Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

A multivariate statistical investigation of background subtraction algorithms for Raman spectra of cytology samples recorded on glass slides

L.T. Kerr, B.M. Hennelly*

Department of Electronic Engineering, Maynooth University, Ireland

ARTICLE INFO

Article history:
Received 24 May 2016
Received in revised form
16 August 2016
Accepted 22 August 2016
Available online 1 September 2016

Keywords:
Glass
Background subtraction
Cytology
Raman spectroscopy
Extended multivariate signal correction

ABSTRACT

Traditional preparation methods for cytology samples pose a significant problem for Raman micro-spectroscopy, with long-established clinical techniques depositing cells on glass slides. Unfortunately, both the signal from the glass slide and the baseline signal from the cell itself obscure the Raman cell spectrum. The intensity of the glass signal varies from cell to cell depending on morphology, and although smooth, the signal is more complex within the fingerprint region than the baseline, and cannot be easily removed from the Raman spectrum using polynomial fitting techniques. It is difficult to accurately remove both background signals, and therefore, the use of standard glass slides compromises the capability of pre-processing methods to extract reliable and reproducible spectra from biological cells. To avoid this signal, Raman spectra are often recorded from cells on expensive substrates, such as calcium fluoride (CaF_2) or quartz, but this practice is impractical for large scale applications of Raman cytology for diagnostics or screening purposes. This study investigates the potential of a number of background subtraction algorithms to remove both the glass signal and the baseline, and investigates the effect of these algorithms on subsequent multivariate analysis for the purpose of cell classification. This study demonstrates that the well-known extended multivariate signal correction (EMSC) algorithm is particularly effective in this regard, and that the results of subsequent multivariate statistical analysis are independent of the reference cell spectrum used in the algorithm. Matlab code is provided for the implementation of the EMSC algorithm.

1. Introduction

The analysis of cytological samples using Raman micro-spectroscopy has the potential of replacing many invasive procedures, such as endoscopy or biopsy. Raman based diagnostics have received growing interest in recent years, particularly for cervical [1–3], urine [4–6], and oral cytology [7,8]. However, the advancement of Raman spectroscopy into the clinic has been hindered by its incompatibility with inexpensive glass slides that are used as a standard consumable within the cytopathological laboratory.

It is necessary to remove background signals insofar as possible from Raman spectra in order to facilitate an accurate comparison of related cell spectra, and in particular for the application of multivariate classification for the purpose of disease diagnostics or screening. In general, Raman spectra of biological samples contain a broad baseline signal that often varies randomly from one recording to the next. The signal is most often ascribed to an autofluorescence from the sample itself [9]. Although it has been suggested by some authors [10] that it may originate from sample morphology and Mie scattering of the source laser wavelength. Regardless of its origins, various algorithms have been developed to identify and remove this baseline signal from Raman cell spectra, with polynomial fitting techniques being the most common technique used today [11–13].

The preparation of cytology samples poses a significant problem for Raman micro-spectroscopy, with current clinical techniques, such as the ThinPrep or SurePath methods, producing cell samples on glass slides for pathological evaluation. Glass is often a necessary consumable in clinical cytopathology due to its low cost. Unfortunately, the background signal from glass adds to the aforementioned baseline signal to further obscure the weak Raman cell spectrum, thus compromising the ability of Raman micro-spectroscopy to produce reliable and reproducible spectra from biological cells [14]. This is particularly evident in the 1050–1150 cm^{-1} region, where the glass signal is often strongest when recording Raman spectra with a 532 nm excitation source. The spectral profile, location, and intensity of the glass signal are

* Corresponding author.
E-mail addresses: lauratheresekerr@gmail.com (L.T. Kerr), bryanh@cs.nuim.ie (B.M. Hennelly).

http://dx.doi.org/10.1016/j.chemolab.2016.08.012
0169-7439/© 2016 Elsevier B.V. All rights reserved.

dependent on the excitation source, as shown elsewhere [15], with the ability to recover Raman spectral peaks decreasing as the source wavelength moves from the visible to the NIR region. Here, we are only interested in 532 nm laser sources, which produce a relatively low, but still problematic glass background signal. Typically, research groups try to avoid the glass signal by recording Raman spectra from cells deposited on expensive substrates, such as CaF_2, which has a relatively flat background within the fingerprint region [16]. In this paper, a number of different background subtraction algorithms are investigated in order to accurately remove the glass contamination present in Raman cell spectra, as well as the baseline signal. It is demonstrated that EMSC is particularly effective in removing both background signals, as well as having the additional benefit of effectively normalising the corrected spectra, a step that is always required in advance of multivariate classification.

Algorithms have been previously developed to remove spectral contaminants from Raman spectra, based on a variety of techniques. Tfayli et al. [17] reported the removal of paraffin signals from Raman spectra using a combination of independent component analysis (ICA) and non-negative constrained least squares; with other research groups applying a similar technique for the removal of known spectral contaminants, such as pharmaceutical drugs [18], or polystyrene nanoparticles [19], that were present within Raman cell spectra. Beier et al. [20] proposed an algorithm that simultaneously removed the baseline as well as the background signal from a known contaminant based on an iterative polynomial subtraction method [21]. This algorithm has previously produced good results in the removal of the glass signal from Raman spectra of epithelial cheek cells [15], however, following extensive testing, it is reported in this paper, that for certain cell lines of particular morphology, this algorithm can result in overfitting and alteration of key spectral information, which has a negative impact on resultant multivariate classification algorithms.

EMSC algorithms are gaining interest in recent times for the removal of spectral interferents from vibrational spectroscopic data [22]. EMSC can be applied to vibrational spectra to separate between different physical effects based on an ordinary least squares fitting approach [23]. This technique has been applied extensively to Fourier Transform Infrared (FTIR) spectroscopic data to correct for Mie scattering effects [24–26]. Liland et al. [27] recently applied EMSC to fit whole datasets to reference spectra, resulting in the removal of an interfering signal from Raman spectra of adipose tissue. This interferent was due to an optical effect resulting from the Raman system design, presenting in various intensities from spectrum to spectrum, and could not be completely removed using traditional background correction algorithms, such as a modified polynomial in combination with the standard normal variate (SNV). A similar approach is applied here for the removal of the glass signal from Raman cytology spectra. It is believed that this algorithm could help with the advancement of Raman cytology into a clinical setting, allowing for the use of current clinical pathology techniques, such as glass slides. Additionally, the results of the EMSC algorithm, and subsequent Principal Component Analysis (PCA) results, are compared with two other background algorithms.

