman bands and CRAs. This led in turn to some weak Raman bands and/or CRAs being misidentified, and only 20 out of 54 CRAs could be properly identified, giving a low *accuracy* of 37.04%. In addition, 17 CRA-free spectra were misidentified as being contaminated.

For the three PC test (Figure 3), 53 CRAs were successfully identified, 29 spectra were corrected, and only one CRA (in spectrum #10) was missed. In that case, PC3 matched almost perfectly the synthetic CRA added to the simulated spectrum. Since this specific CRA had a wide bandwidth (eight channels) similar to a Raman band shape, the over use of one more PC (*i.e.*, PC3) meant that this CRA was misidentified as a real spectral feature, and so the #10 spectrum was not corrected.

As PC number increased to six, method performance decreased, with *accuracy* dropping from 98.15% (3 PCs), to 96.30% (4 PCs), 87.04% (5 PCs), to 85.19% (6 PCs). However, adding more PCs to the procedure caused no significant change to method *precision* and all values were very close to 100%. This indicated that once the CRAs were identified and the most similar spectra found, the curve fit method generated very robust spectrum fits, even when too many PCs were used. In practice, real sample spectra are more complex than simulated spectra, and PC number should be set to a value equal to or slightly larger than the optimal PC number, because the failure to remove a real CRA (when more PCs used) is preferred to the incorrect elimination of a real spectral band (if fewer PCs were used). There are many good and fast methods available for correctly determining PC number [35].

4.3 Method comparison: KPCARD was compared to two alternative methods from the literature in terms of CRA removal performance: dual threshold cosmic spike removal (DTCSR) [12] and

2DCDR [15]. In DTCSR, two thresholds were required: a lower threshold set by the user to ensure sensitive CRA detection and complete removal, and a higher threshold used to avoid clipping Raman bands and to preserve normal Raman band height variability between spectra. By studying a region where CRAs were clearly present, thresholds could be determined; however, this critically depended on user input. Once thresholds were set and CRAs identified, the NN comparison method was used and automatic CRA correction was then performed on the whole Raman dataset. Therefore, this approach can be considered as being semi-automated. It is worth pointing out that an important difference between DTCSR and KPCARD is that with DTCSR only the identified CRAs (and a small spectral region in the vicinity) were replaced by a curve fit, whereas with KPCARD the entire CRA contaminated spectra were replaced.

DTCSR was run on all five simulated datasets and the two thresholds (0.098 and 0.3) were selected after optimization to ensure that as many CRAs as possible were identified and removed, while minimizing artefact generation in the uncontaminated spectra (Table S-2, SI). For the higher threshold, values between 0.2 and 0.4 gave the similar CRA identification results. The method accuracy and precision were then calculated (Table 2). The performance of this parametric method depended mainly on determining an optimized set of thresholds, which can take several iterations (Table S-2, SI). Furthermore, the noise level also influenced DTCSR performance, but, the resultant *precision* was high (99.996%), as this was intrinsically related to the robust curve fit method used.

