Algorithm for optimal denoising of Raman spectra†

Sinead J. Barton[a], Tomas E. Ward[c], and Bryan M. Hennelly*[a,b]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

Raman spectroscopy has been demonstrated to have diagnostic potential in areas such as urine and cervical cytology, whereby different disease groups can be classified based on subtle differences in the cell or tissue spectra using various multi-variate statistical classification tools. However, Raman scattering is an inherently weak process, which often results in low signal to noise ratios, thus limiting the method's diagnostic capabilities under certain conditions. A common approach for reducing the experimental noise is Savitzky-Golay smoothing. While this method is effective in reducing the noise signal, it has the undesirable effect of smoothing the underlying Raman features, compromising their discriminative utility. Maximum Likelihood Estimation is a method for estimating the parameters of a statistical model given an available dataset and *a priori* knowledge of the model type. In this paper, we demonstrate how Savitzky-Golay smoothing may be enhanced with Maximum Likelihood Estimation in order to prevent significant deviation from the 'true' Raman signal yet retain the robust smoothing properties of the Savitzky-Golay filter. The algorithm presented here is demonstrated to have a lower impact on Raman spectral features at known spectral peaks while providing superior denoising capabilities, when compared with established smoothing algorithms; artificially noised databases and experimental data are used to evaluate and compare the performance of the algorithms in terms of the signal to noise ratio. The proposed method is demonstrated to typically provide at least a 50% increase in the signal to noise ratio when compared to the raw data, and consistently out-performs two alternative smoothing filters.

1 Introduction

Raman spectroscopy is a laser based technique that enables the identification and quantification of chemical bonds based on the inelastic scattering of monochromatic light. It is an inherently weak signal with approximately only 1 in 10^7 incident photons being Raman scattered.[1] This low photon count coupled with non-ideal collection efficiencies, e.g. the numerical aperture of the microscope objective or camera quantum efficiency, means that Raman spectroscopy is vulnerable to noise. Noise will decide the detection limit of the recording process as well as the classification potential of multivariate statistical analysis that may be applied to a recorded dataset for classification purposes, making efficient and reliable noise removal a necessity in sensitive applications such as chemical classification or diagnostics.[2]

Noise originates from two primary sources; the camera (dark current and read noise) as well as from the signal itself (shot noise).[3] The effect of these noise sources can be reduced by cooling the camera, using slow read out rates, increasing acquisition times, using a higher power laser, or using a laser wavelength that produces a larger number of photons e.g. 400*nm*. However, these measures can be costly in terms of time and equipment and thus, are impractical in certain applications. Therefore, an efficient denoising algorithm for post-processing of spectra would be advantageous.

Savitzky-Golay (SG) filtering[4] is commonly used to smooth spectra in order to reduce the impact of noise on statistical classification.[5,6] This filtering technique works by dynamically fitting a polynomial to consecutive windows of data points (local least-squares polynomial approximation) in order to follow the shape of the spectrum thereby mitigating the impact of a randomly varying noise signal. Under certain conditions, this can have a negative impact on spectral features; in particular high noise applications that require high levels of smoothing, which may severely affect sharp local features. Maximum Likelihood Estimation (MLE) is a statistical process that enables signal denoising[7] by searching for the most likely value of the signal based on a sequence of measured values and *a priori* knowledge of the noise distribution associated with the collected signal. The proposed algorithm merges the robust smoothing of the SG filter with the restriction that the denoised data must be constrained the noise distribution provided by MLE. In this paper, we demonstrate that this algorithm consistently returns a spectrum with a higher Signal to Noise Ratio (SNR) than SG filtering alone, as well as effec-

† Electronic Supplementary Information (ESI) available: MATLAB code and additional early stopping optimisation figures are given. See DOI: 10.1039/b000000x/

[a] *Electronic Engineering Department, Maynooth University, Co. Kildare, IE.* [b] *Callan Institute, Maynooth University Co. Kildare, IE.* [c] *Faculty of Engineering and Computing, Dublin City University, Glasnevin, Dublin 9, IE. E-mail: bryanh@cs.nuim.ie*

tively preserving the fidelity of sharp peaks.

The proposed denoising algorithm is constructed on the combination of Maximum Likelihood Estimation (MLE), based on estimating the noise in a spectrum, with Savitzky-Golay (SG) smoothing. The algorithm attempts to overcome the classical problem of providing a smooth noise free spectrum, without affecting the fidelity of sharp spectral features in the process. The algorithm is comprised of two competing constraints that are applied to a given spectrum: the first condition, which makes use of SG filtering, assumes that the spectrum is smooth, i.e. that a given sample will not differ significantly from its neighbours; and the second condition, which is based on MLE, requires that the sample does not deviate significantly from the raw value that was recorded, taking into account the noise distribution that exists for that raw sample value.

The paper is split into four main sections. Section 2 provides an overview of the properties of the noise sources, how they are modelled within the context of Maximum Likelihood Estimation, and the integration of Savitzky-Golay smoothing with MLE. Section 3 defines the metrics that are used to evaluate the performance of the algorithm, the creation of artificial data on which to evaluate and compare the data, and the steps taken to optimise the algorithm's input parameters. Section 4 provides the result and illustrates the SNR improvement provided by the proposed algorithm over competing denoising algorithms for experimental and simulated data. Finally, in Section 5 we offer a brief conclusion and propose a number of possible avenues for further improvements.

2 Theory

In this section, a brief overview is provided of the underlying theory for modelling the noise distributions, the basic MLE algorithm, and the utilisation of SG smoothing together with MLE such that the smoothed version of the spectrum can act as its own prior, i.e. it can provide a form of *a priori* information that can be employed in the MLE algorithm.

2.1 Noise

Noise in Raman data originates from four main sources: shot noise, dark current noise, read noise, and cosmic ray artefacts.[3] The latter are random energy pulses that interact with the camera and usually present as intense narrow spikes superimposed on spectra; however, the width and intensity of these spikes may vary significantly. The effect of the proposed algorithm on cosmic rays is not investigated in this paper, and it is assumed that the spectrum under investigation has been preprocessed for cosmic ray removal in advance of denoising.[8,9] Shot noise is inherent in all recorded electromagnetic signals and is the result of inconsistent irradiance on a pixel over consecutive fixed periods of time. Shot noise is governed by a Poisson distribution and is time dependent. Due to the square root relationship of the standard deviation of shot noise with respect to the collected irradiance, shot noise is more apparent at low signal levels and hence at shorter acquisition times. In general, Raman scattering is a weak process, which is particularly true for biological samples, and therefore shot noise is usually a problem in the field of Raman based biophotonics. The weak Raman scattering often necessitates long acquisition times in order to reduce the effect of shot noise. However, this comes at the expense of increased levels of dark current noise, which results from thermal effects in the camera and, like shot noise, it can be modelled by a time dependent Poisson distribution. High levels of dark current are a result of insufficient camera cooling or long acquisition times, and under certain conditions can have a significant impact on spectral features. Due to the reproductive properties of independent Poisson distributions, dark current noise and shot noise can be modelled by a single Poisson distribution[10] with a mean value given by the sum of the spectral irradiance and the dark current, both multiplied by the acquisition time. In some cases the true limit of detection can be said to be read noise, which is a time independent Gaussian distributed noise[11] added by the camera's analogue to digital converter; however, when recording Raman signals from biological samples, shot noise and dark current are usually several orders of magnitude greater than read noise.