2. The EMSC algorithm for removal of glass signal

A raw spectrum, **S**, can be described as a linear superposition of the Raman spectrum of interest, **R**, the baseline signal, **B**, and the glass signal, **G**:

$$\mathbf{S} = \mathbf{R} + \mathbf{G} + \mathbf{B} \tag{1}$$

The goal is to estimate the values of **B** and **G** such that they may be subtracted from the recorded spectrum. Although noise will always be present in the raw spectrum [28], it is assumed that the signal to noise ratio is sufficiently high such that the noise signal may be ignored.

A reference spectrum, **r**, is first obtained such that it may be assumed that **R** can be approximated by the product of this reference spectrum and a certain weight:

$$\mathbf{R} \approx c_r \times \mathbf{r} \tag{2}$$

where c_r is a scalar for a given spectrum.

Similarly, by recording a spectrum directly from a glass slide, **g**, it is possible to represent the spectral contribution of glass in the recorded cell spectrum, **G**, as the product of the pure glass spectrum and a certain weight:

$$\mathbf{G} = c_g \times \mathbf{g} \tag{3}$$

It should be noted that both c_r and c_g are scalar values that are unique to each cell spectrum, and are dependent on experimental parameters such as the Raman acquisition time.

The slowly varying baseline **B** can be represented using an appropriate N order polynomial:

$$\mathbf{B}_N = c_0 + c_1\mathbf{x} + c_2\mathbf{x}^2 + \cdots + c_N\mathbf{x}^N \tag{4}$$

where N is the order of the polynomial, and c_m for $m = 0 \to N$ represents the various coefficients in the polynomial [29].

The raw spectrum, **S**, the reference spectrum, **r**, the glass spectrum, **g**, and the order of the polynomial, N, are all input to the EMSC algorithm, which returns estimates for c_r, c_g, and c_m for $m = 0 \to N$. These estimates are based on an optimal fit of the various vectors in Eq. (5) in an ordinary least squares sense [22,27]:

$$\mathbf{S} \approx [c_r \times \mathbf{r}] + [c_g \times \mathbf{g}] + \left[\sum_{m=0}^{N} c_m\mathbf{x}^m\right] \tag{5}$$

The background corrected cell spectrum, **T**, is given by:

$$\mathbf{T} = \frac{\mathbf{S} - [c_g \times \mathbf{g}] - \left[\sum_{m=0}^{N} c_m\mathbf{x}^m\right]}{c_r} \tag{6}$$

The choice of the reference spectrum, **r**, is a subject of particular interest. It is common to set **r** to be equal to the mean spectrum for a given dataset of interest [22,27]. In this paper, however, in order to omit the glass signal from the reference spectrum, spectra are recorded from similar cells on CaF_2 slides, and **r** is taken to be equal to the mean spectrum. The CaF_2 substrate produces a relatively weak background signal, and it can therefore be assumed that $\mathbf{r} \approx \mathbf{r}_{cell} + \mathbf{b}$, where $\mathbf{r}_{cell}$ denotes the true Raman spectral irradiance of the cell on the CaF_2 substrate, and **b** represents a baseline signal that is inherent in the reference spectrum. All corrected spectra therefore will be fit to a reference that includes this baseline signal. The presence of this constant baseline is only a matter of aesthetics since the qualitative and quantitative data within a dataset will be unaffected so long as all of the spectra in the dataset are processed using the same reference spectrum [22]; it has been shown that this constant baseline, therefore, has no effect on multivariate statistical analysis, such as PCA, that follows after processing using EMSC.

The purpose of this paper is to investigate the application of the EMSC algorithm to pre-process Raman datasets recorded from cells on glass slides, in advance of PCA based classification, with a view to understanding the potential of the method for improving the sensitivity and specificity of cytopathology. For such an application, it is important to fully evaluate the effect of different reference spectra on the multivariate classification of pre-processed data, in order to assess any possible biasing of the results.

3. Methods

3.1. Spectral acquisition

Spectra are recorded with a custom built confocal Raman micro-spectrometer operating with a 532 nm laser (150 mW, Torus; Laser Quantum, Cheshire, UK), 50 × microscope objective (50 ×/ 0.8 Olympus UMPlanFl; Olympus Corporation, Japan), and 100 μm confocal aperture. The laser illuminates a 3 μm diameter spot size that is targeted at cell nuclei. Raman scattered photons are collected with a spectrograph (Shamrock 500; Andor Technology, UK) operating with a 600 lines mm^{-1} grating (spectral resolution of 4 cm^{-1} at the centre), and a cooled CCD camera (DU420A-BR-DD; Andor Technology, UK) operating at −80 °C. Spectra are recorded from each cell nucleus with an acquisition time of 5 s each; two spectra are recorded from the same location within the nucleus, and are averaged together using an algorithm that simultaneously removes cosmic rays. [30].

Four datasets of Raman spectra were recorded using the system described above, with 50 cell spectra present in each dataset; (i) T24 high grade bladder cancer cells recorded on CaF_2 substrates (Raman grade CaF_2, Crystran, UK); (ii) MDA-MB-231 triple negative breast cancer cells recorded on CaF_2 substrates; (iii) T24 bladder cancer cells recorded on glass slides; and (iv) RT112 low grade bladder cancer cells recorded on glass slides. All cell lines were obtained from Cell Lines Services GmBH, Germany. The latter two datasets contain the spectra selected for pre-processing using the EMSC algorithm, while the other datasets are used to generate the reference spectra used in the algorithm.

3.2. Background subtraction algorithms

For the sake of comparison, two other well-known background subtraction algorithms are first applied to the T24 and RT112 datasets recorded on glass slides.

The first algorithm is the modified polynomial baseline correction method, as proposed by Lieber et al. [21]. This is an automated approach in which the spectrum is first fitted with an *N* order polynomial using ordinary least squares. Those values of the spectrum that lie above the polynomial are set equal to the value of the polynomial; the resultant signal is again fitted with an *N* order polynomial, and the process is iteratively repeated, until a point is reached such that the polynomial lies directly underneath the Raman spectral peaks. Here, this method is applied based on a fifth order polynomial and 200 iterations. This algorithm makes no attempt to remove the glass signal, but it is one of the most common approaches used for baseline correction in the literature.