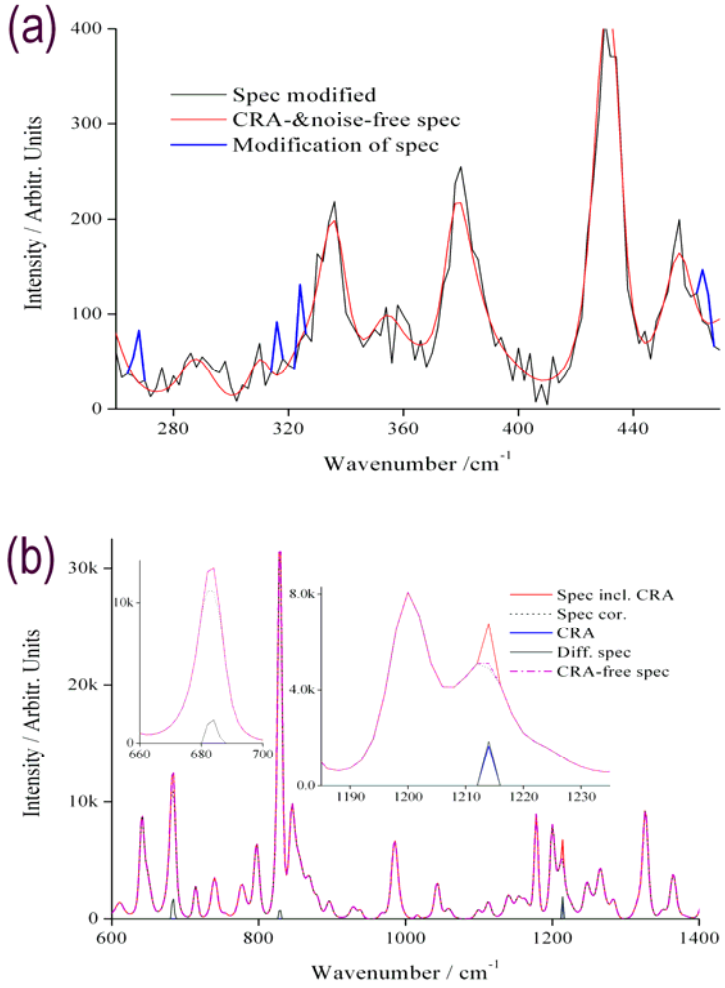

Figure 4: (**a**) Spectra showing modification of spectral noise of a CRA-free spectrum when an overly sensitive lower threshold was used for DTCSR. Red (CRA-free spectrum, no added noise), Black (CRA-free spectrum with added noise and after modification), Blue (Noise features truncated by DTCSR modification). (**b**) Spectra showing the clipping of two Raman bands (~684 and ~830 cm^{-1}) and CRA not fully removed by 2DCDR (0.001 noise level case). Red solid, black dotted, magenta, blue, and dark green curves respectively represent the CRA-included spectrum, the CRA-corrected spectrum, the uncontaminated spectrum, the synthetic CRA, and the difference spectrum between the CRA-included and CRA-corrected spectra.

If the noise level was low and the lower threshold was not small enough, then the weaker CRAs were not detected. As noise increased, the small lower threshold caused some noise to be incorrectly identified as CRAs, which led to more CRA-free spectra being modified: 109/470 for the 0.01 noise and 378/470 for the 0.02 noise datasets respectively. As *precision* did not change much, this was not a major issue, apart from the fact that it increased the computational burden.

Figure 4a shows the modification of one representative CRA-free spectrum. The original CRA- and noise-free spectrum (red curve) was smooth, but the addition of 0.005 level noise caused significant differences (black-plus-blue curve), making it appear as if there were many CRAs present. The use of an overly sensitive lower threshold (*e.g.*, 0.098) in DTCSR led to the modification of some noise interferences by truncating their relatively higher intensity (blue curves). Decreasing the lower threshold could improve CRA removal and accuracy, however, this would also result in modification of more CRA-free spectra (Table S-2, *SI*). For the higher threshold, a value of 0.3 gave the minimum rate of band clipping.

2DCDR, a fully automated method for CRA removal, used second difference spectrum information to locate CRAs and to determine spectral noise. CRA identification was achieved by first comparing the negative intensity peaks in the second difference spectrum to the automatically defined noise threshold, and then spike removal was realized through adding half its negative intensity in the second difference spectrum to its intensity in the original spectrum [15]. This method was suitable for processing large datasets; however, the main issue encountered was the clipping of Raman bands, particularly when these bands were sharp and showed large variations in intensity between spectra. 2DCDR was run on the five simulated datasets, and any CRA-free spectra that were clipped during CRA removal were counted. The precision values for all the clipped spectra were also calculated to quantify the extent that CRA-free spectra were distorted (Table 2).

The ability of 2DCDR to identify CRAs and therefore subsequently remove the artefact was very heavily dependent on noise level, and accuracy got worse as noise increased. This sensitivity was due to the fact that 2DCDR was based on estimated spectral noise. Since in most cases, CRA contamination was only partially removed, high numbers of corrected spectra were obtained for all five simulated datasets. Figure 4b, for example, shows the removal of a CRA at 1214 cm^{-1} (spectrum #7, 0.001 noise case) where the difference spectrum does not match exactly the profile/intensity of the synthetic CRA. This small difference while relatively insignificant was due to the band clipping, also accompanied by the clipping of another two real Raman bands (684 and 830 cm^{-1}). The precision values which were calculated either from the corrected spectra or from the CRA-free but clipped spectra, indicated that there was a certain degree of spectral distortion (*ca*. 1%) compared to the other methods. Even when there was no added

noise, 348 (74%) of the CRA-free spectra were clipped (Figure S-3), and for example at the 830 cm^{-1} band, 234 unnecessary modifications were made.