It is possible to approximate a Poisson distribution as a Gaussian distribution provided the mean photo-electron count registered in the camera pixel is high enough. Therefore, if the spectral irradiance is sufficiently high, the total noise can be estimated with a single additive Gaussian distribution, and this enables the denoising process to be modelled as a decomposition problem, $y = x + d$, where y is a vector of discrete samples that is the recorded spectrum, x is the true spectral intensity in units of photons collected in each pixel area over the full acquisition time, t, and d is the noise signal, which is defined in terms of the following Gaussian probably distribution:

$$p(d_i) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left[\frac{-(d_i-\mu_i)^2}{2\sigma_i^2}\right] \quad (1)$$

where:

- i is an integer index that denotes the i^{th} discrete sample in a spectrum
- $\mu_i = r + tc_i$ i.e. the mean value of the distribution in the i^{th} sample (in electrons per second) is given by the sum of the mean read noise, r (in electrons), and the product of mean dark current, c_i, and time, t.

- $\sigma_i^2 = x_i + tc_i + \sigma_r^2$ i.e. the variance of the noise distribution is given by the sum of the variances of the individual noise terms. [12]

- The spectral intensity can be defined in terms of the spectral irradiance as follows: $x_i = tl_i$, where l_i denotes the irradiance in photons per pixel per second.

It is notable in the above description of the noise term, d, that the dark current noise can vary from sample to sample, which is due to the variable properties of the semiconductor pixels of modern CCD detectors, while the read noise is assumed to have a constant mean value and standard deviation across all pixels in the detector.

2.2 Maximum likelihood estimation

Maximum Likelihood Estimation (MLE) is a statistical method whereby the parameters of a known statistical model can be estimated based on a number of observations; this is achieved by calculating the parameter values for that model, which maximise the likelihood of making the set of observations. In order to use MLE to reduce the noise in a signal, the statistical model for the noise must be known; take for example the above decomposition problem for the noise in a single sample of the spectrum, i.e. the i^{th} sample. Since the values for the dark current and read noise parameters within the Gaussian are known (these can be measured in advance of recording a spectrum), then the only unknown is x_i. If a number of, k, different spectra are recorded, y_{i1}, y_{i2}, ... y_k, then MLE can be applied to determine the most likely value of x_i that would have resulted in this set of observations. However, this approach requires a number of different recordings and the outcome of MLE would be the trivial result that the most likely value is the mean of all the observations minus the mean noise. Here, we set ourselves the problem of applying MLE based only on a single observation. A similar approach has recently been proposed for removing noise from astronomical images, [13] which is of particular relevance to the current discussion due to the similarity between astronomical images and Raman spectra, i.e. areas of dark (flat regions) interspersed with stars (peaks).

We begin the derivation of the algorithm by formally defining the probability of recording an intensity value in the i^{th} sample, y_i, given the true intensity, x_i, as follows:

$$p(y_i;x_i) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left[\frac{-(y_i-\mu_i-x_i)^2}{2\sigma_i^2}\right] \quad (2)$$

The mean noise, μ_i, may be subtracted from y_i by recording a dark frame of sufficiently long acquisition time. The standard deviation of the noise, σ_i, varies across the samples due to the varying dark current contributions, c_i, which are often pixel dependent, and the dependence of shot noise on the varying signal intensity, x_i. However, for simplicity and ease of computation, the algorithm assumes a constant standard deviation, denoted $\bar{\sigma}$, for all samples and is calculated as follows:

$$z = y - \mu - SG(y-\mu, v, q) \quad (3a)$$

$$\bar{\sigma} = \frac{1}{N}\sum_{i=1}^{N} z_i^2 \quad (3b)$$

where N is the total number of samples in the spectrum. The value of $\bar{\sigma}$ that is used in the algorithm is calculated by estimating the mean standard deviation of the global noise term. This is achieved by applying an appropriate SG smoothing filter ($v = 3$, $q = 9$, where v represents the polynomial order and q represents the window size of the filter) to the spectrum, subtracting the smooth from the raw, and finally taking the standard deviation of the remaining signal. Following from this, the negative log likelihood of observing a signal intensity at sample i is:

$$-log(p(y_i;x_i)) = \frac{(y_i-\mu_i-x_i)^2}{2\bar{\sigma}^2} \quad (4)$$

2.3 Maximising an 'a posteriori' estimator

Denoising in this context requires the use of an image prior, x', i.e. a reference signal to allow the user to deduce *a priori* knowledge of a given spectral sample on the basis that the spectrum should not deviate significantly from the image prior. The probability of the true intensity at sample i, can be defined in terms of the intensity values of the samples in the image prior in the neighbourhood around i as follows:

$$p(x_i) = \prod_{j=i-n}^{i+n} exp[-\lambda|x_i - x'_j|^p] \quad (5)$$

where $2n+1$ is the size of the neighbourhood and the λ and p parameters are used to define how closely a sample in x is expected to match the surrounding samples in the image prior; selecting $p > 0$ will impose a constraint that a smooth transition must exist from one sample to the next. In the limiting case, if only one observation is available, we can set $x' = y - \mu$. This approach is used in Gomez-Rodriguez *et al.* [14] whereby the image itself serves as it's own image prior, eliminating the need for multiple acquisitions. The basis of the first MLE model described here makes use of a similar approach, whereby the neighbouring samples in the signal can provide a reference for that sample. In this case the values of λ and p determine how smooth the transition should be from one sample to the next. [13–15] Following from this discussion, the negative log likelihood of Eq.5 can be determined:

$$-log(p(x_i)) = \lambda \sum_{j=i-n}^{i+n} |x_i - x'_j|^p \quad (6)$$

Using Bayes' theorem, the negative log likelihood of $p(x_i;y_i)$ can be expressed as follows:

$$-log(p(x_i;y_i)) = -log(p(y_i;x_i)) - log(p(x_i)) \quad (7)$$

Explicitly:

$$MLE(x_i) = \frac{(y_i - \mu_i - x_i)^2}{2\bar{\sigma}^2} + \lambda \sum_{j=i-n}^{i+n} |x_i - x'_j|^p \quad (8)$$

Therefore, the most likely value of x_i is the one that minimises Eq.8 and this equation is the basis of the first MLE based algorithm that we propose here. The algorithm begins by setting $x' = y - \mu$; this involves subtracting a dark frame from the raw spectrum. The second step is to calculate the most likely estimate of x, which we denote as x^e; this is achieved by performing a brute force search to find the sample values that minimise Eq.8, which we denote as x_i^e. This process is repeated for each sample, i, until the entire spectrum is estimated. The third step is to set $x' = x^e$, and then to repeat the second and third step iteratively until the conditions for stopping are met; an early stopping strategy is important in order to prevent over-smoothing of key signal features. Although the denoising algorithm described by Eq.8 provides meaningful results, we do not investigate it any further in this paper. A superior algorithm is proposed in the section that follows, which employs a similar approach; the detailed development of the first algorithm above is a necessary first step before introducing the algorithm below.

