Dossier

This paper is a part of the hereunder thematic dossier published in OGST Journal, Vol. 69, No. 2, pp. 195-372 and available online here

Cet article fait partie du dossier thématique ci-dessous publié dans la revue OGST, Vol. 69, n°2, pp. 195-372 et téléchargeable ici

DOSSIER Edited by/Sous la direction de : **L. Duval**

Advances in Signal Processing and Image Analysis for Physico-Chemical, Analytical Chemistry and Chemical Sensing

Progrès en traitement des signaux et analyse des images pour les analyses physico-chimiques et la détection chimique

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 69 (2014), No. 2, pp. 195-372

Oil & Gas Science and Technology – Rev. IFP Energies nouvelles, Vol. 69 (2014), No. 2, pp. 207-228

DOI: 10.2516/ogst/2013185

Dossier

Advances in Signal Processing and Image Analysis for Physico-Chemical, Analytical Chemistry and Chemical Sensing
Progrès en traitement des signaux et analyse des images pour les analyses physico-chimiques et la détection chimique

Multivariate Analysis for the Processing of Signals

J.R. Beattie*

Crescent Diagnostics, UCD-NOVA, Belfield, Dublin 4 - Ireland
e-mail: rene@jrenwickbeattie.com

* Corresponding author

Résumé — **Traitement de signaux par analyse multivariée** — L'analyse multivariée, dont l'analyse en composantes principales (ACP), a transformé, dans des contextes concrets, l'étude de mesures complexes, formées de signaux chargés d'informations. Si la réduction de dimension permet de simplifier grossièrement un enchevêtrement de variations multidimensionnelles, elle est également plus robuste aux perturbations que les méthodes d'analyse univariées. Plus récemment, il est apparu que les propriétés des méthodes multivariées les rendaient propices à d'autres usages que statistiques, comme le traitement des signaux pour l'élimination des variations/fluctuations non pertinentes pour une analyse ultérieure. Il a été montré que l'exploitation spécifique de la réduction de dimension permet un débruitage précis (suppression de "non-signaux/perturbation" non reproductibles), la soustraction fiable et consistante de la ligne de base (suppression de "non-signaux/perturbation" reproductibles), l'élimination d'interférences (suppression de "signaux" reproductibles et inutiles), ainsi que la standardisation des fluctuations d'amplitude des signaux.
Si ce champ d'investigation est encore restreint, les possibilités de diffusion de ses applications sont considérables. En effet, ces améliorations, intrinsèquement liées aux signaux eux-mêmes, sont hautement reproductibles entre les répétitions, possèdent une grande capacité d'adaptation et d'application à des situations de bruit, ou de variations complexes dans les signaux. Alors que les disciplines scientifiques sondent des volumes de données toujours plus volumineux, dans des situations de moins en moins étroitement contrôlées, la capacité à apporter des corrections/améliorations précises/hautes résolutions, de manière flexible, devient de plus en plus critique. Aussi les traitements de signaux multivariés offrent un éventail de solutions potentiellement très large.

Abstract *— **Multivariate Analysis for the Processing of Signals** — Real-world experiments are becoming increasingly more complex, needing techniques capable of tracking this complexity. Signal based measurements are often used to capture this complexity, where a signal is a record of a sample's response to a parameter (*e.g. *time, displacement, voltage, wavelength) that is varied over a range of values. In signals the responses at each value of the varied parameter are related to each other, depending on the composition or state sample being measured. Since signals contain multiple information points, they have rich information content but are generally complex to comprehend. Multivariate Analysis (MA) has profoundly transformed their analysis by allowing gross simplification of the tangled web of variation. In addition MA has also provided the advantage of being much more robust to the influence of noise than univariate methods of analysis. In recent years, there has been a growing awareness that the nature of the multivariate methods allows exploitation of its benefits for purposes other than data analysis, such as pre-processing of signals with the aim of eliminating*

irrelevant variations prior to analysis of the signal of interest. It has been shown that exploiting multivariate data reduction in an appropriate way can allow high fidelity denoising (removal of irreproducible non-signals), consistent and reproducible noise-insensitive correction of baseline distortions (removal of reproducible non-signals), accurate elimination of interfering signals (removal of reproducible but unwanted signals) and the standardisation of signal amplitude fluctuations. At present, the field is relatively small but the possibilities for much wider application are considerable. Where signal properties are suitable for MA (such as the signal being stationary along the x-axis), these signal based corrections have the potential to be highly reproducible, and highly adaptable and are applicable in situations where the data is noisy or where the variations in the signals can be complex. As science seeks to probe datasets in less and less tightly controlled situations the ability to provide high-fidelity corrections in a very flexible manner is becoming more critical and multivariate based signal processing has the potential to provide many solutions.

ABBREVIATIONS

DOSC	Direct Orthogonal Signal Correction
EMSC	Extended Multiplicative Scatter Correction
FTIR	Fourier Transform InfraRed spectroscopy
MLR	Multiple Linear Regression
MSC	Multiplicative Scatter Correction
NIR	Near InfraRed spectroscopy
NMR	Nuclear Magnetic Resonance spectroscopy
OSC	Orthogonal Signal Correction
PC	Principal Component
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PLS	Partial Least Squares (regression)
SNR	Signal to Noise Ratio
SVD	Singular Value Decomposition
XPS	X-ray Photoelectron Spectroscopy

INTRODUCTION

As technology progresses analytical science (science performed with the aim to identify and quantify useful parameters of unknown samples) is increasingly being applied to 'real-world' situations, where the overall composition of the samples is not fully controlled. There is a higher risk of variation that is unrelated to the target analytes (the chemical constituents to be identified or quantified) and a higher risk of interfering processes than may occur in highly controlled laboratory analyses. Signal based measurement is ideal for capturing complex information as it records multiple channels of interrelated information. In the context of this manuscript, a signal will be considered as any multi-variable array of information where the relationship between all variables in the array is continuous and interrelated and the ordering of the variables is systematic. Non-signal arrays consist of discrete parameters with no interrelation that is obvious prior to analysis, whose scales may differ widely and whose magnitudes may be independent of each other. Because variables in non-signal arrays are independent, what happens in one variable gives little information about what to expect in another. In signals the interrelatedness of the variables means that what happens in each variable has a bearing on all others allowing manipulation of the variables as an array.

However, it is easy to quickly lose sight of the pertinent information in the midst of the jumble of different signal contributors being recorded. Consequently there has been a growing interest in the use of data analysis approaches that can capture the complexity of the variation observed in the course of analysis and report it in a much more concise and interpretable manner. A regrettable side effect of capturing more complex information on target analytes is that such datasets tend to capture more complex interferences and this can lead traditional signal processing methods to provide suboptimal solutions.

This manuscript is intended to provide an overview of techniques that have exploited multivariate analysis for the purpose of signal processing, exploiting the condensation of complex interferences as well as complex target analytes, into a simpler form. Multivariate data reduction provides a very high fidelity and high specificity summary of a dataset, benefits which can be exploited for signal processing. Such an approach will not work on all types of data, but is applicable where the signal of target analytes *and/or* interfering parameters recur and are summarisable by Principal Component Analysis (PCA) or Partial Least Squares (PLS) regression. As with traditional per-signal processing, the unwanted (non-signal and contaminant signal) variation needs to be distinguishable from the wanted (analyte signal) variation.

This review is intended as an introduction to the possibilities rather than an operating manual and will provide some of the basic underlying principles involved in

multivariate based signal processing. To date the published methods have primarily exploited PCA and PLS regression. This will involve an initial, non-comprehensive, discussion of PCA intended to provide a basic mathematical context for the methodologies. The body of the manuscript will provide examples of the different types of signal interference that the procedures have been used to correct. The majority of the examples is spectroscopic and biomedical in origin, but analytic method and application independent guidance on applicability will be provided. Many of the issues faced in these examples are common to all 'real-world' analyses including samples encountered in the assessment of complex oil and gas mixtures.

The vast majority of the published literature applying multivariate analytical techniques to signals is for the final data analysis stage subsequent to any signal processing that has been carried out by alternative means. However, the calculations underlying multivariate analysis are mathematical in nature and are not necessarily restricted to pure statistical analysis. Because PCA summarises the major waveforms present in the dataset and the scores represent their contribution to each individual signal it is possible to manipulate and exploit these results for signal processing. In this manuscript, some existing uses of multivariate analysis for signal processing will be presented and discussed, falling into the categories of broad-bandwidth background signal and interferant removal, amplitude/intensity normalisation and denoising. The underlying principles of PCA, the most commonly used multivariate data analysis, are described in a technical discussion for readers who are keen to understand the underlying mechanism of the approach, but this section is not essential for those readers who may wish to return to it later.

1 TECHNICAL DESCRIPTION OF PRINCIPAL COMPONENT ANALYSIS

Multivariate analysis is a very common solution to the complexity of 'real-world' analysis and is simply any approach that simultaneously analyses multiple variables to reduce them to a simpler output [1, 2]. One of the most common forms of multivariate analysis is PCA, in which the major signals contributing to the dataset are identified and quantified (or from the perspective of obtaining chemical/physical parameters it is semi-quantified, in that there is a numerical scale that has an underlying basis but needs transformation to correspond to chemical or physically meaningful parameters established by other means). PCA is an unsupervised technique and does not discriminate between 'good' and 'bad' variation and will often retain desirable and undesirable information in a mixed form. However, PCA does find a representation of the simplest possible descriptions (under some assumptions) of the variation in the dataset and so, where it is possible to discriminate desired and undesired variation, these simplified results can allow powerful opportunities for correcting undesired variation.

The equation that underpins PCA is a very simple equation:

$$D = SL^T \tag{1}$$

where D is the data matrix with o observations along rows and v variables accounting for each column ($o \times v$ matrix). If there are more observations than variables ($o > v$) then S is a $o \times v$ matrix of values (known as scores) for each sample lying along the row, and L^T is a $v \times v$ transpose matrix of waveforms known as principal components (PC, also known as latent variables, eigenvectors or as loadings), with intensities for each loading lying along each column. If there are more variables than observations ($v > o$) then S is an $o \times o$ and L^T is a $v \times o$ matrix. The superscript ^{T} indicates a transpose and is included in the L^T term because the vectors for the PC lie perpendicular to the vectors in the original data. Using the principles of matrix multiplication, where each element in a column of the first matrix is multiplied by the corresponding row of the second matrix, Equation (1) can be expanded:

$$D = \sum_{p=1}^{p=\min(o,v)} \left(\mathbf{s}_p \times \mathbf{l}_p^T\right) = \mathbf{s}_1 \times \mathbf{l}_1^T + \mathbf{s}_2 \times \mathbf{l}_2^T + \cdots + \mathbf{s}_p \times \mathbf{l}_p^T \tag{2}$$

where p is the index of a principal component, with its upper limit as the lesser dimension of D, *i.e.* number of samples o or the number of variables, v. $\mathbf{s}_p$ is the vector for the p^{th} row in the S matrix and $\mathbf{l}_p^T$ is the vector for the p^{th} column in the L^T matrix. Each loading is scaled by the score for the corresponding principal component and the data can be reconstructed by simply summing the score weighted loadings. Due to the way PCA is calculated L^TL or LL^T give an identity matrix (the linear algebra equivalent of the digit 1), so by multiplying each side of Equation (1) it can be rearranged [3]:

$$S = DL \tag{3}$$

The scores are the data multiplied by the loadings and thus provide quantitative concentration information on the proportion of each loading within each sample.

In PCA, S^TS does not give an identity matrix, but premultiplying each side of the equation by the pseudo inverse of $S^\dagger$ (denoted as S) allows Equation (1) to be rearranged:

$$L = S^\dagger D \tag{4}$$

This demonstrates that the loadings are simply score weighted averages of the original data matrix, a point which will be discussed below. The pseudoinverse is a least squares approximation equivalent to dividing by the S matrix (when it is neither square nor invertible).

The first step to calculating S and L^T is to calculate the sum of the squared intensities by multiplying the transpose of the data matrix by the untransformed data matrix, *i.e.* D^TD. The result of pre multiplying the D matrix by its transpose gives a diagonal containing the sum across all samples of the squared amplitudes or intensities (in this manuscript, **a** will denote the vector of amplitudes that represents a signal, which means that D^TD gives $\Sigma(\mathbf{a}^2)$, i to represent intensity is not used to avoid confusion with the identity matrix I). This sum of squares is then fitted by determining the weightings for the individual spectra (represented by the vector **a**) such that the sum of the weighted intensities ($\mathbf{w} \times \mathbf{a}$) minimises the sum of the squares, *i.e.*:

$$\sum_{i=1}^{i=0} \left(\mathbf{a}_i^2 - \mathbf{w}_i \times \mathbf{a}_i\right) \rightarrow 0 \tag{5}$$

Because the weights are fitting a squared term, these weights will be related to the square of the original intensities. When variation in the form of the signal is present (*i.e.* each signal is not just a scaled version of the others) the sum of the weighted amplitudes will not equal that of the sum of the squared intensities. This inequality means that a residual of the sum of squares is left behind. This residual sum of squares is then input into (5) in place of $\mathbf{a}_i^2$ and the spectra reweighted to best fit the residual and calculate the next PC. This process is repeated until all PC are calculated (if $o < v$ this is at most o PC, if $o > v$ this is at most v PC, but other limitations can be applied to reduce the number of PC calculated).

