Multivariate analysis of ToF-SIMS images after variable selection

Seetharaman Vaidyanathan[1],
[1] ChELSI, Department of Chemical & Process Engineering, Mappin Street, Sheffield S1 3JD, UK.
S.Vaidyanathan@sheffield.ac.uk

Abstract. Variable selection was explored as a perturbation tool prior to multivariate analysis to extract peaks of relevance from Time-of-flight secondary ion mass spectral images. The information modelled by the multivariate analytical tools used (PCA and MFA) before and after variable selection did not show sufficient difference to allow novel features to be extracted. The scores images generated from regions of the spectrum that did not have a readily recognisable contribution to the total ion image showed similarities to those derived from the more readily recognisable regions of the spectrum. This suggests that sample topography, which is common to both sets of data is being modelled, and needs to be accounted for before extraction of chemically relevant information.

Keywords: Principal component analysis, maximum autocorrelation factor analysis, bacteria, mass spectrometry, image analysis.

1 Introduction

Time of flight secondary ion mass spectrometry (ToF-SIMS) is a surface technique that employs a primary ion beam to desorb and ionise secondary ions from surfaces that can be detected by time-of-flight mass spectrometry (9). The technique enables measurement and monitoring of chemical changes associated with the surface analysed. The advent of polyatomic ion sources has enabled more efficient detection of biochemical species from biological surfaces making it possible to apply the technique to study the associated lateral and depth-wise distribution of (bio)chemical species in such systems (11). Typically, a ToF-SIMS image consists of information at several hundreds or thousands of variables (peaks) at each pixel, many of which can arise from the same chemical species. The image dataset can be expressed as n m x n x p matrix of data that consists of m x n pixels, each with p data points. The application of multivariate analytical tools to such datasets can help in the extraction of useful information.

It has been shown that multivariate analytical tools such as principal component analysis (PCA) and maximum autocorrelation factors (MAF) analysis can be used to improve image contrast and to identify regions of interest within the images, so the associated variables can be extracted and their distribution studied after appropriate

chemical identifications (1, 5, 6, 8). There is however a challenge in understanding the outputs from such analyses and interpreting the changes modelled.

Image data that are scaled have shown to give better performances using PCA (6). This is especially true of datasets with dominating peaks. Unlike PCA, MAF is scaling independent and applied as such without data pre-processing. Whilst PCA has been extensively used in image data analysis, MAF has been shown to perform better (1, 6). The performance and interpretation of the data can be dataset dependent. In this regard selection of variables to analyse can influence the results. In datasets where preliminary information on variables is available it should be possible to use the information to screen for hidden structures in the datset. In this investigation, the effect of removing selected variables on the multivariate analysis of the image and its influence on the resulting scores for the analysis of ToF-SIMS images generated from *Streptomyces coelicolor* (a filamentous bacterium) population on silicon is discussed. Both PCA and MAF were examined before and after removing selected variables

2 Experimental

Positive ion ToF-SIMS images of *Streptomyces coelicolar* cell population dried on a silicon chips were acquired as detailed elsewhere (7). A BioTof instrument equipped with a wien filtered 40 keV C_{60}^+ primary ion source was used for the data acquisition (10).

The image data analysis was carried out in MATLAB (The MathWorks, Natick, MA, USA), using locally written routines, unless mentioned otherwise. The image data was binned to 1 amu mass resolution and converted to Ascii format prior to MATLAB analysis. PCA and MAF were carried out using the routines of Nielsen (4). For both analysis the option of having the eigenvectors to be unit vectors was chosen. Only the data in the mass range m/z 20–80, 100–130, 360–400 and 515–575 was combined and considered for analysis. This was done as most of the relevant information pertaining to the data occur in this mass range and the reduced mass range helped reduce computational memory for data handling for analysis. Selected variables that correspond to known chemical species were removed from the image data one at a time before PCA and MAF analysis and this was compared with the scores generated prior to removal. The corresponding loadings were also studied to examine the changes. At a resolution of 1 amu mass unit, the dataset modelled consisted between 180-200 variables (peaks).

