Adaptive Band Target Entropy Minimization: Optimization for the Decomposition of Spatially Offset Raman Spectra of Bone.

John H. Churchwell,[†,‡,*] Kay Sowoidnich,[†,‡] Oliver Chan[†], Allen E. Goodship,[†] Anthony W. Parker[†,‡], Pavel Matousek[†,‡]

[†]Department of Medical Physics and Biomedical Engineering, UCL, London, WC1E 6BT, UK.

[‡]Central Laser Facility, Research Complex at Harwell, STFC Rutherford Appleton Laboratory, Harwell Campus, OX11 0FA, UK.

[*]Corresponding Author: j.churchwell@ucl.ac.uk

Abstract: We report a novel variant of Band Target Entropy Minimization (BTEM), Adaptive Band Target Entropy Minimization (A-BTEM) that offers an improved ability to accurately decompose mixed spectra obtained from complex multicomponent systems. Several key challenges have existed in the application of the basic BTEM approach to decompose Spatially Offset Raman Spectra (SORS) of bone underneath soft tissues and other multi-layer systems demonstrating high collinearity between the spectra of individual components; these have included instabilities, signal mixing and spectral artefacts which have precluded the reliable use of BTEM in such situations. By using automatic factor selection, an adaptive penalty function in addition to metaheuristic optimization strategies we demonstrate that high quality spectral reconstructions of underlying pure component spectra can be obtained with typical correlation coefficients >0.996. Furthermore, we ascertain the behaviour of A-BTEM with different input datasets, both synthetic and real, displaying varying signal-to-noise ratio, signal composition and numbers of spectra. We thereby identify the multifarious parameter space in the application of A-BTEM and the quality of data required for the most accurate spectral estimates of pure component spectra.

[Keywords: spatially offset Raman spectroscopy, bone disease, self modelling curve resolution, multivariate curve resolution, band target entropy minimization]

1. Introduction

Band target entropy minimization (BTEM) is a multivariate analytical technique for the blind reconstruction of spectra from mixtures where prior knowledge of the mixture composition is unknown. Originally developed using the deconvolution of infrared spectra of organometallic mixtures[1] and mixtures of common laboratory solvents[2] as examples, it has been applied to data from multifarious analytical techniques including Raman spectroscopy, hyperspectral imaging,[3] mass spectroscopy and nuclear magnetic resonance.[4] Despite the publication of these articles, currently no published work has focussed on the systematic *optimization* of the BTEM approach for the deconvolution of spectra of chemical mixtures where some *a priori* information on the spectral target is known. Furthermore, there is no available comprehensive quantitative information in the literature that identifies the quality of Raman spectroscopic data required for optimal spectral reconstruction. Whilst there are many articles describing novel application-specific results, these provide limited information on how best to apply the BTEM method to acquire the required spectral estimates. Overall, there is a need to develop approaches which require little or no supervision that provide trustworthy spectral reconstructions, particularly when specific analytes are known to be present within a multi-component system.

Spatially Offset Raman Spectroscopy (SORS) is a Raman technique where the point of sample illumination on the surface of the specimen is spatially offset from the point of signal collection. The spatial volume from which the major component of scattered light is collected moves deeper into the sample as the offset is increased within the limit of detectable signal of sufficient signal-to-noise ratio (SNR).[5,6] SORS has found numerous applications including explosives detection at airport security screening checkpoints as well as the verification of pharmaceutical raw materials[7] and shows great potential in the biomedical area.[8] Several groups including ours are currently in the process of developing SORS for *in vivo* bone disease diagnostics to discriminate between healthy and diseased bone for selected bone conditions including *osteoarthritis,*[9] *osteoporosis*[10,11] and *osteogenesis imperfecta*[12]. Recently our group has also conducted experiments exploring photon migration properties in bone and investigated the depth origin of the major signal component as a function of spatial offset.[13,14] A key part of applying SORS for *in vivo* bone disease diagnosis is separating bone signal from spectral contributions associated with the soft tissue strata above it.

The routine use of vibrational spectroscopy, particularly Raman spectroscopy, for the *in vivo* diagnosis of bone disease is still yet to be established. It is well known that the chemical composition of bone determines overall bone strength and disease status. However, current imaging modalities such as dual band x-ray absorptiometry (DXA) only provide information of the hydroxyapatite component but not the organic phase that also determines strength, therefore bone strength is only indirectly associated with DXA data.[15] Despite extensive work and some success in detecting abnormal bone composition both *ex vivo* and *in vivo* in both animal models and human subjects key challenges remain in spectral analysis of SORS datasets particularly in the separation of the bone signal component from that of soft tissues.

In simple two-layer systems where light absorption is negligible and a suitable isolated band exists for the surface layer a simple scaled subtraction can be carried out to remove the surface contribution from the SORS spectra. However, in the case of biological tissues pertaining to *in vivo* bone data acquisition where the major spectral components of the surface layer are shared with that of bone, and other layers are present, the problem is more complex and self-modelling curve resolution (SMCR) methods, independent components analysis (ICA)[16] or multiway methods such as parallel factor analysis (PARAFAC) are necessitated.[17] The application of several currently available multivariate techniques for SORS data deconvolution have been investigated.[18,19] Where the data show adequate trilinearity PARAFAC is the optimum choice (for instance when *in vivo* SORS datasets of bone are obtained from multiple anatomical locations), but for systems where the data are not sufficiently trilinear SMCR two-way methods must be employed. Therefore, to acquire meaningful pure bone spectra specific to an anatomical location in an individual, multiway methods are intrinsically precluded.

1.1 Band Target Entropy Minimization (BTEM)

BTEM was based upon the work of Sasaki *et al.* who first described the use of Shannon information entropy to decompose mixed spectra.[20] Sasaki *et al.* used entropy minimization to estimate pure component spectra for whole mixtures simultaneously arguing that by minimizing the information entropy for each component you can accurately and realistically model the spectra for the entire system. Approaches using target bands and dissimilarity functions were developed by Garland and co-workers.[21] Building upon their initial work they suggested the idea of one-at-a-time reconstruction using both information entropy and a target band in addition to the traditional self-modelling-curve resolution (SMCR) constraints of spectral and concentration non-negativity.[2] These additional constraints helped to overcome the problem of rotational ambiguity, providing numerically unique solutions in the limit of high SNR data. Since then, several improvements and modifications have been implemented to automate the decomposition of mixtures composed of an unknown number of component species[22] in addition to generating multiple spectral estimates without the need for band specific targets.[23]

Despite these developments several problems remain in BTEM's application to systems where there is high spectral overlap or collinearity between constituents or where data collection is heavily signal limited and noisy. This is particularly so for the decomposition of *in vivo* SORS datasets of bone, where one seeks to obtain an extremely accurate and stable BTEM spectral estimate of pure bone from datasets where this signal is mixed with contributions of soft tissues and where these components share similar spectral features. Typical example average differences for univariate Raman spectral band ratios between diseased and healthy bone are approximately: *osteoporosis* 5-10%,[10] *osteoarthritis* 7%[9] and *osteogenesis imperfecta* 11%[12] thus placing extreme emphasis on achieving high quality reconstructions. Although we have had some success in ideal situations, the spectral estimates obtained from routine SORS screening of patient volunteers are often of inferior quality using the BTEM method as originally described. This problem is particularly evident for spectra acquired with short integration time as required in a clinical setting where owing to laser safety standards power densities are severely restricted and the SNR of data obtained is often limited.[10]

Recently Maher *et al.* have found that over-constrained library based fitting could provide an alternative means to produce reliable spectral estimates for SORS datasets.[19] They noted specific difficulties in applying BTEM to *in vivo* SORS data to acquire pure bone spectra and assert that BTEM is not a suitable method to extract pure bone signals. More recently de Juan and co-workers have investigated using MCR-ALS and BTEM in succession to improve the spectral reconstructions obtained from a number of data types.[24] However, this does not address the underlying issues in BTEM and the results showed little improvement for Raman spectral estimates as compared to using BTEM alone. A key challenge of generating spectral estimates using optimization techniques is ensuring the stability and accuracy of the spectral reconstructions; there are often differences in the spectral estimates calculated on repeat runs using the same penalty function and optimization routines – these have not been characterized to date.