The fact that the loadings are a weighted average of the original data has important consequences. First, they are weighted average signals and because the scores are both positive and negative they present as the result of subtracting two signals [4], with the positive scores giving the positive signal and the negative scores contributing the negative portion. As alluded to above, the loadings do not discriminate between desirable and non-desirable variation as they are blind to this division. Also, the loadings do not necessarily represent 'pure' units of signal (this depends on experimental design) but they do summarise linear combinations of any signals that consistently occur in tandem. For example, a signal from source X may always appear in conjunction with signal from source Y and so the loading describing X will contain a contribution from Y. In 'real world' datasets the complexity of such co-variation can make interpretation of the loadings back to physically or chemically meaningful parameters tricky. In many cases X and Y might be desirable analyte signals, but often one may be desirable and the other not and this is where correction on the loadings themselves steps in.

To illustrate the basic concepts discussed above, the process is visualised using anonymised signals (since the principles apply to signals generally, the identity of the signals does not matter for this figure) in Figure 1a shows some data from a dataset along with its PC (*Fig. 1b*) and scores derived from PCA of that dataset (*Fig. 1c*). The PC are the score weighted averages of the data (*Eq. 4*) and represent the basis set of signals that can be used to explain all the significant variation in the original data. Because the loadings are weighted averages of real signals, the signal shapes observed in the loadings do have real physical meaning (although in very complex datasets this might still be hard to decipher). Since scores are positive and negative, the loadings take the appearance of subtraction spectra, but it is possible to connect the signal in the loading back to the signals that contribute to the dataset [4]. The contribution of each PC to a signal can be calculated by dividing each channel in the PC by the equivalent channel in the signal and summing the result to give the score, although it is usually calculated as matrices (see *Eq. 3*). Because they capture all the information present in the data, the original data D_i can be reconstructed from the loadings and scores using Equation (1) above.

The score relays quantitative information on its contribution to the signal being analysed. The scores can be used in many ways, including visualisation in figures such as the scatter plot shown in Figure 1c. Score plots can be very powerful tools in assessing datasets and in the figure some selected samples are circled.

These show extreme score values in PC 2 and 3 and it is very instructive to compare the signals that gave these scores to the relevant PCs. Figure 1d (see "i") has a strong positive score in both PC 2 and 3 (*Fig. 1c*, see "i") and this signal exhibits strong bands indicated by the arrows that match positive bands in the loadings for PC 2 (two right hand peaks) and 3 (two left hand peaks). Figure 1d (see "ii") is a signal which has only the right hand peaks but not the left and so it has a high score in PC2 but a low score in PC3. Figure 1d (see "iii") is a signal that has no strong bands at either of these

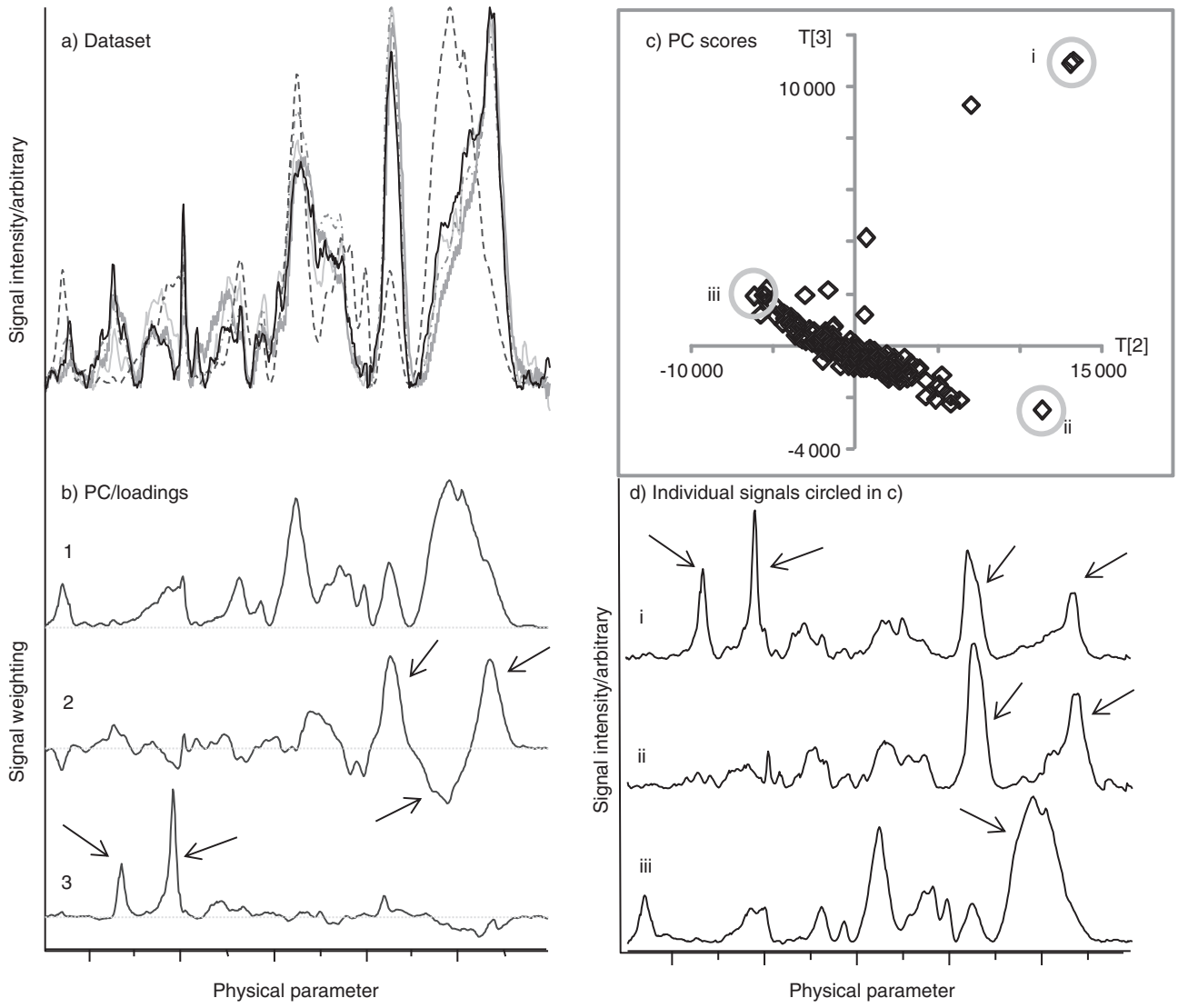

Figure 1

Multivariate analysis of the example dataset a) summarises the main sources of variation within the dataset as b) principal components (PC, also known as loadings and eigenvectors). The proportion of each PC is given by the scores plotting as a scatter plot in c), in which scores from PC 2 (x-axis) and 3 (y-axis) are plotted. Selected samples are circled in c), with the corresponding spectra shown below in d).

positions, but has a strong band that matches the position of the negative feature in PC2 (arrowed). Because the scores are calculated by matrix multiplication of the inverse, the strong bands take on the same sign as the bands in the PC and so strong bands matching the negative bands of a PC give a negative score. It is not the intention to provide a comprehensive insight into the workings of PCA in this manuscript and many more detailed accounts of its operation are available [1, 2, 5-9].

It has been recognised that PCA has an inherent noise cancelling ability due to the fact that the error in each signal channel is averaged out over the full signal. In spectral mapping (*i.e.* a full spectrum acquired at each point in a grid over a defined area) extracting the signal specific information using univariate analyses gives a much more speckled chemical image than the equivalent multivariate determined image [10-12]. Figure 2 illustrates this principle in chemical images of mineral distribution in a sample, with the univariate derived maps exhibiting significantly greater levels of speckle than the multivariate derived maps, with the difference most exaggerated in the poorer quality dataset.

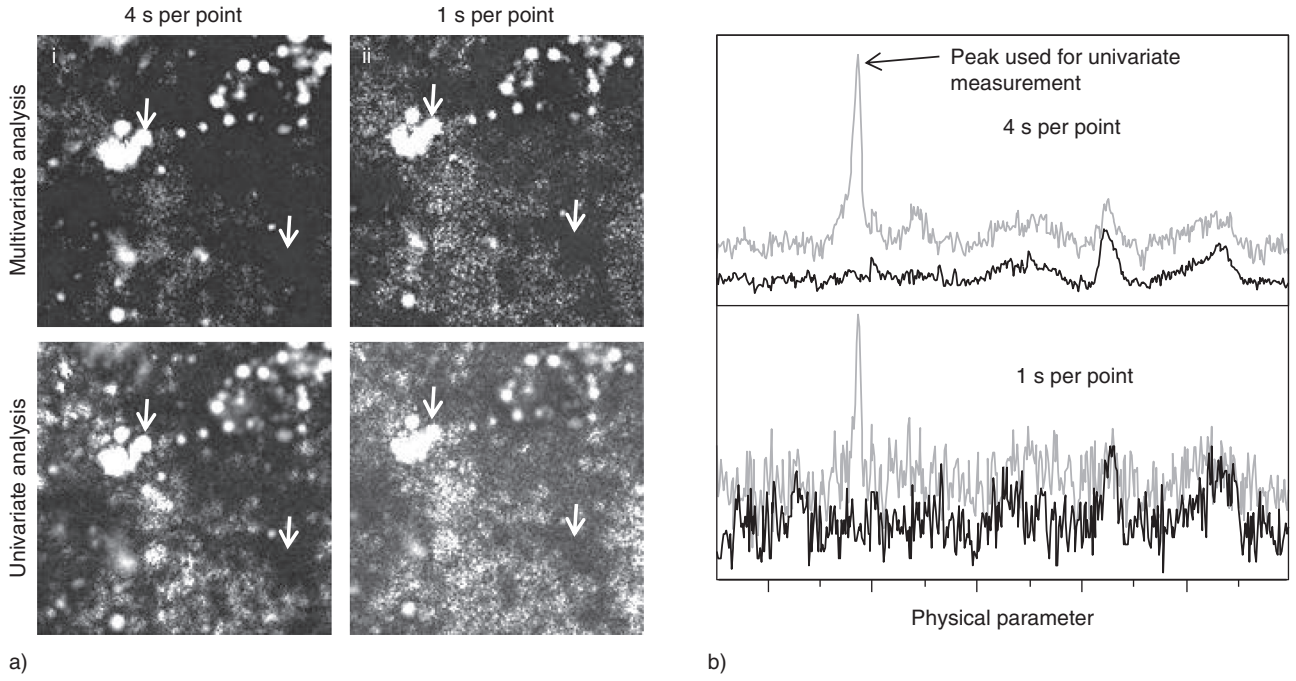

Figure 2

Comparison of chemical images a) calculated using univariate and multivariate methods, the univariate method of calculating intensity produces significantly more speckle in the image, with the difference increased at lower SNR. Signals corresponding to the arrowed positions on the images are shown in b), with the grey signal corresponding to the bright spot, the black signal to the dark region. Adapted from Beattie [44] JRS.

Singular Value Decomposition (SVD) is a technique that is commonly encountered in multivariate analysis and much confusion surrounds how it relates to PCA (the author has been victim to this in the past). SVD is one of several approaches to calculating PCA, involved on the computation of the pseudo-inverse (see above) and is described separately when not carried through to a complete PCA. SVD simplifies the least squares fitting of the data by normalising the scores and loadings such that the sum of their squared values equals 1, constraining the range of values over which the best fit needs to be calculated. This is easily achieved by calculating the sum of the sum of squares, the result of which is a number called the singular value, which also corresponds to the square root of the eigenvalue [3]. PCA is calculated from SVD by multiplying the scores and singular value matrices while the loadings matrix remains unchanged.