The 0.001 noise level data were used to calculate the variability of the 348 CRA-free spectra that were clipped during CRA removal, in terms of the variation of the Raman bands among these spectra. The degree of spectrum clipping (normalized scale, 0–1 scale) was obtained by first computing the difference spectra (raw minus clipped spectra), second dividing the difference spectra by the raw spectra, and finally averaging the resultant quotients at each wavenumber. Figure 5 plots the band variability of these spectra and the degree of spectrum clipping. In most cases clipping occurred with bands that were either sharp or had large intensity variations between spectra.

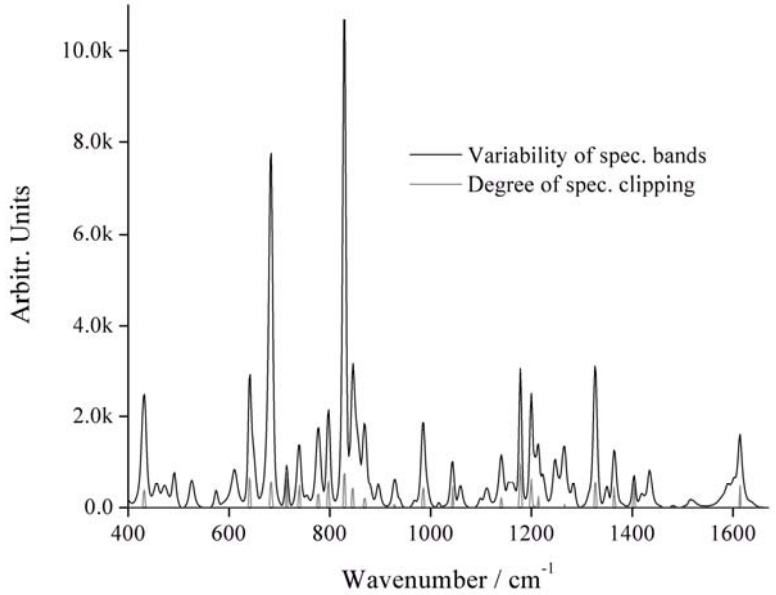

Figure 5: The variability (black curve) of the 348 CRA-free spectra clipped by the 2DCDR CRA removal (0.001 noise level data) and the degree of spectrum clipping (grey, low intensity curve). Most clipping occurred with Raman bands that were narrow and/or highly variable intensity. For easy visualization, the degree of spectrum clipping was magnified 7000 times.

Schulze proposed a compensating scheme to mitigate unnecessary Raman band clipping in 2DCDR [15]. If the frequency of CRA occurrence at a given wavenumber across different spectra, or at a given pixel location across different measurements followed Poissonian statistics, then the detection of a CRA across spectra at a specific wavenumber should have a very low

probability, *e.g.*, a Poisson statistical confidence limit of α=0.01 or 0.05. If the number of modified spectra (or more correctly modifications at a specific pixel) was greater than the established limit for the detector, then that was statistical proof that sharp bands were being clipped and that the spectra should be restored [15]. For example, in our 0.001 noise level case, 234 excessive modifications were made to the real band at 830 cm^{-1} in the 348 clipped spectra. Therefore, these spectra (or more correctly the 830 cm^{-1} band) would have to be restored to their original intensities (*SI*).