In this paper, we highlight the fundamental problems associated with applying the basic BTEM approach to acquiring pure spectra of bone from *in vivo* SORS data including instability and common spectral artefacts. We then describe the development of an advanced variant of the BTEM method, Adaptive-BTEM (A-BTEM) that affords more stable and accurate spectral estimates. In addition, we ascertain the quality of SORS data required to enable high quality reconstructions. The proposed improvements and modifications to the BTEM method include:

1) Automatic selection of the optimum number of retained basis vectors.
2) Enhanced adaptive penalty function that is dynamically adjusted during the optimization process to prevent competing terms inducing spectral artefacts, i.e. distorted spectral estimates, whilst enhancing stability.
3) The use of linked metaheuristic optimization strategies to improve spectral estimate accuracy and stability between repeat runs.

These improvements are general and will be useful to anyone considering using the BTEM method particularly where there is significant collinearity between the Raman spectra of the chemical components in the system being examined. The analytical methods developed and presented here are applicable to many types of spectroscopy where the output data are continuous.

1.2 Theory

The basic BTEM algorithm has been described in detail elsewhere.[2,25] Briefly, BTEM is initiated via the singular value decomposition (SVD) of the pre-processed data matrix $\mathbf{D}_{k \times v}$,

$$\mathbf{D}_{k\times v} = \mathbf{U}_{k\times k}\boldsymbol{\Sigma}_{\boldsymbol{k}\times \boldsymbol{v}}\mathbf{V}^{\mathrm{T}}_{v\times v} \qquad (1)$$

Where $\mathbf{U}_{k\times k}$ are the left singular vectors, $\boldsymbol{\Sigma}_{\boldsymbol{k}\times\boldsymbol{v}}$ are the singular values, $\mathbf{V}^{\mathrm{T}}_{v\times v}$ are the right singular vectors, k is the number of spectra or rows and v is the number of columns or wavenumber variables in the dataset. In general, it is at this stage one inspects the right singular vectors $\mathbf{V}^{\mathbf{T}}$ for characteristic spectral features to target (manually or automatically) but for Raman spectra of bone this is not required as the target peak is well established (~960 cm^{-1} $\nu_1 PO_4^{3-}$).[26]

Factor compression is also carried out following the initial decomposition, whereby only the significant eigenvectors are retained for the subsequent reconstruction. The number of factors retained by BTEM varies considerably in the literature. Typically, the number chosen is system dependent, but usually one of two general strategies is applied for their determination. One is to retain a very large number of basis vectors[3] - even arbitrarily large[22] – the second is to select a relatively small fixed number for a given system type.[10] To date there has only been limited discussion of the rationale for selecting the precise number of factors and no routine implementation of suitable statistical tests to aid appropriate factor compression exists – some authors have chosen the number visually.[27] Selecting the optimum number of retained eigenvectors is important as retaining an excessive number of factors increases the number of decision variables required for the BTEM estimate which when excessive can induce numerical instability during optimization and increase the computational time. In addition, retaining an excess of vectors increases the noise contribution in the final spectral estimate unnecessarily. On the other hand, retaining too few basis vectors removes physically meaningful information that would otherwise contribute to the spectral estimate and increase its accuracy and could prevent reconstruction of very low signal components. For error free measurements, the number of basis vectors required for an adequate model of the data is equal to the number of distinct chemical components in the system.[28] In the presence of errors or noise the physically meaningful information is spread into eigenvectors of lower value and mixed with noise. This implies that factor compression can reduce the errors or noise but cannot entirely remove them as some error is held within the significant eigenvectors. Therefore, the appropriate number of eigenvectors to retain not only depends upon the chemical composition of the precise system under investigation, but also on the quality of data obtained from the system and the number of spectra available.

The choice of test to estimate the required number of basis vectors depends on whether accurate knowledge of the errors associated with a given measurement exists. Where such knowledge is lacking, empirical methods are required. Several such tests are available, and these often provide different numbers of meaningful eigenvectors. One of the most well-known and successful tests in the factor analysis literature is Malinowski's indicator function (IND); for our work here, we have included this as an initial step in the A-BTEM implementation to ensure that factor compression and the selection of the number of basis vectors is carried out as rigorously as possible. From our recent *in vivo* trial, the number of eigenvectors required for a given dataset for an anatomical location varies from patient to patient and using a fixed number will undermine many BTEM decompositions. Malinowski's IND function is given by,

$$\mathrm{IND}_z = \frac{1}{(k-z)^2} \cdot \left(\frac{\sum_{j=z+1}^{k} \lambda_j^0}{v(k-z)}\right)^{\frac{1}{2}} . \qquad (2)$$

Here v is equal to the number of columns in the data matrix $\mathbf{D}_{k\times v}$ which is the number of data channels or variables in each spectrum, k is the number of rows in $\mathbf{D}_{k\times v}$ or the number of spectra in the dataset, z is the number of retained eigenvectors and λ_j^0 is the eigenvalue associated with the j^{th} eigenvector. The correct number of eigenvectors is identified when IND_z reaches a minimum for a given z.

Following factor compression, i.e. the process of retaining only the significant number of eigenvectors, a spectral estimate, $\hat{a}_{k\times v}$, is obtained,

$$\hat{a}_{1\times v} = t_{1\times z}\mathbf{V}_{\mathbf{z}\times\boldsymbol{v}}^{\mathbf{T}} \qquad z \geq s . \qquad (3)$$

Where z or the number of retained factors is greater than or equal to the number of contributing species s. In practice, the elements of the transformation vector $t_{1\times z}$ are determined numerically by a global optimization process (traditionally simulated annealing[2] although simplex[19] and particle swarm[24] have been used), minimizing a penalty function. During each iteration in the optimization the spectral estimate obtained is normalized to the maximum intensity. The basic penalty function used in BTEM,

$$\min F = -\sum_v h_v \ln h_v + P\left(\hat{a}_{1\times v}, \hat{C}_{k\times 1}, \hat{a}_{1\times v}^{max}\right) \qquad (4)$$

includes a term for the Shannon information entropy,[29] in which h_v is defined by,

$$h_v = \left|\frac{d\hat{a}_v}{dv}\right| \Big/ \sum_v \left|\frac{d\hat{a}_v}{dv}\right| . \qquad (5)$$

In addition, it also includes a penalty function, P, containing the terms pertaining to spectral non-negativity, concentration non-negativity and a factor that penalises estimates where the target peak is not the maximum peak in the spectral estimate,

$$P\left(\hat{a}_{1\times v}, \hat{C}_{k\times 1}, \hat{a}_{1\times v}^{max}\right). \qquad (6)$$

Typical numerical weights of each of the non-entropy terms can be found in several papers.[1],[22] Extensions of the basic BTEM approach suggested in an early paper[2] involve the use of the sum of absolute spectral derivatives (first, second, or forth order) and the numerical integrated intensity of the spectral estimate in addition to the basic Shannon information entropy term or sum of higher order derivatives,

$$\min F = \sum_v |ds_v| + \sum_v |\hat{a}_v| + P(\hat{a}_{1\times v}, \hat{C}_{k\times 1}, \hat{a}_{1\times v}^{\max}) \qquad (7)$$

where,

$$\mathrm{d}s_v = \frac{d^n \hat{a}_v}{dv^n} \qquad (8)$$

and n is the order of differentiation.