Multiple Linear Regression (MLR) can be used to exploit PCR (Principal Component Regression), first reducing the signal data matrix then correlating this with independent reference values related to the individual signals. In PCA, the aim is to fit the variance of the signal data with the signal intensities, which requires a squared weighting to scale onto the data. In MLR, the aim is to fit the covariance within the signal data to the covariance between the signal data and the reference data:

$$\sum_{i=1}^{i=0} \mathbf{a}_i^2 - \mathbf{w}_i(\mathbf{a}_i \times \mathbf{y}_i) \rightarrow 0 \qquad (6)$$

where $\mathbf{y}_i$ is the vector of independent reference values that corresponds to the signal $\mathbf{a}_i$ (where regression is against one reference value this will be a scalar). So multiple linear regression also involves the calculation of a sum of squared intensities, as with PCA, but this is fitted to the sum of the weighted amplitudes which have already been scaled by the reference values. This operation allows comprehension of the correlated relationship between the variables within the signal and within the independent reference values. Because there is now a defined relationship between the signal and external data it is now possible to use external data to manipulate the original data by applying appropriate transformations scaled by the regression coefficients.

In MLR, the variance in the signal matrix is minimised, but this gives greater priority to the signal rather

TABLE 1
Summary of properties of each multivariate signal processing approach

Technique	Type of variation eliminated	Type of variation retained
Denoising	Irreproducible non-signal	Recurrent signal and non-signal
Background correction	Recurrent non-signal	Recurrent signal and irreproducible non-signal
Contaminant elimination	Recurrent unwanted signal	Recurrent wanted signals, recurrent non-signals and irreproducible signals
Intensity/amplitude normalisation	Variable magnitude between signals	Relative variation within signals

than the reference data for directing this regression. Partial Least Squares (PLS) regression differs slightly from MLR in that the covariance between the signal data and the reference data is even more intimately entwined in the data reduction in order to better balance the reduction of both the signal and reference data variance. This maximises the description of the covariance between the signal and the reference values during the actual data reduction step. The exact mechanism is considerably more complex and not readily condensed for this context, but readers wishing to understand more fully can refer to the online resource at StatSoft for an in-depth explanation [13]. This has the consequence that the signal matrix can summarise the information in the Y matrix as succinctly as possible based on the chosen limitations and assumptions, which can be exploited for signal processing where some independent measure can provide useful information on the quality of the signals. To date this benefit has been exploited in a negative sense, detecting the lack of correlation between signal and reference parameter and identifying distortions that can be removed to eliminate uninformative portions of the signal, as is described in detail below.

In essence, the term 'multivariate' can mean any method that uses multiple variables for achieving a goal, but in this manuscript the scope has deliberately been restricted specifically to the use of multivariate data analyses based on data reduction through PCA or PLS regression. The principles of transforming the information in the loadings and scaling this back to the original signals should hold for most other multivariate analysis methods.

Signal processing is used across many different fields of scientific investigation and regrettably the terms and definitions used may differ widely. It will not be possible to deal with such variation in terminology efficiently, but Table 1 is proved in the hope of helping readers to understand the terminology used in this manuscript and to understand how it may relate to their own data. Signal refers to any intensity or amplitude that is observed in the data that corresponds to the physical process being measured, but is not necessarily due to the desired target analytes. Non-signal refers to any intensity or amplitude that corresponds to other physical processes besides the one the measurement is intended to collect. This can be due to competing phenomema (which tend to give structured interferences, which are reproducible) or due to random processes (which tend to give unstructured interferences and are irreproducible).

2 MULTIVARIATE DE-NOISING: ELIMINATING IRREPRODUCIBLE NON-SIGNAL

The above noise cancelling effect can be exploited in another way, other than achieving lower prediction errors or better chemical image contrast. The loadings are calculated from linear weighting of the entire dataset and therefore the loadings are less noisy than the original signals. If these low-noise loadings are weighted by the scores, the reverse calculation (*i.e.* by employing *Eq. 1* with truncated matrices) can be made – constructing the linear combination of the loadings that best fits the original signal, effectively denoising the signal (a process known as Factor Analysis Rank Reduction or Reduced Rank Approximation). The example below will illustrate the elimination of pixel-to-pixel shot noise in a signal as this is the most common use.

PCA identifies and (by virtue of exploiting least squares fitting) inherently prioritises the signals that contribute most, in energy or intensity, to a dataset and so favours strong signals that occur reproducibly within the dataset and ranks weaker or inconsistent irreproducible signals lower [14, 15]. The calculation differs slightly from Equation (1) in that S is truncated to an $o \times i$ matrix, and L^T is an $i \times v$ matrix, where i is the number of low noise PC identified.

Figure 3a shows a number of replicates of a sample, acquired at the same point, in a stable environment using a non-destructive technique [16]. The individual repetitions have a average SNR of 3.07 (4.8 dB), the signals for these replicates show no correlation *versus* time,

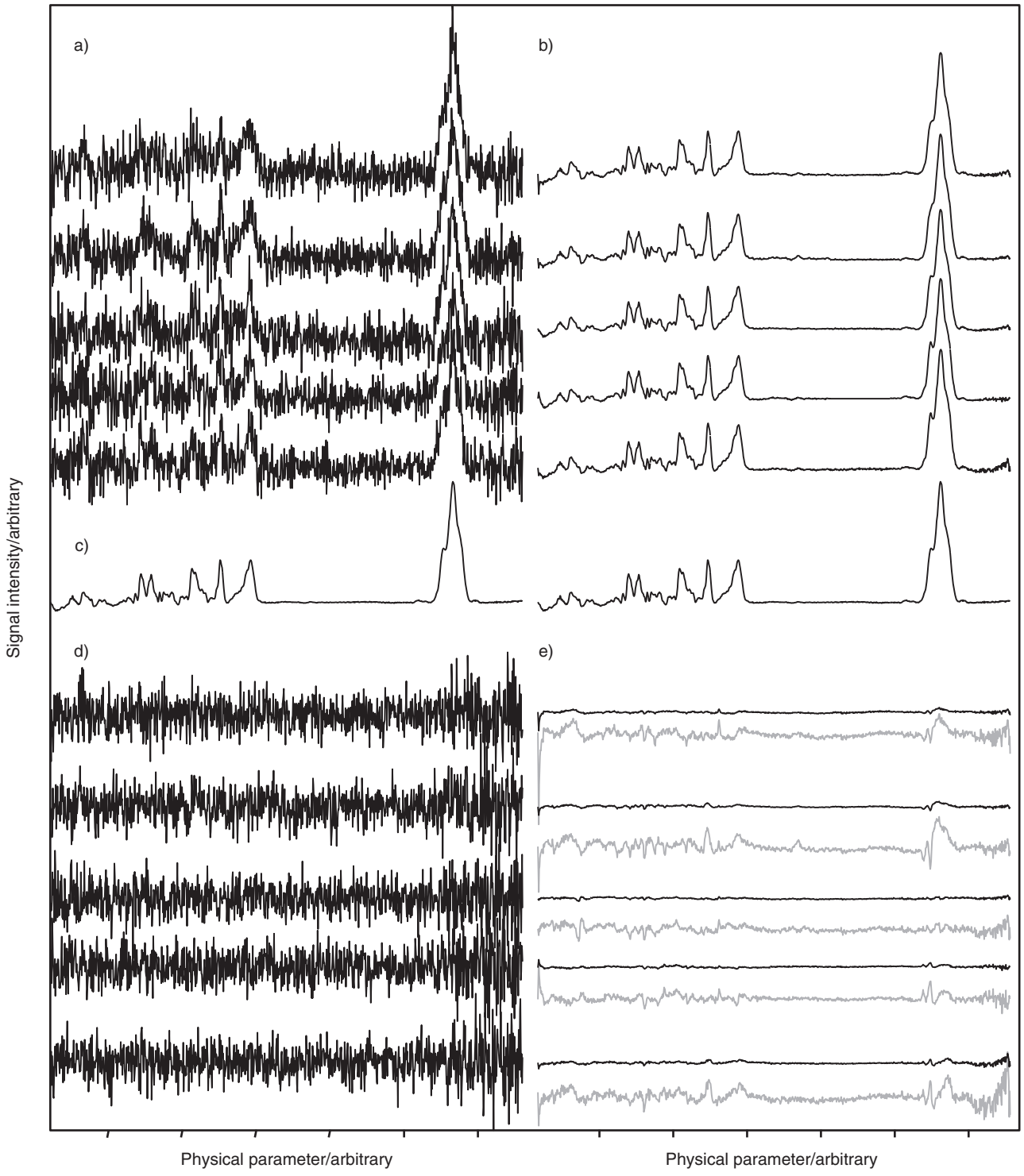

Figure 3

Multivariate denoising of low SNR signals a) 5 replicates of one sample at one location, SNR 3.07 b) the same signals denoised using multivariate methods c) the mean signal for the sample (2 000 replicates) d) the difference between the individual replicates and the average signal e) the difference between the denoised replicates and the average signal, grey lines are the differences scaled by a factor of 5. The signals have been offset for clarity. Constructed from data from Beattie [44].

and PCA of 2 000 replicates returns only one significant PC, so the analyte contribution to the signals is not significantly changing. The signals show very high variability in their pixel-to-pixel variation and so they differ substantially from the much better quality average signal (2 000 replicates at the same point) in Figure 3c. Figure 3b shows the same signal replicates reconstructed from the combination of 8 PC and their respective scores, with the variability between the replicates substantially reduced. Comparing the residual after subtracting the mean spectrum from the original signals for the replicates (*Fig. 3d*) with the residual after subtracting the mean signal from the denoised replicates (*Fig. 3e*) shows a dramatic decline in the variability between the replicates. The variation between the original replicate signals is large and highly unstructured with no consistent pattern and the application of multivariate denoising reduces the level of variation between replicates by a factor of 11. The spectral features are fully preserved with no band broadening that is typical of many moving filters such as Savitsky-Golay.

The powerful denoising while maintaining compositional fidelity that can be achieved using multivariate denoising, is particularly beneficial with samples that change rapidly or are sensitive to overexposure and it is critical that analysis is carried out as quickly and uninvasively as possible. Ghita *et al.* [17] developed a method using Raman spectroscopy to identify live stem cells in a cell culture and in order this application to be viable, it is imperative that the cells remain unaffected by the interrogation and that many cells are probed as possible in a unit of time. The team used the intensities of bands that were unique to specific biochemicals such as RNA to map the differing compositions of stem cells and differentiated cells, but were constrained to used rapid acquisition times resulting in noisy individual spectra. Application of the SVD based denoising allowed chemical images of high contrast to be generated from the noisy data, enabling identification of the cellular regions that could be used to differentiate between stem cells and differentiated cells. The team concluded that RNA was elevated specifically within the cytoplasm of neural stems cells when compared with the cytoplasm of mature glial cells.

The applicability of the multivariate denoising approach is very broad, and it is conceivably possible to use the approach for any digital signal for which multivariate methods can be exploited on the signal of interest. Indeed the published literature abounds with examples of the use of multivariate denoising for a very diverse range of digital signal types. Some further recent examples include X-ray Photoelectron Spectromicroscopy (XPS) [18], ultrasound [19], seismic data [20], Positron Emission Tomography (PET) [21], High Angular Resolution Diffusion Imaging (HARDI) [22], Electronic Nose [23], Image denoising [24], Speech signal [25], length measurements from kymographs [26] and Electron Energy Loss Spectroscopy [27].

It is clear that multivariate based denoising is a powerful technique, however, magnifying the residuals in Figure 3e demonstrates that care must be taken when interpreting the variability of data that has been denoised using the multivariate approach. The variation between denoised signals is highly structured and could easily be misinterpreted as some compositional variation, when in fact it is the noise in the spectrum that induces a small error in the reconstruction. It is important that the level of confidence that can be placed in a reconstructed spectrum be determined.

Regrettably, this aspect of multivariate denoising is underrepresented in the literature (the author has not been able to identify a manuscript) and there are not clear rules defined for determining confidence in reconstructed spectra. Because we know the composition in Figure 3 is as close to constant as we can achieve experimentally, it follows that there should be infinitesimal confidence that the small variations observed are real. However, in real-world datasets we may not have the luxury of repeated measurements on stable samples to determine the magnitude of changes that are unreliable. Instead it is important that limits of detection for each PC (not just overall signal strength) are reliably estimated and only variations in the reconstructed signals that exceed the detection limit are interpreted or used further.