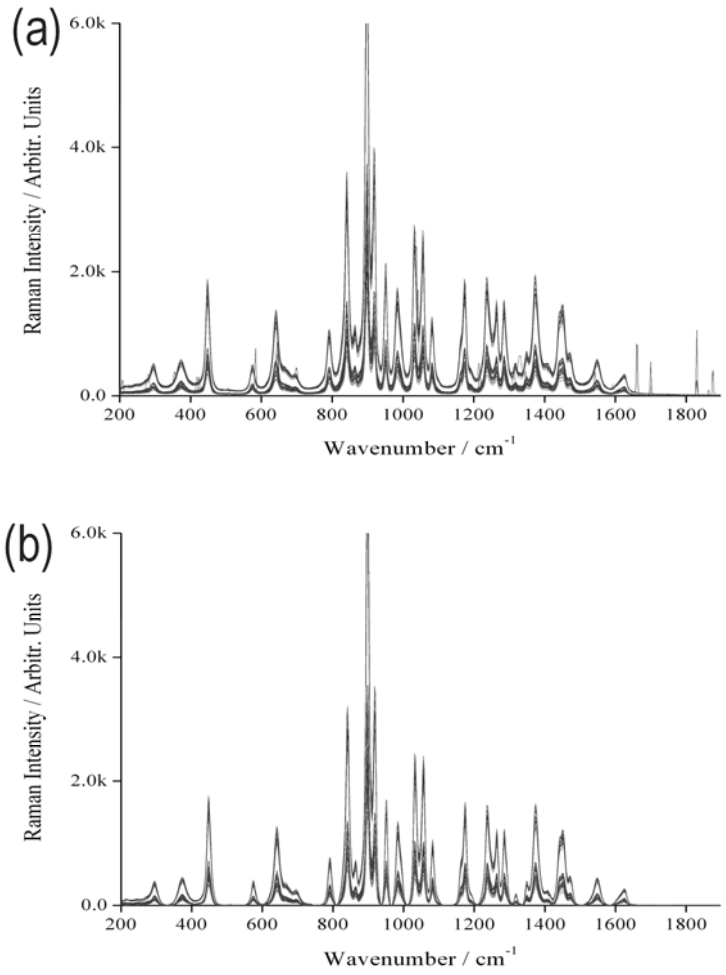

Figure 6: (**a**) *as acquired* Raman spectra show CRAs and small baseline features in a piracetam/proline powder mixture, and (**b**) corrected spectra after application of automated baseline correction using MPLS and CRA removal using KPCARD method.

Overall, compared to DTCSR and 2DCDR, KPCARD was more accurate, precise in terms of both identification and correction of CRA contaminated spectra. KPCARD was also more computationally efficient, and able to automatically accomplish CRA removal on one real dataset of 8410 spectra × 849 variables in 45 seconds on a desktop computer. In contrast, 2DCDR took ~6 minutes and DTCSR needed ~ 6.7 minutes, when performed with the same data set using the same computer configurations (*SI*). Overall, across the full Raman dataset from the 150 samples, KPCARD never took more than 2 minutes to implement, and some sets took less than 40 seconds to complete.

4.4 Piracetam-proline mixture quantification: Ultimately, the rationale for developing the method is for the processing of large volumes of Raman data used for low content quantification. KPCARD was used for CRA correction of a very large Raman mapping dataset generated from 50×3 proline-piracetam binary powder mixtures, with 8410 spectra per sample, 1.3 million spectra in total [7] To show how KPCARD was implemented in practice we took a typical example from a single map measurement. For each map measurement (150 in total), there were typically between 14–56 CRA contaminated spectra (0.17 to 0.67%) as identified by KPCARD. The mean value was 33.6, and standard deviation (STD) was 6.9. This rate of CRA contamination, meant that it was reasonable to use α=0.01 for calculating the upper noise level limit at any given wavenumber across the spectra and thus identifying CRAs in KPCARD. α=0.05 was also tried but this did not yield any improved results (data not shown). Figure 6a shows the 41 CRA contaminated spectra *as acquired* from a single 0.1% piracetam sample, where these spectra were collected at different surface locations *via* the 10 probe channels used [7]. Since each channel sampled a different physical location with possibly a different composition, for each wavenumber position, we get variable signal-to-noise ratios and considerable variability between spectra. All of these particular spectra contained CRAs and a small baseline effect. The baseline can be removed using morphological weighted penalized least-squares (MPLS) [36].

In next step, the mean noise and standard deviation was determined for all 8410 spectra of each individual Raman mapping measurement (Table S-3, *SI)*. The noise was different for each channel, as expected; channels 5 and 6 had the largest noise of 51.83±6.02 and 47.79±5.51 au,

respectively. Since these two channels also gave the best spectra, we can safely conclude that shot noise was the most important factor [7]. The numbers of pixel rows binned for spectral channels 1 to 10 were: 13, 14, 16, 22, 55, 55, 22, 16, 14, and 13 respectively (Table S-3, *SI*), which accounted for the noise differences (larger shot noise for more rows binned). This was confirmed by collecting 8410 spectra from a sample under dark conditions (laser off) with the same exposure settings (Figure S-4, *SI*). The measured dark noise was very low (0.31–0.51) and similar for all channels (Table S-4). Relative to these mean values, the noise variation along the wavenumber axis for each detector channel were small, and therefore, the noise could be regarded as having approximately normal distribution. This fulfilled the requirements for implementing KPCARD, using Equations (6) and (7) to determine the first threshold. CRAs also have to be identified and removed from Raman spectra on a channel-by-channel basis because of the different noise values. The 41 KPCARD and baseline corrected spectra from the 10 channels (Figure 6b) show no obvious spectral distortions.