2.4 Improving the 'a posteriori' estimator by employing SG smoothing

The MLE algorithm defined in the previous section is similar to that defined in Burger *et al.* [13], which was applied to astronomical images and employs a two dimensional neighbourhood of nine pixels around the sample of interest. A Raman spectrum is inherently one dimensional and, therefore, only the samples immediately to the left and right of the sample i can be used in the MLE algorithm. In this section, we propose an improved MLE algorithm that makes use of Savitzky-Golay (SG) filtering. The algorithm is similar to that described in the previous section; however, in this case the first step is to set $x' = SG(y - \mu, v, q)$, where SG denotes the application of an SG filter to the raw spectrum with a dark frame subtracted, $y - \mu$. The second step involves finding the values of x_i that minimise Eq.8 as described for the previous algorithm, which results in the estimate x^e. The third step involves setting $x' = SG(x^e, v, q)$; the second and third steps are repeated iteratively and once again an early stopping strategy is employed to avoid over smoothing. The algorithm investigated in this paper uses a neighbourhood of only 1, i.e. $n = 0$. Therefore, Eq.8 reduces to:

$$MLE(x_i) = \frac{(y_i - \mu_i - x_i)^2}{2\bar{\sigma}^2} + \lambda |x_i - x'_i|^p \quad (9)$$

The algorithm described above, which will be referred to by the acronym MLESG going forward, is essentially a pixel by pixel estimator, that is constrained in two opposing directions. The left term in Eq.9 will increase as x_i deviates from the raw value. Conversely, the right term will increase as the estimated value deviates from the SG smoothed version of the spectrum. It can be expected that the algorithm will perform at least as well as traditional SG filtering, and with the additional constraint that the smoothed spectrum is not permitted to deviate far from the recorded value within the bounds of the noise distribution. We can therefore expect superior results in terms of recovering a truer estimate of the underlying Raman spectrum.

3 Tuning the algorithm

The algorithm outlined in the previous section requires five input variables, namely; λ, p, v, q, and the number of iterations, m. An investigation into the optimal values for these parameters was performed in order to minimise the number of input variables and maximise the denoising capability of the algorithm. The results of this investigation are detailed in this section in terms of Signal to Noise Ratio (SNR) and a metric that is proposed for the first time here, which we refer to as the SNR product.

3.1 Noise metrics for optimisation of parameters

SNR is an important metric for establishing signal quality in all fields of engineering, and is employed here to evaluate the performance of the proposed algorithm; to the best of our knowledge SNR has never previously been applied to evaluate smoothing algorithms in the field of Raman spectroscopy. SNR is commonly defined as the ratio of the true signal intensity to the standard deviation of the noise, [12] i.e. the signal to noise ratio at a single discrete sample in the spectrum is given by $SNR(x_i) = x_i/\sigma_i$, where σ is as previously defined in Eq.1. However, x_i, and by consequence σ_i, cannot be determined, and therefore, a method is required to approximate this definition. A definition of SNR that has previously been applied to Raman spectra, [16] is the ratio of the maximum value in the spectrum to the Root Mean Square Error (RMSE) of a flat region of the spectrum; the noise can be estimated by calculating the Root Mean Square Error (RMSE) of the noisy spectrum with respect to an accurately recorded reference spectrum, which is known to have very low noise. We employ a similar definition to estimate the SNR of a denoised spectrum,

x^e, which is defined as follows:

$$SNR(x^e) = \frac{max(x^e)}{RMSE(x^e, x^{ref})} \tag{10a}$$

$$RMSE(x^e, x^{ref}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i^e - x_i^{ref})^2} \tag{10b}$$

where $max()$ is a function that returns the maximum value in the input vector and *RMSE* calculates the root mean square error of the input vector with respect to the reference signal intensity, x^{ref}, both of length N. However, while smoothing may increase the SNR of a spectrum as a whole it can also negatively affect sharp local features, which may be of importance. In order to monitor the effect of the algorithm, specifically on sharp spectral features, the SNR in the neighbourhood of a peak, $x^e[pk-n:pk+n]$, is calculated. This calculation is based on an $2n+1$ sample subset of x^e centred on a feature located at index pk. The SNR for this peak region is defined as follows:

$$SNR(x^e[pk-n:pk+n]) = \frac{max(x^e)}{RMSE(x^e[pk-n:p+n], x^{ref}[pk-n:pk+n])} \tag{11}$$

This definition uses the same maximum value as for the global spectrum definition given in Eq.10 but *RMSE* is calculated only over the peak region. This ensures a meaningful comparison with SNR values of the global spectrum. An estimate for the SNR of the raw spectrum can be determined by calculating $SNR(y-\mu)$ for the global case and for the peak area $SNR(y[pk-n:pk+n]-\mu[pk-n:pk+n])$, using Eq.10 and Eq.11 respectively. Finally, in order to reflect the overall SNR improvement that is provided by the denoising algorithm, we propose a novel metric called the SNR product which takes into account the enhanced SNR globally as well as in the region of a sharp peak:

$$SNR_{prod} = \frac{SNR(x^e)}{SNR(y-\mu)} \times \frac{SNR(x^e[pk-n:pk+n])}{SNR(y[pk-n:pk+n]-\mu[pk-n:pk+n])} \tag{12}$$

Focusing on the left hand side of the above equation, this term concerns the global SNR and is used to evaluate the mean improvement in signal quality across the entire wavenumber range being examined. This term is primarily influenced by large low frequency regions. The right term focuses on a sharp local feature and is used to monitor whether the algorithm is negatively impacting peaks. If there is no SNR enhancement in the denoised spectrum compared to the raw, then the SNR product will return a value of 1 or lower. The typical range of results for the SNR product is $0 < SNR_{prod} < 4$.