Not to be confused with detection limits for reconstruction, the number of PC that should be used to reconstruct the signal is important to determine. Using too few PC will eliminate true signal information while using too many will incorporate too much noise, weakening the denoising effect. Uzunbajakava *et al.* [15] used the eigenvalues (square of singular values) to determine which components to use in the reconstruction of the spectra. However, the eigenvalues relay information on the contribution of the multivariate loadings to the variation in the dataset and not specifically on the relative signal information of the loadings and as such are not ideally suited to defining significance for denoising. In order to minimise this issue of affecting the information in the signals, Bassan *et al.* [28] first assessed the PC derived from FTIR data for presence of signal information. Details were not provided on how this was performed, but an estimation of noise levels (pixel to pixel variation) in the loadings could prove useful. The authors of the paper then used twice the number of PC in which signal information could be found to perform

the multivariate denoising [28] to ensure that no subtle chemical information was eliminated. It should be noted that because the denoised spectra are constructed from the scores, the scores contain exactly the same information as the reconstructed data. Thus reconstruction of the denoised spectra should only be carried out if subsequent analysis is signal based, such as peak parameters (ratios, widths or positions) or fitting the data against a database of reference standards [15]. If the subsequent steps do not consist of spectral feature analysis it is much more efficient to use the scores for subsequent analysis (this is the common approach in cluster analysis and linear discriminant analysis).

Variants of multivariate denoising involves the use of the MultiScale version of PCA (MSPCA) [29] in combination with wavelet transformations. A wavelet transformation is applied to the data signals to produce detail (D) and approximation (A) matrices, but these are heavily influenced by high shot noise levels in the original data which lowers the reproducibility of the wavelet transform. The matrices are then subjected to a PCA and an appropriate number of PC are chosen to reconstruct the D and A matrices. This creates a simplified set of detail and approximation coefficients that describe the major portions of the signal and ignore the minor (noise) contributions. This simplified set of D and A coefficients can then be reverted to the data space by inverting the wavelet transform [30]. The transformed dataset has then been more consistently transformed from the original than is possible by traditional application of wavelets. Chaux *et al.* [31] demonstrated a neat solution to denoising multichannel images by exploiting correlations between spatially equivalent pixels in each channel. Wavelet transforms were calculated for each channel separately and the transform is initially denoised within each channel as described above. Application of PCA to the multiple channels identified the principal relationships between the wavelet transformations in each of the channels. By reconstructing the wavelet transforms based on the highest eigenvalue components, it was possible to improve the denoising effect further.

3 MULTIVARIATE BACKGROUND CORRECTION: ELIMINATING REPRODUCIBLE NON-SIGNAL

The above approach works well for inconsistent and irreproducible signals or non-signals, but many signals are affected by the opposite problem: persistent and highly reproducible interfering signals. These can take many forms. In the dataset demonstrated earlier in Figures 1 to 3, the main source of reproducible interference is from very broad bandwidth signals. Martens developed a multivariate based improvement to an existing spectral pre-processing method: Multiplicative Scatter Correction (MSC, in which multiplicative scattering errors were corrected through standardising the offset and mean value of the spectra). Martens and Stark [32] introduced the concept of an adaptable MSC based on construction of a database on interfering signals and a database of desired ('target') signals. These databases were then combined and subjected to a multivariate data reduction and multiple linear regression onto the original data performed and the technique was termed Extended Multiplicative Scatter Correction (EMSC). The contribution from undesired signals was removed from the original data while the contributions from the desired signals and the residual were retained. The technique has been applied to a range of signal types, including NIR [33], Raman Spectroscopy [34] and FTIR [35].

Where the cause of spectral distortions can be determined very accurately, such as through mathematical modelling, the EMSC approach can prove very powerful as is illustrated by the case of resonance Mie scattering, which is a perturbation that occurs during an infrared measurement when the physical size of an object being measured is a similar size to the wavelengths of mid-infrared light. The Mie effect involves a relationship between the physical size of the particles in a sample and the wavelength of light being absorbed, when the two are similar. Synchrotron radiation FTIR spectra of biological cells were shown to suffer significantly from this problem [36] in which the spectra are distorted in a non uniform manner, dependant on the absorption of the sample at each wavelength, illustrated in Figure 4. This figure shows the spectra of polymethylmethacrylate of different particle sizes, *i.e.* the composition of the samples is constant, it is just their physical size that varies. The consequence of the wavelength and absorption dependence is that using simple baseline corrections may correct the gross distortions of the broad background but are unable to correct for the subtler changes occurring at peaks in the spectrum such as the shifts in the modes around 1 700-1 770 cm^{-1} in Figure 4a and 4b, leaving peak shifts and distortions uncorrected. The same team developed a method of constructing an appropriate database of Mie signals from a large database of infrared spectra using complex theoretical modelling. This database of Mie backgrounds was then input into the EMSC algorithm, allowing comprehensive correction of the contribution of Mie scattering [28, 35, 37]. The appropriate use of targeted baseline database formation allowed a very accurate and consistent correction to be applied to the data, enhancing the reliability of the resulting data analysis [28]. In contrast, where detailed independent methods of providing a database

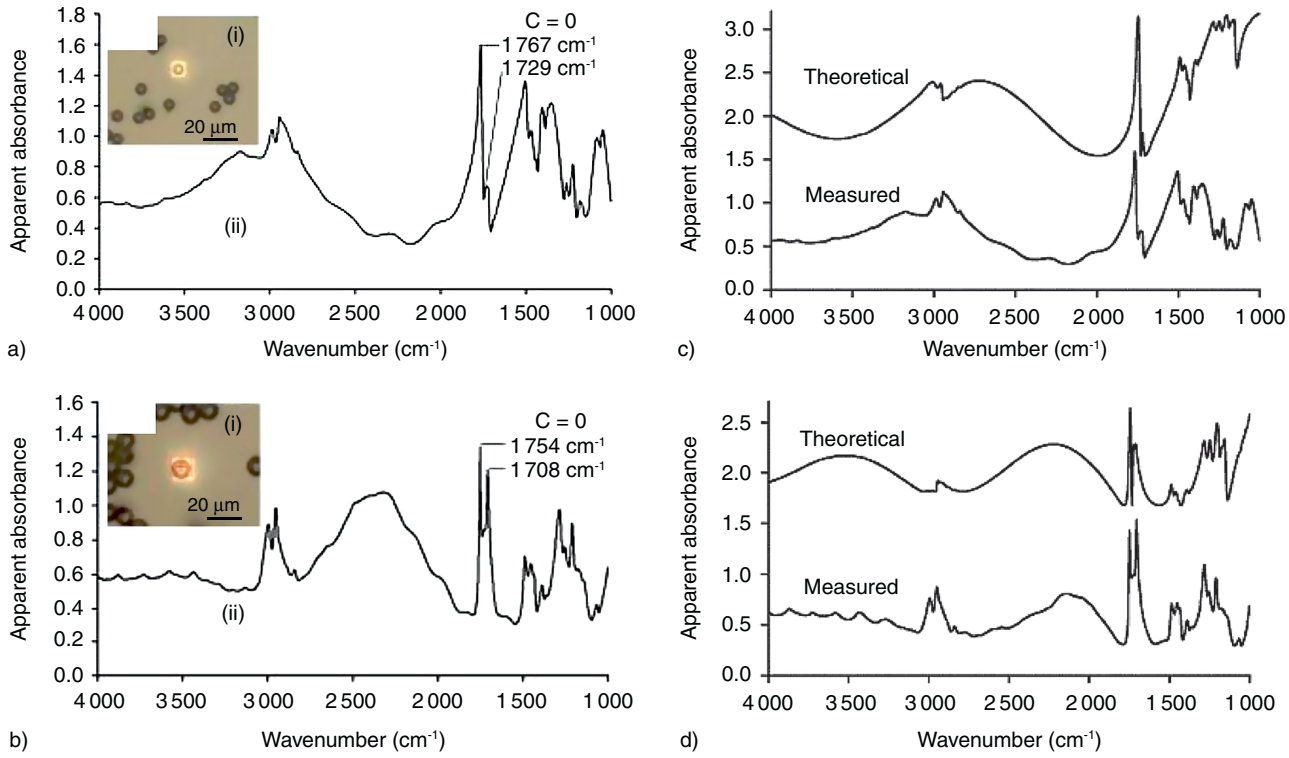

Figure 4

Infrared spectra of polymethylmethacrylate spheres a) 10.8 and b) 15.7 µm in diameter (optical images inset) showing apparent differences in the infrared adsorption bands despite the consistent chemical composition. The measured spectra are compared to the theortically calculated Mie spectra in for the c) 10.8 and d) 15.7 mm spheres, showing that the Mie distortions are accurately modelled. Figure adapted from Bassan *et al.* (2009) [36].

of backgrounds do not exist, a panel of different curves must be prepared blind and it may not be possible to achieve a similar level of consistency without additional optimisation [38, 39].

A number of other approaches has been explored to exploit multivariate analysis for correcting the background signal without the need for prior understanding of the structure of the interfering backgrounds. Balcerowska and Siuda [40] published a method of analysing the loadings of a multivariate analysis (from XPS data) to determine what regions of the spectrum were reliably free of any signal features of interest. Rather than further exploiting the loadings by performing the correction on them, the knowledge gained from assessing the PC was used to define the baseline regions that were used to estimate the backgrounds on the original signals. A closer step to exploiting the potential of multivariate analysis to correct background signals was taken by Marbach *et al.* [41]. This approach took advantage of an experimental method of artificially manipulating the relative contribution of the background to the signal. By manipulating the background, it was possible to magnify its contribution to the dataset variation and allow multivariate methods to separate it from the chemical signals of interest. Once this was done, a modification of the denoising process is used to reconstruct the data without the principal component that accounted for the background. The method is hindered by two main problems; first, not many signals can have their background manipulated easily and secondly the manipulation did not necessarily affect all sources of background variation equally and so, at best, a reduction in background contribution could be achieved.

Glenn *et al.* [43] published a method that exploited the principal components more directly for estimating the broad bandwidth backgrounds that contribute to a dataset where the signal of interest had medium bandwidths. These backgrounds were estimated on the PC and fed into a database that was used to perform a linear combination analysis on the individual spectra. The linear

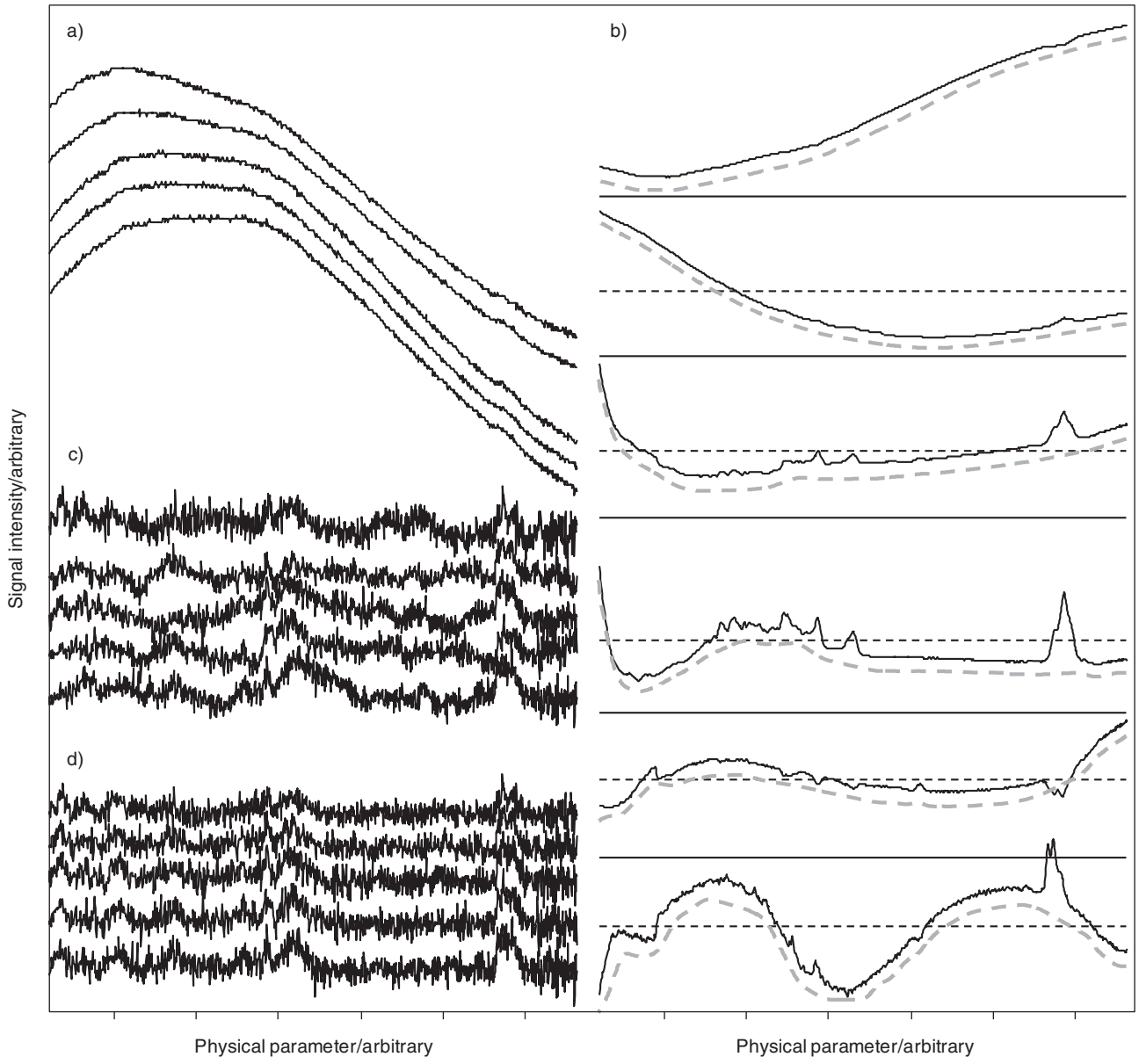

Figure 5

Multivariate based background correction of low SNR signals a) 5 raw signal replicates of one sample at one location, SNR 0.8 b) PCA loadings derived from a dataset of 12 000 signals with the estimate baseline contribution to the loadings indicated by the dashed grey lines, which have been offset for clarity, c) the individual replicates baseline corrected using a linear interpolation of the individual signals, d) the individual replicates baseline corrected using linear interpolation of the PC loadings. The signals have been offset for clarity. Adapted from Beattie [44] JRS.

