A comparison of PCA and MAF for ToF-SIMS image interpretation

DOI:
10.1002/sia.3084

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Published in:
Surface and Interface Analysis

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher’s definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
A Comparison of PCA and MAF for ToF-SIMS Image Interpretation

Alex Henderson*, John S. Fletcher, John C. Vickerman
Surface Analysis Research Centre, Manchester Interdisciplinary Biocentre,
School of Chemical Engineering and Analytical Science,
The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK
*Corresponding author
Tel: +44 (0)161 306 {4462 | 4440 | 4544} Fax: +44 (0)870 762 2743
Email: {alex.henderson | john.fletcher | john.vickerman}@manchester.ac.uk

Abstract
While ToF-SIMS image generation has been useful in the determination of the spatial
distribution of chemistry in a broad application area, the amount of secondary ion signal
available in each pixel remains small, hampering the use of multivariate analysis (MVA)
approaches. The pre-treatment of data commonly comprises two approaches to increase pixel
signal intensity prior to multivariate analysis: mass channel summation and pixel summation.
Recent advances in instrumentation have lead to much greater signal per image pixel such that
a true spectrum can be discerned at each point across the sample. In this study we determine
the outcomes of these two pre-treatments prior to principal components analysis (PCA) and
maximum autocorrelation factor analysis (MAF) in comparison with high signal intensity data.
Both PCA and MAF analyses of the high signal intensity data are presented with MAF being
identified as the most effective approach. Image data was reduced in intensity to determine the
effectiveness of MVA with lower spectral intensity. MAF was found to outperform PCA on such
data, although both techniques were useful in the identification of the chemistry present. The
data were then mass summed, using a number of different approaches, to reduce mass
resolution, leading to a detrimental effect on PCA analysis, but little discernable change to MAF
output. Reducing the spatial resolution by summing spectra from adjacent pixels, however, lead
to a severe blurring of the MAF image structure, with PCA performing well.
Conclusions and recommendation for pre-treatment and MVA technique selection are
presented.

Keywords
ToF-SIMS; Secondary Ion Mass Spectrometry; Multivariate analysis; Principal Components
Analysis; PCA; Maximum Autocorrelation Factor Analysis; MAF; Image processing; HeLa cells;
Introduction

When generating SIMS images with ToF-SIMS instruments equipped with pulsed primary sources of atomic ions there is a limit to the amount of molecular information than can be generated from each pixel.\(^1\) The *static limit* imposes a cut-off in the number of chemically relevant ions that can be extracted from the small analysis area typically defining a pixel. Even with cluster ion sources, where the static limit is relaxed,\(^2,3,4\) the duty cycle of the instrument and/or the low primary ion current may impose a practical limit to the time taken to acquire a significant spectrum at each pixel. In order to maximise the analytical potential of a SIMS image multivariate analysis is often used, both to help in the identification of the chemistry present and also to increase the image contrast and thereby identify regions of interest. With the new instrumentation recently developed, incorporating both small spot cluster ion sources and quasi-continuous analysis, we can now generate images with a high secondary ion yield per pixel in a manageable timeframe.\(^5,6\) This allows us to consider the effectiveness of multivariate analysis strategies in the interpretation of ToF-SIMS images.

It has been shown that multivariate analysis methods, particularly PCA, can provide a useful tool for discriminating between ToF-SIMS data either for classification of proteins based on amino acid fragment variation, classification of tissue spectra from organs in mouse embryos\(^7\) or the identification of microorganisms.\(^8,9\) In the case of ToF-SIMS image analysis such analyses can be used to improve image contrast, to identify regions of interest and in each case to then output the associated variables (\(m/z\) values) for chemical identification.\(^10\)

Principal components analysis (PCA) has come to be extensively used in the analysis of collections of SIMS spectra and SIMS images.\(^11,12,13\) In PCA we seek to convert the many mass channels present in a collection of spectra into a smaller number of independent terms that express the chemical variation of the dataset more clearly. This is done through an analysis of the variance of each mass channel across the dataset, where the combination of mass channels containing the most variance is termed the first principal component (PC 1) and subsequent combinations, with decreasing variance, labelled with increasing number PC 2, PC 3 etc. up to PC \(n\) where \(n\) is the lower value of the number of mass channels (variables) submitted to analysis or the number of spectra (pixels) used. This is a correlation in the mass domain.