The second algorithm applied is the method proposed by Beier et al. [20], which involves the subtraction of a weighted glass signal and an *N* order polynomial in an iterative manner, similar to the modified polynomial method, until a point is reached such that the modelled baseline lies directly below the Raman peaks. The *fminsearch* function in Matlab (Matlab, Mathworks Inc., USA) can be applied to determine which glass weight/polynomial combination results in the smallest residual spectrum, as described in [20]. In the results presented here, *N* is chosen to be 5, and the number of iterations is 200.

The EMSC algorithm is implemented via an adapted version of Matlab's *polyfit* function, which is available in the Appendix. A glass signal was recorded by focusing the laser onto the surface of a glass slide over a 5 s integration time. The glass signal was then smoothed using a Savitsky Golay filter ($w = 3,\ k = 41$). In this study, only a first order polynomial is chosen for all cases (i.e. a straight line), however, in general the order of the polynomial is dependent upon the dataset being analysed, and the associated baseline intensity present.

The choice of reference spectrum is an important consideration as discussed in Section 3. EMSC has previously been shown to work well for Raman spectroscopy; in previous examples the mean spectrum of a dataset has been shown to be a good choice for the reference spectrum. However, in the present study, two datasets are analysed for the purpose of cell classification. It is essential to use the same reference for both bladder cell datasets on glass slides. The reason for this lies in the constant baseline that is inherent in the reference. Therefore, all spectra must be fit to the same reference for the sake of fair comparison. In order to understand if the choice of a particular reference spectrum can introduce any bias into subsequent multivariate analysis, two different reference spectra were applied, and PCA analysis of the resultant dataset pairs was compared. In order to generate a reference spectrum, 50 cell spectra are recorded from cells on CaF_2, and averaged together to form a single spectrum. No pre-processing or baseline correction algorithms are applied to the reference spectrum, unless desired for aesthetical purposes. Here, the first reference spectrum is generated from T24 cells that are recorded on a CaF_2 substrate, and the second reference spectrum is taken from MDA-MB-231 cells that are also recorded on a CaF_2 substrate. These two references have been chosen because they are both from epithelial cells, but have different spectral shapes and intensities to each other. Furthermore, one of the reference spectra is related to one of the two bladder cell lines under investigation, while the other is unrelated to the two cell lines under investigation. In both cases, a Savitsky Golay smoothing filter ($w = 3,\ k = 7$) is applied to the mean CaF_2 spectra, to generate the final reference spectra for use in the EMSC algorithm.

PCA is applied to both datasets following each background subtraction method, and the first three PC coefficients are analysed for residual glass signals. Additionally, the standard deviation across an entire dataset is monitored and compared for each background algorithm.

4. Results

We begin this section with the results of the two well-known background subtraction algorithms discussed in Section 2; the modified polynomial [21], and the Beier method consisting of a modified polynomial with the glass signal [20]. Fig. 1(a) shows the mean spectra for T24 and RT112 cells recorded on glass, averaged over 50 cell spectra in each group respectively, following (i) the application of the modified polynomial baseline correction method, and (ii) normalisation. The shaded region highlights the presence of varying levels of glass signal across both datasets. This varying signal is due to the different cell morphology of the two cell lines; the RT112 cells appear to contain relatively smaller nuclei sizes, and therefore their Raman spectra contain a larger proportion of the glass substrate signal. It is expected that the glass contribution will remain in the processed spectra since the modified polynomial method is designed only to subtract the baseline signal.

Fig. 1(b) shows the PC score plot, and the first three PC coefficients obtained when the data in Fig. 1(a) are subject to PCA. Here, it can be seen that the first PC has a significant glass contribution (see Fig. 1(c)). It is interesting to note that the spectra are separating mainly along the first PC; the physical interpretation of this is that the cells are effectively separation according to differing morphology across the two cell lines, which is manifesting through the varying power of the glass signal component.

