SAND2016-8518J

Removing Cosmic Spikes Using a Hyperspectral Upper-Bound Spectrum Method

Stephen M. Anthony and Jerilyn A. Timlin

Abstract

Cosmic ray spikes are especially problematic for hyperspectral imaging, due to the large number of spikes often present and their negative effects upon subsequent chemometric analysis. Fortunately, while the large number of spectra acquired in a hyperspectral imaging data set increases the probability and number of cosmic spikes observed, the multitude of spectra can also aid in the effective recognition and removal of the cosmic spikes. Dongmao Zhang and Dor Ben-Amotz were perhaps the first to leverage the additional spatial dimension of hyperspectral data matrices (DM). They integrated principal component analysis (PCA) into the upper bound spectrum method (UBS), resulting in a hybrid method (UBS-DM) for hyperspectral images. Here, we expand upon their use of PCA, recognizing that principal components primarily present in only a few pixels most likely correspond to cosmic spikes. Eliminating the contribution of those principal components in those pixels improves the cosmic spike removal. Both simulated

and experimental hyperspectral Raman image data sets are used to test the newly developed UBS-DM-hyperspectral (UBS-DM-HS) method which extends the UBS-DM method by leveraging characteristics of hyperspectral datasets. A comparison is provided between the performance of the UBS-DM-HS method and other methods suitable for despiking hyperspectral images, evaluating both their ability to remove cosmic ray spikes and the extent to which they introduce spectral bias.

INTRODUCTION

Cosmic ray events occur when high energy (gigaelectronvolt) particles impact the earth's atmosphere, creating an avalanche of secondary particles. If the corresponding shower of ionizing radiation intersects a spectroscopic detection element, a large positive cosmic ray spike will be recorded at those pixels. Both charge-coupled devices (CCDs) and complementary metal-oxide semiconductor (CMOS) sensors are vulnerable to cosmic rays.[1-3] While the frequency of cosmic rays depends upon the altitude, latitude, and cross-section of the pixel,[1] for typical spectroscopic detectors the rate of cosmic ray spikes is often reported to be roughly on the order of 10^{-7} s^{-1}/pixel.[4, 5] Shielding against cosmic rays is not practical, as totally protecting detectors requires 20 meters of concrete.[1]

Cosmic ray spikes vary substantially in intensity and morphology,[6] such that many cosmic ray spikes affect multiple adjacent pixels on the detector.[7-10] Depending upon the angle of incidence of the cosmic ray relative to the detector, the cosmic ray may affect adjacent spectra which are acquired simultaneously[7, 9] or cause broader cosmic ray spikes spanning several pixels.[7, 8] Experimental Raman spectra are known to sometimes exhibit cosmic ray spikes with a spectral bandwidth of more than 10 pixels.[7] Cosmic rays also contaminate astronomical images, where the resulting cosmic ray objects can be effectively simulated as Gaussian profiles with full width at half maximum (FWHM) ranging from 0 to 10 pixels.[11] A study using radioactive gamma ray sources to generate simulated cosmic rays found that roughly 30-50% of cosmic rays caused spikes in multiple pixels and approximately 7-12% of cosmic rays caused spikes in 10 or more pixels, regardless of the angle of the detector relative to the source beam.[3] Additionally, software supplied with some spectrometers applies an apodization which can increase the bandwidth of the spikes to 30 pixels,[12] or blooming can occur when a cosmic ray spike saturates the detector pixel. Finally, while rare, two cosmic ray spikes can occur adjacent to each other, effectively creating one broader spike.

Cosmic ray spikes are particularly problematic for hyperspectral imaging. In comparison to traditional spectroscopy, hyperspectral imaging generally requires the

acquisition of a large number of spectra of low signal-to-noise samples and acquisition times can vary from milliseconds to minutes. Both the large number of samples and the longer acquisition times provide more opportunities for cosmic ray events. Additionally, hyperspectral imaging is frequently followed by chemometric analysis of the data, attempting to extract the component spectra and concentrations from the sample.[13] Unfortunately, the presence of cosmic ray spikes can result in very low cross-correlation coefficients between otherwise similar spectra and skew the eigenvectors of the spectral matrix, complicating chemometric analysis.[10] Factor analytic chemometric approaches, including principal component analysis and self-modeling curve resolution are particularly challenged by the presence of cosmic ray spikes.[14]

While both hyperspectral Raman and hyperspectral fluorescence imaging are prone to interference from cosmic rays, cosmic ray spikes are harder to remove from Raman spectra than fluorescence spectra. Cosmic ray spike removal (despiking) algorithms applied to individual spectra are generally limited in operation to cases where the cosmic ray spikes are narrower in bandwidth than the spectral peaks.[15, 16] Cosmic ray spikes typically exhibit quite narrow bandwidths,[1, 2] where the peaks in the fluorescence spectra are much broader than the cosmic ray spikes. In such cases, even simple signal processing can distinguish the fluorescence signals from the cosmic ray spikes. In contrast, Raman

spectral peaks can be quite sharp, and depending upon the resolution of the Raman instrument, cosmic ray spikes and Raman signals often have comparable bandwidths.

Over the years, a variety of despiking algorithms have been developed, which can generally be broken down into four categories: single-spectrum, multiple-acquisition, spatiotemporal-oversampling, and hyperspectral data matrix methods. In addition, a related class of algorithms exist which simply determine the presence or absence of cosmic ray spikes in each spectrum, which we will categorize as detection-only methods. The various categories of algorithms arose, in part, due to the different experimental constraints imposed. Here we are interested in cosmic ray spike removal from hyperspectral Raman and fluorescence images of living cells and tissues, limiting which categories and despiking algorithms can be applied.

Detection-only approaches[17, 18] do not attempt to determine which portions of the spectrum are contaminated by cosmic ray spikes. Instead, these algorithms simply attempt to determine whether a given spectrum is contaminated by cosmic rays and spectra determined to be contaminated can then either be discarded or replaced by adjacent spectra. Detecting and eliminating suspect spectra makes sense when a small fraction of the spectra can be discarded without affecting the analysis. For example, if the desire is simply to determine the underlying spectral components within the sample, eliminating

0.01% of the spectra is acceptable as long as the remaining spectra include the full range of spectral components.[19] However, often discarding spectra is impractical. In particular, performing Raman imaging on samples requiring longer integration times and/or over a larger detector area results in a larger fraction of spectra contaminated with cosmic ray spikes, such that sometimes >1% of spectra are contaminated (see the data presented later in Figure 6a). Additionally, when imaging sub-cellular components, images often contain a high degree of heterogeneity and the features of interest may be present in only a small fraction of the pixels. If some of those pixels are discarded, the population statistics can suffer substantial distortion.

The first category of cosmic ray despiking algorithms corresponds to single-spectrum algorithms. These algorithms perform signal processing on individual spectra, attempting to distinguish spikes from legitimate signals. Single spectrum algorithms include median filtering - detecting outliers compared to a smoothed spectrum[15, 20] - and wavelet techniques.[16, 21] Lacking additional spectra from which the shape of the spectral components can be inferred, a general limitation of these algorithms is that the cosmic ray spikes must be narrower than the peaks in the signal.[15, 16] Since Raman spectra often exhibit sharp peaks with bandwidth comparable to that of cosmic ray spikes, the corrected spectra after single-spectrum despiking can show undesirable spectral distortion including

truncation of legitimate peaks misidentified as spikes.[16] Median filtering is particularly prone to spectral distortion. Despite this limitation, single-spectrum despiking algorithms remain relevant as sometimes only individual spectra are available. For this reason, two single-spectrum despiking algorithms are included in our supplementary comparisons: median-filtering and the algorithm by Katsumoto and Ozaki.[15] Median-filtering is included for comparison because it is well known and easily implemented. The Katsumoto-Ozaki algorithm was selected because while it is an advanced single-spectrum despiking algorithm; it is easy to implement and therefore attractive for many applications.

In contrast, multiple-acquisition despiking algorithms explicitly make use of multiple spectra, exploiting the randomness of cosmic rays.[4, 10, 22, 23] If multiple but-otherwise identical spectra are acquired, the probability of a cosmic ray spike occurring at the same position in all spectra is presumed to be negligible if the cosmic rays are randomly distributed. While cosmic rays are randomly distributed overall, the probability of cosmic rays appearing in adjacent spectra which are taken simultaneously, such as in line-scanning experiments, can be quite high.[3, 7] However, the probability of all spectra being contaminated at a given point can be reduced to negligible by increasing the number of spectra considered and/or requiring the use of spectra acquired at different times.

The necessary spectra can be acquired in two ways, either by performing multiple sequential acquisitions [4, 10] or by using a multi-channel detector.[22, 23] While multiple-acquisition despiking algorithms often exhibit excellent performance, both approaches have drawbacks and are not always practical, especially for hyperspectral imaging of live cells. First, when working with dynamic samples, sequential acquisitions may not be spectrally identical simply due to the dynamic nature of the material in the acquisition volume of the objective. Additionally, acquiring multiple spectra at each spatial location increases the acquisition time required and degrades the maximum temporal resolution. Meanwhile, spreading the signal out across a multi-channel detector can require substantial changes to the optics - particularly for line-scan hyperspectral imaging systems which already use both camera dimensions. Finally, the samples of interest often have limited signal-to-noise and may be prone to photo-bleaching and/or photodegradation, limiting the total amount of signal which can be obtained. When dealing with limited signal, distributing signal out over multiple acquisitions results in a lower overall signal-to-noise ratio. Therefore, while multiple-acquisition despiking algorithms are frequently employed in commercial spectrometer systems, they are not necessarily practical for live-cell hyperspectral imaging. Due to this impracticality, no multiple-acquisition despiking algorithms were evaluated.

The third category of despiking algorithms, spatiotemporal-oversampling, relaxes one of the constraints of the multiple-acquisition algorithms. Instead of requiring the multiple spectra to be virtually identical except for the presence of cosmic rays, spectra which are similar are sufficient, where the required degree of similarity varies between algorithms. Included within this category are nearest neighbor methods, where the sample is presumed to be spatially oversampled.[5, 7, 9, 24] In such cases, for every spectrum contaminated by a cosmic ray spike there is presumed to exist an uncontaminated, highly-correlated adjacent spectrum. Other algorithms have been developed with online or process monitoring in mind, in which case the process is presumed to be temporally oversampled.[8, 12, 25, 26] While at first spatial and temporal oversampling may appear quite different, from an algorithmic standpoint the two are quite similar. When considering hyperspectral imaging of living cells, spatial or temporal oversampling is rarely practical or desirable. As with the multiple acquisition methods, oversampling increases the total acquisition time required while reducing the signal-to-noise ratio.

The algorithm by Behrend et al.[5] was explicitly designed for analyzing hyperspectral data, as were several algorithms[7, 9, 24] similar to it. Among these, the algorithm by Zhang and Henson[7] appeared the most suitable and thus we evaluated it in our supplementary analysis. The other algorithms in this category appear to require a

higher degree of similarity between adjacent spectra, a constraint not justified by live cell imaging data.

The final category of despiking algorithms, hyperspectral data matrix methods, leverage correlations between the large number of spectra available from hyperspectral imaging.[13, 27] In contrast to the prior multi-spectral algorithms, the requirement that any two spectra be similar is completely eliminated. Instead, the hyperspectral data matrix methods merely require that the same underlying spectra be present throughout the image. As a result, these methods are ideally suited for despiking of hyperspectral images, minimizing the imaging constraints and the need for excess acquisitions. The greatest factor limiting the application of hyperspectral data matrix methods is their need for large numbers of correlated spectra, where thousands of spectra are preferred. At present, this condition is readily satisfied for hyperspectral fluorescence live cell imaging,[28, 29] and is becoming increasingly easy to satisfy for Raman due to advances in detection speeds and fast, line-scan Raman systems.[30]

The most notable hyperspectral data matrix method is the UBS-DM method,[13] which suppresses most cosmic ray spikes, particularly when multiple iterations are employed. While UBS-DM virtually eliminates narrow-bandwidth cosmic ray spikes, cosmic ray spikes which spread across multiple pixels in a spectrum are often only

partially suppressed, reducing their intensity but leaving spikes still significantly greater than the spectral noise. This limitation motivated our work extending the UBS-DM algorithm as substantial multiple-pixel spikes persisted in our experimental data even after applying UBS-DM. Very recently, another hyperspectral data matrix method, KPCARD, was developed for pharmaceutical process control.[27] The authors indicate that KPCARD requires that when PCA is applied to the data matrix, the resulting principal components separate cleanly into spectral and noise components. As this condition is rarely satisfied for the biological samples we are interested in, this algorithm was not selected for evaluation. In this paper, we develop a new hyperspectral data matrix method, UBS-DM-HS, which expands upon UBS-DM and addresses these limitations.

Source code for all the tested algorithms, including the newly developed UBS-DM-HS algorithm, is included in the supplementary material.

THEORY

Overview of the Algorithm. The UBS-DM-HS algorithm is largely identical to the UBS-DM algorithm[13] with one important change, the addition of a step in the algorithm procedure. For a more detailed rationale explaining the original algorithm, please consult the UBS-DM paper.[13] In brief, the underlying concept of both algorithms is that PCA

deconstructs the hyperspectral data matrix, **D**, into an orthonormal eigenvector matrix, **V**, and a set of scores, **S**. The entries in the score matrix **S** indicate how much each eigenvector in **V** contributes to a given spectrum in the hyperspectral data matrix **D**. The eigenvectors in **V** are ordered such that the first few eigenvectors represent most of the variance in **D**, where with ideal, noise-free data the set of eigenvectors with non-zero scores would be identical to the underlying set of spectra present in the sample. For experimental data, the presence of both noise and cosmic ray spikes increases the number of eigenvectors necessary to model the data well and may skew the eigenvectors relative to the ideal spectra.

Fortunately, the contributions of the spectral signal, cosmic ray spikes, and random noise vary in accordance with the relative eigenvalue weights. As a result, the eigenvectors can generally be divided into three groups, although the separation is imperfect. Generally, the largest eigenvalues correspond to legitimate spectral components (group I). While cosmic ray spikes may be the most intense features in an individual spectrum, the location of spikes is not correlated across the spectra whereas the legitimate signal is. Therefore, the eigenvalues associated with cosmic ray spikes (group II) are generally weaker since eigenvalues represent the influence of an eigenvector in the image as a whole. Of course, many eigenvectors in group II will have a mix of contributions from cosmic ray spikes and

real spectral components. Finally, the third group of eigenvectors is unlikely to contain significant signal, consisting mainly of random noise and weaker cosmic ray spikes.