combination of the PC backgrounds was then subtracted from the spectra to leave baseline corrected data. This approach was computationally intensive and at risk of local, non-optimal, solutions. Two studies developed this approach to use the scores of the multivariate analysis to calculate the combination of backgrounds required, eliminating the calculation of the linear combination and streamlining the process [44, 45], since applied to two new datasets [16, 46]. As indicated above, the loadings are the score weighted average of the dataset, which means they bear a simple relationship with the original signals. Consequently any signal manipulations

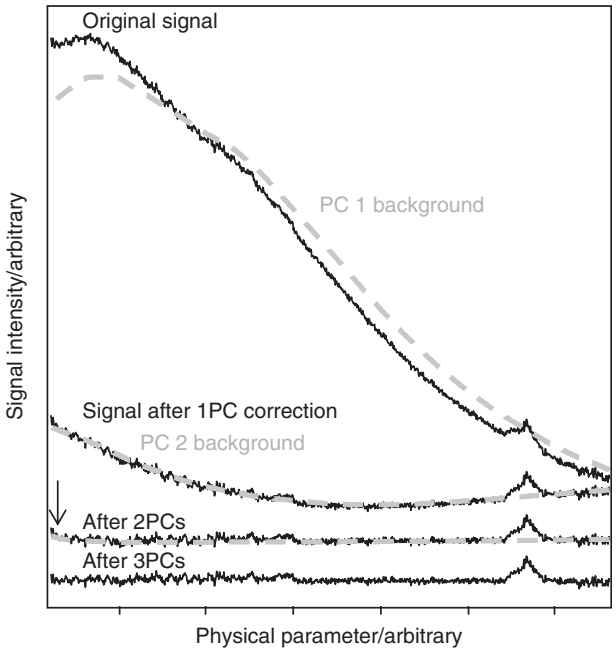

Figure 6

The effect of applying sequential PC based background corrections to a noisy signal containing multiple background sources.

applied to these loadings will have a simple relationship to the original signals, one defined by the scores (these are the measure of the contribution of the loading to the signals). In the papers by Beattie [44], Beattie and McGarvey [39] and by Palacký *et al*. [45] standard baseline estimation methods similar to local minima interpolation were exploited to estimate the background within each loading. Some loadings were clearly peaks in one direction superimposed on a broad bandwidth background and the local minima between the peaks were used to define the baseline points. Some higher order PC showed positive and negative features and the average of maxima and adjacent minima were used to estimate the baseline points.

Figure 5 shows the raw data acquired from the same sample type illustrated in Figure 3, sharing the same signal of interest (*Fig. 3c*) [16]. The informative part of the signal is medium-bandwidth which is superimposed on a variable broad bandwidth background. While the background is variable, it can be described by a small number of PC indicating that it arises from multiple consistent and reproducible sources (*Fig. 5b*). Fitting the background shape to the loadings means that the scores from those loadings will scale the background onto the original data. The process is illustrated in Figure 6; the background estimated from the first PC removes most of the broad bandwidth background but leaves substantial curvature. The second PC estimates the majority of this residual background, with only a minor curvature (arrowed) at the left hand side remaining. Application of the third PC eliminates even this small curvature.

The use of PC to estimate the background confers a number of important benefits compared with estimation of background on individual signals. First, the multivariate nature conveys a significant insensitivity to noise, allowing reliable reproducible extraction of information from much poorer signals, which can clearly be seen comparing the repetitions in Figure 5c (standard linear interpolation on individual signals) and Figure 5d (linear interpolation on the loadings). The data (mean SNR of 3, 4.8 dB) is just barely classified as a signal and not surprisingly estimating the background on each signal leaves signals with severe baseline distortions, with a very high level of irreproducibility between them. In contrast, the SVD baseline corrected data is much more consistent. In fact, the reproducibility of a signal baseline corrected using per-signal linear interpolation is inversely related to the square of the signal to noise ratio, *i.e*. the variability of the baseline estimation on individual signals is directly affected by the noise causing a multiplicative propagation of errors significantly deteriorating the analytic quality of the signal. In contrast the SVD based linear interpolation showed reproducibility that was inversely related to the signal to noise ratio, *i.e*. there was no significant variability associated with the background correction and therefore no propagation of error. As a consequence, the reproducibility of the SVD based approach at a SNR of 3 (4.8 dB) is 16 times better than using the per-signal approach. Because confidence intervals depend on the square root of the number of repetition measured, 256 repetitions would be required to obtain the same level of confidence in per-signal baseline correction as would be found in one repetition processed by SVD baseline correction.

Figure 7 illustrates another benefit of exploiting SVD loadings for the purposes of baseline correction. This dataset, from Glenn *et al*. [43] consists of Raman spectra measured on a thin collagenous membrane from the rear of the retina (light sensitive part of the eye). Beattie and McGarvey [39] used this dataset to compare a range of different background correction procedures (including linear interpolation, sequential polynomial, SVD based baseline correction and EMSC). The most consistent correction method was found to be SVD based baseline correction. The raw data (*Fig. 7a*) exhibit substantial variation in both the broad bandwidth non-Raman background and in the Raman features themselves. Use of per-signal linear interpolation removes much of the broad background but induces a severe dip in one

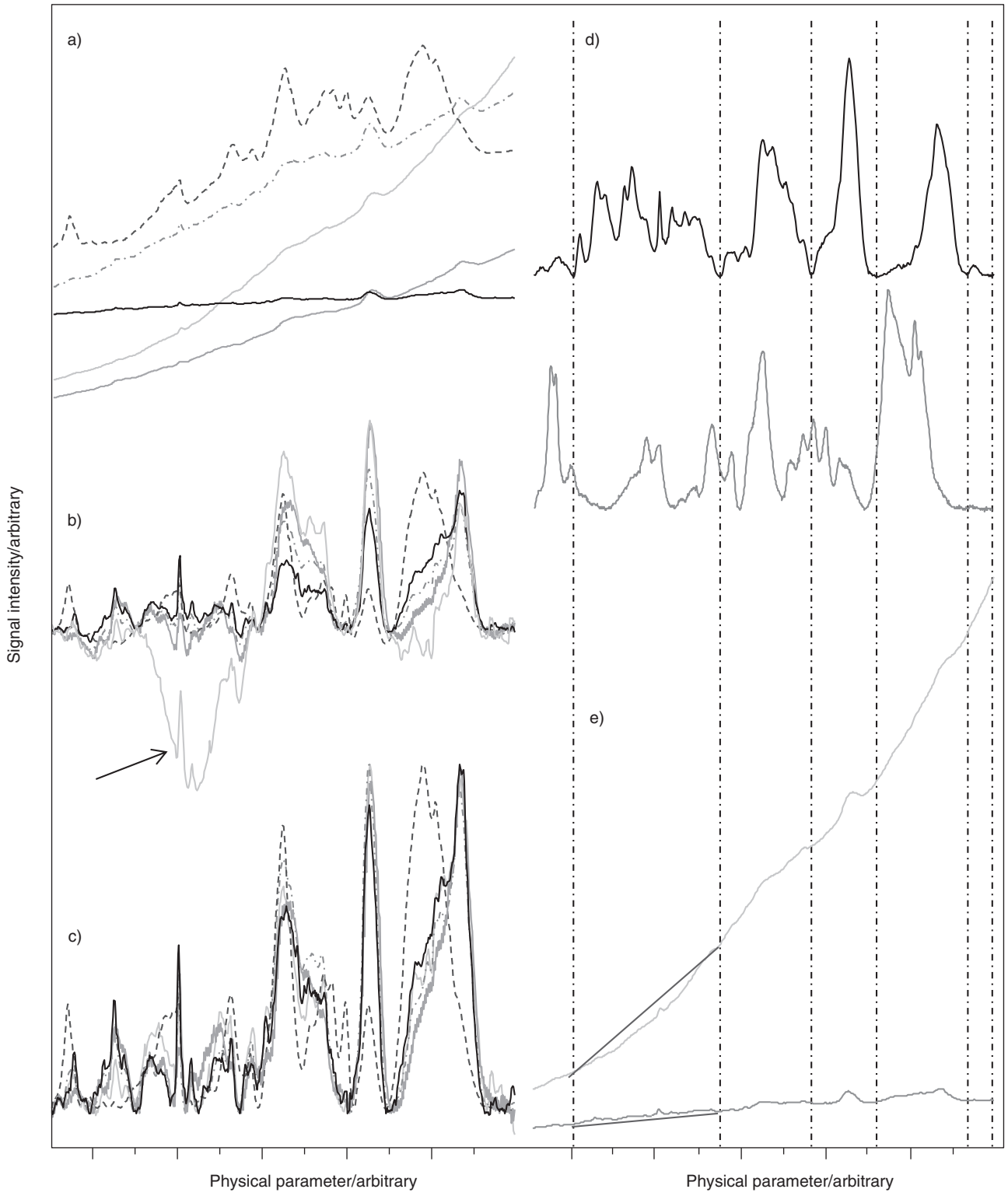

Figure 7

Multivariate based background correction of data with complex variation in both background and signal of interest a) raw signals of five different samples b) signals baseline corrected using a linear interpolation of the individual signals c) the signals baseline corrected using linear interpolation of the PC loadings d) signal obtained from two pure reference compounds of the two major constituents of the samples, vertical dashed lines indicate local minima used in the linear interpolation e) raw signals of samples with similar signals of interest but different backgrounds. Adapted from Beattie and McGarvey [39] JRS.

spectrum (arrowed in *Fig. 7b*). There are also many much smaller dips close to the baseline points, with the result that much inconsistency remains in the 'corrected' data. The mean SNR was 70 (18 dB) so noise could not be blamed for the poor reproducibility of the correction procedure. In contrast, the SVD-baseline corrected data (*Fig. 7c*) shows much more consistent level baselines after correction.

Two reasons were determined to be responsible for this improvement using the SVD based approach. Figure 7d shows the reference spectra for the two main spectral constituents of the dataset, collagen and heme, with the local minima for collagen indicated by the vertical dashed lines. It is clear that the collagen and heme do not share many common baseline points. Where the collagen local minima coincide with the side of a heme peak then the portion of the peak outside these minima is below that intensity and becomes a local dip. The authors tested an adaptive local minima algorithm but found this performed even worse. This however was not sufficient to account for the very large dip in the spectra of Figure 7b, rather Figure 7e compares the corresponding raw signal with another raw signal that contained a very similar Raman pattern, but a very different background shape. The signal that produced the good baseline corrected signal has a much flatter and monotonic background than the one that resulted in the dip. The dip coincides with a region in which the background slope is increasing at an accelerated rate but there is a broad region in which no useful local minima occur.

The SVD based approach is able to handle these more complicated scenarios for the simple reason that the individual PC isolate these variances individually allowing tailoring of the baseline estimation step to both the shape of the background and to the exact signal of interest. This makes it possible to estimate widely differing background shapes but calculate their combination very accurately. In addition, the signals of interest are extracted individually allowing the influence of each one to be handled individually.

4 INTERFERING SIGNAL REMOVAL: ELIMINATING UNDESIRABLE SIGNALS

So far we have looked at using multivariate analytical techniques eliminating low bandwidth (pixel to pixel) and high bandwidth non-signal features that are unrelated to the signal of interest. However, many signals and situations involve interference from features of similar bandwidth to the target signal, *e.g.* contaminants, substrate signals, sample matrix constituents etc. If these signals are more prevalent and variable in the dataset than those of the target analyte, they can readily swamp the target analyte, making its detection difficult and reducing analytical precision. If there are multiple sources of such interferents, then the problem must be addressed using methods capable of handling complex information of multiple variables, for which multivariate analyses offers a logical pathway.