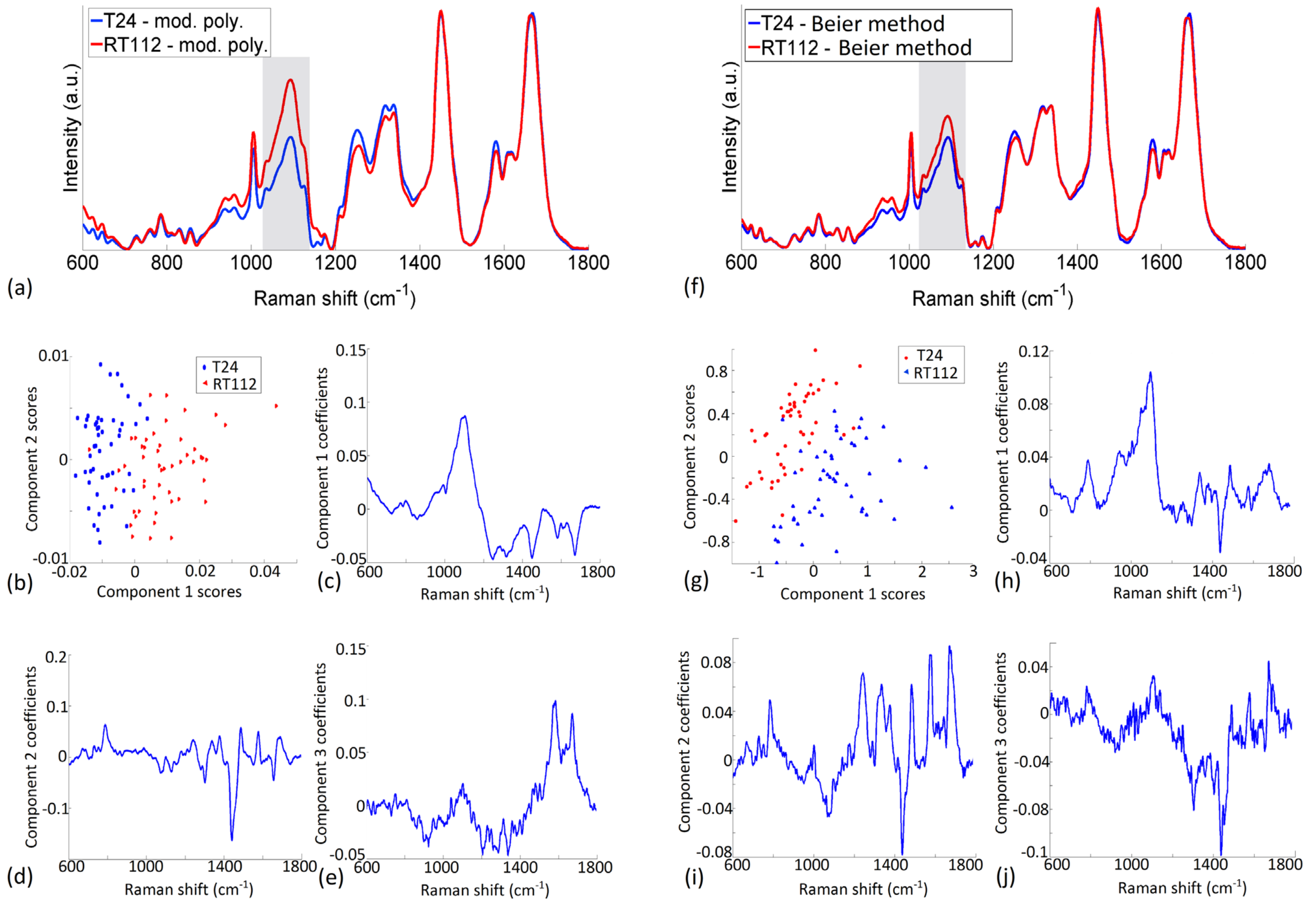

Fig. 1. (a) Mean of T24 and RT112 datasets following a modified polynomial background subtraction method; (b) PC scores, and the first three PC coefficients [(c), (d), and (e), respectively] for the data shown in (a); (f) mean of T24 and RT112 datasets following the background subtraction method proposed by Beier et al., involving a modified polynomial and glass signal subtraction [20]; (g) PC scores, and the first three PC coefficients [(h), (i), and (j), respectively] for the data shown in (f). The shaded areas highlight the region where the glass signal is most present within Raman cell spectra. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

Although interesting, this is not a desirable result. It cannot be expected that a similar morphological difference will occur in all cell classification applications. It is far more reliable in general to base classification on the biochemical variation across the datasets.

The second algorithm that is investigated is the method proposed by Beier et al. [20] which, as described in Section 3, is designed to simultaneously remove both the glass signal and the baseline based on the combination of a modified polynomial and a weighted glass signal. Fig. 1(f) displays the mean spectra for T24 and RT112 cells recorded on glass, averaged over 50 cell spectra in each group respectively, following application of this algorithm and normalisation. This figure demonstrates a significant reduction in the amount of glass signal present in the spectra, particularly for RT112 cells, when compared to spectra that have been baseline corrected with a modified polynomial alone.

Fig. 1(g) shows the PC score plot, and the first three PC coefficients obtained when the processed data is subject to PCA. Similar to the results presented in Fig. 1(c), the first PC contains a signal within the 1050–1150 cm^{-1} region, associated with glass (see Fig. 1(h)). Therefore, while the glass signal has been reduced, it is still a significant component in both spectral datasets, and will remain the dominant factor in any PCA based classification.

Fig. 2 illustrates the removal of the glass signal contribution from a sample T24 cell spectrum recorded on a glass slide based on a reference spectrum generated from T24 cells on CaF_2 using EMSC. The red line represents the raw spectrum recorded from a T24 cell on glass, and the black line represents the reference spectrum to which all other spectra are fitted. The green line is the modelled background consisting of a combination of the glass signal and a first order polynomial, determined using the EMSC algorithm, and the blue line is the corrected spectrum, which has had the glass signal subtracted.

Fig. 3(a) shows the equivalent mean spectra for the same cell lines as shown in Fig. 1, where pre-processing is implemented using the EMSC algorithm; input to this algorithm was the glass signal, a reference spectrum based on T24 cells on a CaF_2 substrate, as well as the chosen polynomial order N. Here, the glass signal has been effectively removed from the Raman cell spectra. The remaining peaks within the 1050–1150 cm^{-1} region are Raman cell peaks, seen in urothelial cell spectra recorded on CaF_2 and other similar substrates that produce low background signals [31]. The standard deviation is also shown for each cell line following EMSC processing. The amplitude of the standard deviation is amplified by a factor of 10 with respect to the mean spectra shown in the same figure.

Fig. 3(b) displays the PC score plot obtained when these datasets are subject to PCA, with Fig. 3(c), (d), and (e) illustrating the first three PC coefficients obtained. In contrast to the PC coefficients using the previously discussed processing methods, the PC coefficients presented in Fig. 3 do not appear to contain a glass signal component. This is an important result as it is essential for

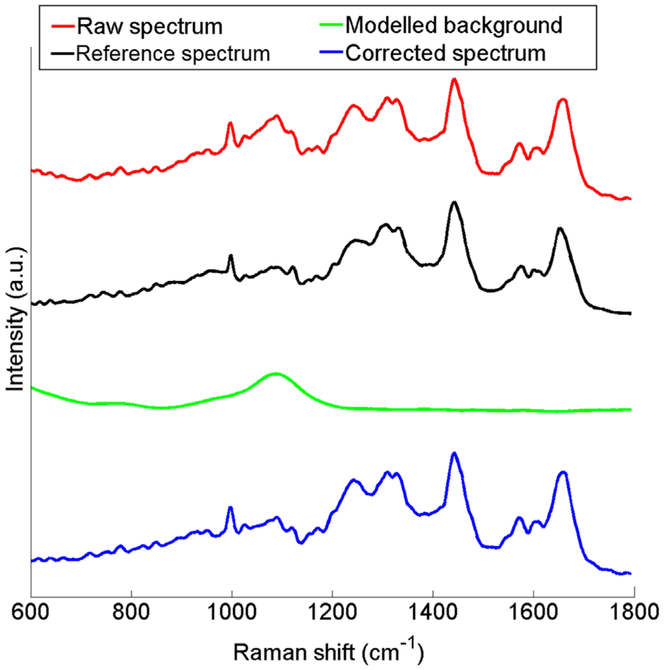

Fig. 2. Raw spectrum of T24 cell (red), recorded on a glass slide, which has been fit to a reference spectrum recorded from T24 cells on CaF_2 (black) using EMSC, resulting in a modelled background signal (green) consisting of a first order polynomial and a weighted glass signal, and the final background corrected spectrum (blue). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

reliable classification algorithms to be based on biomolecular differences across cell groups, and not to be based on the presence, or absence, of a glass signal. Both cell lines separate according to the second PC coefficient, with key biomolecular peak differences observed at 790 (DNA), 938 (proteins), 1170 (tyrosine), 1221 (Amide III), 1340 (CH_2/CH_3 wagging of nucleic acids), 1580 (proteins), and 1676 cm^{-1} (Amide I); similar peak differences have previously been observed in the separation of urothelial cell lines [5,32].