In an ideal chemical system the number of meaningful components is governed by the number of spectra from different chemicals present in the sample, with the remaining components forming a model of the noise present in the data. However, such an ideal system neglects non-linear contributions such as matrix effects or non-linear instrumental parameters such as
detector saturation. For model systems, the number of chemicals present in a sample is known \textit{a priori}. However, for analysis of real world samples, such as those encountered in the analysis of biological systems, the number of separate chemicals is unknown and the concentration of some is likely to be so low that their spectra appear as noise. One of the decisions that an analyst must make therefore is to determine the point at which the principal components no longer explain meaningful chemistry and begin to model noise in the system.

Maximum autocorrelation factor analysis (MAF) has been suggested to be useful in the analysis of SIMS images, however to date there has been little evidence of MAF in the literature.\textsuperscript{14,15} MAF analysis can be considered to be an extension of PCA applicable to images.\textsuperscript{16} MAF also attempts to identify the variance in mass channels corresponding to spectral contributions, but also considers the relationship between pixels in the image. During a MAF analysis a correlation is calculated between the image and the same image diagonally offset by one pixel. MAF is therefore specific to image analysis and identifies a correlation in the spatial domain as well as the mass domain.

\textbf{Experimental}

The ToF-SIMS data were acquired using an Ionoptika J105 \textit{3D Chemical Imager}. This instrument has been described in detail elsewhere.\textsuperscript{5} Briefly it is a ToF-SIMS instrument that allows the primary ion beam(s) to be operated in a DC mode by sampling the resulting secondary ion beam using a linear buncher. The mass spectrometry is thus decoupled from the sputtering event and the duty cycle of the instrument is greatly improved compared to conventional instruments that utilise pulsed ion beams. The instrument is fitted with 40 keV C\textsubscript{60}\textsuperscript{+} and Au\textsubscript{n}\textsuperscript{+} ion sources, incorporates sophisticated sample handling and also has the capability to perform MS/MS analysis.

The high duty cycle of the instrument, coupled to the properties of C\textsubscript{60} that allow collection of molecular information from the sample at primary ion doses far in excess of the traditional SIMS static limit, enables the generation of data that is ideal for the testing of multivariate methods for data interpretation and visualisation. The data set we have used is from a sample of HeLa cells that were entirely consumed during the SIMS analysis, thus the signal in each pixel is much higher than would be observed in a routine static SIMS image. The cells were cultured on poly-L-lysine coated silicon shards for 24 hours at 37 °C with 5% CO\textsubscript{2}. The cells were freeze dried and washed briefly in 0.15 M ammonium formate solution prior to ToF-SIMS analysis. Typical cell diameter is 20 µm. The image was acquired using 40 keV C\textsubscript{60}\textsuperscript{+} primary ions scanned over 128 × 128 pixels and the field of view of the image is 88 × 108 µm\textsuperscript{2} corresponding to a pixel dimension of ~0.7 × 0.8 µm\textsuperscript{2}. The total ion dose used was 3 × 10\textsuperscript{14} ions cm\textsuperscript{-2}. The total
sum of intensity across the 128 × 128 pixel image was 1.6 × 10⁸ counts corresponding to approximately 10⁴ counts per pixel.