3 Results and Discussion

The total ion ToF-SIMS spectrum of the acquired area (Figure 1) shows peaks that can be associated with Na^+ (*m/z* 23), K^+ (*m/z* 39), organic fragments (*m/z* 41, 43, 57), and amino acid fragments (possibly arising from proteins) (*m/z* 70, 84, 86, 110, 120). In addition, two prominent peaks at *m/z* 368 and 394, corresponding to two of the antibiotics the microorganism produces can also be identified, and the peaks at around *m/z* 500-575. The distribution of these signals over the analysed area can be seen in

the respective ion images (each pixel in the image consists of a mass spectrum). Clearly, differences in the distribution of the ions can be identified. The two antibiotics that the microorganism produces can be seen to have differences in their localisation. The ion images at the different *m/z* can show their distributions. However, given the presence of other variables in the spectral information, the application of multivariate analysis of the images can help in exploring the data to seek information about non-obvious and/or co-localised signals . PCA and MAF are two multivariate analytical tools that have been used to explore image datasets generated by ToF-SIMS. PCA relies on the variance in the dataset to generate linear combinations of these in an orthogonal fashion, whilst in MAF analysis, which can be considered to be an extension of PCA, in addition to the variance in the dataset the relationship between adjacent pixels in the image is also considered (3).

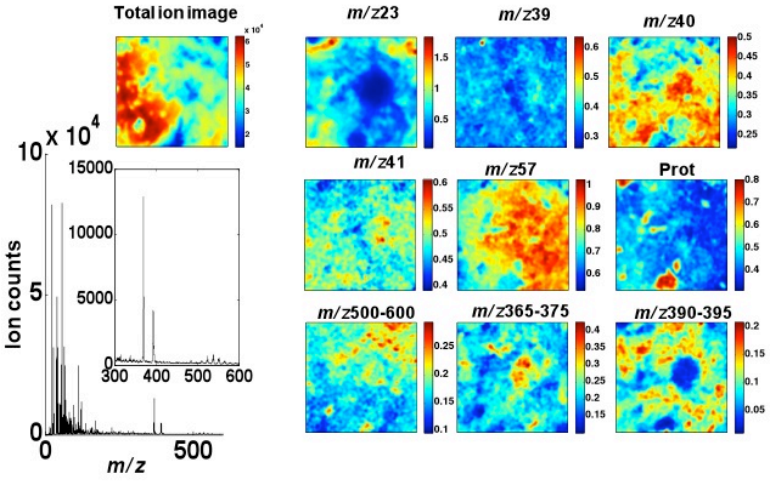

Fig. 1. ToF-SIMS ion images (128 x 128 pixels) in the positive mode derived from *S. coelicolor* filaments on silicon. The analysis field of view is 250 x 320 microns. The total ion image is shown in the top corner, with the corresponding total ion spectrum below. Ion images at selected *m/z* values are shown alongside.

Perturbations introduced into the analysis can highlight changes that can be used to explore the information content of the images. To this effect, variable selection was applied to the image dataset as a perturbation tool to study the resulting changes and assist in interpretations. The variables selectively removed were the prominent sodium signal at *m/z* 23, the organic signal at *m/z* 57 that has a prominent distribution, and the two antibiotic signals at *m/z* 368 and 394, one at a time. An analysis of information content in the parts of the spectrum that were originally discounted (namely *m/z* 80-100, 131-360, 400-515, and 575-600) was also carried out. The data set after the variable selection was subsequently analysed using PCA and MAF. The resultant scores for the first three PCs and MAFs are shown in Figures 2, 3 and 4.