3.2 Data driven parameter optimisation

In order to robustly examine the recovery potential of the algorithm, large datasets with varying SNR were required, as well as *a priori* knowledge of x^{ref}. This requirement meant that artificial datasets were best suited for the initial testing and optimisation phase since large amounts of data can be created with known noise parameters and with knowledge of the underlying signal. Datasets with various SNRs were generated based on a signal in the form of a high quality low noise Raman spectrum recorded from a polymer slide (Ibidi Gmbh) due to its resilience to photo-bleaching, thermal stability, intense and reproducible Raman spectrum. In total, 100 spectra were averaged together following subtraction of the mean dark current and mean read noise, which enabled an accurate estimate of the true irradiance in terms of the mean photons collected per pixel per second. This then enabled the signal intensity, x^{ref} (calculated by scaling the irradiance), and the noise to be simulated based on any acquisition time using Eq.1. In this way, six datasets were generated with the SNR values of 20, 40, 60, 80, 100, 120; each dataset contained 100 spectra. Fig. 1 illustrates four sample spectra of SNR values 120, 80, 60, 20 that are approximately indicative of low, medium, high, and extreme noise cases when recording Raman spectra from biological samples.

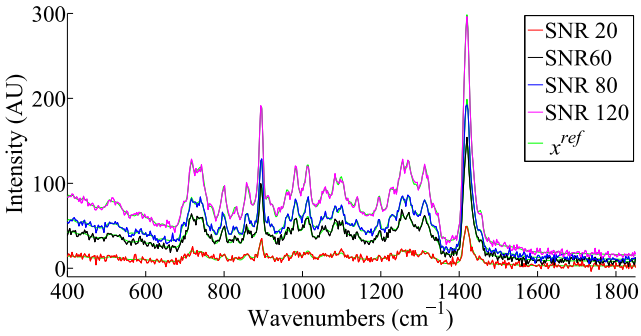

Fig. 1 Illustration of the artificial dataset noise levels and the corresponding x^{ref}

3.3 Optimal parameters

The MLESG algorithm described in Section 2 is dependent on five parameters; namely λ, p, SG parameters (v and q), and finally the number of iterations, m. This section describes the steps taken in order to find the best set of parameters to use for a noisy signal with a given SNR. Using all six artificial datasets described in Section 3.2 a brute force search over a wide range of λ and p was performed and the results were found to be approximately similar for all six datasets. The λ and p values were fixed at 1.8 and 0.4 respectively, which were found to work well for all cases, and the other param-

eters were varied in subsequent investigations. Initial testing revealed that SG input parameter combinations made up of v = 3, 5 and q = 5, 7 showed the most promise for use as spectral priors. All of the parameters were examined in terms of the improvement in the global SNR, peak SNR, and the SNR product of the denoised spectra over a range of iterations ($m = 1, 2, 3...100$). An example of this analysis for an initial SNR of 60 is illustrated in Fig. 2, in which the results that are shown are an average across a dataset of 100 spectra. Results beyond 50 iterations are not displayed in Fig. 2

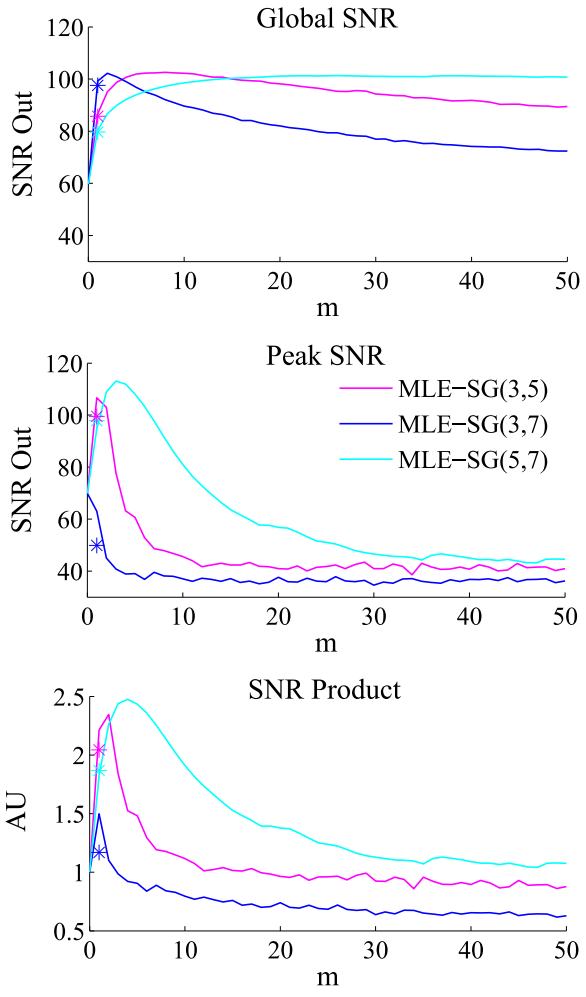

Fig. 2 An illustration of mean SNR recovery for the datasets with an initial global SNR of 60 for three sets of SG input parameters over 50 iterations (The asterisks aligned with the first iteration are representative of the SNR achieved by SG filtering alone)

since the SNR recovery has stabilised or is already in decline. These results were reproduced for all datasets previously mentioned. Two important results become clear, (i) noisier signals require a greater number of iterations in order to achieve optimal denoising for both peak regions and globally; this is discussed further below, and (ii) in general optimal improvement in peak SNR occurs much earlier than for the global spectrum, in terms of the number of iterations. It was determined that all other parameters, other than m, can be fixed to constant values regardless of input SNR, with approximately similar results. This significantly simplifies tuning of the algorithm for a given input SNR to selecting the most appropriate value of m. From examination of the figures it was determined that the most reliable SG input parameters for preserving peaks was SG(5,7); however, in terms of global SNR, SG(3,5) produces a slightly higher result. The aesthetic difference of these two conflicting requirements, i.e. smoothing vs. peak preservation, is illustrated in Fig. 3. Optimal numbers of iterations for global and

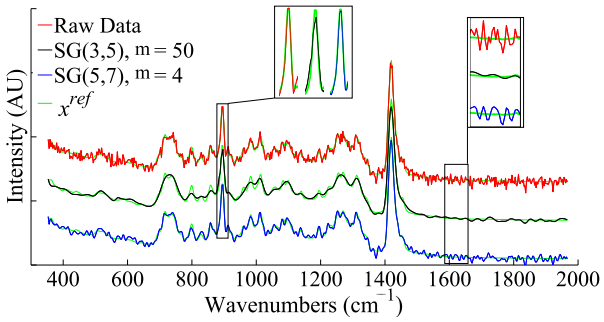

Fig. 3 An illustration of MLESG denoising for different numbers of iterations; denoising of peak regions is optimal at low numbers of iterations, while smooth areas require a significantly larger number of iterations.