Once the eigenvectors are separated into the three groups, appropriate processing schemes can be applied to each group. To begin with, the eigenvectors in group III are simply discarded since these eigenvectors are dominated by noise and contain minimal legitimate spectral information. While the group II eigenvectors may contain some legitimate spectral information, sharp spectral features were most likely assigned to group I. Therefore, median filtering the eigenvectors in group II suppresses the cosmic ray spikes but is expected to have minimal effect upon the true signal. Meanwhile, the eigenvectors in group I receive no further processing in the UBS-DM method as altering these critical eigenvectors runs the risk of introducing spectral distortion. Finally, reconstructed spectra are generated using the remaining eigenvectors and scores. The original spectra are compared to the reconstructed spectra using the UBS method.[10]

By not altering the most critical spectral eigenvectors and employing the UBS method, the UBS-DM algorithm minimizes the introduction of spectral distortion.[13] The UBS-DM algorithm also does an excellent job of suppressing stereotypical cosmic ray spikes where a sharp spike occurs at a single spectral pixel.

Unfortunately, the UBS-DM method can be less effective at filtering broader cosmic ray spikes which span multiple pixels. Generally, eigenvectors dominated by cosmic ray spikes (spike eigenvectors) will not be assigned to group I. If a spike eigenvector were assigned to group I, no further processing is performed on the eigenvector so that cosmic ray spike would not be suppressed. In most cases, spike eigenvectors either have lower eigenvalues (group IIb) or the majority of their contribution to the hyperspectral image is in a few spectra (group IIa). However, the median filtering applied to the eigenvectors in group II is only effective at eliminating cosmic ray spikes contributing substantially to fewer than 3 spectral pixels (group IIa – 5-pixel median filter) or 4 spectral pixels (group IIb – 7-pixel median filter). Therefore, if a spike eigenvector is assigned to group II, broad cosmic ray spikes will not be eliminated.

The UBS-DM-HS algorithm extends the UBS-DM algorithm, further employing correlations across the hyperspectral data matrix to detect and eliminate cosmic ray spikes, particularly broader ones. Eigenvectors corresponding to legitimate spectral components are expected to contribute substantially to numerous spectra throughout the image whereas spike eigenvectors generally contribute substantially to only a few pixels. Therefore, any eigenvectors which contribute significantly to only a few spectra are identified as spike eigenvectors, similar to the fashion that group IIa eigenvectors were differentiated from

group I eigenvectors. While intuitively it might make sense to simply discard such eigenvectors, in practice doing so results in undesirable spectral distortion. While the spike eigenvectors only contribute significantly to a few spectra, they frequently contribute minor amounts to many other spectra, helping to represent the low signal to noise components and noise in the spectra. If any individual spectrum is considered, eliminating the eigenvector will not be seen to significantly perturb the spectrum. However, when the image as a whole is considered, wholesale elimination of a spike eigenvector can result in a negative spectral distortion at that point which would affect subsequent chemometric analysis. Instead of discarding the spike eigenvectors, the UBS-DM-HS algorithm sets the contribution of spike eigenvectors to zero for the limited number of spectra where that eigenvector contributes substantially. In contrast to the UBS-DM method, this step does not rely on a spectral bandwidth dependent median filtering step. Instead, this step of the UBS-DM-HS algorithm depends solely upon correlations across the hyperspectral data matrix and as a result is capable of suppressing broad cosmic ray spikes that the UBS-DM method cannot.

Algorithm Procedure. The notation used is the same as in the UBS-DM paper[13] (except uppercase bold vectors have been converted to lowercase bold to conform to publication guidelines). The UBS-DM-HS algorithm can be broken down into 13 steps,

where the primary difference between the UBS-DM algorithm and the UBS-DM-HS algorithm is the addition of step 10. While the two algorithms are otherwise virtually equivalent, the enumeration of steps differs between the algorithms. For instance, a single step of the UBS-DM algorithm may be multiple discreet steps here for greater clarity. Source code for the UBS-DM and UBS-DM-HS algorithms is included in the supplementary material.

The UBS-DM algorithm requires four input variables, all shared by UBS-DM-HS. $\mathbf{D}^{\text{spike}}$ is the hyperspectral data to be processed and should be arranged as a data matrix $\mathbf{D}(m, n)$ of m spectra, each of which was measured at n wavelengths. For the first iteration, the original data matrix $\mathbf{D}^{\text{spike}}$ will also serve as the input data matrix $\mathbf{D}^{\text{in}}$. The readout noise level σ, is generally part of the specifications of the detector. Another input, the approximate number of spectral components, is generally known at least approximately. This input is used to determine an upper limit to the number of spectra which can be assigned to group I, where the limit is set to 10 times the expected number of true spectra. In practice, fewer spectra than this limit are typically employed since addition of spectra to group I also terminates when the included spectra model a sufficient fraction of the data. As a result, an estimated value is sufficient since altering this parameter generally has little or no effect on the algorithms. The final shared input, the optimum numbed of iterations,

may not be known *a priori*. However, the algorithms can simply be run until the results change negligibly between iterations, as was done in this paper. Here, after each iteration of the algorithms, the correlation coefficient is calculated between each input spectrum and its corresponding output spectrum. If the mean of these correlation coefficients is greater than .999999, no further iterations are done since virtually no changes are occurring between iterations.

The UBS-DM-HS algorithm requires one additional variable, a threshold t, relating to the maximum percentage that any single spectrum is allowed to contribute to a given eigenvalue. The threshold t can take on any value between 0 and 1, where t=1 effectively eliminates the threshold, rendering the UBS-DM-HS algorithm equivalent to UBS-DM. A reasonable value for t can be estimated based upon the expected distribution of the spectral components throughout the image. For each spectral component, determine the maximum amount of that component expected in one spectrum relative to the total amount of that component expected throughout the image. The threshold t should be greater than the value of this ratio for any component. In this paper, t was set as 10 divided by the number of spectra analyzed.

Each iteration of the UBS-DM-HS algorithm consists of 13 steps, where the UBS-DM algorithm involves either eliminating step 10 or setting t=1. The steps are:

(1) Set the maximum number of group I eigenvectors, *c*, to be 10 times the number of suspected components in the sample.

(2) Perform principal component analysis by applying singular value decomposition (SVD) to the data matrix $\mathbf{D}^{\text{in}}$.

$$\mathbf{D}^{\text{in}} = \mathbf{U\Sigma V}$$

(3) Compute the score matrix **S**.

$$\mathbf{S} = \mathbf{U\Sigma}$$

(4) Construct a vector of eigenvalues, **e**, by extracting the diagonal of the diagonal matrix **Σ** and squaring each element.

$$\mathbf{e}_i = \mathbf{\Sigma}_{i,i}^2$$

(5) Compute a vector, **f**, of cosmic spike probability factors.

$$\mathbf{f}_i = \frac{\max(\mathbf{S}_i^2)}{[\text{norm}(\mathbf{S}_i)]^2}$$

where $\mathbf{S}_i$ is the *i*th column vector in **S**, corresponding to the score vector of the *i*th eigenvector.[31]

(6) Evaluate the *i*th eigenvector, starting with *i*=1, to determine whether it should be assigned to group I, IIa, IIb, or III. Stop when sufficient eigenvectors have been assigned to group I (see Step 7d).

a. If $\mathbf{f}_i \leq 0.25$, then assign this eigenvector to group I. If not, assign this eigenvector to group IIa.

b. Set $\mathbf{e}_i = 0$ if the ith eigenvector was assigned to group IIa.

c. Determine the eigenvalue percentage, p, captured by the first i eigenvalues.

$$p = \frac{\sum_{j=1}^{i} \mathbf{e_j}}{\mathrm{sum}(\mathbf{e})}$$

d. If $p > 99.5\%$ or c eigenvectors have been assigned to group I, then groups I and IIa are complete. If not, evaluate the next eigenvector.

(7) Assign up to the next 40 eigenvectors to group IIb, assigning either 40 eigenvectors or all the remaining eigenvectors if fewer than 40 eigenvectors remain. Note: this is a minor deviation from the original UBS-DM algorithm,[13] which always assigned 40 eigenvectors.

(8) Discard the remaining eigenvectors (group III) by creating an abridged eigenvector matrix $\mathbf{V}^{\mathrm{I,II}}$ and an abridged score matrix $\mathbf{S}^{\mathrm{I,II}}$ which only contain entries corresponding to the eigenvectors in groups I and II.

(9) Apply a 7 point median filter to each eigenvector in group IIa and a 5 point median filter to each eigenvector in group IIb to generate the abridged, median-

filtered eigenvector matrix $\mathbf{V}^{I,II*}$. (The eigenvectors in group I are not median-filtered.)

(10) Determine whether any of the spectra in $\mathbf{D}^{in}$ are responsible for an excessive fraction of any of the eigenvectors in group I or II.

a. Construct positive and negative score matrices $\mathbf{S}^{+}$ and $\mathbf{S}^{-}$.

$$\mathbf{S}^{+} = \max(\mathbf{S}^{I,II}, 0)$$

$$\mathbf{S}^{-} = \min(\mathbf{S}^{I,II}, 0)$$

b. For each eigenvector in $\mathbf{V}^{I,II*}$, determine whether any given spectrum in $\mathbf{D}^{in}$ contributes an excessive fraction of either the total positive or negative contribution. If it does, set the contribution of the score for that combination of spectrum and eigenvector to be zero. (If $\mathbf{S}_{i,j}^{+} > t \times \mathrm{sum}(\mathbf{S}_{i}^{+})$ or $\mathbf{S}_{i,j}^{-} < t \times \mathrm{sum}(\mathbf{S}_{i}^{-})$, then set $\mathbf{S}_{i,j}^{I,II} = 0$.)

(11) Generate the reconstructed hyperspectral data matrix $\mathbf{D}^{R}$ using the group I and II eigenvectors and the corresponding scores.

$$\mathbf{D}^{R} = \mathbf{S}^{I,II} \times \mathbf{V}^{I,II*}$$

(12) Construct the upper bound spectrum data matrix $\mathbf{D}^{UBS}$, accounting for the noise variance due to the Poisson noise and the readout noise σ.

$$\mathbf{D}_{i,j}^{\mathrm{UBS}} = \mathbf{D}_{i,j}^{\mathrm{R}} + 4\sqrt{\mathbf{D}_{i,j}^{R} + \sigma^2}$$

(13) The decontaminated hyperspectral data matrix $\mathbf{D}^{\mathrm{DC}}$ is created, where $\mathbf{D}_{i,j}^{\mathrm{DC}} = \mathbf{D}_{i,j}^{spike}$ except where a cosmic ray spike is identified in the (*i,j*)th element. Any element $\mathbf{D}_{i,j}^{spike}$ of the original data matrix which is greater than the corresponding element $\mathbf{D}_{i,j}^{\mathrm{UBS}}$ in the upper bound spectrum data matrix is classified as a spike. Therefore, if $\mathbf{D}_{i,j}^{spike} > \mathbf{D}_{i,j}^{\mathrm{UBS}}$, then $\mathbf{D}_{i,j}^{\mathrm{DC}} = \min(\mathbf{D}_{i,j}^{\mathrm{R}}, \mathbf{D}_{i,j}^{spike})$.

The outlined procedure (steps 1-13) implements a single iteration of the UBS-DM-HS algorithm (or the UBS-DM algorithm if t=1). More complete elimination of the cosmic ray spikes can be achieved by applying multiple iterations.[13] In order to perform multiple iterations, all steps of the algorithm are repeated using the decontaminated hyperspectral data matrix $\mathbf{D}^{\mathrm{DC}}$ as the new input $\mathbf{D}^{\mathrm{in}}$. However, the original hyperspectral data matrix $\mathbf{D}^{\mathrm{spike}}$ continues to be used in step 13.

Practical Implementation Considerations. While a single iteration of either the UBS-DM or the UBS-DM-HS algorithm will reduce the number and magnitude of the cosmic ray spikes, multiple iterations are recommended[13] as each algorithm approaches a

final, stable result. While the original paper suggested 3 iterations of the UBS-DM algorithm was generally sufficient,[13] the ideal number of iterations was found to vary between images. In particular, for the simulated data sets, more iterations were generally found to be desirable, most likely due to the higher rate of cosmic ray spikes in the simulations. Therefore, after each iteration of the UBS-DM-HS and UBS-DM algorithms, the correlation coefficient was calculated between each input spectrum and its corresponding output spectrum. Iterations were stopped once the mean of these correlation coefficients was within 1 part in 1 million of 1.

EXPERIMENTAL

Simulated Data. In order to create realistic simulated hyperspectral data sets, linear superpositions of cellulose, sucrose, and xanthan reference Raman spectra[32] were generated, excluding Raman shifts of less than 100 cm^{-1} (See Supplementary Figure S5). For each simulated spectrum, the concentration of each of the spectral components was randomly determined from a uniform random distribution. The maximum number of photon counts totaled across the spectrum was 5000 for cellulose, 3000 for sucrose, and 1000 for Xanthan. Poisson shot noise and Gaussian readout noise (with a standard deviation of 5 counts) were then added to each noise-free spectrum in order to generate image data with realistic noise.

To investigate the effect the shape of the cosmic ray spike has on the performance of the algorithms, two simulations were generated that differed only in the shape of the cosmic ray spikes. The first employed stereotypical cosmic ray spikes, where each spike affected only a single spectral position. In contrast, the second simulation convolved the spectral profile of each cosmic ray spike with a Gaussian kernel, reflecting the fact that many cosmic ray spikes have been experimentally shown to span multiple pixels and can be modeled by a Gaussian profile. The specific kernel employed ([.13 .37 .37 .13]) consisted of a Gaussian with a width 1 cm^{-1} centered at the middle of the 4-pixel kernel, spreading each cosmic ray spike over 4 adjacent pixels. Since the first simulation corresponds to a delta-function kernel (point spikes) and the second to a Gaussian kernel, for brevity the simulations are referred to as the point-spike simulation and Gaussian simulation. For each simulation, a total of 4096 spectra were generated.