Wold *et al.* [47] introduced Orthogonal Signal Correction (OSC) as a means to eliminate such interfering dominant signals. The principle is simple: it identifies any multivariate signals that are unrelated to the target analyte and removes its contribution from the individual signals, leaving only signals that are correlated with the target analyte. In order to achieve this, it is important to have independent information on the target analyte in the form of preparation quantities or reference concentrations determined by a gold standard method. A PLS data reduction is applied to the data to reduce the data based on covariance between the signal (X) and the reference (Y) data, enabling quantification of the correlation between the two. Dominant basis signals that account for substantial amounts of variation in the original signals but are uncorrelated to the Y values are eliminated from the original signals (using the scores to determine the proportion). The corrected data no longer contains any information on dominant signals that are uncorrelated with the target analyte. The original algorithm was used as a pre-processing step and not subject to cross validation with the consequence that application to new data was not possible with a high degree of certainty, while a more theoretically pure algorithm proposed by Fearn [48] was unable to provide any improvement in regression performance. An updated algorithm from Trygg and Wold [49] incorporates the process more fully into the PLS regression model allowing better cross validation and also an alternative algorithm based solely on least squares steps called Direct OSC by Westerhuis *et al.* [50]. The main benefit of OSC is the simplification of PLS regression models, allowing simpler interpretation. Additionally, it is possible to investigate the uncorrelated signals extracted from the dataset in order to learn more about the sources of interference and thereby implement steps in the analytical process to circumvent the appearance of these interferents. The scope of this manuscript does not permit a detailed comparison of the various algorithmic variants of OSC, but the readers are directed towards the extensive literature [47-52].

Using the DOSC algorithm, Lin *et al.* [55] found that they were able to improve the classification of geographic origin of olive oils based on their NIR spectra. In the study, the group used both leave-one-out cross validation (on 66% of the data) and test set validation

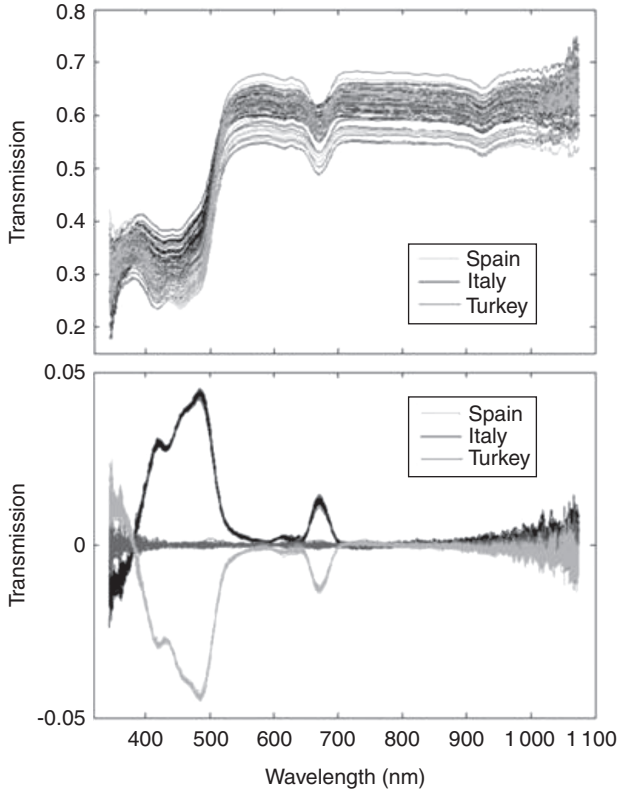

Figure 8

a) Raw NIR signals from olives oils of different geographical origins as indicated, b) OSC corrected NIR spectra of the same oils.

(on the remaining 33% of the data) to assess the impact of DOSC on the analytic performance of the dataset. Figure 8a shows the NIR spectra acquired from 30 spectra of oils from each geographical origin tested. There is considerable overlap between the oils of each country and the correct classification rate based on the uncorrected data is 70% and depends on 4 factors. In contrast the DOSC corrected data exhibited clear differences between each country (*Fig. 8b*) which results in an improved classification rate of 97% and a simplification of the regression model to 1 factor. The authors found that an improvement in performance based on cross validation held true for the test set, but the relative deviation of prediction (test set, RDP = 0.087) was an order of magnitude greater than the relative deviation of cross validation (training set, RDCV = 0.008) indicating that the cross validation was very optimistic of the improvement achieved, underscoring the importance of properly validating the performance of multivariate based methods. This concern for taking care to avoid overfitting the data, was echoed by Zhang *et al.* [53].

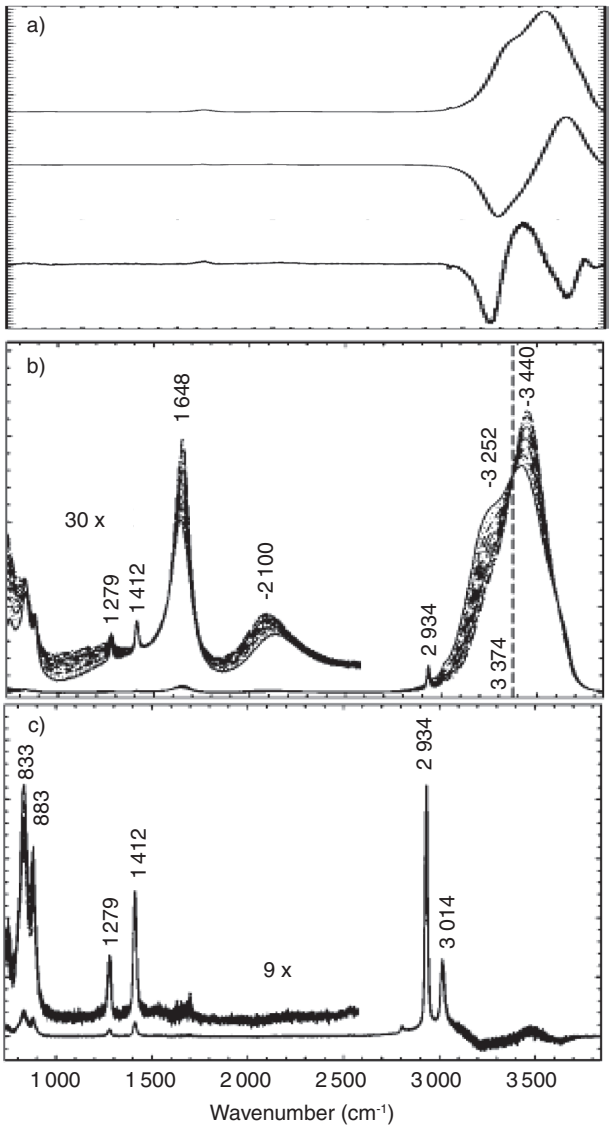

Figure 9

a) PCs extracted from the Raman spectra of deionised water at varying temperature (1-95°C). Raman spectra of aqueous 100 mM Na + -cacodylatein solutions of different ionic strengths adjusted by NaCl a) before and b) after SVD base solvent signal subtraction. Adapted from Palaký *et al.* (2011) [45].

OSC based methods have found widespread use in NMR [54], NIR [55, 56], infrared [57], UV/visible [55, 58], and fluorescence spectroscopies [51], chemical information for Quantitative Structure-Activity Relationship (QSAR) [59] and chromatography [60].

The principle of eliminating a broad featureless interfering signal can quite readily be tweaked to the principle of eliminating a structured signal of interest. In EMSC,

this elimination is achieved by creating an additional database of reference signals, but one of interfering substances that can be combined with the database of background signals and so both structured and unstructured signals can simultaneously be corrected. In the same way, an interfering signal can be added onto the estimated background from the loadings in the SVD based approach, so that the loading-backgrounds now containing the structured signal of the interferent will then be scaled to the original signal by the scores. Palaký *et al*. demonstrated this process for Raman spectra of aqueous cacodylate, using the procedure to reproducibly eliminate the signal of water that dominated the original signals [45] as shown in Figure 9. The problem of subtracting off the spectrum of water from a Raman spectrum is that the shape of the Raman bands is influenced by solvent-solute interactions, ionic strength, pH, temperature etc. and subtracting off an average signal for water is unable to handle the variations induced in the water by these perturbations. By creating a multivariate model of water under a number of perturbed conditions (temperature, pH and ionic strength), it is possible to generate signals corresponding to these perturbations (*Fig. 9a*, shows the PC loadings extracted from water at temperatures between 1 and 95°C) and fit them onto the target signals. Figure 9b shows the Raman spectra for 100 mM solutions of Na^+-cacodylate in which the ionic strength of the solution has been manipulated by NaCl, inducing significant variations in the features that arise due to water in the spectrum. By applying the model based on ionic strength variations, it was possible to very reproducibly eliminate the contribution of water to the signal (*Fig. 9c*).

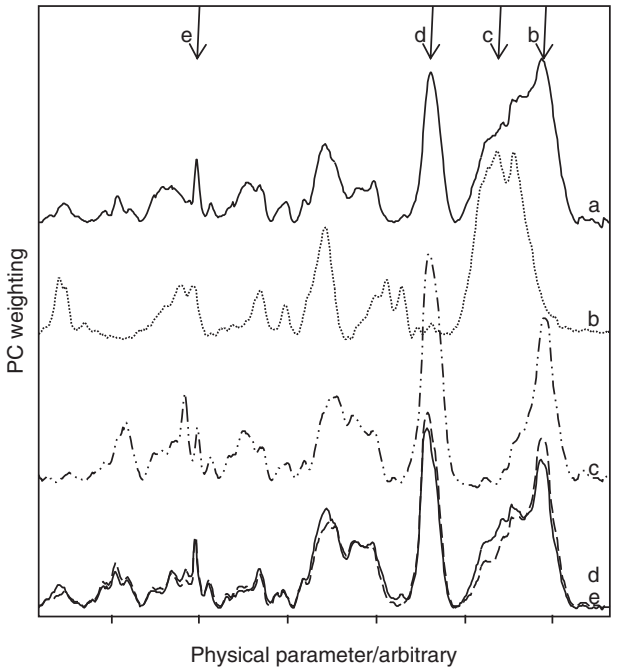

Figure 10

The effect of different normalisation routines on the 1st PC returned from a 'real-world' dataset. a) No normalisation b-e) have been Normalised as indicated by the arrows; d) and e) are overlaid as these bands are used for equivalent normalisations in literature.

5 SIGNAL INTENSITY/AMPLITUDE NORMALISATION/ STANDARDISATION

For many signals the amplitude/intensity of the signal can be calibrated consistently during measurement, *e.g*. through the use of a parallel reference beam, but in many others it is not possible to control the absolute signal intensity with sufficient reproducibility for analytical purposes. In 'real-world' applications, it is typical to encounter complex mixtures of different signals that do not share a common feature suitable for standardising the intensity of the signal [61]. Beattie *et al*. [62] showed that selection of different signal features for intensity normalisation had a very profound effect on the variation within a 'real-world' dataset and the results obtained from the analyses of this data, as illustrated in Figure 10 where the 1st PC loading after different normalisation routines are compared and exhibit wide variation in bandshape. Choosing one channel (or the average of a few adjacent channels) from the signal for calculating the intensity of the signal leads to the risk of very complex scaling unless the feature exists in a channel that exhibits zero contribution from any other constituent. For example, the two normalisation bands chosen for Figure 10d and e are typical bands used to normalise the spectra of proteins, but give noticeably different results for PC1. In situation where no one channel can be attributed to the desired internal standard alone, it would be preferable to exploit methods that can unmix the spectra as a whole, *i.e*. multivariate data reduction.

If a good band area normalisation is known to exist it is possible to use PLS to generate a multivariate regression model between the unnormalized dataset and the scaling factors that are calculated from those spectra. The application of the PLS model to calculate the scaling factors resulted in a halving of the prediction variance compared with using traditional band area [44]. This PLS based approach does not solve the issues raised by Beattie *et al*. [62] regarding the unpredictability of normalisation by band area, but has the advantage of

allowing established normalisation routines to benefit from the noise insensitivity, enhanced specificity etc that comes from using multivariate based processing.

Multiplicative Scatter Correction was originally intended to eliminate errors due to intensity variations by standardising the magnitude of each signal to the mean intensity. As an extension of MSC, EMSC retains this function but has the capacity to perform this correction in a more subtle manner. As an integral part of EMSC the user defines ideal target spectra (in the absence of a specified signal the mean of the dataset is selected as the target). It is possible to define more than one target signal for EMSC, allowing the algorithm to handle signals with widely differing features and simultaneously normalise each signal.