We propose a new adaptive penalty function as part of A-BTEM, where the integrated intensity term is normalized to the number of spectral channels, v, and is 'switched-off' in the limit of a *'reasonable'* solution, i.e. below a threshold normalized integrated intensity. For a given target feature, this numerically emphasizes the region of factor space in which the best spectral estimate is found. Our novel penalty function is shown in equation (9),

$$\min F = \sum_v |ds_v| + \left(\frac{\sum_v |\hat{a}_v|}{v}\right)\beta + P(\hat{a}_{1\times\upsilon}, \hat{C}_{k\times 1}, \hat{a}^{\max}_{1\times v}) \qquad (9)$$

where,

$$\beta = \begin{cases} 1, & \sum_v |\hat{a}_v|/v > \delta \\ 0\,, & \sum_v |\hat{a}_v|/v \le \delta \end{cases} \qquad (10)$$

Here δ is equal to the threshold normalized integrated intensity value. For a known target species, normalized reference spectra can be used to obtain a conservative initial estimate of δ. However, for datasets where no *a priori* knowledge of the system composition is known, it is necessary to obtain appropriate δ for a given targeted feature in $\mathbf{V}^{\mathbf{T}}_{\mathbf{z}\times\boldsymbol{v}}$ by trial-and-error and careful examination of the spectral estimates obtained for each threshold value tested following an initial interpretation of the species likely present.

Owing to the random initialization and probabilistic nature of the simulated annealing algorithm, different spectral estimates will be obtained on different runs of *traditional* BTEM; in fact, simulated annealing is best used where approximate solutions are required. We introduce the use of a metaheuristic optimization method, 'Global Optimum Determination by Linking and Interchanging Kindred Evaluators' (GODLIKE),[30] for minimizing the penalty function in A-BTEM. This optimization method operates by simultaneously running several basic implementations of different population based heuristic optimizers that are linked; namely the genetic algorithm, differential algorithm, particle swarm and adaptive simulated annealing. The members of each method's populations are swapped with some probability during each run which acts to reduce the likelihood of the overarching method returning a secondary minimum thereby increasing robustness and final accuracy of the spectral estimates obtained. Furthermore, it is less sensitive to problem specific tuning compared to traditional heuristic optimization methods as it takes advantage of the strengths of the different underlying methods. A summary description of the A-BTEM algorithm is described in section 2.3 below.

2. Materials and Methods

2.1 Experimental SORS Data Acquisition

A fresh frozen cadaveric human leg from a 93-year-old female donor was obtained from the Vesalius Clinical Training Centre at the University of Bristol with the appropriate MTA and ethical approval (London Stanmore REC - 08/H072/34) and stored in a freezer at -80 °C. The leg specimen was thawed for 24 hours and the upper leg and foot were subsequently removed by an orthopaedic surgeon leaving the tibia and fibula surrounded by original soft tissue including skin. The specimen at room temperature was then used for transcutaneous SORS spectroscopy by selecting a suitable site on the front medial aspect of the tibia. Subsequently the soft tissue was thoroughly removed, and 10 Raman spectra were acquired directly from the exposed bone surface corresponding to the site sampled during transcutaneous measurements.

SORS measurements were performed using an inverse SORS custom-built system (Cobalt Light Systems Ltd, Oxfordshire, UK). A diode laser at 830 nm was used as the excitation source, the nominal optical power at the sample was 320 mW. The instrument delivers the light as a user definable annulus of variable spatial offset or radius. The scattered light is always collected from a single point at the centre of the annulus in a ~1.5 mm diameter spot. In this study, for transcutaneous measurements the chosen spatial offsets or radii were 3, 6 and 9 mm. The ring width of the different annulae was ~1 mm. For each spatial offset 10 separate spectra were acquired for 0.5 s and 360 accumulations (i.e. with total acquisition time of 180 s per single spectrum) to enable the evaluation of our novel BTEM implementation on averaged experimental data of varying signal-to-noise ratio.

On the excised and cleaned tibia, measurements were carried out at 0 mm offset from 10 sites on the bone surface corresponding to an area within the maximal annulus chosen for the transcutaneous measurements. At 0 mm offset, the instrument is effectively acquiring point scan measurements from the surface of the bone, but the probed volume is significantly greater than a standard confocal Raman microscope.[13] These spectra were acquired for 0.5 s with 120 accumulations (i.e. for 60 s integration time).

The scattered Raman light from the central collection zone was focused into a low-loss Optran WF fiber bundle (CeramOptec, East Longmeadow, MA) connected to a spectrograph (Raman Explorer, Headwall, MA) equipped with a CCD detector (Andor iDus 420 BR-DD; Andor, Belfast, Northern Ireland). The detector had a spectral resolution of 8 cm^{-1}. Spurious cosmic ray spikes were removed automatically by the instrument's operating software. To afford high efficiency collection, the fibre optic collection bundle used a round configuration of 33 fibres on the sample side, whilst at the spectrograph end was configured in a linear arrangement to optimally fill the spectrometer slit and the vertical axis of the detector.

2.2 Synthetic Data

Two extremely high SNR archetypal spectra of bone and soft tissue were created from large libraries of excised bone and skin/soft tissue spectra by averaging. These were subsequently smoothed with a Savitzky-Golay filter using a 3rd order polynomial and 11-point frame length and normalized to their respective spectral maxima. After smoothing the SNR of the two spectra was ~150.

Based on these two spectra synthetic SORS datasets where then prepared by generating ranges of combinatory ratios between maximum and minimum bone contributions. Datasets with different predetermined numbers of spectra were prepared for each range. The maximal ratio value was varied to emulate SORS datasets with different overall bone signal contribution levels *e.g.* SORS data acquired from different anatomical locations where the thickness of soft tissue is greater for one location or for the same location for different individuals with widely different body mass index (BMI). The mean maximal ratio of bone signal to soft tissue signal used was 4:5 in line with data obtained as part of our previous *in vivo* measurements.

For each of the smooth SORS datasets specified above a further eight matching datasets were created with various levels of added Gaussian white noise to simulate experimental data acquired with poorer signal-to-noise ratio to compare spectral estimates obtained from these with those obtained from the high SNR versions.