**The test image**

Inspection of the data using conventional methods, i.e. viewing the localisation of signal associated with each peak in the mass spectrum, shows a number of distinct regions of chemical interest in the image. Firstly there is the signal arising from the substrate on which the cells were grown, dominated by silicon containing species. Secondly there is organic signal arising from the cells. Of this organic signal, some peaks are associated with the entire cell, some the periphery of the cell and some the centre of the cell. There is also an area that appears to be debris on the sample. If these features are obvious by simple inspection of the data they should appear within the first few principal components or factors. We have thus decided that the success criteria for our multivariate analysis of the data should clearly capture these features. Some of the peaks in the mass spectrum that localise to these different regions can be assigned to known and expected biological species along with signals characteristic of the silicon substrate. Figure 1 shows the total ion image along with some of the identified species, m/z 184 is a common fragment associated with biological samples as it arises from the phosphatidylcholine head group of phospholipids associated with biological membranes, m/z 136 is assigned to the [M+H]^+ ion of adenine and the species at m/z 112, localised to the substrate, is the signal from the Si₄⁺ ion. Of particular interest is the ability of the MVA to identify differences in chemistry within the individual cells. Such discrimination would be expected to be more difficult than the discrimination of cells from the substrate as the species arising from the cells will be predominantly organic with numerous isobaric interferences. Thus, as an extension of our success criteria, the optimisation of our MVA approach will be assessed by the appearance and position/rank of the principal component or factor that illustrates the differences between the outer cellular region (illustrated by the m/z 184 peak) and the inner cellular region (i.e. co-localised with the m/z 136).
Figure 1 ToF-SIMS images of HeLa cells (scaled and smoothed). Total ion image (a) and single ion images showing easily distinguished chemical variation across the sample; (b) m/z 136.1 [max 686 counts (55 after scaling)], (c) m/z 184.1 [max 699 counts (20 after scaling)]; and (d) m/z 111.9 [max 2067 counts (899 after scaling)]. These ions are assigned to adenine, phosphatidylcholine and silicon respectively.

Mathematical treatment

Computer equipment and software

Processing was undertaken using a PC equipped with two dual core Intel Xeon processors and 8 GB RAM running the Microsoft Windows XP64 operating system. Data was converted from the instrument manufacturer’s file format using in-house routines executed in 64-bit MATLAB (The MathWorks, Inc., Natick, Massachusetts). PCA and MAF algorithms were sourced from the website of Allan Aasbjerg Nielsen,\textsuperscript{17} modified to work with variables already in the MATLAB workspace. All other MATLAB routines were written in-house. While MATLAB is a closed software package, the operations described in this work are non-specific and should be compatible with open source mathematical packages.\textsuperscript{18}

MVA analyses

In order to improve calculation speed and to prevent divide-by-zero errors, variables (masses) that exhibited no variance (perhaps because the particular mass channel contained no data) were removed prior to analysis. In order to allow for comparison between the original spectra and the loadings plots, variables of zero value were inserted into the loadings matrix at the appropriate positions.
Prior to PCA analysis the data was Poisson scaled as detailed below. Since MAF requires a correlation between the image and offset versions of the same image, any scaling performed by way of pretreatment will be cancelled out during the analysis. Therefore, prior to MAF analysis, the data was not scaled. While mean centering is not strictly necessary for multivariate analysis of SIMS data,\textsuperscript{19,20} the algorithms employed here were preconfigured to perform mean centering. No peak selection was performed. All mass channels between 50 and 400 u were used in the analyses.

**Poisson scaling**

Prior to PCA analysis, scaling of the data was performed following the procedure outlined in the draft International Standard 18115\textsuperscript{21} based on work by Keenen and Kotula,\textsuperscript{19} itself developed from the work of Cochran and Horne.\textsuperscript{22} Briefly, data from ToF-SIMS instruments are collected in a pulse counted manner which is subject to an uncertainty explained by Poisson statistics. This uncertainty is equal to the mean of the signal intensity. Multivariate algorithms such as PCA and MAF are designed to account for the variance in the data. Since the data here will have variance related to the signal intensity (are heteroscedastic), the algorithms will perform sub-optimally. The approach of Keenan and Kotula, commonly called Poisson scaling, attempts to account for this by performing a scaling such that the estimated variance due to counting statistics is equal on all variables (intensity values at different masses) so multivariate analysis can better account for other sources of variance, for example those due to sample chemistry. The algorithm used by Keenen and Kotula assumes that row and column sums are preserved. This entails a scaling of each spectrum (row) independently, in addition to the scaling of intensity (column) to compensate for Poisson statistical variance. See Appendix A.