The positions (both spatially and spectrally) and intensities of the cosmic ray spikes were randomly generated. The total intensity (across all wavelengths) of each cosmic ray spike was determined randomly from an exponential distribution with a mean value of 5000 photon counts. The probability of a spike occurring at any given position in a spectrum was 1 in 5000. Therefore, the rate of cosmic ray spikes in the simulation was approximately 3 orders of magnitude higher than typically observed experimentally. The

higher rate of cosmic ray spikes was chosen for several reasons. First, the higher rate of cosmic ray spikes increased the frequency of spikes occurring in close proximity, a situation known to be challenging for some despiking algorithms. Second, the desire was to evaluate the algorithms' performance under worst-case scenario conditions. Finally, as an added benefit, by increasing the incidence of cosmic ray spikes per spectra, smaller simulations could be employed, reducing the computational time required.

Actual Data. Mouse P388 cells were cultured with 5% CO_2 at 37°C in DMEM with 5% FBS on #1.5 coverslips. After the media was removed and the cells were rinsed in HBBS, the coverslip was inverted and sealed to a glass slide with an *in situ* frame (ThermoFisher Scientific) with 25 µL HBSS. Hyperspectral Raman microscopy was performed on a WiTec Alpha 300R confocal Raman microscope equipped with a UHTS300 spectrometer (grating 600 g/mm blazed at 500 nm). In this arrangement, the spectral response is 3 cm^{-1} per pixel. Improved spectral resolution is obtained by fitting the discrete per-pixel data to a continuous Voigt distribution. Employing peak fitting reduces the uncertainty associated with specifying a peak position and increases the spectral accuracy to ±1 wavenumber. The sample was excited with 3 mW of power from a 532 nm laser using a 50X 0.55 NA objective producing an approximately diffraction-limited spot. A 36 by 36 point image was generated with a 0.5 µm step size and an integration time of 2

seconds/point using an Andor DU970 EMCCD camera. The detector was run at -88°C in conventional CCD mode (16-bit) with a preamplifier gain of 1 and full vertical binning.

Algorithm settings. Except when testing the dependence of the algorithms upon their input arguments, both the UBS-DM and UBS-DM-HS algorithms were run using a value of 3 for the number of suspected spectral components in the sample. A readout noise value of $\sigma = 5$ was used throughout, matching the value used when simulating the readout noise. Except when testing the dependence of the UBS-DM-HS algorithm upon t, the threshold t was set as 10 divided by the number of spectra in the data set. Three additional algorithms (median-filtering, Zhang-Henson, and Katsumoto-Ozaki) are evaluated in the Supplementary Information. Details of the settings for those algorithms are provided there.

RESULTS AND DISCUSSION

The performance of the algorithms was evaluated using simulated data for which the ground truth is known. For each simulation, we have not only the hyperspectral image contaminated with cosmic ray spikes (spiky image), but also the image containing noise but without added cosmic ray spikes (noisy image) and the image as it would appear in the absence of any readout or shot noise (noise-free image). The objective when despiking is simply to remove the cosmic ray spikes, recovering the noisy image from the spiky image.

Smoothing or any other operation which disturbs the underlying noise profile is undesirable as it can complicate subsequent chemometric analysis.[33] Therefore, applying each algorithm to the spiky image results in a recovered image for that algorithm, ideally identical to the noisy image. The simulated data was analyzed by subdividing the simulated data into four separate images of 1024 pixels each to allow the standard deviation to be computed. Each algorithm was then run on each of the four images and the values of the metrics determined for each image. Both the mean value and the standard deviation of each metric were then determined.

[Insert Figure 1.]

As would be expected from the algorithm design, while UBS-DM and UBS-DM-HS exhibit comparable performance for point spikes, UBS-DM-HS is much better at suppressing Gaussian spikes (see Figure 1). The underlying rationale is that UBS-DM can eliminate point-spikes which are associated with group II or group III, but can only eliminate larger bandwidth spikes when they are assigned to group III. Any spike which has a bandwidth of greater than 2 pixels (group IIa) or 3 pixels (group IIb) will only be truncated, not eliminated, when group II is median-filtered (Step 9). Successive rounds of median-filtering with the same filter will not truncate them further. Unfortunately, no matter how many iterations are run, the total number of eigenvectors in groups I and II will

generally be at least the number of real spectral components (group I) plus 40 (step 7). At least some of the 40-plus eigenvectors in group II can be expected to correspond to one or more partially truncated cosmic ray spikes which will never be eliminated by UBS-DM. In contrast, UBS-DM-HS has an alternate mechanism to eliminate cosmic ray spikes of any shape, step 10, which applies to all eigenvectors regardless of group. As long as the cosmic ray spikes do not appear disproportionately at a single spectral position, the corresponding eigenvector will have significant weight for only a few spectra. This additional step allows UBS-DM-HS to recognize and eliminate cosmic ray spikes of any bandwidth.

Where the residual intensity (Figure 1) serves as a measure of how effectively the algorithms suppress (or eliminate) cosmic ray spikes, the absolute deviation (Figure 2) serves as a measure of how much bias each algorithm introduces. The average deviation is calculated by first subtracting the ideal result (the noisy image) from the result for each algorithm to obtain the residuals (shown in Figure 1). Next, the average deviation is determined by averaging the residuals for all the spectra in a manner that keeps cosmic ray spikes from contributing. Specifically, although the cosmic ray spikes were randomly distributed (both spatially and spectrally) in the simulations, we know where each simulated spike was placed. Therefore, when averaging the spectra, for each spectrum only the positions which were not contaminated with cosmic ray spikes were included in the

average. See the Supplementary Information for a formal mathematical definition. As a result, the absolute deviation evaluates the amount of bias each algorithm introduces to the spectra, unaffected by any incompletely suppressed cosmic ray spikes.

[Insert Figure 2.]

As can be seen in Figure 2, when applied to the point spike simulation UBS-DM and UBS-DM-HS introduce similar, small amounts of negative bias. However, when applied to the Gaussian spike simulation UBS-DM introduces additional bias appearing as sharp, negative spikes which are not seen with UBS-DM-HS. The reason UBS-DM introduces this bias is that even when an eigenvector is predominantly associated with a cosmic ray spike, they frequently contribute minor amounts to many other spectra, helping to represent the noise in the spectra. Therefore, simply eliminating eigenvectors associated with cosmic ray spikes tends to introduce bias with a similar shape. Initially, the sign of the bias would be expected to vary from spectrum to spectrum. However, since the UBS-DM and UBS-DM-HS algorithms have a tendency to truncate large, positive noise, after multiple iterations the net result would be primarily negative bias. For this reason, when designing UBS-DM-HS we chose not to completely eliminate eigenvectors identified as cosmic ray spikes in step 10, instead only eliminating them from the spectra where they contributed to cosmic ray spikes.

[Insert Figure 3.]

Figure 3 effectively combines the information displayed in Figures 1 and 2, providing quantitative measures of the despiking effectiveness (Figure 1) and the amount of bias introduced (Figure 2). Whereas Figures 1 and 2 showed only the final result for each algorithm, Figure 3 shows the performance after each iteration. The first metric, total residual spike counts, is designed to evaluate how effective each algorithm was at suppressing the cosmic ray spikes. The second metric, spectral bias, is designed to evaluate the extent of the systematic errors each algorithm introduces, biasing the spectral average. The first step in calculating both metrics is to calculate the residuals by subtracting the noisy image from the algorithm results. Next, the spiky image simulations were compared to the noisy image simulations to determine which elements of the hyperspectral data matrices contained spikes. The first metric, total residual spike counts, is evaluated only on those elements containing spikes while the second metric, spectral bias, only uses elements which do not contain spikes. The total residual spike count consists of the sum of the absolute value of the residuals for all locations containing simulated spikes. Meanwhile, the spectral bias is calculated by first determining the average deviation (Figure 2), then taking the sum of the absolute value of all elements of the average deviation. See the

Supplementary Information for formal mathematical definitions. The lower left corner of the plot represents ideal despiking algorithm performance, no residual spikes and no bias.

Figure 3 reveals that UBS-DM and UBS-DM-HS perform comparably for single-point cosmic ray spikes. UBS-DM and UBS-DM-HS require similar numbers of iterations to converge (11 and 12 iterations respectively) and converge to nearly the same point (the differences cannot be declared statistically significant at the $p<0.05$ level). Figure 3 also shows that neither UBS-DM nor UBS-DM-HS perform as well when despiking Gaussian cosmic ray spikes, but for this case UBS-DM-HS significantly outperforms UBS-DM (statistical significance determined using a t-test, where $p = 0.02$ for spectral bias and $p = 0.002$ for the total residual spike counts). Additionally, UBS-DM requires many more iterations to reach that result, requiring 105 iterations where UBS-DM-HS required only 22. Since the computational time per iteration is similar for UBS-DM-HS and UBS-DM, the two algorithms required comparable time for point spikes while UBS-DM-HS was significantly faster when multi-point spikes were included. Both algorithms were capable of processing thousands of spectra per minute, suitable for offline processing of hyperspectral data.

While UBS-DM-HS proved superior to UBS-DM at removing Gaussian spikes, neither algorithm was as effective for the broader spikes as for point spikes. To ensure this

was a general limitation and not simply limited to these algorithms, three other algorithms (median-filtering, Katsumoto-Ozaki, and Zhang-Henson) were also evaluated (see supplementary material). All algorithms were selected for their ease of implementation, and source code for the algorithms is provided in the supplemental material. Collectively, Katsumoto-Ozaki, Zhang-Henson, and UBS-DM provide comparison to the state-of-the-art in three of the four categories of cosmic ray despiking algorithms. Our experimental conditions are not suitable for the final category, multiple-acquisition despiking algorithms. Our analysis (see Supplementary Figure S3) showed that all algorithms have more difficulty with broader cosmic ray spikes. Of these algorithms, only Zhang-Henson proved competitive with UBS-DM and UBS-DM-HS. When multiple pure spectra were provided as inputs, Zhang-Henson proved nearly as effective as UBS-DM-HS at suppressing cosmic ray spikes while introducing less spectral bias. As such, the Zhang-Henson algorithm is an excellent choice when the pure spectra are known *a priori*, such as the pharmaceutical processing scenario for which it was designed.[7] However, when the intention is to perform exploratory chemometric analysis to determine the spectra, the pure spectra are not available as inputs rendering UBS-DM-HS more suitable.

[Insert Figure 4.]

[Insert Table 1.]

Figure 4 demonstrates the importance of despiking when intending to perform chemometric analysis. Scree plots (see Figure 4a) are commonly used after performing principal component analysis during chemometric analysis. In the absence of despiking it is difficult to accurately determine the number of spectral components in the sample. Since these simulations included three distinct spectra, ideally the first three eigenvalues should be large while the remaining eigenvalues should be much smaller (they would be zero in the absence of noise). The presence of cosmic ray spikes results in additional components that have similar eigenvalues to the 3rd component. Despite the substantial reduction in Gaussian cosmic ray spikes provided by UBS-DM, the eigenvalues for the 3rd and 4th components remain similar. In contrast, UBS-DM-HS offers performance nearly indistinguishable from what would be seen for the noisy image. Table 1 reveals that when multivariate curve resolution (MCR)[34] is applied to the output of the UBS-DM-HS algorithm, MCR is able to correctly recover all three spectral components (see supplementary material for more detail). In contrast, MCR only correctly recovers two spectral components for UBS-DM. The third spectral component for UBS-DM only partially models the actual spectrum and several cosmic ray spikes are clearly visible (see Figure 4b).

[Insert Figure 5.]

The most significant factor limiting the applicability of the UBS-DM and UBS-DM-HS algorithms is that they require large numbers of spectra all containing contributions from the same spectral components. Hyperspectral imaging routinely deals with thousands of such spectra, the number of spectra recommended by UBS-DM.[13] However, these algorithms could be more broadly applied if the required number of spectra could be reduced, including possibly performing real-time despiking for large data sets. Therefore, we investigated how the performance of these two algorithms depended upon the number of spectra analyzed. As shown in Figure 5, the performance of the UBS-DM-HS algorithm is relatively stable as long as more than 128 spectra are analyzed simultaneously. The UBS-DM algorithm shows a similar dependence when processing single-point cosmic ray spikes. On the other hand, the performance of the UBS-DM algorithm varies continuously with size when attempting to remove multi-point Gaussian spikes. Note however, these simulations were designed such that every recorded spectrum contained contributions from the various spectral components. Therefore, the recommended minimum number of spectra would have to be increased accordingly if some of the spectra were largely background, for instance measuring positions outside a cell in live cell imaging. The number of spectra required also depends upon the degree of cosmic

ray spike contamination in the data where simulations were designed to test worst-case scenario performance.

[Insert Figure 6.]

Figure 6 provides a qualitative comparison of the performance of UBS-DM and UBS-DM-HS, using an experimental hyperspectral Raman image of a live cell. The spikes observed in Figure 6a were not artificially added but are the naturally occurring cosmic ray spike contamination, thus reflecting the natural distribution of spike shape and size. Both algorithms substantially reduce the cosmic ray spikes but visible spikes remain for UBS-DM that are largely eliminated by UBS-DM-HS. Despiking with the UBS-DM-HS algorithm was also evaluated on hyperspectral fluorescence images (not shown), where it also exhibited excellent performance.