2.3 Data Analysis

Raw experimentally acquired spectra were imported into MATLAB (Mathworks, Natick, MA) and pre-processed using several custom functions. All spectra were first corrected for CCD sensitivity variation and the fluorescence background was removed using an implementation of the mollifier algorithm previously described by Koch *et al.*[31] The resulting corrected spectra were normalized (to the phosphate mode at ~960 cm^{-1} if obtained from excised bone or to the CH_2 wagging mode at ~1450 cm^{-1} if SORS measurements) and either used as is if acquired from excised bone or if transcutaneous SORS measurements, subsequently submitted to BTEM or A-BTEM decomposition. The relevant penalty functions and optimization routines were specified depending on the specific requirements and those chosen are made clear in the following results section. All BTEM and A-BTEM spectral estimates were carried out on a MacBook Pro with a 2.3 GHz quad core Intel i7 processor and 16 GB of RAM. Typical run times for A-BTEM were around 5 minutes per spectral estimate, although on our machine four estimates could be run in parallel effectively reducing the total processing time. For clarity a schematic showing the structure of the new A-BTEM algorithm can be found in the supplementary information (see figure S1). The algorithm begins with the input of the pre-processed Raman spectra, the spectral range (start and end wavenumbers) of the target band and the target spectrum threshold integrated intensity, δ, required for the adaptive penalty function. Subsequently singular value decomposition is performed on the pre-processed spectra and Malinowski's indicator function is calculated for different numbers of retained factors. Following the determination of the minimum number of retained factors the singular vectors and values are truncated according to the number of retained factors. The transformation vector, t, of length equal to the minimum in IND is initialized. The elements of t are then adjusted iteratively by GODLIKE so as to minimize the adaptive penalty function (see equation (9) above) by finding the optimal linear combination of right singular vectors and thereby return the optimal spectral estimate for the target peak and threshold integrated intensity.

3. Results and Discussion

3.1 Selection of the number of basis vectors

As described in the theory section above, we have chosen Malinowski's indicator function to select the optimum number of eigenvectors to retain during factor compression. In general, the number of factors containing physically meaningful information increases for real *in vivo* SORS data of human bone compared to synthetic data. This is due to additional sources of variance in the data. In experimental situations, the number of factors required will change from person to person and at different anatomical sites. Figure 1 shows IND as a function of the number of retained factors for two anatomical sites from the same human tibia, notice that the function's minimum value occurs at different numbers of retained factors, 3 and 4 respectively. This plot highlights intrinsic variation that is inextricable from practical SORS data acquired from human subjects and highlights the importance of choosing the correct number of factors when carrying out BTEM on *in vivo* SORS data. Furthermore, the number of required basis vectors also increases with noise and the number of spectra in the dataset. For a fixed SNR increasing the number of spectra increases the decision variables required for the optimization aspect of BTEM.

3.2 Existing methods

Spectral estimates calculated using traditional BTEM from the synthetic high SNR SORS dataset containing 17 mixed spectra that mimics the mixture ratios of average *in vivo* data, i.e. bone to soft tissue signal ratios ranging from 0:1 to 4:5, are shown in Figure 2. (a) - (d). Each subfigure shows the results obtained by using different but previously described penalty functions and the simulated annealing optimization routine with optimal annealing parameters. For each function, our BTEM program was run fifty times to assess the stability of the spectral estimates obtained.

It is clear from inspection of Figure 2 (a) that using the basic BTEM penalty function, which only includes the Shannon information entropy, does not produce realistic spectral estimates of bone in almost all runs. In addition, the spectral estimates obtained highlight a very high degree of optimization instability as they are markedly different from run to run. In most runs, there was a large degree of signal mixing from subcutaneous adipose tissue. Commonly researchers have used Pearson's correlation coefficient, ρ, to determine the quality of spectral estimates – the mean value obtained in this case was 0.91 ± 0.13. Our attempts to numerically stabilize the solution were unsuccessful, i.e. this situation was found regardless of simulated annealing parameters and cooling/reannealing schedule (data not shown). Figure 2 (b) shows the spectral estimates obtained after the incorporation of the un-normalised integrated intensity term in the penalty function. The addition of this term stabilizes the algorithm and the fifty spectral estimates fall into two sets in each of which all the estimates are essentially identical. The mean correlation coefficient in this case was 0.9930±0.0023. However, obvious deficiencies in the minimal spectral estimate are plain, chiefly dips in the amide I (1640 cm^{-1}) and CH_2 wagging (1450 cm^{-1}) regions of the spectrum. Such features were also observed in Maher et al.[19] N.B. the spectra in the set with greater intensity fall into a secondary minimum which qualitatively appears to be a set of better estimates, but the penalty function values for these are an order of magnitude larger than for the set of estimates with obvious artefacts; they are a manifestation of the optimization routine getting trapped in a secondary minimum.

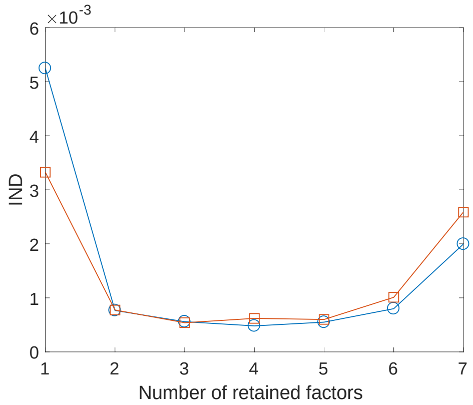

Figure 1. Variation of IND as a function of the number of retained eigenvectors z for SORS datasets obtained from two distinct anatomical sites on the front medial aspect of a human tibia. **[Color in Print]**