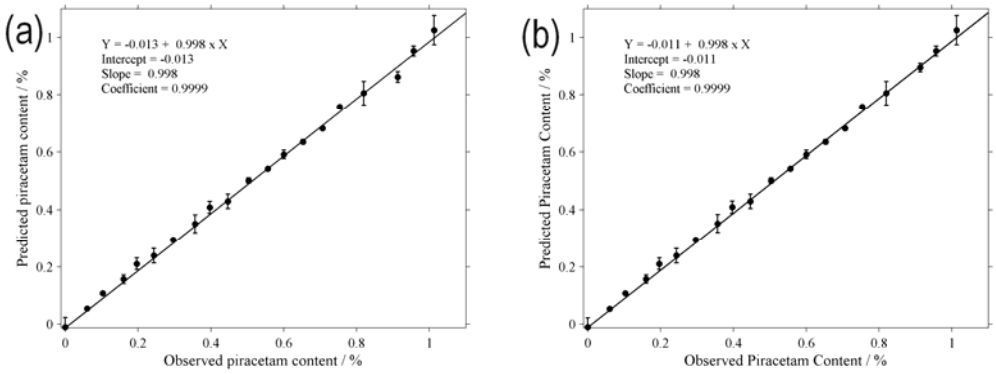

Figure 7: Quantification of piracetam content (0.05−1.0%) in powder mixtures using Raman spectra that were: (**a**) CRA contaminated, and (**b**) KPCARD corrected. Error bars represent standard error for n=3 replicate samples.

Using CRA corrected Raman spectra, accurate quantitative calibration models were established with the powder mixture data, 0−100% piracetam content once the data were appropriately pretreated, as detailed previously [7]. Here, piracetam content in the 0.05−1.0% concentration range was predicted, using both the CRA removed, and CRA contaminated Raman data, and the RMSEP(***REP%***) values were 0.012%(***2.43%***) and 0.016%(***3.28%***), respectively

(Figure 7). The improvement in relative prediction accuracy (*i.e.*, REP%=100×RMSEP/ $\bar{y}$, where $\bar{y}$ = mean value of measured piracetam concentration) was significant. There was also a slight improvement in the intercept value, which again demonstrates the improvement in the quantitative model.

5. Conclusions

KPCARD has been shown to be both rapid (<1 min. for 8400 spectra) and fully automatable. This makes it very applicable to the correction of large volumes of Raman data such as that generated by Raman imaging. When compared to two of the best literature methods, DTCSR and 2DCDR, it was much quicker, more accurate and precise. The method is particularly important for the correction of data used for low content, quantitative Raman analysis [7] where its implementation delivered a ~25% improvement in RMSEP/REP%.

6. Supplemental information available

Supporting information is available at:

http://www.sciencedirect.com/science/article/pii/S0003267016301301#ec1

Acknowledgements

Research undertaken as part of the Synthesis and Solid State Pharmaceutical Centre, funded by Science Foundation Ireland and industry partners, and Enterprise Ireland (Grant No: TC-2012-5106). Kaiser Optical Systems, Inc. and Mr. Harry Owen are thanked for their assistance.

References

[1] R.B. Shah, M.A. Tawakkul, M.A. Khan, Process analytical technology: Chemometric analysis of Raman and near infra-red spectroscopic data for predicting physical properties of extended release matrix tablets, J. Pharm. Sci., 96 (2007) 1356-1365.

[2] G. Fini, Applications of Raman spectroscopy to pharmacy, J. Raman Spectrosc., 35 (2004) 335-337.

[3] A.A. Gowen, C.P. O'Donnell, P.J. Cullen, S.E. Bell, Recent applications of Chemical Imaging to pharmaceutical process monitoring and quality control, European Journal of Pharmaceutics and Biopharmaceutics, 69 (2008) 10-22.