CONCLUSION

We present a new despiking algorithm, the UBS-DM-HS algorithm, specifically for preparing/preprocessing hyperspectral image data for subsequent chemometric analysis, such as multivariate curve resolution. The performance of UBS-DM-HS algorithm was assessed and compared to several popular despiking algorithms using both realistically simulated hyperspectral image data and real live cell Raman image data. The results

conclude the UBS-DM-HS algorithm is well-suited for hyperspectral image data and represents an improvement over the UBS-DM algorithm from which it was derived and other algorithms tested. It is able to detect and remove broader cosmic ray spikes, while retaining the benefit of not requiring knowledge of the pure spectral components *a priori*. Importantly, the UBS-DM-HS method is able to be applied to hyperspectral data with hundreds rather than thousands of spectra, which has direct applicability to new problems requiring real-time multivariate analysis of hyperspectral image data, where a data set may consist of a small portion of the anticipated image data.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Thomas Beechem and Anthony McDonald for the use of the Raman microscope and assistance with the data collection and Bryan Carson for helpful discussion. This work was supported by Sandia National Laboratories Laboratory Directed Research and Development program, under projects "Unknown Pathogen Detection in Clinical Samples: A Novel Hyperspectral Imaging and Single Cell Sequencing Approach" and "Unmasking Hidden Compounds within Hyperspectral Images." Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin

Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

REFERENCES

1. Theuwissen AJP. Influence of terrestrial cosmic rays on the reliability of CCD image sensors - Part 1: Experiments at room temperature. *Ieee T Electron Dev*. 2007; 54: 3260-6.
2. Whiteson D, Mulhearn M, Shimmin C, Brodie K and Burns D. Observing Ultra-High Energy Cosmic Rays with Smartphones. *arXiv preprint arXiv:14102895*. 2014.
3. Cogliati JJ, Derr KW and Wharton J. Using CMOS Sensors in a Cellphone for Gamma Detection and Classification. *arXiv preprint arXiv:14010766*. 2014.
4. Takeuchi H, Hashimoto S and Harada I. Simple and Efficient Method to Eliminate Spike Noise from Spectra Recorded on Charge-Coupled Device Detectors. *Appl Spectrosc*. 1993; 47: 129-31.
5. Behrend CJ, Tarnowski CP and Morris MD. Identification of outliers in hyperspectral Raman image data by nearest neighbor comparison. *Appl Spectrosc.* 2002; 56: 1458-61.
6. Esmonde-White FWL, Schulmerich MV, Esmonde-White KA and Morris MD. Automated Raman spectral preprocessing of bone and other musculoskeletal tissues. *SPIE BiOS: Biomedical Optics*. 2009, p. 716605--10.
7. Zhang L and Henson MJ. A practical algorithm to remove cosmic spikes in Raman imaging data for pharmaceutical applications. *Appl Spectrosc*. 2007; 61: 1015-20.
8. Chew W. Information-theoretic chemometric analyses of Raman data for chemical reaction studies. *Journal of Raman Spectroscopy*. 2011; 42: 36-47.
9. Cappel UB, Bell IM and Pickard LK. Removing Cosmic Ray Features from Raman Map Data by a Refined Nearest Neighbor Comparison Method as a Precursor for Chemometric Analysis. *Appl Spectrosc*. 2010; 64: 195-200.
10. Zhang DM, Jallad KN and Ben-Amotz D. Stripping of cosmic spike spectral artifacts using a new upper-bound spectrum algorithm. *Appl Spectrosc*. 2001; 55: 1523-31.
11. Farage CL and Pimbblet KA. Evaluation of Cosmic Ray Rejection Algorithms on Single-Shot Exposures. *Publications of the Astronomical Society of Australia*. 2005; 22: 249-56.
12. Mozharov S, Nordon A, Littlejohn D and Marquardt B. Automated Cosmic Spike Filter Optimized for Process Raman Spectroscopy. *Appl Spectrosc*. 2012; 66: 1326-33.
13. Zhang DM and Ben-Amotz D. Removal of cosmic spikes from hyper-spectral images using a hybrid upper-bound spectrum method. *Appl Spectrosc*. 2002; 56: 91-8.

14. Morris MD, Timlin JA, Carden A, Tarnowski CP and Edwards CM. Multivariate data reduction techniques for hyperspectral Raman imaging. *Prog Biom O*. 2000; 1: 151-8.
15. Katsumoto Y and Ozaki Y. Practical algorithm for reducing convex spike noises on a spectrum. *Appl Spectrosc*. 2003; 57: 317-22.
16. Maury A and Revilla RI. Autocorrelation Analysis Combined with a Wavelet Transform Method to Detect and Remove Cosmic Rays in a Single Raman Spectrum. *Appl Spectrosc*. 2015; 69: 984-92.
17. Egan WJ and Morgan SL. Outlier detection in multivariate analytical chemical data. *Anal Chem*. 1998; 70: 2372-9.
18. De Groot P, Postma G, Melssen W, Buydens L, Deckert V and Zenobi R. Application of principal component analysis to detect outliers and spectral deviations in near-field surface-enhanced Raman spectra. *Anal Chim Acta*. 2001; 446: 71-83.
19. Jones HDT, Haaland DM, Sinclair MB, Melgaard DK, Collins AM and Timlin JA. Preprocessing strategies to improve MCR analyses of hyperspectral images. *Chemometr Intell Lab*. 2012; 117: 149-58.
20. Phillips GR and Harris JM. Polynomial Filters for Data Sets with Outlying or Missing Observations - Application to Charge-Coupled-Device-Detected Raman-Spectra Contaminated by Cosmic-Rays. *Anal Chem*. 1990; 62: 2351-7.
21. Ehrentreich F and Sümmchen L. Spike removal and denoising of Raman spectra by wavelet transform methods. *Anal Chem*. 2001; 73: 4364-73.
22. Zhang D, Hanna JD and Ben-Amotz D. Single scan cosmic spike removal using the upper bound spectrum method. *Appl Spectrosc*. 2003; 57: 1303-5.
23. Zhao J. Image curvature correction and cosmic removal for high-throughput dispersive Raman spectroscopy. *Appl Spectrosc*. 2003; 57: 1368-75.
24. Sabin GP, de Souza AM, Breitkreitz MC and Poppi RJ. Development of an Algorithm for Identification and Correction of Spikes in Raman Imaging Spectroscopy. *Quim Nova*. 2012; 35: 612-5.
25. James TM, Schlosser M, Lewis RJ, Fischer S, Bornschein B and Telle HH. Automated Quantitative Spectroscopic Analysis Combining Background Subtraction, Cosmic Ray Removal, and Peak Fitting. *Appl Spectrosc*. 2013; 67: 949-59.
26. Li S and Dai LK. An Improved Algorithm to Remove Cosmic Spikes in Raman Spectra for Online Monitoring. *Appl Spectrosc*. 2011; 65: 1300-6.
27. Li B, Calvet A, Casamayou-Boucau Y and Ryder AG. Kernel principal component analysis residual diagnosis (KPCARD): An automated method for cosmic ray artifact removal in Raman spectra. *Anal Chim Acta*. 2016; 913: 111-20.
28. Sinclair MB, Haaland DM, Timlin JA and Jones HDT. Hyperspectral confocal microscope. *Appl Optics*. 2006; 45: 6283-91.

29. Sinclair MB, Timlin JA, Haaland DM and Werner-Washburne M. Design, construction, characterization, and application of a hyperspectral microarray scanner. *Appl Optics*. 2004; 43: 2079-88.
30. Christensen KA and Morris MD. Hyperspectral Raman microscopic imaging using Powell lens line illumination. *Appl Spectrosc*. 1998; 52: 1145-7.
31. Note that the paper describing the UBS-DM method indicates that Si is the ith row vector. However, the corresponding explanation makes clear that the various elements in Si indicate the relative contribution of the ith eigenvector to the different pixels, dictating that a column vector of S must be employed. The elements of a row of S instead correspond to the relative contribution of the different eigenvectors to a single pixel. Therefore, the use of a column vector here corresponds to the intended implementation of UBS-DM.
32. http://www.models.life.ku.dk/~specarb/cellulose.html

http://www.models.life.ku.dk/~specarb/sucr.html

http://www.models.life.ku.dk/~specarb/xanthan.html

33. Ruckebusch C and Blanchet L. Multivariate curve resolution: A review of advanced and tailored applications and challenges. Ana*l Chim Acta. 2*013; 765: 28-36.
34. Haaland DM, Jones HDT, Van Benthem MH, et al. Hyperspectral Confocal Fluorescence Imaging: Exploring Alternative Multivariate Curve Resolution Approaches. App*l Spectrosc. 2*009; 63: 271-9.

FIGURE CAPTIONS

Figure 1. The residual differences between the ideal result and the results obtained from applying UBS-DM and UBS-DM-HS to spectra with simulated a) single-point and b) Gaussian cosmic ray spikes. The residuals are offset for clarity, with the UBS-DM results shifted upwards by 200 counts. Within each algorithm, the results for 1024 simulated spectra are overlaid. Cosmic ray spikes which are incompletely eliminated manifest as large, positive residual values. Negative residuals generally result from truncation of experimental noise. While both algorithms are excellent at suppressing single-point cosmic ray spikes (a), UBS-DM-HS is much better than UBS-DM at suppressing broader spikes such as the simulated Gaussian spikes (b).

Figure 2. The average deviation (see text) when applying UBS-DM and UBS-DM-HS to spectra with simulated a) single-point and b) Gaussian cosmic ray spikes. The residuals are offset for clarity, with the UBS-DM results shifted upwards by 0.3 counts. The average deviation provides a measure of the amount of bias each algorithm introduces. Both algorithms exhibit a tendency to truncate the most extreme, positive noise, resulting in small amounts of negative bias. However, when processing images with simulated

Gaussian spikes (b), UBS-DM introduces additional bias appearing as sharp, negative spikes which are not seen for UBS-DM-HS.

Figure 3. Quantitative evaluation of the performance of the UBS-DM and UBS-DM-HS algorithms depending upon the number of iterations (see text for definitions). The figure includes results for both single-point cosmic ray spikes and Gaussian cosmic ray spikes spanning 4 points (see legend). Within each algorithm, only the symbol for the first iteration is solid and the symbols for successive iterations are connected by line segments. For the final iteration, error bars (std. dev.) are shown in black; in all except one case, they are smaller than the symbols. As shown, running multiple iterations improves the performance of both UBS-DM and UBS-DM-HS as an ideal algorithm would have zero residual spike counts and zero spectral bias. UBS-DM and UBS-DM-HS require similar number of iterations to converge when processing point spikes (11 and 12 iterations respectively) but when processing Gaussian spikes UBS-DM requires many more iterations than UBS-DM-HS (105 and 22 respectively).

Figure 4. Chemometric analysis of the Gaussian spike simulation showing that incomplete suppression of cosmic ray spikes prevents cleanly decomposing the hyperspectral data matrix to obtain the concentrations and spectra. a) Scree plots showing the eigenvalues versus the component number for the UBS-DM and UBS-DM-HS algorithms. The corresponding scree plot for the unprocessed data including simulated Gaussian cosmic ray spikes is also shown. For both the unprocessed data and the UBS-DM scree plot, the 3rd and 4th components have similar eigenvalues making it difficult to determine the number of true components in the sample. In contrast, the presence of 3 true components can easily be determined from the abrupt transition between the 3rd and 4th components in the UBS-DM-HS scree plot (known as the "elbow"), where from the 4th component onward the eigenvalues are much lower and nearly constant. On this scale, the scree plot for the noisy image (not shown) would be identical to the result for UBS-DM-HS. b) Comparison of the 3rd spectral component obtained from MCR chemometric analysis of the outputs of the UBS-DM and UBD-DM-HS algorithms. For clarity, the UBS-DM algorithm is offset upward by 2, UBS-DM-HS by 1, and the actual spectrum is shown at the baseline. UBS-DM-HS allows accurate retrieval of the all three spectral components (weakest spectral component shown). In contrast, UBS-DM only allows accurate retrieval of the strongest two spectral components (not shown) while the 3rd

spectral component poorly models the actual spectrum and remnants of several cosmic ray spikes are clearly visible (red arrows).

Figure 5. To determine the dependence of the algorithm upon the number of spectra analyzed, the same simulated spectra were analyzed by the UBS-DM and UBS-DM-HS algorithms while varying the batch size. Both the simulation including point spikes and the simulation including Gaussian spikes were evaluated. After each algorithm was run, the correlation coefficient was calculated between each resulting spectrum and the corresponding spectrum with no cosmic ray spikes. The average correlation is simply the average value for all 4096 spectra in a simulation. For the single-point cosmic ray spike simulation, the performance of both algorithms began to degrade when processing fewer than ~128 spectra at a time. A similar degradation in performance when processing fewer than ~128 spectra was observed for the UBS-DM-HS algorithm applied to the multi-point Gaussian cosmic ray spike simulation. In contrast, the UBS-DM algorithm's performance varied continuously for the Gaussian simulation.

Figure 6. a) The raw spectra from a 36 by 36 point hyperspectral Raman image of P388 cells were normalized to unit area and superimposed upon each other, where different colors correspond to different spectra. Dozens of cosmic ray spikes are clearly visible. b) The results of the UBS-DM and UBS-DM-HS algorithms are shown for the data in a). The UBS-DM algorithm result is shifted upward by 0.02 for clarity. For each algorithm, all spectra processed with that algorithm are superimposed on each other and share the same color. The rectangles spanning a) to b) highlight some of the cosmic ray spikes which were incompletely suppressed by the UBS-DM algorithm which the UBS-DM-HS algorithm either eliminated or reduced to near the noise-level of the spectrum.

Table 1. Similarity between the actual spectra and the spectra obtained from MCR chemometric analysis of the simulated image contaminated with Gaussian cosmic ray spikes after despiking with UBS-DM and UBS-DM-HS. The similarity is calculated as the correlation between each actual spectrum and the corresponding recovered spectrum. Both algorithms successfully recover the first two spectra, but only UBS-DM-HS recovers the final spectrum (see Figure 4b).