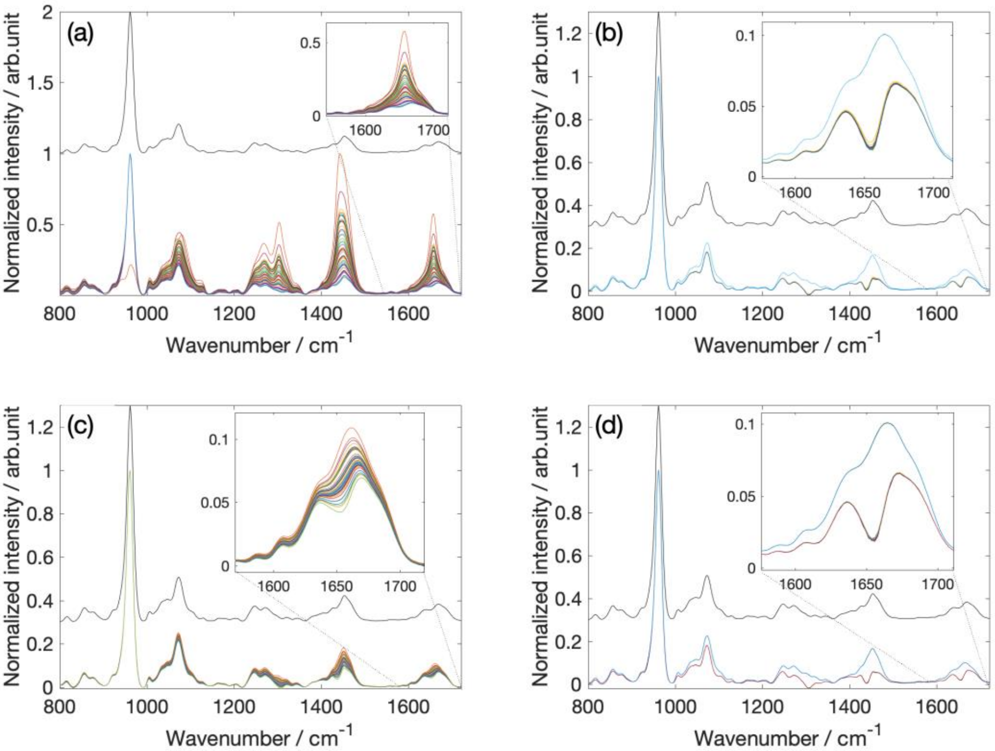

Figure 2. Plots showing 50 spectral estimates of bone obtained from synthetic *in-vivo* SORS datasets using different penalty functions as described in the original BTEM literature with simulated annealing optimization. (a) Shannon information entropy, (b) Shannon information entropy and integrated intensity, (c) sum of absolute second derivatives, (d) sum of absolute second derivatives and integrated intensity. In all figures, the offset black spectrum is the reference spectrum of bone used to generate the underlying synthetic SORS data. Inset figures magnify the amide I region. **[Color in Print]**

Spectral estimates obtained using the sum of absolute second derivatives in place of the Shannon entropy are shown in figure 2 (c). The use of this term improves the spectral estimates obtained from noiseless synthetic data with the instability and signal mixing with soft tissue being dramatically reduced but a visible range of varying estimates is still obtained (ρ = 0.9973±0.0024). However, in figure 2 (d), that highlights spectral estimates where the integrated intensity is included in the penalty function in addition to the sum of second derivatives we can see the same artefactual behaviour as for the estimates obtained with the Shannon entropy and integrated intensity terms shown in figure 2 (b) (ρ = 0.9926±0.0020).

Maher et al. associated these flaws with the fundamental BTEM approach as applied to the deconvolution of transcutaneous SORS data of bone rather than identify them as a consequence of the numerical implementation and penalty function behaviour.[19] When considering the original BTEM penalty function the value of the standard terms, spectral/concentration non-negativity and targeted peak intensity, tend to zero in the limit of a reasonable solution and the only non-zero term in the limit of a global function minimum is the information entropy. However, when integrated intensity is used as further contributing term to “prevent spectral over-resolution” two non-zero terms will be present no matter what the spectral estimate is. Global optimization will always seek the lowest valued solution, this means that the two non-zero terms will be in direct competition with each other, the final global minimum found will minimize the term with the greater numerical weight at the expense of the lesser. The greater term is invariably the normalised integrated intensity as the

Shannon entropy (or sum of second derivatives) cost of the incorporation of an artefactual dip (the indentation to the Raman peak) is far less than the associated integrated intensity gain in its absence. We conclude therefore, that use of the basic information entropy term alone is insufficient to afford a stable separation of bone from soft tissue related signals. Inclusion of the integrated intensity term prevents the mixing of unwanted signals but creates artefacts and distorts the resulting minimal spectral estimate. Furthermore, in the best case of the sum of second derivatives with noiseless data inherent instability in the simulated annealing method leads to an unacceptable variation in spectral estimates over many runs. It will be seen in a later section that these problems are exacerbated by experimental error (noise) as this has a marked effect on the sum of second derivatives and that even in the best case shown above the presence of normal noise levels will lead to signal mixing.

3.3 Adaptive BTEM algorithm

From the problems identified in the previous subsection two issues need to be addressed, the first is the stability of the optimization step and associated spectral estimate reproducibility, the second is the systematic errors or artefacts (spectral distortions) introduced by using additional penalty terms that aim to stabilize solutions but end up reducing reconstruction accuracy in the global minimum.

In the first instance, we make use of the metaheuristic optimization method called 'Global Optimum Determination by Linking and Interchanging Kindred Evaluators' (GODLIKE). which acts to reduce the likelihood of the overarching method returning a secondary minimum, or non-optimal solution, thereby increasing robustness and final accuracy of the spectral estimates derived.[30]

For the second problem of spectral artefacts there are two potential solutions, one is to use multi-objective optimization, the second is to initially normalize the integrated intensity term to the number of data channels and then to use it adaptively, disabling the non-zero integrated intensity term when the objective function is approaching a minimum, thus allowing the Shannon information entropy or sum of higher order derivatives to determine the final estimate but simultaneously providing a numerical bias to eliminate solutions where there is obvious mixing of soft tissue signal with that of bone. The former case is a general solution to the problem of soft constrained penalty functions with multiple potentially competing terms. In multi-objective optimization, each term is treated as an independent objective, but no term is biased. One downside of this is a marked reduction in the speed of optimization. Multi-objective optimization is significantly slower than the standard GODLIKE method or single heuristic methods such as adaptive simulated annealing. In the interest of time efficiency, we have focussed on developing an adaptive penalty function. Figure 3 (a) and (b) show the results obtained when using the standard GODLIKE method and GODLIKE in combination with the adaptive penalty function. It is clear from the data that the fifty spectral estimates in each case match the underlying reference spectrum extremely well ((a) $\rho > 0.9999 \pm 3.2877\times10^{-5}$, (b) $\rho > 0.9994 \pm 8.6310\times10^{-5}$). It should be noted that in this case there is only a slight difference in the highly stable and reproducible spectral estimates obtained whether the additional adaptive integrated intensity term is being used. However, we shall see in the following subsection on the influence of experimental error on our novel BTEM implementation that with the noisier data characteristic of real measurements obtained from patient volunteers the use of the adaptive term becomes necessary. Similar high-quality spectral estimates obtained from experimental SORS data measured from a cadaveric human leg and reconstructed using A-BTEM are shown in figure 4.

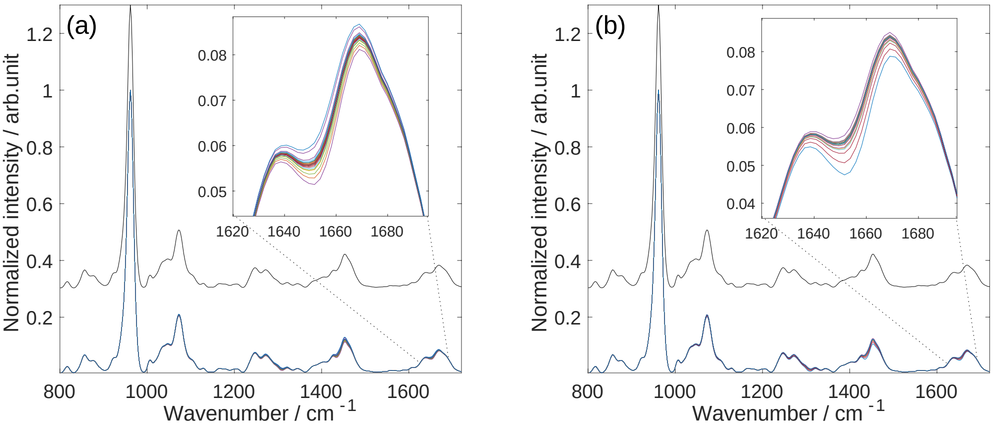

Figure 3. Plots showing 50 spectral estimates of bone obtained from an extreme quality synthetic *in vivo* SORS dataset using different penalty functions as described in the original BTEM literature with GODLIKE optimization. (a) sum of absolute second derivatives, (b) novel adaptive penalty function (A-BTEM). In both figures, the offset black spectrum is the reference spectrum of bone used to generate the underlying synthetic SORS data. The inset highlights part of the amide I region. **[Color in Print]**