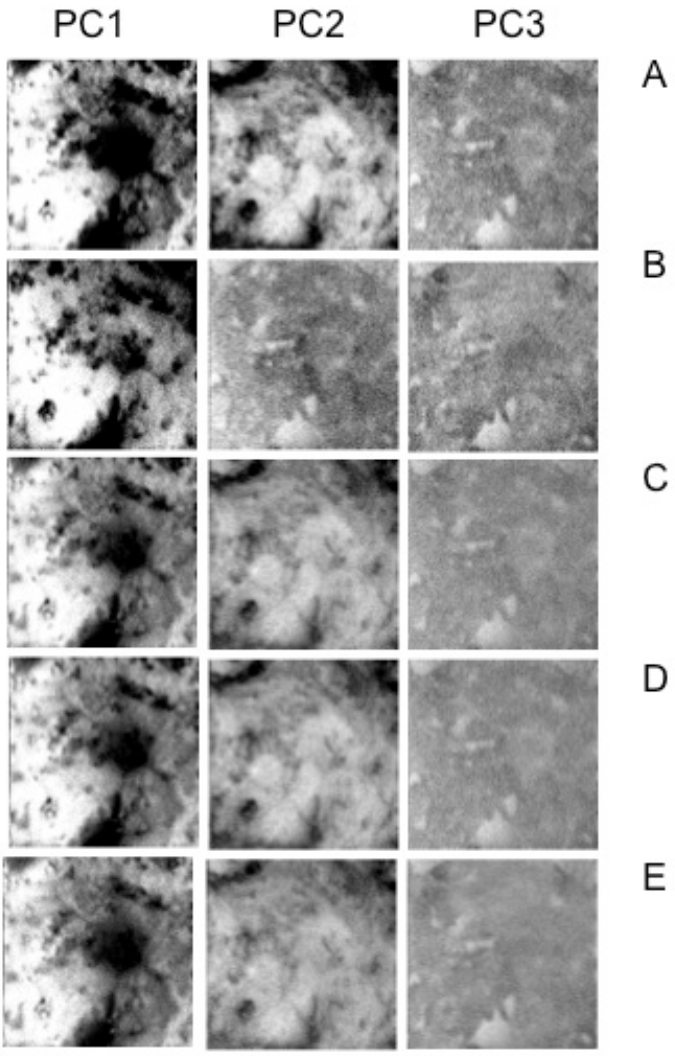

Fig. 1. PCA pseudo-image of scores after PCA on selected *m/z* windows. The first 3 PCs are shown for five different windows: (A) all peaks mentioned in the Experimental section analysed, (B) *m/z* 20-24 excluded (to exclude the Na^+ signal), (C) *m/z* 366-374 excluded (to exclude *m/z* 368 antibiotic signal), (D) *m/z* 391-395 excluded (to exclude *m/z* 394 antibiotic signal), (E) *m/z* 57 excluded. The images are 256 x 256 pixels with a pseudo grey colour code, ranging from -30 to -30 arbitrary units.