Figure 1

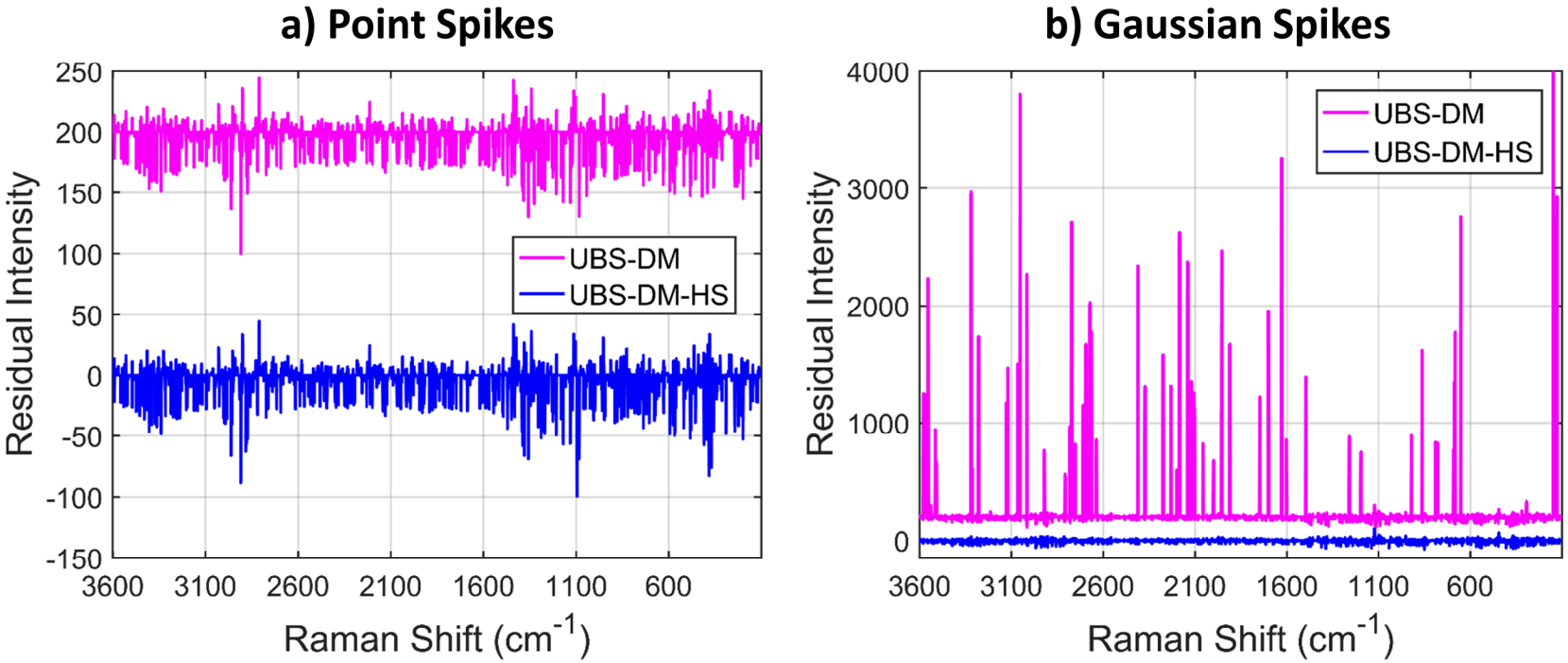

a) Point Spikes
Residual Intensity
250
200
150
100
50
0
-50
-100
-150
3600
3100
2600
2100
1600
1100
600
Raman Shift (cm-1)
UBS-DM
UBS-DM-HS
b) Gaussian Spikes
Residual Intensity
4000
3000
2000
1000
0
3600
3100
2600
2100
1600
1100
600
Raman Shift (cm-1)
UBS-DM
UBS-DM-HS

Figure 2

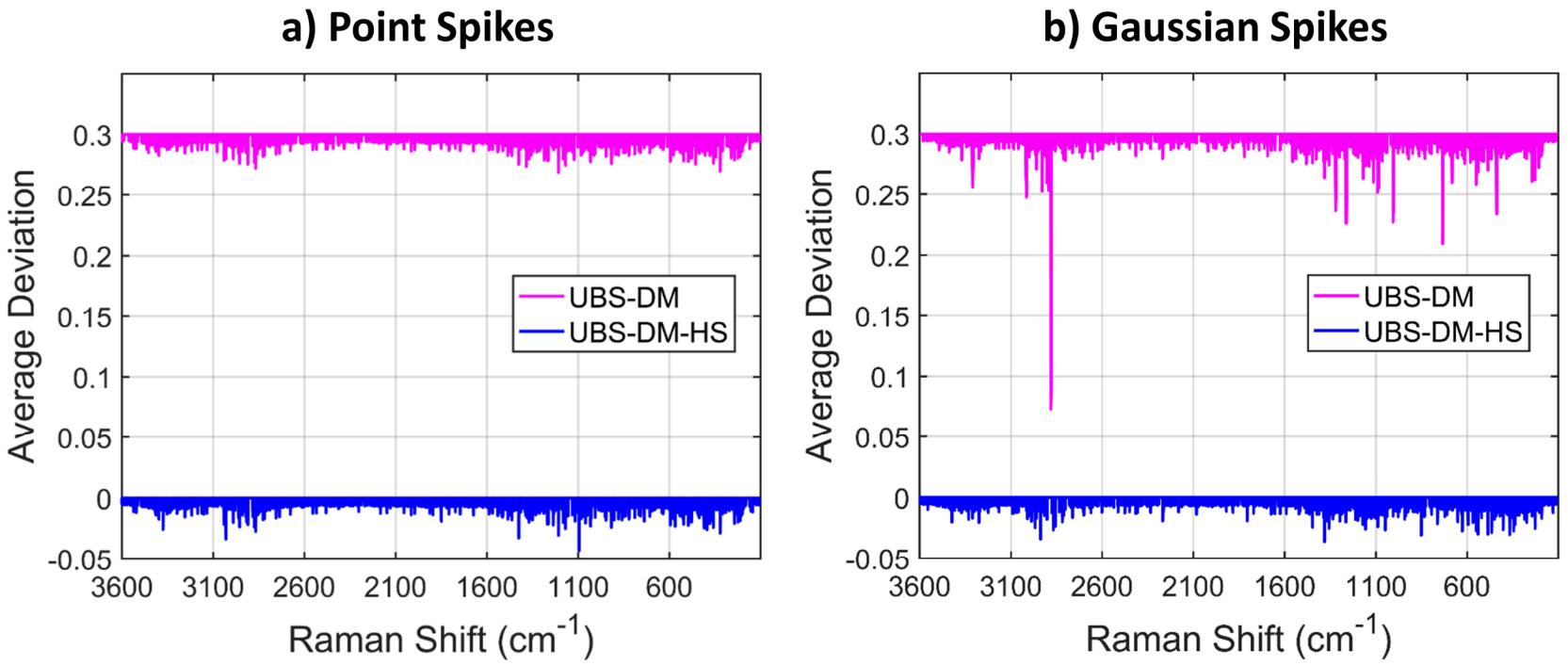

a) Point Spikes
b) Gaussian Spikes
Average Deviation
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
3600
3100
2600
2100
1600
1100
600
Raman Shift (cm-1)
UBS-DM
UBS-DM-HS

Figure 3

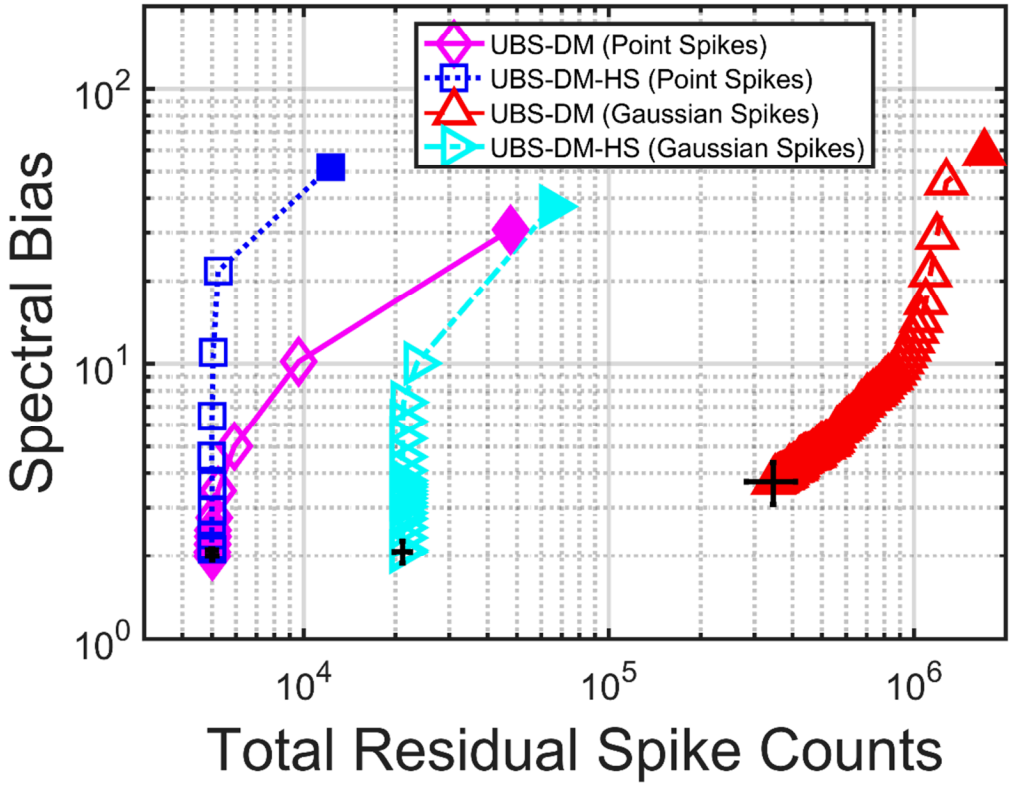
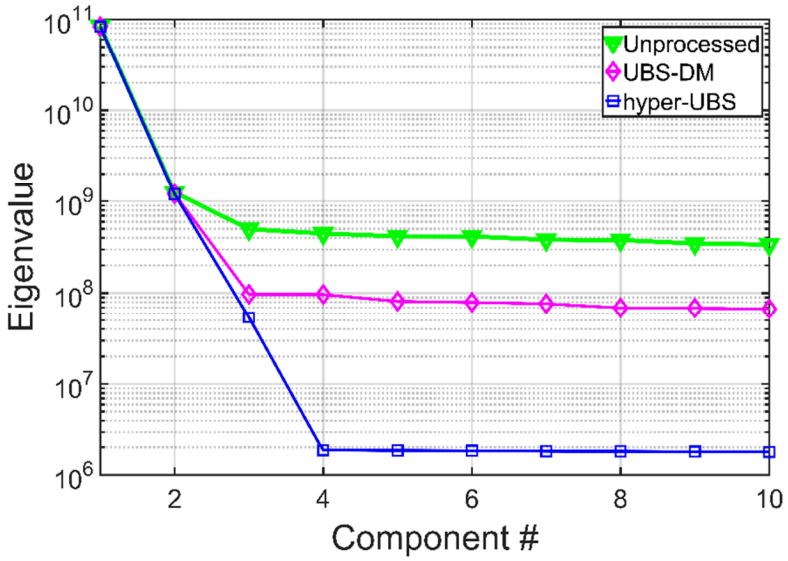

UBS-DM (Point Spikes)
UBS-DM-HS (Point Spikes)
UBS-DM (Gaussian Spikes)
UBS-DM-HS (Gaussian Spikes)
Spectral Bias
10^2
10^1
10^0
10^4
10^5
10^6
Total Residual Spike Counts

Figure 4

a) Scree Plot – Gaussian Spikes



b) 3rd MCR Spectral Component

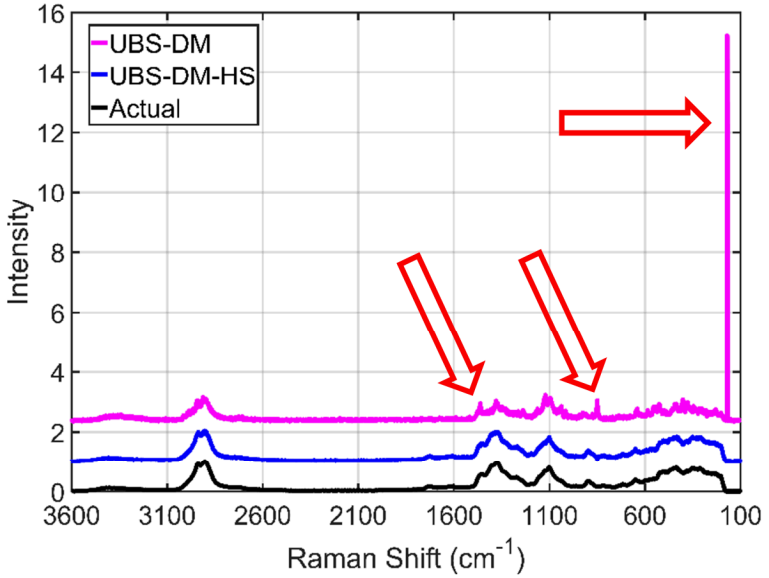


Figure 5

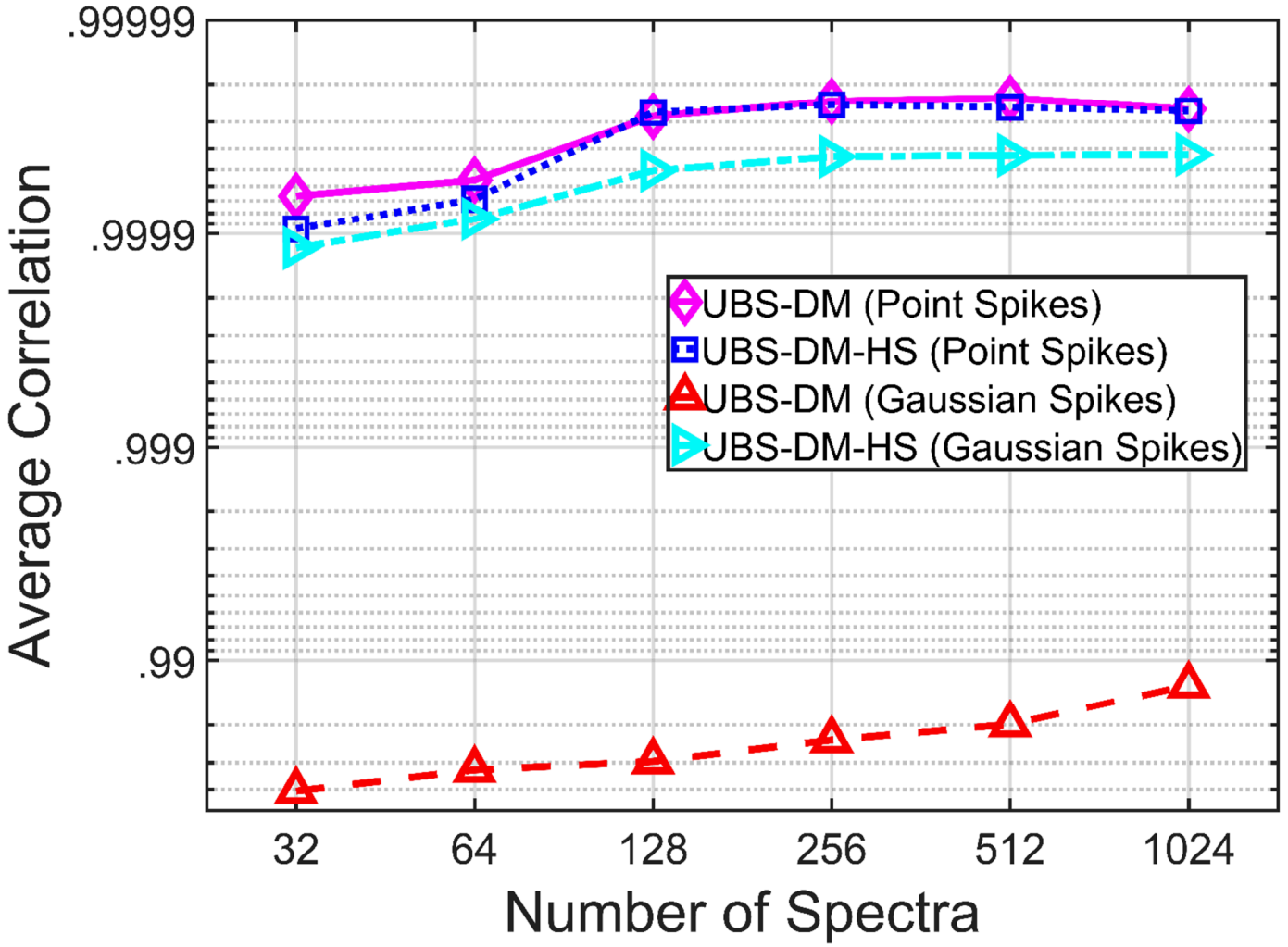

.99999
.9999
.999
.99
Average Correlation
32
64
128
256
512
1024
Number of Spectra
UBS-DM (Point Spikes)
UBS-DM-HS (Point Spikes)
UBS-DM (Gaussian Spikes)
UBS-DM-HS (Gaussian Spikes)