[4] L.L. Simon, H. Pataki, G. Marosi, F. Meemken, K. Hungerbuehler, A. Baiker, S. Tummala, B. Glennon, M. Kuentz, G. Steele, H.J.M. Kramer, J.W. Rydzak, Z. Chen, J. Morris, F. Kjell, R. Singh, R. Gani, K.V. Gernaey, M. Louhi-Kultanen, J. O'Reilly, N. Sandler, O. Antikainen, J. Yliruusi, P. Frohberg, J. Ulrich, R.D. Braatz, T. Leyssens, M. von Stosch, R. Oliveira, R.B.H. Tan, H. Wu, M. Khan, D.

Kernel principal component analysis residual diagnosis (KPCARD): an automated method to remove cosmic ray artefacts in Raman spectra. B. Li, A. Calvet, Y. Casamayou-Boucau, A.G. Ryder, *Analytica Chimica Acta*., 913, (111-120), (2016). DOI: 10.1016/j.aca.2016.01.042

O'Grady, A. Pandey, R. Westra, E. Delle-Case, D. Pape, D. Angelosante, Y. Maret, O. Steiger, M. Lenner, K. Abbou-Oucherif, Z.K. Nagy, J.D. Litster, V.K. Kamaraju, M.-S. Chiu, Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review, Org Process Res Dev, 19 (2015) 3-62.

[5] K.C. Gordon, C.M. McGoverin, Raman mapping of pharmaceuticals, International Journal of Pharmaceutics, 417 (2011) 151-162.

[6] R.L. McCreery, Raman Spectroscopy for Chemical Analysis, Wiley-Interscience, New York, 2000.

[7] B. Li, A. Calvet, Y. Casamayou-Boucau, C. Morris, A.G. Ryder, Low-content quantification in powders using Raman spectroscopy: a facile chemometric approach to sub 0.1% limits of detection, Anal Chem, 87 (2015) 3419-3428.

[8] A.D. Bond, Polymorphism in molecular crystals, Curr Opin Solid St M, 13 (2009) 91-97.

[9] S. Sasic, S. Mehrens, Raman Chemical Mapping of Low-Content Active Pharmaceutical Ingredient Formulations. III. Statistically Optimized Sampling and Detection of Polymorphic Forms in Tablets on Stability, Anal. Chem., 84 (2012) 1019-1025.

[10] S. Sasic, Chemical imaging of pharmaceutical granules by Raman global illumination and near-infrared mapping platforms, Anal. Chim. Acta, 611 (2008) 73-79.

[11] H.S. Lin, O. Marjanovic, B. Lennox, S. Sasic, I.M. Clegg, Multivariate Statistical Analysis of Raman Images of a Pharmaceutical Tablet, Appl. Spectrosc., 66 (2012) 272-281.

[12] L. Zhang, M.J. Henson, A practical algorithm to remove cosmic spikes in Raman imaging data for pharmaceutical applications, Appl. Spectrosc., 61 (2007) 1015-1020.

[13] J.A. Spencer, J.F. Kauffman, J.C. Reepmeyer, C.M. Gryniewicz, W. Ye, D.Y. Toler, L.F. Buhse, B.J. Westenberger, Screening of Heparin API by Near Infrared Reflectance and Raman Spectroscopy, J Pharm Sci, 98 (2009) 3540-3547.

[14] S.E.J. Bell, J.R. Beattie, J.J. McGarvey, K.L. Peters, N.M.S. Sirimuthu, S.J. Speers, Development of sampling methods for Raman analysis of solid dosage forms of therapeutic and illicit drugs, J. Raman Spectrosc., 35 (2004) 409-417.

[15] H.G. Schulze, R.F.B. Turner, A Two-Dimensionally Coincident Second Difference Cosmic Ray Spike Removal Method for the Fully Automated Processing of Raman Spectra, Appl. Spectrosc., 68 (2014) 185-191.

[16] H. Takeuchi, S. Hashimoto, I. Harada, Simple and Efficient Method to Eliminate Spike Noise from Spectra Recorded on Charge-Coupled Device Detectors, Appl. Spectrosc., 47 (1993) 129-131.

[17] D.M. Zhang, K.N. Jallad, D. Ben-Amotz, Stripping of cosmic spike spectral artifacts using a new upper-bound spectrum algorithm, Appl Spectrosc, 55 (2001) 1523-1531.

Kernel principal component analysis residual diagnosis (KPCARD): an automated method to remove cosmic ray artefacts in Raman spectra. B. Li, A. Calvet, Y. Casamayou-Boucau, A.G. Ryder, *Analytica Chimica Acta*., 913, (111-120), (2016). DOI: 10.1016/j.aca.2016.01.042

[18] J. Zhao, Image curvature correction and cosmic removal for high-throughput dispersive Raman spectroscopy, Appl. Spectrosc., 57 (2003) 1368-1375.