peak denoising were derived from maxima in the SNR graphs created for each dataset; from this set of results the optimal number of iterations for both global (m_{max}) and peak (m_{min}) denoising, as a function of input SNR, were found and are illustrated for SG(5,7) in Fig. 4. Ideally, the algorithm should

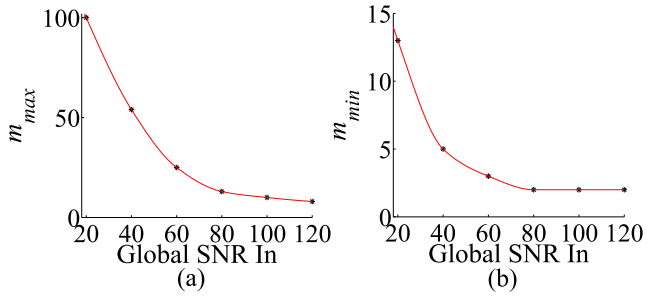

Fig. 4 (a) Number of iterations ($m = m_{max}$) required for optimal denoising of the global spectrum, as a function of input SNR; (b) Number of iterations ($m = m_{min}$) required for optimal denoising of a peak region, as a function of input SNR

provide high levels of smoothing while effectively preserving the integrity of sharp peak features. Thus, it was decided to further develop the algorithm to implement an early stopping procedure (setting $m = m_{min}$) in peak regions while also applying a late stopping procedure ($m = m_{max}$) in smoother regions. In order to avoid sharp discontinuities between regions of early and late stopping, an approach was developed to ensure a gradual change in the number of iterations from one sample to the next. The development of this procedure is discussed in the following section.

3.4 Early stopping at peaks

Equation 9 is comprised of two opposing constraints; the first constraint penalises deviation from the smoothed version of the spectrum, while the second constraint penalises deviation from the raw values. With the first application of Equation 9, an initial denoised estimate of the spectrum is obtained that is more accurate than that produced by SG filtering alone, in terms of SNR. A second application of Equation 9 is likely to produce a second estimate of the denoised spectrum with a further enhanced SNR. This is due to the fact that the smoothed version of the first estimate, which is used in this second iteration, is a more accurate representation of the spectrum than the smoothed version of the raw spectrum that was used in the first iteration. This argument can be applied to each subsequent iteration up to some point for which the spectrum has become over smoothed, and the SNR of the estimate will begin to reduce. In areas where sharp features are present, it is better to apply an "early stopping" strategy, i.e. to use only a few iterations of the algorithm in order to avoid over smoothing, while in areas of the spectrum that contain smooth features, "late stopping", i.e. application of a large number of iterations, will provide higher SNR values.

The number of iterations, m_i, associated with each sample index i, which is imposed by the presence of a peak at a wavenumber given by pk_j is determined using a Gaussian distribution as follows:

$$G(j,i) = exp\left[\frac{-|wavenumber_i - pk_j|^2}{2\sigma_g^2}\right] \quad (13a)$$

$$m_i = min_j[G(j,i)] \times (m_{max} - m_{min}) + m_{min} \quad (13b)$$

where $wavenumber_i$ is the spectrum wavenumber axis as a function of sample index i, and σ_g is the standard deviation of the the Gaussian, all in units of cm^{-1}. The vector pk_j contains a series of k wavenumber peak locations that is input by the user and, therefore, j takes values of 1 to k. m_i denotes the number of iterations that will be applied to the i^{th} sample in the spectrum and m_{max} and m_{min} are as previously described. The $min_j[]$ operator returns the minimum value in the j dimension. The result of applying the algorithm defined by Eq.13 to the polymer spectrum is shown by the red line in Fig. 5, where $\sigma_g = 10$. Samples that are located in large slowly varying regions are associated with high numbers of iterations (m_{max}), while peak locations are associated with a low number of iterations (m_{min}); an appropriate gradient of iterations from m_{max} to m_{min} is calculated by Eq.13 to prevent discontinuities in the denoised spectrum. However, disassociating the flat regions from the peaks in this way allows for greater smoothing in the low frequency regions. Rather than increase m_{max}, and therefore the run-time of the algorithm, the window size of the SG filter and λ are increased for the final 20% of the iterations to produce an improved spectrum by providing an increased rate of smoothing in areas that are relatively flat. A flowchart of the over-all operation of the algorithm is illustrated in Fig. 6

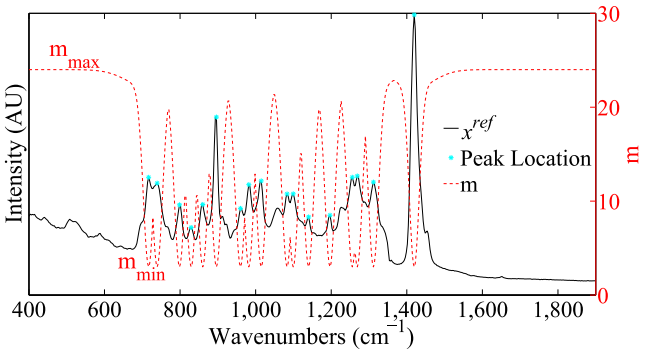

Fig. 5 The number of iterations is determined using Eq.13 and provides an early stopping strategy for the MLESG algorithm in peak regions as well as a smooth transition in iteration numbers from one sample to the next.

Automatic identification of peaks in a noisy spectrum is a challenging process and not within the scope of this paper; therefore, it was decided that the user would input a number of distinct peak wavenumber locations. This is a reasonable approach since many applications involve a set of known peak locations in each recording, e.g. in the case of recording spectra from an epithelial cell, which is discussed in more detail in the following section. This can be achieved by manually inputting a vector of wavenumber locations, which can be time consuming, or by defining a set number of locations and loading them automatically from a text file. In the case of known peak locations this allows the algorithm to be applied as a single post-processing step for individual spectra or as part of a larger, automated process. If this is not the case, a peak for which the wavenumber location is not defined may be subjected to unnecessary smoothing. Two alternative approaches are discussed in the future work section, of the conclusion.

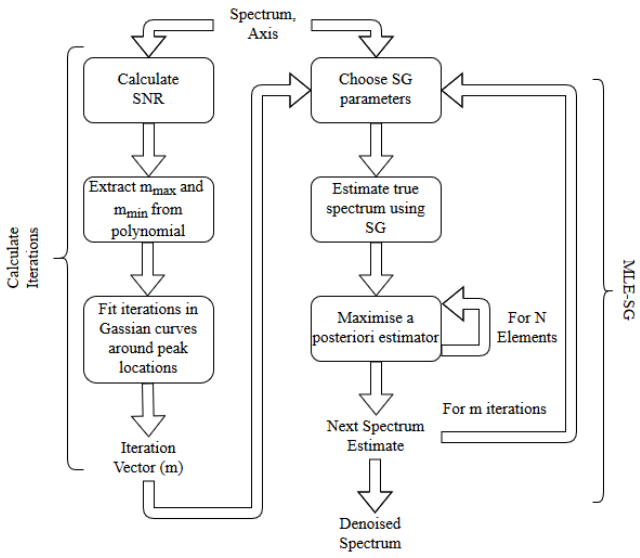

Fig. 6 Flowchart of the algorithm's operation steps.