Figure 6

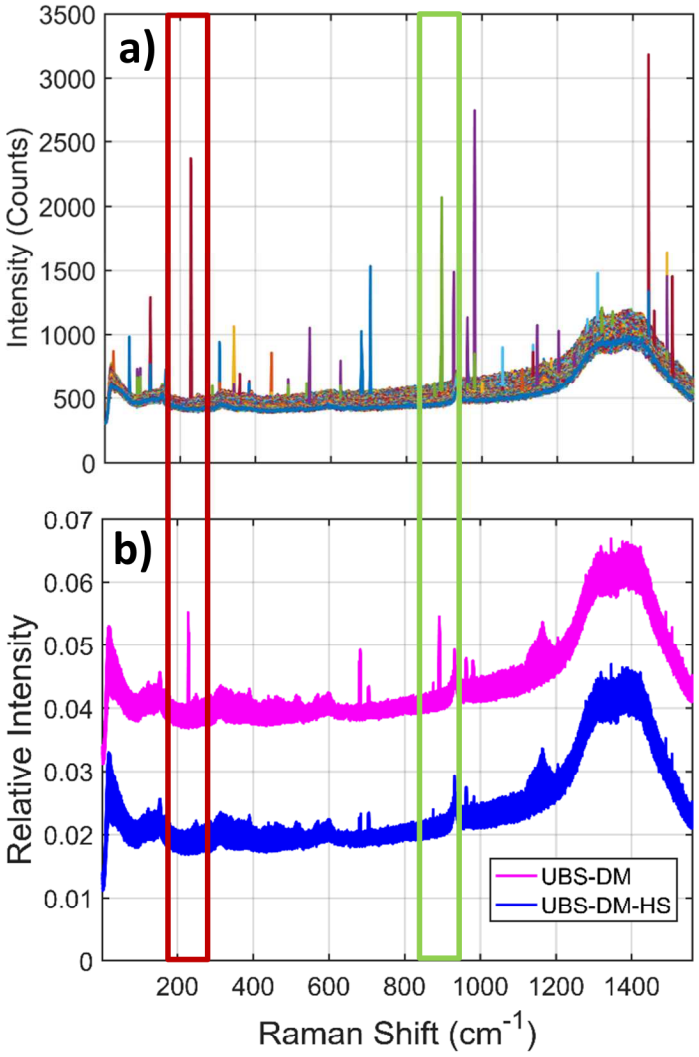

a)
b)
Intensity (Counts)
Relative Intensity
Raman Shift (cm-1)
UBS-DM
UBS-DM-HS
0
500
1000
1500
2000
2500
3000
3500
0.01
0.02
0.03
0.04
0.05
0.06
0.07
200
400
600
800
1000
1200
1400

Figure S1

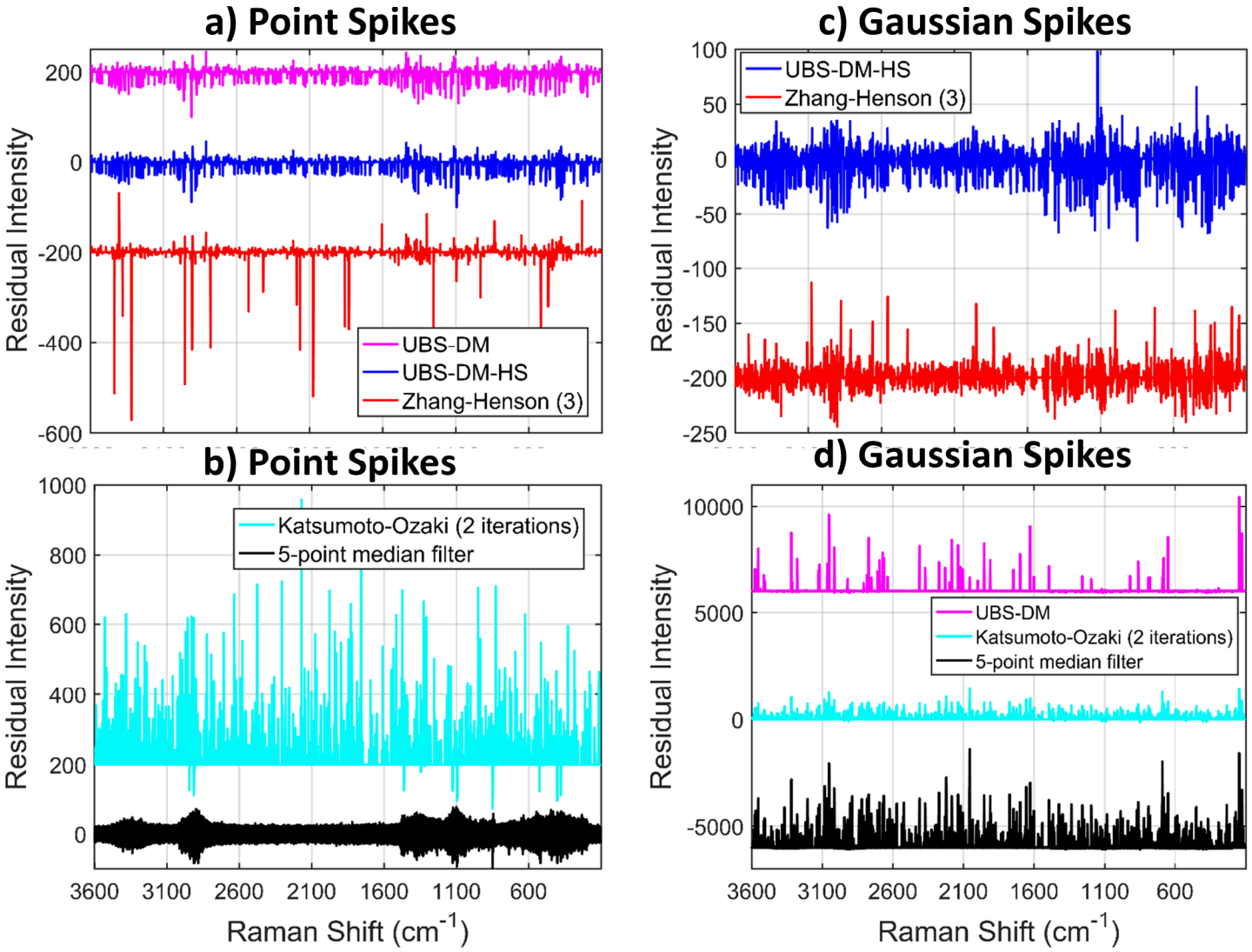

a) Point Spikes
Residual Intensity
200
0
-200
-400
-600
UBS-DM
UBS-DM-HS
Zhang-Henson (3)
c) Gaussian Spikes
Residual Intensity
100
50
0
-50
-100
-150
-200
-250
UBS-DM-HS
Zhang-Henson (3)
b) Point Spikes
Residual Intensity
1000
800
600
400
200
0
Katsumoto-Ozaki (2 iterations)
5-point median filter
3600 3100 2600 2100 1600 1100 600
Raman Shift (cm-1)
d) Gaussian Spikes
Residual Intensity
10000
5000
0
-5000
UBS-DM
Katsumoto-Ozaki (2 iterations)
5-point median filter
3600 3100 2600 2100 1600 1100 600
Raman Shift (cm-1)

Figure S2

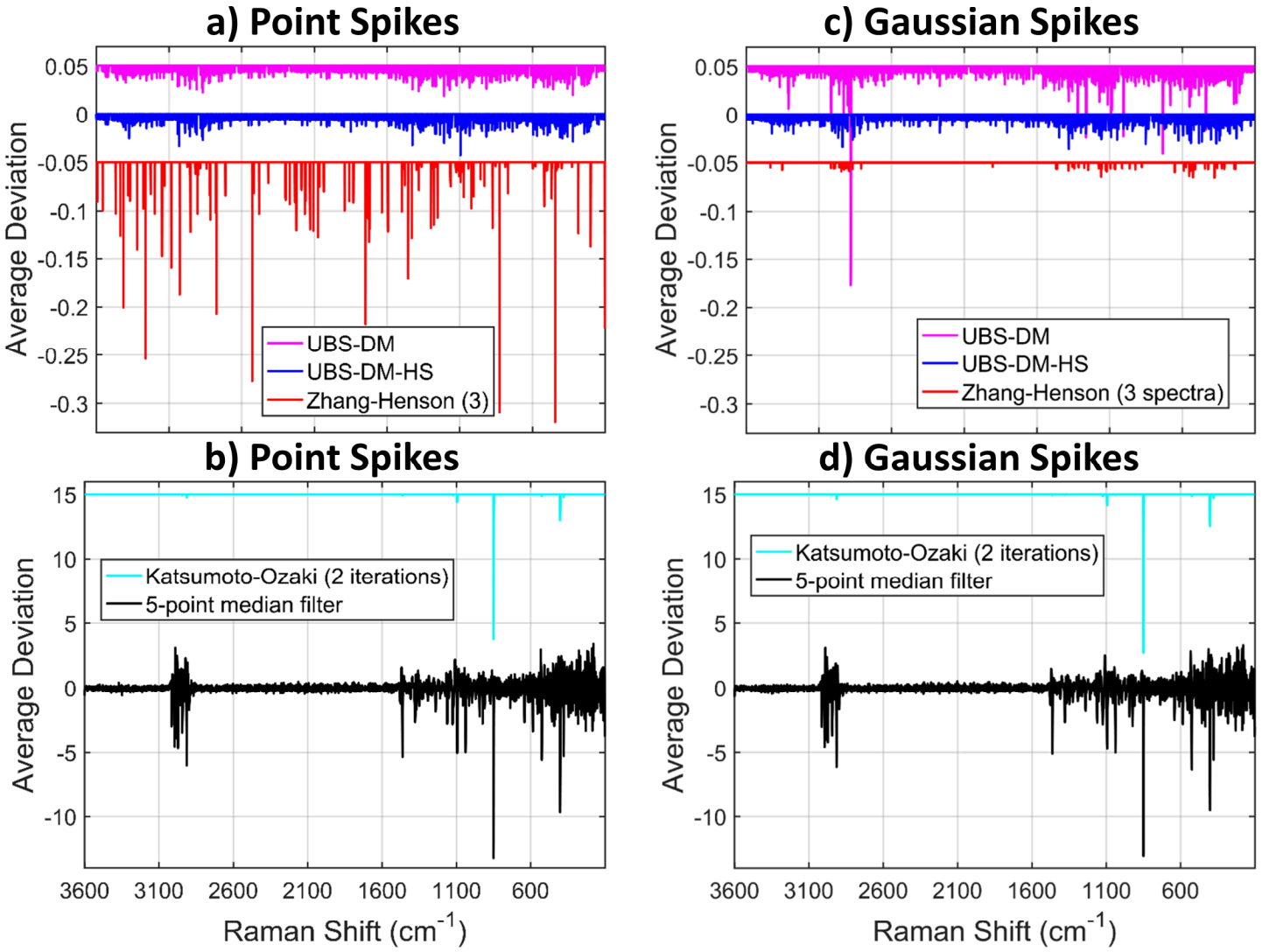

a) Point Spikes
c) Gaussian Spikes
b) Point Spikes
d) Gaussian Spikes
Average Deviation
UBS-DM
UBS-DM-HS
Zhang-Henson (3)
Zhang-Henson (3 spectra)
Katsumoto-Ozaki (2 iterations)
5-point median filter
Raman Shift (cm-1)
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
15
10
5
0
-5
-10
3600
3100
2600
2100
1600
1100
600

Figure S3

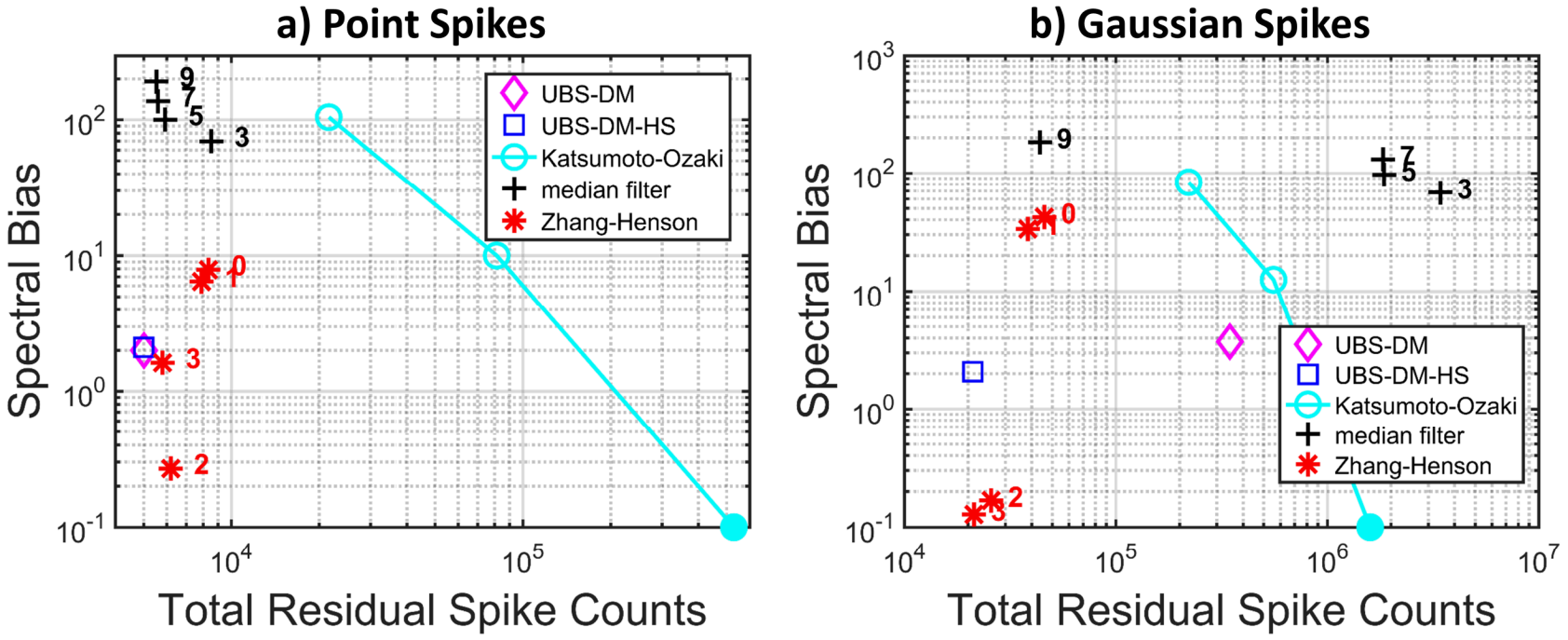

a) Point Spikes
b) Gaussian Spikes
Spectral Bias
Total Residual Spike Counts
UBS-DM
UBS-DM-HS
Katsumoto-Ozaki
median filter
Zhang-Henson