**Channel summation (mass binning)**

The use of channel summation, or mass binning, plays an important role in multivariate analysis. In the treatment of a collection of time-of-flight spectra, one use of such summation is to ‘digitally align’ spectra that may have been mass calibrated differently, perhaps due to sample topography or charging effects. Where two (or more) spectra are calibrated differently, the precise mass corresponding to the \( n \)th time channel in one spectrum will not be the same as that for the \( n \)th time channel in the second spectrum. Therefore, when the data are expressed as a matrix comprised of rows of spectra, any given column will contain intensities from different mass values. For a SIMS image, where the spectrum at each pixel has been acquired using the same instrumental parameters, the data is likely to be consistently mass calibrated. Here, however, bin-summation is performed for two reasons; firstly, in order to reduce the number of variables submitted to the multivariate algorithms. A typical SIMS spectrum contains anywhere between \( 10^5 \) and \( 10^6 \) mass channels. Typical MVA algorithms require this mass domain to be
reduced to a few thousand (or fewer) selected or summed channels as a practical limit for processing on personal computers. Secondly, many ToF-SIMS images are very sparse. That is to say they contain a proportionately large number of zero values. For example, a typical ToF-SIMS image may contain a total of $1 \times 10^8$ counts distributed over 65536 pixels. This gives an average of $\sim 1500$ counts distributed over the $1 \times 10^5$ mass channels making up the pixel’s spectrum. A second reason for bin-summation is therefore to circumvent the sparse nature of the data from inhibiting a meaningful analysis. We examined four mass binning strategies to determine their influence on the outcome of PCA and MAF analyses: 1 u, 0.5 u and 0.1 u all centred on nominal mass and 0.5 u with the bin edge aligned to nominal mass. More details on the exact alignment are given in Appendix B.

**Pixel summation**

Composite datasets were produced to investigate the impact of the pixel dimension with respect to sample feature size on the image contrast produced in the scores output. Images were re-sampled by summing the spectra in adjacent pixels in both the x- and y-directions. This produces a dataset containing one quarter of the original pixels. By summing together the spectra from neighbouring pixels we increase the signal to noise characteristics of the data. In order that the re-sampled image represented the original data more closely, the data was reduced in intensity prior to pixel summation. Here we divided the value of the intensity for each mass channel in each pixel by 4, rounding down to the next lowest integer value to better match the quantised characteristics of pulse counted data. This has the effect of removing any intensity falling below 1 count in the reduced spectrum.

**Resultant image scaling for display**

The colour scale used to display images should ideally capture the major variation in the intensity of signal at each pixel. Pixels with very high values (and in the case of scores plots, very low values) can mask changes in intensity elsewhere in the image. These pixels can be considered as outliers with respect to the rest of the data in the image. To avoid this we have scaled the scores (and secondary ion) data prior to an image being reproduced here. The statistical spread of the scores values was considered and the upper quartile ($q_U$), lower quartile ($q_L$) and interquartile range ($iqr = q_U - q_L$) of the scores data for each principal component/factor were calculated. Upper and lower limits were determined at $Limit_{upper} = q_U + (1.5 \cdot iqr)$ and $Limit_{lower} = q_L - (1.5 \cdot iqr)$ respectively. For each image, any value greater than $Limit_{upper}$ was set to that value and likewise any value lower than $Limit_{lower}$ was set to that lower limit. This scales the image contrast to cover the majority of the variance in the data and prevents outliers obscuring information.
Images were smoothed for display purposes using a one point cosine taper smoothing routine. Examples of the scaling and smoothing procedures are shown in Figure 2. The scaled and smoothed image (Figure 2d) indicates subcellular features most clearly. Scores images are shown with positive scores in green and negative scores in red.

**Figure 2** Total secondary ion image from 0.1 u mass binned data showing the difference in information content between scaled and unscaled data. (a) Unscaled [max: 39979 counts], (b) scaled [max: 24069 counts], (c) unscaled and smoothed; and (d) both scaled and smoothed.