These estimates were made using the advanced optimization routine (GODLIKE) and the adaptive penalty function. The spectral estimates correlate extremely well with most of the reference spectrum obtained from the exposed bone surface. However, we note an important divergence above 1550 cm^{-1} whereby the intensity of the amide I band is reduced in the spectral estimate when compared to the reference spectrum (dashed black line in figure 4). We ascribe this departure to tissue absorption being particularly strong at the amide-I band position. As the photons of Raman scattered light which originate from inside the bone volume need to pass through the overlaying soft tissue to reach the detector, photons of different energy are absorbed to different extents. A recent work by Stone and co-workers has explored the effects of laser excitation wavelength on signal recovery as related to transmission Raman spectroscopy focussing particularly on the effects of tissue absorption.[32] It is clear from the work of Stone that the absorption coefficient of soft tissue dramatically increases at the amide-I band position at our excitation wavelength. The measured Raman signal contributing to our SORS signals is different from that measured from exposed bone, which is consistent with our observations. As the thickness of the soft tissue is likely to vary from patient to patient any calibration models built that include the amide I region must be well controlled through appropriate balancing of the subjects' soft tissue thicknesses between the training set classes, or this region will have to be removed for that type of analysis. This is not ideal in terms of losing spectral information and a preferred alternative would be to alter our excitation wavelength to mitigate the effects of the unwanted absorption.[32] Despite this, the 1660/1690 cm^{-1} ratio, which is indicative of bone matrix crosslinking maturity,[26] remains constant to a first approximation when comparing the exposed bone surface measurement to the BTEM spectral estimate regardless of the overall attenuation in this region, which in its nature is spectrally much broader than the separation between the two overlapped Raman bands.

3.4 Influence of noise on A-BTEM spectral estimates after factor compression

A key issue in the application of any BTEM implementation is the role of experimental error or noise. To explore and explain the effects of noise on our new algorithm we have prepared synthetic and real experimental SORS datasets with different SNRs but maintaining constant overall bone signal contributions and number of spectra per dataset. Figure 5 plots (a) - (f) show spectral estimates obtained when using either: (i) the sum of second derivatives and no integrated intensity term – blue plots or (ii) the sum of second derivatives and our new adaptive integrated intensity term – red plots. The dashed black plot is the reference bone spectrum used to produce the underlying synthetic SORS datasets. Similar real experimental data are shown in figure 6 plots (a) – (d) where the same colour scheme applies except that the reference bone spectrum has been obtained directly from the bone surface corresponding to the area sampled during SORS measurements. For both figures by looking at the blue spectra with decreasing signal-to-noise ratio we immediately notice that as the noise levels increase the spectral estimates worsen and there is significant mixing of lipid signal observed at 1299 cm^{-1} with that of bone. This phenomenon occurs because the sum of absolute second derivatives becomes larger for the 'real' solution than that for the mixed case when noise levels increase; the BTEM algorithm is effectively seeing the mixed situation as a simpler solution. This is a key challenge with BTEM in so-far-as the algorithm has no intrinsic way to discriminate between noise 'information' and physically meaningful information.

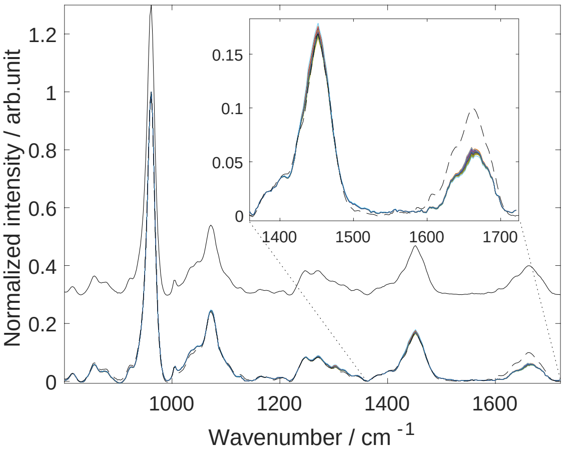

Figure 4. Plot showing 50 spectral estimates of bone obtained from high quality experimental *in vivo* SORS data. Spectral estimates were calculated using the adaptive penalty function and 'GODLIKE' optimization after determining the required number of basis vectors using Malinowski's IND function. The inset highlights the CH_2 wagging and amide I regions. The dashed black curve is the reference exposed bone measurement. **[Color in Print]**

The red plots show the situation for the same starting synthetic and real SORS datasets after the inclusion of the adaptive integrated intensity term (A-BTEM). It is immediately apparent that the use of this penalty term allows the implementation to penalise the mixed cases heavily but not to the extreme as seen earlier where spectral artefacts were introduced – this affords reliable spectral estimates at higher noise thresholds although it does not mitigate the problem entirely. When the signal-to-noise ratio reaches a threshold of around 20-25 signal, as shown in the synthetic examples, mixing becomes evident despite the new term. BTEM performs best with low noise data which is at odds with its application to signal limited techniques such as Raman spectroscopy. Some degree of de-noising/smoothing during data pre-processing would benefit most applications of the BTEM technique to Raman spectra, unfortunately this will always be at the cost of losing real physical information about the system under investigation.

3.5 The role of bone signal contribution on A-BTEM

Figure 7 shows how the correlation coefficients of different spectral estimates obtained from synthetic SORS datasets containing 17 spectra vary as a function of bone signal contribution (expressed as percentage of the most intense lipid Raman band present) and signal-to-noise ratio (correlation coefficient 1 represents a perfect reconstruction).