Because multivariate analysis untangles the constituent signals of a dataset, it opens up the possibility of a much more selective calculation of signal magnitude. Bassan *et al*. [28] exploited SVD to selectively normalise Raman spectra of aqueous solutions to the main band arising from water. The scores calculated upon applying the first PC were used to normalise the data to the water mode (data in *Fig. 9b*). If the PC can readily be interpreted then it is conceivable that a linear combination of relevant PC could be calculated to provide highly targeted normalisation. As an example, in many oil systems a range of different aliphatic and aromatic hydrocarbons and other constituents/contaminants/active molecules with functional groups may be present. Standard analyses may involve calculating the ratio of an analyte to the total hydrocarbon or some similar parameter. By identifying which PC contain contributions from different species of hydrocarbon, it would be possible to calculate the overall contribution of the class of molecule based on the linear sum of the individual species present in the dataset.

6 APPLICABILITY OF MULTIVARIATE ANALYSIS FOR SIGNAL PROCESSING

The multivariate based techniques discussed above can be applied in different ways; the multivariate model can be applied internally to the dataset or it can be applied to external data, or the model can be built on multiple measurements from one subject or measurements from multiple subjects. There is a great deal of flexibility in how one can generate a dataset suitable for multivariate processing, but there are some very important ground rules. First, the variation must be present: multivariate analysis is based on detection of variance and without it there can be no multivariate analysis. Secondly, the variation must be appropriate – the full range of variation that the method is expected to handle should be included in the dataset so that it can be modelled. The second requirement raises the concerns regarding validation of the multivariate model. As was seen in the case of OSC validation of multivariate based operations can be a very real concern as transfer of an established OSC model to new data is not always successful. The same concern applies to all multivariate based methods – if it is to be used on new data then the model needs to be validated for use on new data. If the model is to be used only internally on the dataset then an independent test set may be excessive for validation, but rigorous cross validation should be employed to ensure consistency across the dataset and to prevent outliers from distorting the model.

As indicated above, one traditional use of multivariate analysis is the creation of mathematical models that can be applied to new data. A large dataset of different samples is investigated, taking care to cover as much of the overall population variation as possible. If validation confirms that the dataset comprehensively covers the population variance then the model can readily be used on any new samples within that population to extract the same information or carry out the same signal processing procedures [44] without the additional error associated with new estimations.

Alternatively, the multivariate analysis can be used within-dataset, or even within-sample for very complex sampling procedures such as PET [21]. PET responses are consistent within a patient, but the 'resting-state' responses (in this context 'resting-state' refers to the state before any experiment is commenced) can vary enormously between patients, while the data is a large and complex 4D array of signals. This inter-patient discontinuity makes multivariate analysis across different patients less useful, but multivariate analysis exploiting the variation provided by the spatial and temporal variation within the patient's data can provide a very powerful mechanism for signal processing. Wack and Badgaiyan [19] used SVD based denoising within each patient's scans to very effectively denoise the data since the overall signal to noise in the entire dataset was much greater than the individual signals. Figure 11 shows how effective this process can be for individual samples, with the SVD denoising process showing a large increase in contrast ratio of the PET frames (*Fig. 11*, i *versus* ii and iv *versus* v) and a considerable reduction in the scatter for single voxel measurements (*Fig. 11*, vi).

It is important to understand the need for variation in the dataset if multivariate analysis is to be used for signal processing; if there is no variation, the multivariate analysis will be equivalent to simple averages and no gain will be found. Thus multiple signals, recorded in such a way

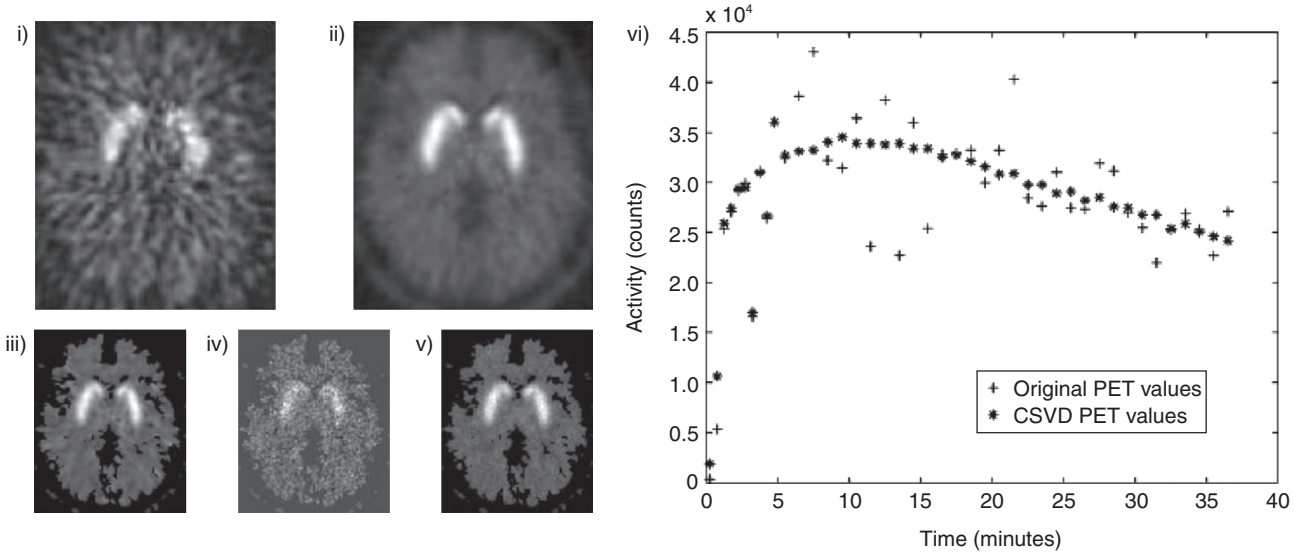

Figure 11

A signal frame of a PET scan i) the original frame ii) after SVD denoising iii) noiseless model of the scan iv) noisy model of the scan (equivalent to iii with noise added) v) SVD denoisednoisy model. vi) the PET values from a single voxelagainst time of scan, the SVD denoiseddata shows significantly less scatter. Adapted from Wack and Badgaiyan (2011) [21].

as to capture different aspects of the samples, are essential. This variation can be obtained in a number of ways. The classical method is to build a sample-set of multiple samples, possibly subjecting these to particular treatments that induce variation or selecting from a naturally variable population. However, it is not necessary to have multiple samples – it is possible to build up variation with multiple measurements of one sample, similar to the PET example above. This variability can be induced from different areas of a heterogeneous sample (variation comes from sampling position), a time series of a dynamic system or measurements of the sample under slightly different conditions. It is important to realise that the individual signals do not need to be of the same quality as for classical analysis; the multivariate analysis will depend on the overall signal to noise ratio of the entire dataset. This makes it possible to achieve sufficient quality of data by retaining the same overall acquisition time but taking multiple shorter acquisitions. In most analytical methods multiple measurements will be the normal course of action and this can be exploited by retaining these repetitions as separate signals until processing is complete.

Table 1 summarises the applicability of the main types of multivariate signal processing to achieve correction of different types of variation. The techniques can be combined in order to achieve elimination of all necessary sources of variation. For example, eliminating the recurrent non-signals with multivariate baseline correction will leave irreproducible non-signals, amplitude variation and recurrent signals (including any unwanted signals). The recurrent unwanted signals can then be eliminated by contaminant elimination, leaving the amplitude variation and irreproducible non-signal. Application of multivariate amplitude normalisation will then eliminate the variation of intensity between signals leaving just the irreproducible non-signal that can be eliminated by multivariate denoising if desired (see above). Note that because the scores contain exactly the same information as multivariate denoised spectra there is no benefit to carrying out multivariate denoising prior to any subsequent multivariate processing and if it is to be utilised, should always be left to the final stage.

The remaining question is what types of signal would multivariate based signal processing be applicable to? Table 1 will give some indication of applicability, but some further expansion is warranted. Clearly any signal that can usefully be analysed by multivariate data reduction (PCA, PLS, etc.) would also be suited to multi variate based signal processing. These signals are characterised by reproducible signals whose features appear at consistent channels (before or after any necessary preprocessing). However, many signals that do not currently exploit PCA and related methods may yet benefit.

Sporadic irreproducible signals may potentially benefit if they meet two criteria:
– they exhibit recurrent sources of interfering signals (and so the interferences can be statistically modelled);
– the sporadic signal accounts for a relatively minor proportion of the signal intensity when it does appear (*i.e.* the presence of the signal will not significantly interfere with applying the corrective model to the signal).

CONCLUSION

A range of studies has demonstrated the profound advantages of multivariate analysis in both quantitative and qualitative analysis; noise cancelling, specificity, adaptability, comprehensiveness. Further studies have demonstrated that these very attributes can be exploited in other ways, not just for quantification. The quantitative information extracted *via* multivariate data reduction can be exploited to direct signal manipulations with high fidelity, high specificity and offering adaptation to widely varying signals.

The nature of multivariate analysis, involving deconstruction of a dataset of signals using only the information present within that dataset, has been historically under-appreciated. There is a very strong link between the original signals and the signals extracted from the multivariate analysis that can be exploited for signal based manipulations. Any transformations applied to the loadings can be transmitted back to the original signals using the inherent scaling information contained in the scores associated with the loadings.

The principle of using multivariate analyses to correct signal perturbations is much more widely applicable that has been exploited to date. The multivariate based approach will not work on all data types, but it can be applied to correct data where either (or both) the signal of interest or the interfering signals are recurrent in nature and can be summarised by a data reduction method such as PCA.

ACKNOWLEDGMENTS

The author would like to thank the considerable number of researchers who provided advice and guidance during the preparation of this manuscript. Profs Svante Wold and Tom Fearn were very accommodating in discussing aspects of OSC, Profs Peter Gardner and Tom Bakker-Schutt provided useful insights into EMSC, Drs Ioan Notingher and Francis Esmond-White provided useful discussions on multivariate denoising. Prof Hannu Toivonen was very accommodating in discussions on the underlying mathematics of PCA. Finally I need to thank Dr Laurent Duval for his considerable assistance as editor in bringing this manuscript to fruition, without his input and many introductions to additional fields of application the manuscript would be a shadow its current self.

REFERENCES

1 Eriksson L., Johansson E., Kettaneh-Wold N., Trygg J., Wikström C., Wold S. (2006) *Multi- and Megavariate Data Analysis Part I: Basic Principles and Applications, Second revised and enlarged edition Vol. 1*, 2nd ed, Umetrics, Umea.

2 Eriksson L., Johansson E., Kettaneh-Wold N., Trygg J., Wikström C., Wold S. (2006) *Multi- and Megavariate Data Analysis Part II: Advanced Applications and Method Extensions, Second revised and enlarged edition Vol. 2*, 2nd ed, Umetrics, Umea.

3 Orfanidis S.J. (accessed 2013) SVD, PCA, KLT, CCA, and All That, in *332:525 Optimum Signal Processing*, available at: http://www.ece.rutgers.edu/~orfanidi/ece525/svd.pdf.

4 Bonnier F., Byrne H.J. (2012) Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems, *Analyst* **137**, 322.

5 Gerbrands J.J. (1981) On the Relationships between Svd, Klt and Pca, *Pattern Recognition* **14**, 1-6, 375.

6 Wall M.E., Rechtsteiner A., Rocha L.M. (2003) Singular value decomposition and principal component analysis, in *A Practical Approach to Microarray Data Analysis*, Berrar D.P., Dubitzky W., Granzow M. (eds), Kluwer, Norwell, MA.

7 Elliott M.A., Walter G.A., Swift A., Vandenborne K., Schotland J.C., Leigh J.S. (1999) Spectral quantitation by principal component analysis using complex singular value decomposition, *Magnetic Resonance in Medicine* **41**, 3, 450.

8 Brand M. (2003) Fast online SVD revisions for lightweight recommender systems, in *Proceedings of the Third Siam International Conference on Data Mining*, San Francisco, CA, 1-3 May.

9 Toivonen H. (2012) Some multivariate signal processing operations, in *Applied Signal Processing*, available at: http://users.abo.fi/htoivone/courses/sbappl/asp_chapter5.pdf.

10 Clark D., Sasic S. (2006) Chemical images: Technical approaches and issues, *Cytometry Part A* **69A**, 8, 815.

11 Sasic S., Clark D.A. (2006) Defining a strategy for chemical imaging of industrial pharmaceutical samples on Raman line-mapping and global illumination instruments, *Applied Spectroscopy* **60**, 5, 494.