**Results and discussion**

**Comparison criteria**

As discussed above we define an improvement in an analysis when the inner and outer compartments of the cellular structure are well defined. This *cell marker*, shown in Figures 3b and 3c, should appear as a low principal component or autocorrelation factor indicating a high degree of anticorrelation between these two states. Good contrast is determined visually where well defined separation of species is apparent.
Figure 3 Scaled, smoothed images of HeLa cells showing (a) the total ion image, (b) PC 6, and (c) MAF 4. MAF captures the distinction between inner and outer cellular regions more clearly than PCA. Comparison with the total ion image shows the additional information available following multivariate analysis.

Comparison between PCA and MAF

Data binned to 0.1 u was analysed by both PCA and MAF. Poisson scaling improved the PCA analysis in comparison to unscaled data. The cell marker component appeared spread between PCs 8 and 9, but appeared at PC 6 following Poisson scaling. Following this result, all further PCA analyses were performed on Poisson scaled data. The scores images for PCA of Poisson scaled data are shown in Figure 6a. MAF analysis indicated the cell marker factor at MAF 4, improving further on the PCA result. The contrast in MAF 4 is also more distinct than in the PCA result as can be seen in Figure 3. The scores images produced by the MAF analysis are shown in Figure 7a. Comparison of the loadings produced by PC 6 and MAF 4, shown in Figure 4, indicate a high level of similarity in the spectral features identified with positive and negative loading. The PCA scores in general exhibit more noise making the MAF easier to interpret. The five most intense positively loaded spectral features (green in scores images) for both PCA and MAF were selected for use in a fingerprint library search. In both cases 3 of these 5 features generated hits for phosphatidylcholine containing molecules. A similar search for the 5 most intense negatively loaded spectral features (red in scores images) generated hits at the 3 from 5 level for deoxyadenosine, deoxyguanosine and spectra from proteins such as fibronectin, thus reinforcing the manually identified chemistry.

Figure 4 Loadings plots from (a) PC 6 and (b) MAF 4 showing the similarity in mass peaks identified and the difference in noise level.

Intensity reduction

Many SIMS images exhibit low secondary ion signal per pixel. Here we address this by determining how the result from an average of $10^4$ secondary ions per pixel is modified by a reduction in intensity. The data, binned to 0.1 u, was reduced in intensity by dividing by factors
of 4 and 10. When dividing by a scalar value we must be careful to preserve the characteristics of the noise. Poisson noise is a function of peak intensity and so dividing by a constant will be inconsistent peak-to-peak. Here we select to Poisson scale the data prior to the division. Each resultant intensity value is then rounded down to the nearest lower integer to simulate the quantised characteristics of data with low signal intensity. This means that reduced signal falling below 1 count was removed entirely. The average secondary ion yield per pixel is shown in Table 1.

It is important to note that this scaling followed by integer rounding is not the same as reducing the primary ion beam current or other instrumental parameters to reduce the measured intensity by the same scalar. The noise structure is very different. It is included here to illustrate different computational effects that could occur during data processing.

### Table 1

<table>
<thead>
<tr>
<th></th>
<th>Original data</th>
<th>Poisson scaled data</th>
<th>Poisson scaled intensity / 4</th>
<th>Poisson scaled intensity / 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average SI per pixel</td>
<td>9990</td>
<td>2050</td>
<td>350</td>
<td>100</td>
</tr>
</tbody>
</table>

The reduction in available intensity information has a marked effect on the PCA analysis. Reducing by a factor 4 moved the cell marker component from PC 6 to a combination of PCs 11 and 12. Further reduction lead to the marker component falling outside the first 20 principal components; unlikely to be noticed in a typical analysis. Gross changes in chemistry, for example the separation between organic and inorganic species remained visible in PCs 1–5. The MAF analysis fared much better in that the reduction in intensity by factors of 4 and 10 had little effect on the result; the cell marker information still appearing in MAF 4.