Figure S4

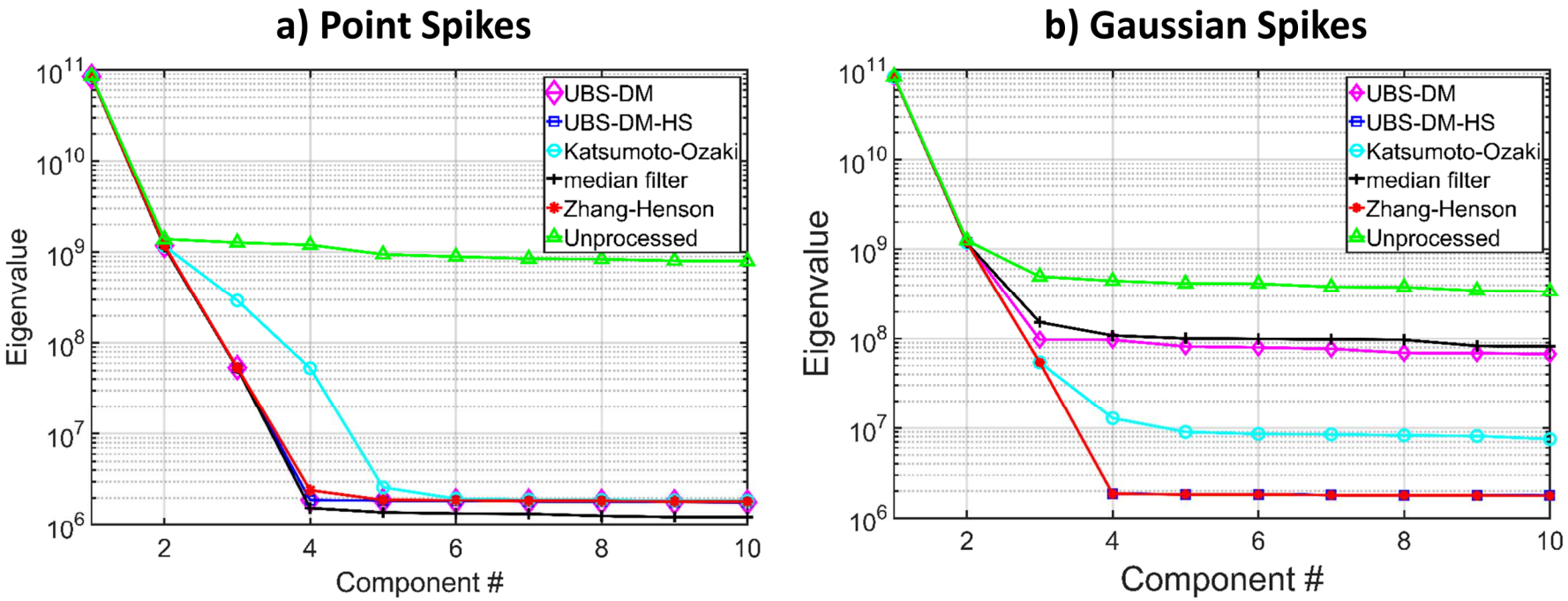

a) Point Spikes
b) Gaussian Spikes
Eigenvalue
Component #
UBS-DM
UBS-DM-HS
Katsumoto-Ozaki
median filter
Zhang-Henson
Unprocessed

Figure S5

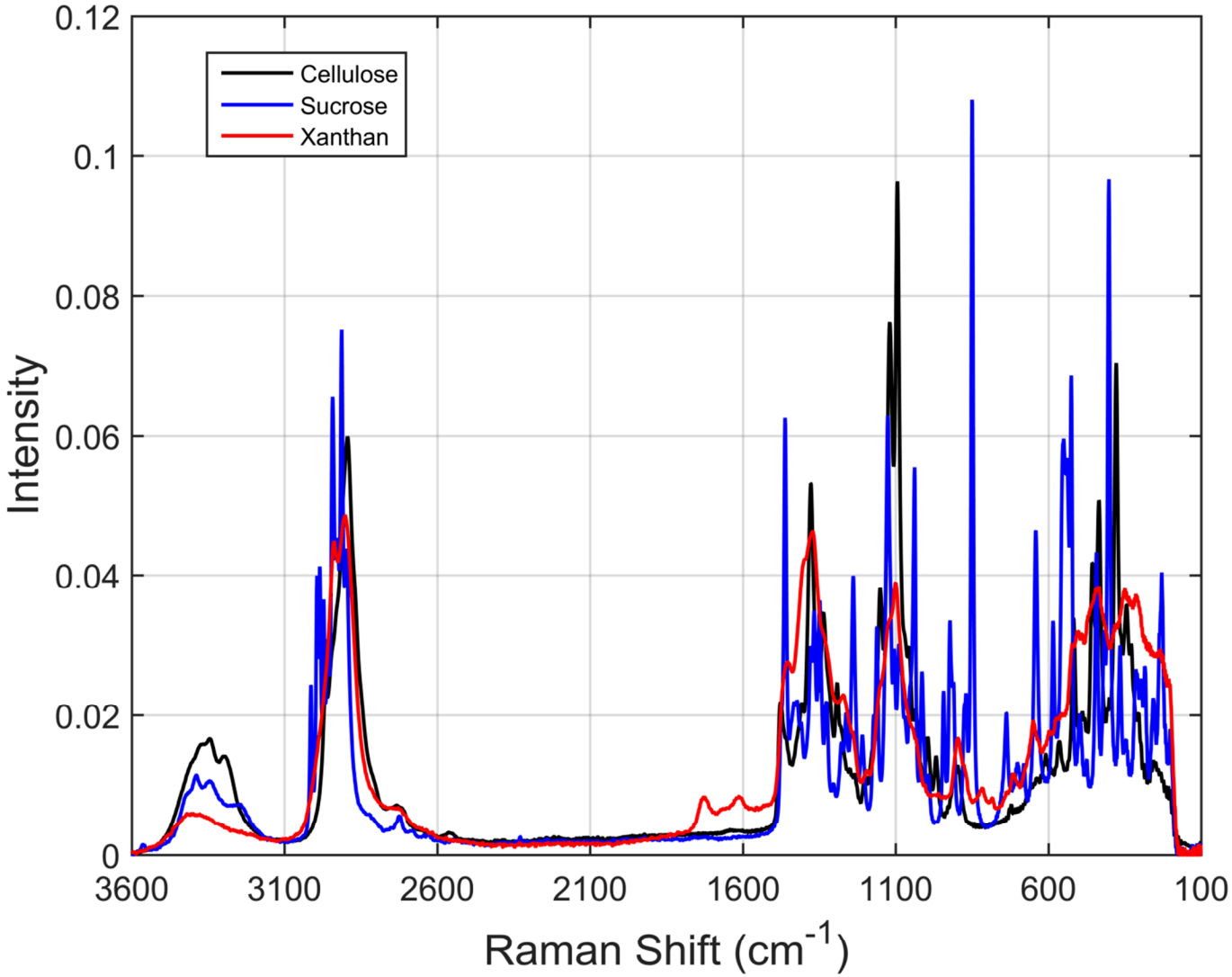

Cellulose
Sucrose
Xanthan
0.12
0.1
0.08
0.06
0.04
0.02
0
Intensity
3600
3100
2600
2100
1600
1100
600
100
Raman Shift (cm-1)

Figure S6

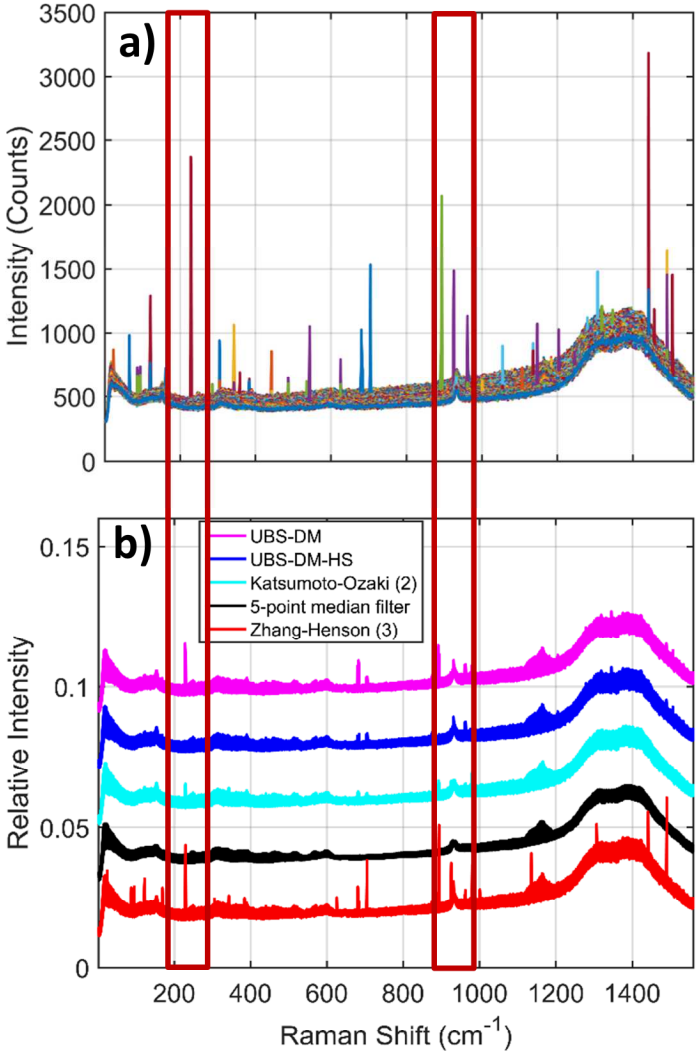

a)
b)
Intensity (Counts)
Relative Intensity
Raman Shift (cm^{-1})
UBS-DM
UBS-DM-HS
Katsumoto-Ozaki (2)
5-point median filter
Zhang-Henson (3)

Table 1

	1st Spectrum	2nd Spectrum	3rd Spectrum
UBS-DM	0.9992	0.9980	**0.2875**
UBS-DM-HS	1.0000	0.9999	0.9985

Supplementary Information: Removing Cosmic Spikes Using a Hyperspectral Upper-Bound Spectrum Method

Stephen M. Anthony and Jerilyn A. Timlin

EXPERIMENTAL

Algorithm settings. When analyzing the simulated data, four different sizes of the median filter were evaluated, spanning 3, 5, 7 and 9 pixels. The data points for the median-filter are labeled accordingly. For the Katsumoto-Ozaki and Zhang-Henson algorithms, appropriate values of the input variables were selected based upon preliminary analysis of which values appeared to perform best. For the Katsumoto-Ozaki algorithm a filter span size of 3 was selected. For the Zhang-Henson algorithm, the median-filtering step was performed using a filter span of 3 pixels, the lower threshold was set at 2.3 and the upper threshold at 10. The benefits of providing known spectra to the Zhang-Henson algorithm were evaluated for the simulated data. In the figures for the simulated data, the results for the Zhang-Henson algorithm are labeled with the number of pure component spectra provided, ranging from all of them (3) to none of them (0). When only a partial set of spectra were provided, the spectra which contributed the most total photons to the image as a whole were provided first.

Mathematical Definitions. When analyzing simulated data, there exist three hyperspectral data matrices associated with each algorithm ($\mathbf{D}^{noise}$, $\mathbf{D}^{spike}$, and $\mathbf{D}^{alg}$). The data matrices should be arranged as a data matrix $\mathbf{D}(m, n)$ of m spectra, each of which was measured

at n wavelengths. The first two hyperspectral data matrices, $\mathbf{D}^{\text{noise}}$ and $\mathbf{D}^{\text{spike}}$, correspond to the simulated dataset before and after the addition of the simulated cosmic ray spikes. Meanwhile, $\mathbf{D}^{\text{alg}}$ corresponds to the result of processing $\mathbf{D}^{\text{spike}}$ with the specified algorithm.

The absolute deviation (Figure S2) serves as a measure of how much bias each algorithm introduces. Calculation of the average deviation for an algorithm, $\mathbf{a}^{\text{alg}}$, consists of 3 steps as follows:

(1) Determine the m by n matrix of the residuals, $\mathbf{R}^{\text{alg}} = \mathbf{D}^{\text{alg}} - \boldsymbol{D}^{\text{noise}}$.

(2) Construct an m by n matrix $\mathbf{Q}$ indicating which elements of $\mathbf{D}^{\text{spike}}$ are not contaminated by the simulated cosmic ray spikes, where $\mathbf{Q}_{i,j} = 1$ if $\mathbf{D}_{i,j}^{\text{spike}} = \mathbf{D}_{i,j}^{\text{noise}}$ and $\mathbf{Q}_{i,j} = 0$ otherwise.