It is important to note that $N = 1$ was used for the EMSC algorithm; in this case, therefore, the EMSC algorithm ultimately amounts to subtracting only a straight line, and a weighted glass signal from each cell spectrum, followed by normalisation. The use of higher orders were also investigated, but the use of $N = 3, 5,$ or 7 appeared to offer no improvement over the results presented here. For this reason, and since it may help to invalidate any suggestion of over-fitting, $N = 1$ is chosen. For some cases, which are not presented here, where strong baseline signals are present within a dataset, higher values of N are needed for accurate modelling of the background signal. It has been shown elsewhere that the use of high values of N (e.g. up to 7th order) with EMSC does not result in over-fitting. [22].

Although these results are positive, and it is clear that the glass signal is removed, and will no longer be a factor in any subsequent classification applied to the PCA results, it could be argued that the EMSC algorithm might inadvertently influence the results of any subsequent multivariate analysis, particularly when the reference is based on the mean spectrum of T24 cells on CaF_2 for both the T24 and RT112 glass datasets. To investigate such effects, a second reference spectrum is used that is unrelated to the two cell lines under investigation. The reference used in this case is based on MDA-MB-231 cells on CaF_2. All other parameters (i.e. glass signal, N) remained the same. The mean spectra of the two processed datasets are shown in Fig. 3(f), where it can be seen that the overall shape of these mean spectra are moderately different from the corresponding result shown in Fig. 3(a) for the T24 reference spectrum. This results from the two references containing differing baselines. This difference is merely a question of aesthetics, and has no impact on any multivariate statistical analysis that is to follow EMSC processing. Indeed, it can be seen that the peak differences between the two spectra are effectively the same for both references. An analysis of the standard deviations of the two datasets also shows a very similar trend to that found for the previous T24 reference.

Fig. 3(g) shows the PC score plot obtained when the two processed datasets are subject to PCA. Remarkably, the PCA score plot appears to be identical to that obtained for the previous reference. Fig. 3(h), (i), and (j) illustrate the first three PC coefficients obtained. The PC coefficients appear to be identical to those presented in Fig. 3(c), (d), and (e), with both cell lines separating across the same regions, thus showing that the EMSC algorithm appears to produce results in the subsequent multivariate analysis that are independent of the reference spectrum used, and that it has no impact on the relative Raman peaks.

The ratio of c_g to c_r corresponds to the weight of the glass signal relative to the weight of the reference that is present in a given spectrum. These ratios are shown in Fig. 4 for both the T24 and RT112 datasets following application of the EMSC algorithm with the T24 reference spectrum. By applying a Gaussian distribution fit to these values, it is possible to estimate both the mean and standard deviation of the glass to reference ratio. It is clear that in general, the RT112 cells contain a stronger glass signal and a larger standard deviation, likely resulting from their smaller morphology. It is possible to choose a threshold ratio value, in between the two mean values, that largely separates the two datasets. By choosing the mid-point as a simple threshold, it is possible to achieve a sensitivity of 97% and specificity of 82% for T24 and RT112 cells. We believe that this classification is based purely on the morphology of the cells being analysed, with the ratio of c_g to c_r being inversely proportional to the cell thickness. It is interesting to note that in the case of the two algorithms investigated in Fig. 1, classification was possible using only the first PC, which was primarily composed of the glass signal, rather than the subtle variation in the Raman spectra owing to varying biochemical concentrations. In the case of EMSC, the glass signal can be extracted, and two approaches exist for classification: (i) analysing the ratio of c_g to c_r to perform an approximate classification, or (ii) analysing the Raman spectra after EMSC for a more accurate classification based on biomolecular variation.

In order to compare the three algorithms investigated here, the standard deviation and confidence intervals, across the various processed datasets, were analysed. Fig. 5 shows the mean, and associated 95% confidence interval, for RT112 cells following each background correction algorithms. Following a modified polynomial correction, there remains a significant deviation across the fingerprint region, with the largest differences seen in the region where the glass signal is present. The Beier method reduces this confidence interval, however, there remains a considerable amount of variance seen across the 1050–1150 cm^{-1} region. The third method, based on EMSC, shows a reduced confidence interval across the entire spectrum, including the 1050–1150 cm^{-1} region. This indicates that the glass signal has been effectively removed from all of the spectra in the dataset.

5. Conclusion

For many years, the advancement of Raman micro-spectroscopy into the clinic has been impeded by its incompatibility with standard clinical protocols, particularly the use of inexpensive glass slides. In this paper, the ability to remove the glass signal