The change in the amount of variance explained by each principal component and autocorrelation factor, for each of the intensity reduced datasets, is shown in Figures 5a and 5b respectively. For PCA, the amount of variance explained by the first PC varies inversely with the amount of data remaining in the dataset. The remaining PCs are not affected by the data reduction. For MAF, there is almost no change in the amount of variance explained by a given PC for reductions in the data up to a factor of 10. This indicates that even with poor signal...
intensity there is enough information in the fragmentation pattern to indicate chemistry.

**Figure 5** Amount of variance explained by each principal component (a) and autocorrelation factor (b) expressed on a log 10 scale. Reducing the intensity up to a factor of 10 shows little effect on MAF data, but PCA varies worse.

**Mass binning**

In order to determine the degree to which mass resolution is involved in the multivariate analysis of images, the data was mass binned to generate new spectra with bin widths of 1 u, 0.5 u and 0.1 u, centred on the nominal mass, and 0.5 u with the bin edge on the nominal mass (centre-split). Both PCA and MAF were performed on these mass resolution reduced data sets; the data being Poisson scaled prior to PCA analysis.

The result from MAF analysis is only slightly influenced by the reduction in mass bin size. For data centred on the nominal mass, the cell marker moved from MAF 3 to MAF 4 as the bin width was reduced from 1 u to 0.1 u. The 0.5 u centred and 0.5 u centre-split showed an inversion between MAF 3 and MAF 4 indicating that the centre-split data will separate any inorganic contribution differently.

For PCA the treatment introduced larger variation. The lack of spatial correlation requires the PCA algorithm to be more reliant on mass resolution. Here the 0.5 u centred data separates the lipid signal into two components, while the 0.5 centre-split data separates the organic from inorganic signals into two components before identifying the cellular material. The 0.1 u data separates the (in)organic signal into four components before, again, identifying the cellular material in PC 6. The indication here is that there is a difference between the two 0.5 u binned datasets in that the inorganic component is enhanced in the centre-split data. The 0.1 u data combines these differences.

**Spatial reduction**

When confronted by low signal intensity in a SIMS image, many analysts would consider combining pixels to produce a lower contrast image with more secondary ion signal per pixel, prior to multivariate analysis. Here we consider the outcome of such a procedure.
Reducing the intensity of a dataset by a factor of 4 and then reducing the spatial resolution by a factor of two (four times fewer pixels) will produce a similar dataset to the original in terms of ions per pixel. Since the reduction in spectral intensity includes a rounding down of any fractional counts there will also be a removal of noise with the removal of reduced spectral intensity < 1 count. The average ion count per pixel is given in Table 2.

### Table 2 Average secondary ion intensity, approximate dimension and percentage of cell sampled per image pixel where both the spectral intensity and spatial resolution have been reduced.

<table>
<thead>
<tr>
<th></th>
<th>Original data</th>
<th>Intensity / 4, pixel number reduced by 4</th>
<th>Intensity / 16, pixel number reduced by 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average SI per pixel</td>
<td>9990</td>
<td>9348</td>
<td>7714</td>
</tr>
<tr>
<td>Pixel dimension</td>
<td>~0.7 × 0.8 µm²</td>
<td>~1.4 × 1.7 µm²</td>
<td>~2.8 × 3.4 µm²</td>
</tr>
<tr>
<td>Percentage of cell taken up by pixel</td>
<td>~0.2%</td>
<td>~0.8%</td>
<td>~3.0%</td>
</tr>
</tbody>
</table>

Note that there is no loss of spectral mass resolution as it is only the intensity and spatial distribution of the data that are being modified. Therefore it is expected that the PCA analysis, which relies solely on mass resolution will outperform MAF analysis where spatial resolution is important. This assumption is partially borne out in the analysis. The PCA outcome shown in Figure 6b contains the cellular information, but this is now spread out over at least 3 principal components – PCs 3, 4 and 7. Visual contrast is enhanced since we now have a single pixel representing the chemistry of 4 original pixels.