12 Sasic S., Clark D.A., Mitchell J.C., Snowden M.J. (2004) A comparison of Raman chemical images produced by univariate and multivariate data processing - a simulation with an example from pharmaceutical practice, *Analyst* **129**, 11, 1001.

13 StatSoft (accessed 2013) Electronic Statistics Textbook: Partial Least Squares (PLS), in, available at: http://www.obgyn.cam.ac.uk/cam-only/statsbook/stpls.html.

14 Shin K., Hammond J.K., White P.R. (1999) Iterative svd method for noise reduction of low-dimensional, chaotic time series, *Mechanical Systems and Signal Processing* **13**, 1, 115.

15 Uzunbajakava N., Lenferink A., Kraan Y., Volokhina E., Vrensen G., Greve J., Otto C. (2003) Nonresonant confocal Raman imaging of DNA and protein distribution in apoptotic cells, *Biophysical Journal* **84**, 6, 3968.

16 Beattie J.R., Pawlak A.M., McGarvey J.J., Stitt A.W. (2011) Sclera as a Surrogate Marker for Determining AGE-Modifications in Bruch's Membrane Using a Raman Spectroscopy-Based Index of Aging, *Investigative Ophthalmology and Visual Science* **52**, 3, 1593.

17 Ghita A., Pascut F.C., Mather M., Sottile V., Notingher I. (2012) Cytoplasmic RNA in Undifferentiated Neural Stem Cells: A Potential Label-Free Raman Spectral Marker for Assessing the Undifferentiated Status, *Analytical Chemistry* **84**, 3155.

18 Walton J., Fairley N. (2005) Noise reduction in X-ray photoelectron spectromicroscopy by a singular value decomposition sorting procedure, *Journal of Electron Spectroscopy and Related Phenomena* **148**, 1, 29.

19 Mauldin F.W., Lin D., Hossack J.A. (2011) The Singular Value Filter: A General Filter Design Strategy for PCA-Based Signal Separation in Medical Ultrasound Imaging, *Ieee Transactions on Medical Imaging* **30**, 11, 1951.

20 Yuan S.Y., Wang S.X. (2011) A local f-x Cadzow method for noise reduction of seismic data obtained in complex formations, *Petroleum Science* **8**, 3, 269.

21 Wack D.S., Badgaiyan R.D. (2011) Complex Singular Value Decomposition Based Noise Reduction of Dynamic PET Images, *Current Medical Imaging Reviews* **7**, 2, 113.

22 Patel V., Shi Y.G., Thompson P.M., Toga A.W. (2011) K-Svd for Hardi Denoising, *2011 8th IEEE International Symposium on Biomedical Imaging: from Nano to Macro*, Chicago, 30 March-2 April.

23 Jha S.K., Yadava R.D.S. (2011) Denoising by Singular Value Decomposition and Its Application to Electronic Nose Data Processing, *IEEE Sensors Journal* **11**, 1, 35.

24 Liu B.Y., Liao X. (2009) Image Denoising and Magnification *via* Kernel Fitting and Modified SVD, *Fifth International Conference on Information Assurance and Security, IAS '09, Xi'An*, China, 18-20 Aug.,Vol. **2**, Proceedings, pp. 521-524, http://dx.doi.org/10.1109/IAS.2009.29.

25 Nazari B., Sarkrni S.M.A., Karimi P. (2009) A Method for Noise Reduction in Speech Signal Based on Singular Value Decomposition and Genetic Algorithm, *Eurocon 2009: International IEEE Conference Devoted to the 150 Anniversary of Alexander S. Popov*, Vol. **1-4**, Proceedings.

26 van't Hoff M., Reuter M., Dryden D.T.F., Oheim M. (2009) Screening by imaging: scaling up single-DNA-molecule analysis with a novel parabolic VA-TIRF reflector and noise-reduction techniques, *Physical Chemistry Chemical Physics* **11**, 35, 7713.

27 Borglund N., Astrand P.G., Csillag S. (2005) Improved background removal method using principal components analysis for spatially resolved electron energy loss spectroscopy, *Microscopy and Microanalysis* **11**, 1, 88.

28 Bassan P., Sachdeva A., Kohler A., Hughes C., Henderson A., Boyle J., Shanks J.H., Brown M., Clarke N.W., Gardner P. (2012) FTIR microscopy of biological cells and tissue: data analysis using resonant Mie scattering (RMieS) EMSC algorithm, *Analyst* **137**, 6, 1370.

29 Bakshi B. (1998) Multiscale PCA with application to MSPC monitoring, *AIChE J* **44**, 1596.

30 Aminghafari M., Cheze N., Poggi J.M. (2006) Multivariate denoising using wavelets and principal component analysis, *Computational Statistics and Data Analysis* **50**, 9, 2381.

31 Chaux C., Duval L., Benazza-Benyahia A., Pesquet J.-C. (2008) A nonlinear Stein-based estimator for multichannel image denoising, *IEEE Transactions on Signal Processing* **56**, 8, 3855.

32 Martens H., Stark E. (1991) Extended Multiplicative Signal Correction and Spectral Interference Subtraction - New Preprocessing Methods for near - Infrared Spectroscopy, *Journal of Pharmaceutical and Biomedical Analysis* **9**, 8, 625.

33 O'Farrell M., Wold J.P., Hoy M., Tschudi J., Schulerud H. (2010) On-line fat content classification of in homogeneous pork trimmings using multispectral near infrared interactance imaging, *Journal of Near Infrared Spectroscopy* **18**, 2, 135.

34 De Gelder J., De Gussem K., Vandenabeele P., De Vos P., Moens L. (2007) Methods for extracting biochemical information from bacterial Raman spectra: An explorative study on Cupriavidus metallidurans, *Analytica Chimica Acta* **585**, 2, 234.

35 Bassan P., Kohler A., Martens H., Lee J., Jackson E., Lockyer N., Dumas P., Brown M., Clarke N., Gardner P. (2010) RMieS-EMSC correction for infrared spectra of biological cells: Extension using full Mie theory and GPU computing, *Journal of Biophotonics* **3**, 8-9, 609.

36 Bassan P., Byrne H.J., Bonnier F., Lee J., Dumas P., Gardner P. (2009) Resonant Mie scattering in infrared spectroscopy of biological materials - understanding the 'dispersion artefact', *Analyst* **134**, 8, 1586.

37 Bassan P., Kohler A., Martens H., Lee J., Byrne H.J., Dumas P., Gazi E., Brown M., Clarke N., Gardner P. (2010) Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples, *Analyst* **135**, 2, 268.

38 Afseth N.K., Segtnan V.H., Wold J.P. (2006) Raman spectra of biological samples: A study of preprocessing methods, *Applied Spectroscopy* **60**, 12, 1358.

39 Beattie J.R., McGarvey J.J. (2013) Estimation of signal backgrounds on multivariate loadings improves model generation in face of complex variation in backgrounds and constituents, *Journal of Raman Spectroscopy* **43**, 2, 329-338.

40 Balcerowska G., Siuda R. (1999) Inelastic background subtraction from a set of angle-dependent XPS spectra using PCA and polynomial approximation, *Vacuum* **54**, 1-4, 195.

41 Marbach R., Tenhunen M., Niemelä P. (2008) Simple and powerful new method for "subtracting" fluorescence backgrounds in Raman spectra, *ICORS*, 1113, London, UK, 17-22 Aug.

42 Beattie J.R., Pawlak A.M., Boulton M.E., Zhang J., Monnier V.M., McGarvey J.J., Stitt A.W. (2010) Multiplex analysis of age-related protein and lipid modifications in human Bruch's membrane, *FASEB Journal* **24**, 12, 4816-4824.

43 Glenn J.V., Beattie J.R., Barrett L., Frizzell N., Thorpe S.R., Boulton M.E., McGarvey J.J., Stitt A.W. (2007) Confocal Raman microscopy can quantify advanced glycation end product (AGE) modifications in Bruch's membrane leading to accurate, nondestructive prediction of ocular aging, *FASEB Journal* **21**, 13, 3542-3552.

44 Beattie J.R. (2011) Optimising reproducibility in low quality signals without smoothing; an alternative paradigm for signal processing, *Journal of Raman Spectroscopy* **42**, 1419.

45 Palacký J., Mojzeš P., Bok J. (2011) SVD-based method for intensity normalization, background correction and solvent subtraction in Raman spectroscopy exploiting the properties of water stretching vibrations, *Journal of Raman Spectroscopy* **42**, 7, 1528-1539.

46 Beattie J.R., Finnegan S., Hamilton R.W., Ali M., Inglehearn C.F., Stitt A.W., McGarvey J.J., Hocking P. M., Curry W.J. (2012) Profiling Retinal Biochemistry in the MPDZ Mutant Retinal Dysplasia and Degeneration Chick: A Model of Human RP and LCA, *Investigative Ophthalmology and Visual Science* **53**, 1, 413.

47 Wold S., Antti H., Lindgren F., Ohman J. (1998) Orthogonal signal correction of near-infrared spectra, *Chemometrics and Intelligent Laboratory Systems* **44**, 1-2, 175.

48 Fearn T. (2000) On orthogonal signal correction, *Chemometrics and Intelligent Laboratory Systems* **50**, 1, 47.

49 Trygg J., Wold S. (2003) O2-PLS, a two block (X-Y) latent variable regression (LVR) method with an integral OSC filter, *J. Chemometrics* **17**, 53.

50 Westerhuis J.A., de Jong S., Smilde A.K. (2001) Direct orthogonal signal correction, *Chemometrics and Intelligent Laboratory Systems* **56**, 1, 13.

51 Eriksson L., Trygg J., Johansson E., Bro R., Wold S. (2000) Orthogonal signal correction, wavelet analysis, and multivariate calibration of complicated process fluorescence data, *Analytica Chimica Acta* **420**, 2, 181.

52 Igne B., Roger J.-M., Roussel S., Bellon-Maurel V., Hurburgh C.R. (2009) Improving the Transfer of Near Infrared Prediction Models by Orthogonal Methods, *Chemometrics and Intelligent Laboratory Systems* **99**, 1, 57.

53 Zhang X., Yuan H.F., Guo Z., Song C.F., Li X.Y., Xie J.C. (2011) Study of the Over-Fitting in Building PLS Model Using Orthogonal Signal Correction, *Spectroscopy and Spectral Analysis* **31**, 6, 1688.

54 Wu Q.F., Guo L.L., Yu S.G., Zhang Q., Lu S.F., Zeng F., Yin H.Y., Tang Y., Yan X.Z. (2011) A (1)H NMR-based metabonomic study on the SAMP8 and SAMR1 mice and the effect of electro-acupuncture, *Experimental Gerontology* **46**, 10, 787.

55 Lin P., Chen Y.M., He Y. (2012) Identification of Geographical Origin of Olive Oil Using Visible and Near-Infrared Spectroscopy Technique Combined with Chemometrics, *Food and Bioprocess Technology* **5**, 1, 235.

56 Zhu W.C., Cheng F. (2012) Analysis of Transgenic and Non-Transgenic Rice Leaves Using Visible/Near-Infrared Spectroscopy, *Spectroscopy and Spectral Analysis* **32**, 2, 370.

57 Versari A., Parpinello G.P., Laghi L. (2012) Application of Infrared Spectroscopy for the Prediction of Color Components of Red Wines, *Spectroscopy* **27**, 2, 36.

58 Khajehsharifi H., Pourbasheer E. (2011) Simultaneous Spectrophotometric Determination of Xanthine, Hypoxanthine and Uric Acid in Real Matrix by Orthogonal Signal Correction-Partial Least Squares, *Journal of the Iranian Chemical Society* **8**, 4, 1113.

59 Andersson P.M., Sjostrom M., Lundstedt T. (1998) Preprocessing peptide sequences for multivariate sequence-property analysis, *Chemometrics and Intelligent Laboratory Systems* **42**, 1-2, 41.

60 Imbert L., Ramos R.G., Libong D., Abreu S., Loiseau P. M., Chaminade P. (2012) Identification of phospholipid species affected by miltefosine action in Leishmania donovani cultures using LC-ELSD, LC-ESI/MS, and multivariate data analysis, *Analytical and Bioanalytical Chemistry* **402**, 3, 1169.

61 Panneton B., Roger J.-M., Guillaume S., Longchamps L. (2008) Effects of Preprocessing of Ultraviolet-Induced Fluorescence Spectra in Plant Fingerprinting Applications, *Applied Spectroscopy* **62**, 7, 747.

62 Beattie J.R., Glenn J.V., Boulton M.E., Stitt A.W., McGarvey J.J. (2009) Effect of signal intensity normalization on the multivariate analysis of spectral data in complex 'real-world' datasets, *Journal of Raman Spectroscopy* **40**, 429.

Manuscript accepted in August 2013
Published online in December 2013