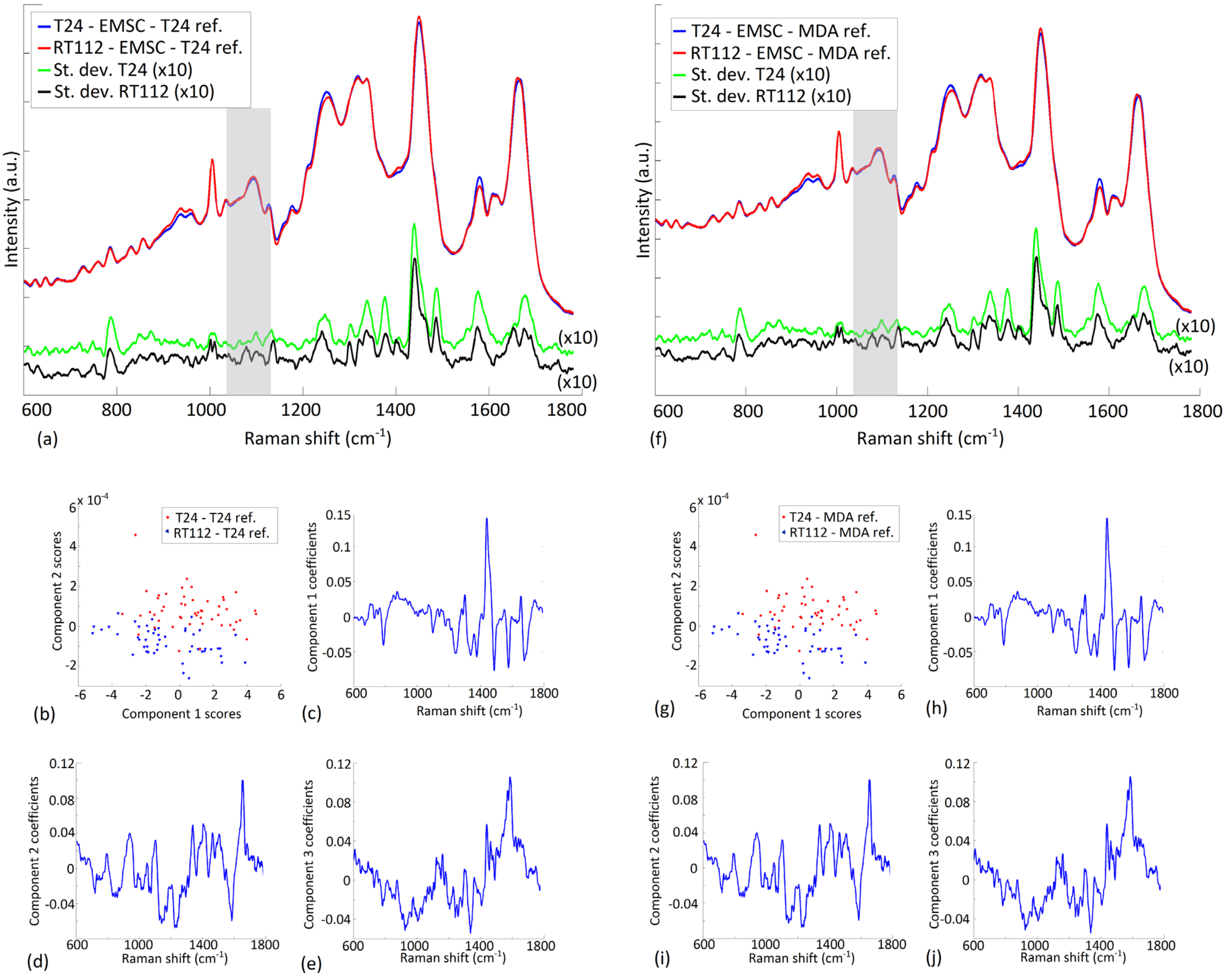

Fig. 3. (a) Mean of T24 and RT112 datasets following EMSC based on a T24 reference spectrum; (b) PC scores, and the first three PC coefficients [(c), (d), and (e), respectively] for the data shown in (a); (f) mean of T24 and RT112 datasets following EMSC based on a MDA-MB-231 reference spectrum; (g) PC scores, and the first three PC coefficients [(h), (i), and (j), respectively] for the data shown in (f). The shaded areas highlight the region where the glass signal is strongest within Raman cell spectra. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

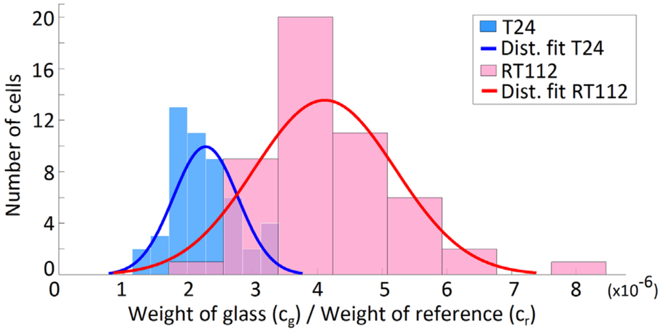

Fig. 4. Histograms, and associated distribution fits, of the weight of the glass signal (c_g) divided by the weight of the reference signal (c_r), obtained following EMSC with a T24 reference for T24 and RT112 cells.

present within Raman spectra, as well as the baseline signal, has been demonstrated, resulting in spectra that are free from glass contamination. The EMSC algorithm takes as input (i) the signal generated from glass slides upon illumination with a 532 nm laser, (ii) a reference spectrum based on a similar cell type recorded on

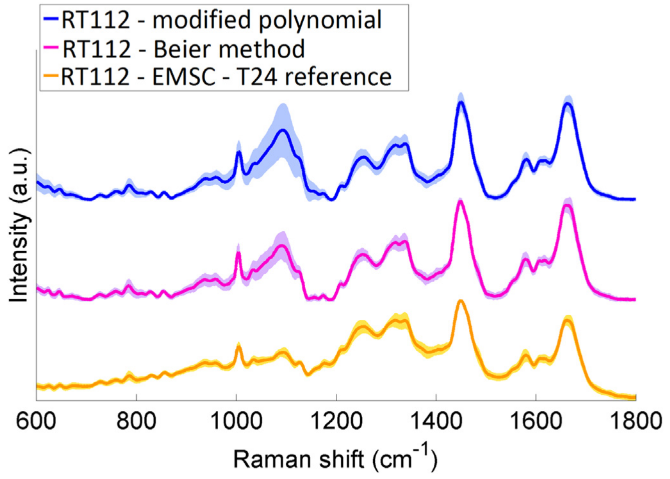

Fig. 5. Mean, and associated 95% confidence intervals, of RT112 cells, recorded on glass slides, following three different background algorithms: (i) modified polynomial, (ii) Beier method based on a modified polynomial plus glass signal, and (iii) EMSC algorithm, based on a reference spectrum of T24 cells on CaF_2.

CaF_2 substrates, and (iii) a chosen baseline polynomial order N; based on these input parameters the algorithm estimates a background consisting of a weighted glass signal and a slowly varying baseline curve, which can be subtracted to produce spectra that are equivalent to those recorded on expensive "Raman-friendly" substrates. The output of the EMSC algorithm is the raw spectrum with this background subtracted, followed by multiplication with a normalisation factor that is related to the reference spectrum.

Additionally, application of PCA to the background corrected spectra indicates that EMSC produces reliable and reproducible results that are independent of the reference spectrum used. This is an important result since objectivity and reproducibility are crucial for providing good diagnostic classification, and it demonstrates that the reference spectrum does not introduce any biasing of PCA based classification. EMSC was also compared with two other well-known background subtraction algorithms, for which it can be seen that the glass signal remained a significant component within their first PCs, thus reducing the reliability of these algorithms for Raman based cytological diagnostics on glass slides.