The MAF analysis produces loadings (Figure 8) that appear to be meaningful in that they have the form of a SIMS spectrum rather than uniformly distributed noise, but the spatial distribution of the scores as shown in Figure 7b is difficult to align with the original data in Figure 7a. Being presented with the scores images in Figure 7b, without prior knowledge such as the MAF results from data that had not been pixel reduced (Figure 7a) would make it difficult for an analyst to draw any conclusions from the analysis.

A further reduction in spatial resolution and intensity was performed by dividing the intensity by 16, rounding down as before, and reducing the data to a 32 × 32 pixel image — a 16-fold decrease in spatial resolution. Here PCA still gave some contrast relating to the inorganic
components now split across PC 2 and PC 3 and the cell structure across PCs 4–6. The MAF algorithm was unable to generate a result, reporting a singular matrix error. Although the secondary ion signal per pixel is similar in the original and reduced cases, the act of summing the spectra from adjacent pixels has markedly changed the PCA and MAF analyses. One explanation as to why the MAF has performed so poorly here is that the percentage of cellular material that is represented by a pixel becomes a larger proportion of the cell. This, coupled with the difference in shape between a square pixel and a rounded cell, leads to a combination of the chemistry in adjacent pixels. Since MAF takes the chemistry in adjacent pixels into account, their similarity will lead to an increased autocorrelation and result in a blurring of the scores image.

**Figure 6** Scores images from PCA of the 0.1 u data, both in the original intensity and 128 × 128 spatial resolution (left ≡ a) and where the data was divided by 4 and spatially reduced to 64 × 64 pixels (right ≡ b). Principal component numbers are given in each image.
Figure 7 Scores images from MAF of the 0.1 u data, both in the original intensity and 128 \times 128 spatial resolution (left \equiv a) and where the data was divided by 4 and spatially reduced to 64 \times 64 pixels (right \equiv b). Autocorrelation factor numbers are given in each image.

Figure 8 Loadings on MAF where both the spectra and the number of pixels have been reduced in intensity by a factor 4 showing that the loadings may still identify chemistry even though the spatial location may be difficult to discern. Autocorrelation factor numbers are indicated.

Conclusion and recommendations

We have seen that when there is sufficient spectral intensity at each image pixel, multivariate analysis can greatly assist the SIMS analyst with data interpretation. However when there is a limited spectral intensity, the steps taken by the analyst in pre-processing the data before multivariate analysis may have a significant impact on the information returned. Maximum autocorrelation factor analysis outperforms principal components analysis where the spatial distribution of chemistry is large with respect to the pixel dimension.

In order to increase spectral intensity, summing mass channels together is preferred to summing pixels, particularly if MAF analysis is planned. The inorganic contribution to a peak will clearly fall into the same bin as the organic contribution when the data is binned to 1 u, and yet the first principal component or autocorrelation factor separates the inorganic substrate from the organic cellular material well. Therefore it is the fragmentation pattern that is the dominant factor in multivariate analysis rather than mass resolution. If a separation between inorganic and organic components is expected, then mass binning using a centre-split approach with the bin edge aligned with nominal mass may be beneficial.

Our recommendations therefore are as follows:

1. Use MAF rather than PCA for images with low spectral intensity.
2. Sum mass channels together to improve multivariate analysis.
3. To separate inorganic material from organic, mass binning using a centre-split approach, with the bin edge aligned with nominal mass, may be beneficial.
4. Summing pixels together prior to MAF analysis may give a misleading result.
5. Image scaling and smoothing can assist in visually identifying regions of interest.
6. Poisson scaling prior to PCA can improve the outcome.

**Acknowledgements**

We thank Allan Aasbjerg Nielsen for making his PCA and MAF algorithms openly available. We thank Dean Jackson and Sadia Rabbani of the University of Manchester for providing and culturing the HeLa cells for analysis. This research was supported by a grant from the UK Engineering and Physical Sciences Research Council (Grant number EP/C008251/1).