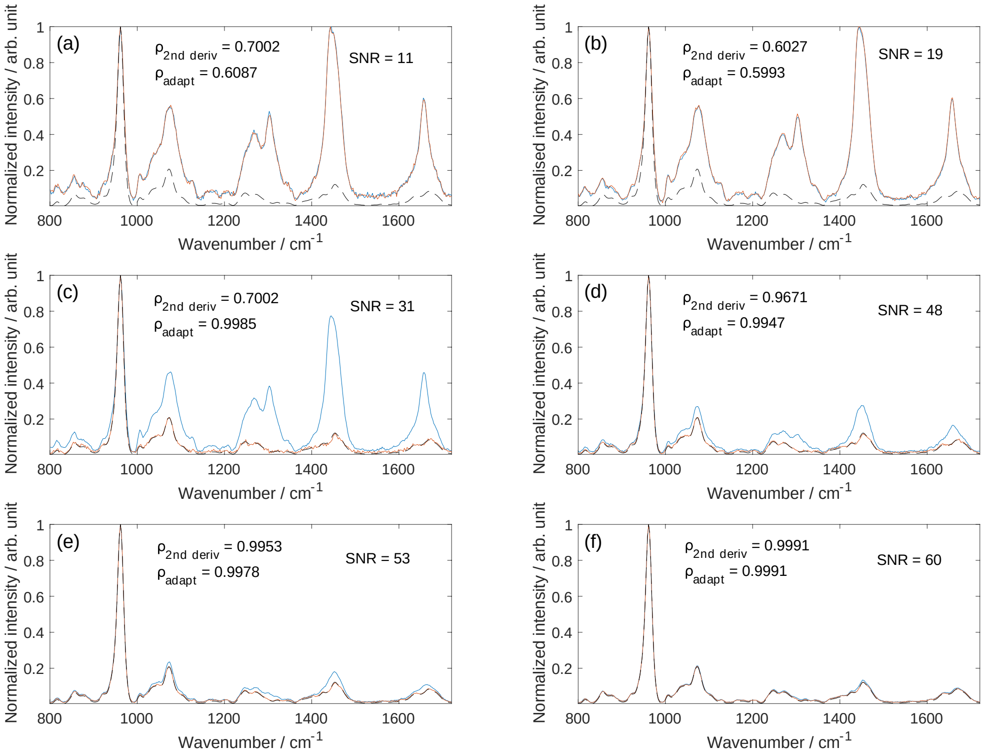

Figure 5. Plots (a) – (f) showing spectral estimates of bone obtained from synthetic *in vivo* SORS datasets with varying SNR using selected penalty functions after the addition of different amounts of Gaussian white noise. Blue plots – spectral estimates obtained using the sum of second derivatives only. Red plots – spectral estimates obtained using the sum of second derivatives and the additional normalized adaptive integrated intensity. Black dashed plots – reference spectrum of underlying bone signal. Correlation coefficients shown for comparison. **[Color in Print]**

In these data, we can see a relationship between signal-to-noise ratio and bone signal contribution. The greater the bone contribution the lower the required SNR for a highly accurate reconstruction. Most importantly when the SNR of the data is greater than ~55 within the range of bone signal contributions investigated the bone signal contribution becomes unimportant and high-quality A-BTEM reconstructions are readily obtainable. Therefore, to obtain the best possible spectral estimates from transcutaneous SORS spectra of human bone, one should pay attention to not just the absolute strength of Raman bone signal but also ensure the SNR is sufficiently high, our results indicate that a SNR of at least 55 is appropriate. For other systems, similar analyses using A-BTEM could be carried out with synthetic data to discover the optimum data necessary for the most accurate spectral reconstructions. When considering the effects of signal contribution and noise on the reconstructed spectrum, it is necessary to understand the relationship between the contributions of noise to a given solution in terms of the different penalty terms in comparison to that of likely signal contaminants. The algorithm treats experimental noise no differently from physically meaningful information, in excess it forces the algorithm to include unwanted signals leading to erroneous features and false estimated spectra.

3.6 The Effect of the Number of Spectra on A-BTEM

Figure 8 shows how the correlation coefficients of bone spectral estimates obtained from synthetic data vary as a function of the number of spectra in each SORS dataset and mean SNR for a fixed relatively low bone contribution of 16.7%. Interestingly, for all but the highest SNR value shown (56.1) increasing the number of spectra per SORS dataset improves the quality of the spectral estimate.

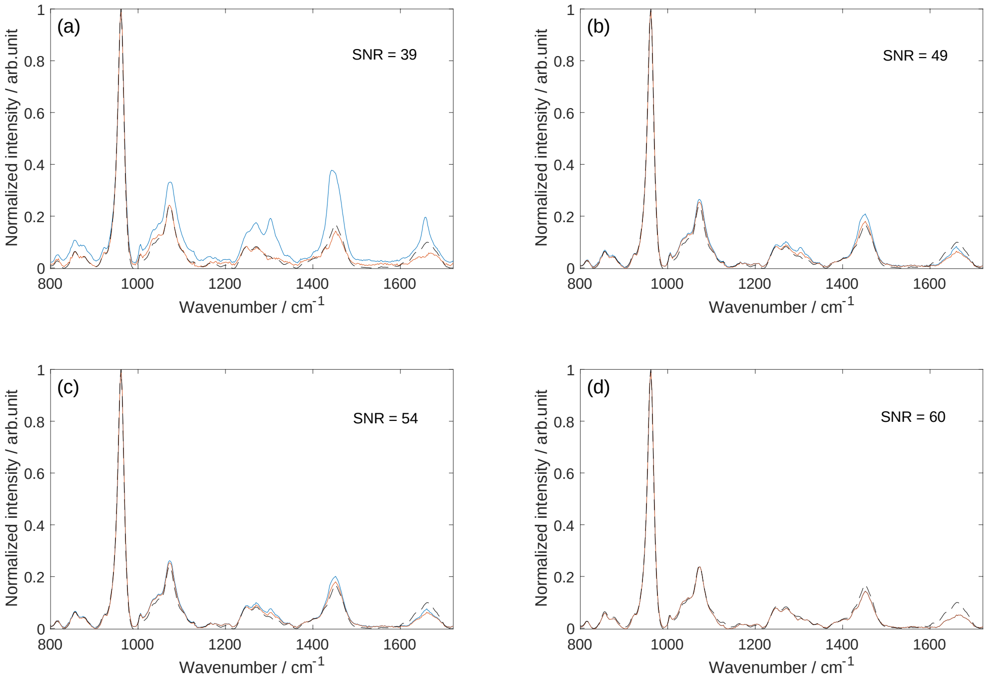

Figure 6. Plots (a) - (d) showing spectral estimates of bone obtained from real experimental *in vivo* SORS datasets with varying SNR using selected penalty functions. Blue plots – spectral estimates obtained using the sum of second derivatives only. Red plots – spectral estimates obtained using the sum of second derivatives and the additional normalized adaptive integrated intensity. Black dashed plots – actual averaged reference spectrum of underlying cortical bone. **[Color in Print]**

However, to achieve the necessary level of accuracy for our diagnostic purposes relating to determining the presence of bone disease we must focus on obtaining datasets of the required SNR, i.e. ~55. For other applications where such high stringency is not required increasing the number of spectra could be used in a compromise situation to further enhance the spectral estimates obtained, potentially, to a satisfactory or compromise level at median SNRs for lower/moderate target species signal contributions. In these situations, there will be a trade-off compromise between the integration time invested in obtaining higher SNRs and the time required to obtain a greater number of spectra.

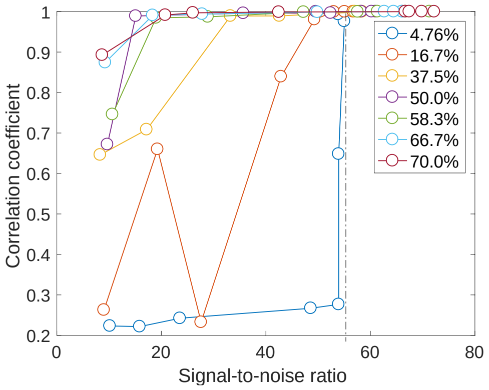

Figure 7. Plots showing how correlation coefficient for deconvoluted bone spectra obtained from synthetic SORS datasets containing 16 spectra varies as a function of signal-to-noise ratio and mean bone signal contribution (given in figure legend). **[Color in Print]**