(3) Calculate the 1 by n vector $\mathbf{a}^{\text{alg}}$ as follows: $\mathbf{a}_j = \sum_i^m \left(\mathbf{Q}_{i,j}\mathbf{R}_{i,j}^{\text{alg}}\right) / \sum_i^m \mathbf{Q}_{i,j}$.

The total residual spike count (Figure S3) is designed to evaluate how effective each algorithm was at suppressing the cosmic ray spikes. Building off the calculation of the average deviation above, the total residual spike count c^{alg} for an algorithm is: $c^{\text{alg}} = \sum_{i,j} \left|\left(1 - \mathbf{Q}_{i,j}\right)\mathbf{R}_{i,j}^{\text{alg}}\right|$. The second metric, spectral bias, is designed to evaluate the extent of the systematic errors each algorithm introduces, biasing the spectral average. Building off the calculation of the average deviation above, the spectral bias b^{alg} for an algorithm is: $b^{\text{alg}} = \sum_j \left|\mathbf{a}_j\right|$.

Multivariate Curve Resolution. MCR is a matrix decomposition algorithm which when used for chemometric analysis of hyperspectral images decomposes the hyperspectral data matrices into matrices of the concentrations and spectra. See Ruckebusch et al.[1] for a review of MCR. MCR analysis was performed on the results of the various despiking algorithms applied to

the Gaussian simulation. To assure that the reported results did not depend upon the specific MCR implementation employed, two different MCR algorithms were tested, Sandia National Laboratories' MCR algorithm[2] and MCR-ALS GUI 2.0.[3] The results from both algorithms were nearly identical, where Figure 4 and Table 1 correspond to the results from Sandia National Laboratories' MCR algorithm. While MCR algorithms solve for both the concentrations and spectra where neither is known *a priori*, the algorithms must be supplied with an initial guess for one of the two which can simply be a random matrix of the appropriate size. Here we initialized the MCR algorithms with the known pure component spectra. As such, any deviation of the output spectra from the pure component spectra cannot be due to trapping in a local minimum where the global minimum would correspond to the pure component spectra. Similarly, while MCR can suffer from rotational ambiguity,[1] initialization with the known spectra should present rotational ambiguity from arising. An additional spectral component constrained to correspond to a constant baseline offset was also included. All concentrations and spectral components were constrained to only non-negative values and the algorithms were allowed to iterate until stable results were achieved.

RESULTS AND DISCUSSION

Algorithm performance. Of the algorithms considered, Zhang-Henson proved to be one of the better performing algorithms. Depending upon the number of pure spectra with which it was provided, Zhang-Henson was almost as effective as UBS-DM-HS at suppressing both types of simulated cosmic ray spikes (Supplementary Figure S1). While neither UBS-DM nor UBS-DM-HS introduced much spectral bias, Zhang-Henson had even lower bias when all the pure spectra were input. While Zhang-Henson exhibited less overall bias, its bias is concentrated in a few wavelengths and for those wavelengths the bias is larger (Supplementary Figure S2).

(Thorough examination revealed that the bias introduced by these algorithms mainly arises from slight truncation of the upper extremes of the noise.) As such, the Zhang-Henson algorithm is an excellent choice when the pure spectra are known *a priori*, such as the pharmaceutical processing scenario for which it was designed.[4] However, when the intention is to perform exploratory chemometric analysis, such as multivariate curve resolution, to determine the spectra, UBS-DM-HS is more suitable. UBS-DM-HS provides better results than Zhang-Henson when the pure component spectra are not known *a priori* and hence are unavailable for input to Zhang-Henson (Supplementary Figure S3). The total residual spike count of the spiked image prior to despiking was 1.4×10^7 for both simulations. Therefore, even the worst performing algorithm reduced the total intensity of the cosmic ray spikes by >95% for the single point spikes and >75% for the Gaussian spikes.

Unsurprisingly, neither single-spectrum despiking algorithm was as effective since neither was designed to leverage the benefits of having multiple spectra. As is common knowledge, median-filtering introduced substantial spectral bias, where the amount of bias increased with the size of the median filter (Supplementary Figure S3). However, larger median-filters were better able to remove multi-point cosmic ray spikes (Supplementary Figure S3b). A single iteration of the Katsumoto-Ozaki algorithm proved capable of suppressing many cosmic ray spikes and was the only algorithm that did not introduce any spectral bias. While perfectly avoiding bias, unfortunately a single iteration of Katsumoto-Ozaki was not nearly as effective as UBS-DM-HS at removing cosmic ray spikes. For both point spikes and Gaussian spikes, the total residual spike count was nearly two orders of magnitude higher for a single iteration of Katsumoto-Ozaki than for UBS-DM-HS. While the Katsumoto-Ozaki paper[5] did not recommend multiple iterations, as an experiment we tested whether multiple iterations of the algorithm

would improve its performance. Unfortunately, while multiple iterations of the algorithm resulted in a significant reduction in cosmic ray spikes, multiple iterations of the algorithm introduced substantial spectral bias.

Supplementary Figure S1c makes clear that both UBS-DM-HS and Zhang-Henson nearly completely suppress all the multi-point cosmic ray spikes, with no spikes greater than 100 counts remaining. In contrast, all the other algorithms compared continue to display spikes on the order of thousands of counts (Supplementary Figure S1c and S1d). Supplementary Figure S2c shows that any spectral bias introduced by UBS-DM-HS is less significant than the uncertainty introduced by the readout noise, which introduces an average of 0.078 counts of spectral bias when the noise-free simulation is compared to the noisy simulation. In contrast, the median-filter (and experimentation with multiple iterations of Katsumoto-Ozaki) shows spectral bias above the level of the readout noise and the bias is most significant at regions where the true spectra have sharp peaks.

Interestingly, the UBS-DM algorithm is shown to produce an inferior scree plot to two iterations of the Katsumoto-Ozaki algorithm when processing Gaussian spikes (Supplementary Figure S4b). At first, this result is surprising since UBS-DM was shown (Supplementary Figure S3b) to leave fewer residual spike counts than two iterations of the Katsumoto-Ozaki algorithm. While at first counter-intuitive, the scree plot result makes sense after accounting for the distribution of the residual counts (see Supplementary Figure S2d). Where Katsumoto-Ozaki incompletely suppresses a large number of spikes, UBS-DM completely suppresses most spikes but the few remaining spikes are much larger, resulting in concentration of the residuals into fewer components.

Supplementary Figure S6 provides a qualitative comparison of the performance of the algorithms, using an experimental hyperspectral Raman image of a live cell. All algorithms substantially reduce the cosmic ray spikes to varying degrees. Note that the Zhang-Henson algorithm was designed to utilize the pure component spectra when available and may be expected to perform better in such cases. For this example, as is often the case when imaging cellular components, pure component spectra were unavailable. The algorithms which appear to have done the best job at removing the cosmic ray spikes in Supplementary Figure S6 are median filtering and the Katsumoto-Ozaki (2 iterations) and UBS-DM-HS algorithms, with 5-point median-filtering appearing to eliminate all spikes. However, in depth examination reveals that median-filtering severely truncated the narrower peaks in the spectrum, nearly eliminating some of them (see Supplementary Figure S6). While the 2 iterations of the Katsumoto-Ozaki algorithm introduced less distortion than median-filtering, it tended to reduce the intensity of the brightest portions of the spectra, presumably truncating the Poisson noise. Therefore, UBS-DM-HS provided the best despiking performance when the pure component spectra were not known *a priori*. UBS-DM-HS was able to handle both point spikes and more complex multi-pixel spikes without introducing noticeable distortion.

Computational Time. Comparing to the computational time of the other algorithms considered is less straightforward since the required time can depend both upon the level of optimization of the code and how the algorithm scales as the data size increases. As a rough guide, using our implementations it appears the Katsumoto-Ozaki algorithm was almost an order of magnitude faster than UBS-DM-HS, with median filtering another 5 times faster. Zhang-Henson was somewhat slower than UBS-DM-HS, taking roughly 50% longer. The relative speed of the algorithms did not change dramatically over a reasonable range of data sizes. All of the

algorithms were capable of processing thousands of spectra per minute, suitable for offline processing of hyperspectral data.

REFERENCES

1. Ruckebusch C and Blanchet L. Multivariate curve resolution: A review of advanced and tailored applications and challenges. *Anal Chim Acta*. 2013; 765: 28-36.
2. Haaland DM, Jones HDT, Van Benthem MH, et al. Hyperspectral Confocal Fluorescence Imaging: Exploring Alternative Multivariate Curve Resolution Approaches. *Appl Spectrosc*. 2009; 63: 271-9.
3. Felten J, Hall H, Jaumot J, Tauler R, de Juan A and Gorzsas A. Vibrational spectroscopic image analysis of biological material using multivariate curve resolution-alternating least squares (MCR-ALS). *Nat Protoc*. 2015; 10: 217-40.
4. Zhang L and Henson MJ. A practical algorithm to remove cosmic spikes in Raman imaging data for pharmaceutical applications. *Appl Spectrosc*. 2007; 61: 1015-20.
5. Katsumoto Y and Ozaki Y. Practical algorithm for reducing convex spike noises on a spectrum. *Appl Spectrosc*. 2003; 57: 317-22.

SUPPLEMENTARY FIGURE CAPTIONS

Figure S1. The analogue to Figure 1 except including three more algorithms. a) The residual differences between the ideal result and the results obtained from applying UBS-DM, UBS-DM-HS, and Zhang-Henson to spectra with simulated single-point cosmic ray spikes. For clarity, the UBS-DM residuals were shifted upwards and the Zhang-Henson residuals were shifted downwards by 200 counts. Within each algorithm, the results for 1024 simulated spectra are overlaid. Cosmic ray spikes which are incompletely eliminated manifest as large, positive residual values. Negative residuals generally result from truncation of experimental noise. The Zhang-Henson algorithm was supplied with all three pure spectra as initial inputs. b) As a), except showing the residual differences for five-point median filtering and two iterations of the Katsumoto-Ozaki algorithm (shifted upwards by 200 counts). Note the difference in scale. c) The residual differences between the ideal result and the results obtained from applying UBS-DM-HS and Zhang-Henson (shifted downwards by 200 counts) to spectra with simulated Gaussian cosmic ray spikes. Note the difference in scale. d) As c) except showing the residual differences for UBS-DM (shifted upwards by 6000 counts), five-point median filtering, and two iterations of the Katsumoto-Ozaki algorithm (shifted downwards by 6000 counts). Note the difference in scale.

Figure S2. The analogue to Figure 2 except including three more algorithms. a) The average deviation (see text) when applying UBS-DM, UBS-DM-HS, and Zhang-Henson to spectra with simulated single-point cosmic ray spikes. For clarity, the UBS-DM residuals were shifted

upwards and the Zhang-Henson residuals were shifted downwards by 0.05 counts. The Zhang-Henson algorithm was supplied with all three pure spectra as initial inputs. b) As a), except showing the average deviation for five-point median filtering and two iterations of the Katsumoto-Ozaki algorithm (shifted upwards by 15 counts). Note the difference in scale. c) The average deviation when applying UBS-DM (shifted upwards by 0.05 counts), UBS-DM-HS, and Zhang-Henson (shifted downwards by 0.05 counts) to spectra with simulated Gaussian cosmic ray spikes on the same scale as a). d) As c), except showing average deviation for five-point median filtering and two iterations of the Katsumoto-Ozaki algorithm (shifted upwards by 15 counts) on the same scale as b). b) and d) look virtually identical except under very close inspection, indicating that the bias introduced by either five-point median filtering or two iterations of the Katsumoto-Ozaki algorithm depends almost exclusively upon the underlying spectra and is largely unaffected by the cosmic ray spikes. In contrast, the average deviation for UBS-DM and Zhang-Henson varies substantially between the point-spike and Gaussian cosmic ray spike simulations.

Figure S3. Conceptually similar to Figure 3, Figure S3 provides a quantitative evaluation of the performance of the various algorithms for (a) single-point cosmic ray spikes and (b) Gaussian cosmic ray spikes spanning 4 points. For UBS-DM and UBS-DM-HS, only the final result is shown (see Figure 3 for all iterations). For Katsumoto-Ozaki, the solid symbol represents result of running one iteration of the algorithm except that the spectral bias was artificially increased from 0 to 0.1 to allow log-scale display. Symbols for successive iterations of Katsumoto-Ozaki are connected by line segments. The median filter was run with four different filter sizes, where each point is labeled with the filter size used. The Zhang-Henson algorithm performs better when

supplied with the pure spectra. The labels for the Zhang-Henson algorithm indicate how many of the three spectral components used in the simulations were supplied. An ideal algorithm would have zero residual spike counts and zero spectral bias.

Figure S4. Analogous to Figure 4a except including three more algorithms and showing the results for a) the single-point cosmic ray spike simulation and b) the Gaussian cosmic ray spike simulation. Incomplete suppression of cosmic ray spikes complicates chemometric analysis. Ideally, an abrupt transition should be seen between the values for the first *n* components and the remaining components, where *n* is the number of real spectra components (in this case 3). a) For the point-spike simulation, all algorithms achieve this other than Katsumoto-Ozaki (2 iterations). On this scale, the results for the UBS-DM and UBS-DM-HS algorithms are indistinguishable and overlap substantially with the result for the Zhang-Henson algorithm. b) For the Gaussian spike-simulation, a sharp transition is observed at the proper location only for the UBS-DM-HS and Zhang-Henson algorithms. On this scale, the scree plot for UBS-DM-HS, Zhang-Henson (3 input spectra), and the noisy image (not shown) are identical.

Figure S5. The reference spectra (without any cosmic ray spikes) used in the simulation.

Figure S6. The analogue to Figure 6 except including three more algorithms. a) The raw spectra from a 36 by 36 point hyperspectral Raman image of P388 cells were normalized to unit area and superimposed upon each other, where different colors correspond to different spectra. Dozens of

cosmic ray spikes are clearly visible. b) The results of the 5 different cosmic ray despiking algorithms evaluated in this paper are shown for the data in a), offset vertically for clarity. The order of the algorithms from top to bottom matches the order in the legend. The bands highlight cosmic ray spikes which were incompletely suppressed by the UBS-DM algorithm which the hyper-UBS algorithm either eliminated or reduced to near the noise-level of the spectrum. Note that while median-filtering appears to eliminate nearly all the cosmic ray spikes, it is also well-known to introduce undesirable spectral bias. Careful examination reveals that median-filtering also truncated real peaks at 59, 193, and 973 cm^{-1}.