**Appendix A**

**Data structure**

In this text an image is defined as a 3D data ‘cube’ with each x-y plane describing the spatial distribution of a particular mass. This 2D matrix can be ‘unfolded’ to produce a 2D array of spectra where each row represents a spectrum from a particular pixel. The rows are aligned such that each column represents the intensity values for any given mass. In order to convert between an image and the matrix of the spectra contained within it, we can use the following MATLAB code:

```matlab
[xpixels, ypixels, massvalues] = size(image);
spectra = reshape(image, xpixels * ypixels, massvalues);
```

In order to reconstruct an image from scores values obtained through PCA or MAF analysis we can use the following MATLAB code:

```matlab
scoresimage = reshape(scores, xpixels, ypixels, massvalues);
```

**Poisson scaling**

Consider a spectral data matrix \( X = m \times n \) with \( m \) spectral rows and \( n \) mass values. The matrix can be Poisson scaled to produce the output variable \( scaler \) using the following MATLAB code:

```matlab
a = sqrt(mean(X,2));
b = sqrt(mean(X));
scaler = a*b;
scaledX = X ./ scaler;
```

Where: \( a \) is the square root of the average pixel; \( b \) is the square root of the average spectrum; \( scaler \) is the outer product of \( a \) and \( b \) and has the same dimensions as \( X \); and \( scaledX \) is the Poisson scaled version of \( X \) where each value is divided by the equivalent value in \( scaler \).

**Appendix B**

Here we define the four mass bin strategies used to align the data prior to analysis.
1 u centred
The channels in each spectrum were summed together such that all channels falling within
± 0.5 u of a nominal mass were assigned to that nominal mass. $m_{i,0} = \sum (m_i \pm 0.5u)$ where $m_i$ is
initially set to the first integer mass value and subsequent masses are incremented by 1.0 u.
Example output: mass bin centres 75 u, 76 u, 77 u etc. See Figure 9a.

0.5 u centred
The channels in each spectrum were summed together such that all channels falling within
± 0.5 u of a nominal mass or half-nominal mass were assigned to that mass value. This
produces narrower mass bins than the 1 u centred data, with additional mass bins ‘between’ the
peaks. $m_{0.5c} = \sum (m_i \pm 0.25u)$ where $m_i$ is initially set to the first integer mass value and
subsequent masses are incremented by 0.5 u. Example output: mass bin centres 75.0 u,
75.5 u, 76.0 u etc. See Figure 9b.

0.5 u centre-split
The channels in each spectrum were summed together such that all channels falling within
± 0.25 u of a nominal mass + 0.25 u, were assigned to that mass value. This produces new
mass bins where the bin edge lies on a nominal mass. The advantage of such a procedure is
that spectral features that are mass deficient (typically inorganic species) will fall into a separate
mass bin to those that are mass sufficient (typically organic species). $m_{0.5s} = \sum (m_i \pm 0.25u)$
where $m_i$ is initially set to 0.25 u above the first integer mass value and subsequent masses
are incremented by 0.5 u. Example output: mass bin centres 75.25 u, 75.75 u, 76.25 u etc. See
Figure 9c.

0.1 u centred
The channels in each spectrum were summed together such that all channels falling within
± 0.05 u of a nominal mass + 0.1 u, were assigned to that mass value. $m_{0.1} = \sum (m_i \pm 0.05u)$
where $m_i$ is initially set to the first integer mass value (nominal mass) and subsequent masses
are incremented by 0.1 u. Example output: mass bin centres 75.0 u, 75.1 u, 75.2 u etc. See
Figure 9d.
Figure 9 Location and width of mass bins. a) 1 u centred, b) 0.5 u centred, c) 0.5 u centre-split, d) 0.1 u centred.

References

18 See for example, Octave (http://www.gnu.org/software/octave/), Scilab (http://www.scilab.org/) or FreeMat (http://freemat.sourceforge.net/).
21 J.L.S. Lee, I.S. Gilmore, M.P. Seah, Draft definition of terms for ISO 18115 Surface Chemical Analysis – Vocabulary. 8 September, 2008
23 Downloaded from MATLAB Central File Exchange (http://www.mathworks.com/matlabcentral/fileexchange/7951)
24 The inverse of the MAF loadings must be taken before comparison with PCA loadings is possible.