[19] S. Li, L. Dai, An Improved Algorithm to Remove Cosmic Spikes in Raman Spectra for Online Monitoring, Appl Spectrosc, 65 (2011) 1300-1306.

[20] B.M. Bussian, W. Hardle, Robust Smoothing Applied to White Noise and Single Outlier Contaminated Raman-Spectra, Appl. Spectrosc., 38 (1984) 309-313.

[21] Y. Katsumoto, Y. Ozaki, Practical algorithm for reducing convex spike noises on a spectrum, Appl. Spectrosc., 57 (2003) 317-322.

[22] G.R. Phillips, J.M. Harris, Polynomial Filters for Data Sets with Outlying or Missing Observations - Application to Charge-Coupled-Device-Detected Raman-Spectra Contaminated by Cosmic-Rays, Anal. Chem., 62 (1990) 2351-2357.

[23] W. Hill, D. Rogalla, Spike-Correction of Weak Signals from Charge-Coupled-Devices and Its Application to Raman-Spectroscopy, Anal. Chem., 64 (1992) 2575-2579.

[24] F. Ehrentreich, L. Summchen, Spike removal and denoising of Raman spectra by wavelet transform methods, Anal. Chem., 73 (2001) 4364-4373.

[25] U.B. Cappel, I.M. Bell, L.K. Pickard, Removing Cosmic Ray Features from Raman Map Data by a Refined Nearest Neighbor Comparison Method as a Precursor for Chemometric Analysis, Appl. Spectrosc., 64 (2010) 195-200.

[26] W. Chew, Information-theoretic chemometric analyses of Raman data for chemical reaction studies, J. Raman Spectrosc., 42 (2011) 36-47.

[27] H.D.T. Jones, D.M. Haaland, M.B. Sinclair, D.K. Melgaard, A.M. Collins, J.A. Timlin, Preprocessing strategies to improve MCR analyses of hyperspectral images, Chemometr. Intell. Lab. Syst., 117 (2012) 149-158.

[28] J.E. Jackson, in: J.E. Jackson (Ed.), JohnWiley & Sons, 2003.

[29] W. Wu, D.L. Massart, S. deJong, The kernel PCA algorithms for wide data .1. Theory and algorithms, Chemometr. Intell. Lab. Syst., 36 (1997) 165-172.

[30] S. Rannar, F. Lindgren, P. Geladi, S. Wold, A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects .1. Theory and Algorithm, J. Chemometr., 8 (1994) 111-125.

[31] F. Lindgren, P. Geladi, S. Wold, The Kernel Algorithm for PLS, J. Chemometr., 7 (1993) 45-59.

[32] H.G. Schulze, M.M.L. Yu, C.J. Addison, M.W. Blades, R.F.B. Turner, Automated estimation of white Gaussian noise level in a spectrum with or without spike noise using a spectral shifting technique, Appl. Spectrosc., 60 (2006) 820-825.

Kernel principal component analysis residual diagnosis (KPCARD): an automated method to remove cosmic ray artefacts in Raman spectra. B. Li, A. Calvet, Y. Casamayou-Boucau, A.G. Ryder, *Analytica Chimica Acta*., 913, (111-120), (2016). DOI: 10.1016/j.aca.2016.01.042

[33] C.J. Behrend, C.P. Tarnowski, M.D. Morris, Identification of outliers in hyperspectral Raman image data by nearest neighbor comparison, Appl. Spectrosc., 56 (2002) 1458-1461.

[34] D.M. Haaland, E.V. Thomas, Partial Least-Squares Methods for Spectral Analyses .2. Application to Simulated and Glass Spectral Data, Anal. Chem., 60 (1988) 1202-1208.

[35] P. Eshghi, Dimensionality choice in principal components analysis via cross-validatory methods, Chemometr Intell Lab, 130 (2014) 6-13.

[36] Z. Li, D.-J. Zhan, J.-J. Wang, J. Huang, Q.-S. Xu, Z.-M. Zhang, Y.-B. Zheng, Y.-Z. Liang, H. Wang, Morphological weighted penalized least squares for background correction, Analyst, 138 (2013) 4483-4492.