4 Evaluation

In this section, the MLESG algorithm is applied to experimental data based on multiple recordings from the aforementioned plastic slide. The algorithm is also applied to a spectrum recorded from a biological cell, which has been artificially noised in order to generate a large amount of data for testing. In all cases the performance of the algorithm is compared to results obtained from SG filtering and an algorithm known as 'the perfect smoother', or pSmooth [17] for short, in terms of the three SNR metrics described in Section 3.1; this latter algorithm is based on penalised least squares and has been shown to provide an improvement over SG filtering. In the case of pSmooth, the algorithm has an inbuilt parameter optimiser provided by the author; however, it was noted by the author that pSmooth is optimised to smooth and does not adapt itself to preserve sharp local features. For the SG smoothing input parameters, in all following cases the smoothing parameters are fixed at $v = 3$ and $q = 7$ since these parameters were biased more towards global SNR and so it lends itself to a reasonably fair representation of SGs capabilities.

4.1 Application to experimental data

The tuning of the algorithm described in Section 3 was performed with simulated datasets based on a spectrum recording from an Ibidi polymer slide. In order to test the performance of the algorithm on experimental datasets, spectra from the same slide were recorded using a confocal Raman microscopy system and the SNR of the recorded datasets was controlled by varying the acquisition time. A reference spectrum with low noise was collected using a long acquisition time and subtraction of a dark frame of equal acquisition time; the value of x^{ref} for a given dataset could then be calculated by scaling the reference spectrum appropriately in order to match the acquisition time used to record that dataset; this was done using an Extended Multiplicative Signal Correction (EMSC) algorithm. [18]

Both simulated and experimental data were processed using the MLESG algorithm with early stopping for peaks at the wavenumbers illustrated in Fig. 5 and the denoised spectrum was evaluated in terms of SNR, and compared with the other smoothing algorithms. The results of this analysis show that the SNR of the collected signals can potentially be doubled through applying the MLESG algorithm, which is advantageous in low light applications or in applications where cost or time constraints exist. Results from the experimental and simulated datasets were similar; however, experimental data had a slightly lower improvement in SNR results, which is to be expected due to experimental variability of the signal intensity and dark current both of which result in a variation in the SNR of the raw spectrum. Another possible cause of the slightly lower SNR improvement is the use of a Gaussian noise model, instead of the more accurate Poisson model, for the experimental noise. A qualitative comparison of signal recovery is illustrated in Fig. 7 where the quality of the denoised spectra by the relevant algorithms is illustrated.

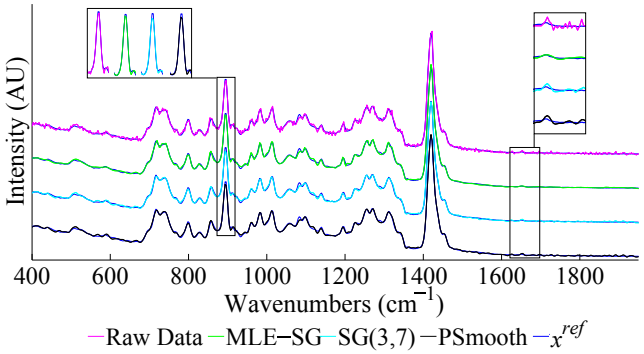

Fig. 7 Qualitative comparison of signal recovery achieved by the three denoising algorithms on an experimental spectrum with an initial SNR of 97. A high resolution image is available online that clearly illustrates the improvement.

In Fig. 7 it can be seen that the flat regions have a significantly lower standard deviation than that of the spectra that have been processed using the other techniques. However, this has had little to no effect on its capacity to preserve the characteristic features of the sharp peak, which has been highlighted in Fig. 7. This is further demonstrated in Fig. 8, which com-

pares the results in terms of the metrics previously discussed in Section 3.1. However, it is difficultly to visually appreciate a small improvement in the SNR of a signal; for example, a spectrum with a SNR of 90 may appear qualitatively similar to a spectrum with an SNR of 100. In order to provide a more rigorous quantitative evaluation, a set of tables that correspond to Fig. 8(a), Fig. 8(b), and Fig. 9 are given in the accompanying ESI.

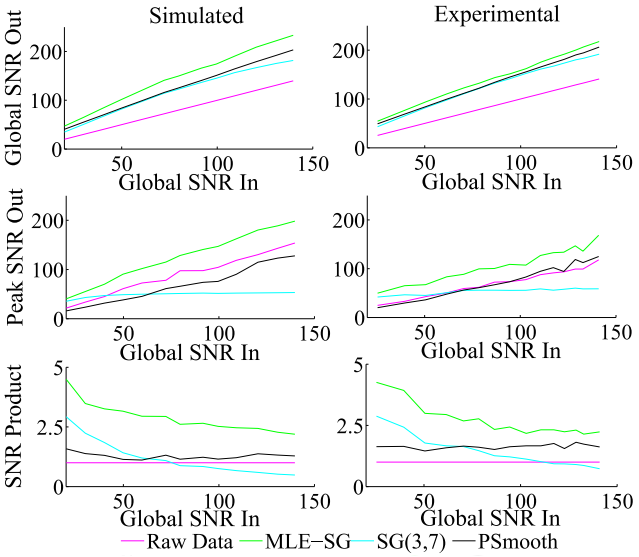

Fig. 8 Comparison of SNR enhancement achieved by denoising algorithms for simulated and experimental datasets of similar SNRs. Corresponding table of values is available in the ESI, see Tables 1 and 2.

Although the experimental datasets were collected with the intention of having matching SNRs to the simulated datasets, this was not strictly possible and so the SNR range for the experimental data is from 33 - 141 approximately; however, the axes ranges have been kept the same for ease of comparison. While the results fluctuate more than the results for the experimental section, they show similar trends; in both cases pSmooth out-performs SG smoothing in terms of global SNR, which is to be expected, and MLESG shows the highest SNR improvement in all contexts. However, in the figures that describe the improvement in peak SNR it is clear that the MLESG algorithm is the only method that consistently and reliably improves the SNR within a peak region, and the SNR product further reflects this with significantly higher values for the MLESG algorithm than for the other methods, in all cases. The algorithms were also evaluated in terms of computational efficiency. The algorithms were implemented using Matlab running on a Dell Inspiron 15 with an Intel Core i7 processor. The average time taken for MLE-SG, SG, and PSmooth was 195 ms, 0.8 ms, and 34.8 ms, respectively. As expected, SG smoothing provides the fastest implementation.