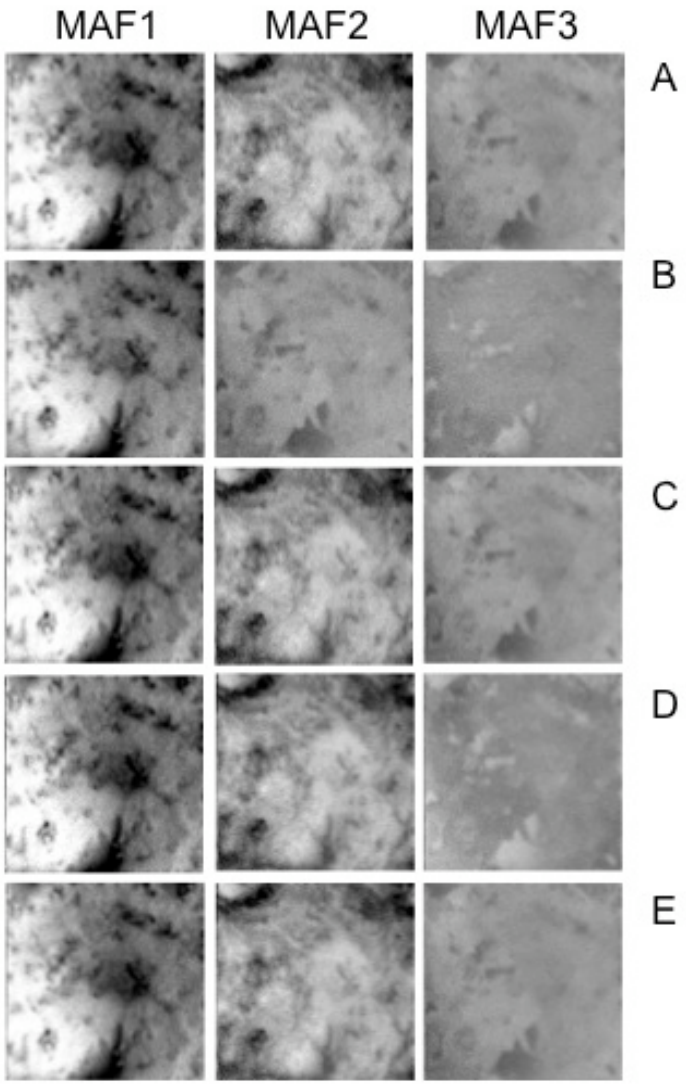

Fig. 2. MAF pseudo-image of scores after MAF on selected *m/z* windows. The first 3 MAFs are shown for five different windows: (A) all peaks mentioned in the Experimental section analysed, (B) *m/z* 20-24 excluded (to exclude the Na^+ signal), (C) *m/z* 366-374 excluded (to exclude *m/z* 368 antibiotic signal), (D) *m/z* 391-395 excluded (to exclude *m/z* 394 antibiotic signal), (E) *m/z* 57 excluded. The images are 256 x 256 pixels with a pseudo grey colour code, ranging from -10 to -10 arbitrary units.

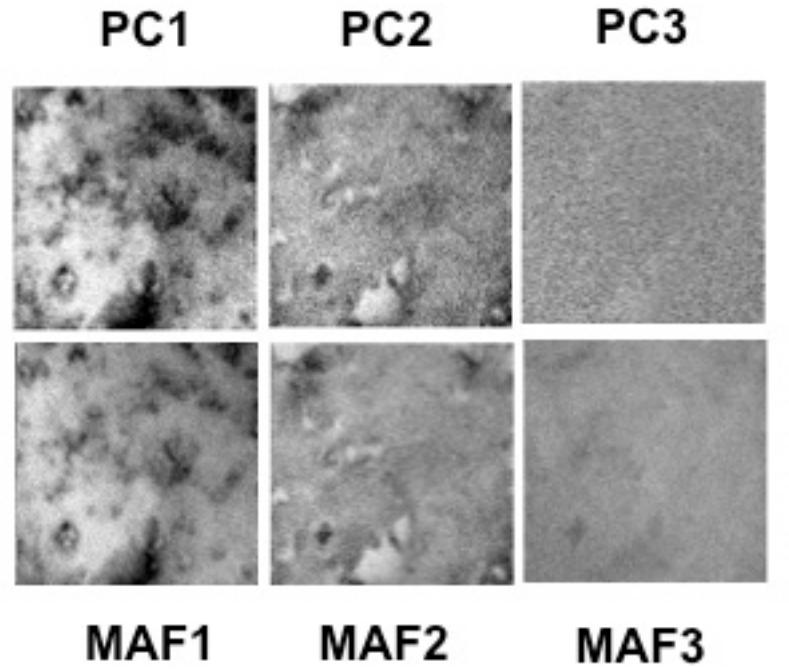

Fig. 3. PCA and MAF scores for masses not included in the ranges in figures 2 and 3 (*m/z* 80-100, 130-360, 400-515, 575-600). The images are 256 x 256 pixels with a pseudo grey colour code, ranging from -3 to -3 arbitrary units.

With both PCA and MAF analysis little difference is observed in the first PC or MAF, but differences can be seen in the second and third PCs/MAFs for the removal of the sodium peak at *m/z* 23 (Figures 2 and 3). An inspection of the loadings (data not shown) showed that the information already identified in Figure 1 was predominantly modelled. The contributions of the rest of the signals were masked by these peaks. With both PCA and MAF analysis, removal of the prominent signals that are known to contribute to the information content still results in similar scores images (Figure 4). Although there may be contributions from fragment ions that have similar origins as those selected in the analysis for Figures 2 and 3, the results in Figure 4 suggests that a more common information is being modelled.