At present, it is not possible to provide a universal reference spectrum that can be applied to Raman datasets recorded on glass slides from any Raman micro-spectrometer. The reasoning for this is due to the lack of accurate system calibration protocols for Raman micro-spectrometers. Such rigorous calibration tools involve wavelength calibration, using a Neon source, intensity calibration, with a NIST calibrated white light source, followed by wavenumber calibration [33–35]. In order to utilise a universal Raman reference spectrum, similar to the Matrigel spectrum applied for FTIR RMie correction [25], every Raman micro-spectrometer would need to be calibrated with a similar calibration tool. Furthermore, even with the application of such calibration tools, variable results are still often recorded across different systems [36]. Therefore, it is advised for the reference spectrum to be recorded from any epithelial cell type on CaF_2 with the user's own Raman micro-spectrometer.

As demonstrated in Fig. 4, the ratio of the weight of the glass signal to the weight of the reference can be applied as a single classification metric for the case of the two cell lines investigated within this paper; using this approach, it is possible to separate low grade and high grade bladder cancer cell lines with 97% sensitivity and 82% specificity. This technique could be used to quickly identify large or abnormal cells on a slide. However, as these results are most likely based on cell thickness, it may not be possible to separate cell groups that are more similar in size. This is an interesting application for Raman micro-spectroscopy, but it should be noted that there are alternative modalities that can provide better information about cell morphology, such as digital holographic imaging [37], scanning near-field microscopy (SNOM) [38], or scanning electron microscopy (SEM) [39]. Raman micro-spectroscopy identifies the biomolecular differences between different cell groups, and by removing the glass signal from Raman spectra with EMSC, it is possible to classify cells based on biochemistry rather than cell morphology, which produces higher classification results and can be applied across all cell types. Therefore, Raman micro-spectroscopy remains the preferred modality for the identification of cancerous or diseased cells. However, it should be noted that it may be possible to include the c_g/c_r metric as an additional variable, together with the processed cell spectrum, for enhanced classification; the benefit, as well as the manner, of such an approach may be the subject of future work.

There are many further advantages of the EMSC algorithm when compared to commonly used baseline correction methods. It is computationally less intensive, which is an important factor when considering the high patient through-put present in cytopathological laboratories worldwide, and produces spectral datasets with significantly smaller standard deviation, which improves the reproducibility of results. We believe that this algorithm will help with the advancement of Raman based cytology into a clinical setting, allowing for the use of current clinical techniques, such as the ThinPrep or SurePath methods, and glass slides.

Acknowledgements

The authors would like to thank Dr. Shirley O'Dea and her lab at the Institute of Immunology, Maynooth University, and Dr. Amanda McCann and her lab in the Conway Institute, University College Dublin, for kindly preparing the cell samples for this study. This research was conducted with the financial support of Science Foundation Ireland (SFI) under Grant no. 11/SIRG/I2140.

References

[1] F. Bonnier, D. Traynor, P. Kearney, C. Clarke, P. Knief, C. Martin, J.J. O'Leary, H. J. Byrne, F. Lyng, Processing ThinPrep cervical cytological samples for Raman spectroscopic analysis, Anal. Methods 6 (2014) 7831–7841.

[2] I.R.M. Ramos, A. Malkin, F.M. Lyng, Current advances in the application of raman spectroscopy for molecular diagnosis of cervical cancer, BioMed Res. Int. 561242 (2015) 1–9.

[3] L.E. Kamemoto, A.K. Misra, S.K. Sharma, M.T. Goodman, H. Luk, A.C. Dykes, T. Acosta, Near-infrared micro-raman spectroscopy for in vitro detection of cervical cancer, Appl. Spectrosc. 64 (3) (2010) 255–261.

[4] E. Canetta, M. Mazilu, A.C. De Luca, A.E. Carruthers, K. Dholakia, S. Neilson, H. Sargeant, T. Briscoe, C.S. Herrington, A.C. Riches, Modulated Raman spectroscopy for enhanced identification of bladder tumour cells in urine samples, J. Biomed. Opt. 16 (3) (2011) 037002.

[5] T.J. Harvey, C. Hughes, A.D. Ward, E.C. Faria, A. Henderson, N.W. Clarke, M. D. Brown, R.D. Snook, P. Gardner, Classification of fixed urological cells using raman tweezers, J. Biophotonics 2 (1-2) (2009) 47–69.

[6] L.T. Kerr, K. Domijan, I. Cullen, B.M. Hennelly, Applications of raman spectroscopy to the urinary bladder for cancer diagnostics, Photonics Lasers Med. 3 (3) (2014) 193–224.

[7] A. Sahu, S. Tawde, V. Pai, P. Gera, P. Chaturvedi, S. Nair, C.M. Krishna, Raman spectroscopy and cytopathology of oral exfoliated cells for oral cancer diagnosis, Anal. Methods 7 (2015) 7548–7559.

[8] L.F.C.S. Carvalho, F. Bonnier, K. O'Callaghan, J. O'Sullivan, S. Flint, H.J. Byrne, F. M. Lyng, Raman micro-spectroscopy for rapid screening of oral squamous cell carcinoma, Exp. Mol. Pathol. 98 (2015) 502–509.

[9] A. Cao, A.K. Pandya, G.K. Serhatkulu, R.E. Weber, H. Dai, J.S. Thakur, V.M. Naik, R. Naik, G.W. Auner, R. Rabah, D.C. Freeman, A robust method for automated background subtraction of tissue fluorescence, J. Raman Spectrosc. 38 (2007) 1199–1205.

[10] H.J. Byrne, P. Knief, M.E. Keating, F. Bonnier, Spectral pre and post processing for infrared and raman spectroscopy of biological tissues and cells, Chem. Soc. Rev. 45 (7) (2016) 1865–1878.

[11] N.K. Afseth, V.H. Segtnan, J.P. Wold, Raman spectra of biological samples: a study of preprocessing methods, Appl. Spectrosc. 60 (12) (2006) 1358–1367.

[12] K.H. Liland, T. Almoy, B.H. Mevik, Optimal choice of baseline correction for multivariate calibration of spectra, Appl. Spectrosc. 64 (9) (2010) 1007–1016.