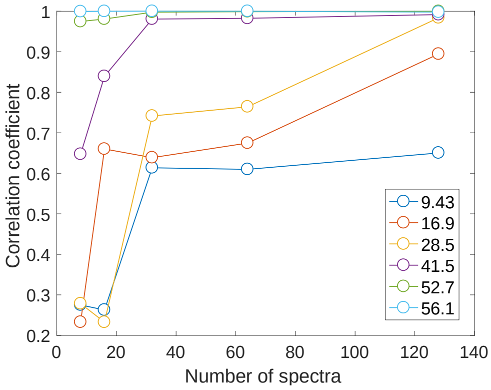

Figure 8. Plots showing how the correlation coefficient for deconvoluted bone spectra obtained from synthetic SORS datasets varies as a function of the number of spectra in each dataset and the mean SNR of each dataset (given in figure legend). **[Color in Print]**

4. Conclusions

We have described and implemented a novel variant of BTEM, Adaptive BTEM (A-BTEM) that provides excellent stability, accuracy and reproducibility when decomposing challenging SORS datasets, not available from the original BTEM algorithms. We have shown that: by using an appropriate empirical test to choose the number of retained basis vectors, the use of an adaptive penalty function and the incorporation of extremely robust metaheuristic optimisation strategies - accurate extremely stable spectral estimates of bone can be acquired. We have also demonstrated that BTEM is sensitive to noise and that accurate spectral estimates can only be obtained with the new implementation (A-BTEM) when the mean signal-to-noise ratio of SORS dataset is sufficiently high (>55) for a given underlying bone signal contribution. In our case, spectral estimates showing a correlation coefficient of > 0.996 are readily available in these instances. The algorithm has been successfully applied to transcutaneous bone SORS spectra although its applicability is general and it can be used for any other system of interest where de-mixing has proved challenging provided some knowledge of the target species is available or can be estimated. We also note that tissue absorption is a fundamental issue in the *in vivo* application of SORS to obtain accurate spectral estimates of human bone. Overall this work provides a much-needed pathway to developing an optimized SORS instrument for the diagnosis of human bone disease *in vivo* by developing a novel BTEM implementation that affords highly accurate and stable spectral estimates as well as highlighting the quality of real data required to be obtained within a brief integration time typical of patient visits simultaneously meeting the requirements of unsupervised use and the reliable highly accurate recovery of clinical data.

Conflicts of Interest

Declarations of interest: none

Acknowledgements

The authors would like to thank the Vesalius Centre at the University of Bristol for providing the cadaveric tissue samples. This work was carried out with the appropriate human tissue authority license (HTA12135) and ethical approval (London-Stanmore REC - 08/H072/34).

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EP/K503745/1).

References

[1] W. Chew, E. Widjaja, M. Garland, *Organometallics* **2002**, *21*, 1982.

[2] E. Widjaja, C. Li, W. Chew, M. Garland, *Anal. Chem.* **2003**, *75*, 4499.

[3] E. Widjaja, N. Crane, T.-C. Chen, M. D. Morris, M. A. Ignelzi, Jr., B. R. McCreadie, *Appl. Spectrosc.* **2003**, *57*, 1353.

[4] L. Guo, P. Sprenger, M. Garland, *Anal. Chim. Acta* **2008**, *608*, 48.

[5] P. Matousek, I. P. Clark, E. R. C. Draper, M. D. Morris, A. E. Goodship, N. Everall, M. Towrie, W. F. Finney, A. W. Parker, *Appl. Spectrosc.* **2005**, *59*, 393.

[6] P. Matousek, *Appl. Spectrosc.* **2006**, *60*, 1341.

[7] K. Buckley, P. Matousek, *Analyst* **2011**, *136*, 3039.

[8] P. Matousek, N. Stone, *Chem. Soc. Rev.* **2016**, *45*, 1794.

[9] J. G. Kerns, P. D. Gikas, K. Buckley, A. Shepperd, H. L. Birch, I. McCarthy, J. Miles, T. W. R. Briggs, R. Keen, A. W. Parker, P. Matousek, A. E. Goodship, *Arthritis Rheumatol.* **2014**, *66*, 1237.

[10] K. Buckley, J. G. Kerns, J. Vinton, P. D. Gikas, C. Smith, A. W. Parker, P. Matousek, A. E. Goodship, *J. Raman Spectrosc.* **2015**, *46*, 610.

[11] J. R. Maher, J. Inzana, M. Takahata, H. A. Awad, A. J. Berger, in *Proc. SPIE*, (Eds: A. Mahadevan-Jansen, W. Petrich), International Society for Optics and Photonics, **2012**, p. 82190P.

[12] K. Buckley, J. G. Kerns, P. D. Gikas, H. L. Birch, J. Vinton, R. Keen, A. W. Parker, P. Matousek, A. E. Goodship, *IBMS Bonekey* **2014**, *11*, doi:10.1038/bonekey.2014.97.

[13] K. Sowoidnich, J. H. Churchwell, K. Buckley, A. E. Goodship, A. W. Parker, P. Matousek, *J. Raman Spectrosc.* **2016**, *47*, 240.

[14] K. Sowoidnich, J. H. Churchwell, K. Buckley, A. E. Goodship, A. W. Parker, P. Matousek, *Analyst* **2017**, *142*, 3219.

[15] S. R. Cummings, D. Bates, D. M. Black, *JAMA* **2002**, *288*, 1889.

[16] H. Ding, G. Lu, C. West, G. Gogola, J. Kellam, C. Ambrose, in *SPIE Proceedings*, **2016**, vol. 9689, p. 96894M.

[17] R. Bro, *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149.

[18] K. Buckley, J. G. Kerns, A. W. Parker, A. E. Goodship, P. Matousek, *J. Raman Spectrosc.* **2014**, *45*, 188.

[19] J. R. Maher, J. A. Inzana, H. A. Awad, A. J. Berger, *J. Biomed. Opt.* **2013**, *18*, 077001.

[20] K. Sasaki, S. Kawata, S. Minami, *Appl. Opt.* **1984**, *23*, 1955.

[21] S. Y. Sin, E. Widjaja, L. E. Yu, M. Garland, *J. Raman Spectrosc.* **2003**, *34*, 795.
[22] S.-T. Tan, H. Zhu, W. Chew, *Anal. Chim. Acta* **2009**, *639*, 29.
[23] H. Zhang, M. Garland, *Appl. Spectrosc.* **2007**, *61*, 1366.
[24] C. G. Bertinetto, A. de Juan, *J. Chemom.* **2018**, *32*, 1.
[25] L. R. Ong, E. Widjaja, R. Stanforth, M. Garland, *J. Raman Spectrosc.* **2003**, *5*, 282.
[26] G. S. Mandair, M. D. Morris, *Bonekey Rep.* **2015**, *4*, doi:10.1038/bonekey.2014.115.
[27] C. K. Chua, Y. Lv, H. J. Zhang, X. Y. Gu, *Anal. Methods* **2017**, *9*, 2667.
[28] E. R. Malinowski, *Factor Analysis in Chemistry*, Wiley, 3rd Ed., **2002**.
[29] C. E. Shannon, *Bell Syst. Tech. J.*, **1948**, 27, 379–423.
[30] R. P. S. Oldenhuis, Trajectory Optimization for a Mission to the Solar Bow Shock and Minor Planets, Delft University of Technology, Netherlands, **2010**.
[31] M. Koch, C. Suhr, B. Roth, M. Meinhardt-Wollweber, *J. Raman Spectrosc.* **2016**, *48*, 336.
[32] A. Ghita, P. Matousek, N. Stone, *Analyst* **2016**, *141*, 5738.