4.2 Application to biological spectra

It may be difficult to record a reliable and low-noise reference Raman spectrum from a biological sample that could subsequently be used in an accurate quantitative evaluation of improvements in SNR for a denoised spectrum. An accurate representation of the irradiance would require a long exposure time on one sample point and would likely result in photo-bleaching/damage. In addition, biological cells are often biochemically heterogeneous at different locations in a single cell, as well as across a group of similar cells; such a heterogeneity presents an additional complexity in terms of finding an accurate reference spectrum that could be used in a quantitative assessment of SNR over a dateset. Therefore, it was decided to test the algorithm on a simulated dataset based on one high quality cell spectrum that is artificially noised; in this case the reference spectrum is the original cell spectrum before the addition of noise. Considering the similarity in the results between the experimental and simulated polymer datasets, it was inferred that the simulated cell spectra would provide a suitable representation of the algorithm's capabilities for this application. A low noise reference spectrum was generated by adding together more than fifty spectra recorded from a high grade bladder cancer cell line, following formalin fixation; cell preparation, recording, and appropriate processing methods.[19] The end result is a reference spectrum of 500s acquisition time recorded from the nucleus of 50 cells from this cell line using a 120mW 785nm laser. This low noise reference was then artificially noised as described in Section 3, and 17 datasets were generated with SNR values from 20 to 200 in steps of 10, each containing 100 spectra. Following this, the spectra were denoised using the MLESG algorithm with early stopping at appropriate wavenumber locations and the average improvement in SNR was measured for each dataset. Among the marked peak locations (in cm^{-1}) are: 785, 1004, 1090, 1127, 1262, 1319, 1341, 1451, 1585, 1619, and 1662. These peak number locations correspond to well known biochemical assignments in epithelial cells, as shown in Table 1.[20,21]

Wavenumber (cm^{-1})	Chemical Bond	Association
785 - 788	Stretching of DNA related bonds and DNA/RNA breathing modes	Nucleic Acid
1004	Phenylalanine	Protein
1090	Stretching of DNA related bonds Stretching of C-N backbone	Nucleic Acid Protein
1127	Stretching of C-N backbone Stretching of C-C	Protein Lipid
1262	DNA/RNA breathing modes Amide III	Nucleic Acid Lipid
1319	CH2, CH3 twisting DNA/RNA breathing modes CH deformation vibration	Lipid Nucleic Acid Protein
1341	DNA/RNA breathing modes CH deformation vibration	Nucleic Acid Protein
1451	CH_2 deformation vibration	Protein/Lipid
1585	DNA/RNA breathing modes	Nucleic Acid
1619	Tyrosine; tryptophan	Protein
1662	DNA/RNA breathing modes Amide I Fatty Acids	Nucleic Acid Protein Lipid

Table 1 A table of common biochemical assignments in epithelial cells

The standard deviation of the Gaussian modelling the peak regions was kept at 10, as in the previous section. The results, together with corresponding results for the PSmooth algorithm, and SG filtering (with polynomial and window sizes of 3 and 7) respectively, are shown in Fig. 9. In almost all cases the MLESG algorithm outperforms the other algorithms, with a minimum of 50% improvement in SNR compared to the raw data. For the low SNR case pSmooth shows a comparable result to MLESG for the global spectrum due to the higher amount of smoothing generated by that algorithm; however, the improvement in the SNR in the region of the phenylalanine peak is significantly higher for MLESG.

Similar results are observed to those in the previous section. The peaks have been effectively preserved and the large flat regions have a lower standard deviation than that provided by the other two denoising algorithms. In all cases the MLESG algorithm preserves the peaks better than the other two methods, while also out-performing the other methods in terms of global smoothing. The SNR product clearly demonstrates the superiority of the algorithm for all cases of input SNR by taking into account both global and peak SNR improvement in a single metric. Despite the tuning of the algorithm using the polymer spectrum, which has a significantly different spectral form the algorithm still provides a superior SNR enhancement over the other algorithms and produces high quality denoised spectra. This indicates that the algorithm performs robustly across different types of spectra. This does not preclude the possibility of improving performance through spectra-specific tuning.

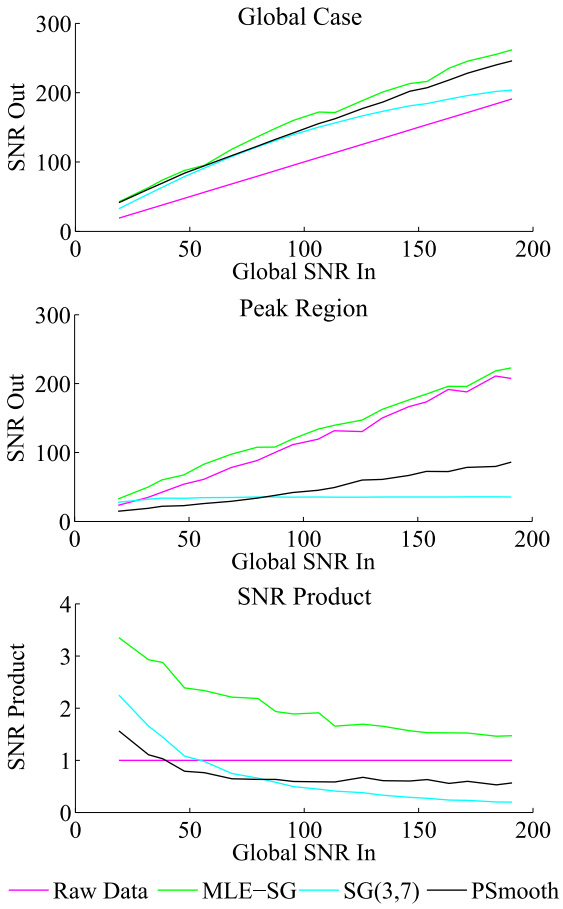

Fig. 9 Comparison of SNR enhancement for simulated T24 datasets achieved by denoising algorithms. A corresponding table of values is available in the ESI, see Table 3.