[13] P. Lasch, Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging, Chemometr. Intell. Lab. 117 (2012) 100–114.

[14] L. Mikoliunaite, R.D. Rodriguez, E. Sheremet, V. Kolchuzhin, J. Mehner, A. Ramanavicius, D.R.T. Zahn, The substrate matters in the Raman spectroscopy analysis of cells, Sci. Rep. 5 (2015) 13150.

[15] L.T. Kerr, H.J. Byrne, B.M. Hennelly, Optimal choice of sample substrate and laser wavelength for raman spectroscopic analysis of biological specimen, Anal. Methods 7 (2015) 5041–5052.

[16] L.M. Fullwood, D. Griffiths, K. Ashton, T. Dawson, R.W. Lea, C. Davis, F. Bonnier, H.J. Byrne, M.J. Baker, Effect of substrate choice and tissue type on tissue preparation for spectral histopathology by Raman microspectroscopy, Analyst 139 (2014) 446–454.

[17] A. Tfayli, C. Gobinet, V. Vrabie, R. Huez, M. Manfait, O. Piot, Digital dewaxing of Raman signals: discrimination between nevi and melanoma spectra obtained from paraffin-embedded skin biopsies, Appl. Spectrosc. 63 (5) (2009) 564–570.

[18] Z. Farhane, F. Bonnier, A. Casey, H.J. Byrne, Raman micro spectroscopy for in vitro drug screening: subcellular localisation and interactions of doxorubicin, Analyst 140 (2015) 4212–4223.

[19] E. Efeoglu, M. Keating, J. McIntyre, A. Casey, H.J. Byrne, Determination of nanoparticle localisation within subcellular organelles in vitro using raman spectroscopy, Anal. Methods 7 (2015) 10000–10017.

[20] B.D. Beier, A.J. Berger, Method for automated background subtraction from

raman spectra containing known contaminants, Analyst 134 (2009) 1198–1202.

[21] C.A. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological raman spectra, App. Spectrosc. 57 (11) (2003) 1363–1367.

[22] N.K. Afseth, A. Kohler, Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, Chemometr. Intell. Lab. 117 (2012) 92–99.

[23] H. Martens, E. Stark, Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy, J. Pharm. Biomed. Anal. 9 (8) (1991) 625–635.

[24] B. Bird, M. Miljkovic, M. Diem, Two step resonant mie scattering correction of infrared micro-spectral data: human lymph node tissue, J. Biophotonics 3 (8-9) (2010) 597–608.

[25] P. Bassan, H.J. Byrne, F. Bonnier, J. Lee, P. Dumas, P. Gardner, Resonant mie scattering in infrared spectroscopy of biological materials—understanding the dispersion artefact, Analyst 134 (2009) 1586–1593.

[26] P. Bassan, H.J. Byrne, J. Lee, F. Bonnier, C. Clarke, P. Dumas, E. Gazi, M.D. Brown, N.W. Clarke, P. Gardner, Reflection contributions to the dispersion artefact in FTIR spectra of single biological cells, Analyst 134 (2009) 1171–1175.

[27] K.H. Liland, A. Kohler, N.K. Afseth, Model-based pre-processing in Raman spectroscopy of biological samples, J. Raman Spectrosc. 47 (6) (2016) 643–650.

[28] J.M. Smulko, N.C. Dingari, J.S. Soares, I. Barman, Anatomy of noise in quantitative biological Raman spectroscopy, Bioanalysis 6 (3) (2014) 411–421.

[29] J.H. Matthews, K.D. Fink, Numerical Methods Using Matlab, third ed., Prentice-Hall, Upper Saddle River, New Jersey, USA, 1999.

[30] T.M. James, M. Schlosser, R.J. Lewis, S. Fischer, B. Bornschein, H.H. Telle, Automated quantitative spectroscopic analysis combining background subtraction, cosmic ray removal, and peak fitting, App. Spectrosc. 67 (8) (2013) 949–959.

[31] P. Crow, J.S. Uff, J.A. Farmer, M.P. Wright, N. Stone, The use of raman spectroscopy to identify and characterize transitional cell carcinoma in vitro, BJU Int. 93 (2004) 1232–1236.

[32] R.O.P. Draga, M.C.M. Grimbergen, P.L.M. Vijverberg, C.F.P. van Swol, T.G. N. Jonges, J.A. Kummer, J.L.H.R. Bosch, In vivo bladder cancer diagnosis by high-volume Raman spectroscopy, Anal. Chem. 82 (2010) 5993–5999.

[33] D. Hutsebaut, P. Vandenabeele, L. Moens, Evaluation of an accurate calibration and spectral standardization procedure for Raman spectroscopy, Analyst 130 (2005) 1204–1214.

[34] J.D. Rodriguez, B.J. Westenberger, L.F. Buhse, J.F. Kauffman, Standardization of raman spectra for transfer of spectral libraries across different instruments, Analyst 136 (2011) 4232–4240.

[35] T. Dorfer, T. Bocklitz, N. Tarcea, M. Schmitt, J. Popp, Checking and improving calibration of raman spectra using chemometric approaches, Z. Phys. Chem. 255 (2011) 753–764.

[36] M. Isabelle, J. Sorney, A. Lewis, G.R. Lloyd, O. Old, N. Shepherd, M. Rodriguez-Justo, H. Barr, K. Lau, I. Bell, S. Ohrel, G. Thomas, N. Stone, C. Kendall, Multi-centre Raman spectral mapping of oesophageal cancer tissues: a study to assess system transferability, Faraday Discuss. 187 (2016) 87–103.

[37] C.J. Mann, L. Yu, C.M. Lo, M.K. Kim, High-resolution quantitative phase-contrast microscopy by digital holography, Opt. Express 13 (22) (2005) 8693–8698.

[38] S. Rieti, V. Manni, A. Lisi, L. Giuliani, D. Sacco, E. D'Emilia, A. Cricenti, R. Generosi, M. Luce, S. Grimaldi, SNOM and AFM microscopy techniques to study the effect of non-ionizing radiation on the morphological and biochemical properties of human keratinocytes cell line (HaCaT), J. Microsc. 213 (1) (2004) 20–28.

[39] S. Passey, S. Pellegrin, H. Mellor, Scanning Electron Microscopy of Cell Surface Morphology, vol. 4, 2007.