Given that the bacterial sample is deposited on silicon and that the bacterial cell mass is a few microns (about 1-5 microns) thick, it is very likely topography of the analysed surface contributes predominantly to the information modelled. Although subsequent PCs and MAFs do contain information that is relevant to the chemical makeup these are largely masked by sample topography. Clearly, there is a requirement to account for sample topography before it is possible to extract chemically relevant information from the images, using multivariate analysis. Keenan and Kotula (2) have suggested accounting for poisson noise prior to multivariate analysis to enhance the applicability of the later to image analysis. Various scaling

options have also been shown to improve analysis, especially with PCA (6). In this investigation, variable selection in itself was not sufficient to tease out the required chemical information, even with scaling independent MAF analysis of the data. It is possible that the application of Poisson correction and accounting for sample topography prior to multivariate analysis will enable variable selection to be of use in teasing out further useful information from TF_SIMS images. This is currently being explored.

4 Conclusions

PCA and MFA analysis of ToF_SIMS images (generated from a bacterial population on silicon) were carried out before and after variable selection and analysed to look for novel extractable features. The resultant scores images were similar and did not show sufficient differences to enable novel features to be extracted. Removal of the recognisable regions of the mass spectrum from the analysis still resulted in similar scores images, suggesting that sample topography, a common feature of both the data sets dominates the information modelled. Data pre-processing to account for sample topography will be needed and is likely to enable variable selection as a useful tool in the multivariate analysis of ToF-SIMS images.

Acknowledgments. The author is grateful to Prof. John Vickerman (University of Manchester) and group, especially with Dr. John Fletcher and Dr. Alex Henderson for help with the ToF-SIMS analysis. Dr. Jason Micklefield (University of Manchester) and his group is also thanked for help with the growth of the bacterial sample. The author is also thankful to members of ChELSI at the University of Sheffield for the support. ChELSI is funded by the EPSRC, UK.

References

1. Henderson, A., J. S. Fletcher, and J. C. Vickerman. A comparison of PCA and MAF for ToF-SIMS image interpretation. Surface and Interface Analysis 41:666-674 (2009).
2. Keenan, M. R., and P. G. Kotula. Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images. Surface and Interface Analysis 36:203-212. (2004).
3. Larsen, R. Decomposition using maximum autocorrelation factors. Journal of Chemometrics 16:427-435. (2002).
4. Nielsen, A. A. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3620. (2009).
5. Park, J. W., H. Min, Y. P. Kim, H. K. Shon, J. Kim, D. W. Moon, and T. G. Lee. Multivariate analysis of ToF-SIMS data for biological applications. Surface and Interface Analysis 41:694-703 (2009).
6. Tyler, B. J., G. Rayal, and D. G. Castner. Multivariate analysis strategies for processing ToF-SIMS images of biomaterials. Biomaterials 28:2412-2423 (2007).

7. Vaidyanathan, S., J. S. Fletcher, R. Goodacre, N. P. Lockyer, J. Micklefield, and J. C. Vickerman. Subsurface biomolecular imaging of Streptomyces coelicolor using secondary ion mass spectrometry. Anal Chem 80:1942-51 (2008).
8. Vaidyanathan, S., J. S. Fletcher, A. Henderson, N. P. Lockyer, and J. C. Vickerman. Exploratory analysis of TOF-SIMS data from biological surfaces. Applied Surface Science 255:1599-1602 (2008).
9. Vickerman, J. C., and D. Briggs. ToF-SIMS: Surface Analysis by Mass Spectrometry. IM Publications, Chichester and Surface Spectra, Manchester, UK (2001).
10. Weibel, D., S. Wong, N. Lockyer, P. Blenkinsopp, R. Hill, and J. C. Vickerman. A C-60 primary ion beam system for time of flight secondary ion mass spectrometry: Its development and secondary ion yield characteristics. Analytical Chemistry 75:1754-1764. (2003).
11. Winograd, N. The magic of cluster SIMS. Analytical Chemistry 77:142a-149a (2005).