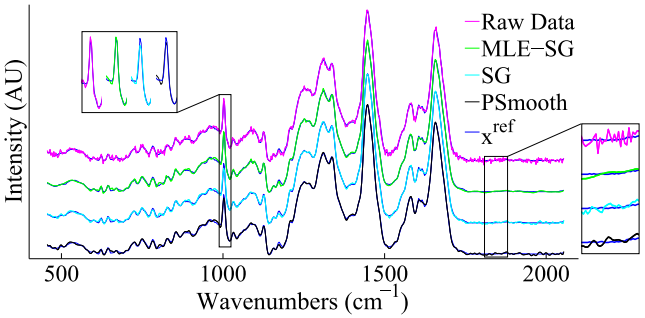

Fig. 10 Comparison of SNR enhancement achieved by denoising algorithms for an initial SNR of 50

5 Conclusion and Future Work

This paper has demonstrated how Savitzky-Golay filtering may be enhanced with Maximum Likelihood Estimation to produce an algorithm that consistently out-performs competing algorithms. MLE provides bounding properties, based on the noise distribution associated with the signal, to prevent the SG smoother from significantly altering the underlying spectral features. The algorithm is iterative, with increased smoothing occurring with each iteration; inclusion of an early stopping procedure, based on a user input of peak locations, further ensures that sharp local features are effectively preserved while allowing further smoothing in low frequency regions. The resulting algorithm provides up to a 100% improvement in SNR when compared to the raw data, consistently out-performing competing algorithms (PSmooth and SG filtering) in terms of all metrics used to evaluate algorithm performance. While inputting all peak locations of interest may not always be possible, particularly if unexpected components exist, many applications are based on recording a known sample repeatedly, therefore, the features of interest are generally well known and will require the same wavenumber locations for subsequent experiments.

The proposed algorithm was optimised and rigorously tested on simulated datasets based on a polymer spectrum before being tested on experimentally collected datasets; a close correspondence was observed in the results for the simulated and experimental datasets. Finally the algorithm was tested on simulated datasets of epithelial cells and the results showed similar trends in SNR improvement despite there being no retuning of the algorithm.

Another contribution of this paper is the development of a rigorous approach to evaluate Raman smoothing algorithms in general in terms of SNR and the proposed SNR product metric. This analysis is based on estimating the RMSE with respect to a known reference and is applied to both the global spectrum, as well as in a peak region; since these regions can be adversely affected by smoothing. The proposed metric known as the SNR product, i.e. the product of the improvement in global SNR multiplied by the improvement in peak SNR, is used to monitor the overall spectral quality provided by the denoising algorithms. This allows the user to evaluate how effectively the algorithm preserves peaks and smooths low frequency regions simultaneously. We believe that this paper constitutes the first attempt to rigorously investigate the effects of smoothing algorithms on Raman spectra in terms of SNR.

Recently, a blind deconvolution algorithm has been proposed that appears to have some similarities to the denoising method presented here[22,23]. Their method also makes use of the maximising a posteriori technique in an iterative manner, and uses a modified Tikhonov regularization model that appears to be similar to the constraint used in our approach that penalises deviation from neighbouring values. Their method also includes a deconvolution process during each iteration in order to take into account, and correct for, the system response function, which is also varied with each iteration. This approach is demonstrated to recover highly degraded and noisy Raman spectra, particularly for cases in which spectral structure is corrupted due to the instrument response. Although this blind deconvolution algorithm has some similarities to the proposed method, both are derived in fundamentally different ways, and each has its own unique characteristics. More work is needed to fully elucidate the relationship between the two algorithms and to compare their results.

A further avenue of investigation to improve performance is to vary λ across the spectrum in a similar manner to that of the adaptive regulariser discussed in the previous paragraph.[22,23] By varying this parameter rather than m it may be viable to produce a result in fewer iterations, perhaps as few as one. It may also limit the requirement for a peak identifier in the algorithm although this method could be challenging to implement for exceptionally narrow spectral features. Another avenue for improvement is to include an automatic peak identifier[24] in the operation of the algorithm to negate the need for user input; as well as having the benefit of including unexpected spectral peaks in the early stopping process.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was conducted with the financial support of the Irish Research Council (IRC) under project ID GOIPG/2013/1434 and Science Foundation Ireland (SFI) under Grant Number 15/CDA/3667. I would like to thank the IRC and SFI for their support.

References

1 B. Saleh and M. Teich, *Fundamentals of Photonics*, Wiley Series in Pure and Applied Optics, 2007.
2 P. Jess, D. Smith and M. M. et al., *International Journal of Cancer*, 2007, **121**, 2723–2728.
3 J. Janesick, *Scientific charge-coupled devices*, SPIE - The international society for optical engineering, 2000.
4 A. Savitsky and M. Golay, *Analytical Chemistry*, 1964, **36**, 1627–1639.
5 P. Heraud, B. Wood, J. Beardall and D. McNaughton, *Journal of Chemometrics*, 2006, **20**, 193–197.
6 N. Afseth, V. Segtnan and J. Wold, *Applied Spectroscopy*, 2006, **60**, 1358–1367.
7 F. Scholz, *Maximum Likelihood Estimation*, Encyclopedia of statistical sciences, 1985.
8 H. Takeuchi and I. Harada, *Applied Spectroscopy*, 1993, **47**, 129–131.

9 L. Zhang and M. Henson, *Applied Spectroscopy*, 2007, **61**, 1015–1020.

10 D. Montgomery and G. Runger, *Applied statistics and probability for engineers*, John Wiley and Sons, Inc., 5th edn, 2011.

11 M. Hirsch and R. W. et al., *PLOS one*, 2013, **8**, e53671.

12 D. Dussault and P. Hoess, *Optical Science and Technology, the SPIE 49th Annual Meeting*, 2004, 195–204.

13 H. Burger, B. Schölkopf and S. Harmeling, *IEEE International Conference on Computational Photography (ICCP)*, 2011, 1–8.

14 M. Gomez-Rodriguez, J. Kober and B. Schölkopf, *IEEE International Conference on Computational Photography (ICCP)*, 2009, 1–9.

15 B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, MA, 1985.

16 T.J. Harvey et al, *Journal of biophotonics*, 2009, **2**, 47–69.

17 P. Eilers, *Analytical Chemistry*, 2003, **75**, 3631–3636.

18 N. Afseth and A. Kohler, *Chemometrics and Intelligent Laboratory Systems*, 2012, **117**, 92–99.

19 L. Kerr, K. Domijan, I. Cullen and B. Hennelly, *Photonics and Lasers in Medicine*, 2014, **3**, 2193–0643.

20 R. Kiselev, I. Schie, S. Aškrabić, C. Krafft and J. Popp, *Biomedical Spectroscopy and Imaging*, 2016, **5**, 115–127.

21 Z. Movasaghi, S. Rehmen and I. Rehmen, *Applied spectroscopy*, 2007, **42**, 493–541.

22 H. Liu, Z. Zhang, J. Sun and S. Liu, *Photon. Res.*, 2014, **2**, 168–171.

23 T. Liu, H. Liu, Z. Chen and A. M. Lesgold, *IEEE Transactions on Industrial Informatics*, 2018, **PP**, 1–1.

24 Y. Tian and K. Burch, *Applied Spectroscopy*, 2016, **70**, year.